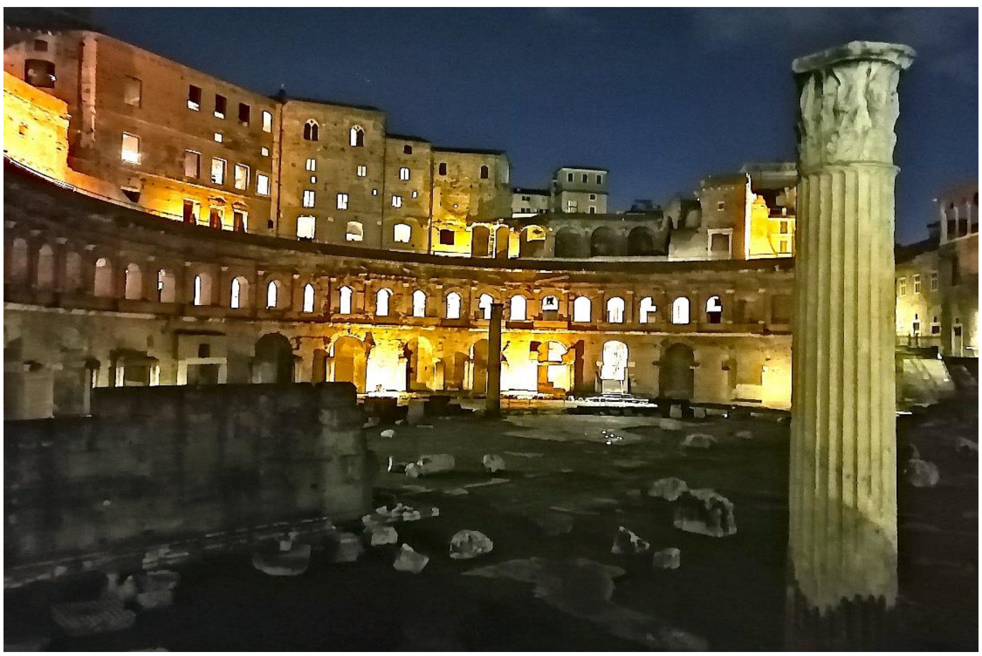
Proceedings of the GRASPA 2021 Conference
Rome, 07-09 June 2021



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

Contents

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

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

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:

$$p(y_i|\theta_i, \epsilon_i) = \prod_{j=1}^{k_n} e^{-(E_i\theta_i\epsilon_j)} \frac{(E_i\theta_i\epsilon_j)^{y_i}}{y_i!} \tag{1}$$

$$log(\theta_i) = \beta_i + \beta_1 X_{1i} + ... + \beta_p X_{pi} \tag{2}$$

$$\epsilon_i \sim iid\ G \tag{3}$$

$$G \sim DP(\alpha, G_0) \tag{4}$$

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.