ICRA9
9th International Conference on Risk Analysis

BOOK OF ABSTRACTS

University of Perugia

Perugia, May 25-27, 2022

Book of Abstracts

9th International Conference
on
Risk Analysis

Editors:

Christos Kitsos
Teresa A. Oliveira
Francesca Pierri
Marialuisa Restaino

Perugia, 25-27 May, 2022

9th International Conference on Risk Analysis

Perugia, 25-27 May, 2022

Book of Abstracts

Editors: Christos Kitsos, Teresa A. Oliveira, Francesca Pierri, Marialuisa Restaino

Conference Website: http://icra9.unipg.it/

Conference E-mail: icra9@unipg.it

ISBN: 978-972-674-919-6

Preface

In principle, Risk is defined as exposure to the chance of injury or loss. Practically it is a hazard or dangerous chance and is wondering about the probability that something unpleasant will take place. Therefore the probability of damage, caused by external or internal factors, has to be evaluated. The essential factors that influence the increment of the Risk are asked to be determined. That is why eventually we are referring to Relative Risk (RR). In epidemiological studies, it is needed to identify and quantitatively assess the susceptibility of a portion of the population to specific risk factors. It is assumed that all the individuals have been equally exposed to the same possible hazardous factors. The difference, at the early stage of the research study, is only due to a particular factor that acts as a susceptibility.

Under this line of thought, we started the ICCRA (International Conference on Cancer Risk Assessment) conferences on August 22, 2003, in Athens and we proceed in Santorini, in 2007 and 2009. We moved to Limassol, Cyprus 2011, with the essential adjustment to ICRA (International Conference to Risk Analysis) opening to other areas besides risks in cancer research. ICRA5 was held in Tomar, a Templar city in Portugal, where actually the extension of RA to Bioinformatics, Management, and Industry, was established. The SPRINGER volume with Eds. Christos P. Kitsos, Teresa A. Oliveira, Alexandros Rigas, and Sneh Gulati, entitled "Theory and Practice of Risk Assessment: ICRA 5, Tomar, Portugal, 2013" was published in 2015, and provides the appropriate evidence. One step forward, further from game theory, towards more fields under risk, was offered by a second SPRINGER volume, in 2018, entitled "Recent Studies on Risk Analysis and Statistical Modeling", with Eds. Teresa A. Oliveira, Christos Kitsos, Amílcar Oliveira and Luìs Grilo.

Meanwhile, as for the meetings, ICRA6 moved to Barcelona-Spain in 2015, ICRA7 to Chicago-USA in 2017, and ICRA8 was held in Vienna-Austria in 2019. A new Springer volume will appear soon, with Eds. Jürgen Pilz, Teresa A. Oliveira, Karl Moder, and Christos P. Kitsos, entitled "Mindful Topics in

Risk Analysis and Design of Experiments - Selected Contributions from ICRA8, Vienna 2019".

We try all these years to extend the field in different scientific areas such as Food Science, Environmental Problems, Management and Economics, and Engineering, among others. Thanks to the interest and active research work of the distinguished participants, many improvements have been succeeded.

Due to the COVID19 pandemic ICRA9, which was supposed to be held in 2021, was postponed one year and thus it will take place NOW in 2022, in the beautiful city of Perugia, Italy. We are looking forward to seeing the ICRA9 presentations, and we are grateful to all the participants for the submitted Abstracts, as well as to all the Session Organizers and Chairs.

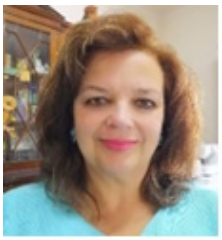
We acknowledge the Magnificent Rector of Universidade Aberta, Professor Carla Oliveira, for the Book of Abstracts (BOA) support and for allowing to allocate it and make it available at the Universidade Aberta Institutional Repository, https://repositorioaberto.uab.pt/. Finally, we are deeply grateful to the Scientific Committee, to the Local Organizing Committee, and, especially to the Executive Committee, Professor Francesca Pierri and Professor Marialuisa Restaino, who worked really hard in order to provide us with a wonderful experience in ICRA9.

We also wish that you can enjoy a lot the Italian environment, music, pasta, ice-creams, coffee and other tasty delights.

WELCOME TO PERUGIA!!!

The Honorary Committee

Christos P. Kitsos
Professor Emeritus
University of West Attica, Athens, Greece
Invited Full Professor
at DCeT-Universidade Aberta
xkitsos@uniwa.gr



Teresa A. Oliveira
Associate Professor with Habilitation
Universidade Aberta-Lisbon and CEAUL
ISI- International Statistical Institute,
Chair of the Committee on Risk Analysis
teresa.oliveira@uab.pt

Greetings from the Executive Committee

Dear Delegates,

thank you for your participation in *The 9^{th} International Conference on Risk Analysis* (ICRA9).

Due to the COVID-19 pandemic, the conference has been organized in a hybrid form by the Department of Economics of the University of Perugia, with the partnership of the Department of Economics and Statistics of the University of Salerno, and is held from 25 to 27 May 2022 in Perugia (Italy).

The aim of ICRA is to provide a forum for presenting and discussing the most emergent topics in theoretical and computational models and methods in Risk Analysis with applications for the risk assessment and the risk management in Life, Biological and Environmental Sciences & Public Health, Economics and Finance, Reliability of Engineering, Technical, Biological & Biomedical Systems.

The conference objective is to assemble researchers and practitioners from universities, institutions and industries from around the world, involved in the same field. Moreover, the meeting encourages and provides opportunities for the participants to exchange ideas and discuss, face to face, new theoretical and practical issues, to establish new collaborations and to contact global partners and research centers leaders for future teamwork.

In this book, the ICRA9 conference abstracts, focused on methods and models applied for analyzing, managing and preventing risks in all research areas, are reported. It is organized into two main parts according to the type of sessions (organized and contributed), and for each session the respective abstracts are shown. Some abstracts are missing, because they are extracted from published papers.

With the Honorary Committees, we are planning to publish the Conference proceedings in a special volume edited by Springer, and extended papers in some special issues of ISI and Scopus-indexed journals.

We thank the Honorary Committee, Professor Christos Kitsos and Teresa A. Oliveira, for giving us

the opportunity to organize this conference and all the members of the Scientific Committee for their availability. A dutiful and sincere thanks is also addressed to the speakers, Professor Chrys Caroni, Professor Paolo Giudici, Doctor Emanuela Raffinetti, and Professor Daniel Farewell, for accepting our invitation to take part in the plenary sessions.

We sincerely thank all session organizers for their contribution to the success of the conference and all speakers for their interesting contributions.

We hope you all enjoy the Conference and Perugia.

Sincerely,

Francesca Pierri
Assistant Professor
University of Perugia, Italy
francesca.pierri@unipg.it

Marialuisa Restaino
Assistant Professor with Habilitation
University of Salerno, Italy
mlrestaino@unisa.it

ICRA9 Committees

Honorary & Executive committees

Christos Kitsos, *University of West Attica* (GR)

Teresa Oliveira, *University of Aberta* (PT)

Francesca Pierri, *University of Perugia* (IT)

Marialuisa Restaino, *University of Salerno* (IT)

Scientific committee

Alessandra Amendola, *University of Salerno* (IT)

Inmaculada Barranco Chamorro, *University of Seville* (SP)

Kristijan Breznik, *International School for Social and Business Studies* (SL)

Chrys Caroni, *National Technical University of Athens* (GR)

Hongsheng Dai, *University of Essex* (UK)

Sergio Destefanis, *University of Salerno* (IT)

Valeria Edefonti, *University of Milan* (IT)

Luz Edler, *German Cancer Research Center* (DE)

Gianna Figà-Talamanca, *University of Perugia* (IT)

Silvia Figini, *University of Pavia* (IT)

Lidia Z. Filus, *Northeastern Illinois University* (USA)

Stefania Galimberti, *University of Milan - Bicocca* (IT)

Beatrice Gasperini, *Marche Polytechnic University* (IT)

Giuseppe Giordano, *University of Salerno* (IT)

Valérie Girardin, *University of Caen Normandy* (FR)

Yvette Gomes, *University of Lisboa* (PT)

Shen Gulati, *Florida International University* (USA)

Annamaria Guolo, *University of Padova* (IT)

George Halkos, *University of Thessaly* (GR)

Samad Hedayat, *University of Illinois Chicago* (USA)

Catherine Huber, *Paris Descartes University* (FR)

Christos Kitsos, *University of West Attica* (GR)

Michele La Rocca, *University of Salerno* (IT)

Nikolaos Limnios, *Sorbonne University* (FR)

Marica Manisera, *University of Brescia* (IT)

Stanislaw Mejza, *Poznań University of Life Sciences* (PL)

Carlo Migliardo, *University of Messina* (IT)

Karl Moder, *University of Natural Resources and Life Sciences* (AU)

Tsuyoshi Nakamura, *Nagasaki University* (JP)

Marcella Niglio, *University of Salerno* (IT)

Amílcar Oliveira, *University of Aberta* (PT)

Teresa Oliveira, *University of Aberta* (PT)

Francesca Pierri, *University of Perugia* (IT)

Jürgen Pilz, *University of Klagenfurt* (AU)

Francesco Porro, *University of Genova* (IT)

Giancarlo Ragozini, *University of Napoli Federico II* (IT)

Marialuisa Restaino, *University of Salerno* (IT)

Juan Eloy Ruiz-Castro, *University of Granada* (SP)

Elena Stanghellini, *University of Perugia* (IT)

Milan Stehlik, *Johannes Kepler University Linz* (AU)

Roberto Tagliaferri, *University of Salerno* (IT)

Mei-Ling Ting Lee, *University of Maryland* (USA)

Maria Prosperina Vitale, *University of Salerno* (IT)

Mariangela Zenga, *University of Milan - Bicocca* (IT)

Paola Zuccolotto, *University of Brescia* (IT)

Keynote Speakers

Chrys Caroni - "National Technical University of Athens, Greece"

The "cure fraction" and other aspects of lifetime data: Analysing students' completion of a degree programme and the risk of failure to graduate.

Paolo Giudici and **Emanuela Raffinetti** - "Department of Economics and management, University of Pavia, Italy"

Artificial Intelligence: principles and risk.

Daniel Farewell - "School of Medicine, Cardiff University, United Kingdom"

Regression by composition.

Abstracts

Contents

Schedule

Wednesday, 25th May 2022

Palazzo Murena - Piazza dell'Università 1		
	Room Dessau	**Room 7**
15:00	Opening Session	
15:45	Keynote Session	
17:10	**OS01** - Statistical Robust Analysis and applications in variable selection, classification and insurance modelling	**OS02** - Risk analysis in public health in collaboration with Italian Society of Medical Statistics and Clinical Epidemiology (SISMEC)
19:00	Welcome Reception at the "Chiostro" of Palazzo Murena	

Thursday, 26th May 2022 - Morning

Department of Economics - Via Giovanni Pascoli, 20		
	Room 4	**Room 5**
08:40	**OS03** - Statistical Approaches Towards Sustainability	**OS04** - Statistical Modelling and Risk Analysis
10:40	Coffee Break - Foyer Room 4 & 5	
11:10	**OS05** - Exploring challenges in analyzing medical data across different study designs and settings, sponsored by the International Biometrical Society - Italian Region	**OS06** - Statistical Modelling for Risk Evaluation in social and economic sciences
12:10	**CS01** - Risk analysis and assessment in health care applications	**CS02** - Modelling in Risk Analysis
13:30	Lunch - 110 Bar Cafè	

Thursday, 26th May 2022 - Afternoon

Department of Economics - Via Giovanni Pascoli, 20			
	Room 4	**Room 5**	**Room Salzano**
14:30	Keynote Session		
15:40	**OS07** - Modeling risks of neoplasia: Chance, environment and genes	**OS08** - Statistics in Modelling	**OS09** - Extreme Value Analysis in Weather Events and the Environment
17:00	Coffee Break - Foyer Room 4 & 5		
17:20	**OS10** - Statistical and Data Science Developments for Risk Assessment in Urban Areas	**OS11** - Statistical Modeling and Inference	**OS12** - New Perspective in Financial Risk Management
20:00	Social Dinner - Ristorante del Sole Living Cafè		

Thursday, 27th May 2022 - Morning

Department of Economics - Via Giovanni Pascoli, 20		
	Room 4	**Room 5**
08:40	**OS13** - High-frequency data in economics and finance	**CS03** - Risk Analysis in new disease development
10:40	Coffee Break - Foyer Room 4 & 5	
11:10	**OS14** - Topics in Financial Econometrics	**CS04** - Statistical and Machine learning models for risk detection
12:10	Keynote Session	
13:30	Lunch - 110 Bar Cafè	

Thursday, 27th May 2022 - Afternoon

Department of Economics - Via Giovanni Pascoli, 20			
	Room 4	**Room 5**	**Room Salzano**
14:30	**OS15** - Advanced Statistical Models for Risk Evaluation	**CS05** - Model and Methods in Risk Analysis	**OS16** - Recent advances in systemic risk assessment
15:30	**OS17** - Risk and opportunities in financial innovation	**CS06** - Risk Analysis in Applied Science	**OS18** - Computational Mathematics and Statistics in Risk Analysis
17:10	Closing Session		

Organized sessions

OS01 - Statistical Robust Analysis and applications in variable selection, classification and insurance modelling. Organizer & Chair: *Hongsheng Dai.*

OS02 - Risk analysis in public health in collaboration with the Italian Society of Medical Statistics and Clinical Epidemiology - SISMEC. Organizers & Chairs: *Rosaria Gesuita and Beatrice Gasperini.*

OS03 - Statistical Approaches Towards Sustainability. Organizer & Chair: *Maria do Rosário Ramos.*

OS04 - Statistical Modelling and Risk Analysis. Organizers & Chairs: *Teresa Oliveira and Amílcar Oliveira.*

OS05 - Exploring challenges in analyzing medical data across different study designs and settings, sponsored by the International Biometrical Society - Italian Region. Organizer & Chair: *Valeria Edefonti.*

OS06 - Statistical Modelling for Risk Evaluation in social and economic sciences. Organizer & Chair: *Francesco Santelli.*

OS07 - Modeling risks of neoplasia: Chance, environment and genes. Organizer & Chair: *Marek Kimmel.*

OS08 - Statistics in Modelling. Organizer & Chair: *Milan Stehlik.*

OS09 - Extreme Value Analysis in Weather Events and the Environment. Organizer & Chair: *Sneh Gulati.*

OS10 - Statistical and Data Science Developments for Risk Assessment in Urban Areas. Organizers & Chairs: *Rodolfo Metulini and Maurizio Carpita.*

OS11 - Statistical Modeling and Inference. Organizer & Chair: *Luis M. Grilo.*

OS12 - New Perspective in Financial Risk Management. Organizer & Chair: *Giovanni De Luca.*

OS13 - High-frequency data in economics and finance. Organizer & Chair: *Alessandra Amendola.*

OS14 - Topics in Financial Econometrics. Organizer & Chair: *Vincenzo Candila.*

OS15 - Advanced Statistical Models for Risk Evaluation. Organizer & Chair: *Silvia Osmetti and Silvia Facchinetti.*

OS16 - Recent advances in systemic risk assessment. Organizers & Chairs: *Francesco Porro and Anna Maria Fiori.*

OS17 - Risk and opportunities in financial innovation. Organizer & Chair: *Gianna Figá Talamanca.*

OS18 - Computational Mathematics and Statistics in Risk Analysis. Organizer & Chair: *Filomena Teodoro.*

OS01 - Statistical Robust Analysis and applications in variable selection, classification and insurance modelling

Feature Screening and Selection for Ultra-high-dimensional Additive Quantile Regression

Daoji Li[1], **Yinfei Kong**[2], **Dawit Zerom**[3]

[1] *California State University, Fullerton, Department of Information Systems and Decision Sciences, USA, dali@fullerton.edu*

[2] *California State University, Fullerton, Department of Information Systems and Decision Sciences, USA*

[3] *California State University, Fullerton, Department of Information Systems and Decision Sciences, USA*

Abstract

This paper is concerned with feature screening and selection for additive quantile regression models in the ultra-high-dimensional setting, i.e. the number of potential variables can grow exponentially with the sample size. We propose a two-step procedure, where in the first step, variable screening is performed using an additive quantile forward regression algorithm, and in the second step we further select the best models by adopting a modified Bayesian information criterion. We establish the screening consistency for the proposed method and examine its finite-sample performance using Monte Carlo simulations.

Keywords

Variable screening, High dimensions, Quantile regression, Sure screening property.

Optimal Reinsurance with Multiple Risks under Dependence Uncertainty

Junlei Hu[1]

[1] *University of Essex, Department of Mathematical Sciences, United Kingdom, j.hu@essex.ac.uk*

Abstract

In this paper we study the optimal reinsurance models for multiple risks, where the marginal distributions are fixed but the dependence structure between these risks are unknown. Due to the unknown dependence structure, the optimal strategy of the worst case of reinsurance models is investigated. We consider two types of risk measures: Value-at-Risk (VaR) and Range-Value-at-Risk (RVaR) including expected shortfall (ES) as a special case, and the general premium principles satisfying certain conditions. It turns out that the capped stop-loss reinsurance treaties are optimal for the worst case scenario under dependence uncertainty for both VaR and RVaR and the general premium principles. Moreover, for the case when the multiple risk $n = 2$, we derive a simpler expression for finding the optimal ceded loss function with VaR by employing a different approach. Finally, applying our obtained results, some numerical studies have been implemented to obtain the optimal reinsurance strategy for some specific multiple risks.

Keywords

Optimal Reinsurance, Dependent Uncertainty, Multiple Risks, Value-at-Risk, Range-Value-at-Risk, Expected Shortfall.

Sequential Estimation for Mixture of Regression Models

Hongsheng Dai[1]

[1] *University of Essex, Department of Mathematical Sciences, United Kingdom, hdaia@essex.ac.uk*

Abstract

Mixture model is a classical statistical model to cluster the heterogeneous population into homogeneous subpopulations. However, for highly heterogeneous population with multiple components, its parameter estimation and clustering results may be ambiguous due to bad local maxima. For subtyping purpose, we work on the finite mixture of regression models with concomitant variable to quantify the mixing probabilities and propose a novel statistical method to identify the components in the mixture sequentially. The presentation will mainly focus on the methodology and precision-medicine application.

Keywords

Mixture model, Clustering, Precision-medicine application.

Feature selection for competing risks model in high and ultra-high dimensions: the case of business failure prediction

Francesco Giordano[1], **Sara Milito**[2], **Marialuisa Restaino**[3]

[1] *University of Salerno, Department of Economics and Statistics, Italy, giordano@unisa.it*

[2] *University of Salerno, Department of Economics and Statistics, Italy, smilito@unisa.it*

[3] *University of Salerno, Department of Economics and Statistics, Italy, mlrestaino@unisa.it*

Abstract

Since the 1930s, and over the last 90 years, both predicting the firms survival and studying the effects of financial ratios on firm's exit has recently attracted new attention from both academics and practitioners. Models Business failure prediction models are important in providing warning for preventing financial distress and bringing about some actions that could have helped to restore firms' financial situation.

Most of the existing literature has mainly focused on only one form of exit, separately investigating any decision to leave the market. Since the seminal paper of Altman (1968), researchers have essentially focused on the binary variable (failing vs. non-failing), and analyzed the companies that actually went bankrupt by means of some models (logit, probit, discriminant analysis, survival analysis and so on) (Ohlson, 1980; Zmijewski, 1984; Lennox, 1999; Shumway, 2001).

However, different exit options exist and may force firms to leave the business. Besides entering in involuntary exit procedure (such as bankruptcy), a firm could opt for merger or acquisition or decide for a voluntary liquidation. Each exit could be driven by different factors (Schary, 1991). Therefore, investigating the determinants leading to the different forms of distressed firm's exit is particularly relevant. In order to examine the effects of explanatory variables across the states of financial distress, a multi-state and competing risks approach can be used (Jones and Hensher, 2004; Rommer, 20015; Jones and Hensher, 2007).

Competing risks data are encountered when firms may fail from multiple causes and the occurrence of one failure event precludes the others from

happening. Two main approaches can be used, to investigate the effects of covariates on the hazard function: cause-specific hazard (CSH) model and subdistribution hazard (SDH) model. The difference between these two approaches relies on the definition of the risk set. In CSH, subjects who experience the competing events are treated as censored, while in SDH they are included in the risk set.

Nevertheless, some issues are still under investigation, such as the selection of financial ratios to define business failure and the identification of an optimal subset of predictors. For this purpose, different methods can be used. Among others we recall the *popularity* (i.e. variables have been chosen among those mostly used in previous studies), the *predictive power* (i.e. variables have been selected among those considered more powerful in predicting corporate failure), statistical tests (e.g. test for differences between means, correlation tests, t test, F test), automatic selection procedure (e.g. Stepwise procedure, Wilks lambda, likelihood ratio), penalized variable selection methods (e.g. Lasso, Lars, Elastic Net, etc.).

In many applications involving competing risks, identifying variables that have effects on CSH or SDH is a critical task.

In the CSH model, screening and variable selection methods developed for Cox model can be easily extended. For the SDH approach, due to the different definitions of the risk set, naive applications of these procedures may be problematic and not suitable. This is particularly true when the number of covariates is larger than the number of observations ($n < p$ and $n << p$) and in presence of multicollinearity between covariates. Thus, the aim is to compare the performance of some existing methods for screening and selecting the most significant variables for predicting business failure, for both CSH and SDH models, for highlighting their main advantages and disadvantages and proposing a new procedure able to identify the relevant covariates in the framework of high and ultra-high dimensions, able to identify the factors that may better influence the risk of leaving the market.

Keywords

Competing risks model, Feature selection, Business failure prediction.

References

Altman, E.I. (1968). The prediction of corporate bankruptcy: a discriminant analysis. *Journal of Finance*, *23*, 193–194.

Jones, S., & Hensher, D.A. (2004). Predicting firm financial distress: A mixed logit model. *The Accounting Review*, *79*, 1011–1038.

Jones, S., & Hensher, D.A. (2007). Modelling corporate failure: a multinomial nested logit analysis for unordered outcomes. *The British Accounting Review*, *39*, 89–107.

Lennox, C. (1999). Identifying failing companies: a re-evaluation of the logit, probit and DA approaches. *Journal of Economics and Business*, *51*, 347–364.

Ohlson, J.A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, *18*, 109–131.

Rommer, A.D. (2005). A comparative analysis of the determinants of financial distress in French, Italian and Spanish firms. *Working paper* no.24, Danmarks Nationalbank, 1–76.

Schary, M. (1991). The probability of exit. *RAND Journal of Economics*, *22*, 339–353.

Shumway, T. (2001). Forecasting bankruptcy more accurately: a simple hazard model. *Journal of Business*, *74*, 101–124.

Zmijewski, M.E. (1984). Methodological Issues Related to the Estimation of Financial Distress Prediction Models. *Journal of Accounting Research*, *22*, 59–82.

OS02 - Risk analysis in public health in collaboration with the Italian Society of Medical Statistics and Clinical Epidemiology - SISMEC

Effect of socio-demographic, environmental and individual factors on SARS-CoV-2 spreading dynamics: endemic-epidemic spatio-temporal point process model*

Katiuscia Di Biagio[1], **Marco Baldini**[1], **Jacopo Dolcini**[2], **Pietro Serafini**[3], **Emilia Prospero**[2]

[1] *Environmental Epidemiology Unit, Regional Environmental Protection Agency of Marche, Ancona (Italy)*
[2] *Department of Biomedical Sciences and Public Health - Section of Hygiene, Polytechnic University, Ancona (Italy)*
[3] *Medical Direction Department, Local Health Authority of Marche, Ancona (Italy)*

Abstract

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) transmission rates and host susceptibility are influenced by several factors, such as, among others, demography, age, gender, socio-economic factors and education. Recently, it has been shown how atmospheric particulate matter (PM) may play an important role in the differential distribution and transmission rates of SARS-CoV-2. For surveillance and prevention in public health, the correct identification of sources and transmission dynamics concerning the endemic and epidemic component of infection, respectively persistent sporadic and rapidly clustered in space and over time, can help to implement intervention strategies to reduce the burden of disease. The aim of this study is to assess the effect of long-term residential exposure to atmospheric particulate matter on COVID-19 incidence and on the dynamics of infectious disease spreading in Marche Region (central Italy) during the first pandemic wave, with a prediction model including both endemic and epidemic components. This study includes all individuals with first positive SARS-CoV-2 nasal/oropharyngeal swab test from February up to May 2020, residents and domiciled in Marche region. All tests are recorded in a healthcare database by the Regional Health Service, along with gender, age, domicile/residence address and employment. Linkage with administrative healthcare data gave information about pre-existing diseases (PED) within the previous 5-years from the first positive SARS-CoV-2 molecular test. Long-term exposure to outdoor fine particulate matter air pollution of 10 microgram or less in

diameter (PM10) concentrations, Temperature and Relative Humidity were estimated at 10 km2 grid cells of Marche region; all residential addresses were geocoded and subjects were assigned pollution and meteorological variables of grid cell containing their addresses. PM10 concentrations was estimated as average of daily concentrations on 2010-2019 years at 10 km2 spatial grid, recorded at the stations of Regional Air Quality Monitoring Networks located across Marche. Daily Temperature and Relative Humidity were provided by the Regional Civil Protection Service. To identify individuals and contextual factors that influence the SARS-CoV-2 spread and its disease rate we used the endemic-epidemic spatio-temporal point process regression model (Meyer et al. 2012; Meyer and Held 2014; Meyer et al. 2017, Paul et al. 2008). The conditional intensity function, that represents the instantaneous infection rate at a specific location and time point given all past infections, was additively decomposed into endemic and epidemic component. The epidemic component regards infected cases directly linked to the previously observed cases, whereas the endemic component concerns new infected cases independent, not directly attributable to the epidemic process, then they do not generate secondary cases. Results showed a significant increment of rate ratio for exposure to increased levels of PM10 both in endemic and epidemic components, in agreement with findings observed by an individual-level study in a city of Northern Italy (Veronesi et al. 2022). Targeted interventions are necessary to improve air quality in most polluted areas to minimize the burden of endemic and epidemic COVID-19 disease and to reduce unequal distribution of health risk.

*The paper is under the peer review process to the Environmental Research Journal.

Keywords

Endemic-epidemic, SARS-CoV-2, PM10, Spatio-temporal point process.

References

Meyer, S., Elias, J., Höhle, M. (2012). A space-time conditional intensity model for invasive meningococcal disease occurrence. *Biometrics*, *68*(2), 607–16.

Meyer, S., Held, L. (2014). Power-law models for infectious disease spread. *The Annals of Applied Statistics*, *8*(3), 1612–1639.

Meyer, S., Held, L., Höhle, M. (2017). Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package surveillance. *Journal of Statistical Software*, *77*(11), 1–55.

Paul, M., Held, L., Toschke, A.M. (2008). Multivariate modelling of infectious disease surveillance data. *Stat Med.*, *27*(29), 6250–67.

Veronesi, G., De Matteis, S., Calori, G., Pepe, N., Ferrario, M.M. (2022) Long-term exposure to air pollution and COVID-19 incidence: a prospective study of residents in the city of Varese, Northern Italy. *Occup Environ Med.*, *79*(3), 192–199.

Environment and disease distribution: from description to analysis of relation

Trerotoli Paolo[1], **Bartolomeo Nicola**[2]

[1] *University of Bari Aldo Moro, Department of Interdisciplinary Medicine, Italy, paolo.trerotoli@uniba.it*

[2] *University of Bari Aldo Moro, Department of Interdisciplinary Medicine, Italy, nicola.bartolomeo@uniba.it*

Abstract

The analysis of the relationship between environmental condition and the distribution of disease is a key point in public health, both to evaluate the effect on health status of the population and to use estimates in risk assessment. Issues for these analyses came for estimation of spatial and time effect. We have dealt with these issues in two different setting: the description of diseases distribution in a polluted area near a big industrial plant and the evaluation of the relationship between environmental and climate condition that could have fostered the spread of COVI19 in Lombardy during 2020.

1. The description of disease distribution in an environmental pollution risk area.

It is well known that Taranto, a city in Apulia, southern of Italy, is considered a high risk environmental area. Local health authorities are always engaged in monitoring health status to have information to use in the risk assessment of population. Data came from continuous surveillance systems such as registry (death and tumor) and hospitalizations (Hospital Discharge Forms). To evaluate time and spatial distribution of many different cancer sites and of a list of chronic and acute non neoplastic conditions in census areas we have applied a bayesian hierarchical model model that took into account the contribution of adjacent areas to estimate the relative risk in small areas and thus manage the spatial structure of the data and their overdispersion. The Besag-York-Mollié model was chosen after comparing its goodness of fit with the Zero Inflated Poisson model and Lawson mixuteres model. The choice of the range of mutual influence between two census areas was made in accordance with the data on the diffusion of air pollutants. The main challenge is in the dynamic assignment of spatial location of each single case and the dynamic location of the population, that are crucial information to determine

rates and relative risks. The risk was estimated for the factor "area of living" and a subject in his life could reside in more than one area, therefore this mobility should be accounted before estimation. To adjust results accounting for this issue we have defined the weight of participation in an area as the ratio between the time of staying in an area divided the total time of living resulting from the registry office data.
The estimation techniques used, and the weight attributed to mobility have reduced the variability of the relative risks between neighboring areas and together with the choice to show them through choroplet maps, it has allowed the identification of risk clusters. The disease mapping has shown a more accurate spatial risk pattern for the areas in the proximity of the implant, identifying the risk by area as expected.
2. The relationship between environmental condition and the spread of acute disease: the case of COVID19 in Lombardy.
In February 2020, after the first case of COVID19 in Lodi, Lombardy became the Italian area with higher number of cases of SARS-CoV2 infection. It's well known the way the virus could be disseminate among the population, but an association between environmental conditions and area of diffusion of the infection could be investigate. The aim of the working group was to apply an explorative analysis to evaluate if there was a direct relationship between air pollution, meteo-climatic factors and spread of the infection.
It is an application of a complex regression model that should hold variability of cases among the county of the region, daily and spatial variability of pollutant in the region and the incubation period. Therefore, a hierarchical mixed model was estimated assuming a Poisson distribution with logarithm transformation as link function; the province information was entered as random intercept. We assumed that the effect of air pollution and climate condition could have an effect some days before the day of the diagnosis, thus a model for each of the 15 days before (lag) the day of the new cases was run.
Regression coefficients, anyway, were different by levels of the factors and lag day, and we need to evaluate, for each lag, the level of the factors that could be considered effective in facilitating the infection. The solution was taken by the analysis of fluctuation test applied in econometrics, that is the structural change regression; we tested the null hypothesis that the regression coefficients remain constant between factor levels compared with the alternative hypothesis that at least 1 coefficient varies. Among relevant results, we detected at lag 10 for the ozone classes <20, 20-24, 25-29, 30-34,

35-39, 40-44, 45-49, 50-59, >=60 $\mu g/m^3$, we found the estimation through least square means of the new cases equal respectively to 4.6, 5.6, 9.5, 11.1, 15.3, 18.0, 12.5, 30.1, 34.1, and therefore a significant break point (BP) for ozone levels above the threshold of 44 $\mu g/m^3$ ($p = 0.028$). BPs were found for temperature below 5°C at lags 1, 4, 7, 8, 9 and 10, and for temperature below 6°C at lags 3 and 5, in all the cases with the trend of cases increasing. Conclusions. In complex settings the process of risk assessment request accurate epidemiological approach. The examples have shown that: in descriptive analysis, even without an individual and causal approach, the accuracy of risk determination should affect estimates; in relationship between air and climate variables as a possible cause to facilitate the spread of the COVID19, a complex approach should be used to have detailed information of the effect on disease.

Keywords

Disease mapping, Environmental pollution, Structural change model, Poisson model, Spatial distribution.

Statistical approaches to assess the impact of pollution in different public health scenarios

Simona Villani[1]

[1] *Department of Public Health, Experimental and Forensic Medicine, University of Pavia, Italy, simona.villani@unipv.it*

Abstract

Risk assessment is largely used in different setting. To better understand risk assessment it is necessary to define risk. According to Royal Society (1972) when the word "risk" is used, in relation to the risk assessment process, it may be defined as: "The combination of the probability, or frequency, of occurrence of a defined hazard and the magnitude of the consequences of the occurrence". Environmental risk assessment (ERA) is the evaluation of risks due to human activities that are dangerous for nature, animals and humans. ERA includes human health risk assessment that may be conducted using different type of studies: descriptive, ecological, and analytical.

Starting from the epidemiological approach it needs to plan the most appropriate statistical methods. The descriptive study requires simple statistical analyses to describe the health profile of exposed people. The ecological design needs statistical methods which permits to verify the first relationship between exposure and health outcome. Among these approaches, in particular, the Bayesian spatial-temporal methodologies have a relevant role in case of municipalities with small dimensions since they ensure to smooth easily the background noise. The analytical studies offer a wide type of statistical approaches to evaluate the association between exposure and health outcome. In some condition might be more interesting to investigate the health risk due to different scenarios of exposure. In this case, the first phase of the analysis will assign to each subject its level of exposure and then on the basis of the cluster of exposure to which it belongs, the association with health outcomes will be estimated. In detail, two different "case studies" will be presented: 1) a spatial-temporal distribution of cardiovascular mortality in the Province of Pavia in 2010 through 2015 in association with environmental pollution exposure by means of an ecological study; 2) the health impacts of the existing facility refinery on people living in the neighbouring areas by means of a case control study after clustering of exposure identification through k-means

model. In the first study different models (hierarchical log-linear model) have been assessed: temporal parametric trend components were included together with some random effects that allowed the accounting of spatial structure of the region. The best model has been selected using Watanabe-Akaike Information Criteria (WAIC) and Leave One Out Information Criteria (LOOIC). The environmental exposure (PM2.5) showed a strong significant effect on cardiovascular mortality accounting for the urbanisation level of each geographical unit. In the second case study, a sample size from eligible people living in two municipalities was selected and individual environmental exposure (using ground level pollution from AERMOD model) was assigned by linking the geocodes of participants' addresses to the modelled surface, applying a bilinear interpolation. The identified clusters using a non-hierarchical K-means were considered as a proxy of individual level of exposure. In order to estimate the impact of a cluster of exposure on health outcome a multivariable logistic regression models have been implemented adjusting for several covariates. The adjusted estimate effect of environmental exposure (PM10) on hospitalization indicated a possible excess in most-exposed people, but the effect was not significant.

Keywords

Spatial analysis, Environmental pollution exposure, Cardiovascular mortality, Case control study.

Use of frailty index in epidemiology and public health

Antonella Zambon[1], **Patrizia Enrico**[2], **Giuseppe Bellelli**[3]

[1] *University of Milano-Bicocca, Department of Statistics and Quantitative Methods, Italy, antonella.zambon@unimib.it*
[2] *ASST San Gerardo Monza, Acute Geriatric Unit, Italy, patrizio.enrica@gmail.com*
[3] *University of Milano-Bicocca, School of Medicine and Surgery, Italy, giuseppe.bellelli@unimib.it*

Abstract

Introduction. The currently most accepted definition of frailty is "a clinically recognizable state of greater vulnerability resulting from the decline in reserve and function associated with aging in multiple physiological systems, including the ability to cope with daily or acute stressors" (Fried et al., 2001). The problem of estimating frailty becomes of increasing importance, from both an etiological and social perspective, if considering that up to a quarter of the older adult population is "frail", depending on the criteria used to define it. (Sepehri et al., 2020). Different approaches can be taken to measure frailty (Clegg et al., 2013): the approach based on the concept of the phenotype of frailty (Fried et al., 2001) and the approach that refers to the accumulation of deficits, developed by Rockwood and operationalized with the Frailty Index (FI) (Rockwood and Mitnitski, 2007). Both approaches have shown to successfully predict the development of adverse events in older people, including mortality. However, while the construction of the FI may include the presence of impairment and/or disability, the frailty phenotype is considered a condition that precedes the development of disability. On the other hand, the FI provides a quantitative value and thus may offer an estimation of the severity of frailty which has substantial advantages. In the context of these two approaches, many indicators of frailty have been proposed. Currently, little information is available on the comparison between predictive performance of these indicators and other measures commonly used to identify patients at risk of poor outcomes.

Aim. To compare the ability of FI and other frailty indicators (i.e. Clinical Frailty scale (CFS)) and measures of comorbidity (i.e. Charlson Comorbidity Index (CCI)) in predicting clinical outcome with real data based on a cohort

of older patients admitted to an Acute Geriatric Unit (AGU) of San Gerardo, Monza (Italy) with a diagnosis of sepsis.

Methods. Real data derived by a retrospective cohort study including patients, aged 70 years, acutely hospitalised at the AGU of the San Gerardo hospital (Monza, Italy) between March 1 st, 2017 and January 31 st, 2020 with a diagnosis of sepsis. Frailty was assessed with two tools: the Clinical Frailty Scale (CFS), a nine-point scale which is based on the clinical evaluation of symptoms, mobility, physical activity, and function and the FI, developed in accordance with the Rockwood's approach and based on 29 variables, including diseases and functional impairments, sensory deficits, malnutrition and blood examinations. Univariate and multiple logistic models were fitted to evaluate the performance in predicting the in-hospital and 6-month mortality of each single indicators and of their combination. Discriminating performance was reported as the Area under the Receiver Operating Characteristic (AUROC) curve and its corresponding 95% Confidence Intervals (CI). All statistical tests were two-sided and significance level was set at 0.05.

Results. In the study period, 290 (8.7%) patients admitted to AGU, were diagnosed with sepsis. Fifty patients were excluded from the analysis because of missing data. Overall, 240 patients were included; the median age was 85 years old (interquartile range, IQR, 80 – 89), and 98 (40.8%) were women. Eighty-one patients (33.8%) died during the hospital stay, and 145 (60.4%) within 6 months from hospital admission. CFS and FI showed similar accuracy to predict in-hospital mortality (CFS - AUROC=0.605; 95% CI 0.531–0.678; FI - AUROC=0.582; 95% CI 0.505–0.659). Comorbidity indexes showed lower predictive capacity. The best predictors of 6-month mortality were FI (AUROC=0.677; 95% CI 0.607–0.746).

