

Book of Short Papers

SIS 2021



Editors: Cira Perna, Nicola Salvati and Francesco Schirripa Spagnolo



Distribuzione Software | Formazione Professionale
Statistica | Economia | Finanza | Biostatistica | Epidemiologia
Sanità Pubblica | Scienze Sociali
www.tstat.it | www.tstattraining.eu

Copyright © 2021

PUBLISHED BY PEARSON

WWW.PEARSON.COM

ISBN 9788891927361

Contents

Preface	XIX
1 Plenary Sessions	1
1.1 Citizen data, and citizen science: a challenge for official statistics. <i>Monica Pratesi</i>	2
2 Specialized Sessions	8
2.1 A glimpse of new data and methods for analysing a rapidly changing population	9
2.1.1 The diffusion of new family patterns in Italy: An update. <i>Arnstein Aassve, Letizia Mencarini, Elena Pirani and Daniele Vignoli</i>	10
2.1.2 Causes of death patterns and life expectancy: looking for warning signals. <i>Stefano Mazzucco, Emanuele Aliverti, Daniele Durante and Stefano Campostrini</i>	16
2.2 Advances in ecological modelling	22
2.2.1 A Bayesian joint model for exploring zero-inflated bivariate marine litter data. <i>Sara Martino, Crescenza Calculli and Porzia Maiorano</i>	23
2.3 Advances in environmental statistics	29
2.3.1 Bayesian small area models for investigating spatial heterogeneity and factors affecting the amount of solid waste in Italy. <i>Crescenza Calculli and Serena Arima</i>	30
2.3.2 A spatial regression model for predicting abundance of lichen functional groups. <i>Pasquale Valentini, Francesca Fortuna, Tonio Di Battista and Paolo Giordani</i>	36
2.4 Advances in preference and ordinal data theoretical improvements and applications	42
2.4.1 Boosting for ranking data: an extension to item weighting. <i>Alessandro Albano, Mariangela Sciandra and Antonella Plaia</i>	43
2.4.2 An Extended Bradley-Terry Model For The Analysis Of Financial Data. <i>Alessio Baldassarre, Elise Dusseldorp and Mark De Rooij</i>	49

2.5	Business system innovation, competitiveness, productivity and internationalization	55
2.5.1	An analysis of the dynamics of the competitiveness for some European Countries. <i>Andrea Marletta, Mauro Mussini and Mariangela Zenga</i>	56
2.5.2	National innovation system and economic performance in EU. An analysis using composite indicators. <i>Alessandro Zeli</i>	62
2.6	Challenges for observational studies in modern biomedicine	68
2.6.1	Data integration: a Statistical view. <i>Pier Luigi Conti</i>	69
2.6.2	Exploring patients' profile from COVID-19 case series data: beyond standard statistical approaches. <i>Chiara Brombin, Federica Cugnata, Pietro E. Cippà, Alessandro Ceschi, Paolo Ferrari and Clelia di Serio</i>	75
2.6.3	On the statistics for some pivotal anti-COVID-19 vaccine trials. <i>Mauro Gasparini</i>	81
2.7	Data Science for Industry 4.0 (ENBIS)	87
2.7.1	Sample selection from a given dataset to validate machine learning models. <i>Bertrand Iooss</i>	88
2.7.2	Reliable data-drive modelling and optimisation of a batch reactor using bootstrap aggregated deep belief networks. <i>Changhao Zhu and Jie Zhang</i>	94
2.8	Integration of survey with alternative sources of data	100
2.8.1	A parametric empirical likelihood approach to data matching under nonignorable sampling and nonresponse. <i>Daniela Marella and Danny Pfeffermann</i>	101
2.8.2	Survey data integration for regression analysis using model calibration.	107
2.8.3	Latent Mixed Markov Models for the Production of Population Census Data on Employment. <i>Daniela Filippini, Ugo Guarnera and Roberta Varriale</i>	112
2.9	Media, social media and demographic behaviours	118
2.9.1	Monitoring the Numbers of European Migrants in the United Kingdom using Facebook Data. <i>Francesco Rampazzo, Jakub Bijak, Agnese Vitali, Ingmar Weber and Emilio Zagheni</i>	119
2.10	New developments in ensemble methods for classification	125
2.10.1	An alternative approach for nowcasting economic activity during COVID-19 times. <i>Alessandro Spelta and Paolo Pagnottoni</i>	126
2.10.2	Assessing the number of groups in consensus clustering by pivotal methods. <i>Roberta Pappadà, Francesco Pauli and Nicola Torelli</i>	132
2.10.3	Clustering of data recorded by Distributed Acoustic Sensors to identify vehicle passage and typology. <i>Antonio Balzanella and Stefania Nacchia</i>	138
2.11	New developments in latent variable models	144
2.11.1	A Hidden Markov Model for Variable Selection with Missing Values. <i>Fulvia Pennoni, Francesco Bartolucci, and Silvia Pandolfi</i>	145
2.11.2	Comparison between Different Likelihood Based Estimation Methods in Latent Variable Models for Categorical Data. <i>Silvia Bianconcini and Silvia Cagnone</i>	151
2.11.3	A Comparison of Estimation Methods for the Rasch Model. <i>Alexander Robitzsch</i>	157

