



Book of the Short Papers

Editors: Francesco Maria Chelli, Mariateresa Ciommi, Salvatore Ingrassia, Francesca Mariani, Maria Cristina Recchioni



UNIVERSITÀ
POLITECNICA
DELLE MARCHE



LIUC | BUSINESS
ANALYTICS AND
DATA SCIENCE HUB



CHAIRS

Salvatore Ingrassia (Chair of the Program Committee) - *Università degli Studi di Catania*

Maria Cristina Recchioni (Chair of the Local Organizing Committee) - *Università Politecnica delle Marche*

PROGRAM COMMITTEE

Salvatore Ingrassia (Chair), Elena Ambrosetti, Antonio Balzanella, Matilde Bini, Annalisa Busetta, Fabio Centofanti, Francesco M. Chelli, Simone Di Zio, Sabrina Giordano, Rosaria Ignaccolo, Filomena Maggino, Stefania Mignani, Lucia Paci, Monica Palma, Emilia Rocco.

LOCAL ORGANIZING COMMITTEE

Maria Cristina Recchioni (Chair), Chiara Capogrossi, Mariateresa Ciommi, Barbara Ermini, Chiara Gigliarano, Riccardo Lucchetti, Francesca Mariani, Gloria Polinesi, Giuseppe Ricciardo Lamonica, Barbara Zagaglia.

ORGANIZERS OF INVITED SESSIONS

Pierfrancesco Alaimo Di Loro, Laura Anderlucci, Luigi Augugliaro, Ilaria Benedetti, Rossella Berni, Mario Bolzan, Silvia Cagnone, Michela Cameletti, Federico Camerlenghi, Gabriella Campolo, Christian Capezza, Carlo Cavicchia, Mariateresa Ciommi, Guido Consonni, Giuseppe Ricciardo Lamonica, Regina Liu, Daniela Marella, Francesca Mariani, Matteo Mazziotta, Stefano Mazzuco, Raya Muttarak, Livia Elisa Ortensi, Edoardo Otranto, Ilaria Prosdocimi, Pasquale Sarnacchiaro, Manuela Stranges, Claudia Tarantola, Isabella Sulis, Roberta Varriale, Rosanna Verde.

FURTHER PEOPLE OF LOCAL ORGANIZING COMMITTEE

Elisa D'Adamo, Christian Ferretti, Giada Gabbianelli, Elvina Merkaj, Luca Pedini, Alessandro Pionati, Marco Tedeschi, Francesco Valentini, Rostand Arland Yebetchou Tchounkeu

Technical support: Matteo Mercuri, Maila Ragni, Daniele Ripanti

Copyright © 2023

PUBLISHED BY PEARSON

WWW.PEARSON.COM

ISBN 9788891935618AAVV

Contents

Preface	XXII
1 Plenary Sessions	1
Inequality indices: accurate simulation-based inference Maria-Pia Victoria-Feser	2
Examples from the Interface of Neural Models and Spatio-Temporal Statistics in Environmental Applications Christopher K. Wikle, Likun Zhang, Myungsoo Yoo and Xiaoyu Ma	7
Demographic change and sustainability: novel approaches from digital and computational demography Emilio Zagheni	n.a.
2 Invited Sessions	14
Machine learning in the design, analysis and integration of sample surveys	
Causal Discovery for complex survey data Paola Vicard	15
Data Integration without conditional independence: a Bayesian Networks approach Pier Luigi Conti, Paola Vicard and Vincenzina Vitale	21
Mass imputation through Machine Learning techniques in presence of multi-source data Fabrizio De Fausti, Marco Di Zio, Romina Filippini and Simona Toti	27
Machine learning: different uses and perspectives	
Evaluation of pollution containment policies in the US and the role of machine learning algorithms Marco Di Cataldo, Margherita Gerolimetto, Stefano Magrini and Alessandro Spiganti	32

Machine Learning for Official Statistics: An Application on External Trade	n.a.
Mauro Bruno, Maria Serena Causo, Alessio Guandalini, Francesco Ortame and Silvia Russo	
Machine learning, data quality and official statistics: challenges and opportunities	n.a.
Stefano Menghinello	

Statistical Machine Learning for environmental applications

Gaussian Processes and Deep Neural Networks for Spatial Prediction	38
Alex Cucco, Luigi Ippoliti, Nicola Pronello, Pasquale Valentini and Carlo Zaccardi	
How can we explain Random Forests in a spatial framework?	42
Natalia Golini, Luca Patelli and Xavier Barber	
Recent approaches in coupling deep learning methods with the statistical analysis of spatial point patterns	48
Jorge Mateu and Abdollah Jalilian	

Statistical Process Monitoring for Complex Data in Industry 4.0

A Kernel-based Nonparametric Multivariate CUSUM for Location Shifts	53
Konstantinos Bourazas, Konstantinos Fokianos, Christos Panayiotou and Marios Polycarpou	
An Approach for Profile Monitoring via Mixture Regression Models	58
Davide Forcina, Antonio Lepore and Biagio Palumbo	
Anomaly Detection in Circular Data	63
Houyem Demni and Giovanni C. Porzio	

Advances in Data Science and Statistical Learning [IMS Invited Session]

Empirical Bayes approximation of Bayesian learning: understanding a common practice	n.a.
Sonia Petrone	
Generalized Fiducial Inference on Differentiable Manifolds - a geometric perspective	n.a.
Jan Hannig	
Model-free bootstrap and conformal prediction in regression	n.a.
Dimitris Politis	

ENBIS Session: System Maintenance, Boosting algorithms for regression, and Research Excellence

Boosting Diversity in Regression Ensembles	69
Mathias Bourel, Jairo Cugliari, Yannig Goude and Jean-Michel Poggi	
How ENBIS has contributed to the UK Universities Research Excellence Framework	71
Shirley Coleman	
Maintenance of degrading systems by dynamic programming or reinforcement learning	75
Antonio Pievatolo	

Population Dynamics, Climate Change and Sustainability

- Climate change impacts on fertility in low- and middle-income countries: An analysis based on global sub-national data n.a.
Côme Cheritel, Roman Hoffmann and Raya Muttarak
- Environmental Exposures and Under-5 Mortality in India: A Survival Analysis of DHS data 79
Vinod Joseph Kannankeril Joseph
- The impact of temperature on expressed sentiment by migration status: Evidence from geo-located Twitter data 84
Risto Conte Keivabu and Jisu Kim

Statistical Learning for health research and omics data

- An alternative to the Dirichlet-multinomial regression model for microbiome data analysis 95
Roberto Ascari, Sonia Migliorati and Andrea Ongaro
- Modelling ordinal response to treatment in a real-world cohort study 101
Marco Alfò, Maria Francesca Marino and Silvia D'Elia
- On the application of the symmetric graphical lasso for paired data 105
Saverio Ranciati and Alberto Roverato

The Economic behaviour of Sustainability

- Airports performances and sustainable practices. An empirical study on Italian data 110
Riccardo Gianluigi Serio, Maria Michela Dickson, Diego Giuliani and Giuseppe Espa
- Sustainability: still an undefined concept for Italians 116
Raffaele Angelone and Andrea Marletta
- Quasi-experimental evidence on COVID-19 lockdown effects on Italian household food shopping basket composition and its sustainability 122
Beatrice Biondi and Mario Mazzocchi

Advances in statistical methods for complex problems

- Inferring multiple treatment effects from observational studies using confounder importance learning n.a.
Omiros Papaspiliopoulos
- Path analysis in Ising models: an application to cyber-security risk assessment 127
Monia Lupparelli and Giovanni M. Marchetti
- Causal Regularization n.a.
Lucas Kania and Ernst Wit

Explainable machine learning models

- Enhancing Markowitz model: inspection of correlations and tail covariances 133
Gloria Polinesi