Conclusions. Although frailty is frequently associated with comorbidity, this study shows that it is an independent nosological entity which provides more relevant information in terms of mortality than comorbidity. Data relative to the predictive performance of different frailty indicators proposed in literature may be useful to provide a setting-specific tool to identify frail hospitalised patients and quickly predict their risk of poor outcomes.

Keywords

Frailty, Death, Predictive performance.

References

Fried, L.P., Tangen, C.M., Walston, J., Newman, A.B., Hirsch, C., Gottdiener, J., et al. (2001). Frailty in older adults: evidence for a phenotype. *J. Gerontol. A. Biol. Sci. Med. Sci.*, *56*, M146–56.

Sepehri, K., Braley, M.S., Chinda, B., Zou, M., Tang, B., Park, G., Garm, A., McDermid, R., Rockwood, K., Song, X., (2020). A Computerized Frailty Assessment Tool at Points-of-Care: Development of a Standalone Electronic Comprehensive Geriatric Assessment/Frailty Index (eFI-CGA). *Front. Public. Health.*, 31, 8–89.

Clegg, A., Young, J., Iliffe. S., Rikkert, M.O., Rockwood, K. (2013). Frailty in elderly people. *Lancet*, *381*, 752–62.

Rockwood, K., Mitnitski, A. (2007). Frailty in relation to the accumulation of deficits. *J. Gerontol. A. Biol. Sci. Med. Sci.*, *62*, 722–7.

OS03 - Statistical Approaches Towards Sustainability

Adjustment of state space models in view the improvement of forecasts of meteorological variables

Marco Costa[1,4], **F. Catarina Pereira**[2,5], **A. Manuela Gonçalves**[3,5]

[1] *University of Aveiro, Águeda School of Technology and Management - ESTGA, Portugal, marco@ua.pt*

[2] *University of Minho, Department of Mathematics, Portugal, up202010700@edu.fe.up.pt*

[3] *University of Minho, Department of Mathematics, Portugal, mneves@math.uminho.pt*

[4] *Centre for Research and Development in Mathematics and Applications - CIDMA, University of Aveiro, Portugal*

[5] *Center of Mathematics, University of Minho, Portugal*

Abstract

State space models have been applied in several areas due to their flexibility, as they can accommodate more than one source of variability. Another added value of these models is the possibility of explaining the response variable from an unobserved variable, the state. In many areas the state is a focus of modeling, that is, the variable of interest is not observed directly but its linear transformation is observed and may have an additive noise. In these cases the main goal is to get the best prediction of the state variable at each instant in order to minimize by filtering out the noise. However, in many other situations, state space models are considered stochastic models that allow establishing stochastic relations between a variable and a quantity, usually between an observed quantity, taken without error, and another, the variable of interest, which is related to the quantity through the equations of the model. In this work it is analyzed the possibility to improvement short-term forecasts from a website for a specific location based on other data more accurate. In this first step, a state space model is adjusted in order to establish a stochastic relation-ship between the observed maximum temperature in the location of interest with a portable station installed temporarily and the website's forecasts for that location. This approach was developed on the basis of data obtained from the ongoing "TO CHAIR: The Optimal Challenges in Irrigation" project. The adjustment of the state space

model in this context raises several issues that need solutions or adequate approaches. In fact, in this case, the state variable is not the main interest or the noise filtering is not the main perspective in the modeling. First issue that need some analysis and attention is the possibility of outliers. This can be a problem because outliers can affect the fit of the model, can introduce bias in parameter estimation, and can also lead to confidence intervals with larger amplitudes. These consequences can influence the forecasts in a very significant way. Usually, outliers are studied in the perspective of ARIMA or GARCH models, (Muller et al., 2009; Hotta and Tsay, 2012). In this case the relationship between the observed temperature, Y_t, at the location and the temperatures predicted, $W_{t,h}$, by the website h =1, 2,...or 6 days in advance was carefully analyzed. It is expected that the ratios $Y_t/W_{t,h}$ will be close to 1. In this perspective the outlier definition has to be adjusted to this case. The outliers' analysis is related, or can be, to some difficulties on the parameter estimation. For this, the parameter estimation is discussed in order to provide suitable models which allow to improve the short-term forecasts of the maximum temperature.

Keywords

State space modeling, Kalman filter, Forecasting, Temperature, TO CHAIR project.

Acknowledgements

This work has received funding from FEDER/COMPETE/NORTE2020/ POCI/FCT funds through grants UID/EEA/- 00147/20 13/UID/IEEA/00147/ 006933- SYSTEC, project and To CHAIR - POCI-01-0145-FEDER-028247. This work was also partially supported by the Portuguese FCT Projects UIDB/00013/2020 and UIDP/00013/2020 of CMAT-UM and the Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), references UIDB/04106/2020 and UIDP/04106/2020. This research was financed by national funds through FCT (Fundação para a Ciência e a Tecnologia) through the individual PhD research grant UI/BD/150967/2021 of CMAT-UM.

References

Costa, M., Pereira, F.C., Gonçalves, A.M. (2021). Improving Short-Term Forecasts of Daily Maximum Temperature with the Kalman Filter with

GMM Estimation. Lecture Notes in Computer Science, 12952 LNCS, pp. 552–562.

Gonçalves, A.M., Costa, C., Costa, M., Lopes, S.O., Pereira, R. (2021). Temperature Time Series Forecasting in the Optimal Challenges in Irrigation (TO CHAIR), *Computational Methods in Applied Sciences*, *55*, pp. 423–435.

Hotta, L. K., Tsay, R. S. (2012). Outliers in GARCH processes. In W. R. Bell, S. H. Holan and T. S. McElroy (Eds.), Economic time series: modeling and seasonality (pp. 337–358). Boca Raton: Chapman Hall.

Muler, N., Peña, D., Yohai, V. (2009). Robust estimation for ARMA models. *The Annals of Statistics*, *37*, 816–840.

A statistical boost to assess water quality

Clara Cordeiro[1], **Sónia Cristina**[2], **Farhat-Un-Nisa Bajwa**[3]

[1] *Faculdade de Ciências e Tecnologia, Universidade do Algarve, and CEAUL, Faculdade de Ciências, Universidade de Lisboa, Portugal, ccordei@ualg.pt*

[2] *CIMA-Centre for Marine and Environmental Research, Universidade do Algarve, Campus de Gambelas, 8005-139, Faro, Portugal, sccristina@ualg.pt*

[3] *Faculdade de Ciências e Tecnologia, Universidade do Algarve, Portugal, a66695@ualg.pt*

Abstract

The high urbanized zones in the coastal regions and maritime economic activities are increasing globally and, consequently, causing pressures on the marine and coastal ecosystems (Cristina et al., 2016; Bertram et al.,2014). Therefore, it is essential to monitor the water quality of the marine and coastal waters to assess the effect of human pressures on these waters, which are a risk to human health and on economic activities (Brito et al., 2020). Due to this concern, European directives such as the Water Framework Directive (WFD, 2000/60/EC) and Marine Strategy Framework Directive were developed to achieve and maintain these water's good ecological and environmental quality status.

Currently, the monitoring programs use in situ data. However, this is limited by time and space and is expensive, unlike the data acquired by ocean colour satellite sensors, which allows for increasing the time scale. One of the main water quality indicators that can be retrieved from space is chlorophyll-a (Chl-a). The 90^{th} percentile, P_{90}, is a commonly used descriptive measure to assess the Chl-a concentration because it is robust in the presence of outliers.

This work uses time series of Chl-a retrieved by satellites, with a time horizon from January 1997 until June 2021, from two stations off the coast of the Algarve: Sagres and Guadiana. These time series have different stochastic behaviours: Sagres is characterised by a strong seasonal pattern, with higher Chl-a values in spring and summer, and Guadiana has a weak seasonal pattern. In general, the concentration of Chl-a in Guadiana has higher values than in Sagres, due to its proximity to the Guadiana estuary. Thus, Sagres is classified as a "High state" once the P_{90} for Chl-a concentration is lower

than the reference value of 8 mg m^{-3}. In the case of Guadiana, the value is closer to the reference value, 1.8 mg m^{-3}, and it is often exceeded.

The objective is to study the pertinence of using satellite data in water quality assessment and contribute with new statistical methodologies, such as time series models. In addition, the results achieved using these methods will be compared with the "traditional" one. Moreover, it is possible to identify the seasons or the months that occur the most significant changes in the Chl-a concentration. This is a valuable information that is useful for different stakeholders, such as the Portuguese Environment Agency, coastal managers and the economic activities sectors (such as aquaculture, fisheries and tourism).

Keywords

Time series, Statistical modeling, Water quality status, Satellite data.

Acknowledgements

Clara Cordeiro is partially financed by national funds through FCT - Fundação para a Ciência e a Tecnologia under the project UIDB/00006/2020. Farhat-Un-Nisa Bajwa would like to thank the CEAUL for support through Grant UIDP/00006/2020 funded by FCT through national funds. Sónia Cristina is financed through the FCT under the grant: CEECIND/01635/2017 and would like to acknowledge the financial support of FCT to CIMA through UIDP/00350/2020.

References

Bertram, C., Rehdanz, K. (2013). On the environmental effectiveness of the EU Marine Strategy Framework Directive. *Marine Policy*, *38*, 25–40.

Brito, A.C., Garrido-Amador, P., Gameiro, C., Nogueira, M., Moita, M.T., Cabrita, M.T. (2020). Integrating In Situ and Ocean Color Data to Evaluate Ecological Quality under the Water Framework Directive. *Water*, *12*, 3443. https://doi.org/10.3390/w12123443.

Cristina, S., Icely, J., Costa Goela, P., Angel DelValls, T., Newton, A. (2015). Using remote sensing as a support to the implementation of the European Marine Strategy Framework Directive in SW Portugal. *Continental Shelf Research*, *108*, 169–177.

Projection of the number of amputations in diabetics: An aid for the planning of sustainable portuguese health services

Elisabete Carolino[1,2], **José Pedro Matos**[3], **M. Rosário Ramos**[4,5]

[1] *H&TRC - Health & Technology Research Center, ESTeSL- Escola Superior de Tecnologia da Saúde, Instituto Politécnico de Lisboa, Portugal, etcarolino@estesl.ipl.pt*

[2] *ISAMB - Instituto de Saúde Ambiental, Faculdade de Medicina da Universidade de Lisboa, Portugal*

[3] *ESTeSL - Escola Superior de Tecnologia da Saúde, Instituto Politécnico de Lisboa, Portugal, fulgenciomatos@estesl.ipl.pt*

[4] *Universidade Aberta and CEG (Centro de Estudos Globais), MariaR.Ramos@uab.pt*

[5] *CEAUL, Faculdade de Ciências, Universidade de Lisboa, Portugal*

Abstract

Amputation is the loss (surgical or traumatic) of a segment of the body, applied in the event of an injury (traumatic, vascular or other) that has irreparably affected the human being, causing functional limitation. Amputation also represents a considerable negative socio-economic impact for families and governments.

Knowing the numbers, establishing the prevalence and future trends in limb loss is important for health care planning and for the rational allocation of resources, as a response to the growing indicators of demand for prostheses and related services.

Diabetes is known as one of the leading causes of morbidity and mortality worldwide. Portugal is the fourth country in the European Union with the highest incidence rate of diabetes, according to the IDF Atlas (International Diabetes Federation, 2021).

In the Annual Report of the National Diabetes Observatory - 2019 Edition, it is estimated that in 2018 there will be between 605 and 618 new cases of diabetes per 100 000 inhabitants.

Although diabetes prevalence data are updated annually in Portugal through the INSA - Médicos Sentinela, with the COVID-19 pandemic, it is estimated that at least 20,000 diabetics have not had access to the necessary conditions for an early diagnosis. Therefore, data from this period underestimate the

prevalence rate of Diabetes. The prevalence of complications related to diabetes, namely diabetic foot, tends to increase with the increasing number of people with the pathology.

Studies show that 1 in 7 diabetics will develop foot ulcers in their lifetime, which is a risk factor for amputation. Amputation is probably the most feared and recognized complication of diabetes. It is estimated that about 50 % of amputations and ulcerations can be prevented by evaluating the foot, degree of risk of ulceration, thus allowing the implementation of preventive strategies.

The main objectives of this work are: i) quantify and characterize the amputations performed in Portugal from the year 2000 to the present; ii) Model, estimate and predict the number of amputations by etiology and by level for the future.

This is a retrospective observational cross-sectional study, designed using the "Hospital Morbidity Database (BDGDH), for episodes with amputations", provided by the Central Administration of the Health System, IP. (ACSS), supervised by the Ministry of Health.

The data refer to hospitalizations related to amputations in public hospitals of the National Health Service (SNS) in the mainland, which occurred between 2000 and 2019. It includes the dependent variables, calendar year, age, district (place of birth), gender, etiology and level of amputation. It consists of the 20 initial records of diagnoses and the respective procedures associated with each episode of hospital admission. The disease and procedure criteria were defined according to the International Statistical Classification of Diseases, Injuries and Causes of Death1–3, in the 9th revision of 1975 (ICD-9), with the respective code limits.

Annual trends were estimated through Poisson regression models as well as the future incidence rates, sex and age group stratified. Incidence rate projections were adjusted to the distribution of the resident population in mainland Portugal, considering the Portuguese statistical projections (National Institute of Statistics - INE).

Keywords

Amputation, Diabetes, Poisson Regression, Projections.

Acknowledgements

Elisabete Carolino was supported by FCT/MCTES (UIDB/05608/2020 and UIDP/05608/2020). M. Rosário Ramos is partially financed by national

funds through FCT - Fundação para a Ciência e a Tecnologia under the project UIDB/00006/2020.

References

Carvalho, J.A. (2003). História das amputaçōes e das próteses. *Carvalho JA, ed. Amputaçōes de Membros Inferiores: Em Busca Da Plena Reabilitaçāo* 2nd Ed. Sãao Paulo - Brasil: Editora Manole Ltda, 365.

Cesar, C.L.G., Laurenti, R., Buchala, C.M., Figueiredo, G.M., Carvalho, W.O., Caratin, C.V.S. (2001). Uso da Classificaçāo Internacional de Doenças em Inquéritos de Saúde. *Rev Bras Epidemiol*, *4*(2), 120–130.

Coffey, L., Gallagher, P., Desmond, D. (2014). Goal Pursuit and Goal Adjustment as Predictors of Disability and Quality of Life Among Individuals With a Lower Limb Amputation: A Prospective Study. *Arch Phys Med Rehabil*, *95*(2), 244–252.

Holman, N., Young, R.J., Jeffcoate, W.J. (2012). Variation in the recorded incidence of amputation of the lower limb in England. *Diabetologia*, *55*(7), 1919–1925.

Kolossváry, E., Ferenci, T., Kováts, T., Kovács, L., Járai, Z., Menyhei, G., Farkas, K. (2015). Trends in Major Lower Limb Amputation Related to Peripheral Arterial Disease in Hungary: A Nationwide Study (2004–2012). *Eur J Vasc Endovasc Surg*, *50*(1), 78–85.

Sousa-Uva, M., Antunes, L., Nunes, B., Rodrigues, A.P., Simōes, J.A., Ribeiro, R.T., Boavida, J.M., Matias-Dias, C. (2016). Trends in diabetes incidence from 1992 to 2015 and projections for 2024: A Portuguese General Practitioner's Network study. *Primary Care Diabetes*, *10*(5), 329–333.

WHO (1980). International Classification of Impairments, Disabilities, and Handicaps. Geneve.

Ziegler-Graham, K., MacKenzi,e E.J., Ephraim, P.L., Travison, T.G., Brookmeyer, R. (2008). Estimating the Prevalence of Limb Loss in the United States: 2005 to 2050. *Arch Phys Med Rehabil*, *89*(3), 422–429.

Extreme value theory in time series analysis to estimate risk measures

Manuela Neves[1], **Clara Cordeiro**[2], **Dora Prata Gomes**[3]

[1] *Instituto Superior de Agronomia and CEAUL, Universidade de Lisboa, Portugal, manela@isa.ulisboa.pt*

[2] *Faculdade de Ciências e Tecnologia, Universidade do Algarve, and CEAUL, Faculdade de Ciências, Universidade de Lisboa, Portugal, ccordei@ualg.pt*

[3] *NOVA School of Science and Technology (FCT NOVA), and Center for Mathematics and Applications (CMA), FCT NOVA, Portugal, dsrp@fct.unl.pt*

Abstract

The analysis of financial returns and associated risks is an important issue. Some statistical methods consider modelling financial returns through the normal distribution. However, as it is known, the kurtosis that measures the degree of peakedness of a distribution relative to the tails usually shows, for that kind of data, higher values than for a normal setup. This means that most of the variance is due to extreme deviations that are not predicted by the normal distribution and can be a signal that a return series has fat tails. Extreme value theory must be then applied to the analysis and modelling of those series. In classical time series modelling, a key issue is to determine statistically how many parameters have to be included in the model. However, special care must be given to extreme events in the series that need specific statistical procedures based on the behaviour of extremes. Extreme value models were initially obtained through arguments that assumed an underlying process consisting of a sequence of independent and identically random variables. However, in the financial area, temporal independence is unrealistic. A stationary setup is the most natural generalisation of a sequence of independent random variables. In the last decades, many signs of progress have been made in parameter estimation of extreme values in time series relevant to asymptotic results. However, for finite samples, limiting results provide approximations that can be poor. Computer intensive methods, among which we refer to Generalised Jackknife and Bootstrap methodologies, have improved results in parameter estimation in statistics

of extremes. Resampling techniques and exponential smoothing methods for modelling and predicting a time series are computational procedures proposed to improve the performance of some risk measures estimators. Our approach will be applied to a set of observed stock market data, using the R software.

Keywords

Computational procedures, Estimation, Extreme value theory, Time series, Risk measures.

Acknowledgements

Manuela Neves and Clara Cordeiro are partially financed by national funds through FCT - Fundação para a Cência e a Tecnologia under the project UIDB/00006/2020. Dora Prata Gomes is financed by national funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects UIDB/00297/2020 and UIDP/00297/2020 (Center for Mathematics and Applications).

References

Coles, S. (2001). An introduction to statistical modeling of extreme values. *Springer Series in Statistics Springer-Verlag, London.*

Hall, P. (1990). Using the bootstrap to estimate mean squared error and selecting parameter in nonparametric problems. *Multivar Anal.*, *32*, 177–203.

Neves, M. and Cordeiro, C. (2020). Modelling (and forecasting) extremes in time series: A naive approach. *: Atas do XXIII Congresso da Sociedade Portuguesa de Estatística.* Fátima Salgueiro, et.al.(edts). *124*, 189–202.

Penalva, H., Gomes, D. Neves, M. and Nunes, S. (2019). Testing conditions and estimating parameters in extreme value theory: application to environmental data. *Revstat-Statistical Journal, 17(2)*, 187–207.

Longterm air temperature series in Europe: a comparative analysis of multiple change point detection approaches

Magda Monteiro[1], **Marco Costa**[2]

[1] *ESTGA - Águeda School of Technology and Management & CIDMA - Center for Research and Development in Mathematics and Applications, University of Aveiro, Portugal, msvm@ua.pt*

[2] *ESTGA - Águeda School of Technology and Management & CIDMA - Center for Research*

Abstract

This work presents a comparative analysis of several approaches for detecting multiple change points in long time series of air temperature in Europe. The modelling of the series is performed through state space models and the detection of change points is done by applying two complementary approaches, one using the innovations series, in parametric and non-parametric perspectives, and the other using Kalman filter smoothers which represent the trend prediction of the series under analysis, in a parametric perspective. In half of the analyzed cities only one change point is detected, mostly at the end of the eighties, with all the applied methods. In the remaining series, there are differences between the number of change points detected by the different approaches. We highlight the cities in western Europe, where two change points are detected for the majority of the cities and methods. In Lisbon case, the smoother approach has detected three change points, being this situation the only case with this number of changepoints. In Prague and Vienna, only the non-parametric method applied to innovations has detected two change points.

Keywords

Air temperature, Climate change, State space modeling, Change point detection.

References

Costa, M., Alpuim, T. (2010). Parameter estimation of state space models for uni- variate observations. *Journal of Statistical Planning and Inference*, *140*(7), 1889–1902. https://doi.org/10.1016/j.jspi.2010.01.036.

Costa, M. and Monteiro, M. (2017). Statistical modeling of an air temperature time series of European cities. In: Advances in Environmental Research, pp. 213–236. Nova Science.

Jarušková, D., Antoch, J. (2020). Changepoint analysis of klementinum temperature series. *Environmetrics*, 31:e2570. https://doi.org/10.1002/env.2570.

Ross, G.J. (2015) Parametric and nonparametric sequential change detection in R: The cpm package.*Journal of Statistical Software*, *66*(3), 1–20. http://www.jstatsoft.org/v66/i03/.

Ross, G. J. (2015). Parametric and Nonparametric Sequential Change Detection in R: The cpm Package. *Journal of Statistical Software*, *66*(3), 1–20. https://doi.org/10.18637/jss.v066.i03.

OS04 - Statistical Modelling and Risk Analysis

Data Science Training for Finance and Risk Analysis: A Pedagogical Approach with Integrating Online Platforms

Afshin Ashofteh

NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal, aashofteh@novaims.unl.pt

Abstract

The main discussion of this paper is a method of data science training, which allows responding to the complex challenges of finance. To create and deploy financial models for risk management, the ability to incorporate new data and Big Data sources, as well as benefit from emerging technologies such as web technologies, remote data collection methods, user experience Platforms, and ensemble machine learning methods, becomes increasingly important. Automating, analysing, and optimizing a set of complex financial systems requires a wide range of skills and competencies that are rarely taught in typical finance and econometrics courses. Adoption of these technologies for financial problems necessitates new skills, and knowledge about processes, quality assurance frameworks, technologies, security needs, privacy, and legal issues. In this paper, I discuss a pedagogical approach for data science training in finance and risk analysis, with a graphical summary of necessary skills. A case study of active learning and learning by doing for financial data science course is presented, following with the results of a teaching experience of this course, online and in-person, with a combination of different technologies and platforms in an integrated manner. The outcomes of an online Q/A on the Kaggle competition platform, an online book club, an online video platform, and an online discussion group for this course are presented with their advantages and disadvantages, and vulnerabilities.

Keywords

Data science, Finance, Risk, Pedagogical, Active learning.

References

Ashofteh, A., & Bravo, J. M. (2021a). A conservative approach for online credit scoring. *Expert Systems with Applications*, *176*, 114835. https://doi.org/10.1016/j.eswa.2021.114835.

Ashofteh, A., & Bravo, J. M. (2021b). Data science training for official statistics: A new scientific paradigm of information and knowledge development in national statistical systems. *Statistical Journal of the IAOS*, *37*(3), 771–789. https://doi.org/10.3233/SJI-210841.

Ashofteh, A., & Bravo, J. M. (2021c). Life Table Forecasting in COVID-19 Times: An Ensemble Learning Approach. *In 16th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1–6). IEEE. https://doi.org/10.23919/CISTI52073.2021.9476583.

Ashofteh, A., Bravo, J. M., & Ayuso, M. (2021). A Novel Layered Learning Approach for Forecasting Respiratory Disease Excess Mortality during the COVID-19 pandemic. *In 21th Portuguese Association of Information Systems Conference, CAPSI 2021*. Retrieved from https://capsi2021.apsi.pt/index.php/en/.

Ashofteh, A., & Bravo, J. M. (2019). A non-parametric-based computationally efficient approach for credit scoring. *In Atas da Conferencia da Associacao Portuguesa de Sistemas de Informacao*. Associacao Portuguesa de Sistemas de Informacao. Retrieved from https://www.scopus.com/record/display.uri?eid=2-s2.0-85086641145&origin=inward&txGid=0e87a8c228db37a09073b1441dfffe9e.

Ashofteh, A. (2018). Mining Big Data in statistical systems of the monetary financial institutions (MFIs). *In International Conference on Advanced Research Methods and Analytics (CARMA)*. Valencia: Universitat Politecnica de Valencia. https://doi.org/10.4995/carma2018.2018.8570.

Computing the distribution of two uncorrelated normally distributed variables

Seijas-Macias, Antonio[1,4], **Olivieira, Amílcar**[2,4], **Oliveira, Teresa A.**[3,4]

[1] *Universidade da Coruña, Departamento de Economia, Spain, antonio.smacias@udc.gal*

[2] *Universidade Aberta, Departmento de Ciência e Tecnologia, Portugal, amilcar.oliveira@uab.pt*

[3] *Universidade Aberta, Departmento de Ciência e Tecnologia, Portugal, teresa.oliveira@uab.pt*

[4] *CEAUL - Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*

Abstract

The procedures for determining the probability density function (PDF) of a product of two distributions are not, in general, straightforward to derive. There are several procedures for different types of distributions and their characteristics (domain, independence, range, ...) In order to determinate the PDF of the product of two variables the Rohatgi's theorem is an important result. Rohatgi (1976) defines the density function of the product of two distributions from the density functions of each of the elements of the product. The determination of the density function of the product requires the use of integrals, which do not always have analytical solutions, which means that in many cases it will be necessary to resort to approximations by means of numerical integration. On the other hand, the theorem presents a discontinuity at the product value zero, since the product density function cannot be defined at that point.

In 2004, Glen et al. adapt the theorem to the product of various types of distributions (continuous and discrete), by applying numerical integration methods and defining the various regions of integration according to the values of the product variables. The application to the product of two uncorrelated normally distributed variables is only defined for the particular case of the product of two standard normal distributions. In this case, we have an exact expression of the integrals via the modified Bessel function. For other situations, however, exact expressions of the product density function could not be determined. The published version has several limitations:

the authors restrict themselves to two distributions that fall entirely in the positive range. However, it is easy to adapt this theorem to other scenarios. One can even consider the variables with positive and negative values. This approach also poses the additional problem that the density function of the product is not defined for the value zero. Another important limitation of this approach is the range of the statistical distribution. This range can't be infinity and this fact could be a problem for normal distribution.

Our study analyses the implementation of a procedure in R to calculate the Probability Density Function (PDF) of the product of two Normally Distributed Random Variables. We consider the analyses of this problem development by Glen et al. (2004), where a similar procedure is implemented in a Maple list-of-list data structure. Our approach using R is numerical calculus oriented one. We develop a procedure in R that allows us to obtain the distribution function of the product of two normal variables, considering different scenarios depending on the range of variation of the variables considered. This procedure uses numerical integration processes to calculate the distribution function.

Keywords

Numerical Integration, Distribution Function, Normal Distributions Product, Rohatgi's Theorem.

References

Glen, A.G., Leemis, L. M., Drew, J. H. (2004). Computing the distribution of the product of two continuous random variables. *Computational Statistics & Data Analysis*, *44*, 451–464.

Rohatgi, V.K. (1976). An Introduction to Probability Theory Mathematical Statistics. *Wiley*, New York.

Scenario-Based Risk Aggregation Modeling with Copulas and Mixture distributions

Jörge Basilio[12], **Amílcar Oliveira**[1]

[1] *Universidade Aberta, Lisbon, Portugal, jorge.basilio@uab.pt*
[2] *National College of Ireland, Dublin, Ireland, jorge.basilio@ncirl.ie*

Abstract

The ability to adequately model risks is crucial for financial companies. The measurement of financial risks has been one of the main preoccupations of actuaries and finance practitioners for a long time. They are required to consider more risk types into their models in order to achieve the goal of integrated risk management. The challenge of integrated risk management is how to aggregate different risk types as well as risk distributions. This objective requires modelling the multivariate dependence between the various risk types to be considered.

The Integrated risk management for financial institutions requires an approach for aggregating risk types (such as market and credit) whose distributional shapes vary considerably. Ignoring the coupling risk's influence often leads to underestimating the financial risk involved.

We introduce an aggregation model as well as copula-based Conditional Value-at-Risk (CVaR) sampling algorithm. The risk aggregation method is a flexible way to aggregate risks since the model does not require a fully specified joint distribution of all risks until an additional assumption is added. By using this technique it will be possible include more realistic marginal distributions that capture essential empirical features of the risks, including accurate skewness and heavy-tails and a rich dependence structure.

A copula is a function that links univariate marginal distributions to the joint multivariate distribution function, which is a defective way to describe joint distributions of two or more random variables. Copulas allow the aggregation of diverse marginal distributions and capture some of the essential features found in risk management such as skewness and heavy tails, as introduced in the seminal paper by Sklar, where is demonstrated that all finite-dimensional probability laws have an associated copula function that describes the dependency of their marginal distributions. The copula approach will be yet enhanced by using mixture distribution to improve the modelling of the marginal distribution.

Keywords

Risk Aggregation, Copulas, Mixture Distribution, Conditional Value-at-Risk.

References

Arbenz, P., Hummel, C., & Mainik, G. (2012). Copula based hierarchical risk aggregation through sample reordering. *Insurance: Mathematics and Economics*, *51(1)*, 122-133.

Côté, M. P. (2014). Copula-based risk aggregation modelling.

Deng, S. J., Jiang, W., & Xia, Z. (2002). Alternative statistical specifications of commodity price distribution with fat tails. *Advanced Modeling and Optimization*, *4(2)*, 1-8.

FASSIER, J. (2014). Quantification of Operational Risks using a Scenario Based Approach. Université Paris-Dauphine, Paris.

Godeiro, L. (2013). Estimating the VaR (Value-at-Risk) of Brazilian stock portfolios via GARCH family models and via Monte Carlo Simulation. Available at SSRN 2309659.

Guegan, D., & Jouad, F. (2012). Aggregation of Market Risks using Pair-Copulas.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Univ. Paris 8*, 229–231.

Skoglund, J., Erdman, D., & Chen, W. (2013). A mixed approach to risk aggregation using hierarchical copulas. *Journal of risk management in financial institutions*, *6(2)*, 188-205.

Tan, K., & Chu, M. (2012). Estimation of portfolio return and value at risk using a class of Gaussian mixture distributions. *The International Journal of Business and Finance Research*, *6(1)*, 97-107.

Yoshiba, T. (2013). Risk aggregation by a copula with a stressed condition *(No. 13-E-12)*. Bank of Japan.

Estimation of moments and density of first passage times by lower and upper risk thresholds

Nuno M. Brites[1]

[1] *ISEG/UL - Universidade de Lisboa, Department of Mathematics; REM - Research in Economics and Mathematics, CEMAPRE, Portugal, nbrites@iseg.ulisboa.pt*

Abstract

In random varying environments, we can describe the evolution of a fished population size using stochastic differential equations. Based on general expressions for the mean and standard deviation of first passage times by lower and upper risk thresholds, we compute such values for the particular case of the logistic model, considering several lower and upper risk threshold values. For a fixed risk value, we also present a way to estimate, by numerical inversion of its Laplace transform, the probability density function of the time to reach the risk thresholds.

Keywords

Risk threshold, First passage times, Stochastic differential equations.

Assessment of Allostatic Load and Periodontal Disease relationship using a Fuzzy Machine Learning method

Pereira, J.A.[1,2,3], **Abreu, F.**[1], **Mendes, L.**[1], **Oliveira, T.**[2,3]

[1] *Universidade do Porto, FMDUP, Portugal, up.241045@up.pt*
[2] *CEAUL, Lisboa, Portugal*
[3] *Universidade Aberta, DCeT, Portugal, teresa.oliveira@uab.pt*

Abstract

Purpose: Allostatic load (AL) refers to the cumulative burden of chronic stress and life events and results in multi-systemic physiological dysregulation. Periodontal Disease (PD) has been associated with various systemic diseases and is impacted by metabolic dysregulation. The aim of this research is to assess the relationship between AL and PD using unsupervised machine learning fuzzy methods.

Methods and Results: Data from the National Health and Nutrition Examination Survey (NHANES) 2011 were used. AL was measured using eleven surrogate biomarkers representing cardiovascular, inflammatory, and metabolic system functioning. A total of 1414 US adults aged 35 years and older were allocated to two fuzzy clusters, using an unsupervised machine learning classification method, the fuzzy k-means clustering algorithm. The cluster 1 presented more advantageous values for the allostatic load surrogate biomarkers. In both clusters, the membership degrees (MD) varied from 0.5 and 1.0, with an average of 0.7. The PD parameters' were compared between both clusters using GAMLSS models, yielding statistically significant differences ($p < 0.05$) for pocket probing depth (PPD) mean and maximum and clinical attachment loss (CAL) rate. The correlation coefficients between PD parameters and cluster 1 MD's ranged from -0.06 and -0.11, being statistically significant. The association of PPD mean with cluster 1 MD's statistical significance did hold up after adjustment for age and gender.

Conclusion: The latent nature of AL together with the absence of an universally accepted AL score poses major difficulties when classification of individuals is needed and to correlate their allostatic burden with other conditions. We propose to tackle this issue using fuzzy clustering methods in combinations with GAMLSS models. This approach allowed us to find an association between AL and PD by measuring individuals AL through the

membership grade to a cluster that's hold after adjustment for age and gender. This methodology appears to be promising to deal with variables that results from a complex combination of surrogate endpoints, and which aggregation is difficult or impossible.

Keywords

Allostatic load, Periodontal disease, Machine learning, Fuzzy logic.

References

Guidi, J.; Lucente M.; Sonino N.; Fava G.A.(2021). Allostatic Load and Its Impact on Health: A Systematic Review. *Psychother. Psychosom.*, *90*, 11–27.

Kim, J.; Amar, S. (2006). Periodontal disease and systemic conditions: a bidirectional relationship.*Odontology*, *94*(1), 10–21.

Chaudhuri, A. (2019). *Fuzzy Machine Learning: Advanced Approaches to Solve Optimization Problems*. De Gruyter.

State-space modeling for improving short-term forecast of meteorological time series: a comparative study

A. Manuela Gonçalves[1,2], **F. Catarina Pereira**[2,3], **Marco Costa**[4,5]

[1] *University of Minho, Department of Mathematics, Portugal, mneves@math.uminho.pt*

[2] *Centre of Mathematics, University of Minho, Portugal*

[3] *Centre of Mathematics, University of Minho, Portugal, id9976@alunos.uminho.pt*

[4] *University of Aveiro, Águeda School of Technology and Management - ESTGA, Portugal*

[5] *Centre for Research and Development in Mathematics and Applications - CIDMA, University of Aveiro, Portugal, marco@ua.pt*

Abstract

This study is developed within the scope of the TO CHAIR - Optimum Challenges in Irrigation project and aims to estimate and predict water losses by evapotranspiration in the context of finding technical solutions to improve the efficiency of daily water use in irrigation systems. In this study, a class of models called calibration models, which admit a state space representation associated with the Kalman filter, were proposed to study meteorological time series. Besides allowing dealing with time series with unstable behaviour (which is a predominant characteristic of environmental data), these models are quite effective from a stochastic point of view.

The state space models have the versatility to incorporate unobserved components such as trends, cycles, and seasonality. Since these models incorporate an unobservable component, the state, it needs to be estimated. For that, the Kalman filter is applied, which is a recursive estimation algorithm that allows both obtaining the optimal estimator of the state vector (based on the information available up to time t) and 1- step-ahead predictions by updating and improving the predictions of the state vector in real time when new observations become available. The linear regression model is a particular case of state space models and has been the most applied approach when a forecasting model is needed. The formulation of a problem in the state space representation seeks to evidence a functional, dynamic, and stochastic dependency between components of a system.

In this study, the aim is to calibrate the short-term forecasts in real time (obtained from the weatherstack.com website) through the state space modeling to improve the accuracy of the initial h-steps ahead, considering a six-day temporal window. In particular, a comparison of the two forecasting methods for the meteorological time series (daily maximum temperature) it is presented: the state space representation associated with the Kalman filter and the linear regression models. The statistical analysis was performed using the dataset that includes both the observations of daily maximum temperature (°C) in farm Senhora da Ribeira in Bragança, Portugal, between February 20 and October 11, 2019, and the forecasts obtained from the weatherstack.com website.

Keywords

State space models, Kalman filter, Linear models, Forecasting, Temperature, TO CHAIR project.

Acknowledgements

This work has received funding from FEDER/COMPETE/NORTE 2020/PO %CI/FCT funds through grants UID/EEA/-00147/20 13/UID/IEEA/00147/ 00%6933-SYSTEC project and To CHAIR - POCI-01- 0145-FEDER-028247. A. Manuela Gonçalves was partially financed by Portuguese Funds through FCT (Fundaçāo para a Ciência e a Tecnologia) within the Projects UIDB/00013 /20%20 and UIDP/00013/2020 of CMAT-UM. Marco Costa was partially supported by The Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology (FCT - Fundaçāo para a Ciência e a Tecnologia), references UIDB/04106/2020 and UIDP/04106/2020. F. Catarina Pereira was financed by national funds through FCT (Fundaçāo para a Ciência e a Tecnologia) through the individual PhD research grant UI/BD/150967/2021 of CMAT-UM.

References

Gonçalves, A.M., Costa, M. (2013). Predicting seasonal and hydro-meteorological impact in environmental variables modelling via Kalman filtering. *Stoch Environ Res Risk Assess*, *27*(5), 1021–1038.

Costa, M., Pereira, F.C., Gonçalves, A.M. (2021). Improving Short-Term Fore- casts of Daily Maximum Temperature with the Kalman Filter with

GMM Estimation. Lecture Notes in Computer Science, 12952 LNCS, 552–562.

Gonçalves, A.M., Costa, C., Costa, M., Lopes, S.O., Pereira, R. (2021). Temperature Time Series Forecasting in the Optimal Challenges in Irrigation (TO CHAIR). *Computational Methods in Applied Sciences*, *55*, 423–435.

Gonçalves, A.M., Baturin, O., Costa, M. (2018.) Time Series Analysis by State Space Models Applied to a Water Quality Data in Portugal. AIP Conference Proceedings Vol. 1978, 470101-1 - 470101-4.

Shumway, R.H., Stoffer, D.S. (2017). Time Series Analysis and Its Applications: With R Examples. 4th edition. Springer.

Hyndman, R.J., Athanasopoulos, G. (2018). Forecasting: principles and practice. 2nd edition. OTexts: Melbourne, Australia.

