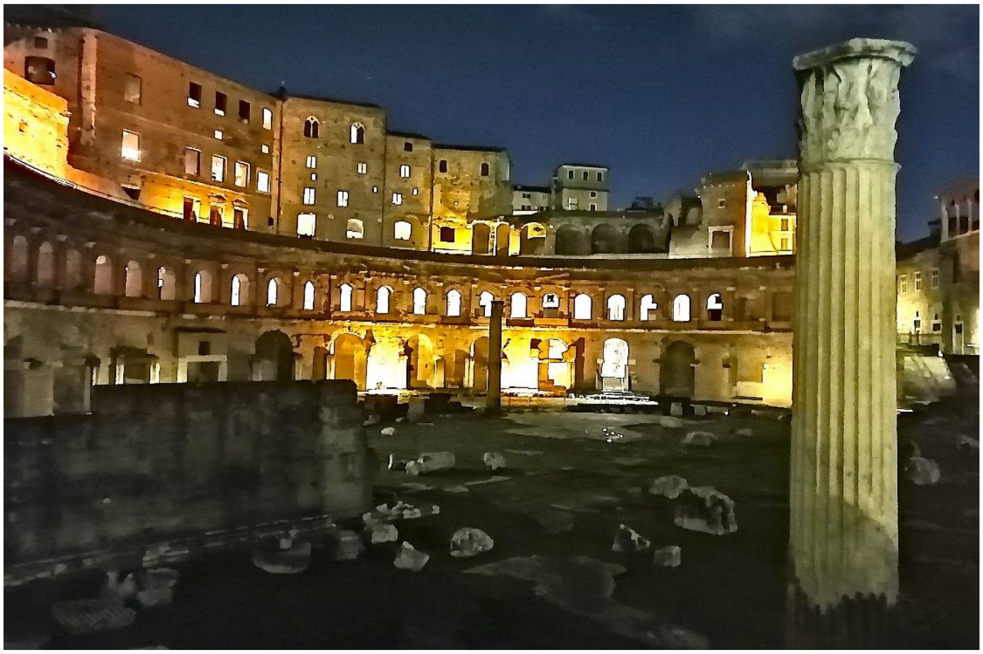
Proceedings of the GRASPA 2021 Conference
Rome, 07-09 June 2021



Edited by: Giovanna Jona Lasinio and Francesco Lagona

2021

Proceedings of the GRASPA 2021 Conference

Rome, 07-09 June 2021

Edited by: Giovanna Jona Lasinio and Francesco Lagona.

University of Rome “La Sapienza”, 2021.

(GRASPA Working Papers)

ISBN: 979-12-200-8496-3

GRASPA 2021

June 7-9 2021

Sapienza University of Rome - Department of Statistics

time schedule

Monday, June 7th

14:30 – 15:00	Conference Opening and Welcome Address
15:00 – 16:00	Invited Session I (3 invited speakers)
16:00 – 16:15	Break with Break-out-rooms
16:15 – 17:15	Keynote speaker I
17:15 – 17:30	Break with Break-out-rooms
17:30 – 18:50	Session II (4 invited speakers)

Tuesday, June 8th

15:00 – 16:00	Invited Session III (3 invited speakers)
16:00 – 16:15	Break with Break-out-rooms
16:15 – 17:15	Keynote speaker II
17:15 – 17:30	Break with Break-out-rooms
17:30 – 18:50	Session IV (4 invited speakers)

Wednesday, June 9th

15:00 – 16:00	Invited Session V (3 invited speakers)
16:00 – 16:15	Break with Break-out-rooms
16:15 – 17:15	Keynote speaker III
17:15 – 17:30	Break with Break-out-rooms
17:30 – 18:50	Session VI (4 invited speakers)
18:50 – 19:10	Best Poster Awards and Closing Cerimony

Monday, June 7th

Session 1: Environmental spatiotemporal statistics (chair: Alessandro Fassò)

15:00-15:20	Song Xi Chen	Effects of COVID-19 Control Measures on Air Quality in North China
15:20-15:40	Mara Bernardi	InSAR data post-processing via functional data analysis with application to the case of Santa Barbara mud volcano
15:40-16:00	Paolo Maranzano	Model selection and inference in spatio-temporal models using a penalised likelihood approach

Keynote Session 1 (chair: Giovanna Jona Lasinio)

16:15-17:15	Sudipto Banerjee	Spatial "Data Science" Meets Bayesian Inference

Session 2: Complex data with spatial dependence (chair: Rosalba Ignaccolo)

17:30-17:50	Stefano Castruccio	Forecasting Air Pollution from Urban Citizen Science Networks with Probabilistically Calibrated Echo State Networks
17:50-18:10	Dyonisios Hristopulos	Boltzmann-Gibbs models for spatio-temporal data
18:10-18:30	Sara Fontanella	Persistent homology in network analysis and its application to environmental data
18:30-18:50	Alessandra Menafoglio	A spatial and distributional analysis of natural background level concentrations in large-scale groundwater bodies

Tuesday, June 8th

Session 3: Marine Ecology (chair: Alessio Pollice)

15:00-15:20	Crescenza Calculli	Bayesian estimation of multiple ecological abundances
15:20-15:40	David Conesa	Incorporating biotic information in Species Distribution Models: a coregionalised approach
15:40-16:00	Sara Martino	The combined use of spatial data derived from conventional research protocols and social media platforms: making this integration possible to predict dolphin distribution

Keynote Session 2 (chair: Francesco Lagona)

16:15-17:15	Ashley Steel	Global Woodfuel Production: A Simple Model?

Session 4: (chair: Massimo Ventrucci)

17:30-17:50	Linda Altieri	Estimation of the biodiversity of a system with covariates and dependence
17:50-18:10	Elias Krainski	Spacetime modeling of ocean carbon in the North Atlantic region
18:10-18:30	Tullia Padellini	Bayesian spatio-temporal model for integrating multiple sources of COVID-19 prevalence
18:30-18:50	Simone Vantini	Nonparametric local inference for functional data with manifold domain and temporal dependence

Wednesday, June 9th

Session 5: Small area estimation (chair: Francesco Finazzi)

15:00-15:20	Serena Arima	Modelling under reporting in small area data
15:20-15:40	Aldo Gardini	Bayesian small area models with log-transformed response
15:40-16:00	Domingo Morales	Poisson mixed models for predicting number of fires

Keynote Session 3 (chair: Luigi Ippoliti)

16:15-17:15	Michela Cameletti	A year of Covid-19 pandemic in Italy: impact on air pollution and mortality

Session 6: Advances in environmental epidemiology (chair: Laura Sangalli)

17:30-17:50	Michela Baccini	Dynamics of SARS-CoV-2 infection in the Italian regions: a descriptive study based on compartmental models
17:50-18:10	Pasquale Valentini	A time varying parameter regression model to investigate the relationship between intensive care occupancies and confirmed COVID-19 deaths in European NUTS-2 Regions
18:10-18:30	Emanuele Giorgi	A geostatistical framework for combining spatially referenced disease prevalence data from multiple diagnostics
18:30-18:50	Maria Franco Villoria	Revisiting space-time disease mapping models

Preface

GRASPA 2021 is the biennial conference of the Italian research group for Environmental Statistics GRASPA-SIS and the major event on Environmental Statistics in Italy. GRASPA 2021 is also the 2021 European regional conference of The International Environmetrics Society and a Satellite Meeting of the 2021 World Statistics Congress. GRASPA 2021 endorses cooperation among statisticians, academics from environmental sciences, and practitioners from government and independent environmental agencies, creating a space for exchanging experiences and ideas on various aspects relevant to protecting the natural environment. GRASPA 2021 is an opportunity to share research interests related to the development and the use of statistical methods in environmental sciences, fostering methodological developments and applications in various environmental fields. This year the online choice for the conference may seem to limit personal exchanges and discussions. We prefer to view it as the occasion to learn how to use new ways to communicate effectively. We imagined a workshop with fewer talks, but all in plenary sessions, posters transformed to short videos and zoom rooms to allow exchanges in small groups. We hope to meet in person very soon and, in the meantime, keep the scientific exchange alive with whatever tool we can use.

Rome,
June 2021

Giovanna Jona Lasinio
Francesco Lagona

Acknowledgements

We wish to thank Pierfrancesco Alaimo Di Loro, Marco Mingione and Marco Pizzuti for the invaluable help provided in the organization of this workshop.

The Scientific Commitee

Rosalba Ignaccolo, University of Torino
Luigi Ippoliti, University of Chieti-Pescara
Alessandro Fassò , University of Bergamo
Francesco Finazzi, University of Bergamo
Giovanna Jona Lasinio, Sapienza University of Rome (chair)
Francesco Lagona, University of Roma Tre
Alessio Pollice, University of Bari
Laura M. Sangalli, Politecnico di Milano
Massimo Ventrucci, University of Bologna

Contents

Part I Keynote sessions

Part II Session 1: Environmental spatio-temporal statistics

Part III Session 2: Complex data with spatial dependence

List of Contributors

Altieri, Linda
University of Bologna, e-mail: linda.altieri@unibo.it

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

Baccini, Michela
Università di Firenze, Italy e-mail: michela.baccini@unifi.it

Banerjee, Sudipto
University of California Los Angeles CA e-mail: sudipto@ucla.edu

Chiara Barbi
Politecnico di Milano e-mail: chiara.barbi@mail.polimi.it

Bernardi, Mara Sabina
Politecnico di Milano, Italy e-mail: marasabina.bernardi@polimi.it

Calculli, Crescenza
University of Bari Aldo Moro, Italy e-mail: crescenza.calculli@uniba.it

Cameletti, Michela
University of Bergamo Italy e-mail: michela.cameletti@unibg.it

Castruccio, Stefano
University of Notre Dame, Notre Dame IN e-mail: scastruc@nd.edu

Conesa, David
University of Valencia e-mail: david.v.conesa@uv.es

D'Angelo, Nicoletta
Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy e-mail: nicoletta.dangelo@unipa.it

Fontanella, Sara
Imperial College London, UK e-mail: s.fontanella@Imperial.ac.uk

Franco-Villoria, Maria
University of Modena e Reggio Emilia, Italy e-mail: maria.francovilloria@unimore.it

Gardini, Aldo
University of Bologna, Italy e-mail: aldo.gardini2@unibo.it

Giorgi, Emanuele
Lancaster University, UK e-mail: e.giorgi@lancaster.ac.uk

Hristopulos, Dionissios T.
Technical University of Crete, Greece. e-mail: dchristopoulos@ece.tuc.gr

Jona Lasinio, Giovanna
Department of Statistical Sciences, Sapienza University of Rome e-mail: giovanna.jonalasinio@uniroma1.it

Krainski, Elias
KAUST, Saudi Arabia, e-mail: elias.krainski@kaust.edu.sa

Maranzano, Paolo
University of Bergamo, Italy e-mail: paolo.maranzano@unibg.it

Martino, Sara Norwegian University if Science and Technology, Norway, e-mail: sara.martino@ntnu.no · Menafoglio, Alessandra
Politecnico di Milano, Italy e-mail: alessandra.menafoglio@polimi.it

Morales, Domingo
Universidad Miguel Hernández de Elche, Spain e-mail: d.morales@umh.es

Moro, Stefano
Department of Environmental Biology, SapienzaUniversity ofRome,Rome, Italy e-mail: stefano.moro@uniroma1.it

Padellini, Tullia
Imperial College London, UK, e-mail: t.padellini@imperial.ac.uk

Panunzi, Greta
Department of Statistical Sciences, Sapienza University of Rome e-mail: greta.panunzi@gmail.com

Pronello, Nicola
Università degli Studi Gabriele d'Annunzio e-mail: nicola.pronello@studenti.unich.it

Song Xi Chen
Peking University, China e-mail: songxichen@pku.edu.cn

Steel, Ashley,
FAO Rome Italy e-mail: Steel@FAO.org

Spoto, Federica
Sapienza University of Rome e-mail: federica.spoto@uniroma1.it

Valentini, Pasquale
Università degli Studi G. d'Annunzio Chieti-Pescara, Italy e-mail: pasquale.valentini@unich.it

Vantini, Simone
Politecnico di Milano, Italy e-mail: simone.vantini@polimi.it

Varini, Elisa
CNR-IMATI Milano, Italy e-mail: elisa.varini@cnr.it

Part I
Keynote sessions

Spatial “Data Science” Meets Bayesian Inference

Sudipto Banerjee

Abstract Geographic Information Systems (GIS) and related technologies such as remote sensors, satellite imaging and portable devices that are capable of collecting precise positioning information, even on portable hand-held devices, have spawned massive amounts of spatial-temporal databases. Spatial "data science" broadly refers to the use of technology, statistical methods, computational algorithms to extract knowledge and insights from spatially referenced data. Applications of spatial-temporal data science are pervasive in the natural and environmental sciences; economics; climate science; ecology; forestry; and public health. With the abundance of spatial BIG DATA problems in the sciences and engineering, GIS and spatial data science will likely occupy a central place in the data revolution engulfing us. This talk will provide an overview of the various challenges data scientists are encountering in analysing massive spatial-temporal data sets in diverse applications. We will begin with a description of different types of spatial data structures and the relevant data analytic questions they pose. We will show, with several examples, the importance of formal statistical inference and, in particular, the many benefits of Bayesian modeling for spatial and spatial-temporal data. We will elucidate recent developments in Bayesian statistical science that harness high performance scientific computing methods for spatial-temporal BIG DATA analysis and emphasize how such methods can be implemented on very modest computing architectures (such as a laptop). The talk will include specific examples of Bayesian hierarchical modeling in Light Detection and Ranging (LiDAR) systems and other remote-sensed technologies; environmental sciences; and public health.

Department of Biostatistics CHS UCLA Fielding School of Public Health, Los Angeles CA e-mail: sudipto@ucla.edu

A year of Covid-19 pandemic in Italy: impact on air pollution and mortality

Michela Cameletti

Abstract The Covid-19 pandemic exploded worldwide in early 2020 has given rise to an enormous amount of research questions across a wide range of research fields, including medicine, economy and psychology among others. In this context, data and statistical models have been playing a major role in the production of scientifically sounds and empirical based results. This talk will discuss two research applications connected with the pandemic. The first part of the talk will regard a comprehensive analysis of the spatio-temporal differences in excess mortality during 2020 in Italy. A spatio-temporal Bayesian model will be presented for the prediction of all-cause weekly deaths and mortality rates at the province level, while adjusting for age, localised temporal trends and the effect of temperature. It will be shown that there are significant differences across gender and geographical areas and that lockdown measures seem to be effective in reducing mortality. The second part of the talk will focus on the effect of the first lockdown (in Spring 2020) on air quality and pollutant concentrations in Lombardia region (Italy). Given the complexity of air pollution in the Po Valley, it is well known that a reduction in emissions, due to the lockdown or other restrictions, does not automatically lead to a significant decrease in concentrations. This is because other factors such as weather conditions and chemical–physical reactions occurring in the atmosphere play an important role in the dynamics of air pollution. Thanks to a spatio-temporal Bayesian model we will be able to show the differences, in terms of NO2 concentrations, between 2020 and the previous years with a weekly temporal resolution. More importantly, this information will be available across the entire region, also where no monitoring stations are available. The effect of the lockdown on air quality will be evident, even if with some geographical peculiarities.

University of Bergamo, Italy e-mail: michela.cameletti@unibg.it

Global Woodfuel Production: A Simple Model?

