Proceedings of the GRASPA 2019 Conference
Pescara, 15-16 July 2019

Edited by: Michela Cameletti, Luigi Ippoliti, Alessio Pollice

Università degli Studi di Bergamo

2019

Proceedings of the GRASPA 2019 Conference
Pescara, 15-16 July 2019

Edited by: Michela Cameletti, Luigi Ippoliti, Alessio Pollice. –

Bergamo : Università degli Studi di Bergamo, 2019.
(GRASPA Working Papers)

ISBN: 978-88-97413-34-9
ISSN: 2037-7738

https://aisberg.unibg.it/handle/10446/142407

Table of Contents

Non-Gaussian spatial and spatio-temporal processes

Alexandra Schmidt

Department of Epidemiology, Biostatistics and Occupational Health, McGill University;E-mail: alexandra.schmidt@mcgill.ca

Abstract. In the analysis of most spatial and spatio-temporal processes in environmental studies, observations present skewed distributions, with a heavy right or left tail. Usually, a single transformation of the data is used to approximate normality, and stationary Gaussian processes are assumed to model the transformed data. Spatial interpolation and/or temporal prediction are routinely performed by transforming the predictions back to the original scale. The choice of a distribution for the data is key for spatial interpolation and temporal prediction. In this talk, I will start discussing the advantages and disadvantages of using a single transformation to model such processes. Then I will discuss some recent advances in the modeling of non-Gaussian spatial and spatio-temporal processes.

Trajectory Functional Boxplots

Marc Genton

[1] *King Abdullah University of Science and Technology (KAUST), Saudi Arabia;E-mail: Marc.Genton@KAUST.EDU.SA*

Abstract. With the development of data-monitoring techniques in various fields of science, multivariate functional data are often observed. Consequently, an increasing number of methods have appeared to extend the general summary statistics of multivariate functional data. However, trajectory functional data, as an important sub-type, have not been studied very well. We proposes two informative exploratory tools, the trajectory functional boxplot, and the modified simplicial band depth (MSBD) versus Wiggliness of Directional Outlyingness (WO) plot, to visualize the centrality of trajectory functional data. The newly defined WO index effectively measures the shape variation of curves and hence serves as a detector for shape outliers; additionally, MSBD provides a center-outward ranking result and works as a detector for magnitude outliers. Using the two measures, the functional boxplot of the trajectory reveals center-outward patterns and potential outliers using the raw curves, whereas the MSBD-WO plot illustrates such patterns and outliers in a space spanned by MSBD and WO. The proposed methods are validated on hurricane path data and migration trace data recorded from two types of birds.

The space environment for satellites orbiting the earth.

John Kent

Department of Statistics, University of Leeds;E-mail: j.t.kent@leeds.ac.uk

Abstract. There are currently about 2000 operational satellites orbiting the earth. The subject of space situational awareness deals with various hazards to these satellites ranging space weather to space debris. There are estimated to be over 30000 pieces of space debris and inactive satellites in orbit bigger than a grapefruit, which can be observed from earth. In this talk I will describe some recent work funded by the US Air Force to develop fast and accurate improved statistical methods to predict the path of the debris so that it can be avoided by active spacecraft. The methodology uses ideas from Kalman filtering, directional statistics and multivariate analysis.

Air quality numerical models. Don't leave them alone.

G. Bonafé[1,*], I. Gallai[1], A. C. Goglio[1,2], D. Giaiotti[1], E. Gianesini[1], F. Montanari[1], A. Petrini[1]

[1] *ARPA-FVG, Agenzia Regionale per la Protezione dell'Ambiente del Friuli Venezia Giulia, via Cairoli 14 - 33057 Palmanova (Italy); giovanni.bonafe@arpa.fvg.it, irene.gallai@arpa.fvg.it, dario.giaiotti@arpa.fvg.it, elena.gianesini@arpa.fvg.it, francesco.montanari@arpa.fvg.it, alessandra.petrini@arpa.fvg.it,*
[2] *CMCC, Centro Euro-Mediterraneo sui Cambiamenti Climatici, viale C. Berti Pichat 6/2 - 40127 Bologna (Italy); annachiara.goglio@gmail.com*
[*] *Corresponding author*

Keywords. *Air Pollution; Universal Kriging; Kalman Filter; Air Quality Numerical Models; Scenario Analysis.*

Air quality management needs different approaches and tools, depending on spatio-temporal scales and on the specific aim. Usually, air quality numerical models (AQMs) are the basis for such tools. AQMs simulate emission, dispersion, transport, chemical and microphysical transformations, wet and dry depositions of gaseous and aerosol species.

Some use cases are described and discussed, covering typical problems we must face in the regional environmental agencies:

- air quality assessment for the past months or years;
- short-term air quality forecast for the next three to five days;
- scenario analysis to evaluate the benefit of mitigation actions in the next five-ten years.

For each use case, a deterministic AQM alone is not enough to give a satisfactory answer, therefore some post-processing is needed, and statistical methods may help.

In order to assess air quality at different spatial scales, focusing on a mixed residential-industrial domain, the output of two different AQMs (one suitable for the regional, the other for the urban spatial scale) and the data observed by the monitoring networks are combined with an approach based on the universal kriging technique.

For short term forecast purpose, the output of a regional scale AQM is corrected with a Kalman filter based on the observed data of the last days. This approach improves the general performance of the concentrations forecast, but for PM10 daily exceedance forecast, sometimes a simple seasonal multiplicative correction can work slightly better.

With the help of an AQM, the effects of a set of emission reduction actions and policies can be evaluated, simulating a "what if" scenario. However, since AQMs and their inputs (meteorology, emissions) are affected by errors, their output is biased as well. Results of some scenario analysis for the Po Valley are presented. Different approaches for scenario unbiasing are discussed.

Acknowledgments. The maintenance and development of the ARPA-FVG air quality modelling suite are partially funded by project EU LIFE-PREPAIR (LIFE15 IPE/IT/000013).

References

[1] Calori, G., Finardi, S., Nanni, A., Radice, P., Riccardo, S., Bertello, A., and Pavone, F.(2008). Long-term air quality assessment: modeling sources contribution and scenarios in Ivrea and Torino areas. *Environmental Modeling Assessment*, **13**(3): 329–335

[2] Delle Monache, L., Nipen, T., Deng, X., Zhou, Y., and Stull, R. (2006). Ozone ensemble forecasts: 2. A Kalman filter predictor bias correction. *Journal of Geophysical Research: Atmospheres*, **111**(D5).

[3] Gladich, I., Gallai, I., Giaiotti, D., Mordacchini, G., Palazzo, A., and Stel, F. (2008). Mesoscale heat waves induced by orography. *Advances in Science and Research*, **2**(1): 139–143.

Integrated Environmental Health Impact Assessment: example of applications in Italy

A. Ranzi[1*]

[1] *Centre for Environmental Health and Prevention, Regional Agency for Prevention, Environment and Energy of Emilia-Romagna, Modena (Italy);aranzi@arpae.it*

[*] *Corresponding author*

Abstract. *A health risk assessment is the scientific evaluation of potential adverse health effects resulting from human exposure to a particular hazard. Traditional methods of risk assessment have provided good service in support of policies, mainly in relation to standard setting and regulation of hazardous chemicals or practices.*

When a population is under risk related to exposure to an environmental hazard, epidemiological studies are often required, to better understand the relation between exposure and health effects and to increase scientific knowledge to derive ERFs (Exposure-Response functions). Nevertheless, an epidemiological study may not provide definite indications, increase uncertainty due to inconclusive results, limited power or other methodological weaknesses or, in an extreme way, can be view as an instrument to hold over actions that would have direct benefits.

A quicker tool can be represented by the integrated environmental health impact assessment (IEHIA), an activity related both to research and public health that is able to evaluate ex-ante scenarios of changes due to harmful interventions or as a result of improvements.

The main purpose of an IEHIA is to answer policy questions about the likely health impacts of planned policies or modifications of exposure scenario.

Here we present two examples of the application of these methodologies to two different regional policies: the air quality plan and the urban solid waste management plan.

An Evaluation of the Health Impact has been carried out in the context of the Strategic Environmental Evaluation belonging to the Regional Air Quality Plan 2014-2020 (PAIR2020) in Emilia-Romagna region; the goal was to estimate the health effects related to the concentration of PM10 in the whole region, based on different policy actions.

Starting from a baseline situation evaluated for 2010, in terms of environmental stressors and health status, three different scenarios were calculated for year 2020.