OS05 - Exploring challenges in analyzing medical data across different study designs and settings, sponsored by the International Biometrical Society - Italian Region

Use of Sequential Multiple Assignment Randomized Trials (SMARTs) in oncology: a systematic review

Giulia Lorenzoni[1]*, Elisabetta Petracci[2]*, Emanuela Scarpi[2], Ileana Baldi[1], Dario Gregori[1], Oriana Nanni[2]

[1] *University of Padua, Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac Thoracic Vascular Sciences and Public Health, Italy, giulia.lorenzoni@unipd.it (G.L.); ileana.baldi@unipd.it (I.B.); dario.gregori@ubep.unipd.it (D.G.)*

[2] *IRCCS Istituto Romagnolo per lo Studio dei Tumori (IRST) "Dino Amadori", Unit of Biostatistics and Clinical Trials, Italy, elisabetta.petracci@irst.emr.it (E.P.); emanuela.scarpi@irst.emr.it (E.S.); oriana.nanni@irst.emr.it (O.N.)*

* Co-primary and co-presenters

Abstract

Sequential Multiple Assignments Randomized Trials (SMARTs) represent a study design where individuals are randomized multiple times at key prespecified decision points and on the basis of their treatment history or characteristics and behaviors or intermediate outcomes. Such designs aim to determine optimal treatment strategies, which are particularly relevant for managing many chronic diseases, including cancer. Chronic conditions often require a sequential approach whereby treatment is adapted and readapted over time in response to the changing state of the individual, which is the basic idea of personalized medicine, i.e., tailor treatments to the individual characteristics of each patient.The present work investigates the state-of-the-art SMART designs in oncology, focusing on the discrepancy between the available methodological approaches in the statistical literature and the procedures applied within clinical trials of solid tumors.

A systematic search in three electronic databases (PubMed, Embase, and CENTRAL - Cochrane Trial Registry) was conducted. No restrictions to publication date were applied, and only English-language publications were considered. Published protocols or results of SMART designs and registrations of SMART designs in clinical trial registries were considered eligible. To be included, the SMART design should be applied to solid tumors research, without restrictions on the intervention type. The criterion to identify SMART designs was the presence of ≥ 2 stages in which patients were

re-randomized to subsequent treatments according to a set of pre-specified decision rules. Conference proceedings, book chapters, systematic reviews, and meta-analyses were excluded, but they were checked for eligible papers. The review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines.

The search of the electronic databases resulted in the inclusion of 14,586 records. After duplicates removal, title/abstract and full-text screening, 33 records were included in the present systematic review. Fifteen were reports of trials' results, four were trials' protocols, and fourteen were trials' registrations. The study design was defined as SMART by only one out of fifteen trial reports. Conversely, except four, all study protocols and trial registrations defined the study design SMART. Only one study reported that determining the optimal treatment regimen was the study's primary objective. Furthermore, only six records, i.e., three reports of trials results and three protocols, employed statistical analyses techniques accounting for patients' characteristics and history through the trial to identify the optimal treatment strategy.

SMART designs in oncology are still limited, but the interest in such methods in solid tumors research is growing.However, study powering and analysis are mainly based on statistical approaches traditionally used in single-stage parallel trial designs, probably because the design and analyses of such trials are challenging, and no formal guidelines are available. Further research in this field is needed to allow for broader use of such designs in oncology.

Keywords

Sequential multiple assignments randomized trial, Dynamic treatment regimens, Multistage intervention strategies, Treatment policies, Individualized treatment sequences.

References

Kosorok, M.R., Moodie, E.E.M. (2016). Adaptive treatment strategies in practice: planning trials and analyzing data for personalized medicine *ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, V.*

Differential methylation regions associated with teen depression and early puberty

Roberta De Vito[1], **Isabella N Grabski**[2], **Barbara E Engelhardt**[3]

[1] *Brown University, Department of Biostatistics, Center for Statistical Sciences, and Data Science Initiative, US, roberta_devito@brown.edu*
[2] *Harvard T. H. Chan School of Public Health, Department of Biostatistics, US, isabellagrabski@g.harvard.edu*
[3] *Princeton University, Department of Computer Science, US, bee@princeton.edu*

Abstract

Background: The Fragile Families Child Wellbeing Study (FFCWS) is a longitudinal study collecting data from 4,898 children and their parents, the majority of whom were unmarried (Reichman et al., 2001).
Methods: We study DNA methylation regions, corresponding to participant ages 9 and 15. Our primary aim is to identify methylation associated with two phenotypes in adolescent health, i.e., early puberty and teen depression, from these two-time points. In our models, we consider counfounders to better identify the association between methylations.
Results: Firstly, we estimate differentially methylated regions (DMRs) associated with depression and early puberty. Secondly, we include interaction terms in our regression models to estimate DMRs associated with the interaction between these two conditions and age. This will allow understanding how age-related changes in methylation are influenced by depression or early puberty. We also identify methylation quantitative trait loci (meQTLs) using genotype data from the participants. Finally, we validate our results by replicating our meQTLs in data from the GoDMC study.
Conclusions: We will be able to identify methylation states associated with both depression and early puberty on low socioeconomic status participants, providing crucial results on how genetic regulation will impact adolescents with these two conditions.

Keywords

Methylation, Depression, Early puberty, Genetic interaction.

References

Reichman, N. E. and Teitler, J. O. and Garfinkel, I. and McLanahan, S. S. (2001). Fragile families: Sample and design. *Children and Youth Services Review*, *23*, 303–326.

Semiparametric approaches to investigate the functional form of a nonlinear relationship in FRAP data

Gioia Di Credico[1], **Valeria Edefonti**[2], **Elena Marcello**[3], **Silvia Pelucchi**[3], **Francesco Pauli**[1]

[1] *University of Trieste, Department of Economics, Business, Mathematics and Statistics "Bruno de Finetti", Italy, gioia.dicredico@deams.units.it (G.D.C.); francesco.pauli@deams.units.it (F.P.)*

[2] *University of Milan, Department of Clinical Sciences and Community Health, Italy, valeria.edefonti@unimi.it*

[3] *University of Milan, Department of Pharmacological and Biomolecular Sciences, Italy, elena.marcello@unimi.it (E.M.); silviapelucchi87@gmail.com (S.P.)*

Abstract

Fluorescence recovery after photobleaching (FRAP) is a method used to study the dynamics of actin in neuronal dendritic spines, responsible for synaptic transmission in vitro neuronal cultures. FRAP data provide recovery trajectories over time within a nested hierarchical structure; namely, our data consists of 64 dendritic spines, each providing 65 measures equally spaced in time over a period of 100 seconds; spines belong to 26 neurons, grouped in 6 cultures. The number of spines within neurons ranges from 1 to 4, and the number of neurons within cultures from 2 to 8.

Traditional approaches to FRAP assume an asymptotic exponential curve, or one phase association curve, to model the observed recovery pattern. The hierarchical data structure is usually not considered in the model procedure, and hypothesis testing is performed either at the spine or the neuron level, averaging the values of different spines within the same neuron. Nonlinear mixed-effects models provide a suitable statistical tool to estimate the curves handling at the same time the hierarchical data structure. Indeed, random effects are added to model the variability at culture, neuron and spine levels.

Our work examines the asymptotic exponential function choice, evaluating if more flexible approaches estimate curves similar to the assumed one. Therefore, we fit two mixed-effects models in which two spline functions replace the asymptotic exponential curve: one is a penalised linear spline, the other is a free knot linear spline without penalisation.

The three models differ in the parameterisation of the fixed and random effect structure, hence parameter estimates are not fully comparable. Notably, three parameters characterise the asymptotic regression function describing the starting point of the curve, the steepness and the fluorescence recovered in time. The linear spline function with penalisation has 11 equally-spaced internal knots (one every five observations), whereas the linear spline function with free knots and without penalisation has 2 knots. For ease of comparison, the random-effects structure of the three models assumes independent and normal-distributed random effects at the three nested levels. The asymptotic regression model includes random effects on the parameters describing the starting point and the fluorescence recovered in time, while the spline models have random intercepts.

Numerical comparisons among the three models are performed approximating the curve parameters estimates. The starting point is compared with the value of the spline functions at time 0, while the fluorescence recovery with the spline functions values at time 100. Lastly, to compare the curve steepness, we introduce the recovery half-time defined as the seconds required to get half of the maximal fluorescence recovery.

Graphic and numerical comparison of the fitted curves highlights a common functional form compatible with the asymptotic exponential curve.

Keywords

FRAP data, Mixed-effects models, Nonlinear models, Splines.

References

Mueller, F., Mazza, D., Stasevich, T. J. and McNally, J. G. (2010). FRAP and kinetic modeling in the analysis of nuclear protein dynamics: what do we really know? *Current opinion in cell biology, 22(3)*, 403–411.

Pinheiro, J. and Bates, D. (2006). Mixed-effects models in S and S-PLUS. Springer science & business media.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). Semiparametric regression. Cambridge university press.

OS06 - Statistical Modelling for Risk Evaluation in social and economic sciences

Exploiting students' inter-degree relocations to assess Italian universities' attractiveness

Ilaria Primerano[1], Francesco Santelli[2], Cristian Usala[3], Giancarlo Ragozini[2]

[1] *University of Salerno, Department of Political and Social Studies, Italy*
[2] *University of Naples Federico II, Department of Political Sciences, Italy*
[3] *University of Cagliari, Department of Social and Political Sciences , Italy*

Abstract

This contribution investigates undergraduate students' attitudes to change the field of study and/or university during their bachelor's careers. Inter-degree students relocation, also defined as churn decisions (La Rocca et al., 2022), can be seen as a measure of universities' attractiveness since it can be related to their ability to meet students' needs and expectations. A churn decision is observed when students enrolled in one university decide to enrol: *i)* at the same university but in a different field of study, *ii)* at another university in the same field, and *iii)* at another university in a different field. Recent studies highlight that students' educational choices, despite depending on their attributes (Impicciatore & Tosi, 2019), are heavily affected by universities' characteristics such as research quality (Bratti & Verzillo, 2019), financial aid (Pigini & Staffolani, 2016), and hosting areas' features (Giambona et al., 2017) by applying a wide range of approaches, from longitudinal analysis (Attanasio et al., 2019) to network analysis (Columbu et al., 2021).

Moving from this framework, the analysis exploits two main sources of data: MOBYSU.IT[1] and USTAT. MOBYSU.IT provides the administrative data regarding all students enrolled in an Italian university between the academic year 2010/2011 and 2016/2017. Specifically, we follow each student during her/his first career for up to four years to examine the variations in her/his chosen university and field of study to define the churn indicator. This information is used to assess the risk of churn associated with each university and field of study in Italy and to identify the main determinants of this phenomenon. At this aim, we exploit the information available in

[1]Data drawn from the Italian 'Anagrafe Nazionale della Formazione Superiore' has been processed according to the research project 'From high school to the job market: analysis of the university careers and the university North-South mobility' carried out by the University of Palermo (head of the research program), the Italian 'Ministero Università e Ricerca', and INVALSI.

the open-data portal USTAT regarding several characteristics of Italian universities: the number of professors by field of study, the amount of tuition fees revenues, and the services aimed to reduce students' costs of attendance (e.g. grants). Moreover, we combine these data with information on hosting areas' attributes such as geographical position, unemployment rate, and the population of students in the area.

This information is used in an integrated methodological approach. First, we define a network in which the combination of universities and fields of study are the nodes, and the churn flows are defined as the number of students that change nodes within their first career. Second, the network centrality measures computed are summarized by applying a Principal Component Analysis to evaluate nodes' attitudes towards students' churn behaviour (i.e. whether nodes are net exporters or importers). Finally, the resulting score of the first dimension is regressed against a set of universities' and hosting areas' determinants to assess the role of these characteristics on nodes' attractiveness. This approach allows us to understand the role played by each node in the network and to understand which dimensions of universities' supply are more related to students' churn behaviour and universities' attractiveness.

Keywords

Mobility Flows, Social Network Analysis, University Churn.

References

Bratti, M., & Verzillo, S. (2019). The 'gravity' of quality: research quality and the attractiveness of universities in Italy. *Regional Studies*, *53*(10), 1385–1396.

Impicciatore, R., & Tosi, F. (2019). Student mobility in Italy: The increasing role of family background during the expansion of higher education supply. Research in Social Stratification and Mobility, 62(June), 100409.

La Rocca, M., Niglio, M., & Restaino, M. (2022). Predicting university students' churn risk, In Book of short Papers IES 2022 Innovation & Society 5.0: statistical and economic methodologies for quality assessment. Rosaria Lombardo, Ida Camminatello, Violetta Simonacci, PKE srl.

Pigini, C., & Staffolani, S. (2016). Beyond participation: do the cost and quality of higher education shape the enrollment composition? The case of Italy. *Higher Education*, *71*(1), 119–142.

Giambona, F., Porcu, M., & Sulis, I. (2017). Students Mobility: Assessing the Determinants of Attractiveness Across Competing Territorial Areas. *Social Indicators Research*, *133*(3), 1105–1132.

Attanasio, M., Enea, M., & Albano, A. (2019). Dalla triennale alla magistrale: continua la 'fuga dei cervelli' dal mezzogiorno d'Italia?. Neodemos, ISSN: 2421-3209.

Columbu, S., Porcu, M., Primerano, I., Sulis, I., Vitale, M.P. (2021). Geography of Italian student mobility: A network analysis approach. *Socio-Economic Planning Sciences*, *73*, 100918.

A bibliometric analysis on credit risk and business failure: foundations and global trends

Kristijan Breznik[1], **Giuseppe Giordano**[2], **Marialuisa Restaino**[3]

[3] *University of Salerno, Salerno, Italy, mlrestaino@unisa.it*
[2] *University of Salerno, Salerno, Italy, ggiordano@unisa.it*
[1] *International School for Social and Business Studies, Celje, Slovenia, kristijan.breznik@mfdps.si*

Abstract

Bibliometric analysis as a subdiscipline of scientometrics represents an effective method to explore the emergence and development of the analyzed scientific field and credit risk and business failure prediction models are no exception. Some previous works on similar topics are those by Alaka et al. (2018), Merigó and Yang (2017), Shi and Li (2019), Zambrano Farias et al. (2021).

This study presents the analysis of scientific publications that investigate the methods and techniques used for studying credit risk and business failure and for evaluating researches about both topics. Therefore, in the current study statistical methods and techniques used for credit risk and business failure analyses are identified and described. Thus, thanks to this bibliometric analysis, it will be possible to address the research trends on both topics and to assess their progress over time and across countries.

Our bibliometric study analyses the time period from the first papers published on these topics to the end of 2021. Data are collected from the Web of Science (WoS) database, one of the larger database in academic research. More specifically, research papers in WoS Core Collection consisting of the terms: Credit, Business, Failure, Risk, Prediction and Models, and their possible combinations were collected. We used some restrictions regarding the documents, e.g. only English language, only journal articles etc., to ensure the quality of the analyzed material.

This analysis uses quantitative techniques to analyze the academic production through citations, co-citations, autorship, co-autorship, keywords and journals as well as through bibliographic distribution, growth and evolution. In particular, some mathematical and statistical tools are used to find publication and authors patterns, through relational networks. Thus, a map

of the conceptual structure and evolution of the performance management is drawn in both of them for addressing scholars in positioning their future research work.

In more detail, the aims of the current study can be summarized in three main points. The first would describes how these areas of research are organized and progressed in terms of publications, authors, and journals, and identify some trends of both topics not only by looking at co-autorship, geographical area of authors, co-citation, co-occurrence, and also by analyzing the abstracts and keywords through text mining. Then, the second point is to present an overview study bringing together research work classified in business, finance, and management fields, so as to have a multidisciplinary research in prediction modelling. Finally, based on results obtained in the first two steps, the third point is to discuss the under-explored areas and reflect on possible future research opportunities to gain some insights and understanding of the research topics.

Keywords

Bibliometrics, Business Failure, Credit risks, Global trends.

References

Alaka, H.A., Oyedel, L.O., Owolabi, H.A., Kumar, V., Ajayi, S.O., Akinade, O.O., and Bilal, M. (2018): Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems with Applications*, *94*, 164–184.

Merigó J.M., Yang J.-B. (2017). Accounting Rsearch: a Bibliometric Analysis. *Australian Accounting Review*, *27* (80), 71–100.

Shi, Y., and Li, X. (2019). A bibliometric study on intelligent techniques of bankruptcy prediction for corporate firms. *Heliyon*, *5*, (12).

Zambrano Farias, F., Valls Martínez, M.D., and Martín-Cervantes, P.A. (2021). Explanatory Factors of Business Failure: Literature Review and Global Trends, *Sustainability*, *13* (18).

Exploring the relational patterns among Italian Firms

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

[1] *University of Salerno, Department of Political and Social Studies, Italy, iprimerano@unisa.it*

[2] *University of Salerno, Department of Economics and Statistics, Italy, mlrestaino@unisa.it*

Abstract

In this contribution we provide a network analysis of Italian firms relations defined according to the positions hold by managers in companies and also by considering their belonging to an Italian Industrial District. The importance of inter-firms networking has been widely addressed in the related literature. The definition of a network system made up of nodes and links is considered very powerful for business and finance to better investigate the main patterns and the type of occurring exchanges. Several empirical studies emphasize the importance of network perspective investigating, for example, the structure of firms ownership revealing a concentration of power among a small number of shareholders controlling many firms (Engel et al., 2021); and knowledge relationships' pattens among firms within an Italian Industrial district demonstrating that the position of a firm in these networks also affects its innovative performance (Boschma et al., 2007).

Industrial districts represent a manufacturing system characterized by a strong industrial specialization whose production is targeted to a specific production sector. Traditionally, these agglomerates were made up of small and tiny industries with solid relationships. Becattini (1991) defines the industrial district as a socio-territorial entity, characterized by the active presence of both a community of people and a population of firms in a natural and historical bounded area, developing the so-called *belong feeling*. This definition underlines the districts communities' tendency to identify themselves with the district. Furthermore, since firms located in industrial districts are always looking for new forms of cooperation, the study of the relationships within industrial district is a very relevant feature for the Italian economy and society. These entities are even more important for the Italian economy since they represented about one-fourth of the Italian economic system (Istat, 2015).

In this work, we consider the Orbis database, which provides detailed information about more than million of firms worldwide. In particular, Orbis ownership data holds information about the management composition of each firm over time. At our aim, we extract data about all small and medium-sized Italian manufacturing enterprises (with less than 250 employees and NACE codes 03-C), active in 2020 and operating over 10 years. We retrieve also information about all current and previous directors and manager, such as gender, age, job title, and nationality. The resulting firms are then grouped into several Industrial Districts according to the NACE codes and the geographical position. In doing so, we refer to the definition given by Istat based on different criteria, mainly referring to firms' Local Market Areas (LMA) of origin, economic specialization, and dimension. Thus, an industrial district is identified where there is a predominantly manufacturing LMA consisting of small and medium-sized enterprises.

Moving from this framework, we aim to read our data into the Social Network Analysis (SNA) framework (Wasserman, 1994). We define a inter-firm network consisting of a set of nodes represented by the Italian firms, and a set of weighted edges linking firms according to the role occupied by each manager in the firms. The presence of the same manager in two or more firms defines a link between the firms. At the same time, the weight depends on the number of different managers that each pair of firms has in common. In addition, since we have specified the firms' belonging to specific industrial districts, we will have different networks for each district. SNA centrality measures allow to explore the role and the position of each firm within the network defined at Italian and district level. Furthermore, the presence of relationships between companies belonging to different districts will highlights exchanges among them and may provide some insights for district governance.

Keywords

Industrial District, Inter-firms relationships, Social Network Analysis.

References

Becattini, G. (1991). Italian industrial districts: problems and perspectives. *International Studies of Management & Organization*, *21*(1), 83–90.

Boschma, R.A., Ter Wal, A.L.J., (2007). Knowledge Networks and Innovative Performance in an Industrial District: the Case of a Footwear District in the South of Italy. *Industry & Innovation*, *14*(2), 177–199.

Engel, J., Nardo, M., Rancan, M. (2021). Network Analysis for Economics and Finance: An Application to Firm Ownership. In: Consoli, S., Reforgiato Recupero, D., Saisana, M. (eds) Data Science for Economics and Finance. Springer, Cham., 331–355.

Istat (2015). 9 Censimento dell'Industria e dei Servizi e Censimento delle Istituzioni Non Profit: I Distretti Industriali 2011. ISBN 978-88-458-1859-2.

Wasserman, S., Faust, K. (1994). Social network analysis: Methods and applications. Cambridge Univ. Press, Cambridge, U.K.

OS07 - Modeling risks of neoplasia: Chance, environment and genes

Modeling risk of lung cancer: Coordination of molecular events and growth characteristics of tumors

Marek Kimmel[1]

[1] *Rice University, Departments of Statistics and Bioengineering, USA, kimmel@rice.edu*

Abstract

Lung cancer is a deadly disease still claiming around 200,000 lives per year in the United States alone. This talk summarizes research on quantitative understanding of progression of lung cancer which was carried out for past 30 years by a group mainly at Rice University, MD Anderson Cancer Center and Baylor College of Medicine. I will first review statistical concepts of screening for early detection of cancer. They culminated in design of a major clinical trial that demonstrated, in 2011, the ca. 20% mortality reduction potentially achievable by screening for lung cancer using computed tomography.

However, earlier than this, mathematical modeling predicted a reduction of similar magnitude. This was accomplished by careful studies of incidence and mortality, by researchers from all over the world, which then informed mathematical models capable of predicting the long-term effects of public health interventions. Building such models involves splitting the natural course of disease into an early phase at which the tumor is growing slowly and remains localized, so that removal may lead to cure, and later phases with faster growth and local and distant aggression, which might be targeted by systemic therapies, but with less certainty of cure.

Due to recent progress in DNA sequencing, including single-cell sequencing, insights can be gained into the timing of the waves of mutations and other genome transformations, which leave trace in the cancer cell genomes. Analysis based on probabilistic models of genetics can help estimate the relative rates of evolution of different clones and hence the relative durations of phases corresponding to small vs. large and slowly vs. fast growing tumors. I will review some of these estimates and discuss their potential importance.

Credits to:

Xing Chen, Shenyan Medical University, PRC; Khanh Ngoc Dinh, Columbia University, USA; Ivan Gorlov, Baylor College of Medicine, USA; Olga Gorlova, Baylor College of Medicine, USA; Roman Jaksik, Silesian University of Technology, Poland; Andrew Koval, Rice University, USA; Amaury Lambert, Sorbonne Universités, France; Simon Tavaré, Columbia University, USA

Keywords

Lung cancer, Screening, Early detection, Mathematical modeling, DNA sequencing, Site-frequency spectrum.

Estimation of timing of past events in cancer, based on DNA sequencing data

Andrew Koval[1], **Khanh Dinh**[2], **Marek Kimmel**[3]

[1] *Rice University, Department of Statistics, USA, alk3@rice.edu*
[2] *Columbia University, Department of Statistics, USA, knd2127@columbia.edu*
[3] *Rice University, Department of Statistics, USA, kimmel@rice.edu*

Abstract

A malignant tumor is often composed of classes (i.e. subclones) of cells that compete for scarce resources in the tumor micro-environment. Previous works have used the variant allele frequency (VAF) distribution from bulk sequencing data to estimate birth times and fitness advantages of these subclones, as well as the overall mutation rate of a tumor. Here, we use some additional theory from population genetics, a set of neutral equations, and the site frequency spectrum (SFS) of bulk sequencing data to estimate the growth rates, mutation rates, and birth times of each subclone. We simulate tumors from a branching process that explicitly incorporates these parameters, and then perform a heuristic search of the space of the neutral components of each subclone to estimate these parameters. We show that incorporating estimates of the neutral and selective mutation components of each subclone can allow us, on average, to accurately predict these subclonal parameters. We show similar results when simulating single-cell sequencing data based on estimates of the neutral and selective mutations from the single-cell data directly.

Keywords

Cancer evolution, Site frequency spectrum, Tumor heterogeneity, Clonal selection, Bulk sequencing.

References

Dinh, Khanh N., et al. (2020). Statistical inference for the evolutionary history of cancer genomes. *Statistical Science*, *35*(1), 129-144.

Integration of clinical and radiomic features for prediction of metastases risk in lung cancer

Agata Wilk[1,4], **Emilia Kozłowska**[1], **Damian Borys**[1], **Andrea D'Amico**[2] ,**Iwona Dębosz-Suwińska**[2], **Rafał Suwiński**[3], **Krzysztof Fujarewicz**[1], **Andrzej Świerniak**[1]

[1] *Silesian University of Technology, Department of Systems Biology and Engineering, Poland*

[2] *Department of Radiotherapy, M. Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch, Gliwice, Poland*

[3] *The 2nd Radiotherapy and Chemotherapy Clinic, M. Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch, Gliwice, Poland*

[4] Department of Biostatistics and Bioinformatics, M. Sklodowska-Curie National Research Institute of Oncology Gliwice Branch, Poland

Abstract

Motivation. Lung cancer is the leading cause of cancer-related death worldwide, responsible for 11 % of all cases and 18 % deaths. Such lethality is associated with a high risk of metastasis (primarily to bones, brain, liver, or another lung), which is one of the most significant negative prognostic factors in cancer due to the vast majority of secondary cancers being treatment-resistant. Thus, metastasis is a turning point in cancer therapy, signifying the transition from curative to palliative care. Prediction of risk of metastasis before its occurrence could provide an opportunity to adjust and intensify treatment. Thus, an effective predictive model would become a valuable aid in therapy planning.

Several attempts have been made to predict metastasis in lung cancer based on clinical characteristics, as well as molecular and imaging data. The latter shows particular promise since medical imaging, usually PET/CT, with relatively low invasiveness and easy acquisition, is currently a routine diagnostic procedure. This allows for model construction and potential application without additional tests.

Data. Data was collected retrospectively at Maria Sklodowska-Curie National Research Institute of Oncology (NRIO), Gliwice Branch, from patients treated for non-small-cell lung cancer between 2009 and 2017. From a cohort of over 800 patients, we selected non-metastatic patients for whom PET/CT

images and clinical data (age, sex, Zubrod score, tumor location, TNM classification, and histopathology) were available, resulting in a total of 125 patients included in this analysis.

Methods. Regions of interest (ROIs) were manually created for the images, with between 1 and 11 ROIs per patient. For each of the 319 ROIs, 105 radiomic features were extracted using pyRadiomics Python package: first-order features, shape features, and higher-order statistics texture features, including Grey-Level Co-occurrence Matrix (GLCM), Grey-Level Dependency Matrix (GLDM), Grey-Level Run Length Matrix (GLRLM), Grey-Level Size Zone Matrix (GLSZM) and Neighboring Grey Tone Difference Matrix (NGTDM).

Cox regression was used to model metastasis-free survival in two scenarios — based on all available ROIs and only the largest one per patient. We tested several feature selection methods: filtering based on univariate analysis, stepwise elimination, and LASSO.

Results. Generally, LASSO allowed for the most effective model reduction, producing models with the comparable predictive ability and fewer variables than the other approaches. In both scenarios, a model integrating radiomic and clinical features performed better (c-index 0.78) than either clinical-only (c-index 0.68) or radiomic-only (c-index 0.71). The final models were similar for both scenarios with regard to predictors, differing slightly in their statistical significance. The statistically significant features for all ROIs were: spread to lymph nodes (N in TNM classification), histopathology, minimum (first-order), normalized run-length non-uniformity (GLRLM), and small area low grey level emphasis (GLSZM). Also for the single-ROI scenario texture features displayed high significance, in particular normalized size zone non-uniformity (GLSZM).

Conclusion. PET/CT imaging data are routinely collected during diagnostics, which makes them perfect candidates as a source of prognostic signatures. Indeed, radiomic features have a high potential for predicting time to metastasis in non-small cell lung cancer. Moreover, adding clinical characteristics to the model improves its performance.

Acknowledgments

We would like to acknowledge financial support of the National Science Center, Poland - grant number 2020/37/B/ST6/01959.

Keywords

Lung cancer, Metastasis, Survival analysis, Radiomics.

Modeling and simulation of cancer evolution in single cells

Khanh N Dinh[1], **Ignacio Vázquéz-García**[2], **Simon Tavaré**[3]

[1] *Irving Institute for Cancer Dynamics, Columbia University, knd2127@columbia.edu*
[2] *Memorial Sloan Kettering Cancer Center, vazquezi@mskcc.org*
[3] *Irving Institute for Cancer Dynamics, Columbia University, st3193@columbia.edu*

Abstract

Recent advances in single-cell whole genome sequencing enable profiling of copy number aberrations at high resolution in thousands of cells. Single-cell genomics data from these technologies has enabled quantitative measurements of tumor dynamics, and measurements of the rate of chromosomal aneuploidy, whole-genome duplications and replication errors in tumors.

We have developed a simulation algorithm for studying single-cell dynamics in a population of cells, incorporating somatic copy number changes, clonal selection of driver mutations and accumulation of neutral passenger mutations. The simulator follows population dynamics as input by the user, generates the clonal evolution forward in time, where clones are defined by their copy number and driver mutation profiles. The phylogeny of a sample is then computed backward in time. The algorithm is designed to be efficient for large cell populations while maintaining statistical accuracy.

We present two examples from the simulator package. The first follows the neutral evolution of copy number events in the population of epithelial cells in the fallopian tube. The second investigates the evolution of high-grade serous ovarian cancer (HGSOC) driven by genomic instability. The simulator may also be used to calibrate clonal reconstruction algorithms used on single-cell DNA sequencing data.

Keywords

Single-cell DNA sequencing, Simulation algorithm, Copy-number event.

References

Laks, E., McPherson, A., Zahn, H., Lai, D., Steif, A., Brimhall, J., . . . , Shah, S. P. (2019). Clonal decomposition and DNA replication states defined by scaled single-cell genome sequencing. *Cell*, *179*(5), 1207–1221.

Dinh, K. N., Jaksik, R., Kimmel, M., Lambert, A., Tavaré, S. (2020). Statistical inference for the evolutionary history of cancer genomes. *Statistical Science, 35*(1), 129–144.

OS08 - Statistics in Modelling

Risk Analysis in Practice and Theory

Christos P. Kitsos[1]

[1] *University of West Attica, Department of Informatics and Computer Engineering, Greece, xkitsos@uniwa.gr*

Abstract

At the early stage risk was involving to political or military games for a decision making with the minimum risk. The pioneering work of Quincy Wright (1964) on the study of war was devoted to this line of thought. The Mathematics and Statistics involved, could be considered in our day as low-level. Wright (1964) applied eventually the differential equation theory with a successful application. Almost two decades later, Megill (1984) in his first edition of the book on Risk Analysis for Economical Data was emphasising that the fundamental in RA was to isolate the involved variables. Still the Statistics background was not too high. But the adoption of the triangle distribution was essentially useful. The triangle distribution has been faced under a different statistical background recently, but still the triangle obtained from the mode, the minimum value and the maximum value of the data can be proved very useful, as a special case of trapezoidal distributions. In Biostatistics and in particular to Risk Analysis for the Cancer problem, the evolution of the Statistical applications can be considered in the over 1000 references in Edler and Kitsos (2005). The development of methods and the application of particular probabilistic models and statistical analyses appear on extended development after 1980. Recently, Stochastic Carcinogenesis Models, Dose Response Models on Modeling Lung Cancer Screening are medical ideas with a strong statistical insight which have been adopted by the scientific community.

The cumulative distribution function (cdf) of the $GN(\mu, \sigma^2; \gamma)$ is:

$$\Phi_G(x) = 1 - \frac{\Gamma(\gamma_0, \gamma_0 z^{\frac{1}{\gamma_0}})}{2\Gamma(\gamma_0)}, \quad \gamma_0 = \frac{\gamma - 1}{\gamma}, \quad \gamma \in \mathbb{R} - [0, 1]$$

In this line of thought Kitsos and Toulias (2015) as well as Toulias and Kitsos (2018) worked on the Generalised Normal Distribution $GN(\mu, \sigma^2; \gamma)$ with $\gamma \in \mathbf{R} - [0, 1]$ being an extra shape parameter.

Moreover most of the evaluated measures, Fisher's Entropy information measure, $J(X)$ say, entropy power $N_p(X)$, Shannon entropy, depend on γ, namely on $\gamma_0 = \frac{\gamma-1}{\gamma}$. therefore a generalisation of RA due to a new general Fisher's information measure was considered as well as a General Normal Distribution.
Under this foundation the commutative hazard function, $H(\cdot)$ say, of a random variable $X \sim GN(\mu, \sigma^2; \gamma)$ can be proved equal to

$$H(x) = -\log \frac{\Gamma(\gamma_0, \gamma_0 z^{\frac{1}{\gamma_0}})}{2\Gamma(\gamma_0)} = -\log \frac{A(\gamma_0, z)}{2\Gamma(\gamma_0)}\,,\ x > \mu$$

$$H(x) = -\log(1 - \frac{\Gamma(\gamma_0, \gamma_0 |z|^{\frac{1}{\gamma_0}})}{2\Gamma(\gamma_0)}) = -\log(1 - \frac{A(\gamma_0, |z|)}{2\Gamma(\gamma_0)})\,,\ x \leq \mu$$

$$z = \frac{x-\mu}{\sigma}$$

where the definition of $A(\gamma_0, z)$ is obvious.
Example: As $\gamma \to \pm\infty$ the Generalised Normal Distribution tends to Laplace, $L(\mu, \sigma)$. Then it can be proved that:

$$H(x) = \log(2 + \frac{x-\mu}{\sigma})\,,\ x > \mu$$

$$H(x) = \log(1 - \frac{1}{2} e^{\frac{x-\mu}{\sigma}})\,,\ x \leq \mu$$

Moreover for the future lifetime rv X_0 at point x_0, $X \sim GN(\mu, \sigma^2; \gamma)$ the density function (d·f), the c.df can be evaluated and the corresponding expected future lifetime is

$$E(X_0) = \frac{2(x-\mu_0)\Gamma(\gamma_0)}{A(\gamma_0, z)}$$

The above mentioned results, among others, provide evidence to discuss, that the theoritical inside is moving faster than the application needed such results. These comments need special consideration and further analysis.

Keywords

Risk Analysis, Hazard function, Generalised Normal Distribution.

References

Edler, L., & Kitsos, C. (Eds.). (2005). Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment, Wiley, Chichester, U.K.

Megill, R. E. (1984). An introduction to risk analysis. Penn. Well. Pub. Co.

Kitsos, C. P., Toulias, T. L. (2015). Generalized Information Criteria for the Best Logit Model. In Theory and Practice of Risk Assessment (pp. 3-20), by Oliveira, T. Kitsos, C. Rigas, A. Gulati, S. (Eds). Springer, Cham.

Toulias, T. L., Kitsos, C. P. (2018). Hazard Rate and Future Lifetime for the Generalized Normal Distribution. In Recent Studies on Risk Analysis and Statistical Modeling (pp. 165-180), by Oliveira, T. Kitsos, C. Oliveira, A. Grilo, L. (Eds.) Springer, Cham.

Wright, Q. (1964). A Study of War. University of Chicago Press.

On fractal based cancer risk assessment

Milan Stehlík[1,2]

[1] *Johannes Kepler University in Linz, Department of Applied Statistics & Linz Institute of Technology, Austria, milan.stehlik@jku.at*
[2] *Universidad de Valparaíso, Institute of Statistics, Chile, milan.stehlik@uv.cl*

Abstract

We consider discrimination between mammary cancer and mastopathy tissues, which plays a crucial role in clinical practice. Based on my recent projects, I will discuss results how stochastic processes can model the both tissue modalities, mammary cancer and mastopathies. I will use exponential families with geometric-topological canonical parameters, like Euler characteristic. Application of such discrimination techniques has been applied to real biopsy images from patients in German clinics with reasonable improvement of false positive rates. Such approach can be considered as stochastic and statistic modelling for image processing which can provide good benchmarks for false-positive rates in most typical cancers assessments. Pure machine learning techniques can benefits from these obtain results and thus contribute to addressing "inherent data ambiguity" in medical imaging. The inter-patient variability of fractal dimension for mammary cancer is high and therefore multi-fractality can serve as a better concept. This is a feasible and parsimonious solution for a more delicate problem of multi-objective aggregation of information, relevant in modeling clinical assessor intervariability. Such problem is well visible if one considers several fractal and texture characteristics of biopsy images. This helps to refine a fractal cancer hypothesis, well applicable for several standard cancers, e.g. mammary or prostate cancer. In this talk I will also introduce a very different type of cancer growth which is not related to fractal cancer hypothesis. Nephroblastoma (given by Wilms tumor) is the typical tumor of the kidneys appearing in childhood, which does not satisfy fractal cancer hypothesis. I will illustrate by recent pre/post clinical study the effect on invasive treatments to Euclidean volumes of such tumors. A special transfer function "SPOCU" for neural network support for image processing has been also developed.

Keywords

Mastopathy, Mammary cancer, Discrimination, SPOCU, Frontality, Stochastic geometry.

References

Hermann, P., Mrkvicka, T., Mattfeldt, T., Minarova, M., Helisova, K., Nicolis, O., Wartner, F., and Stehlík M. (2015), Fractal and stochastic geometry inference for breast cancer: a case study with random fractal models and Quermass-interaction process, *Statistics in Medicine*, *34*, 2636–2661.

Kiselak, J., Lu, Y. , Svihra, J. , Szepe, P., and Stehlík M. (2021) "SPOCU": scaled polynomial constant unit activation function, *Neural Computing and Applications*, *33*, 3385–3401.

On k-Markovian Stochastic Models

Jerzy K. Filus[1], **Lidia Z. Filus**[2]

[1] *Oakton College, Dept. of Mathematics and Computer Science, USA, jkfilus98@gmail.com*

[2] *Northeastern Illinois University, Dept. of Mathematics, USA, L-Filus@neiu.edu*

Abstract

We present two analytic ways of construction of k-Markovian stochastic processes with discrete time. These models are special cases of more general "processes with long memory". Recall, the typically applied Markovian processes catch dependence of modeled phenomena at a given time from the most recent history only, with no regard to what had happened at any time moments earlier than the nearest past. Unlike that, using the triangular transformation or parameter dependence method it is possible to construct nice and useful stochastic processes with arbitrarily old history involved. Those may stand as models of realities with significantly higher accuracy (better fit to data) than the Markovian with their usually strongly limited history. Their drawback, however, is that often too many parameters must be estimated, and their number increases as time goes on. This would require samples with huge amounts of data that often are not available. A remedy for that is to limit the history within the models. This can be done either by introducing the "forgetting factors" of the parameters that quickly decay in time or by limiting the considered amount of past events to the last k ($k = 1, 2, \dots$) time moments. Applying the last paradigm we arrive at the notion of k-Markovian stochastic processes. If $k = 1$ then the k-Markovian process reduces to an ordinary Markovian.