2.12	New issues on multivariate and univariate quantile regression	163
2.12.1	Directional M-quantile regression for multivariate dependent outcomes. <i>Luca Merlo, Lea Petrella and Nikos Tzavidis</i>	164
2.13	Semi-parametric and non-parametric latent class analysis	170
2.13.1	Stepwise Estimation of Multilevel Latent Class Models. <i>Zsuzsa Bakk, Roberto di Mari, Jennifer Oser and Jouni Kuha</i>	171
2.13.2	Distance learning, stress and career-related anxiety during the Covid-19 pandemic: a students perspective analysis. <i>Alfonso Iodice D'Enza, Maria Iannario, Rosaria Romano</i>	177
2.13.3	A Tempered Expectation-Maximization Algorithm for Latent Class Model Estimation. <i>Luca Brusa, Francesco Bartolucci and Fulvia Pennoni</i>	183
2.14	Statistics for finance high frequency data, large dimension and networks	189
2.14.1	The Italian debt not-so-flash crash. <i>Maria Flora and Roberto Reno'</i>	190
3	Solicited Sessions	197
3.1	Advances in social indicators research and latent variables modelling in social sciences	198
3.1.1	A composite indicator to measure frailty using administrative healthcare data. <i>Margherita Silan, Rachele Brocco and Giovanna Boccuzzo</i>	199
3.1.2	Clusters of contracting authorities over time: an analysis of their behaviour based on procurement red flags. <i>Simone Del Sarto, Paolo Coppola and Matteo Troia</i>	205
3.1.3	An Application of Temporal Poset on Human Development Index Data. <i>Leonardo Salvatore Alaimo, Filomena Maggino and Emiliano Seri</i>	211
3.1.4	The SDGs System: a longitudinal analysis through PLS-PM. <i>Rosanna Cataldo, Maria Gabriella Grassia and Laura Antonucci</i>	217
3.2	Changes in the life course and social inequality	223
3.2.1	Heterogeneous Income Dynamics: Unemployment Consequences in Germany and the US. <i>Raffaele Grotti</i>	224
3.2.2	In-work poverty in Germany and in the US: The role of parity progression. <i>Emanuela Struffolino and Zachary Van Winkle Z.</i>	230
3.2.3	Parenthood, education and social stratification. An analysis of female occupational careers in Italy. <i>Gabriele Ballarino and Stefano Cantalini</i>	236
3.3	Composition in the Data Science Era	242
3.3.1	Can we Ignore the Compositional Nature of Compositional Data by using Deep Learning Approaches? <i>Matthias Templ</i>	243
3.3.2	Principal balances for three-way compositions. <i>Violetta Simonacci</i>	249
3.3.3	Robust Regression for Compositional Data and its Application in the Context of SDG. <i>Valentin Todorov and Fatemah Alqallaf</i>	255

3.4	Evaluation of undercoverage for censuses and administrative data	261
3.4.1	Spatially balanced indirect sampling to estimate the coverage of the agricultural census. <i>Federica Piersimoni, Francesco Pantalone and Roberto Benedetti</i>	262
3.4.2	Next Census in Israel: Strategy, Estimation and Evaluation. <i>Danny Pfeffermann</i>	268
3.4.3	Administrative data for population counts estimations in Italian Population Census. <i>Antonella Bernadini, Angela Chieppa, Nicola Cibella and Fabrizio Solari</i>	274
3.4.4	LFS non response indicators for population register overcoverage estimation. <i>Lorella Di Consiglio, Stefano Falorsi</i>	279
3.5	Excesses and rare events in complex systems	285
3.5.1	Space-time extreme rainfall simulation under a geostatistical approach. <i>Gianmarco Callegher, Carlo Gaetan, Noemie Le Carrer and Ilaria Prosdocimi</i>	286
3.6	Hierarchical forecasting and forecast combination	292
3.6.1	Density calibration with consistent scoring functions. <i>Roberto Casarin and Francesco Ravazzolo</i>	293
3.6.2	Forecasting combination of hierarchical time series: a novel method with an application to CoVid-19. <i>Livio Fenga</i>	298
3.7	Household surveys for policy analysis	304
3.7.1	Did the policy responses to COVID-19 protect Italian households' incomes? Evidence from survey and administrative data. <i>Maria Teresa Monteduro, Dalila De Rosa and Chiara Subrizi</i>	305
3.8	Learning analytics methods and applications	311
3.8.1	Open-Source Automated Test Assembly: the Challenges of Large-Sized Models. <i>Giada Spaccapanico Proietti</i>	312
3.8.2	How Much Tutoring Activities May Improve Academic Careers of At-Risk Students? An Evaluation Study. <i>Marta Cannistra, Tommaso Agasisti, Anna Maria Paganoni and Chiara Masci</i>	318
3.8.3	Composite-based Segmentation Trees to Model Learners' performance. <i>Cristina Davino and Giuseppe Lamberti</i>	324
3.8.4	Test-taking Effort in INVALSI Assessments. <i>Chiara Sacco</i>	330
3.9	Light methods for hard problems	336
3.9.1	Fast Divide-and-Conquer Strategies to Solve Spatial Big Data Problems. <i>Michele Peruzzi</i>	337
3.9.2	Application of hierarchical matrices in spatial statistics. <i>Anastasiia Gorshechnikova and Carlo Gaetan</i>	343
3.10	Management and statistics in search for a common ground (AIDEA)	349
3.10.1	Customer Segmentation: it's time to make a change. <i>Fabrizio Laurini, Beatrice Luceri and Sabrina Latusi</i>	350
3.10.2	Multivariate prediction models: Altman's ZScore and CNDCEC's sectoral indicators. <i>Alessandro Danovi, Alberto Falini and Massimo Postiglione</i>	356
3.10.3	Comparing Entrepreneurship and Perceived Quality of Life in the European Smart Cities: a "Posetic" Approach. <i>Lara Penco, Enrico Ivaldi and Andrea Ciacci</i>	362