Objective and subjective dimension of economic well-being: an approach based on statistical matching	139
Daniela Marella, Vincenzina Vitale and Pierpaolo D'Urso	
Sustainable, Accurate, Fair and Explainable Machine Learning Models	n.a.
Paolo Giudici and Emanuela Raffinetti	
Flexible Learning for Environmental Sustainability	
Comparison of traffic flow data sources for air pollution modelling	145
Theresa Smith and Nick McCullen	
Data analysis of photogrammetry-based mapping: the sea cucumbers in the Giglio Island as a case-study	150
Gianluca Mastrantonio, Daniele Ventura, Edoardo Casoli, Arnold Rakaj, Giovanna Jona Lasinio and Alessio Pollice	
Understanding forest damage in Germany: Finding key drivers to help with future forest conversion of climate sensitive	156
Nicole Augustin, Heike Puhlmann and Simon Trust	
Inequalities in higher education outcomes: learning from data	
Inequalities in international students mobility	163
Kristijan Breznik, Giancarlo Ragozini and Marialuisa Restaino	
Uncovering the interplay of territorial, socioeconomic, and demographic factors in high school to university transition	169
Vincenzo Giuseppe Genova, Andrea Priulla and Martina Vittorietti	
Statistical Learning of demographic and health dynamics	
Estimating the impact of a vaccine mandate: the case of measles in Italy	n.a.
Chiara Chiavenna	
Leveraging deep neural networks to estimate age-specific mortality from life expectancy at birth	n.a.
Andrea Nigri	
Nowcasting Daily Population Displacement in Ukraine through Social Media Advertising Data	n.a.
Claire Dooley, Ridhi Kashyap, Douglas Leasure and Francesco Rampazzo	
Challenges towards Fairness and Transparency for Data Processes, Algorithms and Decision-Support Models	
Challenges on Ethics, and Privacy in AI Applications to Fintech	175
Catarina Silva, Joana Matos Dias and Bernardete Ribeiro	
Uncertainty and fairness metrics	180
Anna Gottard	

Educational Data mining: methods for complex data in students' assessment

Analysis of University Grades: An IRT Model for Responses and Response Times with Censoring 186
Michela Battauz

Predicting high schools' students performances with registry's data: a machine learning approach 191
Lidia Rossi, Marta Cannistrà and Tommaso Agasisti

Using response times to identify cheaters in CAT: A simulation study 195
Luca Bungaro, Bernard P. Veldkamp and Mariagiulia Matteucci

Spatial and Spatio-Temporal Modeling: Theory and Applications

A geostatistical investigation of the ammonia-livestock relationship in the Po Valley, Italy 200
Paolo Maranzano, Kelly McConville, Philipp Otto and Felicetta Carillo

Bayesian multi-species N-mixture models for large scale spatial data in community ecology 206
Michele Peruzzi

Minimum contrast for point processes' first-order intensity estimation 212
Nicoletta D'Angelo and Giada Adelfio

Statistical Framework for Measuring the Sustainability of Tourism

Data validity and statistical conformity with Benford's Law: the case of tourism in Sicily 217
Roy Cerqueti and Davide Provenzano

Exploring the level of digitalization of the Italian museums through a multilevel ordered logit model 228
Claudia Cappello, Sabrina Maggio and Sandra De Iaco

Functional Partial Least-Squares via Regression Splines. An application on Italian Sustainable Development Goals data 232
Ida Camminatiello, Rosaria Lombardo, Jean-Francois Durand and Leonardo S. Alaimo

Statistical learning for well-being analysis

Assessing multidimensional poverty of the Italian provinces during Covid-19: a small area estimation approach 238
Mariateresa Ciommi, Chiara Gigliarano, Francesca Mariani and Gloria Polinesi

The fuzzy set approach as statistical learning for the analysis of multidimensional well-being 244
Gianni Betti, Federico Crescenzi, Antonella D'Agostino and Laura Neri

What Makes a Satisfying Life? Prediction and Interpretation with Machine-Learning Algorithms n.a.
Conchita D'Ambrosio

Bayesian contributions to Statistical Learning

A Bayesian framework for early cancer screening 249
Sally Paganin and Jeff Miller

Imputing Synthetic Pseudo Data from Aggregate Data: Development and Validation for Precision Medicine n.a.
Cecilia Balocchi

Linear models with assumptions-free residuals: a Bayesian Nonparametric approach 254
Filippo Ascolani and Valentina Ghidini

Data Visualization for Smart Insights and Advanced Predictive Analytics

Applications of data visualization for industry 259
Martina Dossi, Stefano Sangaletti, Marilena Di Bari and Federica Bruschini

Some Notes on the Use of the Circular Boxplot n.a.
Giovanni Camillo Porzio and Davide Buttarazzi

TERRA: a smart visualization tool for international trade in goods statistics 265
Francesco Amato, Mauro Bruno and Maria Serena Causo

Methods for the analysis of distributional data

Clustering of Distributional Data based on LDQ transformation 271
Gianmarco Borrata and Rosanna Verde

Dynamic learning from data streams through the combined use of probability density functions and simplicial functional principal component analysis 276
Francesca Fortuna, Fabrizio Maturo and Tonio Di Battista

Multivariate Parametric Analysis of Distributional Data n.a.
Paula Brito

Migrants and Refugees in Europe: social, economic and health-related issues

Labor Market Return to Refugees' Human Capital Investment: A Natural Experiment in Sweden n.a.
Eleonora Mussino

Social networks and loneliness among older migrants in Italy 282
Viviana Amati, Eralba Cela and Elisa Barbiano di Belgiojoso

The Italian Decree on Security: An Analysis of the Impact on Asylum Applications 287
Giorgio Piccitto

Modelling and Forecasting High-dimensional time series

Adaptive combinations of tail-risk forecasts 293
Alessandra Amendola, Vincenzo Candila, Antonio Naimoli and Giuseppe Storti

Are Monetary Policy Announcements related to Volatility Jumps? 299
Giampiero Gallo, Demetrio Lacava and Edoardo Otranto

Regularized Estimation and Prediction of the El Nino/Southern Oscillation Cycle	n.a.
Alessandro Giovannelli and Tommaso Proietti	
3 Contributed Sessions	305
Bayesian nonparametric methods	
Bayesian density estimation for modeling age-at-death distribution	306
Davide Agnoletto, Tommaso Rigon and Bruno Scarpa	
Bayesian mixing distribution estimation in the Gaussian-smoothed 1-Wasserstein distance	311
Catia Scricciolo	
Bayesian nonparametric estimation of heterogeneous intrinsic dimension via product partition models	316
Francesco Denti, Antonio Di Noia and Antonietta Mira	
Bayesian nonparametric multiple change point detection for time series of compositional data	322
Edoardo Marchionni and Riccardo Corradin	
Galton-Watson process: a non parametric prior for the offspring distribution	328
Massimo Cannas, Michele Guindani and Nicola Piras	
Hierarchical processes in survival analysis	333
Riccardo Cogo, Federico Camerlenghi and Tommaso Rigon	
Economics and Statistics	
A regression analysis for count data to investigate the effectiveness of incentives on the adoption of 4.0 technologies	339
Stefano Bonnini and Michela Borghesi	
Statistical analysis on SDGs indicators related to environmental sustainability	344
Najada Firza, Anisa Bakiu and Dante Mazzitelli	
Empowering futures adopting a spatial convergence of opinions: a Real-Time Spatial Delphi approach	349
Yuri Calleo, Simone Di Zio and Francesco Pilla	
Stocks price forecasts using Stochastic Differential Equations: an empirical assessment	355
Dario Frisardi and Matteo Spuri	
The Added-Worker Effect within Italian Households	361
Donata Favaro and Anna Giraldo	
Health statistics 1	
A model for the natural history of breast cancer: application to a Norwegian screening dataset	365
Laura Bondi, Marco Bonetti and Solveig Hofvind	