In this presentation we intend to introduce this kind of models for any k. Description of k-Markovian processes involves "k-conditional probability distributions" (see (3)). We will present two methods. The first is the so-called triangular (in particular pseudoaffine) transformation method. Suppose, given any infinite sequence of independent random variables $T_1, T_2, \dots$ each having the probability density $f_j(t_j; j)$, where each j is a scalar or a vector parameter. Applying for $n = 2, 3, \dots$ a triangular transformation $\mathbb{R}^n \to \mathbb{R}^n$ to any random vector $(T_1, T_2, \dots, T_n)$ we obtain as output, the

random vector $(X_1, X_2, \ldots, X_n)$. This procedure is recurrently repeated for $n = 2, 3, \ldots$ with $n \to \infty$.

Recall, a triangular transformation $\mathbb{R}^n \to \mathbb{R}^n$ is characterized by the fact that the underlying Jacoby matrix is triangular which results in a simple form of the jacobians. A particular case of a triangular transformation is the pseudoaffine transformation defined as follows: $X_1 = T_1$, and for $j = 2, 3, \ldots, n$ we have:

$$X_j = \Phi(X_1, X_2, \ldots, X_{J-1})T_j + \Psi(X_1, X_2, \ldots, X_{J-1}), \tag{2.1}$$

where $\Phi(X_1, X_2, \ldots, X_{J-1})$ and $\Psi(X_1, X_2, \ldots, X_{J-1})$ are arbitrary continuous functions with $\Phi(X_1, X_2, \ldots, X_{J-1})$ never being zero.

As a result, for all $n = 2, 3, \ldots$, one obtains the joint probability density of each random vector $(X_1, X_2, \ldots, X_n)$ in the following product form:

$$g(x_1, x_2, \ldots, x_n) = g_1(x_1)g_2(x_2|x_1) \ldots g_n(x_n|x_1, \ldots, x_{n-1}), \tag{2.2}$$

where for $j = 2, \ldots, n$ we have:

$$g_j(x_j|x_1, \ldots, x_{j-1}) = f_j(x_j; \theta_j * (x_1, \ldots, x_{j-1}))$$

with the continues functions $\theta_j * (x_1, \ldots, x_{j-1})$ determined by the functions $\Phi(X_1, X_2, \ldots, X_{J-1})$ and $(x_1, \ldots, x_{J-1})$ from (1) in place of the former numerical values of the parameters θ_j. Realize, this procedure results in creating the stochastic process $\{X_n\}$ $n = 1, 2, \ldots, \infty$, with a "full memory".

This fact is not very convenient in applications. The memory is too long and one has to reduce it. To do this we limit the history at each time epoch $n = k + 1, k + 2, \ldots$ $(k = 1, 2, \ldots)$ to the last k time moments by imposing the condition:

$$X_n = \Phi(X_{n-k}, X_{n-k+1}, \ldots, X_{n-1})T_n + \Psi(X_{n-k}, X_{n-k+1}, \ldots, X_{n-1}),$$

where the functions $\Phi(\cdot)$ and $\Psi(\cdot)$ are assumed not to depend on the variables $X_1, \ldots, X_{n-k+1}$.

The foregoing assumption results in the fact that in (2) for each $j = k + 1, k + 2, \ldots, n > k$ we have

$$g_j(x_j|x_{j-k}, x_{j-k+1}, \ldots, x_{j-1}), \tag{2.3}$$

instead of the former $g_j(x_j|x_1, x_{j-1}, \ldots, x_{j-1})$.

That means k-conditional probability density of any X_j , as considered above, do not depend on realizations of the random variables $X_1, \ldots, X_{j-k-1}$.

The joint densities (2) (but now with the factors $g_j(x_j|x_{j-k}, x_{j-k+1}, \ldots, x_{j-1})$ instead of $g_j(x_j|x_1, x_{j-1}, \ldots, x_{j-1})$, which determine the k-Markovian stochastic process $\{X_n\}$, can also be obtained by the method of parameter dependence in rather simple way. Namely, it is enough to replace the numerical parameters θ_j of the densities of the elements of the original sequence $\{T_n\}$ by continuous (only) functions $\theta_j * (x_{j-k}, x_{j-k+1}, \ldots, x_{j-1})$ which in this second procedure stands for the stochastic model. This, obviously, is a simpler method. However, the triangular transformation such as (1) [in its k-Markovian version] is very useful in simulation problems. Namely, using the triangular transformations one can easily sample realizations of the variables $X_1, X_2, \ldots$ based on a simpler sampling from the independent random variables $T_1, T_2, \ldots$.

Keywords

Stochastic processes, Construction, Triangular transformation method, Parameter dependence method, Application, Simulation.

Berkson's paradox, biased sampling and weighted distributions

P. Economou[1], **A. Batsidis**[2], **G. Tzavelas**[3], **S. Malefaki**[4] **and P. Alexopoulos**[5]

[1] *Department of Civil Engineering, University of Patras, Greece, peconom@upatras.gr*

[2] *Department of Mathematics, University of Ioannina, Greece, abatsidis@uoi.gr*

[3] *Department of Statistics and Insurance Science, University of Piraeus, Greece, tzafor@unipi.gr*

[4] *Department of Mechanical Engineering and Aeronautics , University of Patras, Greece, smalefaki@upatras.gr*

[5] *Department of Psychiatry, Faculty of Medicine, University of Patras, Greece, panos.alexopoulos@upatras.gr*

Abstract

It is not unusual to observe in a sample a false positive or negative correlation between two random variables, which are either not correlated or correlated with a different direction. This is known as Berkson's paradox and is one of the most famous paradox in probability and statistics theory. Berkson's paradox may occur due to the over- or under-representation of individuals with specific properties in the sample. In the current work, the concept of weighted distributions is utilized to describe Berskon's paradox and a proper procedure, based on Approximate Bayesian Computation methods, is suggested to make inference for the population based on a biased sample which possesses all the characteristics of Berkson's paradox. Moreover, we focus on revealing the contribution of each variable to the bias in the bivariate sample by comparing the fit of the models obtained by using different weight functions based on the Deviance Information Criterion. The proposed method is illustrated to real data sets.

Keywords

ABC rejection algorithm, Bias adjustment, Model comparison, Weight functions.

References

P. Economou, A. Batsidis, G. Tzavelas, S. Malefaki (2021). Understanding the sampling bias: A case study on NBA drafts, *Journal of Statistical Theory and Practice* , *15*, Article number: 45.

P. Economou, A. Batsidis, G. Tzavelas, P. Alexopoulos and ADNI (2020). Berkson's paradox and weighted distributions: An application to Alzheimer's disease, *Biometrical Journal*, *62*, 238-249.

OS09 - Extreme Value Analysis in Weather Events and the Environment

Extreme quantile estimation based on the tail single-index model

Wen Xu[1], **Huixia Judy Wang**[2] **and Deyuan Li**[3]

[1] *Talent Today, China, wendy@jinrirencai.com*
[2] *Department of Statistics, George Washington University, USA, judywang@gwu.edu*
[3] *Department of Statistics, Fudan University, China, deyuanli@fudan.edu.cn*

Abstract

It is important to quantify and predict rare events that have significant societal effects. Existing works on analyzing such events rely mainly on either inflexible parametric models or nonparametric models that are subject to the "curse of dimensionality." We propose a new semiparametric approach based on the tail single-index model to obtain a better balance between model flexibility and parsimony. The procedure involves three steps. First, we obtain a $\sqrt{n}$-estimator of the index parameter. Next, we apply the local polynomial regression to estimate the intermediate conditional quantiles. Lastly, these quantiles are extrapolated to the tails to estimate the extreme conditional quantiles. We establish the asymptotic properties of the proposed estimators. Furthermore, we demonstrate using a simulation and an analysis of Los Angeles mortality and air pollution data that the proposed method is easy to compute and leads to more stable and accurate estimations than those of alternative methods.

Keywords

Extreme quantile, Local linear regression, Semi-parametric, Single-index, Tail.

Hydrologic designs for extremes under nonstationarity

Jayantha Obeysekera[1], **Jose Salas**[2]

[1] *Director and Research Professor, Sea Level Solutions Center, Institute of Environment, Florida International University*
[2] *Emeritus Professor, Colorado State University*

Abstract

Probabilistic, statistical, and stochastic methods, as well as physical and numerical models, are commonly used for designing and assessing water infrastructure such as spillways. The classical statistical techniques have assumed that hydrological processes such as precipitation, sea levels, and flow, evolve in an environment where inputs and outputs of the underlying hydrological cycle are stationary over time. It has now become increasingly evident that, in many areas of the world, the foregoing assumptions are no longer applicable, due to the effects of various anthropogenic and climatic induced stressors that cause changes on the environment, i.e. a nonstationary environment. Land-use change due to urbanization and deforestation, natural variability such as climatic cycles, and climate change are among the factors that may cause such nonstationarity. Recent research efforts have led to the development of a new paradigm that includes techniques useful in situations where there is good evidence of significant changes and

This paper will focus on developments of nonstationary counterparts of continuous and discrete distributions that have been typically applied in engineering practice under stationarity. They include new methods for determining return period, risk, reliability, and the temporal frequency of extreme events in a changing environment. The major emphasis is on extreme events of precipitation, floods, and sea levels. The paper presents several alternative techniques proposed in the field, their applications, and points out some of the key challenges ahead in their future development and application.

Keywords

Hydrologic Design, Return Period, Risk, Nonstationarity.

An Analysis of Trends in the Occurrence of Extreme Hurricane Events in the Atlantic

Florence George[1], **Sneh Gulati**[2], **BM Golam Kibria**[3], **Anu Simon**[4]

[1] *Florida International University, Math&Statistics, USA, fgeorge@fiu.edu*
[2] *Florida International University, Math&Statistics, USA, gulati@fiu.edu*
[3] *Florida International University, Math&Statistics, USA, kibriage@fiu.edu*
[4] *NOAA, USA, anu.simon@noaa.gov*

Abstract

South Florida has always been popular as a tourist destination and a place to live. Covid-19 has made it even more popular with people moving to the state in record numbers – it is after all the sunshine state! However, we are also the state that is vulnerable to hurricanes and to sea-level rise. Hurricanes threaten the coast of Florida six months of the year and seem to be getting bigger every year. Perhaps policy changes in development along the coast and stronger building codes could serve to avoid the catastrophic caused by powerful hurricanes that cover a large swath of the state. However before any changes can be proposed, it is imperative that we investigate the risk from such events and quantify it. In this paper, we will investigate whether the odds of observing extreme hurricanes in the Atlantic are increasing and if the hurricanes are getting bigger. To examine this hypothesis, we will be using logistic regression model for our two response variables y1 and y2; where y1 is the type of hurricane based on intensity (extreme vs non-extreme) and y2 is the type of hurricane based on its radius of maximum winds (extreme or not extreme). Based on available data, the covariates to be considered will be recent/past (recent 50 years vs prior to that), sea surface temperature, LaNina/ El Nino year, annual rainfall, amongst others.

Keywords

Atlantic Hurricane, Extreme events, Logistic Regression.

References

Jing, B., Qian, Z., Zareipour, H. Pei, Y. and Wanf, A. (2021). Wind Turbine Power Curve Modelling with Logistic Functions Based on Quantile Regression, *Applied Sciences*, *11*, 3048, https://doi.org/10.3390/app11073048.

Srivastava, N. (2005). A logistic regression model for predicting the occurrence of intense geomagnetic storms. *Annales Geophysicae*, *23*, 2969–2974.

Kovacs, J. M., Blanco-Correa, M. and Flores-Verdugo, F. (2001). A Logistic Regression Model of Hurricane Impacts in a Mangrove Forest of the Mexican Pacific. *Journal of Coastal Research*, *17* (1), 30-37.

Villanueva, D., Feijóo, A. (2018). Comparison of logistic functions for modeling wind turbine power curves. *Electr. Power Syst. Res.*, 155, 281–288.

Optimal pooling and distributed inference for the tail index and extreme quantiles

Abdelaati Daouia[1], **Simone A. Padoan**[2], **Gilles Stupfler**[3]

[1] *Toulouse School of Economics, University of Toulouse Capitole, France,* abdelaati.daouia@tse-fr.eu

[2] *Department of Decision Sciences, Bocconi University of Milan, Italy,* simone.padoan@unibocconi.it

[3] *Ensai & CREST, France,* gilles.stupfler@ensai.fr

Abstract

This paper investigates pooling strategies for tail index and extreme quantile estimation from heavy-tailed data. To fully exploit the information contained in several samples, we present general weighted pooled Hill estimators of the tail index and weighted pooled Weissman estimators of extreme quantiles calculated through a nonstandard geometric averaging scheme. We develop their large-sample asymptotic theory across a fixed number of samples, covering the general framework of heterogeneous sample sizes with different and asymptotically dependent distributions. Our results include optimal choices of pooling weights based on asymptotic variance and MSE minimization. In the important application of distributed inference, we prove that the variance-optimal distributed estimators are asymptotically equivalent to the benchmark Hill and Weissman estimators based on the unfeasible combination of subsamples, while the AMSE-optimal distributed estimators enjoy a smaller AMSE than the benchmarks in the case of large bias. We consider additional scenarios where the number of subsamples grows with the total sample size and effective subsample sizes can be low. We extend our methodology to handle serial dependence and the presence of covariates. Simulations confirm the statistical inferential theory of our pooled estimators. An application to rainfall data is discussed.

Keywords

Extreme values, Heavy tails, Distributed inference, Pooling, Testing.

References

Daouia, A., Padoan, S.A. and Stupfler, G. (2021). Optimal pooling and distributed inference for the tail index and extreme quantiles, arXiv:2111.03173.

OS10 - Statistical and Data Science Developments for Risk Assessment in Urban Areas

Road accident risk using complex networks: the impact of traffic

Gian Paolo Clemente[1], **Francesco Della Corte**[2], **Diego Zappa**[3]

[1] *Universita' Cattolica del Sacro Cuore, Dipartimento di Matematica per le Scienze economiche, finanziarie ed attuariali (DiMSEFA), Italy, gianpaolo.clemente@unicatt.it*

[2] *Universita' La Sapienza - Roma, Italy, francesco.dellacorte@uniroma1.it*

[3] *Universita' Cattolica del Sacro Cuore, Dipartimento di Scienze Statistiche, Italy, diego.zappa@unicatt.it*

Abstract

The assessment of risk related to car crashes in road networks is a relevant topic for both social and political decisions and for insurance companies. To this end, we show how the spatial objects and the information concerning the structure of the roads, that can be collected e.g. from open data sources, along with the crash history can be used to map the risk related to each road. In particular, we follow a combined approach. On the one hand, a statistical model is developed in order to assess the risk on the basis of a set of features related to the characteristics of the streets. On the other hand, from the spatial object we build a weighted network, where vertices and arcs correspond to geographical elements as junctions and roads respectively and where the assessed risk of each segment is used as a weight. We study the topology structure of the graph obtained and we show how classical network indicators can provide meaningful insights about the risk of an area.

To achieve our aim, we need to adapt the current methodology about geospatial modelling to the constraints derived from the maps of the roads of a particular area and to exploit supervised/unsupervised statistical learning algorithms to estimate the local risk of the frequency of accidents (and potentially of the severity). We do not consider here (actually a research in progress) other features that can be detected by telematic data or by adding other data sources (e.g., driving behaviour, driving habits, KM coverage, daytime, weather conditions, etc.).

To consider spatial dependence, spatial models with errors in variables are compared with CAR/SLM/SLX models using distances among edges computed using the minimum path in the directed graph based on the road

network. Moreover different stochastic mechanisms that govern the occurrence of accidents are assumed (as a Poisson or a Negative Binomial random variable). The rich and highly computational intensive experimental framework is due to the need of fitting the model partitioning the overall domain into sub-areas. The aim is indeed to get the model that assures the best adaptation to the peculiar details of an area and, at the same time, is not affected too much by details and characteristics observed far from the spatial domain of interest.

An application refers to Milan (city and province). To apply the proposed approach, we have to overcome a bit of bias in the data, that is somehow unavoidable in this type of research. For instance, an important feature, represented by the number of road crossings, is not directly available in the databases and, therefore, it is computed as the number of segments that have in common one coordinate with that road. This method represents an approximation of the true crossings but it returns an estimate quite close to reality. Moreover, coordinates of accidents are not always strictly in line with its segment. Misalignments are due to proxies implicit into the reverse geocoding algorithms or to errors in the registration of accident locations. In addition for some regions, some features are not available and thus not useful for model identification. The proposed model has been applied introducing the Vehicle Miles Traveled (VMT) as an offset. It is defined as the product of the length of the segment and the volume of traffic and it has the advantage to provide a measure of exposure at risk in each road.

The spatial object and the accident risk assessed by the model for each road are then converted in a directed and weighted graph. In particular, we focus on a "junction graph", where each segment is an arc and nodes are given by junctions (or by termination of closed streets). Each arc is then weighted according to the risk of the segment detected at previous step. Focusing on network topological indicators, we observe a significant correlation between the risk associated to a node and the node betweenness measured on the network. Therefore, the centrality of a node in the topological structure appears related to the risk measured by the model. Additionally, by means of the Louvain methodology, we detect communities in the area. The communities depend on both the arc density and on the weights. We find 287 communities considering the whole area of Milan (city and province). The split of the area into clusters can be used by insurance companies to measure the propensity to get an accidents in the neighbour of a point, and then to fine tune the cost

of premiums to be paid to drive a car.

Keywords

Risk of accidents, Complex networks, Geospatial models.

References

Borgoni, R., Gilardi, A., Zappa, D. (2020), Assessing the Risk of Car Crashes in Road Networks, *Social Indicators Research*.

Tufvesson, O. et al. (2019) Spatial statistical modelling of insurance risk: a spatial epidemiological approach to car insurance, *Scandinavian Actuarial Journal*, (6), 508-522.

A statistical approach for flood risk assessment using mobile phone traffic flows' data

Rodolfo Metulini[1], **Maurizio Carpita**[2]

[1] *University of Salerno, Department of Economics and Statistics, Via Giovanni Paolo II, 132, Fisciano, 84084, Italy. rmetulini@unisa.it*
[2] *University of Brescia, Department of Economics and Management, Contrada Santa Chiara, Brescia, 25122, Italy. maurizio.carpita@unibs.it*

Abstract

Flooding risk exposures maps, traditionally developed by using census and surveys data, have some limitations like high costs and static nature of the data (i.e., they assume amount of presences constant over time). Real-time monitoring and forecasting of people and vehicles mobility is thus a relevant aspect for metropolitan areas subjected to flooding risk, as crowding is a highly dynamic process in metropolitan cities. It is a matter of fact that typical "smart" cities present emerging forms of mobility and time-variations in the use of urban spaces, by both residents and temporary population. Recently, the advent of Information and Communication Technologies (ICT) allows growing availability of data to be used, e.g., in support to the optimization of traffic flows, in tracking real-time citizens' position or in collecting their opinions.

To obtain a dynamic measure for the exposure risk we make use of mobile phone network data, that have already been used in Balistrocchi et al. (2020) and Metulini and Carpita (2021) with the aim of producing dynamic information on people's presences for areas with hydrogeological criticality. More precisely, the methodological approach proposed in this work goes in the direction of developing a framework for an early warning detection of flood exposure risk associated to human presence and people mobility, in support of the broad topic of natural disaster management. We use mobile phone data of Telecom Italia Mobile (TIM) retrieving the Origin-Destination (OD) traffic flows from/to different census areas (ACE of ISTAT, the Italian National Statistical Institute) on hourly basis for twelve months (September 2020 – August 2021). Data are provided by Olivetti (www.olivetti.com/en/iot-big-data) and FasterNet (www.fasternet.it) for the MoSoRe@Unibs Project of Lombardy Region, Italy (https://ricerca2.unibs.it/?pageid=8548, CallHub ID 1180965; bit.ly/2X).

We aim to develop a statistical modelling approach to forecast traffic flows transiting through specific flooding risk areas. In such a way, it may be possible to identify in advances periods of higher than expected exposure risk.

Empirical evidences highlight the presence of a temporal pattern in the time series of OD flows in which a daily and a weekly seasonality are predominant. These evidences motivate the adoption of an Harmonic Dynamic Regression (HDR) for complex seasonality, based on approximate temporal periodicity with the use of Fourier basis (Hyndman and Athanasopoulos, 2018). Since we have reasons to believe that flows in the opposite directions may belong to two dependent processes, to allow such a flows to be contemporaneously correlated to each others and correlated to each others' past values, we include HDR terms into a Vector AutoRegressive with eXogenous covariates (VARX) modelling structure (Hamilton, 1994). A peculiarity of our proposal is that the model does not rely on including the immediately previous hours lag terms among covariates, in such a way a prediction of real-time traffic flows is feasible even if recent data are still not available.

An evaluation of the performance of the adopted model in terms of forecasting accuracy is made by means of an approach based on repeatedly splitting the sample on a training and a validation set. In such a way to be able to circumscribe the prediction accuracy, based on Mean Absolute Percentage Error (MAPE) we detect those year's days for which the performance of the method is not as good as in the rest of the days (i.e. the "outliers"). Finally we evaluate prediction performance by means of the Hit-Rate measure for confusion matrices, after having turned observed and predicted values of the validation set into categories using quintiles representing different levels of exposure risk ("very high", "high", "moderate", "low", "very low").

Possible future directions regard refining the data related to the areas interested by flood risk. A solution may be to match original traffic flows data with additional sources such as mobile phone presences or administrative and online sources (e.g. "OpenStreetMap").

Keywords

Mobility, Forecasting, Vector autoregressive model, Harmonic dynamic regression, Exposure risk.

References

Balistrocchi, M., Metulini, R., Carpita, M. and Ranzi, R. (2020). Dynamic

maps of human exposure to floods based on mobile phone data. *Natural Hazard and Earth System Sciences*, *20*, 3485–3500.

Hamilton, H. (1994), Time Series Analysis. Princeton, New Jersey: Princeton University Press.

Hyndman, R. J. and Athanasopoulos, G. (2018). Forecasting: principles and practice. 2nd ed. Melbourne, Australia: OTexts.

Metulini, R., Carpita, M. (2021). A Spatio-Temporal Indicator for City Users Based on Mobile Phone Signals and Administrative Data. *Social Indicator Research*, *156*, 761–781.

Natural risk assessment of Italian municipalities for residential insurance

Selene Perazzini[1], **Giorgio Gnecco**[2], **Fabio Pammolli**[3]

[1] *IMT Lucca, AXES Research Unit, Italy, selene.perazzini@alumni.imtlucca.it*
[2] *IMT Lucca, AXES Research Unit, Italy, giorgio.gnecco@imtlucca.it*
[3] *Polytechnic University of Milan, Department of Management, Economics, and Industrial Engineering, Italy, fabio.pammolli@polimi.it*

Abstract

The substantial lack of good quality data on losses hinders the assessment and prediction of the financial cost of natural hazards. In order to overcome this issue, insurers are increasingly relying on catastrophe risk models for premium rating and financial planning (Mithcell-Wallace et al., 2017). These models compute expected monetary losses by estimating and combining four fundamental components of risk: hazard, exposure, vulnerability, and loss. In this work we propose a preliminary catastrophe modeling approach to flood and earthquake risks assessment for residential buildings in Italy. This work aims at supporting governors in the definition of a natural risk management strategy to enhance the social and financial resilience of the country. In order to detect the critical areas of the territory, we consider the Italian municipalities and compute the expected losses of the country by aggregation. Moreover, our approach identifies the areas where the exposure strongly shapes the risk profile due to the high inhabited density or the presence of fragile building structural typologies. This information can be useful for urban planning purposes.

Since the two perils can be assumed independent, flood and earthquake risks are assessed separately. Earthquake losses are estimated using the model by Asprone et al. (2013), which we extend by applying the most recent seismic risk maps by INGV for the hazard module. For flood risk assessment, we combine hydrological information from the basin authorities in the "Aree vulnerate italiane" database by the Italian National Research Council (CNR) and the most recent flood risk maps for the hazard module. We use a selection of depth-percentage damage curves from the literature to estimate the vulnerability component. For both the perils, the database "Mappa dei

rischi dei comuni italiani" by ISTAT and information in Agenzia delle Entrate (2015) are used to represent the exposure. Since a set of hypotheses and parameters are needed in catastrophe risk modeling, the uncertainty of the estimate for each component is discussed.

We find that earthquakes in Italy generate annual expected losses approximately equal to 6234.66 million Euros, while flood expected losses amount to about 875.90 million Euros per year. Although earthquakes produce the highest expected losses at the national level, flood losses per square meter often exceed the corresponding earthquake ones. This happens because of the different extent of the areas exposed to the two perils: while almost all the Italian territory is exposed to earthquakes, floods affect a limited area.

Comparing municipal losses and losses per square meter allows to capture the different effects of hazard and exposure. In particular, our analysis shows that the highest earthquake losses per square meter are associated to sparsely inhabited municipalities in the central area of the Appennino mountain chain. This result reflects the high probability of earthquake occurrence in the area. However, the highest municipal expected losses correspond to densely populated cities on the coast. Indeed, the probability that a natural phenomenon will hit these cities is quite low, but their large population densities strongly affect their riskiness. As far as floods concern, Northern Italy is the most flood-prone area, and the highest expected losses per square meter are estimated around the Po river and correspond to municipalities in the Emilia-Romagna, Veneto and Lombardia regions. Most of these municipalities are densely inhabited and are therefore also associated to some of the highest municipal flood losses. In addition to those, high municipal expected losses are also estimated on the north-west coast, in north Sardinia and Rome.

To conclude, we propose an application of the models to the insurance sector. We assume risk aversion of the individuals and estimate the maximum premium that individuals are willing to pay for a full coverage residential policy covering the risk of either earthquakes or floods.

Keywords

Risk assessment, Catastrophe modeling, Earthquake, Flood, Italy.

References

Agenzia delle Entrate (2015). Gli immobili in Italia.

Asprone D., Jalayer F., Simonelli S., Acconcia A., Prota A., Manfredi G. (2013). Seismic insurance model for the Italian residential building stock. *Structural Safety*, *44*, 70-79.

Mitchell-Wallace K., Foote M., Hillier J., Jones M. (2017). Natural catastrophe risk management and modelling: a practitioner's guide. Wiley Blackwell.

Flood risk management using mobile phone data and hydrological modeling in a heavily urbanized area in Lombardy

Babak Razdar[1], **Rodolfo Metulini**[2], **Maurizio Carpita**[3], **Giorgio Paolo Maria Vassena**[1], **Roberto Ranzi**[1]

[1] *University of Brescia, Department of Civil, Environmental, Architectural Engineering and Mathematics (DICATAM), Brescia, Italy, babak.razdar@unibs.it; giorgio.vassena@unibs.it; roberto.ranzi@unibs.it*

[2] *University of Salerno, Department of Economics and Statistics (DISES), Salerno, Italy, rmetulini@unisa.it*

[3] *University of Brescia, Department of Economics and Management, Brescia, Italy, maurizio.carpita@unibs.it*

Abstract

Flood events are one of the natural disasters which cause significant social and economic impacts on human life as, when an area is flooded, many people need to evacuate to a safer place. This phenomena is unpredictable, fast and intense and can develop rapidly with little or no warning. A fast response in evacuating people is one of the main flood risk management issues. Maps of flood risk and exposure generally assume people density constant over time, despite this is not the case in the real world, as crowding is a highly dynamic process in urban areas. Monitoring people mobility is a relevant aspect for urban areas subjected to high risk of flooding. In this study, a combination of flood propagation modeling in three stage of time periods and a time series modeling strategy for traffic flows prediction based on Harmonic Dynamic Regression (HDR, Hyndman, Athanasopoulos, 2021) are performed to obtain dynamic maps of flood risk. To simulate flood propagation, obtained hydrograph from HEC-HMS hydrologic modeling system is utilized. Flood processes, especially in the small catchment, are mainly caused by intense localized precipitation and plane features. To fine describing the effects of land use on the basin response to storm rainfall, the elevation of 350 critical points in the flood plain area were obtained by GNSS technology, and a combination of DEM and field surveying in critical points to present topography is used (see Figure 2.1a). During the physical modeling of overland flows, the highly irregular microtopography is sometimes replaced by a smooth surface. This does not lead to significant differences in discharge hydrographs because the

continuity requirements are met in both cases (Tayfur et al. 1993), unless the micro topography is represented at a very fine scale: in this case all the local irregularities in the soil surface manifest in terms of vastly different slopes for neighboring nodes and mobile phone network data better suits with the aim of develop dynamic exposure to flood risk maps (as done in Balistrocchi et al., 2020). We use data on flows of Telecom Italia Mobile users among different census areas of the Mandolossa (an urbanized area close to Brescia), recorded hourly basis from September 2020 to August 2021 and data coming from Minimization of Drive Tests (MDT) 4G technology, representing the number of signals that mobile phones send to the operator in an interval of time. Both data are provided by Olivetti (www.olivetti.com/en/iot-big-data) and FasterNet (www.fasternet.it) for the MoSoRe Project 2020-2022 (Call-Hub ID 1180965; bit.ly/2Xh2Nfr, https://ricerca2.unibs.it/?page id=8548). People exposure obtained from mobile phone data and processed with the HDR model will be combined to flooding hazard maps at different storm return periods to estimate dynamic flood exposure risk maps (Figure 2.1b giving a provisional idea).

(a) (b)

Figure 2.1: (a) Flooded area and obtained points by the GNSS technology. (b) "Shp_SCE_MDT_22Nov": MDT signals in the interval 7:45am/8:00am (November 22^{nd}, 2021), aggregated at Sezione di CEnsimento (SCE). Amount of signals are expressed by colors' intensity.

Keywords

Flood risk management, Mobile phone data, Hydrological modeling, Flood propagation modeling.

References

Balistrocchi, M., Metulini, R, Carpita, M., Ranzi, R. (2020). Dynamic maps of human exposure to floods based on mobile phone data. *Natural Hazards and Earth System Sciences*, *20*, 3485–3500.

Hyndman, R. J., Athanasopoulos, G. (2021). Forecasting: principles and practice. 3rd edition, OTexts: Melbourne, Australia. OTexts.com.

Tayfur G., Kavvas M.L., Govindaraju R.S., Storm D.E. (1993). Applicability of St. Venant equations for two-dimensional overland flows over rough infiltrating surfaces. *Journal of Hydraulic Engineering*, *119*(1), 51-63.

OS11 - Statistical Modeling and Inference

Operations with iso-structured models with commutative orthogonal block structure: an introductory approach

Carla Santos[1], **Cristina Dias**[2], **Célia Nunes**[3], **João T. Mexia**[4]

[1] *Polytechnic Institute of Beja, Beja, and Center for Mathematics and Applications, New University of Lisbon, Lisbon, Portugal, carla.santos@ipbeja.pt*

[2] *Polytechnic Institute of Portalegre, Portalegre, and Center for Mathematics and Applications, New University of Lisbon, Lisbon, Portugal, cpsd@ipportalegre.pt*

[3] *Department of Mathematics and Center of Mathematics and Applications, University of Beira Interior, Covilhã, Portugal, celian@ubi.pt*

[4] *Department of Mathematics and Center for Mathematics and Applications, Nova University of Lisbon, Lisbon, Portugal, jtm@fct.unl.pt*

Abstract

Linear mixed models have gained prevalence as statistical tools in several fields, as biology, medical research, agriculture, genetics, or industry, due to their suitability in correlated data situations. A particular class of linear mixed models (models with commutative orthogonal block structure) is interesting for the possibility of obtaining least squares estimators giving best linear unbiased estimators for estimable vectors. Aiming at the joint analysis of models obtained independently, the operation of models joining, involving initial models with commutative orthogonal block structure, originates a new model with commutative orthogonal block structure, ensuring that the conditions for the good properties of the estimators are preserved. In this work we focus on the possibility of applying the operation of models joining in the circumstance that the initial models have identical space spanned by their mean vectors and having covariance matrices that are linear combinations of the same pairwise orthogonal projection matrices.

Keywords

Best linear unbiased estimator, Mixed models, Jordan algebra, Models joining.

Acknowledgments

This work was partially supported by national founds of FCT- Fundaçāo para a Ciência e Tecnologia (Foundation for Science and Technology), Portugal, under UIDB/00297/2020 and UIDB/00212/2020.

References

Jordan, P., Von Neumann, J., Wigner, E. (1934). On an algebraic generalization of the quantum mechanical formulation, *Annals of Math*, *35*(1).

Nelder, J.A. (1965a). The analysis of randomized experiments with orthogonal block structure I. Block structure and the null analysis of variance, *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, *283*, 147–162.

Nelder, J.A. (1965b). The analysis of randomized experiments with orthogonal block structure II. Treatment structure and the general analysis of variance, *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, *283*, 163–178.

Santos, C., Nunes, C., Dias, C., Mexia, J. T. (2017). Joining models with commutative orthogonal block structure. *Linear Algebra and its Applications*, *517*, 235–245.

Zmyślony, R. (1978). A characterization of best linear unbiased estimators in the general linear model, *Lecture Notes in Statistics*, *2*, 365–373.

A new approach for analyzing multi-environment trials (MET)

Cristina Dias[1], **Carla Santos**[2], **João T. Mexia**[3]

[1] *Polytechnic Institute of Portalegre, Portalegre, Department of Technologies and Center for Mathematics and Applications (CMA), New University of Lisbon, Lisbon, Portugal, cpsd@ipportalegre.pt*
[2] *Polytechnic Institute of Beja, Beja, Department of Mathematics and Center for Mathematics and Applications (CMA), New University of Lisbon, Lisbon, Portugal, carla.santos@ipbeja.pt*
[3] *New University of Lisbon, Department of Mathematics and Center for Mathematics and Applications (CMA), New University of Lisbon, Portugal, jtm@fct.unl.pt*

Abstract

This study presents a procedure for analyzing networks of randomized block designs often conducted in breeding studies to evaluate new varieties in different environments. These experiments are more frequently called multi-environment trails that consider the same set of varieties in different locations and years. By computing the largest eigenvalues and the corresponding eigenvectors of cross-product matrices (where we work with a randomized block design data matrix), the proposed method can extract the required information from the observed data. On the basis of the orthogonality of the eigenvectors, some F-statistics are then developed for evaluating the hypotheses of interest. The proposed method is illustrated using a wheat dataset.

Keywords

Randomized block designs, Networks, Eigenvalues, Cross Product-matrices.

Acknowledgments This work was partially supported by national founds of FCT-Foundation for Science and Technology, Portugal, under UIDB/00297 /2020 and UIDB/00212/2020.

References

Calinnski, T., Kageyama, S. (1996). Block Designs: A Randomization Approach. Vol. I. *Analysis, Lecture Note in Statistics* vol.150, Springer-Verlag, New York.

Belsley, D. A., E. Kuh and Welsch, R.E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, *John Wiley and Sons, New York.*

Dias, C., Santos, C., Varadinov, M. and J.T. Mexia, J. T. (2016). ANOVA like analysis for structured families of stochastic matrices. *12th International Conference of Computational Methods in Sciences and Engineering, AIP Conf. Proc.1790 (2016), pp.140005-140009.*

Modeling hospital patients no-show: An analysis of cluster dependence

Ana Borges[1], **Mariana Carvalho**[2]

[1] *CIICESI, ESTG, Politécnico do Porto, aib@estg.ipp.pt*
[2] *ESTG, Politécnico do Porto, mrc@estg.ipp.pt*

Abstract

Patients no-show (appointment absenteeism) can severely impact the health system and public health, leading directly to a waste on structural and financial resources, representing additional costs. It affects productivity indicators related to employees, equipment and office. The intensification of queues for procedures (a medium-term effect of patients no-show) may the delay in the provision of care representing a threat to health (Kalb, et al., 2012). Hence, it becomes urgent for hospital management services to understand the risk factors associated with patients no-show, in order to implement actions to mitigate it. In an extent literature review, Carreras-García et al. (2020) resumes and details the factors related to patients no-show tested by several authors in the last decade. Common procedures to predict patients no-shows are the traditional logistic regression methods, neural networks, Naive Bayes classifier and decision trees. The purpose of the present study is to compare the results of the traditional logistic approach with the results of a generalized estimating equations (GEE) analysis that includes the repeated measurements of the outcome of interest (patients' attendance), to find out whether the latter approach produces different parameter and standard error estimates. For the GEE, different "working" correlation structure identification were tested. To attend our main objective, both analytical approaches were applied to real data from a hospital located in the north of Portugal, to examine the relation between selected risk factors related to patient's no-show, extracted from the literature. Data consists of information regarding 61.522 consultations of 2001 patients of the Physical Medicine and Rehabilitation specialty between 03 November 2018 and 15 March 2020. Repeated measures for each patient's consultation, alongside the information if the patients have attended or not (no-show) were registered. For the model's specification the following variables were tested: gender, age, marital Status, month of consultation, weekday of consultation, waiting time between

consultation creation and its date, number of prior no-shows, distance between residence and hospital, mean temperature and season (winter, summer, spring and autumn). The GEE method (Liang & Zeger, 1986), an extension of the quasi-likelihood approach, is used to analyze longitudinal and other correlated data. The GEE logit model estimates a similarly model to the standard logistic regression. However, unlike in logistic regression, it allows for dependence within clusters (in this case the patients), such as in longitudinal data. Since we have repeated measures for patients the standard logistic regression could be insufficient since it does no captures a possible cluster dependence. The GEE model needs the specification of a "working" correlation matrix for the clusters. The "working" correlation matrix is a TxT matrix of correlations, where T is the size of the largest cluster and the elements of the matrix are correlations between within-cluster observations. In this analyze four "working" correlation within clusters were tested, namely: i) independence: the correlation matrix is the identity matrix, hence it assumes no correlation within the clusters, becoming the equivalent to the standard logistic regression; ii) exchangeable: assumes that the correlations are the same for all observations within the cluster; iii) unstructured: no constraints are placed on the correlation which will then be estimated from the data; iv) AR(1): assumes a non-stationary dependence of order 1 between observations. The choice of the appropriate "working" correlation structure was done by comparing the values of the Quasi-likelihood under Independence Model Criterion (QIC) proposed by Pan (2001), where the lowest value corresponds to a better model. Furthermore, the area under the receiver operating characteristic (ROC) curve was calculated in order to compare the accuracy in the distribution models. Results have shown that the GEE model with the independence correlation structure was the one with lower values of QIC and higher values of AUC, providing evidence that in the context of predicting patients no show the standard logistic regression is appropriated.