E. Ashley Steel

Abstract Woodfuel, just sticks or lumps of charcoal, is the core of essential global systems: energy, human health, and ecology. An estimated 2.8 billion people use woodfuel as a primary energy source. The use of woodfuel is a significant contributor to household air pollution and the third leading risk factor of disease burden worldwide. Globally, woodfuel production may also be a driver of deforestation and forest degradation. For statisticians, estimating exactly how much woodfuel is produced or consumed is a particular challenge because so much of woodfuel production and trade is informal and unregistered. Over the past two years, FAO's Forest Products and Statistics Team, which provides the best available global data on production and trade of forest products via FAOSTAT, has taken the lead in updating the model, or collection of models, used to estimate national woodfuel production where countries do not supply official data. I will describe the development of a conceptual model to underpin the work, insights from a systematic country-by-country search for new data, and trade-offs in final model selection. I will conclude with the essential interaction between topical knowledge and statistical knowledge, the importance of mapping information flow, and thoughts on what constitutes a "simple" model.

Food and Agriculture Organization UN, Rome Italy e-mail: Ashley.Steel@FAO.org

Part II
Session 1: Environmental spatio-temporal statistics

Effects of COVID-19 Control Measures on Air Quality in North China

Song Xi Chen

Abstract Corona Virus Disease-19 (COVID-19) has substantially reduced human activities and the associated anthropogenic emissions. This study quantifies the effects of COVID-19 control measures on six major air pollutants over 68 cities in North China by a Difference in Relative-Difference method that allows estimation of the COVID-19 effects while taking account of the general annual air quality trends, temporal and meteorological variations, and the spring festival effects. Significant COVID-19 effects on all six major air pollutants are found, with NO_2 having the largest decline (-39.6%), followed by $PM_{2.5}$ (-30.9%), O_3 (-16.3%), PM_{10} (-14.3%), CO (-13.9%), and the least in SO_2 (-10.0%), which shows the achievability of air quality improvement by a large reduction in anthropogenic emissions. The heterogeneity of effects among the six pollutants and different regions can be partly explained by coal consumption and industrial output data. This is a joint work with Xiangyu Zheng, Bin Guo, and Jing He.

Keywords: Difference in Relative-Difference method, Meteorological adjustment, Treatment effects estimation

Song Xi Chen,
Peking University, Beijing 100871 P.R. China, e-mail: songxichen@pku.edu.cn

Model selection and inference in spatio-temporal models using a penalised likelihood approach

Paolo Maranzano, Alessandro Fassò and Philipp Otto

Abstract In this paper we discuss model selection techniques based on penalised likelihood and LASSO regularisation [2, 1] for geostatistical models. In particular, we are interested in applying feature selection algorithms to the Hidden Dynamics Geostatistical Models, or HDGM, [6] using a multiple-stage approach. HDGM represents the phenomenon of interest using a mixed-effects structure, in which the random-effect term describes the spatio-temporal dynamics and the fixed-effect component models the interaction between the response variable and exogenous phenomena via linear regression. Our focus is on identifying a procedure to select the subset of relevant covariates included in the fixed-effect component. Feature selection issues in geostatistical framework are receiving great attention from researchers due to the increasing availability of georeferenced data. Recent contributions proposed model selection strategies based on graphical-LASSO algorithms ([9]) or penalised likelihood methods with covariance-tapering ([8]). Here we propose a multistage algorithm which integrates model estimation, parameters shrinkage and forecasting performances evaluation. We suggest using a scheme that combines the reduction of the initial number of predictors using a LASSO-like mechanism and the selection of the optimal model via spatio-temporal and spatial cross-validation. Such methods are applied to both simulated and real-world data. The empirical data concerns the case study of airborne pollutant concentrations observed during the lockdown period imposed in 2020 to face the COVID-19 virus spread in Northern Italy. Spatio-temporal models are common tools in statistical air quality control. In this context the number of exogenous factors involved, e.g. weather and seasonal components, can grow very rapidly and negatively affecting the computational demand. For this reason, feature selection covers a remarkable role.

Paolo Maranzano & Alessandro Fassò
University of Bergamo, Dept. of Management, Information and Production Engineering, Via Pasubio 7, Dalmine, Italy, e-mail: paolo.maranzano@unibg.it

Philipp Otto
Institute of Cartography and Geoinformatics, University of Hannover Germany e-mail: philipp.otto@ikg.uni-hannover.de

Key words: Spatio-temporal models, PMLE, LASSO, Feature selection, Spatio-temporal cross-validation, HDGM

References

1. Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties, Journal of the American statistical Association, 2001, 96, 1348-1360
2. Hastie, T., Tibshirani, R. and Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Springer Series in Statistics) Springer New York, 2017
3. James, G., Witten, D., Hastie, T. and Tibshirani, R. An introduction to statistical learning Springer, 2013, 112
4. Zou, H. and Li, R. One-step Sparse Estimates in Nonconcave Penalized Likelihood Models Annals of statistics, 2008, 36, 1509-1533
5. Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. Least angle regression Ann. Statist., 2004, 32, 407-499
6. Calculli, C., Fassò, A., Finazzi, F., Pollice, A. and Turnone, A. Maximum likelihood estimation of the multivariate hidden dynamic geostatistical model with application to air quality in Apulia, Italy Environmetrics, 2015, 26, 406-417
7. Wang, Y., Finazzi, F. and Fassò, A. D-STEM v2: A Software for Modelling Functional Spatio-Temporal Data arXiv preprint arXiv:2101.11370, 2021
8. Chu, T., Zhu, J. and Wang, H. Penalized maximum likelihood estimation and variable selection in geostatistics The Annals of Statistics, 2011, 39, 2607-2625
9. Haworth, J. and Cheng, T. Graphical LASSO for local spatio-temporal neighbourhood selection Proceedings the GIS Research UK 22nd Annual Conference. Presented at the GISRUK, 2020, 425-433

Part III
Session 2: Complex data with spatial dependence

Boltzmann-Gibbs Models for Spatiotemporal Data

Dionissios T. Hristopulos

Abstract Boltzmann-Gibbs (BG) models enforce spatial (temporal) dependence by means of interactions at neighboring sites and/or times. The interactions are incorporated in a scalar energy functional which acts as the exponent of the exponential BG density. For Gaussian BG models, the spatial structure of the interactions determines the model's precision matrix. The latter is sparse by construction if the interactions are local. This property can lead to computationally efficient, estimation, interpolation and simulation algorithms. This contribution reviews Boltzmann-Gibbs models in relation to spatial and space-time data. Three different topics are discussed: (1) The construction of new covariance models. (2) The connection of lattice-based BG models to Gauss-Markov random fields. (3) Extensions to irregular grids by means of interactions modulated via kernel functions.

Keywords: spatial estimation, kriging, space-time models, big data, interpolation, kernel function

1 Introduction

Boltzmann-Gibbs (BG) models have their roots in statistical physics and the early works of Boltzmann and Gibbs on statistical mechanics [Feynman, 1982]. The main idea is that different realizations of "interacting random variables" can be represented in terms of exponential probability density functions $f \sim \exp(-H)$, where H are suitably defined energy functions. Sets of interacting random variables on spatiotemporal domains represent space-time random fields. A famous BG paradigm is the magnetic Ising model [Ising, 1925] which is equivalent to the auto-logistic model of Besag in spatial statistics [Besag, 1972].

Dionissios T. Hristopulos
School of Electrical and Computer Engineering, Technical University of Crete, Chania, 73100, Greece. e-mail: dchristopoulos@ece.tuc.gr

In the following, we refer to both the random field and its realizations (states) by $x(\mathbf{s})$, $\mathbf{s} \in D \subset \mathbb{R}^d$. In continuum domains the states represent functions and on discrete domains (grids) vector variables. The energy H is a functional of the random field states $x(\mathbf{s})$. The functional form of H determines the joint probability density function via $f \sim \exp(-H)$. Gaussian random fields are obtained if H is a quadratic functional of $x(\mathbf{s})$. In contrast with the common definition used in statistics, BG random fields are defined in terms of the interactions incorporated in the energy functional instead of the mean and covariance function. The interactions couple values of the states at different locations (and/or times); for Gaussian fields defined on discrete domains the interactions determine the precision matrix (inverse covariance). In the continuum case, the inverse covariance corresponds to a precision operator [Hristopulos, 2020].

The main reasons for studying BG models are the following: (1) Development of new covariance models that provide parametric flexibility. (2) Generation of efficient and intuitive local models for regular grids. (3) Extensions to irregular grids. In cases (2) and (3) the models are defined in terms of sparse precision matrices leading to computationally efficient approaches for estimation, interpolation and simulation.

2 BG Models defined in Continuum Domains

Spatial covariance models: Let us consider BG scalar Gaussian random fields $\mathbb{R}^d \mapsto \mathbb{R}$ with energy expressed by means of the integral

$$H = \frac{1}{2} \int_{\mathbb{R}^d} x(\mathbf{s})\, p_M(L) x(\mathbf{s})\, d\mathbf{s}, \tag{1}$$

where $p_M(L) = \sum_{m=0}^{M} c_m \nabla^{2m}$ is a polynomial of the Laplacian operator $L = \nabla^2$. Proper definition of the derivatives requires the formalism of generalized random fields [Gelfand et al, 2014]. Furthermore, let the image of $p_M(L)$ under the Fourier transform be $\tilde{p}_M(k) = \sum_{m=0}^{M} (-1)^m c_m k^{2m}$, where k is the Euclidean measure of the wavevector (spatial frequency vector) [Lighthill, 1958]. Assuming that $\tilde{p}_M(k)$ has no real-valued roots, Eq. (1) defines an admissible energy functional which corresponds to a radial spectral density given by rational function $\tilde{C}(k) = 1/\tilde{p}_M(k)$.

BG covariance functions for $M = 2$ were obtained in [Hristopulos, 2003, Hristopulos and Elogne, 2007, Hristopulos, 2015]. Covariances for general M (henceforward, BG-M) were obtained in [Yaremchuk and Smith, 2011]. BG-M ($M > 2$) models were derived from rational spectral densities without connection to BG random fields. BG-M models can have spectral densities with multiple modes which are useful for modeling ocean waves.

The interesting property of $M = 2$ BG covariance functions is that in one dimension (i.e., for time series or drill-hole data), they generate the three different regimes (underdamped, overdamped, critical) corresponding to the covariance of a linear damped harmonic oscillator driven by white noise [Hristopulos, 2020]. In two- and three- dimensional domains they retain the three-regime form and provide functions of the spatial lag with oscillatory and heavily damped dependence. The parameters of BG-2 covariance functions involve, in addition to the variance and characteristic length, a dimensionless rigidity coefficient which determines the respective regime.

Karhunen-Loéve expansion: In one dimension the Karhunen-Loéve (KL) representation of BG-2 covariance functions over a bounded domain has been derived in terms of regime-dependent eigenbases that involve combinations of harmonic and hyperbolic functions [Tsantili and Hristopulos, 2016]. This result can be used for dimensionality reduction in spaces $d > 1$ using separable covariance functions constructed by means of the product rule and the KL basis expansion in $d = 1$.

Anisotropy and vector random fields: It is mathematically trivial to extend BG-M covariance models to include geometric anisotropy. It is also possible to extend Gaussian BG energy functionals to multivariate models and respective matrix-valued covariance functions applicable to vector random fields as is done in [Hristopulos and Porcu, 2014].

Space-time covariance models: Extending Gaussian BG models to space-time is in principle trivial using suitable modifications of the energy function to include time as well as space. However, obtaining explicit covariance functions for such models is not an easy task [**?**]. Nonetheless, an explicit, non-separable, space-time covariance which involves the error function was obtained by combining the BG representation with linear response theory and the turning bands method [Hristopulos and Tsantili, 2017].

Alternatively, space-time covariance functions can be constructed using the spatial BG models and a composite space-time distance with lag $h = \sqrt{\|\mathbf{r}\|^2 + \alpha^2\tau^2}$, where $\mathbf{r}$ and τ are respectively the spatial lag vector and the temporal lag [Varouchakis and Hristopulos, 2019].

3 BG Models defined on Grids

Connection with Markov Random Fields: The continuum BG models defined by the energy functional (1) can be discretized on regular grids by replacing derivatives with finite differences approximations [Hristopulos, 2003, Hristopulos, 2020]. Then the BG models become equivalent to Gaussian Markov random fields (GMRFs), with a specific structure (precision matrix) inherited from the energy functional. The theoretical machinery and computational efficiency of GMRFs [Rue and Held, 2005] can be applied to grid-based BG models.

Irregular Grids: BG models with sparse precision matrices, in the spirit of GMRFs, can also be constructed for irregular spatial [Hristopulos, 2015a, Hristopulos, 2020] or space-time grids [Hristopulos and Agou, 2020]. The main idea is to express the energy in terms of kernel functions that extend the interactions over adaptively determined local neighborhoods. For compactly supported kernel functions the precision matrix is quite sparse (the fraction of non-zero elements is less than 1%). Sparsity allows for direct calculations of the full likelihood for datasets comprising $10^4 - 10^5$ points. In addition, the conditional mean and variance are fully determined from the precision matrix. Hence, matrix inversion is not required for predictive purposes.

4 Conclusions and Future Research

The BG formulation provides a fruitful and flexible framework for the development of space-time models which is complementary to the covariance-based approach. Links of Gaussian BG models with stochastic partial differential equations can be further explored [Hristopulos, 2020]. The development of sparse models for irregular grids is ongoing and involves topics related to model flexibility, estimation and simulation.

References

Besag, 1972. Besag, J.E.: Nearest-neighbour systems and the auto-logistic model for binary data. Journal of the Royal Statistical Society: Series B (Methodological) **34**(1), 75–83 (1972)

Feynman, 1982. Feynman, R.P.: Statistical Mechanics. Benjamin and Cummings, Reading, MA (1982)

Gelfand et al, 2014 . Gelfand, I.M., Vilenkin, N.Y.: Generalized functions: Applications of Harmonic Analysis, vol. 4. Academic Press (2014)

Hristopulos, 2003. Hristopulos, D.: Spartan Gibbs random field models for geostatistical applications. SIAM Journal of Scientific Computing **24**(6), 2125–2162 (2003)

Hristopulos and Elogne, 2007. Hristopulos, D., Elogne, S.: Analytic properties and covariance functions of a new class of generalized Gibbs random fields. IEEE Transactions on Information Theory **53**(12), 4667–4679 (2007)

Hristopulos, 2015. Hristopulos, D.T.: Covariance functions motivated by spatial random field models with local interactions. Stochastic Environmental Research and Risk Assessment **29**(3), 739–754 (2015)

Hristopulos, 2015a. Hristopulos, D.T.: Stochastic local interaction (SLI) model: Bridging machine learning and geostatistics. Computers & Geosciences **85**, 26–37 (2015)

Hristopulos, 2020. Hristopulos, D.T.: Random Fields for Spatial Data Modeling: A Primer for Scientists and Engineers. Springer, Dordrecht, the Netherlands (2020)

Hristopulos and Agou, 2020. Hristopulos, D.T., Agou, V.D.: Stochastic local interaction model with sparse precision matrix for space–time interpolation. Spatial Statistics **40**, 100,403 (2020)

Hristopulos and Porcu, 2014. Hristopulos, D.T., Porcu, E.: Multivariate spartan spatial random field models. Probabilistic Engineering Mechanics **37**, 84–92 (2014)

Hristopulos and Tsantili, 2016. ristopulos, D.T., Tsantili, I.C.: Space-time models based on random fields with local interactions. International Journal of Modern Physics B **30**(15), 1541,007 (2016)

Hristopulos and Tsantili, 2017. Hristopulos, D.T., Tsantili, I.C.: Space–time covariance functions based on linear response theory and the turning bands method. Spatial Statistics **22**, 321–337 (2017)

Ising, 1925. Ising, E.: Contribution to the theory of ferromagnetism. Z. Phys. **31**(1), 253–258 (1925)