Generalized Bayesian Ensemble Survival Trees: an extension to categorical variables to apply it to real data Elena Ballante	370
Joint modelling of hospitalizations and survival in Heart Failure patients: a discrete non parametric frailty approach Chiara Masci, Marta Spreafico and Francesca Ieva	375
Mobility trends in Italy during the first wave of Covid-19 pandemic: analysis on Google data Ilaria Bombelli and Daniele De Rocchi	381
Tracking attitudes towards COVID vaccines: A text mining analysis Leonardo Scarso, Marco Novelli and Francesco Saverio Violante	387
Treatment effect assessment in observational studies with multi-level treatment and outcome Federica Cugnata, Paola Vicard, Paola M.V. Rancoita, Fulvia Mecatti, Clelia Di Serio and Pier Luigi Conti	393
 Indicators: composition, uses and limitations	
Are European consumers willing to pay the true price for sustainable food? Luca Secondi and Mengting Yu	399
Can the reliability of composite indexes be impacted by uncertainty of individual indicators? Caterina Giusti, Stefano Marchetti and Vincenzo Mauro	406
Initial Coin Offerings and ESG: allies or enemies? Alessandro Bitetto and Paola Cerchiello	411
On the impact of intraclass correlation in the ANVUR evaluation of academic departments Giorgio Edoardo Montanari and Marco Doretti	417
Small area estimation of monetary poverty indicators with poverty lines adjusted using local price indexes Luigi Biggeri, Stefano Marchetti, Caterina Giusti, Monica Pratesi, Francesco Schirripa Spagnolo and Gaia Bertarelli	422
Smart Composite Indicators Measuring Corporate Sustainability: A Sensitivity Analysis Camilla Salvatore, Annamaria Bianchi and Silvia Biffignandi	428
 Multivariate data analysis 1	
A note on most powerful tests for right censored survival data Maria Veronica Vinattieri and Marco Bonetti	434
Enhancing Principal Components by a Linear Predictor: an Application to Well-Being Italian Data Laura Marcis, Maria Chiara Pagliarella and Renato Salvatore	439

Proper Bayesian Bootstrap for Bagging tree model in survival analysis with correlated data	445
Farah Naz and Elena Ballante	
ROBOUT: a multi-step methodology for conditional outlier detection	450
Matteo Farnè and Angelos Vouldis	
Robustness of the Efficient Covariate-Adaptive Design for balancing covariates in comparative experiments	456
Rosamarie Frieri, Alessandro Baldi Antognini, Maroussa Zagoraiou, and Marco Novelli	
Separation scores: a new statistical tool for scoring and ranking partially ordered data	462
Marco Fattore	
Statistics in Society 1	
Community detection analysis with robin on hashtag network	468
Valeria Policastro, Francesco Santelli and Giancarlo Ragozini	
Film Tourism Motivation through the lens of Trip Advisor data	474
Nicolò Biasetton, Marta Disegna, Girish Prayag and Elena Barzizza	
Life satisfaction and social activities in later life in Italy: a focus on the Internet use	480
Claudia Furlan and Silvia Meggiolaro	
Social capital endowment's role in the intergenerational transmission of education	485
Alessandra Trimarchi, Maria Gabriella Campolo and Antonino Di Pino Incognito	
Streaming Data from Social Networks to Track Political Trends	490
Emiliano del Gobbo and Barbara Cafarelli	
The scientific production on gender dysphoria: a bibliometric analysis	495
Maria Gabriella Grassia, Marina Marino, Massimo Aria, Rocco Mazza, Luca D'Aniello and Agostino Stavo	
Assessment and Education	
A hierarchical modelling approach to explain differential functioning of mathematics items by student's gender	500
Clelia Cascella	
A latent variable approach to Millennials' knowledge of green finance	506
Maria Iannario, Alessandra Tanda and Claudia Tarantola	
Archetypal analysis and latent Markov models: A step-wise approach	512
Lucio Palazzo, Rosa Fabbriatore and Francesco Palumbo	
From high school to university: academic intentions and enrolment of foreign students in Italy	518
Francesca Di Patrizio, Eleonora Trappolini and Cristina Giudici	
Growth models for the progress test in Italian dentistry degree program	523
Giulio Biscardi, Leonardo Grilli, Carla Rampichini, Laura Antonucci and Corrado Crocetta	

The COVID-19 pandemic and academic E-learning: Italian students and instructors' perceptions	527
Francesco Santelli, Teresa Gentile, Davide Bizjak and Lorenzo Fattori	
Working Students and job market outcomes: Insights from the University of Florence	532
Gabriele Lombardi, Valentina Tocchioni and Alessandra Petrucci	
Bayesian methods and applications 1	
Analyzing RNA data with scVelo: identifiability issues and a Bayesian implementation	538
Elena Sabbioni, Enrico Bibbona, Gianluca Mastrantonio and Guido Sanguinetti	
Approximate Bayesian Computation for Probabilistic Damage Identification	544
Cecilia Viscardi, Silvia Monchetti, Luisa Collodi, Gianni Bartoli, Michele Betti, Michele Boreale and Fabio Corradi	
Estimation of scientific productivity with a hierarchical Bayesian model	550
Maura Mezzetti and Ilia Negri	
Heat waves and free-knots splines	555
Gioia Di Credico and Francesco Pauli	
The Hierarchical Beta-Bernoulli Process as Out-of-Scope Query Detector	560
Marco Dalla Pria and Silvia Montagna	
Health and mortality	
A novel definition of comorbidity based on the Global Burden of Diseases project weights	566
Angela Andreella, Lorenzo Monasta and Stefano Campostrini	
An Age-Period-Cohort model of gender gap in youth mortality	572
Giacomo Lanfiuti Baldi and Andrea Nigri	
Kinlessness in adult and old age across Europe	578
Marta Pittavino, Bruno Arpino and Elena Pirani	
Parameter orthogonalization for Siler mortality model	584
Claudia Di Caterina and Lucia Zanotto	
Pseudo-observations in survival analysis	590
Marta Cipriani, Alfonso Piciocchi, Valentina Arena and Marco Alfò	
Sex Gap in Cancer-Free Life Expectancy: The Association with Smoking, Obesity and Physical Inactivity	595
Alessandro Feraldi, Cristina Giudici and Nicolas Brouard	
Women's Exposure to HIV in Africa: the Role of Intimate Partner Violence	599
Micaela Arcaio and Anna Maria Parroco	

Mixture Models

An extension of finite mixtures of latent trait analyzers for biclustering bipartite networks	605
Dalila Failli, Maria Francesca Marino and Francesca Martella	
Constrained Mixtures of Generalized Normal Distributions	611
Pierdomenico Dutillo, Alfred Kume and Stefano Antonio Gattone	
Mixture-based clustering with covariates for ordinal responses	617
Kemawadee Preedalikit, Daniel Fernández, Ivy Liuc, Louise McMillan, Marta Nai Ruscone and Roy Costilla	
Partial membership models for soft clustering of multivariate count data	623
Emiliano Seri, Thomas Brendan Murphy and Roberto Rocci	
Regression for mixture models for extremes	629
Viviana Carcaiso, Ilaria Prodocimi and Isadora Antoniano-Villalobos	
Robust matrix-variate mixtures of regressions	635
Salvatore Daniele Tomarchio and Michael P. B. Gallagher	

Sampling methods and analysis of survey data

On the use of auxiliary information to define the sampling design for large-scale geospatial data	641
Chiara Bocci and Emilia Rocco	
Optimal joint inclusion probabilities for spatial sampling	n.a.
Giuseppe Arbia, Piero Demetrio Falorsi and Vincenzo Nardelli	
Robustness and Balance of Sampling or Experimental Designs and Mixture of Designs	647
Yves Tillé and Ejub Talovic	
Robustness Bounds for Sampling and Experimental Designs	654
Ejub Talovic and Yves Tillé	
Statistical Matching: Hotdeck or Propensity Score?	661
Elena Dalla Chiara, Marcello D'Orazio and Federico Perali	
The Italian experience on register-based statistics considering measurement, coverage and sampling errors	667
Marco Di Zio, Romina Filippini and Simona Toti	

Space-time statistics

A Hierarchical Spatio-Temporal Model for Time-Frequency Data: An application in bioacoustic analysis	673
Hiu Ching Yip, Gianluca Mastrantonio, Enrico Bibbona, Daria Valente and Marco Gamba	
An approach to cluster time series extremes with spatial constraints	679
Alessia Benevento, Fabrizio Durante and Roberta Pappadà	
An integrated space-time model to evaluate the innovation drivers in Italy	685
Emma Bruno, Rosalia Castellano and Gennaro Punzo	