Environmental data used the Emilia-Romagna Regional Emission Inventory and for neighboring regions Emilia-Romagna the National Emissions Inventory (ISPRA 2005) taking into account the national energy strategy SEN2013 (source GAINS Italy); the selection of optimal emission reduction technologies through a cost-benefit analysis were made using the RIAT+ software (an integrated assessment software tool, developed in OPERA LIFE+ Project). In this way, it was possible to evaluate the evolving scenarios of population exposure to PM10 and related health impacts up to 2020, resulting from the reduction of the average regional concentration levels of pollutants in air..

Three different scenarios were provided: the CLE (Current Legislation), that took into account the changes in emissivity due to regional plans and actions already approved or adopted. The Target

Scenario plan (TS), with the main goal of respecting the maximum number of daily excesses of PM10 in almost all the regional territory. The last scenario (Maximum Feasible Reduction scenario –MFR) was the result of a theoretical simulation that applied all currently available technologies, without considering costs and practical feasibility.

Using ERFs provided by WHO, IEHIA applied to these 3 scenarios provided an estimation of avoidable deaths in each situation. Results were expressed both as attributable cases and gain in life-expectancy. The application of the TS plan would provide a gain of about 3 months for the population living in Emilia-Romagna region.

Another Italian experience focalized attention on providing guidelines and methods for HIA of population exposed to pollution due to waste management plants. SESPIR project (REF)

The SESPIR Project (Epidemiological Surveillance of Health Status of Resident Population Around the Waste Treatment Plants) assessed the impact on health of residents nearby incinerators, landfills and mechanical biological treatment plants in five Italian regions (Emilia-Romagna, Piedmont, Lazio, Campania, and Sicily). The assessment procedure took into account the available knowledge on health effects of waste disposal facilities. Within the project, suitable ERFs were calculated, reviewing existing literature. Gains in health related to the application of different scenarios were expressed by DALYs (Disability Adjusted Life Years); a significant reduction of landfills, as indicated by European Legislation, provided the most relevant improvements in health status of residents near waste management plants. Simulations on 5 regions demonstrates that the differences in DALYs between a baseline scenario (calculated for 2008) and a "green" scenario, that applies completely all indications from EU, could reduce up to 90% the impact on health due to landfills.

In this assessment time and uncertainty represent important aspects to be considered, in order to help in deciding when the epidemiological study is to be carried out for the growth of scientific knowledge and when for public health purposes.

The intrinsic probabilistic nature of this approaches involves a load of uncertainty, the nature and distribution of which is often difficult to evaluate. The uncertainty of an assessment is related to a lack of knowledge about one or more components of the assessment.

A further relevant aspect is related to the communication of results to different stakeholders. Experts cannot always express the uncertainty on their results in statistical terms; when it is only possible for them to identify that scientific knowledge is limited in a given area, the potential for surprise is therefore large and it is related to the severity of the uncertainty from the viewpoint of the decision maker. The inclusion of decision makers and stakeholders since the early stages of the study and impact assessment processes could help this aspect.

Keywords. Health Impact Assessment; Air pollution; Exposure-response functions; Attributable cases

PM10 prediction of daily data at 1 km grid using satellite data

N. Scortichini[1], N. Renzi[1], and Stafoggia[1] *

[1] *Department of Epidemiology of the Lazio Region Health Service / ASL Roma 1, Rome, Italy;* *m.scortichini@deplazio.it; m.renzi@deplazio.it; m.stafoggia@deplazio.it*

**Corresponding author*

Abstract.

Background: Health effects of air pollution, especially particulate matter (PM), have been widely investigated. However, most of the studies rely on few monitors located in urban areas for short-term assessments, or land-use/dispersion modelling for long-term evaluations, again mostly in cities. Recently, the availability of finely resolved satellite data provides an opportunity to estimate daily concentrations of air pollutants over wide spatio-temporal domains. Italy lacks a robust and validated high resolution spatio-temporally resolved model of particulate matter. The complex topography and the air mixture from both natural and anthropogenic sources are great challenges difficult to be addressed.

Materials and Methods: We combined finely resolved data on Aerosol Optical Depth (AOD) from the Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm, ground-level PM_{10} measurements, land-use variables and meteorological parameters into a four-stage mixed model framework to derive estimates of daily PM_{10} concentrations at 1-km2 grid over Italy, for the years 2006-2015. We checked performance of our models by applying 10-fold cross-validation (CV) for each year. A similar project is ongoing to estimate daily $PM_{2.5}$ concentrations nationwide.

Results: PM_{10} average concentration over the whole country and study period was 30.2 µg/m3, with higher values in winter, Northern Italy and in proximity of sites influenced by traffic sources. Air pollution levels decreased over the years, from 35 to 28 µg/m3, with similar drops across the different macro-areas. Our models displayed good agreement between observed and predicted PM concentrations, with mean CV-R2=0.65 and little bias (average slope of predicted VS observed PM10 ~ 0.99). Out-of-sample predictions were more accurate in Northern Italy (Po valley) and large conurbations (e.g. Rome), for background monitoring stations, and in the winter season. Resulting concentration maps showed highest average PM_{10} levels in specific areas (Po river valley, main industrial and metropolitan areas) with decreasing trends over time, and a clear seasonality with highest concentrations in winter and lowest in summer.

Conclusions: The results of this study will allow us to investigate short-term and long-term health effects of PM_{10} countrywide, even in areas poorly covered by routine monitoring networks, such as rural and suburban settings or municipalities influenced by industrial emissions sources. In addition, within this project we have built a large geodatabase by characterizing each 1-km2 grid cell of Italy and each day in 2006-2012 in terms of many different spatial and temporal predictors with regard to

satellite retrievals, meteorology, land cover characteristics, street and population density, orography, industrial emissions. On this regard, this represents a powerful tool on its own, which will be made available on request for applications in environmental epidemiology at the local or regional level. We are currently following up this study with a similar approach to predict daily concentrations of $PM_{2.5}$ for the period 2013-2015.

Keywords. *Aerosol optical depth; Machine learning; Particulate matter; Random Forest; Satellite.*

References

[1] de Hoogh, K., Héritier, H., Stafoggia, M., et al., 2018. Modelling daily PM2.5 concentrations at high spatio-temporal resolution across Switzerland. Environ. Pollut. 233, 1147–1154.

[2] Stafoggia, M., Schwartz, J., Badaloni, C., et al., 2016. Estimation of daily PM10 concentrations in Italy (2006–2012) using finely resolved satellite data, land use variables and meteorology. Environ. Int. 99, 234–244.

[3] Kloog, I., Nordio, F., Coull, B.A., et al., 2012. Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal PM2.5 exposures in the Mid-Atlantic states. Environ. Sci. Technol. 46, 11913–11921.

Change Detection of 4D Spatiotemporal Data Using a LASSO-Gaussian Process Approach: Preliminary results

Alessandro Fassó[1,*], Igor Valli[1] and Fabio Madonna[2]

[1] *University of Bergamo, Dalmine, BG, Italy; alessandro.fasso@unibg.it, igor.valli@unibg.it*
[1] *CNR-IMAA, C.da S. Loja, Tito Scalo, PZ, Italy; fabio.madonna@imaa.cnr.it.*
[*] *Corresponding author*

Abstract. *This talk will discuss the Gaussian Process modelling and change detection of temperature profiles from the Integrated Global Radiosonde Archive (IGRA) which consists of global radiosonde observations dating back to 1905. Change detection methods developed for radiosonde have a long history. In this paper, a locally stationary 4D geostatistical model coupled with fused LASSO is used for identifying changes.*

Keywords. *Radiosonde data; Integrated Global Radiosonde Archive (IGRA); Climate series harmonisation.*

Analysis of Data Over Complex Regions

F. Ferraccioli[1,*], L. Finos[2] and L. M. Sangalli[3]

[1]*Federico Ferraccioli, Dipartimento di Scienze Statistiche, Via Cesare Battisti, 241, 35121 Padova (Italy), ferraccioli@stat.unipd.it*
[2] *Livio Finos, Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Via Venezia, 8, 35131 Padova (Italy), livio.finos@unipd.it*
[3]*Laura M. Sangalli, MOX-Dipartimento di Matematica, Piazza L. da Vinci, 32, 20133 Milano (Italy), laura.sangalli@polimi.it*
[*]*Corresponding author*

Abstract. *We present nonparametric method for data distributed over complex spatial domains. In particular, we consider hypothesis testing procedures in the case of spatial regression models with differential regularization. We also consider a nonparametric penalized likelihood approach to density estimation over planar domains with complex geometry. The model formulation is based on a regularization with differential operators and it is made computationally tractable by means of finite element method. The performances of the proposed methods are presented through extended simulation studies.*

Keywords. *Regularization; Permutation test; Penalized likelihood; Finite element method.*

The analysis of data distributed over complex domains represent a fascinating statistical challenge, and it has stimulated recent advances in the statistical literature. The complex spatial or spatio-temporal dependencies and the presence of complicated boundaries make standard inferential tools inappropriate. A possible solution is to consider spatial regression models with differential regularization, such as [2] and [1]. Although the linear nature of these estimators allows the derivation of some distributional and asymptotic properties, the study of inferential procedures is still ongoing. In particular, we address the problem of hypothesis testing in the presence of covariates through a nonparametric approach based on the score contributions.
Moreover, analogous spatial models with differential regularization can be used to tackle density estimation problems. We propose a nonparametric penalized likelihood method for density estimation that can deal with data scattered over complex multidimensional domains, characterized by boundaries or by non-Euclidean geometries. The proposed methods leverages on advanced numerical analyses techniques, such as finite element analysis. The strong synergy between statistical and numerical approaches and tools ensures the high flexibility and computational efficiency of the method. Simulation studies are proposed to illustrate the performance of the estimators with respect to competing methods.