Keywords

Generalized estimating equations, Logistic model, Patients no-show.

Acknowledgments

This work has been supported by national funds through FCT - Fundação para a Ciência e Tecnologia through project UIDB/04728/2020.

References

Carreras-García, D., Delgado-Gómez, D., Llorente-Fernández, F., & Arribas-Gil, A. (2020). Patient no-show prediction: A systematic literature review. *Entropy*, *22*(6), 675.

Kalb, L. G., Freedman, B., Foster, C., Menon, D., Landa, R., Kishfy, L., & Law, P. (2012). Determinants of appointment absenteeism at an outpatient pediatric autism clinic. *Journal of Developmental & Behavioral Pediatrics*, *33*(9), 685-697.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13-22.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, *57*(1), 120-125.

Modeling psychosocial risks of workers using PLSc-SEM estimator

Luís M. Grilo[1,2,3], **Miguel Lopes**[3], **Vanda Lima**[3], **Aldina Correia**[3], **Ana Martins**[3]

[1] *Instituto Politécnico de Tomar (IPT), Tomar, and Centro de Matemática e Aplicações (CMA), FCT, Universidade Nova de Lisboa, Portugal, lgrilo@ipt.pt*

[2] *Centro de Investigação em Cidades Inteligentes (Ci2), IPT, Portugal*

[3] *CIICESI, ESTG, Politécnico do Porto, Portugal*

aml@estg.ipp.pt, vlima@estg.ipp.pt, aic@estg.ipp.pt, 816003@estg.ipp.pt

Abstract

To assess psychosocial risks of workers in a Portuguese industrial company the medium version of the Copenhagen Psychosocial Questionnaire (COPSOQ II), a reliable and internationally validated instrument with 76 questions (variables expressed in a Likert-type scale), was applied in a survey. Considering the latent constructs available in COPSOQ II and based on the specialized literature and also on some of the authors' experience, a theoretical path model was proposed. The Structural Equations Modeling (SEM) is an advanced and powerful multivariate statistical technique that has been used to deal with complex models in several scientific areas, namely in the social and health sciences, and the Partial least squares (PLS) is a variance-based SEM estimator that maximizes the explained variance of the endogenous constructs while it simultaneously relaxes the demands on data. Thus, with the sample collected from the company's workers about their perception of their health and well-being and using the consistent PLS (PLSc) version, that corrects for bias, to consistently estimate SEM's with common factors, an estimated model that meets all statistical criteria (for outer and inner submodels) was obtained. All statistically significant effects among the latent constructs were expected: the exogenous construct "justice" has a positive direct effect on "job satisfaction" and this, in turn, has a negative direct effect on "stress", which has a positive direct effect in the target construct "burnout"; the exogenous construct "quantitative demand" also has a positive direct effect on "stress" and, through this, an indirect effect on "burnout". Considering the model estimated with the full sample ($n = 268$), two submodels were also estimated by gender (categories: female and male), but a

multigroup analysis, to test whether the differences between both groups are statistically significant in the population, was not developed because one of the requirements of its application is that the sizes of the groups are similar, which is not the case in this study. However, in a comparative analysis, only at the sample level, and regarding the values of the path coefficients, the direct effect of "justice" on "job satisfaction" is the same for both gender categories, the direct effect of "quantitative demands" on "stress" is greater in men and the direct effect of "stress" on "burnout" is greater in women, but the biggest difference found between two groups was in the direct effect of "justice" on "job satisfaction", which is greater for women. As for the value of the coefficient of determination in the endogenous construct "burnout", it was 76.2% for women and 59.9% for the men's model. These results can help to better understand the phenomenon and, thus, lead to the development of prevention mechanisms to reduce absences from work due to illness, as well as lead to a better well-being of workers in their workplace and, consequently, to a higher increase in its productivity.

Keywords

Burnout syndrome, Latent constructs, Mental health, Multivariate statistics, Path model.

Acknowledgments This work was partially supported by national founds of FCT- Fundação para a Ciência e Tecnologia (Foundation for Science and Technology), Portugal, under UIDB/00297/2020 and UIDB/00212/2020.

References

Benitez, J., Henseler, J., Castillo, A. & Schuberth, F. (2020). How to perform and report an impactful analysis using partial least squares: Guidelines for confirmatory and explanatory IS research, *Information & Management*, *57* (2), 103168, ISSN 0378-7206. https://doi.org/10.1016/j.im.2019.05.003.

Dijkstra, T. & Henseler, J. (2015). Consistent partial least squares path modeling, *Mis Q.*, *39* (2), 297–316.

Dijkstra, T. & Henseler, J. (2015). Consistent and asymptotically normal PLS estimators for linear structural equations, *Comput. Stat. Data Anal.*, *81* (1), 10–3.

Grilo, L. M., Grilo, H. L. & Martire, E. (2018). SEM using PLS Approach to Assess Workers Burnout State. 14th International Conference of Computational Methods in Sciences and Engineering (ICCMSE 2018), *AIP Conf. Proc.* 2040, 110008-1-110008-5. https://doi.org/10.1063/1.5079172.

Hair, J. F., Hult, G. T. M., Ringle, C. M. & Sarstedt, M. (2017). *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM).* 2nd Ed., Thousand Oakes, CA: Sage.

Hair, J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M., Danks, N. P. & Ray, S. (2022). *Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R. A Workbook*, Springer.

Hair, J. F., Risher, J. J., Sarstedt, M. & Ringle, C. M. (2019). When to use and how to report the results of PLS SEM, *European Business Review*, 31, 1: pp. 2-24. https://doi.org/10.1108/EBR-11-2018-0203.

OS12 - New Perspective in Financial Risk Management

Extracting implied volatilities from bank bonds

Michele Leonardo Bianchi[1], **Gian Luca Tassinari**[2]

[1] *Bank of Italy, Financial Stability Directorate, Italy, micheleleonardo.bianchi@bancaditalia.it*
[2] *University of Bologna, Department of Economics, and Department of Management, Italy, gianluca.tassinari2@unibo.it*

Abstract

In this work we explore the information content of senior, subordinated and additional tier 1 (or contingent convertible) bonds issued by euro area banks. We analyze both the asset volatility implied in senior and subordinated bonds and credit default swap market spreads, and the CET1 ratio volatility extracted from additional tier 1 bonds secondary market spreads in the period from December 31, 2012 to March 31, 2021. Furthermore, we jointly consider the following important bank variables: asset, equity and CET1 ratio volatilities. In doing so, we can obtain the market view on credit spreads, banks balance sheet and capital ratio dynamics on a daily basis even if bank data are released quarterly. The approach can be used to monitor the risk of each bank, as perceived by the market, and to investigate banking fragility at a stand-alone or at a country level. Finally, we compare our estimated equity implied volatilities with the volatilities implied in equity option quotes and we show that this indicator depends on the model and the financial instruments considered in the calibration.

Keywords

Subordinated bonds, AT1 bonds, CoCo bonds, Credit default swaps, Capital requirements, CET1 ratio, Implied CET1 volatility, Firm value models.

An investigation into the incidence of time-varying quantiles in well-diversified portfolios

Fabrizio Cipollini[1], **Giampiero M. Gallo**[2], **Alessandro Palandri**[3]

[1] *DiSIA, Università di Firenze, Italy,* `fabrizio.cipollini@unifi.it`
[2] *Italian Court of Audits, and New York University in Florence, Italy,* `giampiero.gallo@nyu.edu`
[3] *DiSIA, Università di Firenze, Italy,* `alessandro.palandri@unifi.it`

Abstract

Moving from the customary conditional mean (μ_t) – standard deviation (σ_t) decomposition of the return r_t of a well-diversified portfolio, we model time-varying conditional quantiles of the standardized residuals $z_t = (r_t - \mu_t)/\sigma_t$: the ultimate goal is to investigate their incidence on the (possibly time-varying) conditional Value-at-Risk, say $\text{VaR}_t(\tau)$ for $\tau \in (0,1)$, of r_t.

We provide a new methodological framework and we propose several model specifications which we group in two categories: *Direct Dynamics* (those driven by lagged "actual" quantiles) and *Indirect Dynamics* (those driven by lagged empirical "rejection frequencies"). The Conditional Autoregressive VaR (CAViaR) popularized by Engle and Manganelli (2004) is a member of the latter group.

Estimation is done by minimizing the asymmetric Mean-Absolute Deviation ($\text{MAD}(\tau)$), similarly to Koenker and Bassett (1978) for quantile regression. Due to the discontinuity of this loss function, such minimization is performed by adaptive simulated annealing.

The proposed specifications are compared in the tracking of time-varying quantiles of simulated data (10^7 observations) and evaluated in terms of: 1) in sample RMSE between predicted (1-step ahead) and true quantiles; 2) in-sample $\text{MAD}(\tau)$ (distance between data and predicted quantiles); 3) in-sample coverage of the predicted quantiles in the full period; 4) in-sample coverage, again, but within sub-periods (for example by fiscal year). It emerges that: 1) Model specifications in the *Indirect Dynamics* category tend to track better the time-varying quantiles (lower RMSE between predicted and true quantiles); 2) despite being a good estimation criterion, $\text{MAD}(\tau)$ is not a reliable loss function for the evaluation of in-sample predictions and out-of-sample forecasts (its model rankings are completely different from those produced by RMSE); 3) coverage in the full period is not guaranteed (despite

the length of the time series) except for one model specification; 4) coverage in the subperiods is again better matched by *Indirect Dynamics* specifications but is not necessarily aligned with quantile tacking ability.

Motivated by such evidence, we propose a post-estimation correction procedure that we named *fail-safe*. The target of such a correction is to match the coverage in some reference time window (for example by fiscal year). The *fail-safe* adjustment does not require estimation of parameters and can be applied to any of the specifications proposed.

The empirical application analyzes the conditional quantiles of daily standardized returns of representative equity portfolios (25 Fama-French value-weighted portfolios and 19 European stock market value-weighted indices, both on 2010-2019) on a rolling window, 5-year period: the first 4 years are used for model fitting and the last one for out-of-sample, 1-step ahead evaluation. Preliminarily, we test for the in-sample presence of time-varying conditional quantiles by performing a signed-likelihood ratio test of independence. After estimation, we run forecasts backtesting based on the randomness of the hit sequences, the demeaned violations $I(r_t < \text{VaR}_t(\tau)) - \tau$ ($I(\cdot)$ is the indicator function) based on the estimated model. The main empirical results are that: 1) the in-sample hit sequences do not exhibit the ubiquitous violations of independence that are typical of conditional variances; 2) the *fail-safe* correction provides a marked improvement of the forecasts in terms of prediction coverage; 3) the lack of a suitable measure of prediction accuracy (in the sense of Hansen and Lunde, 2006) makes difficult to rank quantile forecasts.

Keywords

Nonlinear regression quantile, Risk management, Specification testing, VaR, CAViaR.

References

Engle, R. and Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles, *Journal of Business and Economic Statistics*, *22*, 367–381.

Hansen, P. R. and Lunde, A. (2006). Consistent ranking of volatility models, *Journal of Econometrics*, *131*, 97–121.

Koenker, R. and Bassett, G. (1978). Regression quantiles, *Econometrica*, *46*, 33–50.

OS13 - High-frequency data in economics and finance

A mixed-frequency combination approach to forecast covariance matrices of asset returns

M. Marchese[1], F. Di Iorio[2], M. Tamvakis[3], R. Payne[4]

[1] *City University of London, Bayes Business School (formerly Cass), UK, Malvina.Marchese@city.ac.uk*

[2] *University of Naples Federico II, Department of Political Sciences, Italy, fdiiorio@unina.it*

[3] *City University of London, Bayes Business School (formerly Cass), UK, m.tamvakis@city.ac.uk*

[4] *City University of London, Bayes Business School (formerly Cass), UK, richard.payne.1@city.ac.uk*

Abstract

To improve on the forecasting accuracy of conditional covariance matrices of asset returns, we propose a novel forecast combination approach based on mixed information from high and low frequency data. The combination strategy relies on an economic loss function based on portfolio selection. Specifically, it identifies the mixing weights using portfolio diversification optimality criteria. Our approach does not require a proxy for the latent conditional covariance matrix and facilitates the economic interpretation of the combination strategy for decision maker. Two empirical applications involving energy futures and S&P100ETs show that the proposed combination strategy leads to minimum variance portfolios with lower risk on an out-of-sample basis with respect to a number of alternative specifications based on pure statistical loss functions. The results suggest that low-frequency data improve volatility forecasting even when high frequency data is available at medium and long horizons.

Keywords

Forecasting accuracy, Conditional covariance matrices, Economic loss function, Portfolio selection.

References

Amendola, A.; Braione, M.; Candila , V.; Storti, G. (2020). A model confidence set approach to the combination of multivariate volatility forecasts, *International Journal of forecasting, 36*(3), 873–891.

Hansen, P.R.; Lunde, A. ; Nason, J.M. (2011). The model confidence set, *Econometrica*, *79*(2), 453–497.

Laurent, S.; Rombouts, J.V.K. ; Violante, F. (2012). On the forecasting accuracy of multivariate GARCH models, *Journal of Applied Econometrics*, *27* (6), 934–955.

A Structural Analysis of the Competitive Landscape of Brick-And-Mortar Pharmacies

Gail, Maximilian M.[1], **Götz, Georg**[2], **Herold, Daniel**[3], **Schäfer, Jan T.**[4]

[1–4] *Justus-Liebig-University Giessen, Licher Strasse 62, 35394 Giessen (Germany), Contact: maximilian.m.gail@wirtschaft.uni-giessen.de*

Abstract

This paper presents a structural model introduced in the seminal paper of Berry, Levinsohn & Pakes (1995) and implemented in a recent Python-Package by Conlon & Gortmaker (2020) that is used to study price competition for OTC-drugs among pharmacies. Data on OTC-sales of over one third of Germany's Brick-and-Mortar pharmacies is used to recover information on elasticities and markups for drugs that are used to treat infections of the upper respiratory tract. Preliminary analysis yields that margins in the federal states and across ZIP postal codes of Germany are quite different. This finding can be traced back to socio-economic factors.

Keywords

Pharmaceuticals, Structural models, Markups, Demand estimation.

References

Berry, S., Levinsohn, J., & Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, 841–890.

Conlon, C., & Gortmaker, J. (2020). Best practices for differentiated products demand estimation with pyblp. *The RAND Journal of Economics*, *51*(4), 1108–1161.

Leadership Communication and COVID-19 Vaccination Hesitancy

Phil-Adrian Klotz

[1] *Chair for Industrial Organization, Regulation and Antitrust, Department of Economics, Justus Liebig University Giessen. Licher Strasse 62, 35394 Giessen, Germany. E-mail: phil.a.klotz@wirtschaft.uni-giessen.de*

Abstract

This paper empirically analyzes the impact of leadership communication on the COVID-19 vaccination rate using a quasi-experimental design. Based on a speech by the President of France, Emmanuel Macron, we examine how political leaders can influence the willingness to get vaccinated of a country's citizens by transmitting scientific insights into a clear and vivid message as well as by threatening credibly with future restrictions for unvaccinated people. In a Difference-in-Differences (DiD) framework it is shown that a televised address of Macron has increased the vaccination rate in France by roughly 5%. We test the robustness of this result by applying an event study design. Our findings imply that leadership communication is an effective weapon to change the beliefs of unvaccinated citizens and to overcome COVID-19 vaccination hesitancy.

Keywords

COVID-19, Leadership communication, Vaccination, DiD.

Long and Short run dynamics in Realized Covariance Matrices: a Robust MIDAS Approach

Luca Scaffidi Domianello[1], **Edoardo Otranto**[2]

[1] *University of Messina, Department of Economics, Italy, e-mail: lscaffidi@unime.it*

[2] *University of Messina, Department of Economics, Italy, e-mail: eotranto@unime.it*

Abstract

Forecasting time-varying conditional (co)variances is an interesting research topic, due to the importance of asset returns correlation for financial applications: hedging, asset allocation, pricing, risk management, and so on.

Early multivariate volatility models (e.g. the BEKK of Engle and Kroner, 1995) were based on daily cross-product returns and assumed a constant average (or long-run) level of (co)variances, though empirical evidence suggests that it is time-varying (see, for example, the results in Gallo and Otranto, 2015, for the S&P500 volatility). In the last decade, a great deal of effort was put into the development of models based on Realized Covariance (see, for example, the Conditional Autoregressive Wishart (CAW) model of Golosnoy et al., 2012), modeling directly a nonparametric estimation of the covariance matrices, based on intra–daily returns.

A recent stream of literature is devoted to detect long and short–run components that characterize, with different dynamics, the realized covariance series (see, for example, Bauwens et al., 2016). By decomposing the covariance matrix into a short-run and a long-run component, it is possible to capture, in a parsimonious way, the long–memory behavior of (co)variances. The short-run component is aimed to capture daily fluctuations and transitory effects; conversely, the long-run component represents the average level that varies over time according to economic conditions. However, dynamic component models are based on the Cholesky decomposition, which makes the short-run component potentially sensible to asset order.

We propose a new additive component model belonging to the MIDAS family, in the spirit of Colacito et al. (2011), with features that help overcome some drawbacks:

- it does not depend on the Cholesky decomposition to the covariance matrix, so that the order of the series is not relevant in the estimation of the model parameters;
- the multiplicative decomposition of the covariance matrix, adopted in other models, requires the calculation at each time of the inverse of the Cholesky factor, thus slowing down the optimization algorithm. Our additive specification does not require this step, with a clear computational gain;
- multivariate volatility models, to overcome the *curse of dimensionality problem*, usually assume a scalar specification of the conditional (co)variances, imposing the same dynamics for each series. This hypothesis is very strong and not supported by empirical evidence. The model we propose adopts the Hadamard exponential function proposed by Bauwens and Otranto (2022), which allows asset-pair-specific and time-varying parameters. This specification offers a more flexible dynamics with only one more parameter than the baseline specification, thus preserving the parsimony of the model.

In the empirical analysis we fit our set of models to the Realized Covariance series of 9 assets belonging to the Dow Jones Industrial Average (DJIA) index. Then, we compare the in-sample fitting of the estimated models through some information criteria and statistical loss functions. Finally, we verify the superior forecasting performance of our model with respect to the competitive ones in terms of statistical and economic loss functions.

Keywords

Realized Covariance, MIDAS, Dynamic component models, Hadamard exponential parameterization.

References

Bauwens, L., Braione, M., and Storti, G. (2016). Forecasting comparison of long term component dynamic models for realized covariance matrices. *Annals of Economics and Statistics, (123/124)*, 103–134.

Bauwens, L., and Otranto, E. (2022). Modelling realized covariance matrices: A class of Hadamard exponential models. *Journal of Financial Econometrics*, forthcoming.

Colacito, R., Engle, R.F., and Ghysels, E. (2011). A component model for dynamic correlations. *Journal of Econometrics*, *164*(1), 45–59.

Engle, R.F., and Kroner, K.F. (1995). Multivariate simultaneous generalized ARCH. *Econometric Theory*, *11*(1), 122–150.

Gallo, G.M., and Otranto, E. (2015). Forecasting Realized Volatility with changing average volatility levels. *International Journal of Forecasting*, *31*, 620–634.

Golosnoy, V., Gribisch, B., and Liesenfeld, R. (2012). The conditional autoregressive Wishart model for multivariate stock market volatility. *Journal of Econometrics*, *167*(1), 211–223.

On the estimation of Value-at-Risk and Expected Shortfall at extreme levels

Emese Lazar[1], **Jingqi Pan**[2], **Shixuan Wang**[3]

[1] *University of Reading, Henley Business School, ICMA Centre, UK, e.lazar@icmacentre.ac.uk*

[2] *University of Reading, Henley Business School, ICMA Centre, UK, jingqi.pan@pgr.reading.ac.uk*

[3] *University of Reading, Department of Economics, UK, shixuan.wang@reading.ac.uk*

Abstract

The estimation of risk at extreme levels of significance (such as 0.1%) can be crucial to capture the losses during market downturns, such as the global financial crisis and the COVID-19 market crash. For many existing models, it is challenging to estimate risk at extreme levels of significance. In order to improve such estimations, we extend the one-factor GAS model and the hybrid GAS/GARCH model to estimate Value-at-Risk and Expected Shortfall for two levels of significance simultaneously, namely for an extreme level and for a more common level (such as 10%). Our simulation results indicate that the proposed models outperform the GAS model benchmarks in terms of in-sample and out-of-sample loss values, as well as backtest rejection rates. We apply the proposed models to oil futures (WTI, Brent, Gas oil and Heating oil) and compare them with a range of parametric, nonparametric and semiparametric alternatives. The results show that our proposed models are generally superior to the alternatives.

Keywords

Risk models, Value-at-Risk, Expected shortfall, Semiparametric model, Oil futures.

The effect of networking on bad loans: evidence from mutual banks

Carmelo Algeri[1], **Antonio Fabio Forgione**[2], **Carlo Migliardo**[3]

[1] *University of Messina, Economics, Italy, calgeri@unime.it*
[2] *University of Messina, Economics, Italy, fforgione@unime.it*
[3] *University of Messina, Economics, Italy, cmigliardo@unime.it*

Abstract

Since the territoriality of small cooperative banks is limited, network effects can prevail across the bad loans of mutual banks that operate in the same geographical district. It is worth noting that the credit recovery policies of local banks can have a cascading effect on the quality of neighboring banks' loan portfolios. The current study provides compelling evidence that bad loans are geographically contemporaneous and lagged dependent. The two spatial terms, in particular, have opposed effects: the contemporaneous spatial lag term has a direct effect, whereas the space-time autoregressive coefficient has a negative effect. In contrast to the contemporaneous effect, which can be attributed to business cycle fluctations, the time-spatial lag effect confirms the insight that neighboring financial institutions' credit recovery policies can have a detrimental effect on the credit recovery capacities of local banks. Finally, the empirical model depicts a significant relationship between the Lerner index and non-performing loans. This compelling evidence supports the hypothesis of competition-stability.

Keywords

Mutual banks, Spatial Dynamic Data Panel Model, Bad Loan Recovery, Bank market power.

References

Anselin, L. (2013). Spatial econometrics: methods and models. *Springer Science & Business Media*, v.4. Arellano, M.; Bover, O. (1995). Another look at the instrumental variable estimation of error–components models. *Journal of Econometrics*, v.68, n.1, 29–51.

Ari, M.A., Chen, S., and Ratnovski, M.L. (2019). The dynamics of non–performing loans during banking crises: A new database. *International Monetary Fund*.

Beck, R., Jakubik, P., and Piloiu, A. (2015). Key determinants of non–performing loans: new evidence from a global sample. *Open Economies Review*, v.26, n.3, 525–550.

Bernini, C.; Brighi, P. (2018). Bank branches expansion, efficiency and local economic growth. *Regional Studies*, v.52, n.10, 1332–1345.

Bernstein, S., Colonnelli, E., Giroud, X., and Iverson, B. (2019). Bankruptcy spillovers. *Journal of Financial Economics*, v.133, n.3, 608–633.

Cainelli, G., Montresor, S., and Marzetti, G.V. (2014). Spatial agglomeration and firm exit: a spatial dynamic analysis for italian provinces. *Small Business Economics*, v.43, n.1, 213–228.

Campbell, J.Y., Giglio, S., and Pathak, P. (2011). Forced sales and house prices. *American Economic Review*, v.101, v.5, 2108–31.

Chiesa, G.; Mansilla–Fernández, J. M. (2021). The dynamic effects of non–performing loans on banks' cost of capital and lending supply in the Eurozone. *Empirica*, v.48, n.2, 397–427.

Coccorese, P.; Ferri, G. (2020). Are mergers among cooperative banks worth a dime?Evidence on efficiency effects of M&A in Italy. *Economic Modelling*, v.84,147–164.

Coccorese, P.; Santucci, L. (2020). Banking competition and bank size: Some evidence from italy. *Journal of Economics and Finance*. v.44, n.2, 278–299.

Coccorese, P., Ferri, G., Lacitignola, P., and Lopez, J. (2016). Market structure, outer versus inner competition: the case of Italy's credit coop banks. *International Review of Economics*. v.63, n.3, 259–279.

Dimitrios, A., Helen, L., and Mike, T. (2016). Determinants of non–performing loans: Evidence from euro–area countries. *Finance Research Letters*, v.18 116–119.

Frame, W.S. (2010). Estimating the effect of mortgage foreclosures on nearby property values: A critical review of the literature. *Economic Review*, v.95.

Gerardi, K., Rosenblatt. E., Willen. P.S., and Yao, V. (2015). Foreclosure externalities: New evidence. *Journal of Urban Economics*, v.87, 42–56.

Ghosh, A. (2015). Banking–industry specific and regional economic determinants of non–performing loans: Evidence from US states. *Journal of Financial Stability*, v.20, 93–104.

Kuzucu, N.; Kuzucu, S. (2019). What drives non–performing loans? Evidence from emerging and advanced economies during pre– and post–global financial crisis. *Emerging Markets Finance and Trade*, n.55, v.8, 1694–1708.

Louzis, D.P., Vouldis, A.T., and Metaxas, V.L. (2012). Macroeconomic and bank–specific determinants of non–performing loans in Greece: A comparative study of mortgage, business and consumer loan portfolios. *Journal of Banking & Finance*, v.36, n.4, 1012–1027.

Naili, M.; Lahrichi, Y., 2022. The determinants of banks' credit risk: Review of the literature and future research agenda. *International Journal of Finance & Economics*, v.27, n.1, 334–360.

Pino, G.; Sharma, S.C. (2019). On the contagion effect in the US banking sector. *Journal of Money, Credit and Banking.* v.51, n.1, 261–280.

Zhang, X., Guo, D., Xiao, Y. and Wang, M. (2017) Do spatial spillover effects of non–performing loans for commercial banks exist? Evidence from Chinese provinces. *Emerging Markets Finance and Trade*, v.53, n.9, 2039–2051.

OS14 - Topics in Financial Econometrics

Testing the significance of the MIDAS variable in mixed-frequency volatility models through a bootstrap approach

Vincenzo Candila[1], **Lea Petrella**[2]

[1] *DISES Department, University of Salerno, Fisciano, Italy,*
vcandila@unisa.it
[2] *MEMOTEF Department, Sapienza University of Rome, Italy,*
lea.petrella@uniroma1.it

Abstract

Standard likelihood–based inference suffers from the presence of nuisance parameters. In the context of Mixing–Data Sampling (MIDAS) methods dedicated to the volatility estimation and forecasting (Engle et al., 2013, Conrad and Lock, 2015 and Amendola et al., 2021), this problem is of great importance. In such a framework, testing the null of no significance of the MIDAS terms is complicated by the presence of nuisance parameters that under the null hypothesis are not identifiable. This alters the asymptotic distribution of the common statistical tests employed. The present paper proposes a bootstrap likelihood ratio (BLR) test to overcome this problem, simulating the likelihood ratio test distribution. Using two Monte Carlo experiments, the proposed BLR test presents quite good performances in terms of size and power, beating the standard LR and Wald tests.

Keywords

Likelihood ratio test, MIDAS, Nuisance parameter, Bootstrap.

References

Amendola, A., V. Candila, and G.M. Gallo (2021). Choosing the frequency of volatility components within the Double Asymmetric GARCH-MIDAS-X model. *Econom. Stat.*, *20*, 12–28.

Conrad, C. and K. Lock (2015). Stock Market Volatility and Macroeconomic Fundamentals. *J. Appl. Econom.*, *30*(8), 1090–1114.

Engle, R.F., E. Ghysels, and B. Sohn (2013). Anticipating Long-Term Stock Market Volatility. *Rev. Bus. Econ. Stat.*, *95*(3), 776–797.

Time-varying graphical models for financial markets: a quantile approach

Beatrice Foroni[1], **Luca Merlo**[2], **Lea Petrella**[3]

[1] *Sapienza University, MEMOTEF Department, Rome, Italy, beatrice.foroni@uniroma1.it*

[2] *Sapienza University, Department of Statistical Sciences, Rome, Italy, luca.merlo@uniroma1.it*

[3] *Sapienza University, MEMOTEF Department, Rome, Italy, lea.petrella@uniroma1.it*

Abstract

Financial networks are recently emerging as useful tools for describing the propagation of systemic risk inside the financial global system. While most applications focus on time-constant networks, it is widely known that interactions among financial assets change over time, especially during crisis periods and system-wide extreme events. To address these issues, in this paper we introduce a time-varying quantile graphical model for inferring dynamic graphical structures in multivariate data at different quantile levels of interest. The proposed methodology relies upon the Multivariate Asymmetric Laplace distribution, which allows to jointly model the conditional quantiles of multivariate random variables, whose covariance matrix changes over time using a kernel smoothing approach. To fit the model we exploit its location-scale mixture representation and we implement a fused lasso penalized ADMM-based algorithm for estimating the sparse temporally dependent inverse correlation matrices at each time point. The introduced method is applied to financial returns of a set of market indexes, cryptocurrencies and commodities.

Keywords

Time-Varying Graphical Models, Multiple Quantiles, Multivariate Asymmetric Laplace, EM Algorithm.

References

Boyd, Parikh, Chu, Peleato, Eckstein, (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine learning*, *3*(1), 1-122.

Finegold, Drton,(2011). Robust graphical modeling of gene networks using classical and alternative t-distributions. *The Annals of Applied Statistics*, *5*(2A), 1057-1080.

Green, (1990). On use of the EM algorithm for penalized likelihood estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, *52*(3), 443-452.

Kotz, Kozubowski, Podgorski, (2012). The Laplace distribution and generalizations: A Revisit with Applications to Communications. *Economics, Engineering, and Finance*, *183*.

Monti, Ricardo Pio, et al. (2014). Estimating time-varying brain connectivity networks from functional MRI time series. *NeuroImage*, *103*, 427-443.

Petrella, Raponi, (2019). Joint estimation of conditional quantiles in multivariate linear regression models with an application to financial distress. *Journal of Multivariate Analysis*, *173*, 70-84.

Tibshirani, Robert, et al. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*.1, 91-108.

Yang, Peng, (2020). Estimating time-varying graphical models. *Journal of Computational and Graphical Statistics*, *29*(1), 191-202.

The HD(1) process: properties and application to unit root testing in interest rates

Alessandro Palandri[1]

[1] *Università degli Studi di Firenze, Dipartimento di Statistica, Informatica, Applicazioni (DiSIA) G. Parenti, Italy,* alessandro.palandri@unifi.it

Abstract

This paper presents and discusses the properties of the HD(1), a Markovian process of order one with hyperbolic reversion toward the long-run equilibrium. A first-order approximation aHD(1) is introduced to allow for an estimation-calibration procedure based on ARMA routines. Limiting distributions of unit root tests are derived for the aHD(1) specification of the alternative and the corresponding critical values tabulated. The empirical study revisiting the non-stationarity of interest rates finds that overall the aHD(1) is preferred to ARMA and SETAR specifications and rejects the unit root hypothesis for all rates and yields considered.

Keywords

Markov Process, Hyperbolic Reversion, Unit Root Test, Critical Values, Bond Yields.

References

Anderson, H. (1997), Transactions costs and non-linear adjustments towards equilibrium in the US treasury bill market, *Oxford Bulletin of Econometrics and Statistics*, *59*, 465-484.

Baillie, R. (1996), Long memory processes and fractional integration in econometrics, *Journal of Econometrics*, *73*, 5-59.

Balke, N.S. and T.B. Fomby (1997), Threshold cointegration, *International Economic Review*, *38*, 627-645.

Bec, F., M. Ben Salem and M. Carrasco (2004), Tests for unit-root versus threshold specification with an application to the purchasing power parity relationship, *Journal of Business and Economic Statistics*, *22*, 382-395.

Caner, M. and B. Hansen (2001), Threshold autoregression with a unit root, *Econometrica*, **69**, 1555-1596.

Chan, K-S. and H. Tong (2001), Chaos: a statistical perspective, Springer-Verlag, New York.

Chang, S. Y. and P. Perron (2017), Fractional unit root tests allowing for a structural change in trend under both the null and alternative hypotheses, *Econometrics*, *5*, 1-26.

Cho, C-K., C. Amsler and P. Schmidt (2015), A test of the null of integer integration against the alternative of fractional integration, *Journal of Econometrics*, *187*, 217-237.

Dickey, D. A. and W. A. Fuller (1979), Distribution of the estimators for autoregressive time series with a unit root, *Journal of the American Statistical Association*, *72*, 427-431.

Dickey, D. A. and W. A. Fuller (1981), Likelihood ratio statistics for autoregressive time series with a unit root, *Econometrica*, *49*, 1057-1072.

Dolado, J. J., J. Gonzalo and L. Mayoral (2002), A fractional Dickey-Fuller test for unit roots, *Econometrica*, *70*, 1963-2006.

Dolado, J. J., J. Gonzalo and L. Mayoral (2008), Wald tests of I(1) against I(d) alternatives: some new properties and an extension to processes with trending components, *Studies in Nonlinear Dynamics and Econometrics*, *12*, 1-35.

Enders, W. and C. Granger (1998), Unit root test and asymmetric adjustment with an example using the term structure of interest rates, *Journal of Business and Economic Statistics*, *16*, 304-311.

Engle, R. and C. Granger (1987), Co-integration and error correction: representation, estimation and testing, *Econometrica*, *55*, 251-276.

Gali, J. (1992), How well does the IS-LM model fit postwar U.S. data?, *Quarterly Journal of Economics*, *107*, 709-738.

Garcia, R. and P. Perron (1996), An analysis of the real interest rate under regime shifts, *Review of Economics and Statistics*, *78*, 111-125.

Geweke, J., S. Porter-Hudak (1983), The estimation and application of long memory time series models, *Journal of Time Series Analysis*, *4*, 221-238.

Granger, C. W., R. Joyeux (1980), An introduction to long-memory time series models and fractional differencing, *Journal of Time Series Analysis*, *1*, 15-30.

Hansen, B.E. (1996), Inference when a nuisance parameter is not identified under the null hypothesis, *Econometrica*, *64*, 414-430.

Hosking, J. R. M. (1981), Fractional differencing, *Biometrika*, *68*, 165-176.

Lobato, I. N. and C. Velasco (2007), Efficient Wald tests for fractional unit roots, *Econometrica*, *75*, 575-589.

Lo, A. (1991), Long-term memory in stock market prices, *Econometrica*, *59*, 1279-1313.

Lo, M.C. and E. Zivot (2001), Threshold cointegration and nonlinear adjustment to the law of one price, *Macroeconomic Dynamics*, *5*, 533-576.

MacDonald, R. and P. Murphy (1989), Testing for the long-run relationship between nominal interest rates and inflation using cointegration techniques, *Applied Economics*, *21*, 439-447.

MacKinnon, J. G. (2010), Critical values for cointegration tests, Working Paper 1227, Economics Department, Queen's University.

Michael, P., R.A. Nobay and D.A. Peel (1997), Transaction costs and nonlinear adjustment in real exchange rates: an empirical investigation, *Journal of Political Economy*, *105*, 862-879.

Mishkin, F. S. and J. Simon (1995), An empirical examination of the Fisher effect in Australia, *Economic Record*, *71*, 217-229.

Pesaran, M.H. and S.M. Potter (1997), A floor and ceiling model of US output, *Journal of Economic Dynamics and Control*, *21*, 661-695.

Phillips, P. C. B. (1999), Discrete Fourier transforms of fractional processes, *Cowles Foundation Discussion Papers*, *1243*.

Pippenger, M. and G. Goering (2000), Additional results on the power of unit root and cointegration tests under threshold parameters, *Applied Economics Letters*, *7*, 641-644.

Rapach, D. E. and C. E. Weber (2004), Are real interest rates really nonstationary? New evidence from tests with good size and power, *Journal of Macroeconomics*, *26*, 409-430.

Robinson, P. M. (1994), Efficient tests of nonstationary hypotheses, *Journal of the American Statistical Association*, *89*, 1420-1437.

Robinson, P. M. (1995), Log-periodogram regression of time series with long range dependence, *Annals of Statistics*, *23*, 1048-1072.

Rose, A. K. (1988), Is the real interest rate stable?, *Journal of Finance*, *43*, 1095-1112.

Sercu, P., R. Uppal and C. van Hulle (1995), The exchange rate in the presence of transaction costs: implications for tests of purchasing power parity, *Journal of Finance*, *50*, 1309-1319.

Shapiro, M. and M. Watson (1988), Sources of business cycle fluctuations, in Fischer, S. (Ed.), NBER Macroeconomics Annual, MIT Press, Cambridge, MA, 111-148.

Sowell, F. (1992), Maximum likelihood estimation of stationary univariate fractionally integrated time series models, *Journal of Econometrics*, *53*, 165-188.

Tanaka, K. (1999), The nonstationary fractional unit root, *Econometric Theory*, *15*, 549-582.

Taylor, A. (2001), Potential pitfalls for the PPP puzzle? Sampling and specification biases in mean reversion tests of the LOOP, *Econometrica*, *69*, 473-498.

Tong, H. (1990), Nonlinear time series: a dynamical system approach, Oxford University Press, Oxford.

OS15 - Advanced Statistical Models for Risk Evaluation

Environmental Risk: A Bayesian Non-linear State Space Copula Model to Predict Air Pollution in Beijing

L. Dalla Valle[1], **A. Kreuzer**[2], **C. Czado**[3]

[1] *University of Plymouth, School of Engineering, Computing and Mathematics, UK, luciana.dallavalle@plymouth.ac.uk*
[2] *Technische Universität München, Munich Data Science Institute, Germany, a.kreuzer@tum.de*
[3] *Technische Universität München, Munich Data Science Institute, Germany, cczado@ma.tum.de*

Abstract

Air pollution is a serious issue that currently affects many industrial cities in the world and can cause severe illness to the population. In particular, it has been proven that extreme high levels of airborne contaminants have dangerous short-term effects on human health, in terms of increased hospital admissions for cardiovascular and respiratory diseases and increased mortality risk. For these reasons, an accurate estimation of airborne pollutant concentrations is crucial.