Revealing the dynamic relations between traffic and crowding using big data from mobile phone network	691
Selene Perazzini, Rodolfo Metulini and Maurizio Carpita	
SMaC: Spatial Matrix Completion method	697
Giulio Grossi, Alessandra Mattei and Georgia Papadogeorgou	
The impact of traffic flow and road signs on road accidents: an approach based on spatiotemporal point pattern analysis on linear networks	702
Andrea Gilardi and Riccardo Borgoni	
Clustering and classification 1	
A clustering model for flow data: an application to international student mobility	708
Cinzia Di Nuzzo and Donatella Vicari	
Contingency tables with structural zeros and discrete copulas	713
Roberto Fontana, Elisa Perrone and Fabio Rapallo	
Levels Merging in the Latent Class Model	719
Christophe Biernacki	
Model-based clustering of count processes with multiple change	725
Shuchismita Sarkar and Xuwen Zhu	
Similarity Measures and Internal Evaluation Criteria in Hierarchical Clustering of Categorical Data	729
Jana Cibulková, Zdeněk Šulc, Hana Řezanková and Jaroslav Horníček	
Spectral clustering of mixed data via association-based distance	735
Alfonso Iodice D'Enza, Francesco Palumbo and Cristina Tortora	
Dynamic models and time series	
A graph based convolution Neural Network approach for forecast reconciliation	741
Andrea Marcocchia and Pierpaolo Brutti	
A multivariate hidden semi-Markov model for the analysis of multiple air pollutants	747
Marco Mingione, Pierfrancesco Alaimo Di Loro, Francesco Lagona and Antonello Maruotti	
A smooth transition autoregressive model for matrix-variate time series	753
Andrea Bucci	
Dynamic network models with time-varying nodes	759
Luca Gherardini, Mauro Bernardi and Monia Lupparelli	
Time lapse analysis of nuclear calcium spiking in plant cells during symbiotic signaling	765
Ivan Sciascia, Andrea Crosino and Andrea Genre	
Two-stage weighted least squares estimator of multivariate conditional mean observation-driven time series models	770
Mirko Armillotta	

Environmental learning and indicators

- Assessing the performance of nuclear norm-based matrix completion methods on CO₂ emissions data 776
Rodolfo Metulini, Francesco Biancalani, Giorgio Gnecco and Massimo Riccaboni
- Deep Learning for smart and sustainable agriculture 782
Amalia Vanacore, Armando Ciardiello, Annalisa Izzo, Pierdomenico Zaffino, Carolina Vecchio, Gennaro Pio Auricchio and Luigi Uccelli
- Do green transition, environmental taxes and renew-able energy promote ecological sustainability in G7 countries? Evidence from panel quantile regression 788
Aamir Javed, Agnese Rapposelli and Asif Javed
- Doubly Robust DID for National Parks evaluation: “just” environmental benefits, or socioeconomics impacts as well? 795
Riccardo D’Alberto, Francesco Pagliacci and Matteo Zavalloni
- On the gap between emitted and absorbed carbon dioxide. Are trees enough to save us? 801
Lorenzo Mori and Maria Rosaria Ferrante
- Small scale analysis of energy vulnerability in the municipality of Palermo 806
Giuliana La Mantia

Health statistics 2

- A test for non-differential misclassification error in database epidemiological studies 812
Giorgio Limoncella, Leonardo Grilli, Emanuela Dreassi, Carla Rampichini, Robert Platt and Rosa Gini
- Is the COVID-19 ‘color code’ of Italian regions subjected to political manipulation? 816
Giovanni Busetta and Fabio Fiorillo
- Modelling multilevel ordinal response under endogeneity: application to DTC patients’ outcome 822
Silvia D’Elia
- Monitoring drugs-based diagnostic therapeutic paths in heart failure patients using state-sequence analysis techniques 827
Nicole Fontana, Laura Savaré and Francesca Ieva
- Optimal two-stage design based on error rates under a Bayesian perspective 833
Susanna Gentile and Valeria Sambucini

Migrants in Italy and return migration

- Comparing migrant and “native” Italian adolescents in risky behaviours from FSS and SHARE Corona surveys n.a.
Daniela Foresta
- EU-Border crisis on Twitter: sentiments and misinformation analysis 839
Elena Ambrosetti, Cecilia Fortunato and Sara Miccoli

Graduates' interregional migration in times of crisis: the Italian case Thaís García-Pereiro, Ivano Dileo and Anna Paterno	843
Intentions to stay: The experience of return migrants in Albania Maria Carella, Thaís García-Pereiro, Roberta Pace and Anna Paterno	848
Return migration to home country: a systematic literature review with text mining and topic modelling Cecilia Fortunato, Andrea Iacobucci and Elena Ambrosetti	853
The allocation of time within native and foreign couples living in Italy Giovanni Busetta, Maria Gabriella Campolo and Antonino Di Pino Incognito	860
Ειλεΐθυια comes from afar: The foreigners' contribution to fertility by Italian provinces Eleonora Miaci, Cristina Giudici, Eleonora Trappolini, Marina Attili, Cinzia Castagnaro and Antonella Guarneri	866
 Sustainability assessment	
ESG, sustainability and stock market risk Michele Costa	871
Exploring the effect of consumer motivation and perception of sustainability on food choices with a Discrete Choice Experiment Gloria Solano-Hermosilla, Jesus Barreiro-Hurle and Iliaria Amerise	875
Sustainability explained by ChatGPT artificial intelligence in a HITL perspective: innovative approaches Vito Santarcangelo, Angelo Lamacchia, Emilio Massa, Saverio Gianluca Crisafulli, Massimiliano Giacalone and Vincenzo Basile	881
Measuring economic and ecological efficiency of urban waste systems in Italy: a comparison of SFA and DEA techniques Massimo Gastaldi, Ginevra Virginia Lombardi, Agnese Rapposelli and Giulia Romano	887
Profile based latent distance association analysis for sparse tables. Application to the attitude of EU citizens towards sustainable tourism Francesca Bassi, José Fernando Vera and Juan Antonio Marmolejo Martin	893
Sustainable tourism: a survey on the propensity towards eco-friendly accommodations Claudia Furlan and Giovanni Finocchiaro	899
 Bayesian methods and applications 2	
A comparison of computational approaches for posterior inference in Bayesian Poisson regression Laura D'Angelo	903
Bias-reduction methods for Poisson regression models Luca Presicce, Tommaso Rigon and Emanuele Aliverti	908
Finite Mixture Model for Multiple Sample Data Alessandro Colombi, Raffaele Argiento, Federico Camerlenghi and Lucia Paci	913

On Bayesian power analysis in reliability	918
Fulvio De Santis, Stefania Gubbiotti and Francesco Mariani	
Power priors elicitation through Bayes factors	923
Roberto Macri Demartino, Leonardo Egidi and Nicola Torelli	
Predictive Bayes factors	929
Leonardo Egidi and Ioannis Ntzoufras	
Clustering and classification 2	
A Clusterwise Regression Method for Distributional-Valued Data	935
Antonio Balzanella, Rosanna Verde and Francisco de A.T. de Carvalho	
A novel statistical-significance based semi-parametric GLMM for clustering countries standing on their innumeracy levels	939
Alessandra Ragni, Chiara Masci, Francesca Ieva and Anna Maria Paganoni	
Introducing a novel directional distribution depth function for supervised classification	945
Edoardo Redivo and Cinzia Viroli	
Clustering alternatives in the preference-approval context	950
Alessandro Albano, José Luis Garcia-Lapresta , Mariangela Sciandra and Antonella Plaia	
Computational assessment of k-means clustering on a Structural Equation Model based index	955
Mariaelena Bottazzi Schenone, Elena Grimaccia and Maurizio Vichi	
Handling missing data in complex phenomena: an ultrametric model-based approach for clustering	961
Francesca Greselin and Giorgia Zaccaria	
Economics and labour markets	
A multivariate ranking analysis on the employability of young adults	967
Rosa Arboretti, Elena Barzizza, Nicolo Biasetton, Riccardo Ceccato, Monica Fedeli and Concetta Tino	
Analysis of the Gender Pay Gap in the Italian Labour Market	973
Giulia Cappelletti and Daniele Toninelli	
Evaluating the effect of home-based working employing causal Bayesian networks and potential outcomes	979
Lorenzo Giammei	
Patterns of flexible employment careers. Does measurement error matter?	985
Mauricio Garnier-Villarreal, Dimitris Pavlopoulos and Roberta Varriale	
Staying or leaving? A nonlinear framework to explore the role of employee well-being on retention	991
Ulpiani Kocollari, Fabio Demaria and Maddalena Cavicchioli	
The CAP instruments impact on GVA and employment: a multivalued treatment approach	997
Montezuma Dumangane and Marzia Freo	