3.10.4	The Relationship between Business Economics and Statistics: Taking Stock and Ways Forward. <i>Amedeo Pugliese</i>	368
3.11	Mathematical methods and tools for finance and insurance (AMASES)	373
3.11.1	On the valuation of the initiation option in a GLWB variable annuity. <i>Anna Rita Bacinello and Pietro Millosovich</i>	374
3.11.2	Modern design of life annuities in view of longevity and pandemics. <i>Annamaria Olivieri</i>	380
3.11.3	Risk Management from Finance to Production Planning: An Assembly-to-Order Case Study. <i>Paolo Brandimarte, Edoardo Fadda and Alberto Gennaro</i>	386
3.11.4	Some probability distortion functions in behavioral portfolio selection. <i>Diana Barro, Marco Corazza and Martina Nardonthors</i>	392
3.12	Multiple system estimation	398
3.12.1	Multiple Systems Estimation in the Presence of Censored Cells. <i>Ruth King, Oscar Rodriguez de Rivera Ortega and Rachel McCrea</i>	399
3.12.2	Bayesian population size estimation by repeated identifications of units. A semi-parametric mixture model approach. <i>Tiziana Tuoto, Davide Di Cecco and Andrea Tancredi</i>	405
3.13	Network sampling and estimation	411
3.13.1	Targeted random walk sampling. <i>Li-Chun Zhang</i>	412
3.13.2	Estimation of poverty measures in Respondent-driven sampling. <i>María del Mar Rueda, Ismael Sánchez-Borrego and Héctor Mullo</i>	418
3.13.3	Sampling Networked Data for Semi-Supervised Learning Algorithms. <i>Simone Di Zio, Lara Fontanella, Francesco Pantalone and Federica Piersimoni</i>	423
3.13.4	A sequential adaptive sampling scheme for rare populations with a network structure. <i>Emilia Rocco</i>	429
3.14	New perspectives on multidimensional child poverty	435
3.14.1	Estimating uncertainty for child poverty indicators: The Case of Mediterranean Countries. <i>Ilaria Benedetti, Federico Crescenzi and Riccardo De Santis</i>	436
3.14.2	Child poverty and government social spending in the European Union during the economic crisis. <i>Angeles Sánchez and María Navarro</i>	442
3.14.3	The Children's Worlds Study: New perspectives on children's deprivation research. <i>Caterina Giusti and Antoanneta Potsi</i>	448
3.14.4	The impact of different definition of "households with children" on deprivation measures: the case of Italy. <i>Laura Neri and Francesca Gagliardi</i>	454
3.15	Perspectives in social network analysis applications	460
3.15.1	A comparison of student mobility flows in Erasmus and Erasmus+ among countries. <i>Kristijan Breznik, Giancarlo Ragozini and Marialuisa Restaino</i>	461
3.15.2	Network-based approach for the analysis of LexisNexis news database. <i>Carla Galluccio and Alessandra Petrucci</i>	467
3.15.3	A multiplex network approach to study Italian Students' Mobility. <i>Ilaria Primerano, Francesco Santelli and Cristian Usala</i>	473
3.15.4	Ego-centered Support Networks:a Cross-national European Comparison. <i>Emanuela Furfaro, Elvira Pelle, Giulia Rivellini and Susanna Zaccarin</i>	479