In this paper, we propose a flexible novel approach to model hourly measurements of fine particulate matter and meteorological data collected in Beijing in 2014. We show that the standard state space model (Durbin and Koopman, 2000 and 2002), based on Gaussian assumptions, does not correctly capture the time dynamics of the observations. Therefore, we propose a non-linear non-Gaussian state space model where both the observation and the state equations are defined by copula specifications (Kreuzer *et al.*, 2022). The observation and state variables are coupled using two bivariate copulas (Czado, 2019). Since the copula approach allows for separate modeling of the margins and dependence, the observation variables are allowed to follow any time invariant statistical model. In the application we utilized a GAM to allow for non-linear effects of covariates. Once the marginal distribution of the response variables is specified, they can be transformed to the uniform scale using the probability integral transform. The resulting value on the uniform scale at time t, U_t, is then coupled with a [0,1] valued state variable for time t using a bivariate copula. Therefore, the observation equation of the copula based state space formulation is given by the conditional distribution

of U_t given the value of the state variable at time t. The time dynamics of the state variables is then similarly modeled as the conditional distribution of the state variable at time t given the state variable at time $t-1$, where these two state variables are jointly modeled by a bivariate copula. We first show that, in the case of bivariate Gaussian copula, standard linear state space models result. Since many different parametric bivariate copulas exist, the flexibility of the copula-based state space model is evident and thus a significant extension of linear Gaussian state space models is possible.

Of course, such an extension has its price. In our case this means we cannot follow a standard estimation approach as provided by the Kalman filter for linear state space models. Therefore we propose and develop a Bayesian approach based on Hamiltonian Monte Carlo. Further we deal with some identifiability issues of the copula state space, which we solve by restricting the strength of the dependence among the lag-one state space variables to be at least as high as the one of the observation variable U_t and the state variable at time t.

The state variables can be interpreted as a way to capture non-measured effects and thus are very appropriate for the Beijing air contaminants data set. It allowed us to identify unusual high levels of pollution, which were not captured by the measured variables.

In conclusion, the proposed copula state space approach is very flexible, since it allows us to separately model the marginal distributions and to accommodate a wide variety of dependence structures in the data dynamics. The proposed approach allows us not only to accurately estimate particulate matter measurements, but also to capture unusual high levels of air pollution, which were not detected by measured effects.

Keywords

Air Pollution, Bayesian Models, Hamiltonian Monte Carlo, State Space Models.

References

Czado, C. (2019). Analyzing dependent data with vine copulas. *Lecture Notes in Statistics, Springer.*

Durbin, J. and S. J. Koopman (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62* (1), 3–56.

Durbin, J. and S. J. Koopman (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, *89* (3), 603–616.

Kreuzer, A., Dalla Valle, L. and Czado, C. (2022). A Bayesian non-linear state space copula model to predict air pollution in Beijing. *Journal of the Royal Statistical Society – Series C*, in press.

A modelling framework for projections of equity portfolio returns under climate transition scenarios

Lorenzo Prosperi[1], **Luca Zanin**[2]

[1] *Prometeia, Financial Markets and Intermediaries Analysis, Italy,*
lorenzo.prosperi@prometeia.com
[2] *Prometeia, Wealth & Asset Management, Italy,*
luca.zanin@studio.unibo.it

Abstract

Climate change due to anthropogenic emissions is affecting nature, the economy, society, and the financial system (e.g., Calabrese et al., 2021; Magnan et al., 2021). The low carbon transition and the mitigation of damages from extreme weather events are among the most important challenges of our century and an opportunity to build a new economic model based on sustainability. In considering climate-related risks and opportunities, a crucial role for investors is the evaluation of the financial impacts (e.g., stock returns) deriving from firms' exposure to climate policy risks (e.g., a carbon price policy to encourage the divestment from fossil fuels and reduce emissions). Investors and regulators have a growing awareness that climate transition risks (such as carbon pricing policy) are among the new risk drivers for the financial system. One of the relevant implications of a shock in carbon price is the increase in business costs which may diminish profitability and affect capital market returns, especially of the most pollutant firms.

We propose a modelling framework for medium-term projections of returns of stocks and portfolios under different transition scenarios. Firstly, we suggest evaluating the firms' exposure to climate policies through a green factor, by constructing a market-weighted portfolio with long positions in firms labeled as green and short positions in firms labeled as brown. After building the green factor, we estimate an extended capital asset pricing model (CAPM; market factor + green factor) using the classic ordinary least square estimator. Then, we compare the results with those obtained using a robust MM-estimator (e.g., Yohai, 1987) for accounting for possible outliers or extreme events in stock returns. To evaluate the transmission channel from carbon price to the stock market, we propose estimating a Large Bayesian Vector Autoregression (LBVAR) model (e.g., Bańbura et al., 2010;

Cimadomo et al., 2021). This model estimates the relationship between carbon price and factor returns (i.e., the market factor and the green factor) included in the extended CAPM by considering some macro-financial variables according to relevant literature in the field (e.g., Bjørnland and Leitemo, 2009; Kumar et al., 2012). After projecting the factors of the CAPM conditional to different NGFS carbon price pathways using the LBVAR model, we project the stock returns of each firm into the portfolio under different transition scenarios up to 2030 (that are: Hot House World (Current policies; Nationally Determined Contributions), Orderly (Below 2°C; Net zero 2050), and Disorderly (Divergent net zero; Delayed transition)).

From the projections of returns by sectoral portfolios, firms of the mining and quarrying sector would suffer the most from introducing a carbon policy, especially assuming to anticipate a delayed transition scenario (that is, an immediate and high shock in carbon price). This negative performance is explained both by a strongly negative exposure to the green factor (highly sensitive sector to direct transition risk) and large systemic risk, due to the strong correlation of the revenues of this sector to business cycle fluctuations. In general, it emerges that in the short run the portfolios tend to both underperform the current policies scenario, with a worse performance in the disorderly scenario compared to the orderly one. However, in the medium term, they both outperform the current policies scenario, with a better performance in the disorderly transition scenario compared to the orderly one.

Keywords

Capital Asset Pricing Model, Carbon price policies, Green factor, Large Bayesian Vector Autoregression Model.

References

Bańbura M., Giannone D., Reichlin L. (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, *25*, 71–92.

Bjørnland H.C. and Leitemo K. (2009). Identifying the interdependence between US monetary policy and the stock market. *Journal of Monetary Economics*, *56*, 275–282.

Calabrese R., Dombrowski T., Mandel A., Pace R.K., Zanin L. (2021). Impacts of extreme weather events on mortgage risks and their evolution under climate change: A case study on Florida. *Working Paper* Available at SSRN: http://dx.doi.org/10.2139/ssrn.3929927.

Cimadomo J., Giannone D., Lenza M., Monti F., and Sokol A. (2021). Nowcasting with large Bayesian vector autoregressions. *Journal of Econometrics*, https://doi.org/10.1016/j.jeconom.2021.04.012.

Magnan A.K., Pörtner HO., Duvat V.K.E. et al. (2021). Estimating the global risk of anthropogenic climate change. *Nature Climate Change*, *11*, 879–885.

Yohai V.J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, *15*, 642–656.

Kumar S., Managi S., and Matsuda A. (2012). Stock prices of clean energy firms, oil and carbon markets: A vector autoregressive analysis. *Energy Economics*, *34*, 215–226.

A Regularized Ordinal Regression approach for estimation and cluster identification in perceived risk data analysis

Maria Iannario[1], **Claudia Tarantola**[2], **Ioannis Ntzoufras**[3]

[1] *University of Naples Federico II, Department of Political Sciences, Italy, maria.iannario@unina.it*

[2] *University of Pavia, Department of Economics and Management, Italy, claudia.tarantola@unipv.it*

[3] *Athens University of Economics and Business, Department of Statistics, Greece, ntzoufras@aueb.gr*

Abstract

The proportional odds model, which was introduced by McCullagh (1980), is probably the most commonly used ordinal regression model. The assumption that effects of covariates are not category-specific makes it a simply structured model that allows to interpret parameters in terms of cumulative odds. However, in many field of study, the model shows poor goodness-of-fit and does not adequately represent the underlying probability structure. As alternatives, parallel (non-proportional) and semi-parallel (partial proportional) odds models were proposed. Regularization techniques such as the lasso (Tibshirani 1996) and elastic net (Zou and Hastie 2005), among others, can be used to improve regression model coefficient estimation and prediction accuracy, as well as to perform variable selection. In the contribution we consider the elementwise link multinomial-ordinal class, which includes ordinal regression models, obtaining results with improved prediction accuracy and estimation results. For further details we refer to Yee (2010) and Tutz and Gertheiss (2016) about penalized regression models for ordinal response data that allow the parallelism assumption to be relaxed. Estimation is implemented in the R package *ordinalNet* (Wurm et al. 2021).

The analysis concerns the role of some individual characteristics and attitudes on the perceived risk for Covid-19 analysing data collected in Italy during 2020. The perceived risk has been assessed by means of a rating variable on 5-point Likert scale (ranging from 1= not at all risky to 5 = very risky). Details on the survey were reported in Bacci et al (2021). The study contributes to determine if existing control measures are perceived as adequate and how they impact on the perception of the riskiness for for the

Italian society. Furthermore, the interest for new media with related impact on the person's reaction has been examined, and also the perception of individuals with weak immune system and infected by Covid-19 was explored. A comparison among different models with an increasing level of restrictions was assessed. The cumulative model with parallel assumption represents the best candidate model with the lowest BIC index. Results show that respondents with high levels of interest for news from media and of agreement with Government guidelines were associated with an increasing perceived risk for Covid-19. High perceived risk has been also recorded for respondents with high level of education, affected by Covid-19, who perceived themselves as correctly informed and with a weak immune system. On the contrary people who frequently quit in contrast with the provisions of the government turned out to be less insightful/perceptual with respect the risk perception. Clusters of respondents were composed and displayed to identify the the adherence to the government containment measures and address alternative policies.

Keywords

Covid-19, Elastic net, Lasso, Ordinal regression, Variable selection, Risk perception.

References

Bacci B, Fabbricatore R, Iannario M. (2021) Latent trait models for perceived risk assessment using a Covid-19 data survey, *Journal of Applied Statistics*, DOI: 10.1080/02664763.2021.1937584.

McCullagh P (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society B*, *42*(2), 109–142.

Tibshirani R (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society B*, 267–288.

Tutz G, Gertheiss J (2016). Regularized Regression for Categorical Data. *Statistical Modelling*, *16*(3), 161–200.

Wurm, M. J., Rathouz, P. J., Hanlon, B. M. (2021). Regularized Ordinal Regression and the ordinalNet R Package. *Journal of Statistical Software*, *99*(6), 1–42.

Yee TW (2010). The VGAM Package for Categorical Data Analysis. *Journal of Statistical Software*, *32*(10), 1–34.

Zou H, Hastie T (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society B*, *67*(2), 301–320.

OS16 - Recent advances in systemic risk assessment

Systemic risk assessment through higher order clustering coefficient

Roy, Cerqueti[1], Gian Paolo Clemente[2], Rosanna Grassi[3]

[1] *Sapienza University of Rome, Department of Social Sciences and Economics, Italy, roy.cerqueti@uniroma1.it*

[2] *Università Cattolica del Sacro Cuore di Milano, Dipartimento di Matematica per le Scienze economiche, finanziarie ed attuariali, Italy, gianpaolo.clemente@unicatt.it*

[3] *University of Milano - Bicocca, Department of statistics and Quantitative Methods, Italy, rosanna.grassi@unimib.it*

Abstract

Global financial markets can be seen as part of a strongly interconnected system, so modelling them using a network tool can be useful to understand how systemic risk rises and how shocks propagate, thus preventing future financial crises. This work moves from this premise and proposes a novel measure of systemic risk in the context of financial networks. To this aim, we provide a definition of systemic risk based on the structure of neighbours around the nodes of the network. In the literature of financial networks, there are several measures of systemic risk based on the clustering coefficient (see Battiston et al. (2012), Billio et al. (2012) and Markose et al. (2012), Castellano et al. (2021)).

Indeed, being formally constructed on the number of triangles a node belongs to, these local coefficients generate synthetic global indicators that well includes the interconnections between elements of the system. However, the classic clustering coefficient takes into account only the neighbours of a node (Watts and Strogatz (1998)). Our aim is to consider the interconnections of the entire network simultaneously, opening the view beyond the adjacent. To this end, we introduce the concept of local l-adjacency clustering coefficient of a node i as an opportunely weighted mean of the clustering coefficients of the nodes at geodesic distance l from i. Then, we define a global adjacency clustering coefficient of i as the mean of the l-local adjacency clustering coefficients of i and we explore its properties in terms of systemic risk assessment. Empirical experiments on the time-varying global banking network show the effectiveness of the presented systemic risk measure and provide insights on

how systemic risk has changed over the last years, also in light of the recent financial crisis.

Keywords

Systemic risk, Clustering coefficient, Community structures, Network analysis, Cross-border banking.

References

Battiston, S., Gatti, D. D., Gallegati, M., Greenwald, B., Stiglitz, J. E. (2012). Liaisons dangereuses:Increasing connectivity, risk sharing, and systemic risk. *Journal of Economic Dynamics and Control*, *36*(8), 1121–1141.

Billio, M., Getmansky, M., Lo, A. W., Pelizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics*, *104*(3), 535–559.

Castellano, R., Cerqueti, R., Clemente, G. P., Grassi, R. (2021). An optimization model for minimizing systemic risk. *Mathematics and Financial Economics*, *15*(1), 103–129.

Markose, S., Giansante, S., Shaghaghi, A. R. (2012). "Too interconnected to fail" financial network of US CDS market: topological fragility and systemic risk. *Journal of Economic Behavior & Organization*, *83*(3), 627–646.

Watts, D. J, Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature*, *393*(6684), 440–442.

The Volcker Rule and the hedge fund liquidity circle

Michael Bowe[1], **Olga Kolokolova**[2], **Lijie Yu**[3]

[1] *Alliance Manchester Business School, the University of Manchester, UK, michael.bowe@manchester.ac.uk*

[2] *Alliance Manchester Business School, the University of Manchester, UK, olga.kolokolova@manchester.ac.uk*

[3] *Alliance Manchester Business School, the University of Manchester, UK, lijie.yu@manchester.ac.uk*

Abstract

A key component of the US government's response to the 2008 financial crisis is the enactment of the 2010 Dodd-Frank Wall Street Reform and Consumer Protection Act, among the most all-encompassing financial regulations of the current millennium. Section 619 of this Act incorporates one of its core regulatory directives, known as the Volcker Rule, which aims to reduce banks' overall risk profile. This key provision restricts banking entities from proprietary trading, as well as either sponsoring or maintaining an ownership interest in covered funds, a class of entities which includes both hedge funds and private equity funds. The Act became law in July 2010, granting banks a five-year time-frame to achieve full regulatory compliance.

This paper examines two key questions. First, do Volcker Rule regulations, targeted at the banking sector, also influence the risk profile and liquidity transmission channels in the hedge fund industry? Second, is the nature of any influence mediated by prime broker's connections to funds?

We find that following the Volcker Rule's enactment, which constitutes an exogenous shock to the financial system, the funding liquidity of hedge fund's deteriorates. Average hedge fund flows declines while their flow-performance sensitivity increases. The effect is especially discernible in funds with high operational risks. Such an environment of tighter funding liquidity leads to a reduction in hedge funds' exposure to market liquidity risk, inducing a drift towards more liquid investments and a further reduction in hedge funds' liquidity provision to illiquid market segments. One implication of this finding is that overall market efficiency may be adversely affected by the collective response of hedge funds to the changing trading environment after the Volcker Rule. These effects are mitigated for funds with low operational risks

and prime brokerage connections to large and complex financial institutions (LCFIs) in the US. US LCFIs appear to facilitate trading by funds with low operational risks, enabling them to retain their pre-Volcker levels of exposure to market liquidity and to maintain their liquidity provision to more illiquid segments of the equity market. The results are robust when we use: direct hedge fund holdings as reported to 13f, matched sample analysis, as well as within-fund changes in funding liquidity, liquidity risk exposure and comparing liquidity provision in periods before and after the Rule.

The significance of hedge funds in global financial markets is becoming increasingly noteworthy. According to BarclayHedge, global hedge fund assets under management increased 75-fold within 30 years, from $40 billion in 1990 to nearly $3 trillion in 2017. Hedge fund trades account for at least one-third of the total daily trading volume on the New York Stock Exchange (NYSE) (Cao at al., 2017), acting as crucial providers of liquidity and drivers of price formation (Mügge, 2014), thereby attenuating aggregate equity market mispricing (Akbas at al., 2015). At the same time, their relatively high use of leverage leaves hedge funds particularly vulnerable to market and funding liquidity risk, while their close relationships with LCFIs potentially enhances financial instability risks, as became evident following the collapse of Long-Term Capital Management L.P. This network of institutional connectivity has initiated increasing calls for the imposition of controls on hedge fund activity, with indirect regulation of their counterparties often considered to be the most effective mechanism for restraining funds' operations (King and Maier, 2009; Dardanelli, 2011). Our results provide empirical evidence to inform such policy debates.

Keywords

Volcker Rule, Hedge funds, Liquidity risk, Liquidity provision Fund flows.

References

Akbas, F., Armstrong, W. J., Sorescu, S., and Subrahmanyam, A. (2015). Smart money, dumb money, and capital market anomalies. *Journal of Financial Economics*, *118*(2), 355-382.

Cao, C., Liang, B., Lo, A. W., and Petrasek, L. (2017). Hedge fund holdings and stock market efficiency. *The Review of Asset Pricing Studies*, *8*(1), 77-116.

Dardanelli, G. T. (2011). Direct or indirect regulation of hedge funds: A European dilemma. *European Journal of Risk Regulation*, *2*(4), 463-480.

King, M. R. and Maier, P. (2009). Hedge funds and financial stability: Regulating prime brokers will mitigate systemic risks. *Journal of Financial Stability*, *5*(3), 283-297.

Mügge, D. (2014). *Europe and the governance of global finance*. Oxford University Press, USA.

An Extreme Value Approach to CoVaR Estimation

Natalia Nolde[1], **Chen Zhou**[2], **Menglin Zhou**[3]

[1] *University of British Columbia, Department of Statistics, Canada, natalia@stat.ubc.ca*

[2] *Erasmus University, Erasmus School of Economics, the Netherlands, zhou@ese.eur.nl*

[3] *University of British Columbia, Department of Statistics, Canada, menglin.zhou@stat.ubc.ca*

Abstract

The global financial crisis of 2007-2009 highlighted the crucial role systemic risk plays in ensuring stability of financial markets. Accurate assessment of systemic risk would enable regulators to introduce suitable policies to mitigate the risk as well as allow individual institutions to monitor their vulnerability to market movements. One popular measure of systemic risk is the conditional value-at-risk (CoVaR), proposed in Adrian and Brunnermeier (2011). We develop a methodology to estimate CoVaR semi-parametrically within the framework of multivariate extreme value theory. According to its definition, CoVaR can be viewed as a high quantile of the conditional distribution of one institution's (or the financial system) potential loss, where the conditioning event corresponds to having large losses in the financial system (or the given financial institution). We relate this conditional distribution to the tail dependence function between the system and the institution, then use parametric modelling of the tail dependence function to address data sparsity in the joint tail regions. We prove consistency of the proposed estimator, and illustrate its performance via simulation studies and a real data example.

Keywords

Systemic risk, Multivariate extreme value theory, Tail dependence function, Regular variation, Heavy tails, Method of moments.

References

T. Adrian and M.K. Brunnermeier, (2011) CoVaR. Technical report, National Bureau of Economic Research.

OS17 - Risk and opportunities in financial innovation

A sentiment analysis through Fuzzy transfrom to manage risks in tourism

Maria Letizia Guerra[1], **Laerte Sorini**[2]

[1] *Department of Statistical Sciences, University of Bologna, Italy, mletizia.guerra@unibo.it*

[2] *Department of Economics, Society and Politics, University of Urbino, Italy, laerte.sorini@uniurb.it*

Abstract

The connection between tourism flows and the level of perceived risk in the world wide web is analyzed through two types of regressions: quantile and expectile through inverse and direct F-transform.

In particular, Fuzzy transform has proved to capture stylized facts and mutual connections between time series having different nature.

Tourism flows and Google Trends are analyzed in order to establish new perspectives in the description of their dynamics.

Keywords

F-transform, Tourism, Clustering, Sentiment Analysis.

References

Guerra M.L., Sorini L., Stefanini L. (2019). Quantile and Expectile Smoothing based on L1-norm and L2-norm F-transforms, *International Journal of Approximate Reasoning*, *107*, 17–43.

Guerra M.L., Sorini L., Stefanini L. (2020). On the approximation of a membership function by empirical quantile functions, *International Journal of Approximate Reasoning*, *124*, 133–146.

Guerra M.L., Sorini L., Stefanini L. (2020). Bitcoin forecasting through Fuzzy Transform, *Axioms*, *9*(4), 139.

Extrapolation procedures to enhance the accuracy of numerical methods for derivative pricing

Luca Vincenzo Ballestra[1]

[1] *Alma Mater Studiorum University of Bologna, Department of Statistical Sciences "Paolo Fortunati", Italy, luca.ballestra@unibo.it*

Abstract

The calculation of the prices of financial derivatives often requires some numerical approximation. We show how the so-called repeated Richardson extrapolation (Ballestra, 2014, 2018, 2021) can be a very effective tool to obtain accurate and fast approximations of derivative prices. We consider both European vanilla options and barrier options on underlying assets described by the popular Black-Scholes model. Moreover, we also consider commodity derivatives under a model with several stochastic factors accounting for the so-called convenience yield (Schwartz, 1997). In the case of double barrier options, the repeated Richardson extrapolation is used in conjunction with a suitable change of variables, which allows us to align the computational mesh with the options' payoff. Numerical results will be presented showing that the repeated Richardson extrapolation, performed in both space and time, allows us to achieve a high level of computational efficiency. In particular, in the one-dimensional case, errors of the order 10e-6 or even smaller can be obtained in a CPU time smaller than a second. Very satisfactory performances are achieved also in the case of several stochastic factors.

Keywords

Option pricing, Richardson extrapolation, Finite difference scheme, Barrier option, Commodity derivative.

References

Ahmadian, D., Ballestra, L.V., Karimi N. (2021). An extremely efficient numerical method for pricing options in the Black-Scholes model with jumps. *Mathematical Methods in the Applied Sciences*, *44*, 1843–1862.

Ballestra, L.V. (2014). Repeated spatial extrapolation: An extraordinarily efficient approach for option pricing. *Journal of Computational and Applied Mathematics*, *256*, 83–91.

Ballestra, L.V. (2018). Fast and accurate calculation of American option prices. *Decisions in Economics and Finance*, *41*, 399–426.

Ballestra, L.V. (2021). Enhancing finite difference approximations for double barrier options: mesh optimization and repeated Richardson extrapolation. *Computational Management Science*, *18*, 239–263.

Schwartz, E. (1997). The stochastic behavior of commodity prices: implications for valuation and hedging, *Journal of Finance*, *52*, 922–973.

Sentiment-based regimes for stock price volatility

Alessandra Cretarola[1], **Gianna Figà-Talamanca**[2], **Marco Patacca**[3]

[1] *Department of Mathematics and Computer Science, University of Perugia, Italy, e-mail: alessandra.cretarola@unipg.it*
[2] *Department of Economics, University of Perugia, Italy, e-mail: gianna.figatalamanca@unipg.it*
[3] *Department of Economics, University of Verona, Italy, e-mail: marco.patacca@univr.it*

Abstract

With the availability of social networks, specialized forums and online news, sentiment analysis has become a common and useful technique for the analysis of economic and financial scenarios. Several data-providers, such as Bloomberg and Thomson Reuters have also started computing proprietary sentiment indexes on financial assets to be delivered together with traditional figures such as price and trading volume. Inspired by Papanicolaou and Sircar (2014), we propose and analyze a modified version of the Heston model, where the price volatility also varies according to regime changes related to a sentiment indicator and that can be extended to allow for jumps in the asset price process. Under the suggested model specification we will attempt to derive distributional characteristics of asset returns, a numerical procedure for its estimation/calibration on market data as well as pricing formulas for European-style derivatives.

Keywords

Market sentiment, Regime-switching model, Stochastic volatility.

References

Papanicolaou, A., Sircar, R. (2014). A regime-switching Heston model for VIX and S&P 500 implied volatilities. *Quantitative Finance, 14*(10), 1811-1827.

Feynman-Kac formula for BSDEs with jumps and time delayed generators associated to path-dependent nonlinear Kolmogorov equations

Luca Di Persio[1], **Matteo Garbelli**[1,2], **Adrian Zălinescu**[3]

[1] *University of Verona, Department of Computer Science, Italy,* luca.dipersio@univr.it

[2] *University of Trento, Department of Mathematics, Italy,* matteo.garbelli@unitn.it

[3] *Alexandru Ioan Cuza University, Department of Computer Science, Romania,* adrian.zalinescu@info.uaic.ro

Abstract

Motivated by financial considerations, we study a system of forward-backward stochastic differential equations (FBSDEs) in the spirit of the theory developed by Pardoux and Peng. The forward process evolves according to a jump-diffusion dynamic while the BSDE presents a time delayed generator and it is driven by a Lèvy-type noise. We provide a probabilistic representation for a (mild) solution of a path-dependent non-linear Kolmogorov equation by establishing a non-linear Feynman–Kac representation formula to match the FBSDE solution to the one for the associated path dependent non-linear Kolmogorov equation. The obtained results are then applied to an insurance problem based on a step process whose risk is quantified by means of a dynamic measure.

Keywords

BSDEs, Feynman-Kac formula, Jump-diffusion process, Path-dependence.

References

Barles, G., R. Buckdahn, and E. Pardoux (1997). Backward stochastic differential equations and integral-partial differential equations. *Stochastics and Stochastics Reports 60*.

Cordoni, F., L. Di Persio, L. Maticiuc, and A. Zălinescu (2020). A stochastic approach to path-dependent nonlinear Kolmogorov equations via BSDEs with time-delayed generators and applications to finance. *Stochastic Processes and their Applications*.

Cordoni, F., L. Di Persio and I. Oliva, (2017). A nonlinear Kolmogorov equation for stochastic functional delay differential equations with jumps. *Nonlinear Differ. Equ. Appl.*, *24*(16).

Delong, Łukas, Backward Stochastic Differential Equations with Jumps and Their Actuarial and Financial Applications: BSDEs with Jumps. Springer London. EAA Series, 2013.

Delong, Ł, P. Imkeller (2010). Backward stochastic differential equations with time delayed generators - results and counterexamples. *Ann. Appl. Probab.*, *20*(4).

Fuhrman, M., G. Tessitore (2005). Generalized Directional Gradients, Backward Stochastic Differential Equations and Mild Solutions of Semilinear Parabolic Equations. *Appl Math Optim.*, *51*.

Masiero, F., C. Orrieri, G. Tessitore and G. Zanco (2021). Semilinear Kolmogorov equations on the space of continuous functions via BSDEs. *Stochastic Processes and their Applications, 136(June 2021), 1-56 [10.1016/j.spa.2021.01.009].*

Pardoux, E., S. Peng (1992). Backward SDEs and quasilinear PDEs. *Stochastic partial differential equations and their applications, Springer.*

Peng, S. (2010). Backward Stochastic Differential Equation, Nonlinear Expectation and Their Applications. *Proceedings of the International Congress of Mathematicians Hyderabad, India, 2010.*

Investor sentiment as driver of financial stock market

Gianna Figá-Talamanca[1], **Marco Patacca**[2]

[1] *University of Perugia, Department of Economics, Italy,*
gianna.figatalamanca@unipg.it
[2] *University of Verona, Department of Economics, Italy,*
marco.patacca@univr.it

Abstract

Sentiment analysis is the field of machine learning that uses text mining techniques to infer opinions and moods expressed in messages and posts and then classify their polarity. In this paper, we propose an empirical analysis of the relative impact of sentiment measures on the returns of 150 components of the S&P 500 index by using the quantile regression to the historical data. Specifically, we consider companies with higher market capitalization among the 11 sectors of the S&P 500 index, and we measure sentiment through Bloomberg News and Twitter company-level indexes. Empirical results show that sentiment indicators mainly affect the quantiles on the left tail of returns distribution, underlying the different behaviors of investors according to moods and states of the market itself.

Keywords

Sentiment Analysis, Investor attention, Quantile regression.

OS18 - Computational Mathematics and Statistics in Risk Analysis

Modeling the Forest Fire Occurrence in Some Regions of Portugal

M. Filomena Teodoro[1,2]

[1] *CINAV - Center of Naval Research, Portuguese Naval Academy, Universitary Military Institute, Almada, Portugal,*
[2] *CEMAT - Center for Computational and Stochastic Mathematics, Instituto Superior Técnico, Lisbon University, Portugal*
mteodoro64@gmail.com

Abstract

The forest surveillance carried in Portugal presents some difficulties. To improve and solve some of these problems the use of new technologies such as unmanned aerial vehicle systems (UAVS) cam be implemented in an efficient way. To do so we need to determine the risk of occurrence of a forest fires at a certain time in a certain region. With such goal, we have used several statistical techniques, such as ARIMA or GLM approaches. We built adequate models in certain regions but for others it was impossible to determine a good quality model. The work is still going on, we expect to enlarge the area that we can get a good quality prevision of fire risk.

Keywords

Environmental risk assessment, Risk on decision making, Forest surveillance, Statistical approach.

References

Adou, J. K., Brou, A. D. V., Porterie, B. (2015). Modeling wildland fre propagation using a semiphysical network model. *Cases Studies in Fire Safety* [S.l.], *4*, 11–18.

Almeida, R. M. et al. (2015). Autômatos celulares probabilísticos aplicados à modelagem da propagação de incêndios de vegetação. *Proceeding series of the Brazilian Society of Applied and Computational Mathematics*, São Carlos, 3, *1*, 010393-1–010393-7.

Santos, E., Turci, L. F. R., Almeida, R.M. (2019). Evaluation of the probabilistic model of fire propagation using cellular automata applied to small areas. *Ciências Florestais*, *29*(4), 1685–1700. (In Portuguese).

Evaluating the impact of the use of the Model of excellence reference model by Brazilian electricity distributors

Alexandre Carrasco[1], **Marina A. P. Andrade**[2], **Álvaro Rosa**[3], **Filomena Teodoro**[4,5]

[1] *ISCTE, Instituto Universitário de Lisboa, Lisboa, Portugal*
[2] *ISTAR, ISCTE, Instituto Universitário de Lisboa, Lisboa, Portugal*
marina.andrade@iscte.pt
[3] *BRU, ISCTE, Instituto Universitário de Lisboa, Lisboa, Portugal*
[4] *CINAV - Center of Naval Research, Portuguese Naval Academy, Universitary Military Institute, Almada, Portugal,*
[5] *CEMAT - Center for Computational and Stochastic Mathematics, Instituto Superior Técnico, Lisbon University, Portugal*
mteodoro64@gmail.com

Abstract

Disaster situations, of natural origin or caused by man, in general, are emergency situations of great demand, both in terms of human issues and as to the spatial-geographical needs that require a quick response, whether to meet the first assessments of affected sites or to discontinue the process.

The main objective of this project is to build and implement a set of auxiliary tools to different decision support systems that allow, in each process, define priorities for scaling teams, taking into account the importance of each team in action, and what should be the sequence of tasks and orders to be carried out from which the alert is given until the final action is considered.

The definition of a decision support system (DSS) that introduces the possibility of redefine and adjust in real time the weights assigned to the experts involved in the solution obtained by the THEMIS - Distributed Holistic Emergency Management project Intelligent System shall be considered as an improvement. To redefine the weights, it was intended to use techniques multi-criteria, namely the multi-criteria decision analysis approach (Multiple-Criteria Decision Analysis - MCDA).

Keywords

Model of excellence, Costumer satisfaction, Electricity distributors, Parametric and nonparametric approach.

References

ANEEL. Agência Nacional de Energia Elétrica. (2017). Evolução Iasc e Benchmarks Internacionais. http://www.aneel.gov.br/metodologia-iasc/-/asset_publisher/ri7lpr3r2ykt/content/evolucao-iasc-e-benchmarks-internacionais/655804?inheritredirect=false&redirect=http%3a%2f%2fwww.aneel.gov.br%2fmetodologia. Accessed in October 10, 2021.

Carrasco, A. (2018). Dez anos de estudos sobre o impacto do uso modelos de excelência na qualidade do fornecimento (DEC e FEC) e satisfação de clientes no setor de distribuiçâo de energia elétrica brasileiro. Master thesis. ISCTE – Instituto Universitário de Lisboa.

Studying Portuguese Pediatric Hypertension

Carla Simão[1,2], **Filomena Teodoro**[3,4]

[1] *Faculty of Medicine , Lisbon University, Lisboa, Portugal*
[2] *Pediatric Department, Santa Maria's Hospital, Centro Hospitalar Lisboa Norte, Lisboa, Portugal.*
[3] *CINAV - Center of Naval Research, Portuguese Naval Academy, Universitary Military Institute, Almada, Portugal,*
[4] *CEMAT - Center for Computational and Stochastic Mathematics, Instituto Superior Técnico, Lisbon University, Portugal*
mteodoro64@gmail.com

Abstract

The objective of the present study is to characterize the blood pressure (BP) profile of the Portuguese pediatric population at school age and to assess the prevalence of pediatric arterial hypertension (PAH), normal BP, high-normal BP, and analyze the relationship between normal-BP, high-normal BP, PAH and some demographic characteristics. A representative sample of the pediatric population was drawn up at the national level and data collection was completed recently. The statistical approach evidences that the results obtained are in agreement with some literature confirming a high prevalence of PAH among children and adolescents of the Portuguese population.

Keywords

Pediatric hypertension, Statistical approach, General linear model.

References

Rao, C. R., Miller, J. P., Rao, D. C. (2010). Essential Statistical Methods for Medical Statistics. *Elsevier, Nort-Holland.*

Teodoro, M.F. and Simão, C. (2017). Perception about Pediatric Hypertension. *Journal of Computational and Applied Mathematics.* 312, *209–215.*

Build Efficient Response Techniques to Catastrophes

Marina A. P. Andrade[1], **Mário Simões Marques**[2,3], **Filomena Teodoro**[3,4]

[1] *ISTAR, ISCTE, Instituto Universitário de Lisboa, Lisboa, Portugal*
marina.andrade@iscte.pt
[2] *Instituto Hidrográfico, Portuguese Navy, Lisboa, Portugal*
[3] *CINAV - Center of Naval Research, Portuguese Naval Academy, Universitary Military Institute, Almada, Portugal,*
[4] *CEMAT - Center for Computational and Stochastic Mathematics, Instituto Superior Técnico, Lisbon University, Portugal*
mteodoro64@gmail.com

Abstract

Disaster situations, of natural origin or caused by man, in general, are emergency situations of great demand, both in terms of human issues and as to the spatial-geographical needs that require a quick response, whether to meet the first assessments of affected sites or to discontinue the process.

The main objective of this project is to build and implement a set of auxiliary tools to different decision support systems that allow, in each process, define priorities for scaling teams, taking into account the importance of each team in action, and what should be the sequence of tasks and orders to be carried out from which the alert is given until the final action is considered.

The definition of a decision support system (DSS) that introduces the possibility of redefine and adjust in real time the weights assigned to the experts involved in the solution obtained by the THEMIS - Distributed Holistic Emergency Management project Intelligent System shall be considered as an improvement. To redefine the weights, it was intended to use techniques multi-criteria, namely the multi-criteria decision analysis approach (Multiple-Criteria Decision Analysis - MCDA).

Keywords

Disaster situation, Decision support systems, Weights assigned to the experts.

References

Teodoro, M.F., Marques, M.J.S., Nunes, I., Calhamonas, G., Andrade, M.A.P. (2022). New Refinement of an Intelligent System Design for Naval Operations. In: Machado, J., Soares, F., Trojanowska, J., Yildirim, S. (eds) Innovations in Mechatronics Engineering. icieng 2021. *Lecture Notes in Mechanical Engineering*, 164–177. Springer, Cham. https://doi.org/10.1007/978-3-030-79168-1_16

Teodoro, M.F., Marques, M.J.S., Nunes, I., Calhamonas, G., Andrade, M.A.P. (2020). Using MDS to compute the contribution of the experts in a Delphi forecast associated to a naval operation's DSS. In: Gervasi, O et al. *Lecture Notes in Computer Sciences*, *12251*, 446–454. Springer, Cham. ttps://doi.org/10.1007/978-3-030-58808-3_32

Contributed sessions

CS01 - Risk analysis and assessment in health care applications. Chair: *Paolo Trerotoli*.

CS02 - Modelling in Risk Analysis. Chair: *Christos Kitsos*.

CS03 - Risk Analysis in new disease development. Chair: *Teresa Oliveira*.

CS04 - Statistical and Machine learning models for risk detection. Chair: *Chrys Caroni*.

CS05 - Advanced Statistical Models for Risk Evaluation. Chair: *Amílcar Oliveira*.

CS06 - Risk Analysis in Applied Science. Chair: *Rosaria Gesuita*.

CS01 - Risk analysis and assessment in health care applications

Analysis of acute exacerbation and mortality in idiopathic pulmonary fibrosis using secondary sources

Marica Iommi[1], Andrea Faragalli[1], Alessandro Fontanarosa[1], Luigi Ferrante[1], Martina Bonifazi[2], Lara Letizia Latini[2], Edlira Skrami[1], Flavia Carle[1], Rosaria Gesuita[1]

[1] *Center of Epidemiology, Biostatistics and Medical Information Technology, Marche Polytechnic University, Ancona, Italy, m.iommi@staff.univpm.it*

[2] *Department of Biomedical Sciences and Public Health, Marche Polytechnic University, Ancona, Italy*

Abstract

Introduction: Idiopathic pulmonary fibrosis (IPF) is a rare, chronic, and progressive disease, causing irreversible decline of lung function over time which dramatically impairs the quality of life. It mainly occurs in male, older people and it is characterized by a poor prognosis, with a median survival time ranging between 3 and 5 years. The aim of the study is to evaluate the risk of acute exacerbation and the determinants of death in new cases of IPF using secondary sources, during the period 2014-2019. Results of the Motive project (PRIN 2019-2021, code 2017728JPK).