The determinants of leaving the parental home in Italy: 2012-18 Ilaria Rocco and Gianpiero Dalla Zuanna	1003
Environmental modeling	
A Bayesian weather-driven spatio-temporal model for PM10 in Lombardy Michela Frigeri, Alessandra Guglielmi and Giovanni Lonati	1109
A preliminary study on shape descriptors for the characterization of microplastics ingested by fish Greta Panunzi, Tommaso Valente, Marco Matiddi and Giovanna Jona Lasinio	1015
Artificial neural network in predicting odour concentrations: a case study Veronica Distefano and Gideon Mazuruse	1021
Bayesian analysis of PM10 concentration by spatio-temporal ARIMA and STS models Michela Frigeri and Ilenia Epifani	1026
Functional ANOVA to monitor yearly Adriatic sea temperature variations Annalina Sarra, Adelia Evangelista, Tonio Di Battista and Nicola Di Deo	1032
New perspectives in the measurement of biodiversity Linda Altieri, Daniela Cocchi and Massimo Ventrucci	1038
Multivariate data analysis 2	
Feature Selection via anomaly detection autoencoders in radiogenomics studies Alessia Mapelli, Michela Carlotta Massi, Nicola Rares Franco, Francesca Ieva, Catharine West, Petra Seibold, Jenny Chang-Claude and the REQUITE and RADprecise Consortia	1044
Further considerations on the Spectral Information Criterion Luca Martino	1050
How to increase the power of the test in sparse contingency tables: a simulation study Federica Nicolussi and Manuela Cazzaro	1057
Latent event history models for quasi-reaction systems Matteo Framba, Veronica Vinciotti and Ernst Wit	1063
Quantile-based graphical models for continuous and discrete variables Luca Merlo, Marco Geraci and Lea Petrella	1069
The logratio Student t distribution Gianna Monti and Gloria Mateu-Figueras	1075
Statistics in Society 2	
A decomposition of the changes in tourism demand in Tuscany over the 2019-2021 period Mauro Mussini	1079
Bayesian networks as a territorial gender impact assessment tool Flaminia Musella, Lorenzo Giammei, Fulvia Mecatti and Paola Vicar	1084

Can statistics be helpful in detecting electoral fraud? Massimo Attanasio, Vincenzo G. Genova and Michele Tumminello	1088
Companies' sustainability disclosure and contrast to hunger: the role of social inclusion Chiara Di Maria and Rodolfo Damiano	1093
Passing network-based performance indicator in football: evidence from UEFA Champions League 2016-2017 Riccardo Ievoli, Lucio Palazzo and Giancarlo Ragozini	1099
Topic Modeling for the travel and tourism industry: classical and innovative methods compared Fabrizio Di Mari	1105
Bayesian methods and applications 3	
An Importance Sampling Algorithm For Bayesian Logistic Regression with Independent Gaussian Scale Mixture Prior Paolo Onorati and Brunero Liseo	1111
Bayesian analysis of Amazon's best-selling books via finite nested mixture model Laura D'Angelo and Francesco Denti	1117
Binomial Extended Stochastic Block Model for Brain Networks Valentina Ghidini, Sirio Legramanti and Raffaele Argiento	1121
Detecting latent spatial patterns in mass spectrometry brain imaging data via Bayesian mixtures Giulia Capitoli, Simone Colombara, Alessia Cotroneo, Francesco De Caro, Riccardo Morandi, Chiara Schembri, Alfredo G. Zapiola and Francesco Denti	1127
Efficient expectation propagation for high-dimensional probit models Augusto Fasano, Niccolo Anceschi, Beatrice Franzolini and Giovanni Rebaudo	1133
Model-based clustering of non-stationary time series with common historical change times Riccardo Corradin, Luca Danese, Wasiur KhudaBukhsh and Andrea Ongaro	1139
Functional Data Analysis	
A functional Ground Motion Model for Italy built with a weighted analysis of reconstructed seismic curves Teresa Bortolotti, Riccardo Peli, Giovanni Lanzano, Sara Sgobba and Alessandra Menafoglio	1145
Conditional Gaussian Graphical Models for Functional Variables with Partial Separable Operators Rita Fici, Gianluca Sottile and Luigi Augugliaro	1149
Does the Inflation Factor need tuning? Simulation-based adjustment for Outlier Detection via the Functional Boxplot Annachiara Rossi, Andrea Cappozzo and Francesca Ieva	1155
Functional Graphical Models to map Brexit debate on Twitter Nicola Pronello, Emiliano del Gobbo, Lara Fontanella, Rosaria Ignaccolo, Luigi Ippoliti and Sara Fontanella	1160

Measuring Dependence in Multivariate Functional Datasets Francesca Ieva, Michael Ronzulli and Anna Maria Paganoni	1166
Robust Statistical Process Monitoring of Multivariate Functional Data Christian Capezza, Fabio Centofanti, Antonio Lepore and Biagio Palumbo	1173
The effects of mobility restrictions on public health: a functional data analysis for Italy over the years 2020 and 2021 Veronica Mazzola, Giovanni Bonaccorsi, Piercesare Secchi and Francesca Ieva	1179
Machine Learning and text mining	
A vocabulary-based approach for risk detection in textual annotations of contracts of public procurement Giulio Giacomo Cantone, Simone Del Sarto and Michela Gnaldi	1185
Explainable Machine Learning based on Group Equivariant Non-Expansive Operators (GENEOs). Protein pocket detection: a case study Giovanni Bocchi, Alessandra Micheletti, Patrizio Frosini, Alessandro Pedretti, Andrea R. Beccari, Filippo Lunghini, Carmine Talarico and Carmen Gratteri	1191
Hedging global currency risk with factorial machine learning models Paolo Pagnottoni and Alessandro Spelta	1197
InstanceSHAP: An instance-based estimation approach for Shapley values Golnoosh Babaei and Paolo Giudici	1203
Networks & Nature Based Solutions: an application for Milan hydric resources Alessia Forciniti and Emma Zavarrone	1209
The Roe v. Wade sentence: an analysis of tweets trough Symmetric Non-Negative Matrix Factorization Maria Gabriella Grassia, Marina Marino, Rocco Mazza and Agostino Stavolo	1215
Multivariate data analysis 3	
A comparison of different techniques for handling missing covariate values in propensity score methods Anna Zanovello, Alessandra R. Brazzale and Omar Paccagnella	1219
A New Penalized Estimator for Sparse Inference in Gaussian Graphical Models: An Adaptive Non-Convex Approach Daniele Cuntrera, Vito M.R. Muggeo and Luigi Augugliaro	1224
A tool for assessing weak identifiability of statistical models Antonio Di Noia, Francesco Denti and Antonietta Mira	1230
Computing Highest Density Regions with Copulae Nina Deliu and Brunero Liseo	1235
Parameter estimation via Indirect Inference for multivariate Wrapped Normal distributions Francesca Labanca and Anna Gottard	1241