References

[1] Bernardi, M. S., Sangalli, L. M., Mazza, G., & Ramsay, J. O. (2017). A penalized regression model for spatial functional data with application to the analysis of the production of waste in Venice province. Stochastic

environmental research and risk assessment, 31(1), 23-38.

[2] Sangalli, L. M., Ramsay, J. O., & Ramsay, T. O. (2013). Spatial spline regression models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75(4), 681-703.

Analysis of replicated spatially correlated functional data

S. Ray[1,*] and S. Alghamdi[2]

[1] *School of Mathematics and Statistics, University of Glasgow, Glasgow, UK; surajit.ray@glasgow.ac.uk*
[2] *School of Mathematics and Statistics, University of Glasgow, Glasgow, UK; s.alghamdi.1@research.gla.ac.uk*
[*] *Corresponding author*

Abstract. *We propose a model for analyzing replicated functional data which are spatially correlated. Our research, stems from the need for accurate estimation of spatio-temporal fields by summarising information observed over several replicates. Our framework generalizes the existing framework of spatio-temporal regression model with partial differential equations regularisation (ST-PDE) approach proposed by Bernardi et al. (2017) and thus can accommodate spatially dependent functions or time dependent surfaces embedded in manifolds and irregular boundaries. This need has emerged for a study on classification of brain signals based on the difference in visual stimulus. Analytically, we show that the estimators of composite spatio-temporal field is relatively more efficient than existing estimators. The proposed method is thoroughly compared via simulation studies to existing spatio-temporal functional techniques and is applied to the analysis of the EEG data on brain signals to provide a composite temporally varying brain map over several replications.*

Keywords. *Spatial Correlation; Functional Data; Replicated functional data; Space-time model; Differential regularization; Finite elements*

References

[1] Bernardi, M. S., Sangalli, L. M., Mazza, G. and Ramsay, J. O. (2017) A penalized regression model for spatial functional data with application to the analysis of the production of waste in venice province. *Stochastic Environmental Research and Risk Assessment*, **31**, 23–38.

Smooth ANCOVA models for circular regression

Rosa M. Crujeiras[1*], María Alonso-Pena[1] and Jose Ameijeiras-Alonso[2]

[1] *Department of Statistics, Mathematical Analysis and Optimization, Universidade de Santiago de Compostela, Spain; rosa.crujeiras@usc.es, maria.alonso.pena@rai.usc.es*
[2] *Department of Mathematics, KU Leuven, Belgium; jose.ameijeirasalonso@kuleuven.be*
[*] *Corresponding author*

Abstract. *Nonparametric ANCOVA methods for regression models involving circular variables (response and/or covariates) will be proposed. The finite sample performance of the different approaches will be explored by simulation. Some illustrative real-data examples will be also provided.*

Keywords. *Smooth regression; Circular data; ANCOVA model.*

1 Introduction

Smooth regression allow to model the relation between random variables without imposing a specific (and possibly not flexible enough) parametric form for the regression function. Beyond the classical scenario, where both the response and the explanatory variable are real-valued, regression models are also required in other settings as those involving circular random variables, which can be considered as responses and/or covariates.

Circular random variables can be viewed as points in a unit circle, and smooth approaches to regression modeling involving circular covariates and responses have been proposed by Di Marzio et al. (2009) and Di Marzio et al. (2013). In this setting, a categorical covariate may be also included in the model. Anderson-Cook (1999) propose a parametric ANCOVA for circular covariate and linear response, but there is not a nonparametric alternative. It should be noted that, in the classical real-valued scenario, Young and Bowman (1995) propose a nonparametric ANCOVA model, allowing for two types of tests: equality of regression curves among groups and parallelism tests.

In this work, nonparametric ANCOVA regression models involving circular data (as response and/or covariate) will be introduced. Testing proposals for assessing equality and parallelism of regression curves will be also provided. Finite sample performance of the tests (analyzing their empirical size and power) is addressed in a simulation study. In addition, real data illustrations in the different scenarios will be also provided.

2 Some ideas

Let Y and Θ denote two linear and circular variables, and consider Δ a linear or circular covariate, depending on the model. Assume that a categorical variable with I groups may influence the response. With these premises, let us define the following nonparametric regression models:

$$Y_{ij} = m_i(\Theta_{ij}) + \varepsilon_{ij}, \quad i = 1,\dots,I,\ j = 1,\dots,n_i \tag{1}$$

$$\Theta_{ij} = (m_i(\Delta_{ij}) + \varepsilon_{ij}) \bmod(2\pi), \quad i = 1,\dots,I,\ j = 1,\dots,n_i. \tag{2}$$

Model (1) assumes a linear response and a circular covariate, whereas for a circular response, model (2) comprises two cases: linear and circular covariates. For the sake of simplicity, we will just show here the formulation of the equality test for model (1), which can be stated as:

$$H_0 : Y_{ij} = m(\Theta_{ij}) + \varepsilon_{ij}, \quad \forall\, i \in \{1,...,I\},$$
$$H_1 : Y_{ij} = m_i(\Theta_{ij}) + \varepsilon_{ij}, \quad m_i \neq m_k \text{ for some } i,k \in \{1,...,I\}.$$

Following the ideas of Young and Bowman (1995) for the equality test, the following statistic is proposed:

$$C_1 = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{I} \sum_{j=1}^{n_i} [\hat{m}_i(\Theta_{ij}) - \hat{m}(\Theta_{ij})]^2,$$

where $\hat{m}$ and $\hat{m}_i$ are, respectively, the nonparametric circular-linear estimators of m and m_i proposed by Di Marzio et al. (2009). The variance estimator $\hat{\sigma}^2$ is obtained through a new approach that adapts the estimator proposed by Gasser et al. (1986) to the periodic nature of the predictor. Note that if the errors are assumed to be iid with a normal distribution, then the distribution of C_1 under H_0 is approximated to a $a + c\chi^2_b$ distribution with the parameters calculated as a function of the cumulants of the real distribution (Young and Bowman, 1995). An analogue statistic is proposed for the parallelism test , but estimating the shift parameter of the model under H_0 through a semiparametric approach. The distribution of such statistic under H_0 is also approximated by a shifted and rescaled χ^2 distribution. When considering model (2), the test statistics formulation must be adapted to handle the circular nature of the response, considering a cosine distance and calibration will be approached by bootstrap methods.

Acknowledgments. This work has been supported by project MTM2016-76969P from the Ministry of Economy and Competitiveness and they European Regional Development Fund (ERDF).

References

[1] Anderson-Cook, C.M. (1999) A tutorial on one-way analysis of circular-linear data. *Journal of Quality Technology*, **31**, 109-119.

[2] Di Marzio, M., Panzera, A. and Taylor, C.C. (2009) Non-parametric regression for circular predictors. *Statistics and Probability Letters*, **798**, 2066-2075.

[3] Di Marzio, M., Panzera, A. and Taylor, C.C. (2013) Non-parametric regression for circular responses. *Scandinavian Journal of Statistics*, **40**, 238-255.

[4] Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986) Residual variance and residual pattern in nonlinear regression. *Biometrika*, **73**, 625-633.

[5] Young, S.G. and Bowman, A.W. (1995) Nonparametric analysis of covariance. *Biometrics*, **51**, 920-031.