Methods: This observational prospective study was based on administrative databases of hospital discharges, drug prescriptions, regional beneficiaries' database, and outpatient care database; all adult residents in the Marche Region with a first prescription of antifibrotic drugs or a first hospitalization with diagnosis of IPF occurred between 01/01/2014 and 30/06/2019, were included in the incident cohort of IPF. Subjects had to have been resident for at least 3 years prior to the date of inclusion and had no hospitalizations for IPF or antifibrotic prescriptions in the 2011-2013 period. Outcomes were acute exacerbation, defined as any acute respiratory-related hospitalization occurred after the date of IPF incidence, and all-cause mortality. Survival analysis was used to estimate the two outcomes, using the Kaplan-Meier procedure, with 95% Confidence Interval (95%CI). IPF cases were followed from the date of inclusion to the date of the outcome of interest, migration, or December 31, 2019, whichever came first. All patients were followed up for at least 180 days. Kaplan-Meier curves were estimated stratifying by sex,

age (dichotomized at 75 years), and comorbidities at baseline, evaluated using the Multisource Comorbidity Score (MCS) (score 0-4, good or fair health conditions; score $\geq$55, poor health conditions). Log-rank test was used to compare the curves. Multiple Cox regression was applied to estimate the risk of death adjusted by sex, age groups, MCS classes and the number of acute exacerbations (0, 1, 2, or more than 2) during the follow-up considered as time dependent. Since in clinical practice IPF patients with a severe disease are generally not eligible for antifibrotic treatment, subjects were stratified in patients with no antifibrotic prescription (never treated) and those with at least one prescription (treated patients) in all statistical analyses.
Results: During the study period, 676 new cases of IPF were identified (66.6% males), the median age was 75 years (1st-3rd quartile: 68-80) and 57.7% had poor health conditions at diagnosis; 271 patients (40.1%) had at least one antifibrotic prescription. During the follow-up, 276 deaths (225 of patients never treated) and 248 acute exacerbations (175 of patients never treated) occurred. The median time to death was 6 months in never treated patients and 21 months in treated patients; the median time to acute exacerbation was 5.4 months and 16 months in never treated and treated patients, respectively. After five years from the IPF diagnosis, the survival was 55.1% (95%CI: 43.3-70.1) in the treated group and 34.2% (95%CI: 28.8-40.7) in the never treated group ($p<0.001$); whereas the probability of being free from acute exacerbation was of 49.6% (95%CI: 38.6-63.8) and of 37.6% (95%CI: 31.0-45.6), respectively in treated and in never treated patients ($p<0.001$). In never treated patients, the probability of no acute exacerbation was significantly higher in younger patients or in subjects with a MCS score between 0-4. No significant differences were observed in treated patients. In never treated patients, the risk of death was significantly higher in males (HR=1.43; 95%CI: 1.08-1.90), in patients aged $\geq$75 (HR=2.14; 95%CI: 1.58-2.91) and in those with poor health conditions at baseline (HR=2.24; 95%CI: 1.58-3.17). Compared to patients with no exacerbation, having 1, 2, or more than 2 episodes increased the risk of death by 7.02, 8.28, and 7.03-fold, respectively. In treated patients, older age (HR=1.88; 95%CI: 1.07-3.31) and having 1, 2, or more episodes of exacerbations increased the risk of death (by 16.9, 16.5, and 17.9-fold, respectively).
Conclusions: Using healthcare administrative databases, it was possible to evaluate the risk of acute exacerbation and the determinants of death in new IPF cases in the Marche region. IPF patients were characterized by high

mortality and a high risk of exacerbation, already within the first year of diagnosis. Older age was an independent and significant determinant of survival both in not treated and treated patients; the general clinical condition at IPF diagnosis was associated with the risk of exacerbation in not treated patients and independently impact the risk of death in both groups.

Keywords

Healthcare administrative databases, Survival analysis, Cox regression, Real-world evidence, Idiopathic pulmonary fibrosis.

Adherence and tolerance of Idiopathic Pulmonary Fibrosis treatment in real world

Andrea Faragalli[1], **Marica Iommi**[1], **Alessandro Fontanarosa**[1], **Edlira Skrami**[1], **Luigi Ferrante**[1], **Martina Bonifazi**[2], **Lara Letizia Latini**[2], **Flavia Carle**[1], **Rosaria Gesuita**[1]

[1] *Marche Polytechnic University, Center of Epidemiology, Biostatistics and Medical Information, Ancona, Italy, a.faragalli@staff.univpm.it*

[2] *Marche Polytechnic University, Department of Biomedical Sciences and Public Health, Ancona, Italy*

Abstract

Introduction: Idiopathic pulmonary fibrosis (IPF) is a rare, devasting, fibrosing lung disease of still unknown aetiology. The recent approval in Italy of Pirfenidone (2014) and Nintedanib (2016) for IPF treatment, has changed the therapeutic management and improved patient prognosis. The aim of the study was to investigate the adherence and tolerance to Pirfenidone and Nintedanib in patients with newly IPF diagnosis using healthcare administrative databases, during 2014-2019. Results of the Motive project (PRIN 2019-2021, code 2017728JPK).

Study population: The target population was adult inhabitants in Marche Region; included IPF incident cases were all individuals at their first hospital discharge with ICD-9-CM code 516.3 in primary or secondary diagnosis fields (diagnosis of IPF), or at their first drug prescription of Pirfenidone or Nintedanib, between 01/01/2014 and 30/12/2018. To exclude prevalent cases, subjects had to have been resident for at least 3 years prior to the date of inclusion and had no hospitalizations for IPF or antifibrotic prescriptions in the 2011-2013 period.

Outcome: The adherence to antifibrotic prescription was estimated using the Proportion of Days Covered (PDC), i.e., the proportion of days in which a person has access to the antifibrotic drug, over a period of 12 months after the first prescription. Patients were adherent if PDC $\geq$ 75%. In case of two consecutive prescriptions of different drug (switch from Pirfenidone to Nintedanib and vice versa) or of Nintedanib in reduced dosage (switch from 150 mg to 100 mg) within 12 months of the first prescription, the patient was defined as intolerant.

Statistical Analysis: The cumulative probability of receiving the first antifibrotic prescription after 12 months from IPF diagnosis was evaluated considering IPF cases identified by hospital discharge record applying the Kaplan Meier method. Adherence to antifibrotic prescription was estimated as the proportion of patients with PDC $\geq$ 75%, and 95% Confidence Interval (95% CI), considering two sub-periods: 2014-2015 (only Pirfenidone was available), and 2016-2018. Patients were followed up until the first event between a drug switch, end of health care, death, or because at the end of the 12-month period from the first prescription. Logistic regression model was used to estimate the association between the adherence to antifibrotic prescription and type of antifibrotic therapy (Pirfenidone vs. Nintedanib), adjusting by sex, age classes (<75 years, $\geq$ 75 years), and classes of Multisource Comorbidity Score (MCS) (health conditions: 0-4 good or fair, 5-14 slightly poor, $\geq$ 15 poor). The cumulative probability of antifibrotic intolerance at 12 months from the first prescription was estimated using the Kaplan-Meier method, considering new IPF cases detected between 2016-2018.
Results: During the study period, 600 new IPF cases were detected; the median age was 75 years (1^{st}-3^{rd} quartile: 68-80), 68% were males, and 13% had poor health condition at baseline. During follow up, 234 (39%) patients received at least one drug prescription (only-users). Of 471 (78.5%) IPF cases firstly identified from hospital discharge, 92 patients received at least one antifibrotic prescription within 12 months from discharge and the cumulative probability was 23.3% (95%CI: 18.9-27.3). The adherence to Pirfenidone prescription was 70.3% (95% CI: 52.8- 83.6) in the cohort of only-users identified in 2014-2015 (n=37) and 65.9% (95% CI: 57.2-73.7) in those of 2016-2018 (n=135), while the adherence to Nintedanib prescription (n=104) was 63.5% (95% CI: 53.4-72.5). The adherence to treatment was not significantly different between Pirfenidone and Nintedanib (p=0.165). Among all the patients treated with Pirfenidone (n=101), the cumulative probability of intolerance, i.e. switching to Nintedanib, was 13% (95%CI: 6.1-19.3). Only 2 patients out of 75 switched from Nintedanib to Pirfenidone. There were 33 patients out of 71 that reduced the Nintedanib daily dosage from 150mg to 100mg, with intolerance cumulative probability of 31.3% (95%CI: 19.5-41.3).
Conclusion: The use of healthcare administrative databases allowed to describe and analyze the adherence and tolerance to antifibrotic prescription in patients with IPF and both are in line with real world data derived from cohort hospital-based studies. New IPF cases detected by hospitalizations had

a low probability of receiving antifibrotic treatment, because they might have a severe onset of the disease and, hence, not eligible for treatment in clinical practice. High adherence and a fair level of tolerance were observed for both Pirfenidone and Nintedanib. Despite the efficacy of antifibrotic treatment in reducing adverse IPF outcomes has been already shown, the current percentage of treated patients remains low.

Keywords

Estimation of adherence and tolerance, Secondary sources, Real world evidence, Antifibrotic treatment, Idiopathic pulmonary fibrosis.

Copula Link-Based Additive Models for Bivariate Time-to-Event Outcomes with General Censoring Scheme

Danilo Petti[1], **Alessia Eletti**[2], **Giampiero Marra**[3], **Rosalba Radice**[4]

[1] *University of Salerno, Department of Economics and Statistical Science, Italy, dpetti@unisa.it*

[2] *University College London, Department of Statistical Science, London, alessia.eletti.19@ucl.ac.uk*

[3] *University College London, Department of Statistical Science, London, giampiero.marra@ucl.ac.uk*

[4] *Bayes Business School, Faculty of Actuarial Science and Insurance, London, rosalba.radice@city.ac.uk*

Abstract

Bivariate survival outcomes arise frequently in many research areas such as health and epidemiology. For example, bivariate survival data are often used in clinical trials studying diseases concerning paired organs, where the outcomes of interest are measured on the same individual and as a consequence are associated. The main feature of survival data is censoring. We propose a general and flexible copula regression approach that can handle bivariate survival data subject to various censoring mechanisms, which include a mixture of uncensored, left−, right−, and interval-censored data. The proposal permits to specify all model parameters as flexible functions of covariate effects, flexibly model the baseline survival functions by means of monotonic P-splines, characterise the marginals via transformations of the survival functions which yield, e.g., the proportional hazards and odds models as special cases, and model the dependence between events using a wide variety of copulae. The algorithm is based on a computationally efficient and stable penalised maximum likelihood estimation approach with integrated automatic multiple smoothing parameter selection. Despite the proposed framework is complex in that it allows for many layers of structure, there is no price to pay in terms of usability and interpretability. The potential of the approach is illustrated via a simulation study as well as using data from the Age-Related Eye Disease Study (AREDS), a multi-center randomised clinical trial exploring the development and progression of age-related macular degeneration (AMD), sponsored by the National Eye Institute. The analysis

aims to quantify the effect of clinical risk factors on the joint risks of AMD progression as well as to predict the progression profiles of AMD patients with different characteristics. The modelling framework has been incorporated in the newly-revised `R` package `GJRM`, hence allowing any user to fit the desired model(s) and produce easy-to-interpret numerical and visual summaries.

Keywords

Additive predictor, Bivariate survival data, Copula, Link function, Mixed censoring scheme, Simultaneous penalised parameter estimation.

References

Brechmann, E. C.& Schepsmeier, U. (2013). Modelling dependence with c- and d-vine copulas: The R package CDVine. *Journal of Statistical Software*, *52*(3), 1–27.

Chen, M., Chen, L., Lin, K., & Tong, X. (2014). Analysis of multivariate interval censoring by diabetic retinopathy study. *Communications in Statistics-Simulations and computation*, *43*(7), 187–200.

Chen, M., Tong, X., & Sun, J. (2009). A frailty model approach for regression analysis of multivariate current status data. *Statistics in Medicine*, *28*(27), 3424–3436.

Clayton, D. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, *65*(1), 141–151.

Cook, R. & Tolusso, D. (2009). Second-order estimating equations for the analysis of clustered current status data. *Biostatistics*, *65*(1), 141–151.

Group, A. (1999). The age-related eye disease study (areds): Design implications. *AREDS report no.1. Controlled Clinical Trials*, *20*(6), 573–600.

Hu, T., Zhou, Q., & Sun, J. (2017). Regression analysis of bivariate current status data under the proportional hazards model. *Canadian Journal of Statistics*, *45*(4), 410–424.

Kor, C., Cheng, K., & Chen, Y. (2013). A method for analysing clustered interval-censored data based on cox model. *Statistics in Medicine*, *32*(5), 822–832.

Leitenstorfer, F. & Tutz, G. (2006). Generalized monotonic regression based on B-splines with an application to air pollution data. *Biostatistics*, *8*(3), 654–673.

Liu, X.-R., Pawitan, Y., & Clements, M. (2018). Parametric and penalized generalized survival models. *Statistical Methods in Medical Research*, *27*(5), 1531–1546.

Marra, G. & Radice, R. (2020). Copula link-based additive models for right-censored event time data. *Journal of the AmericaN Statistical Association*, *115*(530), 886–895.

Marra, G. & Radice, R. (2022). *GJRM: Generalized Joint regression Modelling. R package version 0.2-6.*

Martins, A., Aerts, M., Hens, N., Wienke, A., & Ambrams, S. (2019). Correlated gamma frailty models for bivariate survival time data. *Statistical Methods in Medical Research*, *28*(10–11), 3437–3450.

Nelsen, R. (2006). *An introduction to Copulas.* Second Edition, Springer, New York.

Oakes, D. (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society: Series B*, *44*(3), 414–422.

Oakes, D. (1986). A class of multivariate failure time distributions. *Biometrika*, *73*(3), 671-678.

Pya, N. & Wood, S. (2015). Shape constrained additive models. *Biometrika*, *73*(3), 671-678.

R Development Core Team. (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Reid, N. (1994). A conversation with sir David Cox. *Statistical Science*, *9*(3), 439–455.

Romeo, J., Meyer, R., & Gallardo, D. (2018). Bayesian bivariate survival using power variance function copula. *Lifetime Data Analysis*, *24*(2), 355–383.

Sun, T. & Ding, Y. (2021a). Copula-based semiparametric regression method for bivariate data under general interval censoring. *Biostatistics*, *22*(2), 315–330.

Sun, T. & Ding, Y. (2021b). *CopulaCenR: Copula-Based Regression Models for Bivariate Censored Data*. R package version 1.1.3.

Swaroop, A., Chew, E., Rickman, C., & Abecasis, G. (2009). Unraveling a multifactorial late-onset disease: from genetic susceptibility to disease mechanisms for age-related macular degeneration. *Annual Review of Genomics and Human Genetics*, *10*(1), 19–43.

Vatter, T. & Chavez-Demoulin, V. (2015). Generalized additive models for conditional dependence structure. *Journal of Multivariate Analysis*, *141*(C), 147–167.

Wang, L., Sun, J., & Tong, X. (2008). Efficient estimation for proportional hazards model with bivariate current status data. *Biometrika*, *14*(2), 134–153.

Wang, N., Wang, L., & McMahan, C. (2015). Regression Analysis of bivariate current status data under gamma- frailty proportional hazards model using the em algorithm. *Computational Statistics & Data Analysis*, *83*(C), 140–150.

Wen, C. & Chen, Y. (2013). A frailty model approach for regression analysis of bivariate interval-censoring survival data. *Statistica Sinica*, *23*(1), 383–408.

Wood, S. N. (2017). Generalized Additive Models: An Introduction with R. Second Edition, Chapman & Hall/CRC, London.

Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, *111*(516), 1548–1563.

Zeng, D., Gao, F., & Lin, D. (2017). Maximum likelihood estimation for semiparametric regression models with multivariate interval-censored data. *Biometrika*, *104*(3), 505–525.

Zhou, Q., Hu, T., & Sun, J. (2017). A sieve semiparametric maximum likelihood approach for regression analysis of bivariate interval-censored failure time data. *Journal of the American Statistical Association*, *112*(518), 664–672.

CS02 - Modelling in Risk Analysis

Bivariate vine copula based quantile regression

Marija Tepegjozova[1], **Claudia Czado**[2]

[1] *Technische Universität München, Department of Mathematics, Germany, m.tepegjozova@tum.de*

[2] *Technische Universität München, Department of Mathematics and Munich Data Science Institute, Germany, cczado@ma.tum.de*

Abstract

The statistical analysis of univariate quantiles is a well developed research topic. However, there is a profound need for research in multivariate quantiles. We tackle the topic of bivariate quantiles and bivariate quantile regression using vine copulas. They are graph theoretical models identified by a sequence of linked trees, which allow for separate modelling of marginal distributions and the dependence structure. We introduce a novel graph structure model (given by a tree sequence) specifically designed for a symmetric treatment of two responses in a predictive regression setting. We establish computational tractability of the model and a straight forward way of obtaining different conditional distributions. Using vine copulas the typical shortfalls of regression, as the need for transformations or interactions of predictors, collinearity or quantile crossings are avoided. We illustrate the copula based bivariate quantiles for different copula distributions and provide a data set example. Further, the data example emphasizes the benefits of the joint bivariate response modelling in contrast to two separate univariate regressions or by assuming conditional independence, for bivariate response data set in the presence of conditional dependence.

Keywords

Multivariate quantiles, Bivariate response, Bivariate conditional distribution functions.

Mean Hitting Time Approximation for Rare Events

Nikolaos LIMNIOS[1]

[1] *Applied Mathematics Laboratory*
Université de Technologie de Compiègne, Sorbonne University Alliance,
nlimnios@utc.fr).

Abstract

The probability of a multi-state system to reach a "bad" state (death, failure, negative performance, etc.) multiplied by the cost of this event provides the risk. The time to reach such a state for the first time is the so called *hitting time.*

For a process $Z^\epsilon(t), t \geq 0$ indexed by the small parameter $\epsilon > 0$, with general state space E, and D a measurable subset of E, the bad states for which transition probabilities, from $E \setminus D$, are small and depend on the parameter ϵ. Define now hitting time τ^ϵ of D, by the process $Z^\epsilon(t)$, i.e., $\tau^\epsilon := \inf\{t \geq 0 : z_t^\epsilon \in D\}$.

We can consider D as a single merged state since we are interested just by the hitting time to D. The problem we are considering now is the asymptotic behaviour of mean hitting time in regards of the asymptotic merging of states $E \setminus D$. For a large family of stochastic processes, as Markov chains, Markov processes, semi-Markov processes, semi-Markov chains, diffusion processes, etc., the expectation function of τ^ϵ, $\eta^\epsilon(x) := \mathrm{E}_x\tau^\epsilon$, $x \in E$, satisfies the operator equation $\mathbf{L}_0^\epsilon\eta^\epsilon = -1$, where $\mathbf{L}_0^\epsilon$ is a partial operator on $E \setminus D$ which is specialized for each particular family of processes. This is a singular perturbation problem which we solve by V.S. Koroliuk's method in order to obtain limit results.

Keywords

Hitting time, Rare event, Functional asymptotic, Singular perturbation problem.

References

Koroliuk V.S., Limnios N., (2005). *Stochastic systems in merging phase space*, World Scientific, Singapore.

Barbu V., Limnios N., (2008). Semi-Markov Chains and Hidden Semi-Markov Models. Toward Applications. Their use in Reliability and DNA Analysis, *Lecture Notes in Statistics*, *191*, Springer.

Limnios N., (2012). Reliability measures of semi-Markov systems with general state space, *Meth. Comput. Appl. Probab.*, *14*, 895–917. Doi.org/10.1007/s11009-011-9211-5

Edgeworth series expansion for option prices

Guillaume Leduc[1]

[1] *American University of Sharjah, Department of Mathematics and Statistics, United Arab Emirates, gleduc@aus.edu*

Abstract

Edgeworth series are a powerful tool in probability used to describe the asymptotic behavior of standardized sum of independent and identically distributed random variables. However, when evaluating options using tree methods, the standard Edgeworth expansion method does not apply directly because only triangular arrays of independent and identically distributed random variables arise. We show how the extension to triangular arrays of the Kolassa-McCullagh approach to Edgeworth series for lattice distributions combined with recent advances in Edgeworth expansion for triangular arrays yield Edgeworth series for digital and standard European put and call options. We develop closed form formula for the coefficients of $1/\sqrt(n)$ and $1/n$ under a general framework. We provide examples with the most popular trinomial models.

Keywords

Edgeworth series, Trinomial models, Option Pricing.

References

Leduc, G. (2013). A European option general first-order error formula. *The ANZIAM Journal*, *54*(4), 248–272.

Leduc, G. (2016). Can high-order convergence of European option prices be achieved with common CRR-type binomial trees? *Bulletin of the Malaysian Mathematical Sciences Society*, *39*, 1329–1342.

Leduc, G. (2016). Option convergence rate with geometric random walks approximations. *Stochastic Analysis and Applications*, *34*(5), 767–791.

Leduc, G., and Nurkanovic Hot, M. (2020). Joshi's split tree for option pricing. *Risks*, *8*(3), 81.

CS03 - Risk Analysis in new disease development

Mutation Patterns in SARS-CoV-2 Alpha, Beta and Delta Variants of Concern Indicate Non-neutral Evolution

Monika Kurpas[1], **Marek Kimmel**[2]

[1] *Silesian University of Technology, Department of Systems Biology and Engineering, Poland, monika.kurpas@polsl.pl*
[2] *Rice University, Department of Statistics, USA, kimmel@rice.edu*

Abstract

Due to the emergence of new variants of the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), the question of how the viral genomes evolved, leading to the formation of highly infectious strains, becomes particularly important. Three early emergent strains, Alpha, Beta and Delta, characterized by a significant number of missense mutations, provide natural testing samples.

In this study we are exploring the history of each of the segregating sites present in Alpha, Beta and Delta variants of concern (VOC), to address the question whether defining mutations were accumulating gradually leading to the formation of sequence characteristic of these variants or whether this phenomena can be explained by recombination of two genomes with subsets of mutations. We also check if mutation patterns observed in whole genome samples of viral variants classified as VOCs indicate the presence of stronger selection than in non-VOC samples.

The analysis was carried out using nucleotide sequences of SARS-CoV-2 genomes, downloaded from the Global Initiative on Sharing Avian Influenza Data (GISAID) database (Shu and McCauley, 2017). Segregating sites, characteristic of Alpha, Beta and Delta SARS-CoV-2 variants (see Background section) were identified from the Multiple Sequence Alignment based on comparison with reference sequence (NC_045512.2, the first sequenced SARS-CoV-2 genome from Wuhan). We quantified the change in the abundance of individual mutations over time, and studied possible combinations of 2, 3, 4 etc. mutations present together in one genome as well as the dates when such combinations arose.

For all variants, we observe that there is a large number of genomes carrying only one or two from VOC-defining mutations but – especially in case of Alpha variant – there is also a lot of sequences containing the complete

set. The least numerous are genomes having mutations in 4-6 out of all segregating sites. Moreover, we observe that genomes carrying combinations of higher number of mutations (even full set) emerge earlier than genomes carrying only some of them (e.g. combinations of 5 or 6 mutations).

We compared observed counts of unique combinations of mutations in tested samples with expected number of combinations, given the count of segregating sites. Obtained results significantly depart from expectations.

In order to check whether there is higher selection pressure among genomes belonging to the VOC strain, for each week we compared site frequency spectra (SFS) of all mutations present in VOC versus all genomes sequenced in that week of the pandemic. The results presented in the form of log-log cumulative tails indicate that the shape of SFS tails differs between sample with VOC genomes and the sample with all genomes sequenced in given week.

In this study we analysed SARS-CoV-2 genomes to see how the individual mutations that define the Alpha, Beta and Delta variants were appearing over time. Our analyses showed that these mutations did not arose gradually, but rather co-evolved rapidly leading to the emergence of the full VOC strain. We do not observe transient states which would be expected under neutral evolution. These results seem to indicate that segregating sites in Alpha, Beta and Delta variants evolved under strong positive selection. Another possible explanation might be recombination event between viruses carrying subsets of VOC-defining mutations or the possibility that such genomes avoided collection and sequencing.

Keywords

SARS-CoV-2, Mutations, Evolution, Variants of concern, Site frequency spectrum.

References

GISAID database. https://www.gisaid.org/.

Shu, Y, and McCauley, J. (2017). Gisaid: Global initiative on sharing all influenza data - from vision to reality. *Eurosurveillance*, *22*(13), 30494.

Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome. NCBI Reference Sequence: NC_045512.2.

On pitfalls in statistical analysis for assessing workplace-specific risk of severe Covid-19

Tomomi Yamada[1], **Hiroyuki Mori**[2], **Todd Saunders**[3], **Tsuyoshi Nakamura**[4]

[1] *Department of Medical Innovation, Osaka University Hospital, Osaka, Japan, tomomi.yamada@dmi.med.osaka-u.ac.jp*

[2] *Department of Life and Creative Sciences, Nagasaki Women's College, Nagasaki, Japan, mori@nagasaki-joshi.ac.jp*

[3] *Graduate School of Biomedical Science, Nagasaki University, Nagasaki, Japan, styer2000@hotmail.com*

[4] *Faculty of Environmental Science, Nagasaki University, Nagasaki, Japan, naka@nagasaki-u.ac.jp*

Abstract

Here we describe pitfalls encountered in reviewing epidemiological literature dealing with occupation or demographic risk of severe Covid-19. Following are the items to be discussed:

1. Loss-to-follow, Self-selection, and Detection biases in Cohort data
2. Measurement errors in endpoint and explanatory variables
3. Odds ratio and Relative risk
4. Significance in variable selection
5. Merging different stages in logistic models
6. Confounder and Intermediate factors in the workplace
7. Age-dependent effects

Examples from the literature will be used to illustrate the pitfalls outlined above. Using our previously published study, which used PCR-positive cohort data of all COVID-19 cases in Osaka prefecture, Japan that were not subject to selection or follow-up biases, as a reference. We will also describe appropriate, standard statistical methods to solve these problems and describe exact workplace-specific risks of severe Covid-19 in Osaka, Japan between February and September, 2020. The emphasis here is the necessity to apply appropriate statistical methods to correctly assess workplace-specific risks of severe Covid-19.

Keywords

Covid-19, Epidemiology, Workplace, Risk, Bias.

References

Drefahl, S., Wallace, M., Mussino, E., et al. (2020). A population-based cohort study of socio-demographic risk factors for COVID-19 deaths in Sweden. *Nature Communications*, *11*, 5097.

Williamson, E.J., Walker, A.J., Bhaskaran, K., et al. (2020). Factors associated with COVID-19-related death using Open SAFELY. *Nature*, *584*, 430–436.

Nguyen, L.H., Drew, D.A., Graham, M.S., et al. (2020). Risk of COVID-19 among front-line health-care workers and the general community: a prospective cohort study. *Lancet Public Health*, *5*, e475–83.

Mutambudzi, M., Niedwiedz, C., Macdonald, E.B., et al. (2020). Occupation and risk of severe COVID-19: prospective cohort study of 120 075 UK Biobank participants. *Occup Environ Med*, *78* (5), 307–314.

Nakamura, T., Mori, H., Saunders, T., et al. (2022). Impact of Workplace on the Risk of Severe COVID-19. *Front. Public Health*, *9*, 731239.

Identification of risk factors for Post-Covid-19 Syndrome using Elastic Net logistic regression

Maria De Martino[1], **Maddalena Peghin**[2], **Alvisa Palese**[3], **Carlo Tascini**[4], **Miriam Isola**[5]

[1] *Division of Medical Statistic, Department of Medicine (DAME), University of Udine, Italy, maria.demartino@uniud.it*
[2] *Infectious and Tropical Diseases Unit, Department of Medicine and Surgery, University of Insubria-ASST-Sette Laghi, Italy, maddalena.peghin@gmail.com*
[3] *Department of Medical Sciences, University of Udine, Italy, alvisa.palese@uniud.it*
[4] *Infectious Diseases Division, Department of Medicine, University of Udine and Azienda Sanitaria Universitaria Friuli Centrale (ASUFC), Italy, carlo.tascini@uniud.it*
[5] *Division of Medical Statistic, Department of Medicine (DAME), University of Udine, Italy, miriam.isola@uniud.it*

Abstract

Introduction. The "post-COVID-19 syndrom" or "chronic COVID-19 syndrom", which describes the experience of persistent symptoms after recovering from the initial acute COVID-19, is increasingly attracting attention. The assessment of the risk factors associated with this syndrome one year after the recovery is of particular interest.
Aim. Evaluate risk factors for the presence of symptoms one year after the recovery from COVID-19 applying elastic net logistic regression. This is a regularized regression method which combines the L_1 and L_2 penalties from the Lasso and Ridge regression.
Methods. The study includes 479 in- and out-patients ($\geq$ 18 years) attending the Infectious Disease Department with a diagnosis of COVID-19 from March 1st to May 30th 2020. Their demographic, clinical and laboratory information were collected. One year after acute disease onset, between March and May 2021, patients were interviewed about specific persistent or emerging symptoms potentially associated with COVID-19. Factors included in the analysis were: sex, age, BMI, smoke habit, alcohol habit, baseline comorbidities, number and type of symptoms during acute onset disease, management

of the patient (outpatients, hospital-ward and intensive care unit), acute COVID severity (WHO), viral shedding, COVID vaccination after recovery, positive IgG and IgM response during infection (IgG and IgM ≥ 10 kAU\L).

The regression was analysed on a train set, estimating the regression coefficient β for each variable appearing in the final model. A positive β denotes that the event is more likely to happen. The performance was evaluated on a test set, using the AUC index, with its 95% confidence interval, obtained through ROC analysis. Train and test sets were obtained using a 1: 1 split.
Results. Risk factors for post-Covid-19 syndrome were: number of symptoms during acute onset disease ($\beta = 0.41$), rheumatological symptoms at the onset ($\beta = 0.26$), asymptomatic infection ($\beta = -0.13$), dyspnoea at the onset ($\beta = 0.26$), female sex ($\beta = 0.07$), viral shedding ($\beta = 0.06$), gastrointestinal symptoms at the onset ($\beta = 0.26$), positive IgM response during infection ($\beta = 0.02$). The model reported an area under the curve of 0.763 (95% CI $0.702 - 0.823$).
Conclusions. Results suggest that elastic net logistic regression can represent a good method to assess risk factors for the Post-Covid-19 Syndrome. In particular this regression shows a great performance also in the case of a large number of covariances.

Keywords

Post-Covid-19 Syndrome, Elastic Net Logistic regression, L_1 penalty, L_2 penalty.

References

Huang C, Huang L, Wang Y et al. (2021). 6-month consequences of COVID-19 in patients discharged from hospital: a cohort study *Lancet*, *397*, 220–232.

Peghin M et al. (2021). Post-COVID-19 symptoms 6 months after acute infection among hospitalized and non-hospitalized patients *Clin. Micro and Inf.*, *10*, 1507–1513.

Friedman J et al. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent *J Stat Softw.*, *33*, 1–22.

Accounting for family risk in models for disease development

Maria Veronica Vinattieri[1], **Marco Bonetti**[2]

[1] *Bocconi University, Department of Decision Sciences, Italy, maria.vinattieri@phd.unibocconi.it*

[2] *Bocconi University, Dondena Research Center and Department of Social and Political Sciences, Italy, marco.bonetti@unibocconi.it*

Abstract

We are interested in a specific aspect related to disease development, i.e. the study of family-specific risk. To fix ideas, we consider breast cancer development and thus, only female family members. We focus on the genetic "from birth" risk component as opposed to the environmental component. Indeed, the genetic family risk R may be assumed to be latent and unchanged from birth, and we construct a simple model that allows for R to be discrete with two ordered risk levels (0/1 for low/high risk of breast cancer development), and we consider families as clusters within which individuals share the same risk, or frailty.

We compare the true latent R to the observed subject-specific time-varying covariate $FH(t)$, where "FH" stands for family history, that is a function of the collection of disease onset experiences in the family (among the subject and the three closest relatives: grandmother, mother and sister) at time t. We use i to identify the subject (i.e. the family) out of G and $\mathbf{i} = \{i, i_g, i_m, i_s\}$ to identify the family members: subject, grandmother, mother and sister, respectively. The observed data are $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_G)^T$, with $\mathbf{Z}_i = (\mathbf{z}_i, \mathbf{z}_{i_g}, \mathbf{z}_{i_m}, \mathbf{z}_{i_s})^T$ where, for the generic subject j, $\mathbf{z}_j = (z_j = \min(t_j, c_j), \delta_j)^T$, $\delta_j = \mathbb{1}(t_j \leq c_j)$ following the usual notation, that has t_j indicate the survival time (from birth b_j) and c_j indicate the (independent) censoring time from birth. We let c_j be the distance between b_j and the time when the data are collected (so that it is administrative right censoring).

We assume a proportional hazard (PH) structure, such that the true hazard function in the two risk groups can be written as $\lambda_0(t)$ and $\lambda_1(\mathrm{t}) = \alpha\lambda_0(\mathrm{t})$ for $R = 0$ and 1, respectively. The baseline hazard function $\lambda_0(t)$ follows a distribution characterized by a parameter θ. More generally, we assume the Lehmann structure $S_1(t) = [S_0(t)]^\alpha$, with $S_0(t)$ an improper survival function defined by the "cure rate" model $S_0(t) = p + (1-p)S^*(t)$ with

$S^*(t)$ a proper survival function. We make some additional assumptions, i.e.: (i) if the ith subject does not have a sister, $b_{i_s} = \infty$ and $t_{i_s} = \infty$; (ii) $b_{i_s} = b_i$, $b_{i_m} = b_i - 30$, $b_{i_g} = b_i - 60$ (this can be readily generalized) [1]. The model that we have described above, with conditionally independent survival times within each family, is an example of a shared frailty multivariate survival model. We call it the complete model.

An alternative, simpler model for the survival function of subject i can be built such that it is based on the use of $FH(u) = \mathbb{1}$(one or more relatives of the subject have experienced the disease by calendar time u) as follows:

$$S(T_i = t|FH_i(b_i + t)) = [S_0(t)]^{e^{\beta_F \cdot FH_i(b_i+t)}} \tag{3.1}$$

with $FH_i(b_i+t) = 1-\mathbb{1}(b_{i_g}+t_{i_g} \geq b_i+t)\mathbb{1}(b_{i_m}+t_{i_m} \geq b_i+t)\mathbb{1}(b_{i_s}+t_{i_s} \geq b_i+t)$. Note that $FH(b_i + t)$ is known given our assumption on follow-up. The parameter β_F is the observed family history modifier, and it is typically used to account for the increased family risk for subjects that have "family history" of the disease at subject's age t, i.e. calendar time $b_i + t$. In other words, $FH_i(b_i + t)$ is meant to estimate the latent genetic risk R from the observed onset histories at calendar time $b_i + t$ (in at least one of the family members). Importantly, while R takes value zero or one from birth and does not change over time, $FH(t)$ is a counting process that takes value zero at calendar time b_i and until the first onset occurs among the family members. Note that, when compared to the complete model above, Model (3.1) refers only to the survival time of the subject, adjusted for the estimated (by $FH(u)$) family-specific frailty R. In addition, Model (3.1) only requires information about the simpler quantity $FH(u)$, as opposed to all birth dates and diagnosis dates for the family members (as required by the complete model).

We first explore how $FH(t)$ and R differ through some calculations and simulated data. We then focus on the comparison between the true two-risk-group model (complete model) with its consequent estimation of α and Model (3.1) with the time-varying covariate $FH(t)$, and its consequent estimation of β_F. The relationship between the interpretation of α and β_F is clearly of interest. For example if $FH(t)$ predicts R well, i.e. they coincide with high probability, then Model (3.1) is such that the estimation of its parameter β_F is almost equivalent to the estimation of the parameter α in the complete model. In general, the predictive power of $FH(u)$ impacts how well β_F can represent the real target of inference, α.

[1] If one were interested in survival and not in disease development, one could incorporate improved survival across generations by assuming, say, $S_{T_{i_s}}(t) = S_{T_i}(t)$, $S_{T_{i_m}}(t) = [S_{T_i}(t)]^{\beta_m}$, and $S_{T_{i_g}}(t) = [S_{T_{i_m}}(t)]^{\beta_m} = [S_{T_i}(t)]^{\beta_m^2}$ with $\beta_m > 1$ to make a mother's survival less favorable than her daughter's survival.

Keywords

Survival analysis, Cancer risk analysis, Family history, Censored data, Shared frailty.

References

American cancer society. Key Statistics for Breast Cancer in Men. *https://www.cancer.org/cancer/breast-cancer-in-men/about/key-statistics.html*

Duchateau, L., Janssen, P. (2007). The frailty model. *Springer Science & Business Media*

Hougaard, P. (2012). Analysis of multivariate survival data. *Springer Science & Business Media*

Kaplanis, J., Gordon, A., Shor, T., et al. (2018). Quantitative analysis of population-scale family trees with millions of relatives. *Science*, *360*(6385), 171–175

Assessment of perceived risk for Covid-19

Silvia Bacci[1], **Rosa Fabbricatore**[2], **Maria Iannario**[3]

[1] *University of Florence, Department of Statistics, Computer Science, Applications "G. Parenti", Italy, silvia.bacci@unifi.it*

[2] *University of Naples Federico II, Department of Social Sciences, Italy, rosa.fabbricatore@unina.it*

[3] *University of Naples Federico II, Department of Political Sciences, Italy, maria.iannario@unina.it*

Abstract

Covid-19 represents a new hazard that may damage lives of individuals and societies from several perspectives (e.g., physical health, psychological well-being, life styles). In this contribution we aim at comparing the people's judgment about the perceived risk for the individual (personal riskiness) and for the community (social riskiness) of Covid-19 with risk perceived for other hazards belonging to multiple domains, such as health (Aids, Cancer, Infarction), environment (Climate change), behaviours (Serious car accidents), and technology (Nuclear weapons).