Lighthill, 1958. Lighthill, M.J.: An Introduction to Fourier Analysis and Generalised Functions. Cambridge University Press (1958)

Rue and Held, 2005. Rue, H., Held, L.: Gaussian Markov Random Fields: Theory and Applications. Chapman and Hall/CRC (2005)

Tsantili and Hristopulos, 2016. Tsantili, I.C., Hristopulos, D.T.: Karhunen–Loeve expansion of Spartan spatial random fields. Probabilistic Engineering Mechanics **43**, 132–147 (2016)

Varouchakis and Hristopulos, 2019. Varouchakis, E.A., Hristopulos, D.T.: Comparison of spatiotemporal variogram functions based on a sparse dataset of groundwater level variations. Spatial Statistics **34**, 100,245 (2019)

Yaremchuk and Smith, 2011. Yaremchuk, M., Smith, S.: On the correlation functions associated with polynomials of the diffusion operator. Quarterly Journal of the Royal Meteorological Society **137**(660), 1927–1932 (2011)

Persistent homology in network analysis and its application to environmental data

L. Fontanella, S. Fontanella, R. Ignaccolo, L. Ippoliti, P. Valentini

Abstract Recent years have seen a growing interest in network modelling applied to many scientific fields, such as social sciences, medicine, physics and environmental science. In the conceptualization of complex systems as networks, the system components are represented as nodes and interactions between components are represented as links. Correlation-based networks and graph-theory based properties can be successfully used to investigate the relational structures among the components. As correlation weighted networks are difficult to interpret and visualize, a common approach is to construct a sparse threshold network, that is the binary network obtained by setting a threshold weight λ and keeping only connections with weights higher than λ. Several approaches have been proposed to choose the threshold; however, different choices of the filtration parameter λ will result in different solutions in the statistical model. In this work, to overcome this issue, we discuss a framework that performs statistical inferences over the whole parameter space using persistent homology [Horak et al., 2009]. In this context, persistent homology allows to investigate the structural properties of the network and their changes for a collection of threshold values. This procedure results in the definition of a collection of nested networks over every possible threshold using *graph filtration*, which is a threshold-free framework for analysing a family of hierarchical graphs. Using persistent homology,

Lara Fontanella
Department of Legal and Social Science, G. d'Annunzio University Chieti-Pescara, Italy, e-mail: lara.fontanella@unich.it

Sara Fontanella
National Heart and Lung Institute, Imperial College London, Lndon, UK e-mail: s.fontanella@Imperial.ac.uk

Rosaria Ignaccolo
Department of Economics and Statistics "Cognetti de Martiis, University of Torino, Torino, Italy, e-mail: rosaria.ignaccolo@unito.it

Luigi Ippoliti - Pasquale Valentini
Department of Economics, University G. d'Annunzio, Chieti-Pescara, Italy , e-mail: ippoliti@unich.it, e-mail: pvalent@unich.it

topological features, such as the connected components and cycles of a graph, can be expressed in terms of the Betti numbers which will be used for differential network analysis. As an application of the proposed method, we apply the technique to quantify the spreading and diffusion of NO_2 concentration patterns via correlation networks, where sites are considered as nodes and link weights express the similarity between the time series of sites. Since there is no prior study on the statistical distribution of Betti numbers, a discussion on how to perform inferential analysis using MCMC algorithm is also provided.

Keywords: Network analysis, Persistent homology, Graph filtration, Differential correlation analysis

References

Horak et al., 2009. Horak, D., Maletić, S., and Rajković, M. (2009). Persistent homology of complex networks.*Journal of Statistical Mechanics: Theory and Experiment*, 2009(03).

An object-oriented approach to the analysis of natural background level concentrations of chemical species in large-scale groundwater bodies

A. Menafoglio, L. Guadagnini, A. Guadagnini, and P. Secchi

Abstract Natural background levels (NBLs) of a chemical species in a groundwater body represents concentrations corresponding to natural conditions which can be considered as unaffected by anthropogenic activities. Characterization of the NBL of aquifer systems is key to identify significant trends of contaminant concentrations and to plan possible actions to reverse undesired trends.

Here, we propose a novel framework to assess the NBL of target chemical species in large-scale groundwater bodies. As atom of the analysis, we consider the probability density function (PDF) of the chemical species of interest. Setting the latter as the object of the statistical analysis is a critical element of innovation with respect to previous works (e.g., [Molinari et al, 2012]), which are mainly focused on the median or on selected quantiles of chemical concentrations. The PDFs of the chemical species are here modeled as random points in a Bayes Hilbert space [Boog et al, 2014] and analyzed in the context of Object Oriented Spatial Statistics (O2S2, [Menafoglio and Secchi, 2017]). In this contribution, we present our novel approach proposed in [Menafoglio et al, 2021], which enables us to take advantage of the entire information content provided by these objects for the purpose of *(i)* identification of central and outlying observations, through empirical prediction bands

A. Menafoglio
MOX, Department of Mathematics, Politecnico di Milano, Milano, Italy; e-mail: alessandra.menafoglio@polimi.it

L. Guadagnini
Department of Civil and Environmental Engineering, Politecnico di Milano, Milano, Italy; e-mail: laura.guadagnini@upc.edus

A. Guadagnini
Department of Civil and Environmental Engineering, Politecnico di Milano, Milano, Italy; e-mail: alberto.guadagnini@polimi.it

P. Secchi
MOX, Department of Mathematics, Politecnico di Milano, Milano, Italy; Center for Analysis, Decisions and Society, Human Technopole, Milano, Italy; e-mail: piercesare.secchi@polimi.it

based on depth measures; *(ii)* prediction, e.g., via kriging, *(iii)* uncertainty quantification based on the generation of multiple scenarios (i.e., conditional stochastic simulation), or *(iv)* assessment of indicators inferable from the PDFs, including, e.g., the probability of exceeding reference thresholds.

This work is motivated by the analysis of the NBL of Ammonium (NH_4), based on the temporal observation records available at 61 monitoring locations in an alluvial aquifer of the Apennines and Po river plains in the Emilia-Romagna region, Northern Italy. We describe the results of the proposed methodological and theoretical framework with reference to this case study, and illustrate the potential of the new approach in comparison to previous works.

Keywords: Kriging; conditional simulations; probability density functions; spatial depth measure.

References

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

Menafoglio et al, 2021. Menafoglio, A,, L. Guadagnini, A. Guadagnini, and P. Secchi (2021). Object oriented spatial analysis of natural concentration levels of chemical species in regional-scale aquifers. *Spatial Statistics* 43, 100494.

Molinari et al, 2012. Molinari, A., L. Guadagnini, M. Marcaccio, and A. Guadagnini (2012). Natural background levels and threshold values of chemical species in three large-scale groundwater bodies in Northern Italy. *Science of the Total Environment* 425, 9–19.

Boog et al, 2014. van den Boogaart K.G., J.J. Egozcue, and V. Pawlowsky-Glahn (2014). Bayes Hilbert spaces. *Australian and New Zealand Journal Statistics* 56, 171–194.

Part IV
Session 3: : Marine Ecology

Bayesian estimation of multiple ecological abundances

C. Calculli and S. Martino and P. Maiorano

Abstract Ecological processes driving the spatial and spatio-temporal distribution of marine species are complex to assess. In fact in several ecological studies, counts, abundances or biomass of interacting species are collected from different sites resulting in sparse datasets that include highly correlated responses. The analysis of relationships among such responses requires a suitable statistical framework to globally study the ecosystem, including relevant variables and combining in a single step environmental and community information. Inspired by Joint species distribution models, we propose a Bayesian model-based approach to deal with the zero-inflation issue, common to semi-continuous data, and with the spatial (and spatio-temporal) structure of abundance monitoring data. The proposal takes its cue from a case study concerning marine litter data collected by fishery surveys in the central Mediterranean. To jointly infer different litter categories, a multiple response Hurdle-model is proposed. This model allows to combine both information on occurrence and conditional-to-presence abundance of litter categories and the effects of environmental potential drivers. Shared spatial effects that link abundances and probabilities of occurrences, together with temporal effects, are efficiently implemented using the SPDE-INLA approach. Results support the possibility of better understanding the spatio-temporal dynamics of marine litter in the study area.

Keywords: spatio-temporal data, species distribution models, hurdle models, marine litter, INLA

Crescenza Calculli
Department of Economics and Finance, University of Bari Aldo Moro, Largo Abbazia S. Scolastica - 70124 Bari, Italy e-mail: crescenza.calculli@uniba.it

Sara Martino
Department of Mathematical Sciences, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway e-mail: sara.martino@ntnu.no

Porzia Maiorano
Department of Biology, LRU CoNISMa, University of Bari Aldo Moro, 70125 Bari, Italy e-mail: porzia.maiorano@uniba.it

Incorporating biotic information in Species Distribution Models: a coregionalised approach

X. Barber, David Conesa, A. Ló pez-Quí lez, J. Martí nez-Minaya, I. Paradinas and M.G. Pennino

Abstract In this work, published in Barber et al (2021), we discuss the use of a methodological approach for modelling spatial relationships among species by means of a Bayesian spatial coregionalized model. Inference and prediction is performed using the integrated nested Laplace approximation methodology to reduce the computational burden. We illustrate the performance of the coregionalized model in species interaction scenarios using both simulated and real data. The simulation demonstrates the better predictive performance of the coregionalized model with respect to the univariate models. The case study focus on the spatial distribution of a prey species, the European anchovy (*Engraulis encrasicolus*), and one of its predator species, the European hake (*Merluccius merluccius*), in the Mediterranean sea. The results indicate that European hake and anchovy are positively associated, resulting in improved model predictions using the coregionalized model.

X. Barber
Center of Operations Research (CIO), Universidad Miguel Hernández, 03202 Elche, Alicante, Spain, e-mail: xbarber@umh.es

David Conesa
Department of Statistics and Operations Research, University of Valencia, 46100 Valencia, Spain, e-mail: david.v.conesa@uv.es

A. Ló pez-Quí lez
Department of Statistics and Operations Research, University of Valencia, 46100 Valencia, Spain, e-mail: antonio.lopez@uv.es

J. Martí nez-Minaya
Data Science Area, Basque Center for Applied Mathematics (BCAM), 14 E48009 Bilbao, Basque Country, Spain, e-mail: jomartinez@bcamath.org

I. Paradinas
Scottish Ocean's Institute, University of St Andrews, St Andrews KY16 9AJ, UK, e-mail: paradinas.iosu@gmail.com

M.G. Pennino
Instituto Español de Oceanografía, Centro Oceanográfico de Vigo, Subida a Radio Faro, 50-52, 36390-Vigo, Spain, e-mail: grazia.pennino@ieo.es

Keywords: Bayesian hierarchical models; coregionalized models; fisheries; INLA; predation; SPDE; species interaction.

References

Barber et al, 2021. X. Barber, D. Conesa, A. Ló pez-Quí lez, J. Martí nez-Minaya, I. Paradinas and M.G. Pennino (2021). Incorporating Biotic Information In Species Distribution Models: A Corregionalised Approach. *Mathematics*, **9**(4), 417.

Combining spatial data derived from conventional research protocols and social media platforms: making the integration possible to predict dolphin distribution

Sara Martino, Daniela Silvia Pace, Giovanna Jona Lasinio

Abstract Presence-only data are typically used in species distribution modeling. Data are often originated from different sources: their integration is both a resource and a challenge. We propose a new protocol for presence-only data fusion, where information sources include social media platforms, with the aim to investigate several possible solutions to reduce uncertainty in the modeling outputs. As a case study, we use spatial data on two dolphin species with different ecological characteristics and distribution, collected in the central Mediterranean through traditional research campaigns and derived from a careful selection of social media images and videos. We build a Spatial Log-Gaussian Cox Process that incorporates different detection functions and thinning for each data source. The Bayesian framework, allows us to specify slightly informative priors to avoid identifiability issues when estimating both the animal intensity and the observation process. We compare different types of detection functions and accessibility explanations. We show how the shape of the detection function affects ecological findings on two species under examination. Our findings allow for a sound understanding of the species distribution in the study area, confirming the proposed approach's appropriateness. Besides, fast inference and the straightforward implementation in the R software, make the proposed approach widely functional and easy to apply on different species and ecological contexts. This work has been made with the collaboration of: Stefano Moro, Edoardo Casoli,Daniele Ventura, Alessandro Frachea, Margherita Silvestri, Antonella Arcangeli, Giancarlo Giacomini, Giandomenico Ardizzone

Sara Martino
Norwegian University if Science and Technology, Trondheime-mail: sara.martino@ntnu.no

Daniela Silvia Pace and Giovanna Jona Lasinio
Sapienza University of Rome Italy

Part V
Session 4: Modelling data with complex dependencies

Estimating the biodiversity of a system with covariates and dependence structures

Linda Altieri and Daniela Cocchi

Abstract Entropy is widely used in ecological and environmental studies, with data often presenting complex interactions. Difficulties arise when the interest is to link such measure to available covariates or data dependence structures, as all existing approaches to entropy estimation assume independence. We focus on improving the estimation of the probabilities of the species in biodiversity studies, accounting for any data dependence and correlation. Estimating entropy is then straightforward and may be an informative index of a system's biodiversity. An application is presented about the biodiversity of rainforest tree data.

Keywords: Entropy estimation, biodiversity, Bayesian multinomial regression, correlated data, rainforest trees.

1 Introduction

Let us consider the estimation of the biodiversity of a system: the number of species is I, the total number of observations is n and the number of observations for each species is n_i, with $i = 1, \dots, I$. Let us take $\widehat{p}(x_i) = n_i/n$ as the ML estimate of $p(x_i)$, the probability of occurrence of the i-th species; then, $\widehat{H}_{ML}(X)$ is the plug-in estimator of Shannon's entropy of the number of species [3]

$$H(X) = \sum_{i=1}^{I} \log p(x_i) \log \frac{1}{p(x_i)}. \tag{1}$$

Linda Altieri, Daniela Cocchi
Department of Statistical Sciences, University of Bologna,
e-mail: linda.altieri@unibo.it;daniela.cocchi@unibo.it

$\widehat{H}_{ML}(X)$ is the maximum likelihood (ML) estimator [5]; it is known to be biased and cannot include covariates or correlation.

To our knowledge, in biodiversity studies, the ML estimator only considers the number of observed species for each site; it cannot account for absent species and for the possibility that species are not the same across observation sites, nor for covariates or other structures. No studies are currently able to include dependence.

We take a new perspective to entropy estimation. If the probability mass function (pmf) of the variable of interest can be properly estimated accounting for dependence to covariates or spatial/temporal factors, such information can be used to enrich an estimator of entropy. In the case of categorical variables, a Bayesian model-based approach for multinomial data allows to derive the pmf parameters including dependence structures. The posterior distribution of entropy may be obtained as a transformation of the posterior estimated pmf.

Our motivating dataset documents the presence of tree species over Barro Colorado Island, Panama. Barro Colorado Island has been the focus of intensive research on lowland tropical rainforest since 1923 (http://www.ctfs.si.edu). Research identified several tree species over a rectangular observation window of size 1000x500 metres; information is available about the soil elevation and slope. We focus on 4 tree species with different spatial configurations, where it is plausible to hypothesize dependence on the covariates and/or other spatial structures. Data is shown in Figure 1.