Sequential marginal likelihood selection for the estimation of sparse correlation matrices	1246
Claudia Di Caterina and Davide Ferrari	
Nonparametric statistical methods	
A Comparison of Distribution-Free Control Charts	1252
Michele Scagliarini	
Characterizing Heterogeneity of Causal Effects in Air Pollution in Florida	1257
Dafne Zorzetto	
Comparing three robust procedures for CANDECOMP/PARAFAC estimation	1262
Valentin Todorov, Violetta Simonacci, Michele Gallo and Nikolay Trendafilov	
How active is a genetic pathway? Comparative analysis of post-hoc permutation-based methods	1268
Anna Vesely and Angela Andreella	
Non Parametric Combination methodology: a literature review on recent developments	1274
Elena Barzizza, Nicolò Biasetton and Riccardo Ceccato	
Regression modeling	
A Quantile Regression Model to Evaluate the Performance of the Italian Courts of Law	1280
Carlo Cusatelli, Massimiliano Giacalone and Eugenia Nissi	
A variable selection procedure based on predictive ability: a preliminary study on logistic regression	1285
Rosaria Simone and Mariarosaria Coppola	
Comparison of binary regressions with asymmetric link function for imbalanced data	1291
Michele La Rocca, Marcella Niglio and Marialuisa Restaino	
New advances in Regression Forests	1297
Mila Andreani, Lea Petrella and Nicola Salvati	
On the Optimal Non-Convexity of Penalty in Sparse Regression Models	1303
Daniele Cuntreza, Vito M.R. Muggeo and Luigi Augugliaro	
Using expectile regression with latent variables for digital assets	1309
Beatrice Foroni, Luca Merlo and Lea Petrella	
4 Program	1315

How can we explain Random Forests in a spatial framework?

Natalia Golini^a, Luca Patelli^b, and Xavier Barber^c

^aUniversity of Torino, Department of Economics and Statistics Cognetti de Martiis, Lungo Dora Siena, 100A, Torino; natalia.golini@unito.it

^bUniversity of Pavia, Department of Economics and Management, Via San Felice al Monastero, 5, Pavia; luca.patelli01@universitadipavia.it

^cUniversidad Miguel Hernández de Elche, Centro de Investigación Operativa, Avenida de la Universidad, Elche; xbarber@umh.es

Abstract

Random Forest (RF) is a Machine Learning algorithm, very popular in environmental applications thanks to its flexibility and predictive performances. Even if its working mechanism is simple and intelligible, RF is considered a *black box* model since it prevents grasping how predictors are combined to generate the response variable predictions. This lack of interpretability represents a limitation of RF, especially when some knowledge is required on the response-predictors relationship from the decision-making perspective. In this work, we aim to explain RF using a Post-Hoc approach, i.e. by extracting a compact and simple list of rules from an estimated RF focusing on a spatial regression context. By means of a spatial dataset, we compare the final sets of rules and discuss the predictive accuracies of the standard RF and its *gold standard* for the case of spatially correlated data.

Keywords: Explainable Machine Learning, inTrees, Post-hoc methods, Rule extraction, RF-GLS

1. Introduction

The Machine Learning (ML) era has given rise to complex and powerful methods that can process vast amounts of data and make predictions with remarkable accuracy. However, the inherent *black-box* nature of some of these techniques has raised concerns about their lack of interpretability. Often the term *interpretability* is used as a synonym for *explainability*, but actually they refer to two different concepts. According to Rudin *et al.* (11), interpretability is referred to models that are built to be interpretable, while explainability is obtained by applying further techniques to non-interpretable models in order to extract information. On the topic of explainable ML methods, the recent paper by Wikle *et al.* (14) is worth to be mentioned. In particular, the authors discuss the use of explainability techniques in spatial ML to understand the role of specific inputs in predicting environmental variables. Even if from a statistical point of view the gold standard would be to use interpretable ML methods, when this is not possible it is a good practice to try to extract information from non-interpretable ML methods that have proven good performance.

In this work, among ML techniques, we consider Random Forest which is well known for its high prediction accuracy. It is a non-parametric supervised algorithm that, thanks to its flexibility, can model complex non-linear relationships between the response variable (categorical or continuous) and the predictors (3). RF is defined as an ensemble model as the result of aggregating a set of decision trees. Each tree is the result of a recursive binary splitting process obtained using re-sampled data and a random

set of predictors evaluated at each node as splitting candidates. Given its adaptability, RF has also been widely applied in the spatial framework with different strategies to deal with the spatial autocorrelation of the data. Patelli *et al.* (10) have recently proposed a literature review and a novel taxonomy of the existing strategies adopted to adjust RF for spatially correlated data. In particular, the authors highlight that the most interesting strategy is the RF-GLS method proposed by Saha *et al.* (12), who extend the RF by estimating trees using generalized least squares (GLS). It was proven that RF-GLS outperforms the classical RF in the presence of spatial correlation, thus representing the gold standard to be used in the spatial framework.

In any case, spatially aware or not, RF remains a non-interpretable algorithm. However, it is possible to use specific methods to explain the RF resulting model, as described in the review by Haddouchi and Berrado (7). In particular, “Internal Processing” (IP) methods try to get “insights that are inherent to internal processing” providing a global overview of the model. “Post-Hoc” (PH) methods instead are based on RF post-processing, such as for example the “Rule Extraction” (RE) approaches (see e.g. inTrees (5), SIRUS (2), Node harvest (9) and RuleFit (6) among others). These methods aim to find a limited set of rules (each defined as the combination of predictors and split values) that is common to many trees in the RF and that allow representing the prediction mechanism of RF.

The main aim of this contribution is to verify if, for a spatial regression problem, there exist differences in the rules obtained by using - so far - the inTrees approach applied to two different cases: trees grown by RF-GLS and by a classical RF. We expect that taking or not into account the spatial correlation when implementing RF will have an impact also in its extracted rules. The analysis is carried out by using a dataset regarding daily meteorological records measured by 159 monitoring stations in Croatia. We present here preliminary results followed by a discussion on further steps.

2. Data and methods

The explainability of RF in the spatial framework is illustrated using meteorological daily data from the national network of 159 stations in Croatia for the year 2008, provided by the Croatian National Meteorological Service (available at <https://github.com/AleksandarSekulic/RFSI>). At this stage of the work, we have not considered the temporal dimension of the data confining the analysis to a single day: 14th June 2008. The locations of the 151 stations working at this date are shown in Fig. 1. In particular, dots and crosses represent training and test data considered to implement the RF-GLS and RF algorithms. For this dataset, we randomly selected 90% of the data (i.e., 135 observations) for training the algorithms and used the remaining 10% of the data (i.e., 16 observations) for testing the algorithms. Croatia is a country located in southeastern Europe, bordering the Adriatic Sea. It has a diverse topography with flat plains in the east, a hilly central region, and mountainous terrain in the west. The response variable is the mean daily temperature¹ [TEMP], measured in degrees Celsius (°C). The minimum and maximum observed mean daily temperature values are 1.8°C and 21.5°C, respectively. The highest temperatures are recorded along the coast and at low altitudes. The variables used as predictors are latitude [lat (in meters)], longitude [lon (in meters)], distance-to-coastline [HRdsea (in km)], elevation [HRDdem (in meters)], wetness index [HRtwi], seasonal fluctuation [ctd (in days)], insolation (total incoming solar radiation) [INSOL (in Joules)], and Moderate Resolution Imaging Spectroradiometer land surface temperature [MODIS.LST] images. The dataset and predictors are detailed in (8) and references therein. In particular, this dataset was used by Sekulić *et al.* (13) to evaluate and compare the performance of a spatial interpolation method they proposed, i.e. the Random Forest Spatial Interpolation.

With the aim of obtaining simple, stable and accurate rules, we implemented the inTrees approach proposed by Deng (5) and implemented in the homonym R package inTrees². The set of algorithms proposed in the work of Deng (5) can be applied to all tree ensemble methods to perform different tasks: extract, prune, select and summarize the rules. Each step is not mandatory, and the procedure can be tailored based on the specific explanatory necessity.

¹On most meteorological stations TEMP is measured three times a day: at 7 am, 1 pm and 9 pm.