Tracking the Magnetic North Pole

M. Di Marzio[1], S. Fensore[1] A. Panzera[2] and C.C. Taylor[3,*]

[1] *University of Chieti-Pescara; marco.dimarzio@unich.it, stefania.fensore@unich.it*
[2] *University of Florence; agnese.panzera@unifi.it*
[3] *University of Leeds; charles@maths.leeds.ac.uk*
[*] *Corresponding author*

Abstract. *We discuss the problem of forecasting the location of the the magnetic north pole. Based on recent evidence, last movements appear to be not completely explainable by the consolidated knowledge on the subject. Then, it could be desirable to make predictions under very mild assumptions. To this end, we propose a nonparametric approach based on sphere-sphere regression by providing some promising experimental evidence.*

Keywords. *Magnetic north pole; Nonparametric rotations; Sphere-sphere regression.*

1 Outline

There have been several recent reports about the movement of the magnetic north pole, and concern about the recent apparent increase in drift - which have led to an early update of the World Magnetic Model (WMM). The phenomenon has very recently been treated even in educational magazines like Nature, see [2]. The location of the north pole at time t can be represented by a point on the 3-d sphere, i.e. $\boldsymbol{x}_t \in \mathbb{S}^2$. We consider an AR(1) model of the form $\hat{\boldsymbol{x}}_t = f(\boldsymbol{x}_{t-1})$ with various representations for the function f — for example, a naive estimator could simply use the identity function. An obvious starting point might be to consider rotation models, but a rigid rotation would lack the flexibility to capture the observed wandering nature. So, in this work we focus on nonparametric rotation regression models, as initially proposed in [1], of the form $\hat{\boldsymbol{x}}_t = \hat{\boldsymbol{R}}^T_{\boldsymbol{x}_t}\boldsymbol{x}_{t-1}$, where the rotation depends on the location on the sphere, and uses local information. In nonparametric regression the weighted least squares solution would usually depend on distances to nearby observations, using a smoothing parameter to trade off bias against variance. However, in the time series context, we additionally consider tapering weights so they depend also on time.

Given that the movements of the pole are typically quite small (less than 50 Km per year), we consider an alternative smoothing approach, based on the fact that a rotation matrix can be expressed as the exponential of a skew symmetric matrix. So, by considering the skew symmetric matrix which corresponds to the great circle rotation of the north pole in consecutive years, we can then similarly obtain a nonparametric prediction of each component of this matrix, then mapping back to rotation space to obtain $\hat{\boldsymbol{x}}_t$.

We use yearly data on the location of the north pole (some of which are raw observations, but mostly derived from the WMM), which starts from 1590, choosing smoothing parameters by cross-validation — which can be trained each year — in order to predict the location in the following year. The results show that, for these data, the approach based on smoothing the skew symmetric matrix generally works best.

References

[1] Di Marzio, M., Panzera, A., Taylor, C.C. (2019). Nonparametric rotations for sphere-sphere regression. *Journal of the American Statistical Association* **114**, 466-476

[2] Witze, A. (2019). Earth's magnetic field is acting up and geologists don't know why. *Nature* **565**, 143-144.

Classification of directional data through data depth

G.C. Porzio[1,*], H. Demni[1,2], and A. Messaoud[2]

[1] *University of Cassino and Southern Lazio; porzio@unicas.it, houyem66@gmail.com*
[2] *Tunis Business School; amor.messaoud@gmail.com*
[*] *Corresponding author*

Abstract. *Nonparametric classification of directional data has received a lot of attention on the last few years. Directional data arise in many areas where observations are recorded as directions, rotations, clock, axes and are represented as angles relative to a fixed reference point or as unit vectors. Given that they lie on a non-linear manifold, directional data requires specific methods. Data depth can be successfully exploited to classify directional data since it provides center-outward ordering of points in any dimension. This work provides an overview on how data depth can deal with classification problems for directional data. Different depth functions and classification procedures will be investigated.*

Keywords. *DD plot; Max depth; Distribution depth classifier.*

1 Introduction

Non-parametric tools for supervised classification of directional data have been recently investigated within the literature. For instance, Di Marzio *et al.* (2019) [2] suggested to adopt a technique based on kernel density estimators. This approach seems to work properly in many settings. However, results are somehow dependent on the choice of the bandwidth parameters, and they are limited to the case of spherical data (i.e., in a low dimensional directional space).

On the other hand, data depth based methods are available for classification purposes as well. They do not require the choice of any specific parameter, and are suitable for higher dimensional classification problems. For instance, Pandolfo *et al.* (2018) [3] investigated the use of the directional max-depth classifier, while Pandolfo *et al.* (2018) [4] discussed the use of the DD-plot to classify circular data. Finally, Demni *et al.* (2019) [1] suggested a depth based distribution classifier.

However, while some results are already available, a comprehensive analysis is lacking. For this reason, this work aims at illustrating all the potential of depth based methods when adopted in supervised classification of directional objects, and at evaluating under which conditions one of the method should be preferred over the others.

2 Non-parametric depth based classifiers for directional data

The depth of a point relative to a given data set measures how deep that point lies in the data cloud with respect to its belonging distribution or multivariate sample. Depth functions have been extensively used for classification purposes given that they do not assume any particular type of probability distribution neither consider any specified parametric form for the separating surface.

Here, the main techniques are evaluated within the directional domain: the max depth classifier, the DD-plot classifier and the depth distribution classifier. The max depth classifier assigns the new data point to the class with respect to which it attains the highest depth value. In case of two populations, and with x the new point to be assigned to group j, $j = 1, 2.$, the associated classification rule is given by

$$\begin{cases} D(x,\hat{H}_1) > D(x,\hat{H}_2) \Longrightarrow \text{assign } x \text{ to population 1} \\ D(x,\hat{H}_1) < D(x,\hat{H}_2) \Longrightarrow \text{assign } x \text{ to population 2,} \end{cases}$$

where $D(x,\hat{H}_j), j = 1, 2.$, is the depth of the point x with respect to the j-th directional sample.

The depth vs. depth (DD) classifier aims to find the best polynomial separating function in a depth vs depth Euclidean space. Consequently, the generic form of the DD classifier is as follows. Let $r(.)$ be some real increasing function. Then, the classification rule is defined by

$$\begin{cases} D(x,\hat{H}_1) > r(D(x,\hat{H}_2)) \Longrightarrow \text{assign } x \text{ to population 1} \\ D(x,\hat{H}_1) < r(D(x,\hat{H}_2)) \Longrightarrow \text{assign } x \text{ to population 2.} \end{cases}$$

The distribution depth based classifier is based on the cumulative distribution functions of a depth: $F_D^H(x) := P(D(X,H) \leq D(x,H))$, where $D(x,H)$ is the depth of the point x with respect to the distribution H. Accordingly, the directional depth distribution classification rule can be thus given by ([1])

$$\begin{cases} F_D^{\hat{H}_1}(x) > F_D^{\hat{H}_2}(x) \Longrightarrow \text{assign } x \text{ to population 1} \\ F_D^{\hat{H}_1}(x) < F_D^{\hat{H}_2}(x) \Longrightarrow \text{assign } x \text{ to population 2.} \end{cases}$$

In all these methods, in case of equality, the classification rule will randomly assign the observation to one of the two groups with equal probability. These methods will be evaluated one against the other through a simulation study and they will be illustrated by means of a real data example.

References

[1] Demni, H, Messaoud, A, Porzio, G.C. (2019). The Cosine depth distribution classifier for directional data. In: Ickstadt K, Trautmann H, Szepannek G Lbke K, N, Bauer (eds), *Applications in Statistical Computing: From Music Data Analysis to Industrial Quality Improvement.* Springer-Verlag, *to appear*.

[2] Di Marzio M., Fensore S., Panzera A., Taylor C. (2019). Kernel density classification for spherical data. *Statistics & Probability Letters* **144**, 23–29.

[3] Pandolfo, G, Paindaveine, D, Porzio, G. C. (2018). Distance-based depths for directional data. *Canadian Journal of Statistics* **46(4)**, 593–609.

[4] Pandolfo, G, D'ambrosio, A, Porzio, G.C. (2018). A note on depth based classification of circular data. *Electronic Journal of Applied Statistical Analysis* **11(2)**, 447–462.