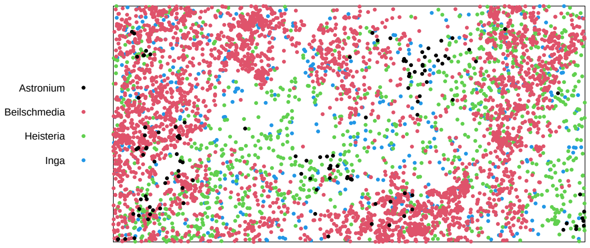


Fig. 1 Rainforest tree data

2 Model-based estimation of entropy components

Let X be a categorical response variable with I outcomes; consider a series of n realizations, which are independent given the distribution parameters, indexed by $u = 1, \ldots, n$, each presenting a value $x_u \in \{x_1, \ldots, x_I\}$. When $I = 2$ the well established class of Bayesian logistic regression models may be used, but difficulties arise when $I > 2$. The natural option is an extension to the multinomial logit model, and the model becomes more complicated. For each location u we have $n_{u1}, \ldots, n_{uI}$, that may be equal or greater than 0, and we use them as a starting point to estimate

$p_{u1}, \ldots, p_{uI}$, with $\sum_{i=1}^{I} p_{ui} = 1$; to ensure that all probabilities are proper, we model them as [4]

$$p_{ui} = \frac{g_{ui}(\theta)}{G_u(\theta)} \text{ with } g_{ui}(\theta) = \exp\{z_{ui}^{'}\beta + \phi_u\} \text{ and } G_u(\theta) = \sum_{i=1}^{I} g_{ui}(\theta) \tag{2}$$

where θ is shorthand notation for all model parameters, $\beta \sim N(0, 10^{-6})$ and the vector $z_u^{'}$ contains the covariates associated to the u-th unit. The vector of random effects is $\phi \sim IGMRF(0, \tau_\phi K)$, with $\tau_\phi \sim Gamma(a_\phi, b_\phi)$. The Intrinsic Gaussian Markov Random Field structure matrix K defines the type of dependence between the random effects, such as temporal or spatial correlation.

Fitting a multinomial model is very complicated in practice, due to the presence of $G_u(\beta)$. We exploit the multinomial-Poisson transform [1] which turns the multinomial likelihood into a Poisson likelihood with extra parameters. It is established that the transform returns the same estimates and asymptotic variances as the original distribution. When covariates and/or dependence structures are detected as relevant, p_{ui} varies across sites; therefore, we can compute one entropy value for each observation site. In such general case, the system's entropy may be plotted as a two dimensional surface over the observation area.

A simulation study (here not reported) assesses the validity of our approach: we generated data under scenarios of independence, binary covariates, continuous covariates, temporal effects and spatial effects. Running a comparison to a selection of existing estimators (frequentist, non parametric and Bayesian), our method reported by far the best RMSE in all departures from independence.

3 Results on rainforest tree data and concluding remarks

Our dataset is a marked point pattern with $n = 5639$ over a rectangular window of $500000m^2$, where the tree species is X and constitutes the point data categorical mark. The four species are x_1 =*Inga sapindoides*, x_2 =*Heisteria concinna*, x_3 = *Beilschmiedia pendula*, x_4 =*Astronium graveolens*, with $n_1 = 487$, $n_2 = 1141$, $n_3 = 3887$, $n_4 = 124$. The ML estimator for Shannon's entropy (1) is $\widehat{H}_{ML}(X) = 0.875$, and is 63% of the maximum possible entropy (log(4)). Such low value hints at an underlying structure in the data, but nothing can be said with the ML estimator, nor with any other correction proposed in the literature.

In order to evaluate the biodiversity of the system, we partition the area into 20×40 cells of size 25×25 metres. For each cell, we know the average values of the elevation and slope covariates and the counts of all species: our multinomial response variable is a table of 800×4 counts. We fit four versions of model (2) that differ as regards the covariate: one with covariate elevation, one with slope, one with both and one with no covariates. All models allow the coefficients to be species-specific, and all include a smooth RW2d spatial effect (i.e. a CAR model with a 12 neighbourhood). The four models are compared by Information Criteria and Likelihood ratio tests and the best model is the one with covariate elevation. Based on the model we obtain, for each species, a smooth surface estimating its probability of occurrence over the observation area, and derive the final surface for the estimated entropy, which is displayed in Figure 2, left panel. The right panel of the figure shows the ML estimator of the entropy computed for each cell, i.e. the standard existing competitor of our approach. Both entropies are in relative terms, i.e. they range $\in [0, 1]$ where 1 is the maximum heterogeneity of the system.

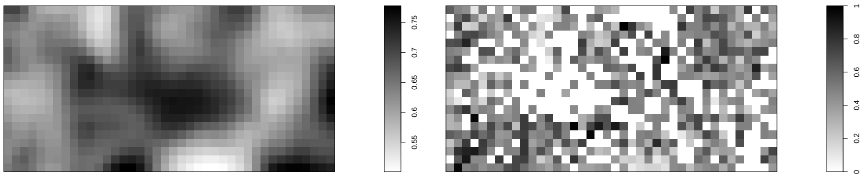


Fig. 2 Model-based entropy estimator (left) vs cell-specific ML estimator (right)

In biodiversity studies, the richness of species may depend on several factors, such as environmental covariates (altitude, soil slope, temperature...), spatial location, temporal structures. The main drawback of traditional Shannon's entropy is that it cannot account for data dependence. Even when the estimator is allowed to vary over a series of small locations such as in Figure 2, right panel, it looks like white noise and is not able to grasp a behaviour in the data; a further limit is that the entropy value is not sensitive to *which* species are present at a location, but only to their number and relative abundance. Our model-based approach allows to include information about which species are present/absent at each location, available covariates and smooth spatial effects: the result is an entropy surface that captures the biodiversity of the system. As regards our application, we can conclude that the biodiversity of the rainforest tree system depends on the covariate elevation, whose effect is particularly strong on the *Beilschmiedia pendula* species, based on the regression coefficient; entropy also shows an underlying spatial structure, and ranges from 50% to 80% of the maximum possible heterogeneity.

References

1. Stuart G. Baker. The multinomial-Poisson transformation. *The Statistician*, pages 495–504, 1994.
2. J. E. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36:192–236, 1974.
3. T. M. Cover and J. A. Thomas. *Elements of Information Theory. Second Edition*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.
4. N. A. C. Cressie. *Statistics for spatial data (rev. ed.)*. New York, Wiley, 2015.
5. L. Paninski. Estimation of entropy and mutual information. *Journal of Neural Computation*, 15:1191–1253, 2003.

Space-time modeling of Ocean Carbon in the North Atlantic

Elias Krainski, Mike Dowd, Claire Boteler, Eric Oliver, Doug Wallace

Abstract In this study, we aim to build maps of ocean carbon in the North Atlantic area using available ocean data from this region. To do this, we propose a statistical model that accounts for covariate effects as well as spatial and temporal variations, all the while working within the constraints imposed by the limitations of the available data. A spatio-temporal mixed model is proposed to estimate ocean carbon, and is then fitted for the observed data. To build high resolution synoptic monthly carbon maps over the full north Atlantic domain, we further make use of numerical ocean model outputs of temperature and salinity for prediction scenarios.

Keywords: Spatio-temporal, Bayesian, Ocean Carbon

1 Introduction

The ocean plays an important role in global carbon cycling and is a major sink for anthropogenic CO2 emissions [20]. The north Atlantic is a key area for carbon uptake due to wintertime deep convection which acts to transport atmospheric CO2 to the deep ocean [22].

Direct measurements of core variables in the ocean carbonate system from water samples obtained during research cruises have been compiled in the comprehensive, quality-controlled, global data archive GLODAPv2 [14]. However, these data are

Elias Krainski
KAUST, Saudi Arabia, e-mail: elias.krainski@kaust.edu.sa

Mike Dowd
Dalhousie University, Canada e-mail: michael.dowd@dal.ca

Claire Boteler
Dalhousie University, Canada e-mail: Claire.Boteler@dal.ca

Eric Oliver,
Dalhousie University, Canada e-mail: Eric.Oliver@dal.ca

Doug Wallace,
Dalhousie University, Canada e-mail: Douglas.Wallace@dal.ca

sparse in space and time and the opportunistic sampling results in a data distribution with a sampling bias towards summer months and along repeated transects.

We provide an approach for estimating surface ocean carbon in the northwest Atlantic for the region 40° – 60°N, 10° – 60°W using observations from the GLODAPv2 data archive. (Figure 1). To do this, we use direct ocean carbon measurements (total inorganic carbon, or TIC) by exploiting the robust relationship that TIC has with other well-sampled covariates, sea surface temperature and salinity. The approach taken is to fit a spatio-temporal mixed model to provide monthly estimates of TIC over the north Atlantic domain. We also seek to separate natural carbon variability from that due to anthropogenic sources.

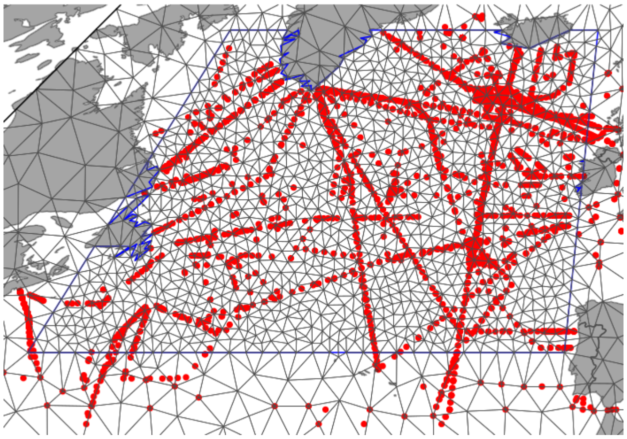


Fig. 1 The computational mesh used for the spatial-temporal analysis. Locations of observations used in the analysis are given as red dots.

2 Methods

In the mixed model, the surface TIC is the response variable y, while temperature (T) and salinity (S) observations are used as fixed effects covariates. The model takes the form

$$y(\mathbf{s}, t) = \beta_0 + \beta_1 T + \beta_2 S + \beta_3 t + u(\mathbf{s}, t) + e(\mathbf{s}, t) \qquad (1)$$

where $\mathbf{s}$ designates space (2-D geographic coordinates) and t designates time. β_0 is the intercept, and β_1 and β_2 are the regression coefficients for T, S respectively and β_3 captures a temporal trend. The random effects are given by $u(\mathbf{s}, t)$, and $e(\mathbf{s}, t)$ represents the error or noise term. The random effect is further split into $u(\mathbf{s}, t) = u_s(\mathbf{s})_a + u_t(t)$, where $u_s(\mathbf{s})_a$ is the annual spatial effect, and u_t the temporal effect.

The seasonal variation in u_t is modelled according to an auto-regressive order 2 process with coefficients chosen so it has a periodicity at the annual period. Note that since we have not enough data to estimate well u_s for every month, we consider u_s to be replicated for each year (annual spatial effect). Specifically, $u_s(\mathbf{s})_a \sim N(0, \mathbf{Q}^{-1})$ follows a Gaussian random field where $\mathbf{Q}^{-1} = \Sigma(\theta)$ is a Matérn covariance. The estimation of the posterior marginal distributions is carried out using the R-INLA software ([7]), with u_s being modeled using the Stochastic Partial Differential Equations (SPDE) approach [6].

In order to build comprehensive TIC fields, we make use of an ocean reanalysis product from Copernicus, [3], in order to provide full temperature and salinity fields over the domain. The mean TIC is estimated using the fitted model parameters as

$$\mu(s,t) = \hat{\beta}_0 + \hat{\beta}_S S(s,t) + \hat{\beta}_T T(s,t) + \hat{\gamma} t + \hat{v}(t) + \hat{u}(s)_a$$

where the domain is now defined over a spatial grid and at monthly intervals.

The variance of the estimated fields, $\mathrm{V}(\mu(s,t))$, is computed considering independent samples from the posterior joint marginal for all the model parameters. This is performed following three steps: (i) sample from the model parameters joint distribution, (ii) compute $\mu(s,t)$ for each joint sample, and (iii) compute the sample variance for each $\mu(s,t)$. We can then add this to the noise variance in order to have an estimate of the overall prediction variance: $\mathrm{V}(\mathrm{y}(s,t)) = \mathrm{V}(\mu(s,t)) + \hat{\sigma}_e^2$.

3 Preliminary results

We are interested in temporal trends in ocean carbon due to anthropogenic processes (atmospheric C02 increase) as well as spatial variations across our study domain. This is expressed as an anomaly of TIC about a space-time mean value. The spatial random effect for each year may be viewed as the spatial pattern of the annual TIC anomalies over the NW Atlantic that cannot be explained by temperature and salinity, and hence is due to the ocean's uptake of anthropogenic CO2 from the atmosphere.

The monthly surface fields for TIC over the 23 years duration of the study period (1993-2016) were estimated. Figure 2 shows the posterior mean of surface TIC for the year 1993. In the TIC mean field in Figure 2 indicates the seasonal progression of TIC moving from higher late winter values to lower values in the summer.

In this initial work, we have assumed a linear relationship for the relationship between the covariates T and S and the response variable TIC. For future work, this could be made more sophisticated for enhanced realism, e.g. we could use lagged relationships, include interaction terms, or even other environmental covariates. Another possible option is to consider variable regression coefficients. Finally, when considering the random effects another possible improvement is to work towards a space-time seasonal model.

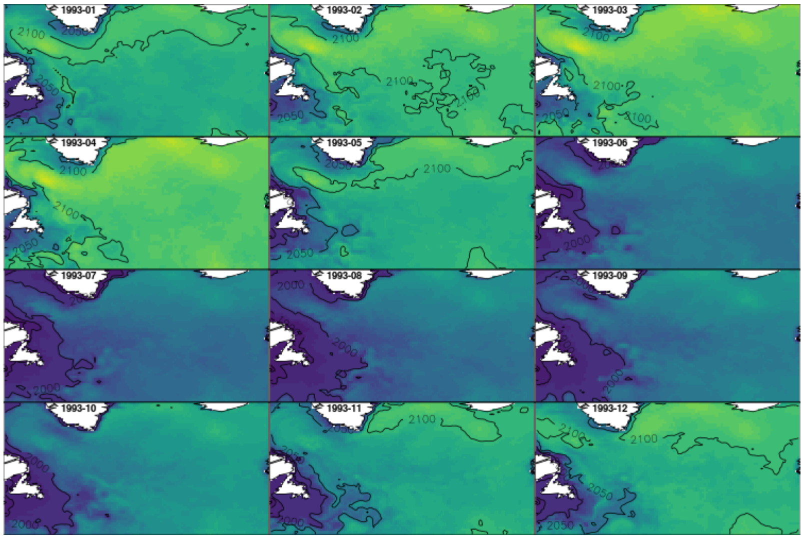


Fig. 2 TIC mean field for each month of the year 1993.