3.16	Statistical analysis of energy data	485
3.16.1	Machine learning models for electricity price forecasting. <i>Silvia Golia, Luigi Grossi, Matteo Pelagatti</i>	486
3.16.2	The impact of hydroelectric storage in the Italian power market. <i>Filippo Beltrami</i>	492
3.16.3	Jumps and cojumps in electricity price forecasting. <i>Peru Muniaín, Aitor Ciarreta and Ainhoa Zarraga</i>	498
3.17	Statistical methods and models for the analysis of sports data	507
3.17.1	Football analytics: a Higher-Order PLS-SEM approach to evaluate players' performance. <i>Mattia Cefis and Maurizio Carpita</i>	508
3.17.2	Bayesian regularized regression of football tracking data through structured factor models. <i>Lorenzo Schiavon and Antonio Canale</i>	514
3.17.3	A dynamic matrix-variate model for clustering time series with multiple sources of variation. <i>Mattia Stival</i>	520
3.17.4	Evaluating football players' performances using on-the-ball data. <i>David Dandolo</i>	526
3.18	The social and demographic consequences of international migration in Western societies	532
3.18.1	Employment and job satisfaction of immigrants: the case of Campania (Italy). <i>Alessio Buonomo, Stefania Capecci, Francesca Di Iorio and Salvatore Strozza</i>	533
3.18.2	Social stratification of migrants in Italy: class reproduction and social mobility from origin to destination. <i>Giorgio Piccitto, Maurizio Avola and Nazareno Panichella</i>	539
3.19	Well-being, healthcare, integration measurements and indicators (SIEDS)	545
3.19.1	A Composite Index of Economic Well-being for the European Union Countries. <i>Andrea Cutillo, Matteo Mazziotta and Adriano Pareto</i>	546
3.19.2	Poverty orderings and TIP curves: an application to the Italian regions. <i>Francesco M. Chelli, Mariateresa Ciommi and Chiara Gagliarano</i>	552
4	Contributed Sessions	558
4.1	Advances in clinical trials	559
4.1.1	Quantitative depth-based [18F]FMCH-avid lesion profiling in prostate cancer treatment. <i>Lara Cavinato, Alessandra Ragni, Francesca Ieva, Martina Sollini, Francesco Bartoli and Paola A. Erba</i>	560
4.1.2	Modelling longitudinal latent toxicity profiles evolution in osteosarcoma patients. <i>Marta Spreafico, Francesca Ieva and Marta Fiocco</i>	566
4.1.3	Information borrowing in phase II basket trials: a comparison of different designs. <i>Marco Novelli</i>	572
4.1.4	Q-learning Estimation Techniques for Dynamic Treatment Regime. <i>Simone Bogni, Debora Slanzi and Matteo Borrotti</i>	578
4.1.5	Sample Size Computation for Competing Risks Survival Data in GS-Design. <i>Mohammad Anamul Haque and Giuliana Cortese</i>	584

4.2	Advances in neural networks	590
4.2.1	Linear models vs Neural Network: predicting Italian SMEs default. <i>Lisa Crosato, Caterina Liberati and Marco Repetto</i>	591
4.2.2	Network estimation via elastic net penalty for heavy-tailed data. <i>Davide Bernardini, Sandra Paterlini and Emanuele Taufer</i>	596
4.2.3	Neural Network for statistical process control of a multiple stream process with an application to HVAC systems in passenger rail vehicles. <i>Gianluca Sposito, Antonio Lepore, Biagio Palumbo and Giuseppe Giannini</i>	602
4.2.4	Forecasting air quality by using ANNs. <i>Annalina Sarra, Adelia Evangelista, Tonio Di Battista and Francesco Bucci</i>	608
4.3	Advances in statistical methods	614
4.3.1	Robustness of Fractional Factorial Designs through Circuits. <i>Roberto Fontana and Fabio Rapallo</i>	615
4.3.2	Multi-objective optimal allocations for experimental studies with binary outcome. <i>Alessandro Baldi Antognini, Rosamaria Frieri, Marco Novelli and Maroussa Zagoraiou</i>	621
4.3.3	Analysis of three-way data: an extension of the STATIS method. <i>Laura Bocci and Donatella Vicari</i>	627
4.3.4	KL-optimum designs to discriminate models with different variance function. <i>Alessandro Lanteri, Samantha Leorato and Chiara Tommasi</i>	633
4.3.5	Riemannian optimization on the space of covariance matrices. <i>Jacopo Schiavon, Mauro Bernardi and Antonio Canale</i>	639
4.4	Advances in statistical methods and inference	645
4.4.1	Estimation of Dirichlet Distribution Parameters with Modified Score Functions. <i>Vincenzo Gioia and Euloge Clovis Kenne Pagui</i>	646
4.4.2	Confidence distributions for predictive tail probabilities. <i>Giovanni Fonseca, Federica Giummolè and Paolo Vidoni</i>	652
4.4.3	Impact of sample size on stochastic ordering tests: a simulation study. <i>Rosa Arboretti, Riccardo Ceccato, Luca Pegoraro and Luigi Salmaso</i>	658
4.4.4	On testing the significance of a mode. <i>Federico Ferraccioli and Giovanna Menardi</i>	664
4.4.5	Hommel BH: an adaptive Benjamini-Hochberg procedure using Hommel's estimator for the number of true hypotheses. <i>Chiara G. Magnani and Aldo Solari</i>	670
4.5	Advances in statistical models	676
4.5.1	Specification Curve Analysis: Visualising the risk of model misspecification in COVID-19 data. <i>Venera Tomaselli, Giulio Giacomo Cantone and Vincenzo Miracula</i>	677
4.5.2	Semiparametric Variational Inference for Bayesian Quantile Regression. <i>Cristian Castiglione and Mauro Bernardi</i>	683
4.5.3	Searching for a source of difference in undirected graphical models for count data – an empirical study. <i>Federico Agostinis, Monica Chiogna, Vera Djordjilovic, Luna Pianesi and Chiara Romualdi</i>	689
4.5.4	Snipped robust inference in mixed linear models. <i>Antonio Lucadamo, Luca Greco, Pietro Amenta and Anna Crisci</i>	695