²<https://cran.r-project.org/web/packages/inTrees/index.html>

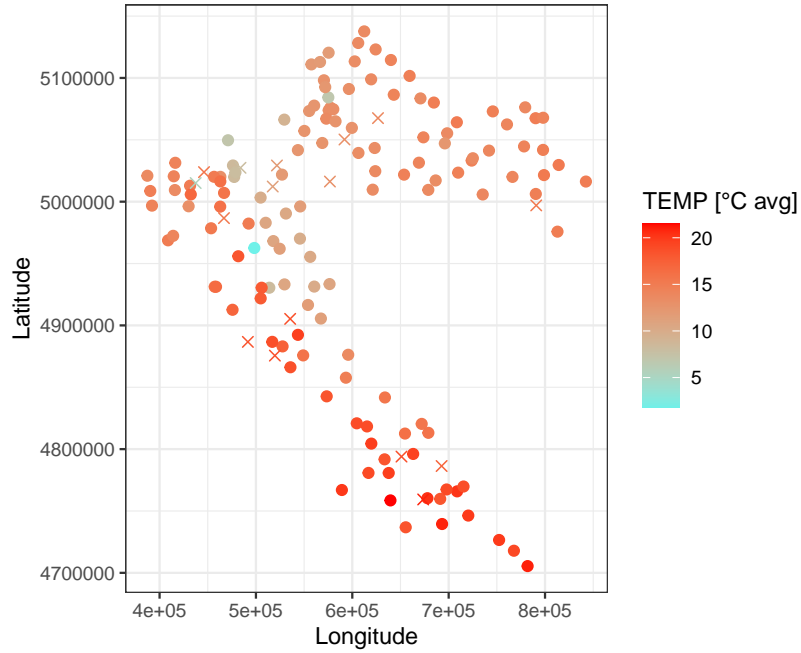


Figure 1: Mean daily temperature recorded on 2008-06-14 in 151 Croatian meteorological stations. Dots represent the mean daily temperature registered in the 135 training sites; crosses represent the mean daily temperature measured in the 16 test sites.

In order to extract and analyze rules by means of `inTrees`, the first step consists in running the chosen RF algorithm to have a collection of trees grown over a set of training data. Each tree results in the combination of all its splits, i.e. the conditions that permit splitting of the predictor space and getting predictions in the final regions. Then the obtained rules can be evaluated by using the relative “frequency” of occurrence, the prediction “error” and their “length” representing the rule complexity.

Using these metrics and considering opportune (decay) functions, the rules can be further simplified by pruning irrelevant predictor-split values. In order to have a compact rule set containing relevant and non-redundant rules, a complexity-guided condition selection method can be used, e.g. guided regularized Random Forest (GRRF) (4). In the end, the extracted rules can also be summarized by a rule-based learner that should be comparable in terms of prediction accuracy to the standard RF but more interpretable, named Simplified Tree Ensemble Learner (STEL). Note that in `inTrees` it is possible to build a STEL only for classification problems.

3. Preliminary empirical results

This section shows our preliminary results by applying the `inTrees` approach to extract insights from the RF-GLS and RF algorithms applied to the temperature spatial dataset.

We started by training the regression RF-GLS and RF on the same training set, by means of the R packages `randomForestGLS`³ and `randomForest`⁴, respectively. We used the same setting for the hyperparameters. In particular, we have set to 1000 the number of trees (`ntree` in R) and to 3 (one-third of the total number of predictors) the number of the variables randomly sampled as candidates at each split (`mtry` in R). For the RF-GLS, the covariance function used in modelling the spatial dependence structure among the observations was the default value, i.e. the exponential covariance function (`cov.mat` in R). Note that the coordinates [`lat`, `long`], measured in meters, have also been considered as predictors in both algorithms. In order to stabilize the forest structure, we followed the strategy pro-

³<https://cran.r-project.org/web/packages/RandomForestGLS/index.html>

⁴<https://cran.r-project.org/web/packages/randomForest/index.html>

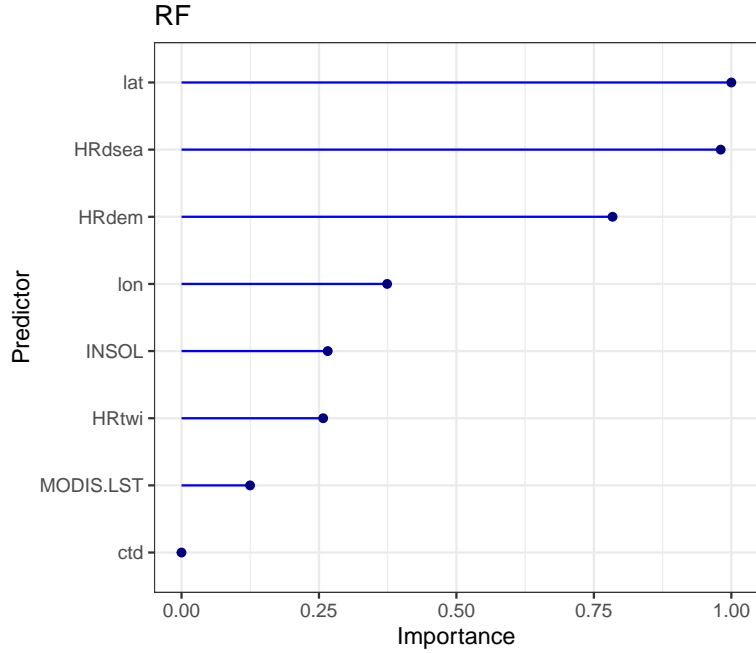


Figure 2: Variable importance plot for RF. The importance index is scaled to a maximum of 1.

posed in B nard *et al.* (2) for rule generation consisting in restricting the node splits to the q -empirical quantiles of the predictors. This modification to Breiman’s original regression tree algorithm is expected to have a small impact on predictive accuracy but is essential for stability.

Table 1 shows the test accuracy in terms of root mean square error and percentage of explained variance of the two algorithms when the node splits are restricted to the 10-empirical quantiles of the predictors. Different values of q will be considered in the next steps of the work. As expected, RF-GLS shows a better predictive performance than RF because it is able to capture the spatial autocorrelation of the response variable.

Table 1: Root Mean Square Error (RMSE) and percentage of explained variance (Var explained) values evaluated for the test dataset.

Algorithm	RMSE [�C]	Var explained [%]
RF-GLS	1.057	93.52
RF	1.357	89.32

Latitude, distance-to-coastline and DEM are the most important predictors for RF (see Fig. 2). This information is not reported for RF-GLS since the R package `randomForestGLS` does not provide the variable importance as output yet.

Given the two forests, we applied the `inTrees` approach described in Section 2. For both algorithms, RF-GLS and RF, we used the same setting for the tuning parameters of the `inTrees` functions. We extracted the rule conditions from the set of trees with a maximum length of 3 (`maxdepth` in R) from each tree. The distinct rule conditions extracted from the 1000 trees of RF-GLS and RF were 2,836 and 3,007, respectively. Then, we assigned the outcome values (mean of the response variable values of the training observations that satisfy the condition) [`pred`] to the conditions and measured the quality of the rules by “frequency” [`freq`], “error” [`err`], and “length” [`len`]. We pruned the extracted rules’ irrelevant or redundant variable-value pairs considering the metric “error” and the “relative” decay function. With the irrelevant variable-value pairs being removed, the pruned rules have shorter conditions and a frequency that increases without an increase in error. Finally, we applied the complexity-guided regularized random forest (GRRF) to the set of distinct pruned rules in order to have two compact lists of stable rules (≤ 30)

able to explain the results of both algorithms. We grew 1000 trees, setting the importance threshold to 0.1 and using the default values for the other tuning parameters of the function `selectRuleRF` in R. From a run of this function we obtained a list of 19 and 25 rules starting from the forests grown by the RF-GLS and RF algorithms, respectively. By applying both these lists of rules to test data we obtained a very good predictive performance: the percentage of variance explained was 92.01 and 90.17, respectively.