References

1. Drévillon, M., Bahurel, P., Bazin, D., et. al. Report on validation guidelines for ocean reanalyses of the Copernicus Marine Environment Monitoring Service MetOcean Mercator Océan, Ramonville St Agne, France, MetOcean Mercator Océan, Ramonville St Agne, France, 2018
2. Key, R., Olsen, A., van Heuven, S. et al GLOBAL OCEAN DATA ANALYSIS PROJECT,VERSION 2 (GLODAPv2) Carbon Dioxide Information Analysis Center, OAK RIDGE NATIONAL LABORATORY, Carbon Dioxide Information Analysis Center, OAK RIDGE NATIONAL LABORATORY, 2015
3. Fernandez, E. and Lellouche, J. PRODUCT USER MANUAL For the Global Ocean Physical Reanalysis product GLOBAL REANALYSIS PHY 001 030 COPERNICUS Marine Environment Monitoring Service, COPERNICUS Marine Environment Monitoring Service, 2018
4. Priestley, M. B. Spectral Analysis and Time Series AP, 1982
5. Rue, H. and Held, L. Gaussian Markov random fields: Theory and applications Boca Raton: Chapman and Hall, 2005
6. Lindgren, F., Rue, H. and Lindström, J. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (with discussion) JRSS-B, 2011, 73, 423-498
7. Rue, H., Martino, S. and Chopin, N. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion) Journal of the Royal Statistical Society, Series B, 2009, 71, 319-392
8. Simpson, D. P., Rue, H., Martins, T. G., Riebler, A. and Sørbye, S. H. Penalising model component complexity: A principled, practical approach to constructing priors Statistical Science, 2017, 32, 1-28
9. Krainski, E. T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F. and Rue, H. Advanced spatial modeling with stochastic partial differential equations using R and INLA CRC Press, 2018
10. Sørbye, S. and Rue, H. Scaling intrinsic Gaussian Markov random field priors in spatial modelling Spatial Statistics, 2014, 8, 39-51

11. Send, U., Fowler, G., Siddall, G. et al. A moored open-ocean profiling system for the upper ocean in extended self-contained deployments Journal of Atmospheric and Oceanic Technology, 2013, 30(7), 1555-1565
12. Longhurst., A. Seasonal cycles of pelagic production and consumption Progress in Oceanography, 1995, 36
13. Lauvset, S. K., Key, R. M., Olsen, A. et al. A new global interior ocean mapped climatology: the 1ĉirc x 1ĉirc GLODAP version 2 Earth System Science Data Discussions, 2016, XX, 1-30
14. Olsen, A., Key R. M. and van Heuven, S., Lauvset, S. K. et al. The Global Ocean Data Analysis Project version 2 (GLODAPv2) – an internally consistent data product for the world ocean Earth Syst. Sci. Data, 2016, 8, 297-323
15. Turk, D., Dowd, M., Lauvset, S. et al Can Empirical Algorithms Successfully Estimate Aragonite Saturation State in the Subpolar North Atlantic? Frontiers in Marine Science, 2017, 4, 385
16. Tanhua, T., Körtzinger, A., Friis, K., Waugh, D. and Wallace, D. An estimate of anthropogenic CO2 inventory from decadal changes in oceanic carbon content Proceedings of the National Academy of Sciences, 2007, 104, 3037-3042
17. Friis, K., Körtzinger, A., Pätsch, J. and Wallace, D. W. On the temporal increase of anthropogenic CO2 in the subpolar North Atlantic Deep Sea Research Part I: Oceanographic Research Papers, Elsevier, 2005, 52, 681-698
18. Clement, D. and Gruber, N. The eMLR (C*) method to determine decadal changes in the global ocean storage of anthropogenic CO2 Global Biogeochemical Cycles, Wiley Online Library, 2018, 32, 654-679
19. Thacker, W. C. Regression-based estimates of the rate of accumulation of anthropogenic CO2 in the ocean: A fresh look Marine Chemistry, Elsevier, 2012, 132, 44-55
20. Sabine, C. L., Feely, R. A., Gruber, N. et al. The oceanic sink for anthropogenic CO2 science, American Association for the Advancement of Science, 2004, 305, 367-371
21. Watson, A. J., Schuster, U., Bakker, D. C., et al Tracking the variable North Atlantic sink for atmospheric CO2 Science, American Association for the Advancement of Science, 2009, 326, 1391-1393
22. Ruhs, S., Biastoch, A., Böning, C. W.et al. Changing spatial pattern of deep convection in the subpolar North Atlantic Journal of Geophysical Research, 2021
23. McGarry, K., Siedlecki, S., Salisbury, J. and Alin, S. Multiple linear regression models for reconstructing and exploring processes controlling the carbonate system of the northeast US from basic hydrographic data Journal of Geophysical Research: Oceans, Wiley Online Library, 2021, 126, e2020JC016480
24. Leseurre, C., Monaco, C. L., Reverdin, G. et al. Ocean carbonate system variability in the North Atlantic Subpolar surface water (1993–2017) Biogeosciences, Copernicus GmbH, 2020, 17, 2553-2577
25. Cai, W.-J., Xu, Y.-Y., Feely, R. A. et al. Controls on surface water carbonate chemistry along North American ocean margins Nature communications, Nature Publishing Group, 2020, 11, 1-13
26. Macovei, V. A., Hartman, S. E., Schuster, U., et al Impact of physical and biological processes on temporal variations of the ocean carbon sink in the mid-latitude North Atlantic (2002–2016) Progress in Oceanography, Elsevier, 2020, 180, 102223
27. Wallace, D. W. Monitoring global ocean carbon inventories Ocean Observing System Development, 1995
28. Gregor, L. and Gruber, N. OceanSODA-ETHZ: A global gridded data set of the surface ocean carbonate system for seasonal to decadal studies of ocean acidification Earth System Science Data Discussions, Copernicus GmbH, 2020, 1-42
29. Keppler, L., Landschützer, P., Gruber, N., Lauvset, S. and Stemmler, I. Seasonal carbon dynamics in the Southern Ocean based on a neural network mapping of ship measurements EGU General Assembly Conference Abstracts, 2020, 18205
30. Landschuetzer, P., Gruber, N. and Bakker, D. C. Decadal variations and trends of the global ocean carbon sink Global Biogeochemical Cycles, Wiley Online Library, 2016, 30, 1396-1417

A Bayesian spatio-temporal model for integrating multiple sources of covid burden

Tullia Padellini, Brieuc Lehmann and Marta Blangiardo
on behalf of Turing – RSS SMML Lab

Abstract COVID-19 tests are commonly used to provide an estimate of the progress of the pandemic. However, community testing programs are typically biased due to the differential uptake across the population, for instance by symptoms or occupation. We propose an integrative framework to provide an estimate of the burden of COVID-19 at high spatio-temporal resolution. Anchored in a Bayesian hierarchical modeling perspective, our modular framework allows, if appropriate, to incorporate different sub-models for each data source and to include spatial and temporal dependencies as well as adjusting for covariate effect. At the same time the joint formulation means that uncertainty is propagated throughout the model. We apply our framework to integrate two different types of information on the daily number of cases at the lower tier local authority level in England: direct estimates coming from randomized surveys and testing programs and indirect estimates coming from hospital admission numbers. We show how this integrated framework is able to estimate metrics of disease progression such as incidence or prevalence, and how favorably our estimates compare to those based on unadjusted test counts only.

Keywords: COVID-19, spatio-temporal model, data integration, Bayesian hierarchical model

Tullia Padellini and Marta Blangiardo
MRC Centre for Environment and Health, Department of Epidemiology and Biostatistics, Imperial College London, Norfolk Place, London, W2 1PG, United Kingdom,
e-mail: t.padellini@imperial.ac.uk,m.blangiardo@imperial.ac.uk

Brieuc Lehmann
Department of Statistics, University of Oxford, 29 St Giles', Oxford, OX1 3LB, United Kingdom
e-mail: brieuc.lehmann@bdi.ox.ac.uk

The Alan Turing Institute – Royal Statistical Society Statistical Modelling and Machine Learning Laboratory e-mail: healthprogramme@turing.ac.uk.

Nonparametric local inference for functional data with manifold domain and temporal dependence

Niels Lundtorp Olsen, Alessia Pini, Simone Vantini

Abstract A topic which is becoming more and more popular in Functional Data Analysis is local inference, i.e., the continuous statistical testing of a null hypothesis along the functional domain. The principal issue in this topic is the infinite number of tested hypotheses, which can be seen as an extreme case of the multiple comparisons problem. During the talk we will define and discuss the notion of False Discovery Rate (FDR) in the setting of functional data defined on a manifold domain. We will then introduce a continuous version of the Benjamini-Hochberg procedure able to control the functional FDR over the functional domain, and finally describe its inferential properties in terms of control of the Type-I error probability and of consistency. The proposed general method will be then applied to the analysis of global satellite measurements of Earth's temperature with the aim of identifying the regions of the planet where temperature distribution has significantly changed in the last decades. In detail, yearly Earth's temperature maps are modelled as an instance of a functional concurrent auto-regressive process with the Earth's surface acting as the functional domain. Inference is performed fully nonparametrically relying on a joint use of a functional version of the Freedman and Lane permutation scheme, of the Phipson and Smyth method for the computation of the unadjusted p-values maps, and of the newly proposed approach to spatially adjust the p-value maps. The approach is shown to be very convenient in real space-time applications since it both allows a simple modeling of temporal dependence through standard time series tools and it does not require an explicit modeling of spatial dependence.

Niels Lundtorp Olsen
Department of Applied Mathematics and Computer Science, Technical University of Denmark, Copenhagen, Denmark. e-mail: nalo@dtu.dk

Alessia Pini
Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy. e-mail: alessia.pini@unicatt.it

Simone Vantini
MOX - Modelling and Scientific Computing Laboratory, Department of Mathematics, Politecnico di Milano, Milan, Italy. e-mail: simone.vantini@polimi.it

Keywords: Functional data, local inference, false discovery rate, Benjamini–Hochberg procedure, regression, time series.

References

1. Olsen, N.L., Pini, A., Vantini, S. (2021): False discovery rate for functional data, *TEST*, in press. Available at https://doi.org/10.1007/s11749-020-00751-x .

Part VI
Session 5: Small area estimation

Bias correction for underreported data in small area mapping.

Serena Arima, Loreto Gesualdo, Giuseppe Pasculli, Francesco Pesce, Silvia Polettini and Deni Aldo Procaccini.

Abstract Data quality is emerging as an essential characteristics of all data driven processes. The problem is particularly severe when health or vital statistics are concerned, with important consequences on government intervention policies and distribution of financial resources. In this paper, we deal with the underreporting issue with particular attention on its effects on the estimation of the prevalence of a phenomenon. We propose a non parametric compound Poisson model that allows for the estimation of underreporting probabilities. We will apply the proposed model to original data about the incidence of Chronic Kidney Disease (CKD) in Apulia.

Keywords: Underreporting probability disease mapping, non parametric model, MCMC.

1 Introduction

Data quality is an essential prerequisite for taking appropriate data driven decisions. As experienced in the last year, inaccurate data collection leads to inappropriate conclusions even when accurate and complex statistical methodologies had been performed. The problem is particularly severe when health or vital statistics are concerned, with important consequences on government intervention policies and distribution of financial resources. For example, the area-specific prevalence of a particular disease is the first criterion considered for distributing financial resources to hospitals and health devices. However, most often the number of individuals affected by the disease is inferred from patient registers, usually compiled upon registration by the health services, e.g. in hospitals when the medical examination occurs. A similar argument applies when it is of interest the geographical distribution

Serena Arima
Department of history, society and social sciences, University of Salento, Lecce, Italy, e-mail: serena.arima@unisalento.it

of different crimes or the regional distribution of forest fires. In all these cases, data are affected by underreporting and estimation of the prevalence of the phenomenon under study is substantially biased. The problem of underreporting is well known in the literature. [2] propose a bias correction method based on a compound Poisson model for count data that includes an area-specific reporting probability, whose uncertainty is accounted for in the model. In the proposal, areas are clustered according to their data quality. Since underreporting probabilities might reflect socio-economic, political and/or demographic characteristics of each region, we focus on modelling the underreporting rates in different areas. Extending the idea in [2], we consider the compound Poisson model. Following the approach described in [4], we propose to introduce covariates to define the clustering structure for the underreporting probability ϵ_i.
We will apply the proposed method to unpublished data coming from a retrospective study conducted between the 1st of January 2011 and the 31st of December 2013 for evaluating the incidence of Chronic Kidney Disease (CKD) in Apulia. To our knowledge, no prior studies have investigated the underreporting issue with respect to CKD prevalence in Italy and specifically in Apulia. But since the seminal paper by [3], it is clear that the availability of medical care tends to vary inversely with the need for it in the population served. Hence the need to further investigate possible underreporting in Apulia also considering that it is one of the most deprived regions in Italy. Moreover, Apulia is also characterized by very different geographical as well social conditions that we will consider in accounting for underreporting in CKD disease mapping.

2 Modelling underreported data

Consider a region consisting of m areas and denote by Y_i the observed counts in area i $(i = 1, ..., m)$. Let E_i denote a known offset representing the expected number of events in the $i-$th area. The observed counts are modeled as a compound Poisson model (CPM)

$$Y_i|\theta_i\epsilon_i \sim Poisson(E_i\theta_i\epsilon_i)$$

and the relative risks are related to a set of covariates $X_1, ..., X_p$:

$$log(\theta_i) = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi}$$

The parameter ϵ_i defines the reporting probability in the $i-$th area: low values of ϵ_i indicate areas whose observed counts are underreported. As in [2], we assume that areas can be clustered according to their data quality. Moreover, we assume that the reporting probabilities are equal for areas where the covariates related to the reporting process take similar values. In the aforementioned paper, this goal is achieved by using a-priori information that induces a clustering structure among the areas: in particular, they fix the number of cluster and model the probability of underreporting according to a-priori knowledge of data quality. Very informative priors are defined especially for the areas that are supposed to be characterized by the best data quality. In this work we consider an alternative approach: we specify a clustering structure for the underreporting probability ϵ_i following a non parametric approach based on a dependent Dirichlet process, that allows the aggregating property of the DP to depend on covariates. Although such a specification is more complex from a theoretical as well as computational point of view, it significantly increases the flexibility of the model since it does not require a-priori knowledge of the number of clusters and it defines the clustering structure in a complete nonparametric way. Indeed, the clustering is induced by introducing covariates in the stick breaking construction of the Dirichlet process.

2.1 The proposed model

Let $\epsilon^n = (\epsilon_1, ..., \epsilon_m)$ and $Z^n = (z_1, ..., z_m)$ denote, respectively, the entire vector of the underreporting probabilities and the covariate Z used as predictor for ϵ.

A simple nonparametric model can be defined by introducing a DP model on the underreporting probabilities:

$$\begin{aligned} p(y_i|\theta_i, \epsilon_i) &= \prod_{j=1}^{k_n} e^{-(E_i\theta_i\epsilon_j)} \frac{(E_i\theta_i\epsilon_j)^{y_i}}{y_i!} && (1) \\ log(\theta_i) &= \beta_i + \beta_1 X_{1i} + ... + \beta_p X_{pi} && (2) \\ \epsilon_i &\sim iid\ G && (3) \\ G &\sim DP(\alpha, G_0) && (4) \end{aligned}$$

For any measurable set B, the DP process has the well known stick-breaking representation [5]

$$G(B) = \sum_{j=1}^{\infty} w_j \delta_{\eta_j}(B)$$

where $\delta_{\eta_j}(\cdot)$ is the Dirac measure at η_j and $w_j = V_j \prod_{l<j}[1 - V_l]$ with $V_j|\alpha \overset{i.i.d.}{\sim} Beta(1, \alpha)$

[1] propose a modification of the well known stick-breaking representation of the DP in which the weights are made dependent on covariates, this is achieved replacing the Beta random variables by normally distributed random variables transformed through the normal cdf. The resulting measure is defined as the probit-stick breaking (PSB) process, see also []. As described by [4], [1] allow for dependence on covariates via the introduction of independent Gaussian processes indexed by the covariates as specified in the following formula:

$$G_z(\cdot) = \sum_{j=1}^{\infty} \left\{ \Phi(\eta_j(z)) \prod_{l<j} [1 - \Phi(\eta_l(z))] \right\} \delta_{\epsilon_j}(\cdot)$$

where $\eta_j(z) = z'\gamma$,

The proposed model will be applied to the data described in Section 1 and results will be presented during the conference.

References

1. Chung, Y. and Dunson, D. B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association* 104,1646–1660
2. de Oliverira, G.L., Argiento, R., Loschi, R.H. (2020) Bias correction in clustered underreported data, *Bayesian Analysis*, TBA, 1–32.
3. Hart,J.T. (1971) The inverse care law, *Lancet*, (7696): 405–412
4. Quintana, F., Mueller, P., Jara, A., MacEachern, S. (2021) The dependent Dirichlet process and related models. arXiv:2007.06129
5. Sethuraman, J. (1994). A constructive definition of Dirichlet prior. *Statistica Sinica*, 2, 639–650.