4.6	Advances in time series	701
4.6.1	A spatio-temporal model for events on road networks: an application to ambulance interventions in Milan. <i>Andrea Gilardi and Riccardo Borgoni and Jorge Mateu</i>	702
4.6.2	Forecasting electricity demand of individual customers via additive stacking. <i>Christian Capezza, Biagio Palumbo, Yannig Goude, Simon N. Wood and Matteo Fasiolo</i>	708
4.6.3	Hierarchical Forecast Reconciliation on Italian Covid-19 data. <i>Andrea Marconcchia, Serena Arima and Pierpaolo Brutti</i>	714
4.6.4	Link between Threshold ARMA and tdARMA models. <i>Guy M��lard and Marcella Niglio</i>	720
4.7	Bayesian nonparametrics	726
4.7.1	Bayesian nonparametric prediction: from species to features. <i>Lorenzo Masoero, Federico Camerlenghi, Stefano Favaro and Tamara Broderick</i>	727
4.7.2	A framework for filtering in hidden Markov models with normalized random measures. <i>Filippo Ascolani, Antonio Lijoi, Igor Pr��nster and Matteo Ruggiero</i>	733
4.7.3	On the convex combination of a Dirichlet process with a diffuse probability measure. <i>Federico Camerlenghi, Riccardo Corradin and Andrea Ongaro</i>	739
4.7.4	Detection of neural activity in calcium imaging data via Bayesian mixture models. <i>Laura D'Angelo, Antonio Canale, Zhaoxia Yu and Michele Guindani</i>	745
4.8	Clustering for complex data	751
4.8.1	Clustering categorical data via Hamming distance. <i>Edoardo Filippi-Mazzola, Raffaele Argiento and Lucia Paci</i>	752
4.8.2	Penalized model-based clustering for three-way data structures. <i>Andrea Cappelozzo, Alessandro Casa, and Michael Fop</i>	758
4.8.3	Does Milan have a smart mobility? A clustering analysis approach. <i>Nicola Cornali, Matteo Seminati, Paolo Maranzano and Paola M. Chiodini</i>	764
4.8.4	A Fuzzy clustering approach for textual data. <i>Irene Cozzolino, Maria Brigida Ferraro and Peter Winker</i>	770
4.8.5	Valid Double-Dipping via Permutation-Based Closed Testing. <i>Anna Vesely, Livio Finos, Jelle J. Goeman and Angela Andreella</i>	776
4.9	Data science for complex data	782
4.9.1	Text mining on large corpora using Taltac4: An explorative analysis of the USPTO patents database. <i>Pasquale Pavone, Arianna Martinelli and Federico Tamagni</i>	783
4.9.2	Emotion pattern detection on facial videos using functional statistics. <i>Rongjiao Ji, Alessandra Micheletti, Natasa Krklec Jerinkic and Zoranka Desnica</i>	789
4.9.3	The spread of contagion on Twitter: identification of communities analysing data from the first wave of the COVID-19 epidemic. <i>Gianni Andreozzi, Salvatore Pirri, Giuseppe Turchetti and Valentina Lorenzoni</i>	795
4.9.4	Composition-on-Function Regression Model for the Remote Analysis of Near-Earth Asteroids. <i>Mara S. Bernardi, Matteo Fontana, Alessandra Menafoglio, Alessandro Pisello, Massimiliano Porreca, Diego Perugini and Simone Vantini</i>	801
4.9.5	Determinants of football coach dismissal in Italian League Serie A. <i>Francesco Porro, Marialuise Restaino, Juan Eloy Ruiz-Castro and Mariangela Zenga</i>	805
4.10	Data science for unstructured data	810
4.10.1	Identification and modeling of stop activities at the destination from GPS tracking data. <i>Nicoletta D'Angelo, Giada Adelfio, Antonino Abbruzzo and Mauro Ferrante</i>	811

Hierarchical Forecast Reconciliation on Italian Covid-19 data

Riconciliazione gerarchica su time-series: applicazione ai dati Covid-19 italiani

Andrea Marcocchia, Serena Arima and Pierpaolo Brutti

Abstract Due to the spread of the Covid-19 pandemic, many different actions to limit personal freedom have been decided according to the trend of the epidemiological situation. Often these decisions were made at the national level or, in a second phase, at the regional one. In order to ensure the implementation of actions and behaviours more in line with the real epidemiological situation, it would be useful to decide and give specific information at a more granular territorial level (provinces) even if one of the main problem is the quality of the data. This paper presents how the use of forecast reconciliation techniques can help forecasts at a detailed territorial level, while exploiting the more robust information available at the aggregate level.