Perceived risk is a not directly observable construct (latent variable), thus its measurement requires ad hoc instruments (questionnaires) based on multiple-choice items, each of which represents observable attributes determining hazard's perceived risk. Item Response Theory (IRT; Hambleton et al., 1991; Van der Linden and Hambleton, 1997) provides the statistical methodology to summarise item responses in a unique measurement of perceived risk along a continuum scale. In this contribution we first estimate a Graded Response Model (GRM; Samejima, 1969) to analyse the positioning of Covid-19 with respect to the other hazards in terms of perceived risk. Among the considered attributes determining hazard's perceived risk, there are the knowledge of risks from the hazard, media attention, irreversibility of effects, and the harmfulness for children. Moreover, to investigate the role of some individual characteristics in affecting the level of perceived risk we also estimate a latent regression version of the GRM model (Zwinderman, 1991; Rijmen et al., 2003). Among the considered individual covariates, we take into account, among others, gender, education, economic condition, and political orientation. The latter are correlated with risk perceptions according to the literature (Lanciano et al., 2020).

Analyses are based on a dataset collected during the first Italian Covid-19 lockdown (March 18 - May 3, 2020) on a sample of $N = 2,224$ voluntary respondents coming from all Italian regions. The risk perception is measured on fourteen hazards through a self-reported questionnaire made of eight attributes rated on a 7-point Likert response scale.

Among the main results, we outline how the Covid-19 perceived riskiness is comparable to that of Aids, whereas Cancer, Nuclear weapons, and Serious car accidents are judged as riskier than Covid-19 on one side and Vaccines, Influenza and Diabetes are perceived as less risky on the other side.

As concerns the role of attributes in determining the level of perceived riskiness of the hazards, the most relevant contribution to the perceived risk come from long-term effects, irreversibility of effects, and harmfulness for children. Thus, we can explain the intermediate position of Covid-19 perceived risk with the fact that its effects are perceived as less long-terming, irreversible and harmful for children than the effects of other hazards.

In summary, the contribution aims at improving knowledge on the way people perceive the riskiness of Covid-19. The comparison with other hazards as well as the analysis of association with individual characteristics allow us to better understand how the Covid-19 risk perception is framed in the people's cognitive representation of other hazards' risk perception. The study provides insights for policymakers on individual susceptibility to the Covid-19 that can be useful to develop effective risk communication strategies and control policies oriented to enhance the citizens' awareness and sensitivity and, thus, to favor the use of protective measures.

Keywords

Covid-19, Graded response model, Item response theory, Latent regression model, Risk-perception analysis.

References

Hambleton, R.K., Swaminathan, H., and Rogers, H.J. (1991). *Fundamentals of Item Response Theory.* Sage Publications, Inc., Newbury Park.

Lanciano, T. , Graziano, G. , Curci, A., Costadura, S., and Monaco, A. (2020) Risk perceptions and psychological effects during the Italian COVID-19 emergency. *Front. Psychol.*, *11*, 2434.

Rijmen, F., Tuerlinckx, F., De Boeck, P., and Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychol. Methods*,

8, 185–205.

Samejima, F. (1969). Estimating of latent ability using a response pattern of graded scores. *Psychol. Monogr. Suppl.*, *1*, i–169.

Van der Linden, W.; Hambleton, R.K. (1997). *Handbook of Modern Item Response Theory*. Springer-Verlag, New York.

Zwinderman, A.H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika*, *56*, 589–600.

Experience of Stroke Survivors During the COVID-19 Pandemic (ESS-COV): a Phenomenological Investigation

Barbara Kimmel[1], **Jane Anderson**[2]

[1] *Baylor College of Medicine, United States, bkimmel@bcm.edu*
[2] *Baylor College of Medicine, United States, janea@bcm.edu*

Abstract

Introduction and Study Aim: Nearly 800,000 Americans suffer a stroke each year, one every 40 seconds; and stroke costs the United States 38 billion a year. Stroke is the fifth leading cause of death in the United States (US) and leads to major disability (AHA/ASA). Many risk factors are associated with stroke care. Support of treatment of acute and follow up care involves complex, evidence-based interventions to reduce patient's death, post-stroke disability, risk factors modification and stroke recurrence. Since March of 2020, COVID-19 pandemic created specific challenges on established stroke care systems worldwide. Risk associated with COVID-19 pandemic and multiple lockdowns transformed established stroke care systems and created specific challenges for patients when to seek medical care for stroke. In April of 2020, temporary emergency guidance to US stroke centers providing stroke care during the COVID-19 pandemic was published on behalf of the American Heart Association (AHA). However, little is known to what extent the individual patients experience of post-stroke care was following the implementation of these guidelines. To close the gap, we conducted a research study to describe lived experience of stroke survivors after being discharged home and to identify the risk perception associated with patients seeking post-stroke care during pandemic. The study was conducted in cooperation with participating facilities within the Lone Star Stroke Research Consortium (LSS) in Texas during the COVID-19 pandemic.

Methods and Analysis: A qualitative phenomenological methodology is employed to conduct the study. Patients who experienced stroke and were discharged from the hospital during the COVID-19 pandemic were enrolled. Using the interview guide, we conducted in-depth interviews via ZOOM, digitally recorded and transcribed sessions. Data was analyzed using Giorgi descriptive method which focuses on intentionality and is seeking the essence of the lived experience.

Preliminary Results: We enrolled 11 patients into the study and conducted in-depth interviews. First round of analysis revealed several themes. Patients expressed frustration and anger with their poststroke care coordination and their recovery process during the COVID-19 pandemic. Risk perception-seeking healthcare services were mixed with some patients seeking care via telemedicine and others preferring face-to face follow up visits. We will discuss in details additional emerging themes and specifics regarding access, telemedicine positives and negatives, and technology challenges of using telemedicine during pandemic.

Conclusion: Preliminary results of the study revealed major challenges that patients surviving stroke experience. It also demonstrated that seeking care during pandemic is associated with additional risk that patients must face for successful recovery and to obtain additional stroke prevention guidance and self-care.

Implications: We hope that completion of this study will help clinicians and policy makers to better understand stroke survivors lived experience during the COVID-19 pandemic. Knowledge gained from a pilot study of these individuals can be used in a wider context of limitations imposed by further civilization-related epidemics.

Keywords

Stroke experience, Telemedicine, Stroke risk recurrence, Post-acute stroke recovery, COVID-19 pandemic.

CS04 - Statistical and Machine learning models for risk detection

From Prediction of an event to Interpretation of Risk Factors for Neural Networks

Catherine Huber[1]

[1] *Université Paris Descartes, France, catherine.huber@parisdescartes.fr*

Abstract

Neural Networks were considered, for a long time, as "black boxes" having good prediction performances but unable to give an interpretation of the relative impact of their entry variables on the response. In our case, the entries are risk factors for the occurrence of a specific event which, in the medical field, is a disease D. However we can reach some insight into the respective impact of each risk factor by distorting the data set, doing sequential permutations of the risk factors. If the prediction performance of the Neural Network is stable, this means that the corresponding factor is irrelevant. We present examples of this phenomenon using simulations and a real data set of patients suffering from Dementia.

Keywords

Interpretation, Neural Networks, Prediction, Risk factors, Survival Analysis.

A New Metric to Deal with Off-Diagonal Elements in Confusion Matrices

I. Barranco-Chamorro[1]

[1] *University of Seville, Faculty of Mathematics, Department of Statistics and Operations Research, Spain, chamorro@us.es*

Abstract

Confusion matrices are a standard way of summarizing the performance of a classification method. This issue is of crucial interest in a variety of applied scientific disciplines, such as Geostatistics, mining data, mining text, Economy, Biomedicine or Bioinformatics, to cite only a few. A confusion matrix is obtained as the result of applying a control sampling on a dataset to which a classifier has been applied. Provided that the qualitative response to be predicted has $r \geq 2$ categories, the confusion matrix will be a $r \times r$ matrix, where the rows represent the actual or reference classes and the columns the predicted classes (or vice versa). So the diagonal elements correspond to the items properly classified, and the off-diagonal to the wrong ones. Most papers dealing with confusion matrices focus on the assessment of the overall accuracy of the classification process, such as kappa coefficient, and methods to improve these measurements, see for instance Grandini et al. (2020) and references therein. However, a scarce number of papers consider the study of the off-diagonal cells in a confusion matrix. In this paper it is shown that this topic is of interest for a better definition of classes and the improvement of the global process of classification.

Based on the results given in Barranco-Chamorro and Carrillo-García (2021), the problem of *classification bias* is introduced. This is a kind of systematic error, which happens between categories in a specific direction. If a classifier is fair or unbiased, then the errors of classification between two given categories A and B must happen randomly, that is, it is expected that they occur approximately with the same relative frequency in every direction. Quite often, this is not the case, and a kind of systematic error or bias occurs in a given direction. The classification bias can be due to deficiencies in the method of classification. For instance, it is well known, Goin (1984), that an inappropriate choice of k in the k-nearest neighbor (k-nn) classifier may produce this effect. In case of being detected, the method of selection of k must

be revised. On the other hand, the classification bias may be caused by the existence of a unidirectional confusion between two or more categories, that is, the classes under consideration are not well separated. Anyway, in case of being detected this problem, the process of classification should be improved. To identify this problem in a global way, first marginal homogeneity tests are proposed. The tests are based on Stuart–Maxwell test, (Black and Gonen, 1997), and Bhapkar test, (Sun and Yang, 2009). If the marginal homogeneity is rejected, a *One versus All* methodology is proposed, in which Mc-Nemar tests, (McNemar, 1947), are carried out for every pair of classes. Second a Bayesian method based on the Dirichlet-Multinomial distribution is developed to estimate the probabilities of confusion between the classes previously detected. So it can be assesed in a formal way, if certain classes suffer from a problem of overprediction or underprediction. To illustrate the use of our proposal, several real applications are considered. As computational tools, we highlight that the R Software and R packages are used.

Keywords

Confusion matrix, Bias of classification, Misclassification, Marginal homogeneity tests, Dirichlet distribution, Posterior density, Overprediction, Underprediction.

References

Barranco-Chamorro I., Carrillo-García R.M. (2021). Techniques to Deal with Off-Diagonal Elements in Confusion Matrices. *Mathematics*, *9*(244), 3233. https://doi.org/10.3390/math9243233

Black, S., Gonen, M. (1997). A Generalization of the Stuart-Maxwell Test. In *SAS Conference Proceedings: South-Central SAS Users Group 1997*; Applied Logic Associates, Inc.: Houston, TX, USA.

Franco M., Vivo J.M. (2021). Evaluating the Performances of Biomarkers over a Restricted Domain of High Sensitivity. *Mathematics*. *9*(21):2826. https://doi.org/10.3390/math9212826

Goin, J.E. (1984). Classification Bias of the k-Nearest Neighbor Algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.* PAMI-6, 379–381.

Grandini, M., Bagli, E., Visani, G. (2020) Metrics for Multi-Class Classification: An Overview. arXiv 2020, arXiv:2008.05756.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*, 153–157.

Pérez, C.J.; Girón, F.J.; Martín, J.; Ruiz, M.; Rojano, C. (2007). Misclassified multinomial data: A Bayesian approach. *Rev. Real Acad. Cienc. Exactas Fís. Nat. Ser. A Mat. (RACSAM)*, *101*, 71–80.

Sun, X.; Yang, Z. (2008). Generalized McNemar's Test for Homogeneity of the Marginal Distributions. *In Proceedings of the SAS Global Forum Proceedings. Statistics and Data Analysis*, San Antonio, TX, USA, 16–19 March. *382*, pp. 1–10.

Variable selection for modelling bankruptcy risk

Francesca Pierri[1]

[1] *University of Perugia, Department of Economics, Italy,*
francesca.pierri@unipg.it

Abstract

Under the adverse circumstances that have continuously tested the world economy from 2005 onwards, credit risk forecasting and bankruptcy prediction have become among the most interesting topics in the modern economic and financial field. A major challenge in constructing predictive failure models is the effective selection of variables from the large number that may have been collected because of their perceived importance or widespread use in the literature. A variety of selection methods has been developed in statistical modelling, for example, traditional stepwise procedures and more modern approaches such as the LASSO. Moreover, given the increasing availability of massive data sets, new high-performance procedures have been developed to take advantage of parallel processing in both multithreaded single-machine mode and distributed multiple-machine mode.

In this paper, various techniques were explored using a balanced data set that included only firms active or at the end of their life (that is, firms which are in the bankruptcy process or have been declared insolvent and are therefore under the protection of the law). In reality, there may be several causes of the end of a firm's life and different variables may be associated with each outcome (Caroni and Pierri, 2020; Pierri and Caroni 2020). Therefore, the choice for the current study fell upon a single adverse event, the most severe, which fortunately in real life is rare. In order to overcome the problem of misclassification errors caused by rare units (King and Zeng, 2001), a balanced data set was built by randomly extracting four controls (firms still active) for each bankruptcy case (Pierri, Stanghellini and Bistoni, 2016). Logistic regression for binary data, the most widely used technique in credit scoring models (Balcaen and Ooghe, 2006), was applied and several selection methods were compared using standard and high-performance procedures available in SAS software.

The model derived using stepwise selection was compared with those coming from the LASSO and an unsupervised variable selection method that

identifies variables that jointly explain the maximum amount of data variance. Furthermore a non-parametric approach was considered and the selections of variables coming from a single decision tree and a random forest are discussed and compared.

Keywords

Variable selection, Default probability, Decision tree, Random forest tree, Logistic regression, Balanced data, ROC Curve.

References

Balcaen, S., and H. Ooghe. (2006). 35 years of Studies on Business Failure: An Overview of the Classic Statistical Methodologies and their Related Problems. *British Accounting Review*, *38*:63–3.

Caroni, C. and Pierri, F. (2020). Different causes of closure of small business enterprises: alternative models for competing risks survival analysis. *Electronic Journal of Applied Statistical Analysis* , *13*(1), 211–228.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis* , *9*, 137-163.

Pierri, F. and Caroni, C.(2020), Analysing the Risk of Bankruptcy of Firms: Survival Analysis, Competing Risks and Multistate Models, 385-394. 10.1007/978-3-030-44695-6_25

Pierri, F., Stanghellini, E., and Bistoni, N. (2016). Risk analysis and retrospective unbalanced data. *Revstat* , *14*, 157–169.

CS05 - Advanced Statistical Models for Risk Evaluation

Mortality smoothing and forecasting under deterministic opinions

Viani Biatat Djeundje[1]

[1] *University of Edinburgh, Business School, viani.djeundje@ed.ac.uk*

Abstract

Modelling and forecasting mortality plays a crucial role in longevity risk quantification. Various mortality projection models have been proposed and developed successfully, including Lee-Carter model and its extensions, P-splines smooth models, Cairns-Blake-Dowd model and its extensions, etc. In the first part of this presentation, I will introduce a tool that allows to fit and explore (through few clicks) all the major mortality forecasting models published in the literature.

In practice nonetheless, forecasts produced by many mortality models are extrapolations of past trends seen in the data. As such, these models are unable to account for some external or expert opinions. In the second part of the talk, I will then present a way to incorporate deterministic opinions into smooth mortality projection models. Not only does this approach yield a smooth transition from the past into future, but also, the shapes of the resulting forecasts are governed by a combination of the speed of improvements observed in the data and the opinion inputs. In addition, this approach offers the possibility to compute the amount of uncertainty around the projected mortality trends conditional on the opinion inputs, and this allows to highlight some of the pitfalls of deterministic projection methods.

Keywords

Mortality modelling, Forecasting, Expert opinions, Smoothing, Conditional uncertainty.

References

Djeundje, B. V. (2022). On the integration of deterministic opinions into mortality smoothing and forecasting. *Annals of Actuarial Science*, https://doi.org/10.1017/S1748499521000282.

Taxonomy-based Risk Analysis with a Digital Twin

Giovanni Paolo Sellitto[1], **Tanja Pavleska**[2], **Massimiliano Masi**[3], **Helder Aranha**[4]

[1] *Independent Scholar, Italy, gogiampaolo@gmail.com*
[2] *Jozef Stefan Institute, Slovenia, atanja@e5.ijs.si*
[3] *Independent Scholar, Italy, max@mascanc.net*
[4] *Independent Scholar, Portugal, hmspider@gmail.com*

Abstract

Although the concept of Digital Twin is gaining ground as a tool to perform risk analysis in various sectors, the transition from a set of practices to a methodology to perform risk-analysis using a Digital Twin of a cyber-physical system is still ongoing. Many aspects remain open to a complete systematization, like how to ensure the fidelity of the digital twin with respect to the physical part, the physical-to-virtual connection, and its maintenance over the system lifecycle. This is especially true in the case of critical infrastructures and pervasive smart environments, where some times performing security tests on the real systems is not an option.

There are several definitions for a Digital Twin: we will borrow the definition of the Digital Twin as a "virtual replica of the system that accompanies its physical counterpart during its life-cycle, consumes real-time data if required, and has the sufficient fidelity to allow the implementation, testing, and simulation of desired security measures and business continuity plans".

This kind of representations can be used to perform simulations and model-based risk analysis, to assess whether the system is secure or if a recovery plan is effective. The methodology that we illustrate here supports cost-effectiveness analysis to find the balance between security and usability, safety, functionality and pay-off. In this manner, through the creation of a digital twin, the application of risk and cost-effectiveness analyses leads to quantitative results and allows to improve the system, applying the devised mitigation measures.

In this work, we propose a methodology that starts from a map of the system and produces a taxonomy of the assets to be protected, enabling the production of a Digital Twin as a translation of this taxonomy into the domain-specific language used in a Visual Threat Modeling and Simulation

environment to perform simulations and devise countermeasures. This approach enables the design of a digital twin that supports safety and security evaluations and the identification of countermeasures without any outage of the operational system.

Quantitative risk analysis is performed using the digital twin, allowing to propose and to evaluate risk mitigation from a cost-effectiveness point of view. This approach allows the identification of a set of adequate countermeasures depending on the chosen security posture for the real-world system. Using the information on the security countermeasures to be implemented, we can harden the real-world twin and close the cycle.

To demonstrate the practical usefulness of the proposed methodology, we will present a practical application of the digital twin based risk analysis, devising and evaluating some countermeasures in the case of a Smart Environment. For this purpose, in order to perform quantitative risk analysis, we will use a Visual Threat Modeling Environment, SecuriCAD. The relevant assets and the relationships among them will be represented using a domain specific modeling language, which is suited for security application and, more specifically, for visual attack simulation and threat modeling.

Keywords

Risk Analysis, Digital Twin, Threat Modeling, Simulations, Smart Environments.

References

Johnson, C., T.J. Laffey, and R. Loewy (1996). The real and the symmetric nonnegative inverse eigenvalue problems are different. *Proc. Amer. Math. Soc.*, *124*, 3647–3651.

Toler, J.E.; Burrows, P.M. (1998). Genotype performance over environmental arrays: a non-linear grouping protocol. *Journal of Applied Statistics*, *25*, n.1, 131–143.

Alternative reliable ways to manage risks of extreme events

M. Ivette Gomes[1,2], **Fernanda Figueiredo**[2,3], **Lígia Henriques-Rodrigues**[4]

[1] *DEIO, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[2] *Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL)*
[3] *Faculdade de Economia, Universidade do Porto*
[4] *Departamento de Matemática e CIMA, Universidade de Évora, Portugal*

Abstract

In the field of statistical extreme value theory, a great variety of alternative methodologies are available to deal with the management of risks of extreme events. Indeed, an important situation in risk management is the risk of a big loss that occurs very rarely. The risk is generally expressed either by the *value at risk* at a level q (VaR_q), the size of the loss occurred with a fixed small probability, q, defined for a random variable X, with a *cumulative distribution function* (CDF) $F(x) = \mathbb{P}(X \le x)$, as the q-quantile $\mathrm{VaR}_q := Q(q)$, with $Q(q) := \inf\{x \ge 0 : F(x) \ge q\}$, $q \in (0,1)$, or by the *conditional tail expectation* (CTE), defined as $\mathrm{CTE}_q = \mathbb{E}(X|X > Q(q))$, $q \in (0,1)$. The CTE measure is more conservative than VaR. Moreover, and contrarily to the VaR, the CTE is a coherent risk measure (Artzner *et al.*, 1999). The value of q is often smaller than $1/n$, where n denotes the size of the available sample, $\underline{\mathbf{X}}_n = (X_1, \ldots, X_n)$. Let us use the notation $X_{1:n} \le \cdots \le X_{n:n}$ for the associated ascending order statistics and assume that the extremal types theorem holds for $X_{n:n}$ (Gnedenko, 1943), i.e. the limiting CDF of $X_{n:n}$, linearly normalized, is necessarily of the type of the *general extreme value* (GEV) CDF, $\mathrm{EV}_\xi(x)$, $\xi \in \mathbb{R}$. The CDF F is then said to belong to the max-domain of attraction of GEV_ξ, and we write $F \in \mathcal{D}_{\mathcal{M}}(\mathrm{GEV}_\xi)$.The parameter ξ is the *extreme value index* (EVI), the primary parameter of extreme events. We consider heavy-tailed models, i.e. Pareto-type underlying CDFs, with a positive EVI, working in ${\mathcal{D}_{\mathcal{M}}}^+ := \mathcal{D}_{\mathcal{M}}(\mathrm{GEV}_{\xi>0})$. These heavy-tailed models are quite common in many areas of application, like biostatistics, finance, insurance and telecommunications, among others. For these Pareto-type models, the classical EVI-estimators are the Hill (H) estimators (Hill, 1975), $\mathrm{H}(k) \equiv \mathrm{H}(k; \underline{\mathbf{X}}_n) := \frac{1}{k}\sum_{i=1}^{k} \ln X_{n-i+1:n} - \ln X_{n-k:n}$, $1 \le k < n$. Necir *et al.*

(2010) considered the CTE-estimators

$$\widetilde{\mathrm{CTE}}_q(k) := \frac{1}{1-q}\int_q^{1-k/n} Q_n(s)ds + \frac{kX_{n-k:n}}{n(1-q)(1-\mathrm{H}(k))},$$

where $Q_n(s)$ is the empirical quantile function, which is equal to the ith order statistic $X_{i:n}$ for all $s \in ((i-1)/n, i/n)$, and for all $1 \leq i \leq n$ (see also, Laidi *et al.*, 2020). Since $\mathrm{H}(k)$ can be replaced in the previous formula by any consistent EVI-estimator, we now suggest an improvement in the performance of the aforementioned CTE-estimators, through the use of a reliable EVI–estimator based on generalized means, and dependent on an extra tuning real parameter (see Caeiro *et al.*, 2016; Penalva *et al.*, 2016, 2020a,b; Paulauskas and Vaičiulis, 2017, and references therein).

Keywords

Conditional tail expectation, Generalized means, Heavy-tailed parents, Risk modeling, Semi-parametric estimation.

References

Artzner, P., F. Delbaen, J-M. Eber, and D. Heath (1999). Coherent measures of risk. *Math. Finance*, *9*, 203–228.

Caeiro, F., M.I. Gomes, J. Beirlant, and T. de Wet (2016). Mean-of-order p reduced-bias extreme value index estimation under a third-order framework. *Extremes*, *19*:4, 561–589.

Gnedenko, B.V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. Math.*, *44*, 423-453.

Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist*, *3*, 1163-1174.

Laidi, M., A. Rassoul, and H. Old Rouis (2020). *Improved Estimator of the Conditional Tail Expectation in the case of heavy-tailed losses.* arXiv:2002.o3414v1.

Necir, A., A. Rassoul, and R. Zitikis (2010). Estimating the conditional tail expectation in the case of heavy-tailed losses. *Journal of Probability and Statistics*, doi:10.1155/2010/596839

Paulauskas, V. and M. Vaičiulis (2017). A class of new tail index estimators. *Annals Institute of Statistical Mathematics*, *69*:2, 461–487.

Penalva, H., F. Caeiro, M.I. Gomes, and M.M. Neves (2016). *An Efficient Naive Generalisation of the Hill Estimator: Discrepancy between Asymptotic and Finite Sample Behaviour.* Notas e Comunicações CEAUL 02/2016.

Penalva, H., M.I. Gomes, F. Caeiro and M.M. Neves (2020a). A couple of non reduced bias generalized means in extreme value theory: an asymptotic comparison. *Revstat–Statist. J.*, *18*:3, 281–298.

Penalva, H., M.I. Gomes, F. Caeiro, and M.M. Neves (2020b). Lehmer's mean-of-order-p extreme value index estimation: a simulation study and applications. *J. Applied Statistics*, *47*:13-15 (*Advances in Computational Data Analysis*), 2825–2845.

CS06 - Risk Analysis in Applied Science

Environmental, Social, Governance scores and the Missing pillar - Why does missing information matter?

Özge Sahin[1], **Karoline Bax**[2], **Claudia Czado**[3], **Sandra Paterlini**[4]

[1] *Department of Mathematics, Technical University of Munich, Germany, ozge.sahin@tum.de*

[2] *Department of Economics and Management, University of Trento, Italy, karoline.bax@unitn.it*

[3] *Department of Mathematics, Technical University of Munich, Germany, cczado@ma.tum.de*

[4] *Department of Economics and Management, University of Trento, Italy, sandra.paterlini@unitn.it*

Abstract

Environmental, Social, and Governance (ESG) scores measure companies' performance concerning sustainability and are organized on three pillars: Environmental, Social, and Governance. These complementary non-financial ESG scores should provide information about companies' ESG performance and risks. However, the extent of not yet published ESG information makes the reliability of ESG scores questionable. To explicitly capture the not yet published information on ESG category scores, a new pillar, the so-called Missing (M) pillar, is proposed and added into the new definition of the Environmental, Social, Governance, and Missing (ESGM) scores. By relying on the data provided by Refinitiv, we show that the ESGM scores strengthen the companies' risk relationship. These new scores could benefit investors and practitioners as ESG exclusion strategies using only ESG scores might exclude assets with a low score solely because of their missing information and not necessarily because of a low ESG merit.

Keywords

Disclosure, ESG investment, ESG methodology, Missing data, Sustainable finance, Value-at-Risk.

Estimation of Conditional Value at Risk under Measurement Errors

J. Jurečková[1,2], **Jan Večeř**[3]

[1] *The Czech Academy of Sciences, Institute of Information Theory and Automation, Czech Republic, jureckova@utia.cas.cz*

[2] *Charles University, Faculty of Mathematics and Physics, Czech Republic, jurecko@karlin.mff.cuni.cz*

[3] *Charles University, Faculty of Mathematics and Physics, Czech Republic, vecer@karlin.mff.cuni.cz*

Abstract

We consider estimating a specific coherent risk measure, namely the alpha-conditional value at risk, in the situation that the incurred loss X is overshadowed by an unobservable additive measurement error with an unknown probability distribution. The repeated experiment leads to independent values X, but they are not directly observable. The only available observations are values Z contaminated by an error of an unknown intensity. In the nonparametric situation, we approximate risk measure using the quantile criterion derived in Bassett et al. (2004). If only the distribution of measurement errors is unknown, we try to cover the family of distribution of Z by a suitable Choquet capacity, and base the risk measure on the capacity, or on the least favorable distribution of the family. Specifically, we use the capacity derived in Broniatowski et al. (2018) which turns out being a probability measure.

Keywords

Coherent risk measure, Conditional value at risk, Choquet capacity, Least favorable distribution, Distortion function.

References

Basett, G.W., Jr., Koenker, R., Kordas, W. (2004). Pessimistic portfolio allocation and Choquet Expected Utility. *Journal Financial Economics*, *2/4*, 477–492.

M. Broniatowski, J. Jurečková and J. Kalina (2018). Likelihood Ratio Testing Under Measurement Errors. *Entropy*, *20*, 966.

Choquet, G. (1953-4). *Theory of Capacities*, *Annales de I' Institut Fourier* (Grenoble), 131–295.

Huber, P., Strassen, V. (1973). Minimax tests and the Neyman-Pearson lemma for capacities. *Ann. Statist*, *2*, 251–273.

Moving average options: Machine Learning and Gauss-Hermite quadrature for a double non-Markovian problem

[1] *Fédération de Mathématiques de CentraleSupélec, France, ludovic.goudenege@math.cnrs.fr*
[2] *Università degli Studi di Udine, Dipartimento di Scienze Economiche e Statistiche, Italy, andrea.molent@uniud.it*
[3] *Università degli Studi di Udine, Dipartimento di Scienze Economiche e Statistiche, Italy, antonino.zanette@uniud.it*

Abstract

Evaluating moving average options is a tough computational challenge for the energy and commodity market as the payoff of the option depends on the prices of a certain underlying observed on a moving window so, when a long window is considered, the pricing problem becomes high dimensional. We present an efficient method for pricing Bermudan style moving average options, based on Gaussian Process Regression and Gauss-Hermite quadrature, thus named GPR-GHQ. Specifically, the proposed algorithm proceeds backward in time and, at each time-step, the continuation value is computed only in a few points by using Gauss-Hermite quadrature, and then it is learned through Gaussian Process Regression. We test the proposed approach in the Black-Scholes model, where the GPR-GHQ method is made even more efficient by exploiting the positive homogeneity of the continuation value, which allows one to reduce the problem size. Positive homogeneity is also exploited to develop a binomial Markov chain, which is able to deal efficiently with medium-long windows. Secondly, we test GPR-GHQ in the Clewlow-Strickland model, the reference framework for modeling prices of energy commodities. Then, we investigate the performance of the proposed method in the Heston model, which is a very popular model among the non-Gaussian ones. Finally, we consider a challenging problem which involves double non-Markovian feature, that is the rough-Bergomi model. In this case, the pricing problem is even harder since the whole history of the volatility process impacts the future distribution of the process. The manuscript includes a numerical investigation, which displays that GPR-GHQ is very accurate in pricing and computing the Greeks with respect to the Longstaff-Schwartz method and it is able to handle options with a very long window, thus overcoming the problem of high dimensionality.

Keywords

Moving average options, Gaussian Process Regression, Gauss-Hermite quadrature, Hedging.

An Application of D-vine Regression for the Identification of Risky Flights in Runway Overrun

Hassan Alnasser[1], **Claudia Czado**[2]

[1] *Technical University of Munich, Faculty of Mathematics, Germany, h.alnasser@tum.de*

[2] *Technical University of Munich, Faculty of Mathematics and Munich Data Science Institute, Germany, cczado@ma.tum.de*

Abstract

In aviation safety, runway overruns are of great importance because they are the most frequent type of landing accidents. Identification of factors which contribute to the occurrence of runway overruns can help mitigate the risk and prevent such accidents. Methods such as physics-based and statistical-based models were proposed in the past to estimate runway overrun probabilities. However, they are either costly or require experts' knowledge. We propose a statistical approach to quantify the risk probability of an aircraft to exceed a threshold at the speed of 80 knots given a set of influencing factors. This copula based D-vine regression approach is used because it allows for complex tail dependence and is computationally tractable. Data obtained from Quick Access Recorder (QAR) for 711 flights are analyzed. We identify 41 flights with an estimated risk probability $> 10^{-3}$ for a chosen threshold and rank the effects of each influencing factor for these flights. Also, the complex dependency patterns between some influencing factors for the 41 flights are shown to be non symmetric. The D-vine regression approach, compared to physics-based and statistical-based approaches, has an analytical solution, is not simulation based and can be used to estimate very small or large probabilities efficiently.

Keywords

Conditional risk probability estimation, D-vine regression, Non Gaussian dependency, Runway overrun.

References

Au, S. K., Beck, J. L. (2001). Estimation of small failure probabilities in high dimensions by subset simulation. *Probabilistic engineering mechanics*, *16*(4), 263–277.

Ayra, E. S., Ríos Insua, D., Cano, J. (2019). Bayesian network for managing runway overruns in aviation safety. *Journal of Aerospace Information Systems*, *16*(12), 546–558.

Czado, C. (2019). Analyzing dependent data with vine copulas. *Lecture Notes in Statistics, Springer.*

Czado, C., Nagler, T. (2022). Vine copula based modeling. *Annual Review of Statistics and Its Application, 9.*

Drees, L. (2016). *Predictive Analysis: Quantifying Operational Airline Risks* (Doctoral dissertation, Technische Universität München).

Kraus, D., Czado, C. (2017). D-vine copula based quantile regression. *Computational Statistics & Data Analysis*, *110*, 1-18.

Wang, X., Fang, X., Beller, L., Holzapfel, F. (2020). Calibration of Contributing Factors for Model-Based Predictive Analysis Algorithm using Polynomial Chaos Expansion Methods.

Robo-advisors: A big data challenge

Federico Severino[1], **Sébastien Thierry**[2]

[1] *Université Laval, Department of Finance, Insurance and Real Estate, Canada, federico.severino@fsa.ulaval.ca*

[2] *Université Laval, Faculty of Business Administration, MBA Finance, Canada, sebastien.thierry.1@ulaval.ca*

Abstract

At the frontier between personal finance and Fintech, robo-advisors aim to provide customized portfolio strategies without human intervention. They typically propose passive strategies that can match the investor's objectives and risk profile at low cost. However, digital advisors feature a lack of precision in capturing clients' attitude towards risk and a (not always suitable) low risk exposure. In this context, leveraging big data and artificial intelligence techniques can improve the main strength of robo-advisors, that is their ability to automatically provide personalized investment solutions. Text data from dialogue systems, such as chatbots, can be employed to improve the client's profiling, while recommendation systems can rely on big data from financial social networks to propose targeted investment strategies. Analysis of big data through machine learning methods can also improve the performance of the optimization algorithms employed by digital advisors. The potential for the exploitation of big data and artificial intelligence in automated asset management is still enormous.

Keywords

Robo-advisors, Fintech, Portfolio management, Big data, Artificial intelligence.

Participant list

Maher Ahmed - Top engineers (Egypt)

Hassan Alnasser - Technical University of Munich (Germany)

Alessandra Amendola - University of Salerno (Italy)

Luca Vincenzo Ballestra - Alma Mater Studiorum University of Bologna (Italy)

Inmaculada Barranco-Chamorro - University of Seville (Spain)

Kristijan Breznik - International School for Social and Business Studies (Slovenia)

Vincenzo Candila - University of Salerno (Italy)

Flavia Carle - Marche Polytechnic University (Italy)

Elisabete Carolino - Lisbon School of Health Technology (Portugal)

Chris Caroni - National Technical University of Athens (Greece)

Clara Cordeiro - University of Algarve and CEAUL (Portugal)

Marco Costa - University of Aveiro (Portugal)

Alessandra Cretarola - University of Perugia (Italy)

Roberta De Vito - Brown University (United States)

Katiuscia Di Biagio - Regional Environmental Protection Agency of Marche (Italy)

Gioia Di Credico - University of Trieste (Italy)

Francesca Di Iorio - University of Naples Federico II (Italy)

Valeria Edefonti - University of Milan (Italy)

Silvia Facchinetti - Catholic University of Milan (Italy)

Andrea Faragalli - Marche Polytechnic University (Italy)

Daniel Farewell - Cardiff University (United Kingdom)

Gianna Figà-Talamanca - University of Perugia (Italy)

Lidia Z. Filus - Northeastern Illinois University (United States)

Jerzy K. Filus - Oakton College (United States)

Anna Maria Fiori - University of Milan-Bicocca (Italy)

Alessandro Fontanarosa - Marche Polytechnic University (Italy)

Maximilian Maurice Gail - Justus Liebig University Giessen (Germany)

Stefania Galimberti - University of Milan-Bicocca (Italy)

Beatrice Gasperini - Marche Polytechnic University (Italy)

Rosaria Gesuita - Marche Polytechnic University (Italy)

Giuseppe Giordano - University of Salerno (Italy)

Paolo Giudici - University of Pavia (Italy)

Rosanna Grassi - University of Milan-Bicocca (Italy)

Sneh Gulati - Florida International University (United States)

Lewitschnig Horst - Infineon Technologies (Austria)

Catherine Huber - Paris Descartes University (France)

Marica Iommi - Marche Polytechnic University (Italy)

Eman Khattab - Cairo University (Egypt)

Barbara Kimmel - Baylor College of Medicine (United States)

Marek Kimmel - Rice University (United States)

Christos P. Kitsos - University of West Attica (Greece)

Phil-Adrian Klotz - Justus Liebig University Giessen (Germany)

Andrew Koval - Rice University (United States)

Monika Kurpas - Silesian University of Technology (Poland)

Guillaume Leduc - American University of Sharjah (United Arab Emirates)

Magda Monteiro - University of Aveiro (Portugal)

Manuela Neves - School of Agriculture and CEAUL (Portugal)

Amílcar Oliveira - University of Aberta (Portugal)

Teresa A. Oliveira - University of Aberta (Portugal)

Silvia Angela Osmetti - Catholic University of Milan (Italy)

Alessandro Palandri - University of Firenze (Italy)

Elisabetta Petracci - IRCCS Istituto Romagnolo per lo Studio dei Tumori (IRST) "Dino Amadori" (Italy)

Danilo Petti - University of Salerno (Italy)

Claudia Piechl - Infineon Technologies (Austria)

Francesca Pierri - University of Perugia (Italy)

Francesco Porro - University of Genova (Italy)

Emanuela Raffinetti - University of Pavia (Italy)

Giancarlo Ragozini - University of Naples Federico II (Italy)

Maria do Rosário Ramos - University of Aberta (Portugal)

Marialuisa Restaino - University of Salerno (Italy)

Özge Sahin - Technical University of Munich (Germany)

Francesco Santelli - University of Naples Federico II (Italy)

Mahmoud Shaban - Cairo University (Egypt)

Milan Stehlik - Johannes Kepler University Linz (Austria)

M. Filomena Teodoro - CINAV, Portuguese Naval Academy (Portugal)

Marija Tepegjozova - Technical University of Munich (Germany)

Paolo Trerotoli - University of Bari (Italy)

Cristian Usala - University of Cagliari and CRENoS (Italy)

Simona Villani - University of Pavia (Italy)

Maria Veronica Vinattieri - Bocconi University (Italy)

Maria Prosperina Vitale - University of Salerno (Italy)

Agata Wilk - Silesian University of Technology (Poland)

Antonella Zambon - University of Milan-Bicocca (Italy)

Antonino Zanette - University of Udine (Italy)