Bayesian small area models with log-transformed response

A. Gardini, E. Fabrizi and C. Trivisano

Abstract When unit-level small area models must be fitted with a positively skewed response, the log-normal linear mixed model represents a common choice. The feature of this model are explored within the hierarchical Bayes framework, proposing prior distributions for the variance components that guarantee the existence of the area means posterior moments, that are not finite under the most widespread prior settings. The theoretical findings are corroborated with a simulation study where the proposed estimators resulted competitive with the empirical Bayes estimates.
Keywords: Generalized Inverse Gaussian distribution, Log-normal distribution, Posterior moments, Unit-level models.

1 Introduction

This work pertains to unit-level small area models estimated with Hierarchical Bayes (HB) methods [4]. Small area estimation is surely widespread in economic statistics but examples of environmental applications are also available: land management [1], soil erosion monitoring [3], and pollutants exposure assessment [7], among the others.

When positive right-skewed responses are analyzed, specifying the popular random intercept model by Battese, Harter and Fuller [1] on the logarithmic transformation of the variable is a common choice (e.g., see [3, 2]). These papers target the log-normal mixed model from the Empirical Bayes perspective, whereas some issues

Aldo Gardini
Università di Bologna, Bologna, Italy, e-mail: aldo.gardini2@unibo.it

Enrico Fabrizi
Università Cattolica del Sacro Cuore, Italy

Carlo Trivisano
Università di Bologna, Bologna, Italy

specific of the fully Bayesian setting have not received attention yet. In particular, it is known that the posterior moments related to quantities in the original data scale under a log-normal model might be not defined [6, 5].

In this paper, after introducing the notation and the considered model (Section 2), we show the conditions for the existence of the posterior moments of quantities of interest in the small area context (Section 3). In Section 4, some details about the simulation study and the real data application are presented.

2 The model

Let's assume we target a finite population U of size N, partitioned into D sub-populations $U_1, ..., U_D$ whose sizes $N_1, ..., N_D$ are such that $N = \sum_{d=1}^{D} N_d$. A random sample s of size n is drawn according to a possibly complex design from U. The domain-specific sub-samples $s_1, ..., s_D$ with sizes $n_1, ..., n_D$, $n_d \geq 0$ with $\sum_{d=1}^{D} n_d = n$. The notation $\bar{s}_d$ represents the non-sampled part of U_d. The unit-level values of the response variable are denoted as y_{di}, $i = 1, \ldots, N_d$, $d = 1, \ldots, D$.

If we are interested in a positive, positively skewed variable for which it is reasonable the log-normality assumption, then it is possible to define $w_{di} = \log y_{di}$ and specify the random intercept model for the transformed variable. Under the additional assumption of non-informativeness of the sampling design, implying that the same model holds at both the population and the sample level, we can write:

$$w_{di} = \mathbf{x}_{di}^T \boldsymbol{\beta} + u_d + e_{di};$$

$$u_d \sim \mathcal{N}(0, \tau^2),\ e_{di} \sim \mathcal{N}(0, \sigma^2);\ d = 1, ..., D;\ i = 1, ..., N_D.$$

To estimate the model parameters, only the sample information can be used and we define the design matrix as $\mathbf{X}_s \in \mathbb{R}^{n \times p}$, where p is the number of observed covariates, and the vector containing the responses id $\mathbf{y}_s$. Since we focus on the HB framework, prior distributions on the model parameters must be specified. A flat improper prior is assumed for the coefficients vector β, whereas the priors for the variance components σ^2 and τ^2 will be discussed in Section 3.

In the small area framework, a target quantity to estimate is the area mean:

$$\bar{Y}_d = N_d^{-1} \sum_{i=1}^{N_d} y_{di}.$$

In the HB setting, the expectation of the posterior predictive distribution represents a natural prediction for the response variable of unsampled units, inducing the following estimator for the area mean:

$$\hat{\bar{y}}_d^{HB} = N_d^{-1} \left[\sum_{i \in s_d} y_{di} + \sum_{i \in \bar{s}_d} E\left[y_{di} | \mathbf{y}_s \right] \right]; \tag{1}$$

We remark that, to evaluate the previous estimator, all the covariates values must be available for the unsampled units. Alternatively, if only area-level covariates (i.e. $\mathbf{x}_{di} = \mathbf{x}_d, \ \forall i$) are included in the model, it is possible to estimate the area mean through the functional:

$$\hat{\bar{y}}_d^{HB'} = E\left[\mathbf{x}_d^T\boldsymbol{\beta} + u_d + e_{di}|u_d, \mathbf{y}_s\right] = \exp\left\{\mathbf{x}_d^T\boldsymbol{\beta} + u_d + \frac{\sigma^2}{2}\right\}, \tag{2}$$

that can be deduced using the properties of the log-normal distribution.

3 Main results

Both estimators (1) and (2) can be easily evaluated if MCMC posterior samples of the parameters are available. Nonetheless, it can be proved that the prior distribution of the variance components is crucial to assure the existence of the posterior moments of (1) and (2).

Focusing on the posterior predictive distribution $p(y_{di}|\mathbf{y}_s)$, it can be proved that its moment of order r is well defined if the prior of σ^2 has a density containing an exponential term $\exp\left\{-c\sigma^2\right\}$, where

$$c > \frac{r^2 + r^2\mathbf{x}_{di}^T\mathbf{X}_s^T\mathbf{X}_s\mathbf{x}_{di}}{2}.$$

Note that this condition must be fulfilled for any unsampled y_{di} to assure the existence of the moments of estimator (1). A similar condition can be derived for the posterior moments of functional (2):

$$c > \frac{r + r^2\mathbf{x}_{di}^T\mathbf{X}_s^T\mathbf{X}_s\mathbf{x}_{di}}{2}.$$

A possible prior for the variance components is the generalized inverse Gaussian (GIG) prior, in line with [6]. If $W \sim GIG(\lambda, \delta, \gamma)$, then the density is:

$$p(w) = \frac{w^{\lambda-1}}{2K_\lambda\left(\delta\gamma\right)}\exp\left\{-\frac{1}{2}\left(\frac{\delta^2}{w} + \gamma^2 w\right)\right\}, \quad w \in \mathbb{R}^+;$$

where $K_\nu\left(z\right)$ is the modified Bessel function of the second kind. The value of γ must be fixed in agreement with the previous existence conditions for the posterior moments. To preserve the prior balance between the variance component, the same prior is also used for τ^2. Eventually, we fix $\lambda = 1$ and $\delta \rightarrow 0$ to induce a uniform prior distribution on the intraclass correlation coefficient $\rho = \tau^2/\left(\tau^2 + \sigma^2\right)$.

4 Empirical findings summary

To investigate the frequentist properties of the Bayes estimators obtained under the proposed prior setting, an extended simulation study has been performed. The empirical Bayes estimator for the area mean is used as benchmark and the HB method with the proposed prior strategy outperformed it in terms of root means square error. In particular, the gain in efficiency is evident when the model is fitted on small samples.

Eventually, the discussed methodology has been applied to estimate the Radon concentration level in the Counties of Minnesota [7]. In this example, the response was recorded at the household level (unit), whereas the logarithm of the Uranium concentration, available at the County level, was considered as auxiliary variable. Hence, the estimator (2) is proposed.

References

1. Battese, G. E., Harter, R. M. and Fuller, W. A. An error-components model for prediction of county crop areas using survey and satellite data Journal of the American Statistical Association, Taylor & Francis Group, 1988, 83, 28-36
2. Molina, I., Martin, N. et. al. Empirical best prediction under a nested error model with log transformation Annals of Statistics, Institute of Mathematical Statistics, 2018, 46, 1961-1993
3. Berg, E. and Chandra, H. Small area prediction for a unit-level lognormal model Computational Statistics & Data Analysis, Elsevier, 2014, 78, 159-175
4. Rao, J. and Molina, I. Small Area Estimation John Wiley & Sons, 2015
5. Manandhar, B. and Nandram, B. Hierarchical Bayesian models for continuous and positively skewed data from small areas Communications in Statistics-Theory and Methods, Taylor & Francis, 2021, 50, 944-962
6. Fabrizi, E. and Trivisano, C. Bayesian estimation of log-normal means with finite quadratic expected loss Bayesian Analysis, International Society for Bayesian Analysis, 2012, 7, 975-996
7. Burris, K. C. and Hoff, P. D. Exact adaptive confidence intervals for small areas Journal of Survey Statistics and Methodology, Oxford University Press, 2020, 8, 206-230

Poisson mixed models for predicting number of fires

M. Boubeta, M.J. Lombardía, M. Marey-Pérez and D. Morales

Abstract Ecological studies consider wild fires to be one of the main causes of forest destruction. In the early years of the 21st century, the number of forest fires has increased in Europe, especially in the Mediterranean regions where the burned areas are very extensive. This problem particularly affects Galicia (north-west of Spain). Modeling wildfires in small areas can have a high error if temporal correlation structures are not taken into account. For this reason, four area-level Poisson mixed models with time effects are proposed. The first two models contain independent time effects, whereas the random effects of the other models are distributed according to an autoregressive process AR(1). A parametric bootstrap algorithm is given to measure the accuracy of the plug-in predictor of fire number under the temporal models. A significant prediction improvement is observed when using Poisson regression models with random time effects. Analysis of historical data from Galicia finds significant meteorological and socioeconomic variables explaining the number of forest fires by area and reveals the presence of a temporal correlation structure captured by the area-level Poisson mixed model with AR(1) time effects.

Keywords: bootstrap, empirical best predictor, forest fires, mean squared error, method of moments, plug-in predictor, time dependency.

Miguel Boubeta
Universidade da Coruña, Facultad de Informática, Spain, e-mail: miguel.boubeta@udc.es

María José Lombardía
Universidade da Coruña, CITIC, Spain, e-mail: maria.jose.lombardia@udc.es

Manuel Marey-Pérez
Universidad de Santiago de Compostela, Escuela Politécnica Superior de Ingeniería, Spain, e-mail: manuel.marey@usc.es

Domingo Morales
Universidad Miguel Hernández de Elche, Centro de Investigación Operativa, Spain, e-mail: d.morales@umh.es

1 Introduction

This communication summarizes some results of Boubeta et al. (2019) related to Poisson mixed models for predicting number of fires. The presentation contains some mathematical developments and an application to real data.

With regard to the statistical methodology, these authors considered four area-level Poisson mixed models with time random effects. The first, Model 1, assumes that the time effects are independent whereas the second, Model 2, assumes that they are AR(1)-correlated within the areas. Simplified versions of Model 1 and Model 2, Model 12 and Model 22, with only area-time random effects are also considered. A Newton-Raphson algorithm was implemented for calculating the method-of-moments estimators of the model parameters. Plug-in predictors of the Poisson mean parameters were proposed in both contexts: independence across time and AR(1) time correlation. The empirical best predictors for the area-time random effects under Poisson mixed models were provided. The mean squared error of predictors is estimated by a parametric bootstrap. The statistical methodology is adapted to obtain predictions for out-of-sample data. The method is of a general nature and is demonstrated against the Galician wild fire data sets.

The developed methodology is applied to forest fires data in Galicia, by month, in the period 2007-08. Further, an application to predict the number of fires in 2009 is given. As the meteorological variables change over time, different scenarios for predicting the number of fires are assumed. On the other hand, the auxiliary variables related to type of vegetation, human activities and land ownership do not vary much over time and depend only on the forest areas. The values of these variables in a near future are easy to establish and facilitates the profitability of deriving predictions. As an example of application, true meteorologic variables for 2009 were used.

Acknowledgements Supported by the Instituto Galego de Estatística, by the grants MTM2017-82724-R and PGC2018-096840-B-I00 of the MINECO and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015, Centro Singular de Investigación de Galicia ED431G/01, Unidades de investigación competitivas ED431C 2020/14), all of them through the ERDF.

References

1. Boubeta, M., Lombardía, M.J. , Marey-Pérez, M.F. , Morales D. (2019). Poisson mixed models for predicting number of fires. *International Journal of Wildland Fire* 28, 3, 237-253.

Part VII
Session 6: Advances in environmental epidemiology

Dynamics of SARS-CoV-2 infection in the Italian regions: a descriptive study based on compartmental models

Michela Baccini, Giulia Cereda and Cecilia Viscardi

Abstract Compartmental models are widely used to model epidemic dynamics. In their simplest form, they assume an Exponential distribution on the transition times between compartments. This assumption is quite unrealistic for the waiting time in the infection status because it assigns not negligible probability to very short and very long times, while a less dispersed distribution could be more appropriate. Additionally, assuming an Exponential waiting time implies that the probability of recovering or dying is constant over the infectivity time, thus independent of the time passed since infection onset. A known solution to this problem consists of assuming an Erlang distribution on the waiting times. From a practical point of view, this is achieved by including in the model several contiguous sub-compartments that the infected person must pass through before exiting the status of infection.
In this work, we propose an Erlang-modified SIRD model to describe the epidemic dynamics of the SARS-CoV-2 infection in the Italian regions. We specify for each region a deterministic SIRD model under the assumption that the time of infection has an Erlang(k, k/T) distribution. We model the time-varying pattern of the infection reproduction number $R_0(t)$ through a parametric cubic regression spline.
We focus on the second wave of the SARS-CoV-2 epidemic in Italy. Setting k to 5, the average time of infection T to 14 days, and the infection fatality rate to values reported by the literature, we calibrate the SIRD model on the notified daily deaths reported by the Protezione Civile, via the minimization of a loss function under a positive constraint on $R_0(t)$. We obtain bootstrap confidence intervals both for the

Michela Baccini
Dipartimento di Statistica, Informatica, Applicazioni (DISIA), Universitá di Firenze, Firenze, e-mail: michela.baccini@unifi.it

Giulia Cereda
Dipartimento di Statistica, Informatica, Applicazioni (DISIA), Universitá di Firenze, Firenze, e-mail: giulia.cereda@unifi.it

Cecilia Viscardi
Dipartimento di Statistica, Informatica, Applicazioni (DISIA), Universitá di Firenze, Firenze, e-mail: cecilia.viscardi@unifi.it

model parameters and for the compartments sizes.
The results from the Erlang-modified SIRD are compared with those obtained by using a standard SIRD model ($k = 1$). Finally, by varying the infection fatality rate within a range of plausible values, the model is used to get insights into the submerged portion of the SARS-CoV-2 epidemic in the Italian regions.

Keywords: SARS-CoV-2, COVID19, Compartimental models, SIRD, Erlang distribution, splines

A time varying parameter regression model to investigate the relationship between intensive care occupancies and confirmed COVID-19 deaths in European NUTS-2 Regions

Valentini P. and Ippoliti L. and Bucci A.

Abstract The impact of the COVID-19 pandemic varied significantly across different countries, with important consequences in terms of population health status and medical resources allocation. In this paper, to investigate the relation between the occupancy of intensive care units by COVID-19 patients and the number of confirmed deaths through time and space, we apply a Bayesian approach for multivariate time series. The model provides a flexible framework for the analysis of time series data, allowing the analysis of different features of the series, such as spatial correlations, time varying parameters and clustering. We evaluate the effect of intensive care units occupancy on the death counts recorded at regional level for several European countries in the period from March 2020 to April 2021.