Abstract Con la comparsa della pandemia da Covid-19 si sono rese necessarie azioni di limitazione delle libertà personali, decise in funzione dell'andamento della situazione epidemiologica. Tali decisioni sono state prese a livello nazionale o, in una seconda fase, regionale. Per garantire la messa in atto di azioni maggiormente in linea con la reale situazione epidemiologica di un territorio, sarebbe utile agire ad un livello di maggiore granularità territoriale (province). Uno dei principali problemi nel prendere decisioni a un livello più dettagliato risiede nella qualità dei dati. In questo lavoro si presenta come l'uso di tecniche di forecast reconciliation possono aiutare a migliorare le previsioni ad un livello granulare, sfruttando contemporaneamente le, più robuste, informazioni disponibili a livello aggregato.

Key words: forecasting, time series, forecast reconciliation, hierarchical time series, Covid-19 data

Andrea Marcocchia

Sapienza University of Rome, e-mail: andrea.marcocchia@uniroma1.it

Serena Arima

University of Salento, e-mail: serena.arima@unisalento.it

Pierpaolo Brutti

Sapienza University of Rome, e-mail: pierpaolo.brutti@uniroma1.it

1 Introduction

The main idea behind forecast reconciliation in a hierarchy is that observed demands at each level will always add up to the observed demands at higher levels. It is usually desirable that the same holds true also for forecasts (that is the “forecasting coherence”). If forecasting at the different levels is done independently, we usually have forecast incoherence, meaning that the bottom level forecasts do not add up. The various components of the hierarchy can interact in a variety of complex ways. A change in one series at one level, can have an impact on other series at the same level, as well as on series at higher and/or lower levels. By modeling the entire hierarchy of time series simultaneously, we obtain better forecasts of each of the component series. *Reconciliation* is the process that fix incoherent forecasts. Another important benefit of forecast reconciliation is that, although usually the research interest lays in the most disaggregated data, these are also noisier than the others; in contrast, the “total” series, although less interesting, is typically way more resilient to noise: forecast reconciliation is capable to extract the most relevant information from both series [1].

2 Models

This section presents the used models: the first part explains the forecast algorithms, while the second subsection adds the description of the reconciliation step.

2.1 Forecast models

Any type of prediction methods can be used in the first phase because the reconciliation models have no limits in this regard. The used forecast methods are:

- **ARIMA with and without external covariates:** considering the difficulty in identifying a single parameterization that guarantees a good performance on all the series of the hierarchy, we proceeded using an automatic selection method of the best set of parameters for each series. In an other case, an external variable has been added as additional information. The considered variable indicates for each date whether it is in a lockdown condition or not. The variable is the same at all levels of the hierarchy, therefore restrictions were not considered at the regional or the provincial level, but only at the national level. It should be borne in mind that in the considered period (up to the beginning of November 2020), the closure measures imposed by the Italian government were mainly national, with few measures at the regional level, as lately more frequently decided.
- **Exponential Smoothing (ETS):** also with regard to this method, the best model was selected for each series.

- **Segmented regression:** this approach allows to divide the time series into different segments and to use a simple linear regression to estimate the trend of each segment. It was decided to use this approach after noting that there were purely linear trends within the time series. The number of segments that have been created is equal to 3. The reason is that, until November 2020, there was an initial period of severe growth in the number of daily infections, then a decrease during the summer and then a new growth phase after summer.

2.2 Reconciliation models

The reconciliation problem can be formalize as follows:

- Consider a multi-level hierarchy, where level 0 denotes the completely aggregated series and level k contains the most disaggregated time series and assume that the observations are recorded at times $t = 1, 2, \dots, n$ and that we are interested in forecasting each series at each level at times $t = n + 1, n + 2, \dots, n + h$.
- The additive structure of the time-series involved can be exemplified in a three level hierarchy: $Y_t = \sum_i Y_{i,t}$ represents the most aggregated series, obtained as sum of all the series that belong to the first level ($Y_{i,t}$), that is $Y_{i,t} = \sum_j Y_{ij,t}$ so that the first level series is obtained starting from the second level. As final stage, $Y_{ij,t} = \sum_z Y_{ijz,t}$ represents the aggregation of the lowest level of the hierarchy (the third level, with $k = 3$). The notation refers to $Y_{ij,t}$ as the values of series ij at time t and to Y_t as the aggregation of all the series at time t .
- More generally, let $\mathbf{Y}_{i,t}$ be the vector that collects the values of all the series for $i = 1, \dots, k$, observed at a specific level i at time t . Now define $\mathbf{Y}_t = [\mathbf{Y}_{1,t}, \dots, \mathbf{Y}_{k,t}]'$. According to this notation, we can define $\mathbf{Y}_t = \mathbf{S}\mathbf{Y}_{k,t}$, where $\mathbf{S}$ is a summing matrix used to aggregate the lowest level series. If we assume that the number of bottom level time series is m_k and that m is the total number of series (considering all the levels), then $\mathbf{S}$ is an $(m \times m_k)$ matrix. The $\mathbf{S}$ matrix can be partitioned by the levels of the hierarchy: the top row is a unit vector of length m_k and the bottom section is a $m_k \times m_k$ identity matrix. The symbols m and m_i , are the total number of series in the total hierarchy and at level i , so it is always true that $m_i > m_{i-1}$ and $m = m_0 + m_1 + \dots + m_k$.
- If we compute forecasts for each period $n + 1, n + 2, \dots, n + h$ for a generic series X , we can define this forecasted series $\hat{Y}_{X,n}(h)$. Applying the same logic, $\hat{Y}_n(h)$ denotes the h -step ahead prediction of the total. The final step is to define $\hat{\mathbf{Y}}_n(h)$ as the vector consisting of these base forecasts, stacked in the same series order as for $\mathbf{Y}_t$;
- Bearing all this in mind, all existing hierarchical forecasting methods can be written as $\tilde{\mathbf{Y}}_n(h) = \mathbf{S}\mathbf{P}\hat{\mathbf{Y}}_n(h)$. The effect of the $\mathbf{P}$ matrix is to extract and combine the relevant elements of the base forecasts $\hat{\mathbf{Y}}_n(h)$, which are then summed by $\mathbf{S}$ to give the final revised hierarchical forecasts, $\tilde{\mathbf{Y}}_n(h)$.