Table 2 and Table 3 show the two lists of the first ten rules output for the meteorological dataset. The scores [impRF] of the selected conditions are calculated by building an RF on the selected rules. In general, the two lists of selected rules have 17 rules in common. An example is represented by the first rule in Table 2 and Table 3. More specifically, the first rule in both lists shows that the interaction of a low latitude with a low elevation and a low distance to the coastline induces a higher mean daily temperature. The third rule in Table 2 (and then the fifth rule in Table 3) displays that the interaction of low longitude and a high elevation induces a mean daily temperature of about 9°C. This is composed of two conditions ($\text{lon} \leq 589199.5$ & $\text{HRdem} > 317.40$), and satisfied by the 14.8% of the observations in the training dataset and has an RMSE (the square root of “err”) of about 2.2°C. One can notice that rule scores (importance values) and the rules metrics are not related. For instance, the fourth rule in Table 2 (and then the second rule in Table 3) has a larger frequency than the three most important ones.

Table 2: First ten rules extracted, measured, pruned and selected via GRRF, generated by RF-GLS. The rules are ordered by scores (importance value - ImpRF)

rule	len	freq	err	condition	pred	impRF
1	3	0.252	3.082	$\text{lat} \leq 4931735.37$ & $\text{HRdem} \leq 609.20$ & $\text{HRdsea} \leq 26.14$	18.534	1
2	3	0.289	3.998	$\text{lon} > 457787.2$ & $\text{HRdem} \leq 609.20$ & $\text{HRdsea} \leq 26.14$	18.160	0.893
3	2	0.148	4.882	$\text{lon} \leq 589199.5$ & $\text{HRdem} > 317.40$	9.325	0.673
4	2	0.681	5.564	$\text{lat} > 4780743.3$ & $\text{HRdsea} > 1.34$	12.885	0.602
5	2	0.230	2.472	$\text{lon} > 457787.2$ & $\text{HRdsea} \leq 1.34$	18.714	0.586
6	3	0.148	4.882	$\text{lon} \leq 620344.9$ & $\text{lat} > 4873835.0$ & $\text{HRdem} > 317.40$	9.325	0.577
7	3	0.230	2.690	$\text{lat} \leq 4931735.37$ & $\text{HRdem} \leq 317.40$ & $\text{HRdsea} \leq 26.14$	18.743	0.505
8	3	0.148	4.882	$\text{lat} > 4873835.0$ & $\text{HRdem} > 317.40$ & $\text{HRdsea} \leq 195.44$	9.325	0.489
9	2	0.259	5.886	$\text{lat} \leq 4931735.37$ & $\text{HRdsea} \leq 26.14$	18.242	0.365
10	2	0.237	2.905	$\text{lat} \leq 4931735.37$ & $\text{HRdem} \leq 317.40$	18.645	0.347

Table 3: First ten rules extracted, measured, pruned and selected via GRRF, generated by RF. The rules are ordered by scores (importance value - ImpRF)

n	len	freq	err	condition	pred	impRF
1	3	0.252	3.082	$\text{lat} \leq 4931735.37$ & $\text{HRdem} \leq 609.20$ & $\text{HRdsea} \leq 26.14$	18.534	1
2	2	0.681	5.564	$\text{lat} > 4780743$ & $\text{HRdsea} > 1.34$	12.885	0.560
3	3	0.148	4.882	$\text{lon} \leq 620344.9$ & $\text{lat} > 4873835$ & $\text{HRdem} > 317.40$	9.325	0.487
4	3	0.148	4.882	$\text{lat} > 4873835$ & $\text{HRdem} > 317.4$ & $\text{HRdsea} \leq 195.44$	9.325	0.486
5	2	0.148	4.882	$\text{lon} \leq 589199.5$ & $\text{HRdem} > 317.40$	9.325	0.475
6	3	0.267	5.879	$\text{lon} > 457787.2$ & $\text{lat} \leq 4982676$ & $\text{HRdsea} \leq 26.14$	18.212	0.439
7	2	0.230	2.472	$\text{lon} > 457787.2$ & $\text{HRdsea} \leq 1.34$	18.714	0.433
8	3	0.230	2.690	$\text{lat} \leq 4931735$ & $\text{HRdem} \leq 317.4$ & $\text{HRdsea} \leq 26.14$	18.743	0.366
9	3	0.207	1.914	$\text{lon} > 457787.2$ & $\text{HRdsea} \leq 1.34$ & $\text{INSOL} > 8.082524$	18.994	0.359
10	3	0.23	2.963	$\text{lon} > 503554.1$ & $\text{HRdem} \leq 609.2$ & $\text{HRdsea} \leq 26.14$	18.723	0.294

4. Discussion and next steps

This work represents a first attempt to “open” an RF that is specifically designed for spatially dependent data, i.e. RF-GLS. This algorithm should be the gold standard in a spatial framework. We compared

the predictive performance and explainability of RF-GLS and RF applied to a Croatian meteorological dataset. Both algorithms have shown high and similar predictive performance in our application. A cross-validation procedure will be implemented to confirm this result. Among the different approaches existing in the literature to obtain explainability from RF, we focused on the rule extraction methods. In particular, we considered the approach proposed by Deng (5) applying the same constraints to the node splits proposed in Bénard *et al.* (2). We found two compact lists of rules with high predictive performance sharing a large number of rules in common. However, the shared rules have different scores (importance values) within their respective membership lists. As next step, we aim to tune the GRRF hyperparameters to reduce the number of rules in the two lists while maintaining their predictive performance. Moreover, we aim to set up a comparison study considering the main competitors of inTrees, i.e. SIRUS (2), Node harvest (9) and RuleFit (6). Unfortunately, the R functions implementing RF-GLS (`RFGLS_estimate_spatial` and `RFGLS_predict_spatial`) return objects that are not valid inputs for the R functions implementing the competitor rule extraction methods. This will require further investigation.

References

- [1] Aria, M., Cuccurullo, C., Gnasso, A.: A comparison among interpretative proposals for Random Forests. *MLWA* **6**, 100094 (2021)
- [2] Bénard, C., Biau, G., Da Veiga, S., Scornet, E.: Interpretable random forests via rule extraction. In: *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pp. 937–945 (2021)
- [3] Breiman, L.: Random forests. *Machine Learning* **45**, 5–32 (2001)
- [4] Deng, H., Runger, G.: Gene selection with guided regularized random forest. *Pattern Recognit.* **46**, 3483–3489 (2013)
- [5] Deng, H.: Interpreting tree ensembles with intrees. *Int J Data Sci Anal.* **7**, 277–287 (2019)
- [6] Friedman, J. H., Popescu, B. E.: Predictive learning via rule ensembles. *Ann Appl Stat.* **2**, 916–954 (2008)
- [7] Haddouchi, M., Berrado, A.: A survey of methods and tools used for interpreting random forest. In: *Proceedings of the 2019 1st International Conference On Smart Systems And Data Science (2019)* doi:10.1109/ICSSD47982.2019.9002770
- [8] Hengl, T., Heuvelink, G.B.M., Perčec Tadić, M., Pebesma, E.J.: Spatio-temporal prediction of daily temperatures using time-series of MODIS LST images. *Theor Appl Climatol.* **107**, 265–277 (2012)
- [9] Meinshausen, N.: Node harvest. *Ann Appl Stat.* **4**, 2049–2072 (2010)
- [10] Patelli, L., Cameletti, C., Golini, N., Ignaccolo, R.: A path in regression Random Forest looking for spatial dependence: a taxonomy and a systematic review. *arXiv* **2303.04693** (2023)
- [11] Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., Zhong, C.: Interpretable machine learning: Fundamental principles and 10 grand challenges. *Stat Surv.* **16**, 1–85 (2022)
- [12] Saha, A., Basu, S., Datta, A.: Random forests for spatially dependent data. *JASA* **118**, 665–683 (2023)
- [13] Sekulić, A., Kilibarda, M., Heuvelink, G. B.M., Nikolić, M., Bajat, B.: Random forest spatial interpolation. *Remote Sens.* **12**, 1687 (2020)
- [14] Wikle, C., Datta, A., Hari, B., Boone, E., Sahoo, I., Kavila, I., Castruccio, S., Simmons, S., Burr, W., Chang, W.: An illustration of model agnostic explainability methods applied to environmental data. *Environmetrics* **34**, e2772 (2023)