Keywords: COVID-19, Dynamic linear models, Spatio-temporal series analysis

Valentini P.
Department of Economics, Università degli Studi “G. d’Annunzio” Chieti-Pescara, Italy, e-mail: pasquale.valentini@unich.it

Ippoliti L.
Department of Economics, Università degli Studi “G. d’Annunzio” Chieti-Pescara, Italy e-mail: luigi.ippoliti@unich.it

Bucci A.
Department of Economics, Università degli Studi “G. d’Annunzio” Chieti-Pescara, Italy e-mail: andrea.bucci@unich.it

1 Introduction

The coronavirus disease (COVID-19) pandemic has affected the health of the entire world population and has put health systems under stress in terms of needed staff and hospital beds in intensive care units (ICU). High-quality supportive care is the more effective solution for ensuring that people with COVID-19 who present severe symptoms have the best chance of surviving. In the early stages of the spread of COVID-19, hospitals have been experimenting capacity saturation from increased patient volume that may have resulted in an increased number of confirmed deaths by COVID-19 [1]. For example, when the pandemic spread at February 2020, Italy had a total of 5 000 ICU beds throughout the Italian territory. However, due to the asynchronous insurgence, in time and space, of the outbreaks, local hospitals and ICUs were rapidly saturated in the most hit Regions, such as Lombardy and Veneto [5].

Generally, ICUs are characterized by a low number of beds with high turnover [2] and most of the patients stay in intensive care only for few days at maximum. However, a study on ICUs occupancy by COVID-19 patients in Lombardy [3] reports a median length of stay of 9 days (6-13 [95% CI]). Longer stays imply higher occupancy rate, higher risk of collapse of the national health system and higher mortality rate in patients who do not manage to arrive in a ICU.

Lockdown measures, implemented at country level, and possibly other factors, allowed to alleviate the burden of the pandemic on the ICU system, with mostly of the patients with early symptoms being treated at home or in COVID-19 dedicated structures. However, further waves of the pandemic have put again under pressure the health system with higher geographical heterogeneity throughout the territory of several European countries. Despite the analysis of mortality during an outbreak may not be an easy feat, some studies [4] show that regions where fewer patients could be admitted into an ICU, due to scarce availability of beds, presented higher rates of mortality by COVID-19, and that this was especially true in the first phase of the pandemic when the countries were not ready in terms of personnel and ICU beds. For this reason, it is appealing to understand and investigate the existence of a direct relation of ICU beds occupancy and mortality rate, and it is of further interest to analyse this effect through time and space.

To this extent, we implement a time varying parameter model on weekly data on COVID-19 confirmed deaths and ICU occupancy from 92 regions of France, Germany, Italy, Spain, Switzerland and United Kingdom. We adopt a Bayesian approach with time varying parameters that foresees the use of a spatial effect in the error term of the model. The objective of such specification is to provide a flexible framework for estimation and interpretation of time variation in the effect of ICU occupancy on confirmed deaths by COVID. The dynamic linear model we propose also allows for the use of Dirichlet process priors in a mixture of Gaussian processes to identify latent common structure among the time series. Moreover, the introduction of a spatial effect can help entailing proximity effects in terms of spread of the pandemic and occupancy of beds in neighbour hospitals.

References

1. Bravata, D.M., Perkins, A.J., Myers, L.J., et al. (2021) Association of Intensive Care Unit Patient Load and Demand With Mortality Rates in US Department of Veterans Affairs Hospitals During the COVID-19 Pandemic. *JAMA Network Open*, 4(1), e2034266. doi:10.1001/jamanetworkopen.2020.34266
2. Farcomeni, A., Maruotti, A., Divino, F., Jona-Lasinio, G. and Lovison, G. (2021) An ensemble approach to short-term forecast of COVID-19 intensive care occupancy in Italian regions. *Biometrical Journal*, 63, 503–513.
3. Grasselli, G., Zangrillo, A., Zanella, A. et al. (2020) Baseline Characteristics and Outcomes of 1591 Patients Infected With SARS-CoV-2 Admitted to ICUs of the Lombardy Region, Italy. *JAMA*, 323(16), 1574–1581.
4. Immovilli, P., Morelli, N., Antonucci, E. et al. (2020) COVID-19 mortality and ICU admission: the Italian experience. *Critical Care*, 24, 228.
5. Olivieri, A., Palù, G. and Sebastiani, G. (2021) COVID-19 cumulative incidence, intensive care, and mortality in Italian regions compared to selected European countries. *International Journal of Infectious Diseases* 102, 363-368.

A geostatistical framework for combining spatiallyreferenced disease prevalence data from multiple diagnostics

Benjamin Amoah, Peter J. Diggle, Emanuele Giorgi

Abstract Multiple diagnostic tests are often used due to limited resources or because they provide complementary information on the epidemiology of a disease under investigation. Existing statistical methods to combine prevalence data from multiple diagnostics ignore the potential overdispersion induced by the spatial correlations in the data. To address this issue, we develop a geostatistical framework that allows for joint modelling of data from multiple diagnostics by considering two main classes of inferential problems: (a) to predict prevalence for a gold-standard diagnostic using low-cost and potentially biased alternative tests; (b) to carry out joint prediction of prevalence from multiple tests. We apply the proposed framework to two case studies: mapping Loa loa prevalence in Central and West Africa, using miscroscopy, and a questionnaire-based test called RAPLOA; mapping Plasmodium falciparum malaria prevalence in the highlands of Western Kenya using polymerase chain reaction and a rapid diagnostic test. We also develop a Monte Carlo procedure based on the variogram in order to identify parsimonious geostatistical models that are compatible with the data. Our study highlights (a) the importance of accounting for diagnostic-specific residual spatial variation and (b) the benefits accrued from joint geostatistical modelling so as to deliver more reliable and precise inferences on disease prevalence. The work presented is based on the publication of Amoah et al. [1].

Keywords: disease mapping, geostatistics, malaria, multiple diagnostic tests, neglected tropical disesaes, prevalence. . . .

Benjamin Amoah
Lancaster Medical School, Lancaster University, e-mail: b.amoah@lancaster.ac.uk

Peter J. Diggle
Lancaster Medical School, Lancaster University, e-mail: p.diggle@lancaster.ac.uk

Emanuele Giorgi
Lancaster Medical School, Lancaster University, e-mail: e.giorgi@lancaster.ac.uk

References

1. Amoah, B, Diggle, PJ, Giorgi, E. (2020) A geostatistical framework for combining spatially referenced disease prevalence data from multiple diagnostics. *Biometrics*, Biometrics, 2020; 158– 170.

Revisiting space-time disease mapping models

Maria Franco-Villoria, Massimo Ventrucci and Håvard Rue

Abstract The linear predictor of a spatio-temporal disease mapping model can be expressed as a sum of main and interaction terms, each of these specified by smooth functions of time, space and time and space respectively. We present the use of Penalized Complexity Priors (PC priors) for spatio-temporal smoothing models, where the interaction model shrinks to the model with only main effects.

Keywords: base model, Gaussian Markov random field, interaction, penalized complexity

1 Introduction

Disease mapping models [4] aim to estimate the relative risk of a particular disease over time and space. In doing so, smoothing models are used to borrow information across neighbouring areas and time points. In some practical applications, an interaction term is required to model the complex space-time relationship in the data. Assume data collected at space locations $s = 1, \dots, S$ and time $t = 1, \dots, T$, the linear predictor takes the following form:

$$\eta_{st} = \alpha + \underbrace{f(\texttt{space})_s + f(\texttt{time})_t}_{\texttt{marginal}} + \underbrace{f(\texttt{space}, \texttt{time})_{st}}_{\texttt{interaction}} \tag{1}$$

Maria Franco-Villoria
University of Modena e Reggio Emilia e-mail: maria.francovilloria@unimore.it

Massimo Ventrucci
University of Bologna, e-mail: massimo.ventrucci@unibo.it

Håvard Rue
King Abdullah University of Science and Technology e-mail: haavard.rue@kaust.edu.sa

We regard each smooth component as a process modelled by a Gaussian Markov Random field (GMRF) prior [7], conditional on one or more hyper-parameters responsible for the complexity introduced in the model. In particular, we focus on cases where marginal effects are modelled with intrinsic GMRFs and the interaction term is modelled with a Kronecker product GMRF (for examples of Kronecker product GMRFs see [6] sec 3.2).

2 Penalized Complexity (PC) Priors

In this section we briefly outline the four principles underpinning the construction of PC priors, namely: support to Occam's razor (parsimony), penalisation of model complexity, constant rate penalisation and user-defined scaling. For a more detailed presentation of these principles the reader is referred to [8].

Let f_1 denote the density of a model component w where τ is the parameter for which we need to specify a prior. The base model, corresponds to a fixed value of the parameter $\tau = \tau_0$ and is characterized by the density f_0.

1. The prior for τ should give proper shrinkage to τ_0 and decay with increasing complexity of f_1 in support of Occam's razor, ensuring parsimony.
2. The increased complexity of f_1 with respect to f_0 is measured as $d(\tau) = \sqrt{2KLD(f_1||f_0)}$, where $KLD(f_1||f_0)$ is the Kullback-Leibler divergence (KLD) [5].
3. The PC prior is defined as an exponential distribution on the distance, $\pi(d(\tau)) = \lambda \exp(-\lambda d(\tau))$, with rate $\lambda > 0$, ensuring constant rate penalization. The PC prior for τ follows by a change of variable transformation.
4. The user must select λ based on his prior knowledge on the parameter of interest (or an interpretable transformation of it, e.g. $T(\tau)$). This knowledge can be expressed in terms of a probability statement, e.g. $P(T(\tau) > U) = a$, where U is an upper bound for $T(\tau)$ and a is a (generally small) probability.

3 Space-time interaction models

Model (1) can be expressed using random effects modelled by intrinsic GMRFs. Consider the model

$$\eta_{st} = \alpha + \beta_s + \delta_t + \gamma_{st} \tag{2}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_S)^T$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_T)^T$ are vectors of spatial and temporal random effects, respectively. For the spatial effects we assume an intrinsic CAR model [1], while a second order random walk [2] is assumed for the time effects. Vector $\boldsymbol{\gamma} = (\gamma_{11}, \ldots, \gamma_{ST})$ contains interaction random effects modelled as a Kronecker GMRF with structure matrix given by the Kronecker product of the structure matrices of the spatial and time effects.

The random effects in Eq. (2) are GMRFs depending on precision (hyper-)parameters to which an (hyper-)prior has to be assigned. A popular approach consists in assigning conjugate Gamma priors to both the main effect and interaction precision terms. However, the individual precision parameters are difficult to interpret in terms of the total variance explained by the interaction and marginal terms. Instead, we re-parametrize model (2) introducing a *mixing parameter* ϕ, for which we consider a PC prior. This provides a simple way for setting the prior to the user who has prior information on the balance of the interaction and marginal components; to do so one could use the scaling approach described above requiring choice of U and a so that $P(\phi > U) = a$. This PC prior guarantees the interaction model shrinks towards a model with only main effects, thus avoiding overfitting.

4 Simulation study

We perform a small simulation study to compare the performance of the PC prior with a Gamma prior on the precision parameters of main effects and interaction. We set $S = T = 10$ and simulate 100 datasets from a Gaussian response corrupted by noise (standard deviation of the noise is equal to 1). We focus on three different scalings (choice of λ) of the PC prior, varying according the relative weight assigned to the interaction terms $\boldsymbol{\gamma}$: *PC tight* assigns very small weight to $\boldsymbol{\gamma}$ basically saying all the variance is explained by the additive model, while *PC moderate* and *PC flexible* are more liberal towards the interaction model. We compare against two specifications of the Gamma prior.

Figure 1 shows log mean square errors (MSE) in two scenarios, with interaction and no interaction (additive model). The behaviour of the PC prior is stable with respect to the choice of λ, whereas Gamma performance strongly depends on the choice on shape and rate. In general, PC priors show better behaviour compared to the Gamma family, except when an interaction is present in the data and the choice of λ is strongly unbalanced in favour of the model having no interaction.

5 Summary and Work in Progress

One major advantage of PC priors is that they prevent overfitting by construction, as they guarantee shrinkage towards the base model. Results from a preliminary simulation study suggest that PC priors give stable results in the context of an interaction model as in (2) for Gaussian data. We are currently working on extendign simulation results to Poisson and Binomial responses and deriving the PC prior for different types of interaction models as described in [3].

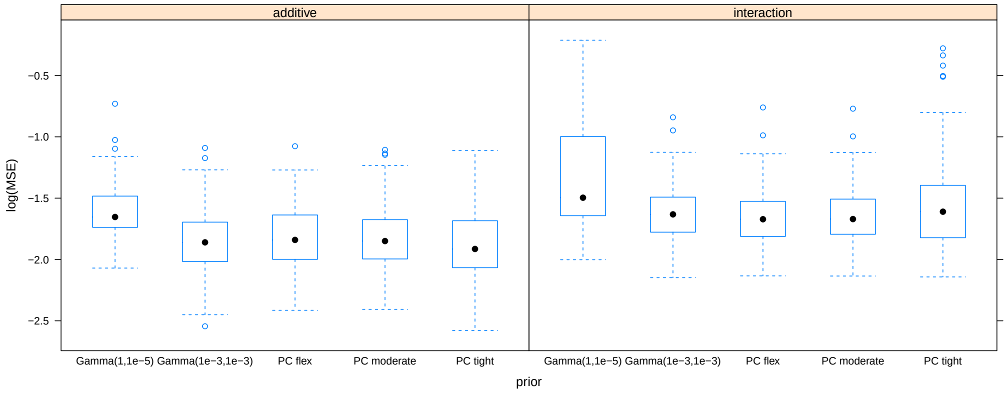


Fig. 1 Boxplots of log(MSE) for the simulation study

References

1. Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical society, series b*, 36(2):192-225.
2. Franco-Villoria, M., Ventrucci, M. and Rue, H. (2019). A unified view on Bayesian varying coefficient models. *Electronic Journal of Statistics*, 13(2):5334-5359.
3. Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19:2555-2567.
4. Knorr-Held, L. and Besag, J. (1998). Modelling risk from a disease in time and space. *Statist. Med.*, 17: 2045-2060.
5. Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79-86.
6. Lindgren, F. and Rue, H. (2015). Bayesian Spatial Modelling with R-INLA. *Journal of Statistical Software, Articles*, 63(19):1–25.
7. Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*. Chapman and Hall/CRC.
8. Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statist. Sci.*, 32(1):1-28.

Part VIII
Poster session

Scalable Gaussian Processes on Physically Constrained Domains

Jin Bora, Herring Amy H. and Dunson David B.

Abstract Providing safe drinking water is a globally imperative issue. In order to do so, it is necessary to properly understand the spatial distribution of aqueous pollutants in groundwater, since many water systems use groundwater resources. One of the distinctive characteristics of pollutants in groundwater is that the measurements are collected and meaningful only in a constrained domain, i.e., groundwater bodies with intrinsic geometry. Typical spatial Gaussian Process (GP) models ignore the unique geometry of the domain, which may lead to inappropriate smoothing over physical barriers. We focus on developing a scalable GP method that incorporates the constrained domain, motivated by modeling of spatial variability of pollutants in groundwater. One way to construct a scalable GP is via sparsity-inducing directed acyclic graphs (DAGs) that limit neighbors and impose conditional independence to the rest given the neighbors. A main contribution of this paper is the development of the Barrier Overlap-Removal Acyclic Directed Graph GP (BORA-GP) that constructs neighbors conforming to barriers. It removes an edge in a DAG if a linear path between two points overlaps the barriers, which enables characterization of dependence in constrained domains. We analyze water pollutant measurements in California collected through the Groundwater Ambient Monitoring and Assessment Program.