Forest Statistics from Local to Global Scales

The SEEA AFF forest accounts application in Senegal

Silvia Cerilli [a], Francesco Tubiello [a], Insa Sadio [b]

a- Statistics Division, Food and Agriculture Organization of the United Nations (FAO), Rome, Italy
b- Chef de Division des Statistiques Economiques (Agence nationale de la Statistique et de la Démographie) ANSD, Dakar, Sénégal

Abstract

Highly quality economic and environmental statistics are important inputs into evidence-based policy formulation and decision-making. Policies of conservation of forestry potential and ecological balances, or of satisfaction of national demand of timber and non-timber products need accurate information, relying on a robust statistical framework, as the SEEA CF, endorsed in 2012 by UNSC as the first UN environmental – economic statistical framework.
The System for Environmental-Economic Accounting for Agriculture Forestry and Fisheries (SEEA AFF) applies the environmental economic structures and principles described in the System of National Accounts (SNA) and in the System of Environmental Economic Accounting - Central Framework (SEEA-CF) to the activities of Agriculture, Forestry and Fisheries. It has been endorsed by the UNCEEA in March 2016 as "Internationally Agreed Methodological Document in support of the SEEA CF". The SEEA AFF includes accounting structure for Land Use, Land Cover, Forest and other wooded Land national and international data (FAOSTAT). In particular, it includes Environmentally Extended Supply and Use Table on Forestry Products, whose application and implementation in Senegal is main scope of this paper.

The SEEA AFF Physical flow account for wood forestry products records the flows in physical terms of wood products (timber) deriving from economic activities (ISIC A 021) and logging activities (ISIC A 022). The SEEA AFF expands its analysis to forestry products other than wood such as for instance resins and gums, mushrooms, wild honey, edible insects, which are derived from economic activities classified under ISIC A 023 "Gathering of non-wood forest products" (NWFP). This accounting table has been selected by the Agence Nationale de La statistique et the la Démographie (ANSD) in the framework of the UNECA Phase III of the capacity building on EE-SUTs for national implementation. The objective of this phase of the UNECA programme is to provide technical support for those pilot countries to compile one account of their selection in the coming six to nine months. This is after completing the first two phases (i.e. Phase I in "e-Training" and Phase II in "face-to-face regional seminar").

The paper aims to measure the forest assets and flows of forest-related services such as timber, fuelwood and charcoal provisioning services. It shows physical and monetary information (hectares, m3 of wood, US$) linked to traditional indicators such as GDP. The information produced can help design and monitor strategies for implementing SDG 15 Life on land, (SDG Indicator 15.1.1: Forest area as a proportion of total land area and SDG Indicator 15.2.1: Progress towards sustainable forest management), SDG 7 affordable and clean energy (sustainable energy from fuelwood) and SDG 13, climate action (reduction of climate change threats).

Keywords. *Natural Capital; SEEA CF; SEEA EEA; Forest Accounts; NWFP*

References

Eurostat. 2000. *Manual on the Economic Accounts for Agriculture and Forestry,* rev 1.1. Luxembourg, Office for Official Publications of the European Communities. Available at: http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-27-00-782/EN/KS-27-00-782-EN.pdf;

FAO, UNSD, 2016. The System of Environmental Economic accounting for Agriculture, Forestry and Fisheries. Available at: http://www.fao.org/fileadmin/templates/ess/ess_test_folder/Publications/Agrienvironmental/SEEA_AFF_White_Cover.pdf

United Nations. 2015. *Framework for the Development of Environment Statistics*. New York.

A design-based approach for mapping the diversity of forest attributes

Rosa Maria Di Biase[1,*], Caterina Pisani[2]

[1] *Department for Innovation in Biological, Agro-food and Forest systems, University of Tuscia, Via San Camillo de Lellis s.n.c., 01100 Viterbo, Italy; rmdibiase@unitus.it*
[2] *Department of Economics and Statistics, University of Siena, Piazza San Francesco 8, 53100 Siena, Italy; caterina.pisani@unisi.it*
[*] *Corresponding author*

Abstract. *Forest attributes such as volume or basal area are concentrated at tree locations and are absent elsewhere. Therefore, it is more meaningful to consider the amount of forest attributes at a pre-fixed spatial grain, within regular plots of pre-fixed size centered at the points of the study area. In this way, also the diversity of attributes within plots can be considered and quantified by suitable indexes, giving rise to a diversity surface defined on the continuum of points constituting the area. We analyze the estimation of diversity surfaces when a sample of plots is selected by a probabilistic sampling scheme and diversity within non-sampled plots is estimated using an inverse distance weighting interpolator. We discuss the design-based asymptotic properties of the resulting maps when the survey area remains fixed and the number of sampled points increases.*
Because diversity surfaces share suitable mathematical properties, if the schemes adopted to select sample points ensure an even coverage of the study areas avoiding large portions of non- sampled zones, it can be proven that the estimated maps approach the true maps.

Keywords. *diversity maps, inverse distance weighting interpolator, design-based consistency, spatial simulation, case study*

1 Introduction

Spatially explicit estimates are needed in many environmental and ecological applications for obtaining the spatial distribution of forest attributes within the area of interest [2]. These attributes, such as volume and basal area, are concentrated at tree locations and absent elsewhere. Therefore, it is more meaningful to consider the amount of forest attributes at a pre-fixed spatial grain, i.e. within regular plots of pre-fixed size centered at the points of the study area, rather than to consider the attribute amounts at single points [4]. In this way, the diversity of these attributes within plots can be quantified by suitable indexes.
Any point of the survey area can be considered as the center of a plot of pre-fixed radius, in such a way that there exists a diversity index value for any point, giving rise to a diversity surface defined on the continuum of points constituting the area. In most cases, the available resources and the continuous nature of the survey area render impossible to completely census the entire region. Therefore, the diversity indexes are recorded only within those plots centered on a sample of points and an estimation criterion is adopted to estimate the index values for those plots centered at non-sampled points, obtaining a wall-to-wall map depicting the spatial pattern of diversity throughout the whole survey area.

2 Design-based prediction

Usually, methods adopted to reconstruct population maps lie in the realm of model-dependent inference, i.e. the sampled sites are held fixed (as if they were purposively selected) and the diversity index values at these sites are supposed to be random variables generated from a continuous spatial process (super-population). Under model-dependent approaches, uncertainty stems from the super-population that has been supposed to generate the surface, conditional on the sampled sites.
However, here we follow an alternative criterion proposed by [1] for attempting the diversity map reconstruction in a design-based framework, i.e. the diversity surface is viewed as constant and uncertainty stems from the probabilistic sampling scheme adopted to select points. It follows that diversity index values at single non-sampled points are estimated by means of a spatial interpolation usually referred to as inverse distance weighting (IDW). The interpolator adopts a weighted average of diversity values recorded at sampled points with weights decreasing with the distance of the sampled points from the point under estimation. The estimation criterion is motivated by the Tobler's first law of geography [3], for which points close in space are more similar than those far apart. It should be noticed that the diversity surfaces under estimation are piecewise Lipschitz functions almost everywhere, a feature of relevant importance for the design-based estimation of these surfaces. Moreover, the asymptotic design-based properties of IDW are derived by [1] as the number of sampled points increases, outlining the conditions ensuring design-based unbiasedness and consistency. An easily computable estimator of the mean squared errors at any points of the estimated surface is adopted. It is worth noting that if the schemes adopted to select sample points ensure an even coverage of the study areas avoiding large portions of non- sampled zones, the estimated maps approach the true maps.
A simulation study on a real population of trees has been performed to estimate the maps of Shannon diversity index of tree stem diameter at breast height, whereas an application of the described method for estimating the map of the basal area diversity in the forest of the Bonis watershed (Southern Italy) has been considered as a case study.

3 Conclusions

Assumptions ensuring design-based asymptotic unbiasedness and consistency of the IDW interpolator of diversity surfaces in environmental studies concern the sampling design to allocate plots on the survey area, the distance function to be used in the interpolation, and the mathematical features of the true diversity surfaces. While the sampling design and the distance function are controllable by the researcher to a large extent, the continuity of the surfaces besides set of measure zero is guaranteed by the nature of these surfaces themselves.
The approach as here applied only exploits geographical distances, while model-based approaches, especially regression kriging, allow exploiting the information provided by various full cover and inexpensive auxiliary variables, usually derived from remote sensing sources. For this reason, our next research goal is to include auxiliary variables in the technique, once again avoiding model constraints and assumptions.

References

[1] Fattorini, L., Marcheselli, M., Pisani, C., and Pratelli, L. (2018). Design-based maps for continuous spatial populations. *Biometrika* 105, 419-429.

[2] Kanevski, M. (2008). *Advanced mapping of environmental data*. Wiley, Hoboken.

[3] Tobler, W.R. (1970). A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* **46**, 234–240.