The choice of $\mathbf{P}$ is a crucial step: the effect of its choice is to extract and combine the relevant elements of the base forecasts according to: $\mathbf{SP}\hat{\mathbf{Y}}_n(h)$.

According with this formulation, the following reconciliation approaches have been used in order to reconcile the base forecasts introduced in the previous paragraph. Net of some exceptions, all the reconciliation methods were applied for each prediction method:

- **Top-Down:** this method entails the forecasting of the completely aggregated series (higher level of the hierarchy), and then the disaggregation of the forecasts based on historical proportions [5]. In this approach $\mathbf{P} = [\mathbf{p} \mid \mathbf{0}_{m_k \times (m-1)}]$, where $\mathbf{p}$ is a vector of proportions that sum to one. Different methods of top-down forecasting lead to different proportionality vectors $\mathbf{p}$. Some possible choices are the *Average historical proportions* that reflects the average of the historical proportions of the bottom-level series, the *Proportions of the historical averages* that captures the average historical value of the bottom-level series $Y_{j,t}$ relative to the average value of the total aggregate Y_t or the *Forecast proportions* that uses proportions based on forecasts rather than historical data.
- **Bottom-Up:** the idea is to forecast each of the disaggregated series at the lowest level of the hierarchy, and then to obtain forecasts at higher levels of the hierarchy by using simple aggregations. In this case $\mathbf{P} = [\mathbf{0}_{m_k \times (m-m_k)} \mid \mathbf{I}_{m_k}]$, so the $\mathbf{P}$ matrix extracts only bottom level forecasts from $\hat{\mathbf{Y}}_n(h)$, which are then summed by $\mathbf{S}$ to give the bottom-up forecasts.
- **Mint:** the forecast reconciliation through trace minimization is an approach that incorporates the information coming from the full covariance matrix of forecast errors in order to obtain a set of coherent forecasts. It minimizes the mean squared error of the coherent forecasts across the entire collection of time series under the assumption that they are unbiased [3].
- **Bayesian approach:** as typical of this inferential paradigm, the central point is to explore and exploit the posterior distribution of the reconciled forecast. The distribution of the reconciled forecasts can be obtained by sampling from the posterior predictive distribution using a Gibbs sampler. Using such techniques it is possible to keep into account some prior belief about the component time series, so that during the reconciliation step, more importance is given to the more robust time series by taking into account the in-sample variance of each base forecast [2].
- **Ordinary or Weighted Least Square:** the revised forecast at each node will be a weighted average of the forecasts from all nodes. The weights are obtained using linear regression, where all the independent forecasts (from all nodes) are regressed against a set of dummy variables indicating which of the bottom-level series contribute to each node[4]. The mean squared reconciliation error, computed using the differences between the reconciled and independent forecasts, is as small as possible. In the OLS case the $\mathbf{P}$ matrix is equal to $(\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'$.

3 Data

The data used are the Italian Covid-19 incidence (new daily confirmed cases) obtained from the Github repository of the *Protezione Civile* agency. The data are daily updated and the information are released on a province basis. The first considered observation is recorded on 24th February 2020 and the last one on 3rd November 2020 for a total of 254 observations. The data are divided into two groups: the training set (from the first day until 14th October 2020 for 234 observations) and the test set (from 15th October 2020 until the last day for 20 observations). The dataset is made by 107 bottom-level time series, related to the Italian provinces.

Starting from these bottom level time series, a hierarchical structure has been created looking at the Italian administrative organization, so the provinces have been aggregated into regions, the regions into the five macrozones partition and the macrozones summed up to obtain the total. The original data are transformed applying a logarithmic function. Due to the fact that the data are forecasted using models not well performing on count data. In detail, for a single value x_i the following transformation is applied: $x_i = \log(x_i + \epsilon)$.