Jin Bora
Statistical Science, Duke University, USA e-mail: bora.jin@duke.edu

Herring Amy H., Dunson David B.
Statistical Science, Duke University, USA

Some properties and applications of local second-order characteristics for spatio-temporal point processes on networks

Nicoletta D'Angelo, Giada Adelfio and Jorge Mateu

Abstract Point processes on linear networks are increasingly being considered to analyse events occurring on particular network-based structures. In this work, we extend Local Indicators of Spatio-Temporal Association (LISTA) functions to the non-Euclidean space of linear networks, allowing to obtain information on how events relate to nearby events. In particular, we propose the local version of two inhomogeneous second-order statistics for spatio-temporal point processes on linear networks, the K- and the pair correlation functions. We also show that these LISTA functions are useful for diagnostics of models specified on the networks, and can be helpful to assess the goodness-of-fit of different spatio-temporal models fitted to point patterns occurring on linear networks. Our methods do not rely on any particular model assumption on the data, and thus they can be applied for whatever is the underlying model of the process. Furthermore, we use the LISTA functions to build a test for the identification of differences in the local spatio-temporal structure of two point patterns: a pattern of interest and a background one, both occurring on the same linear network. We finally present a real data analysis of traffic accidents in Medellin (Colombia).

Key words: Linear networks, Local properties, Residual analysis, Second-order characteristics, Spatio-temporal point patterns

Nicoletta D'Angelo
Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy e-mail: nicoletta.dangelo@unipa.it

Giada Adelfio
Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy e-mail: giada.adelfio@unipa.it

Jorge Mateu
Department of Mathematics, University Jaume I, Castellon, Spain, e-mail: mateu@mat.uji.es

A water level prediction using ARMA and ARIMA models: A case study of the river Niger

Umar Nura and Alison Gray

Abstract Flooding is one of the most frequently occurring natural hazards globally. Studies on how to reduce or prevent this extreme event have used various approaches. Recent studies have proposed water level forecasting as an important technique which enables effective water resources management, helping to prevent flood disasters. In this work, as well as descriptive analysis, we use time series modelling to predict monthly water level discharges using data for the period 2010-2016 collected from Nigeria Hydrological Services Agency for three different water stations along the river Niger, namely Baro, Jebba and Kainji water stations, using Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA) models. The performance of these time series models was tested using three different performance measures, including mean absolute error, root mean square error and Nash-Sutcliffe efficiency, to find an appropriate model which will be utilised to predict the water level discharges from the three water stations. This will provide information to the populace and water management authorities of what to expect in future, so as to mitigate the impact of flooding when it occurs. To the best of our knowledge, this is a novel application of this approach on water level discharge data from Nigeria.

Key words: ARMA; ARIMA; climate change; environment; flooding; modelling; water level discharges

Umar Nura
Department of Mathematics and Statistics, University of Strathclyde, Glasgow, United Kingdom & Department of Mathematics and Statistics, Umaru Musa Yar'adua University. Katsina, Nigeria
e-mail: nura.umar@strath.ac.uk

Alison Gray
Department of Mathematics and Statistics, University of Strathclyde, Glasgow, United Kingdom

Penalized functional clustering with environmental applications

Nicola Pronello, Sara Fontanella, Rosaria Ignaccolo, Luigi Ippoliti

Abstract With the advance of modern technology, and with data being recorded continuously, functional data analysis has gained a lot of popularity in recent years. Working in a model-based framework, we develop a flexible clustering technique embedding dimensionality reduction schemes for functional data. The proposed procedure results in an integrated modelling approach where shrinkage techniques (L1 type penalty) are applied to enable sparse solutions in both the means and the covariance matrices of the mixture components, while preserving the best clustering structure. This leads to an entirely data-driven methodology suitable for simultaneous dimensionality reduction and clustering. In order to compare cluster structures resulting from different model specifications, as well as choosing the number of components, measures of cluster validity assessment and suitable information criteria are considered for model validation. A comparison of the proposal with other existing clustering algorithms is carried out both in a Monte Carlo study and by empirical analysis of real-world environmental data.

Nicola Pronello
Università degli Studi Gabriele d'Annunzio e-mail: nicola.pronello@studenti.unich.it

Sara Fontanella
Imperial College London

Rosaria Ignaccolo
Università degli Studi di Torino

Luigi Ippoliti
Università degli Studi Gabriele d'Annunzio

Predicting Water Temperature Profiles in the Middle Fork River: a Geostatistical Approach for Functional Data Over a Stream Network Domain

Chiara Barbi, Alessandra Menafoglio and Piercesare Secchi

Abstract The analysis of data collected on river networks is becoming increasingly frequent in ecological and environmental settings. The complex reticular nature of the domain requires to use geostatistical methods based on the specific concept of Stream Distance, which captures the spatial configuration of points in a river, the network branching and connectivity. Moving Average models based on Stream distance have been proposed in literature to describe the structure of spatial dependence in a network. However, these models remain mostly restricted to the context of scalar data. This work aims to extend geostatistical methods for spatially distributed data on a river network to the functional case. The contribution will deal specifically with this issue, first proposing a strategy for variographic analysis and estimation of the spatial covariance structure, then adapting the functional Kriging predictor to the Stream distance scenario. In particular, empirical semivariograms able to deal with Stream Distance and functional data are proposed. To illustrate the procedure, a comprehensive geo-statistical analysis on real data is conducted, aiming to (spatially) predict the summer water temperature profiles in the Middle Fork River, Idaho, USA.

Chiara Barbi
Politecnico di Milano e-mail: chiara.barbi@mail.polimi.it

Sara Menafoglio
Politecnico di Milano

Piercesare Secchi
Politecnico di Milano

Model-based clustering for monitoring cetaceans population dynamics

Greta Panunzi, Gianmarco Caruso, Marco Mingione, Pierfrancesco Alaimo di Loro, Stefano Moro, Edoardo Bompiani, Caterina Lanfredi, Daniela Silvia Pace, Luca Tardella and Giovanna Jona Lasinio

Abstract We introduce a Bayesian multivariate framework to investigate the site-fidelity patterns and estimate the population size of bottlenose dolphins (Tursiops truncatus) at the Tiber River estuary (central Mediterranean, Tyrrhenian Sea, Rome, Italy) between 2017 and 2020. In order to compare the results obtained through a distance-based clustering (Pace et al., 2021), a model-based clustering is performed using the same site-fidelity metrics: in particular, a multivariate finite mixture model is assumed for the vector of metrics (McLachlan et al., 2019). The proposed approach consists of two steps. We start with a Bayesian model-based classification of individuals in three different clusters labeled resident, part-time and resident using 347 unique individuals identified. Each individual is allocated to the group with the greatest estimated posterior probability. Finally, for each group, we estimate the corresponding population size via a capture-recapture analysis based on the Jolly-Seber model (Schwarz and Arnason, 1996): this kind of model allows to take into account the apparent survival probability of the animal in the population along with the capture probability. The results are compared to those obtained by the distance-based classification provided by Pace et al. (2021).

Key words: Jolly-Seber model, capture-recapture analysis, wildlife population, finite mixture models

Greta Panunzi
Department of Statistical Sciences, Sapienza University of Rome e-mail: greta.panunzi@gmail.com

Gianmarco Caruso, Marco Mingione, Pierfrancesco Alaimo di Loro, Edoardo Bompiani, Caterina Lanfredi, Luca Tardella and Giovanna Jona Lasinio
Department of Statistical Sciences, Sapienza University of Rome

Stefano Moro, Daniela Silvia Pace
Department of Environmental Biology, Sapienza University of Rome, Italy

References

1. Giordani P, Ferraro MB, and Martella, F (2020) Hierarchical clustering. Pages 9–73 of: An Introduction to Clustering with R. Springer.
2. McLachlan GJ, LEE SX and Rathnayake SI (2019) Finite mixture models. Annual review of statistics and its application, 6, 355–378.
3. Pace DS, Di Marco C, Giacomini G, Ferri S, Silvestri M, Papale E, Casoli E, Ventura D, Mingione M, Alaimo Di Iorio P, et al (2021) Capitoline Dolphins: Residency Patterns and Abundance Estimate of Tursiops truncatus at the Tiber River Estuary (Mediterranean Sea). Biology, 10(4), 275.
4. Schwarz CJ and Arnason AN (1996) A general methodology for the analysis of capture-recapture experiments in open populations. Biometrics, 860–873.

Joint Cetacean Database and Mapping (JCDM) in Italian waters: a tool for knowledge and conservation

Pierfrancesco Alaimo Di Loro, Edoardo Bompiani, Gianmarco Caruso, Giovanna Jona Lasinio, Caterina Lanfredi, Marco Mingione, Stefano Moro, Daniela Silvia Pace, Greta Panunzi, Luca Tardella

Abstract Today, besides carrying out new research, it is pivotal for science to make data and results directly accessible to the public and, in particular, to decision-makers for conservation purposes. In this light, statistical methods, data visualization, and data sharing tools make it easy to visualize and share raw data (databases) as well as data elaborations. Cetaceans are strictly protected in many Mediterranean Sea sectors, though our understanding of species abundance, spatial distributions and habitat use is still deficient. These represent basic information to understand cetaceans' ecology better and identify zones overlapping with human activities. However, these mammals are elusive and highly mobile, able to migrate for thousands of kilometers. Hence, targeted surveys are expensive, and occurrence data are sparse in space and time. We aim at building a comprehensive database of cetaceans' occurrences in the Italian waters to estimate the distribution patterns and abundance of cetaceans, with novel data modeling approaches. The JCDM database will archive presence-only records provided by different sources (existing data geoportals, social media, and scientific surveys) to develop, in the R framework, new analytical mapping and informative tools to reduce uncertainty in the modeling outputs. The JCDM will be nestled in a standardized framework controlling for heterogeneous observation efforts yielded by data sources. The statistical methodology will view these sources of variability and the specific features of presence-only data. Elaborations will be presented in a web app in the form of interactive descriptive maps and tables. Users could navigate the map and simulate different predictive scenarios of occurrence probability based on environmental predictors. Elaborated data will be available to

Giovanna Jona Lasinio
Department of Statistical Sciences, Sapienza University of Rome e-mail: giovanna.jonalasinio@uniroma1.it

Pierfrancesco Alaimo Di Loro, Edoardo Bompiani, Gianmarco Caruso, Caterina Lanfredi, Marco Mingione, Greta Panunzi, Luca Tardella
Department of Statistical Sciences, Sapienza University of Rome

Stefano Moro, Daniela Silvia Pace
Department of Environmental Biology, Sapienza University of Rome, Italy

the scientific community for further investigations through direct downloads from the app. The latter will be regulated according to different access levels defined according to legal regulation on data ownership. The JCDM modeling outputs/outcomes will represent an easily readable tool for managers and decision-makers to plan conservation actions.

Markov modulated Poisson processes for stochastic modelling of background seismicity

Elisa Varini, Antonella Peresan, Amel Benali

Abstract Declustering a seismic catalog is a relevant preliminary step in many applications, such as earthquake forecasting and seismic hazard assessment. Declustering aims at partitioning an earthquake catalog into background seismicity, which is supposed to reflect the steady tectonic loading, and clustered seismicity, which is formed by dependent events. We decluster two Italian earthquake catalogs by applying two different data-driven declustering algorithms, namely the nearest-neighbor method (Zaliapin and Ben-Zion, J. Geophys. Res., 2013) and the stochastic declustering method (Zhuang et al., J. Geophys. Res., 2004). We verify the general assumption according to which the temporal sequence of background seismicity is suitably modelled by the stationary Poisson model. Whenever the Poissonian hypothesis is rejected, we get evidence of certain heterogeneity in the background sequence, which leads us to rule out the Poisson process for background seismicity modeling in favor of the Markov Modulated Poisson Process (MMPP), which allows the Poisson seismicity rate to change over time according to a finite (unknown) number of Markovian states (Benali et al., Stoch. Environ. Res. Risk Assess., 2020). The MMPP model turns out suitable for identifying and quantifying heterogeneities in background seismicity, as well as for comparing them against the two considered declustering algorithms.

Elisa Varini
CNR-IMATI Milano, Italy e-mail: elisa.varini@cnr.it

Antonella Pesaran
OGS-CRS Udine, Italy

Amel Benali
USTHB Département de Probabilités Et Statistiques, Algiers

Spherical autoregressive change-point detection with applications

Federica Spoto, Alessia Caponera and Pierpaolo Brutti

Abstract Spatio-temporal processes arise very naturally in a number of different applied fields, like Cosmology, Astrophysics, Geophysics, Climate and Atmospheric Science. In most of these areas, the detection of structural breaks or regime shifts in the data stream is key. To this end, in the present work, we aim at generalizing the recently introduced SPHAR(p) process by allowing for temporal changes in its functional parameters and variability structure. Our approach, which intrinsically integrates the spatial and temporal dimensions, could give multiscale insights into both the global and local behavior of changes, and its performance will be tested on a real dataset of global surface temperature anomalies.

Federica Spoto
Sapienza University of Rome e-mail: federica.spoto@uniroma1.it

Alessia Caponera
Ecole Polytechnique Federale de Lausanne

Pierpaolo Brutti
Sapienza University of Rome

Using high-resolution density models to predict white sharks' hot spots in the Mediterranean Sea

Stefano Moro, Mattia Palma, Giovanna Jona-Lasinio, Francesco Colloca, Chiara Gambardella and Francesco Ferretti

Abstract The white shark is an apex predator widely distributed between the sub-polar and tropical oceanic regions in both hemispheres (Compagno, 2001). In Mediterranean Sea it is a rare but persistent inhabitant. Here, its ecology is largely unknown and we still lack information about its migratory behaviors and habitat use, aspects that are critical for conservation and management. Moreover, in the last 50 years, Mediterranean white sharks suffered a strong decline in abundance, linked to a contraction of their spatial distribution toward the central sectors of the Mediterranean Sea, such as the Strait of Sicily (Moro et al., 2019). In this study, we estimated monthly high-resolution abundance density surfaces within the Strait of Sicily, starting from opportunistic data. To account for spatial dependence, data were modeled using Point Process in its log-Gaussian Cox Processes (LGCP) variant (Renner et al., 2015). A thinning procedure was also implemented to handle and standardize the different observation processes characterizing the different data sources (i.e. sightings, catches) (Martino et al., 2021). These abundance maps will inform, for the first time, a scientific expedition aiming at collecting unprecedented high-quality ecological data on this species in the Strait of Sicily.

Stefano Moro
Department of Environmental Biology, Sapienza University of Rome, Rome, Italye-mail: stefano.moro@uniroma1.it

Mattia Palma
Department of Environmental Biology, Sapienza University of Rome, Rome, Italy

Giovanna Jona-Lasinio
Department of Statistical Sciences, Sapienza University of Rome, Italy

Francesco Colloca
Stazione Zoologica Anton Dohrn, Naples, ITALY

Chiara Gambardella
Department of Life and Environmental Sciences, Polytechnic University of Marche, Ancona Italy

Francesco Ferretti
Department of Fish and Wildlife Conservation, Virginia Tech, Blacksburg VA, USA

References

1. Compagno L.J., 2001. Sharks of the world: an annotated and illustrated catalogue of shark species known to date, vol 2. Bullhead, mackerel and carpet sharks (Heterodontiformes, Lamniformes and Orectolobiformes). FAO Species Catalogue of Fishery Purposes, 1-269.
2. Martino, S., Pace, D. S., Moro, S., Casoli, E., Ventura, D., Frachea, A., ... & Lasinio, G. J. (2021). Integration of presence-only data from several sources. A case study on dolphins' spatial distribution. arXiv preprint arXiv:2103.16125.
3. Moro, S., Jona-Lasinio, G., Block, B., Micheli, F., De Leo, G., Serena, F., ... & Ferretti, F. (2020). Abundance and distribution of the white shark in the Mediterranean Sea. Fish and Fisheries, 21(2), 338-349.
4. Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., ... & Warton, D. I. (2015). Point process models for presence-only analysis. Methods in Ecology and Evolution, 6(4), 366-379.