[4] West, P.W. (2004). *Tree and Forest Measurement*. Springer, Berlin.

Mapping the intensity function of a point process in unobserved windows

E. Gabriel[1,*], J. Coville[2] and J. Chadœuf[3]

[1] *Laboratory of Mathematics, Avignon University, F-84916 Avignon, France; edith.gabriel@univ-avignon.fr*
[2] *Biostatistics and Spatial Processes Unit, INRA, F-84911 Avignon, France ; jerome.coville@inra.fr*
[3] *Statistics, UR1052, INRA, F-84911 Avignon, France; joel.chadoeuf@inra.fr*
[*] *Corresponding author*

Abstract. *Mapping species distribution is a challenging issue in ecology as exhaustive point patterns are usually unreachable at the survey scale. Our aim is thus to predict the spatial distribution of species from the knowledge of presence locations and environment relationships, taking into account any spatial interactions between individuals. Namely, we aim to estimate the intensity function of a point process in windows where it has not been observed, conditionally to its realization in observed windows, as in geostatistics for continuous processes. Spatial interactions are modeled through the pair correlation function. Our method is illustrated on simulations and used to map the spatial distribution of Cedar trees in South of France.*

Keywords. *Intensity function; Prediction; Spatial point process.*

1 Motivation

Mapping is a key issue in environmental science. A common example lies in ecology when mapping species distribution. When the location of individuals is known, we estimate the local density (usually by kernel smoothing), so-called intensity in point process theory. However, point locations are usually unreachable at the survey scale, so that sampling methods are used: distance sampling or quadrat sampling approaches to only mention the most common [1]. When no covariate is available, a global density estimation is then performed. But species distribution spatially varies as it is governed by environmental data. Several approaches have been developed in that way, for species data formed by reported presence locations also called occurrence-only records (pure records of locations where a species occurred), as Species Distribution Models (SDM) including the popular Maxent [3] and Maxlike [4]. However, point process models offer a natural framework for species distribution modelling and present many advantages. Because they operate at the individual level, they can incorporate interaction (competition, cooperation or mixed effects) between individuals and dependence to environmental covariates.

We aim to map the local intensity of a spatial point process accounting for the individual relationships modeled by the pair correlation function (which is related to the probability to find a second point of the process at a given distance from a known point of the process) and considering environmental covariates (and thus that the local intensity of the individual might spatially vary at large scale). The interest of our method is that it estimates the local intensity outside the observation window, hereafter called prediction.

2 Method

The prediction of the local intensity $\lambda(\cdot)$ of a point process Φ is obtained conditionally to the records in the observation window W_{obs}. We define a predictor as the best linear unbiased combination of the point pattern. For $x_o \notin W_{obs}$,

$$\widehat{\lambda}(x_o|\Phi_{W_{obs}}) = \sum_{x\in\Phi_{W_{obs}}} w(x;x_o)$$

We show that the weight function $w(\cdot)$ associated to the predictor is the solution of a Fredholm equation of second kind [2]. Both the kernel and the source term of the Fredholm equation are related to the second order characteristics of the point process through the pair correlation function $g(\cdot)$.

$$w(x) + \int_{W_{obs}} w(y)\lambda(y)\,(g(x-y)-1)\,\mathrm{d}y - \frac{1}{\nu(W_{obs})}\left[\int_{W_{obs}} w(x)\mathrm{d}x + \int_{W_{obs}^2} w(y)\lambda(y)\,(g(x-y)-1)\,\mathrm{d}x\mathrm{d}y\right]$$
$$= \lambda(x_o)\,(g(x_o-x)-1) - \frac{\lambda(x_o)}{\nu(W_{obs})}\int_{W_{obs}} (g(x_o-x)-1)\,\mathrm{d}x$$

We obtained similar equations for multitype proceses and/or in the presence of covariates. In order to obtain practical solutions, we restrict the solution space to that generated by linear combinations of elementary functions of a finite element basis [2].

3 Illustrations

We illustrate the method both on simulations and real data. Here, we consider a cluster process in $[2,8]\times[2,8]$, with a hardcore process with interaction radius 0.5 for the parent points and normal distribution of the offspring ($\sigma = 0.1$). The simulated pattern is plotted on the left panel of Figure 1. Grey bands correspond to unobserved windows. The middle panel of Figure 1 shows the empirical and estimated pair correlation function. The right panel of Figure 1 illustrates the prediction of the local intensity in the unobserved windows and kernel smoothing in the observed windows.

The method is used to map the spatial distribution of Cedar trees in South of France. The sampled pattern is plotted on the left panel of Figure 2 over the elevation which is used as covariate. We also consider the slope and the distance to the introduction site of Cedar trees. The related pair correlation function is plotted on the right panel of Figure 2. The prediction will be showed during the conference!

References

[1] Buckland, S. T. (2004). *Introduction to Distance Sampling: Estimating Abundance of Biological Populations*. Oxford University Press.

[2] Gabriel, E., Coville, J. and Chadœuf, J. (2017). Estimating the intensity function of spatial point processes outside the observation window. *Spatial Statistics* **22**(2), 225–239.

[3] Phillips, S., Anderson, R. and Schapire, R. (2006). Maximum entropy modeling of species geographic distributions, *Ecological Modelling* **190**, 231–259.

[4] Royle, J. A., Chandler, R. B., Yackulic, C. and Nichols, J. D. (2012). Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution* **3**, 545–554.

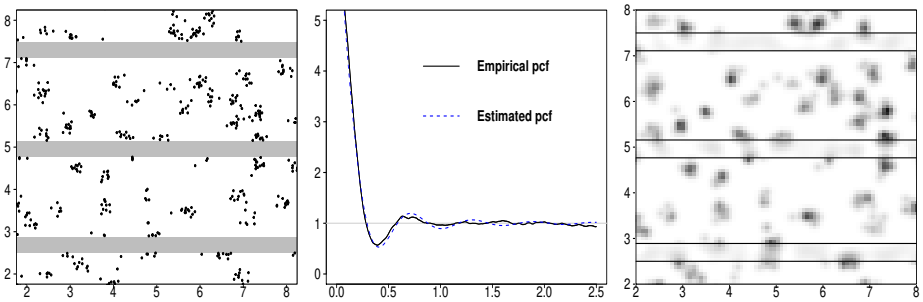

Figure 1: Left: simulated pattern. Middle: Pair correlation function. Right: Prediction.

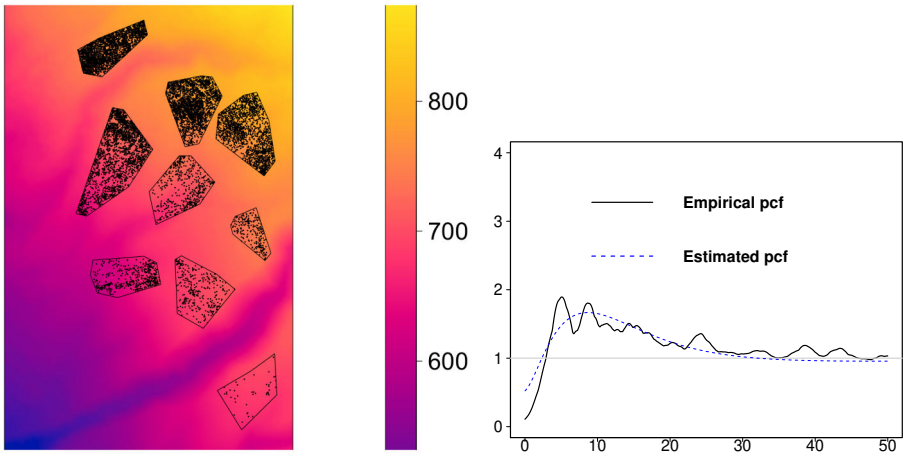

Figure 2: Left: Elevation and locations (dots) of cedar trees. Right: Pair correlation function.