4 Results

Each forecast and reconciliation method leads to different performances depending on the level of the hierarchy on which we focus. The most interesting level, for the purposes of the research, is the bottom-level series, that is the provinces. In over 70% of the 107 Italian provinces, the prediction adjusted by the reconciliation step is more accurate, in terms of RMSE, than the basic prediction.

Table 1 Count of bottom-level series where a method over-performs the others

Method	Count	Method	Count	Method	Count
Segmented	27	Ets + TD-GSF	4	Arima with cov. + TD-FP	2
Ets + OLS	13	Segmented + WLS	4	Segmented + TD-FP	1
Arima + TD-FP	10	Ets + WLS	3	Arima with cov. + OLS	1
Arima + Mint	9	Arima with cov.	3	Arima + TD-GSA	1
Ets + Mint	8	Arima	3	Arima + OLS	1
Ets + TD-FP	8	Ets + TD-GSA	3	-	-
Arima with cov. + Mint	4	Ets	2	-	-

Analyzing Table 1, it is possible to observe that the prediction method that most often has the best performance is the segmented regression. However, it must be considered that there are some Italian provinces in which the trend of Covid-19 daily cases in the considered period has remained stable and close to 0 even in the test period. Focusing on some of the larger Italian provinces, such as Rome (see Figure 1), the best score obtained without the reconciliation was of an RMSE equal to 0.59

(obtained applying segmented regression), while by inserting the reconciliation step, a value of 0.15 is reached (the ETS method was used for the base forecast). This behaviour is true for most of the bigger Italian provinces.

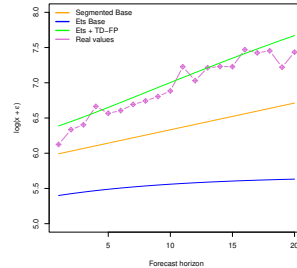


Fig. 1 Forecasted and real values for Covid-19 time series in Rome

It is important to note that among the reconciliation approaches there is no technique that systematically performs better than others. It must be borne in mind that the performance of the various series is profoundly different. Even at a higher level of aggregation (regions, macrozones, ...) there are important differences in the behaviour, which make it difficult to identify a single method valid all over the series.

5 Future works

Given the high variability of methods that over-perform others on the bottom-level time series, the construction of a method that, using a validation frame, is able to identify the best model would be very useful. Reconciliation techniques that make machine learning have recently been presented in the literature, and this could help in solving the problem by providing more accurate predictions.

References

1. G. Athanasopoulos and R. Ahmed and R. J. Hyndman: Hierarchical forecasts for Australian domestic tourism. *International Journal of Forecasting* **25**, 146-166 (2009)
2. F. Eckert and R. Hyndman and A. Panagiotelis: Forecasting Swiss Exports using Bayesian Forecast Reconciliation
3. S. Wickramasuriya and G. Athanasopoulos and R. Hyndman: Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization. *Journal of the American Statistical Association* - **114**, 1-45 (2018)
4. G. Athanasopoulos and R. Hyndman and R. Ahmed and H.L. Shang: Optimal combination forecasts for hierarchical. *Computational Statistics & Data Analysis* - **55**, 2579-2589 (2011)
5. Charles W. Gross and Jeffrey E. Sohl: Disaggregation methods to expedite product line forecasting (1990)

Link between Threshold ARMA and tdARMA models

Relazione tra modelli a soglia ARMA e modelli ARMA con parametri dipendenti dal tempo

Guy Mélard and Marcella Niglio

Abstract In the present contribution, we propose a link between Threshold Autoregressive Moving Average (TARMA) and Time-Dependent ARMA (*tdARMA*) models. We show that a proper parametrization allows to include the TARMA model in the large class of *tdARMA* structures. The main advantage that can be obtained from this result is the derivation of the asymptotic properties of the estimators of TARMA parameters that can be obtained under weaker conditions with respect to those in the available literature.

Abstract *Nel presente contributo proponiamo una relazione tra modelli a soglia autoregressivi media mobile (TARMA) e modelli ARMA con parametri dipendenti dal tempo (tdARMA). Diamo evidenza che una opportuna parametrizzazione consente di includere il modello TARMA nell'ampia classe dei modelli tdARMA. Il principale vantaggio che può essere ottenuto da questo risultato, è la derivazione delle proprietà asintotiche degli stimatori dei parametri TARMA che possono essere ottenuti sotto condizioni più deboli di quelle disponibili in letteratura.*

Key words: Threshold model, time-dependent ARMA model

1 Introduction

In time series analysis, the dynamic of data has often been modelled through the introduction of time-dependent coefficient structures. Starting from [8] different proposals have been given in this domain.

Guy Mélard
Université Libre de Bruxelles, Solvay Brussels School of Economics and Management, ECARES,
Bruxelles, Belgique e-mail: gmelard@ulb.ac.be

Marcella Niglio
Department of Economics and Statistics, University of Salerno, Via Giovanni Paolo II, 132, 84084
Fisciano (SA), Italy e-mail: mniglio@unisa.it