Tracking daily mean flows through a river network

A. Brady[1,*], J. Faraway[1] and I. Prosdocimi[2]

[1] *Department of Mathematical Sciences, University of Bath, Bath, United Kingdom; a.brady@bath.ac.uk, j.j.faraway@bath.ac.uk*
[2] *Department of Environmental Sciences, Computer Science and Statistics, Ca' Foscari University of Venice, Venice, Italy; prosdocimiilaria@gmail.com*
[*] *Corresponding author*

Abstract. *Water-related natural hazards can have tremendous impacts on the well-being of communities; water levels severely below average can bring periods of drought and water scarcity, while those severely above average can be connected to floods. These types of hazards are typically spatial in nature. In the case of rivers, in which flows can only move downstream, the system can be considered as a network. Learning how river flows evolve throughout a network is beneficial when it comes to accurately estimating the probability at each location of exceeding some threshold for flooding. We investigate two approaches to modelling the daily mean flow at each gauging station in a network, both of which exploit the network structure of rivers. One method exploits the network structure to model the covariance between stations, making use of conditional independence and directed graphs to map out these relationships. The other uses the structure in the mean, modelling the directional behaviour of daily mean flows at each gauging station as a weighted combination of flows from stations immediately upstream (using rainfall as a predictor for those most upstream stations). An analysis is undertaken to compare how well these approaches can predict these daily flows at different points on the network. This in turn will help to estimate how often either one or a cluster of stations within the network will exceed a certain high or low threshold under different scenarios. The methods discussed will be showcased using the network of station in the river Eden catchments in the northwest of England, which has experienced a series of devastating floods in the last 15 years.*

Keywords. *Flooding; Bayesian network; Conditional independence; Directional methods.*

Acknowledgments. Aoibheann Brady is supported by a scholarship from the EPSRC Centre for Doctoral Training in Statistical Applied Mathematics at Bath (SAMBa), under the project EP/L015684/1. The authors thank the UK National River Flow Archive (http://www.ceh.ac.uk/data/nrfa) for making the river flow data available.

Modelling the spatial extent and severity of extreme European windstorms

P. Sharkey[1,*], J. Tawn[2] and S. Brown[3]

[1] *BBC, Bridge House, MediaCityUK, Salford, M50 2BH, UK; pgshky@gmail.com*
[2] *Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, UK; j.tawn@lancs.ac.uk*
[3] *Met Office Hadley Centre, Fitzroy Road, Exeter, EX1 3PB, UK; simon.brown@metoffice.gov.uk*
[*] *Corresponding author*

Abstract. *Windstorms are a primary natural hazard affecting Europe that are commonly linked to substantial property and infrastructural damage. Extreme winds are typically generated by extratropical cyclone systems originating in the North Atlantic, which are often characterised by a track of local vorticity maxima. While there have been numerous statistical studies on modelling extreme winds, little has been done to model the influence of the extratropical cyclone on the wind speeds that they generate. By modelling the development of windstorms in a Lagrangian frame of reference, we can assess the joint risk of severe events occurring at multiple sites.*

In this talk, we present a novel approach to modelling windstorms that preserves the physical characteristics linking the windstorm and the cyclone track by exploring the dependence structure of these characteristics in a Lagrangian frame of reference. We explore a combined copula/spatio-temporal filtering approach to identify and extract the spatial footprint of extreme windstorm events, before using a Markov process to propagate the characteristics of the footprint in time relative to the cyclone track.

Our model allows simulation of synthetic windstorm events, which one can use to quantify the risk associated with previously unobserved events at different sites, thus representing a useful tool for practitioners with regard to risk assessment. In particular, we show, for case studies in the northwest of England and eastern Germany, that the spatial extent of windstorms become more localised as its magnitude increases, while our model captures the varying degrees of spatial dependence at different sites.

Keywords. *Climate extremes; Extreme value analysis; Spatial dependence; Windstorms*

Extreme weather, ensemble prediction and postprocessing: from forecast to evaluation

M. Taillardat

Météo-France and CNRM UMR 3589, France; maxime.taillardat@meteo.fr

Abstract. *Forecast and verification of ensemble prediction systems for extreme events remain a challenging question in the numerical weather prediction community. The general public as well as the media pay particular attention on extreme events and conclude about the global predictive performance of ensembles, which are often unskillful when they are needed. In this talk, an overview of the postprocessing methods implemented by the statistical community for the ensemble forecast of extremes is presented. Thus, the paradigm of maximizing the sharpness subject to calibration can be associated with the paradigm of maximizing the value for extreme events subject to a good overall performance.*

Keywords. *Extremes ; Weather ; Forecast ; Calibration ; Postprocessing.*

References

[1] Taillardat, M., Mestre, O., Zamo, M., & Naveau, P. (2016). Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, **144(6)**, 2375–2393.

[2] Taillardat, M., Fougères, A.-L., Naveau, P., & de Fondeville, R. (2019). Extreme events evaluation using CRPS distributions. *arXiv preprint arXiv:1905.04022*.

[3] Taillardat, M., Fougères, A.-L., Naveau, P., & Mestre, O. (2019). Forest-based and semi-parametric methods for the postprocessing of rainfall ensemble forecasting. *Weather and Forecasting*, https://doi.org/10.1175/WAF-D-18-0149.1

A functional kriging approach to multi-fidelity modelling

O. Grujic[1], A. Menafoglio[2,*], G. Yang[1] and J. Caers[1]

[1] *Department of Geological Sciences, Stanford University, USA;*
[2] *MOX, Dipartimento di Matematica, Politecnico di Milano, Italy; alessandra.menafoglio@polimi.it*
[*] *Corresponding author*

Abstract. *The geostatistical analysis of complex data has recently received much attention in the literature, motivated by the increasing availability of massive and heterogenous georeferenced datasets in varied industrial and environmental contexts. Data fusion plays an important role in these settings, because of the need to merge and integrate data from diverse sources – possibly associated with different degrees of uncertainty – while accounting for the possible dependence among the data. In this communication, we consider the problem of fusing the functional responses of a reservoir model, when these are obtained at different degrees of fidelity (i.e., multi-fidelity modeling) and refer to a range of input parameters in a given design of experiment (DoE). Here, the proximity among input parameters in the DoE indeed induces a dependence in the functional responses.*

The context of our study is that of kriging meta-modeling. This is a classical scheme for statistical emulation of numerical models which consists of (i) using statistical DoE on the input space, (ii) employ the numerical model (high-fidelity) to compute a set of outputs for the DoE, and finally (iii) fit a statistical model to predict (via kriging) the numerical models output corresponding to a given set of (new) inputs. In this work, we extend these ideas to perform kriging meta-modeling in the presence of complex outputs, such as functional solutions. For this purpose, we here follow the approach of Object Oriented Spatial Statistics (O2S2, [2]), a recent system of ideas and methods that allows the analysis of complex data when their (spatial) dependence is an important issue. The foundational idea of O2S2 is to interpret the data as objects: the atom of the analysis is the entire object, which is seen as an indivisible unit rather than a collection of features. In this view, the model outputs are interpreted as random points within a space of objects – called feature space – whose dimensionality and geometry should properly represent the data features and their possible constraints.

In this mathematical framework, in [1] we propose a novel object-oriented kriging meta-modeling method that allows to integrate the complex responses obtained from a high-fidelity model with that of a low-fidelity model, in a co-kriging setting. The developed emulator exactly reproduces the high-fidelity training data, but also allows to take advantage of the additional information contained in the low-fidelity solutions.

Although the presented approach is completely general and in principle allows the analysis of very general types of objects, for illustrative purposes we will give emphasis to the case of numerical models with a functional response. We shall describe our developments on an example concerning an oil-water reservoir model, and test on this case its performances.

Keywords. *Object oriented spatial statistics; Statistical meta-modeling; Uncertainty quantification.*

References

[1] Grujic, O., Menafoglio, A., Yang, G., Caers, J. (2018) Cokriging for multivariate Hilbert space valued random fields. Application to multifidelity computer code emulation. *Stochastic Environmental Research and Risk Assessment*, 32(7), 1955–1971.

[2] 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.

Functional approaches for spatiotemporal satellite data

C. Miller[1], M. Scott[1], A. Sehn[1], R. O'Donnell[1], M. Gong[2], C. Wilkie[1],

[1]*School of Mathematics and Statistics, University of Glasgow;*
[2]*British Geological Survey*

Abstract.

Developments in satellite retrieval algorithms continually extend the extraordinary potential of satellite platforms such as the MEdium ReSolution Imaging Spectrometer (MERIS), the Advanced Along-Track Scanning Radiometer (AATSR) and the Ocean Land Colour Instrument (OLCI) to retrieve information across the Earth at finer spatial resolution. Such instruments now enable water quality proxies (such as chlorophyll-a, coloured dissolved organic matter) to be retrieved for lakes and rivers at a global scale.

Challenges associated with these new environmental data streams are the large volumes of data in space and time, collected as images, missing data and uncertainty induced in the production of the water quality proxy measurements. Functional data analysis provides an attractive approach for efficient dimensionality reduction to investigate spatiotemporal images for lake water quality proxies and associated reflectance data, and provides a context for data fusion, linking the satellite data to in-situ monitoring.

The GloboLakes (www.globolakes.ac.uk) project has developed functional clustering methods and nonparametric downscaling to investigate temporal coherence globally for water quality parameters, and data fusion to enable satellite data to be bias-corrected using in-situ data of differing spatiotemporal support. Recent work extends these methods by investigating the gain in accuracy and information by working at the reflectance data level for satellite retrievals and developing approaches to attribute functional clustering water quality proxy results to surrounding catchment information. Additionally, functional data analysis of reflectance data will enable the identification of the variability and bias induced through using different sensors, so that further development of appropriate data fusion approaches to combine outputs from multiple satellite sensors will be needed.

Methods will be illustrated using data from the AATSR, MERIS and OLCI on the European Space Agency EnviSat and Sentinel 3A satellite platforms.

D-STEM: functional hidden dynamic geostatistical model

Y. Wang[1], F. Finazzi[2,*], and A. Fassò [2]

[1] *Guanghua School of Management, Peking University, China;*
[2] *Department of Management, Economics and Quantitative Methods, University of Bergamo, Italy;*
francesco.finazzi@unibg.it
[*] *Corresponding author*

Abstract. *With the increase of multidimensional data availability and modern computing power, statistical models for spatial or spatio-temporal data are developing at a fast pace. Although some of the above software packages consider both space and time, to our knowledge at the time of writing, none of them handles data in a 3D space $\times$ time, nor handles profile data indexed in space and time. This kind of data arises considering, for example, global atmospheric time series, where the information is related to latitude, longitude and altitude in the atmosphere. For example, temperature atmospheric profiles are measured by radiosondes which fly from ground level up to the stratosphere. Atmospheric profile data are also produced by interferometric sensors aboard remote sensing satellites or laser based methods, e.g. LIDAR.*
The software D-STEM (distributed spatiotemporal expectation-maximization) is a statistical tool, implementing three spatial-temporal models: the dynamic coregionalization model (DCM), the hidden dynamic geostatistic model (HDGM), and the functional hidden dynamic geostatistic model (f-HDGM). The dynamic coregionalization model (DCM) is introduced by Finazzi and Fassò. Calculli et al. (2015) propose the hidden dynamic geostatistic model (HDGM) with application to air quality in Apulia, Italy. The main development in this version is f-HDGM, which fits a four-dimension spatiotemporal model based on functional data approach.
Exploring atmospheric data from a 3D spherical shell requires understanding from both the spherical domain and the atmospheric vertical dynamics. Thus, based on the functional representation of atmospheric vertical profiles, the sphere $\times$ time statistical model is proposed. To note, 3D $\times$ T functional data is not restricted to the three-dimensional space. In the case study of Ozone data, we show how to make full use of the software by changing the third dimension - altitude to the domain we are interested. Therefore, the f-HDGM is applied even with ground level data.
The results of model estimation consist of the estimated parameters, the variance-covariance matrix and the observed data log-likelihood. Besides, cross-validation can be conducted to assess the accuracy of model prediction. And kriging is used to map the environmental variable over a region or obtain the atmospheric profile at a certain site if the corresponding covariates are provided.

Keywords. *Functional data analysis; 4D data; Climate data; Environmetrics; EM algorithm.*

Variable selection in small area models

S. Arima[1,*], S. Polettini[2]

[1] *Department of Methods and Models for Economy, Territory and Finance, Sapienza University of Rome; serena.arima@uniroma1.it*
[2] *Department of Social and Economic Sciences, Sapienza University of Rome; silvia.polettini@uniroma1.it*
[*] *Corresponding author*

Abstract. *Model based small area estimation relies on mixed effects regression models that link the small areas and borrow strength from similar domains. The variability of the random effects, while accounting for lack of fit, affects uncertainty of both point and interval estimators of small area means. In the presence of good covariates, low variation of the random small area effects is expected, but when measurement error is present it has been proved that parameter estimates may be dramatically biased and the variability of the random effects and, consequently, of the small area means significantly increases. Adopting a fully Bayesian approach, we define a mixture model that allows us, using spike and slab priors, to infer the presence or not of measurement error in the covariates. We empirically evaluate the accuracy of the estimates in different simulation scenarios. We also apply the proposed procedure to the benthic study carried out by the Dutch Institute RIKZ and analyzed in Zuur et al. (2009) to investigate species richness.*

Keywords. *Small area models; measurement error; hierarchical models.*

Model based small area estimation relies on mixed effects regression models that link the small areas and borrow strength from similar domains. The variability of the random effects, while accounting for lack of fit, affects uncertainty of both point and interval estimators of small area means. Random effects models play an important role in model-based small area estimation; indeed, they account for any lack of fit of a regression model for the population of small areas on a set of explanatory variables. While the inclusion of the random effect may improve the adaptivity and flexibility of the Fay-Herriot model, it also increases the uncertainty of both point and interval estimators of small area means. Because of that, several tests and variable selection procedures have been developed in order to verify the presence or not of the random effects in such models. In general, the reliability of model-based small area estimates is closely related to the availability of good covariates, implying small variation of the random small area effects. However in practice, one may define a model with poor covariates because they are affected by measurement error, an ubiquitous problem ([2]) also studied within the small area literature (see [4], [1]). In the presence of poor covariates, the variability of the random effects increase, with shrinkage to the sampling component of the small area estimator. Correcting for measurement error reduces the random effects variability and improves the resulting estimates, provided the induced uncertainty is small compared to the sampling variation. Working on the model described in [4], we propose a unit-level small area model with measurement error in auxiliary variables that, borrowing from [3], includes a "spike and slab" distribution for modelling the inclusion of the covariate measured with error in each area.

Such model allows us to infer the presence or not of measurement error in the covariates for each area. We empirically evaluate the accuracy of the estimates in different simulation scenarios, varying the percentage of areas in which covariates are measured with error. We apply the proposed model to data from the benthic study, carried out by the Dutch Institute RIKZ and analyzed in [5] to investigate the benthic species richness. Samples at 45 stations along the coastline were taken and benthic species were counted. To measure diversity, species richness (the different number of species) per site was calculated. A possible factor explaining species richness is Normal Amsterdams Peil (NAP), which measures the height of a site compared to average sea level, and represents a measure of food for birds, fish, and benthic species. Since NAP is obtained as a combination of several variables, it might be prone to measurement error and such error might differ from site to site. We estimate the proposed model accounting for the presence/absence of measurement error and assess the small area mean richness. Comparison with the standard nested error model is performed.

References

[1] Arima, S., Datta, G. S., and Liseo, B. (2015). Bayesian estimators for small area models when auxiliary information is measured with error. *Scandinavian Journal of Statistics,* **42**, 518 –529

[2] Carroll, R.J., Ruppert, D., Stefanski, L. and Crainiceanu, C. (2006) *Measurement error in nonlinear models: a modern perspective.* 2nd edn. Chapman & Hall, CRC.

[3] Datta, G.S. and Mandal, A. (2015). Small area estimation with uncertain random effects *Journal of the American Statistical Association*, **110 (512)**, 1735 – 1744.

[4] Ghosh, M., Sinha, K. and Kim, D. (2006). Empirical and hierarchical Bayesian estimation in finite population sampling under structural measurement error model, *Scandinavian Journal of Statistics*, **33**, 591 – 608.

[5] Zuur A., Ieno E., Walker N. , Saveliev, A., Smith G.(2009) *Mixed effects models and extensions in ecology with R*. Statistics for Biology and Health. Springer, New York, NY

Autoregressive random effects models for circular longitudinal data using the embedding approach

A. Maruotti[1,2,*] and M. Ranalli[3]

[1] *Dipartimento di Giurisprudenza, Economia, Politica e Lingue Moderne. Libera Università Maria Ss Assunta, Via Pompeo Magno 22 - 00192 Roma; a.maruotti@lumsa.it*
[2] *Deparment of Mathematics. University of Bergen; Antonello.Maruotti@uib.no*
[3] *Dipartimento di Economia e Finanza, Università degli Studi di Roma Tor Vergata, Rome, Italy*
[*] *Corresponding author*

Abstract. *Some conditional models to deal with circular longitudinal responses are proposed, extending random effects models to include serial dependence of Markovian form, and hence allowing for quite general association structures between repeated observations recorded on the same unit. The presence of both these components implies a form of dependence between them, and so a complicated expression for the resulting likelihood. To handle this problem, we introduce an approximate conditional mode and a full conditional model, with no assumption about the distribution of the time-varying random effects. All of the discussed models are estimated by means of an EM algorithm for nonparametric maximum likelihood.*

Keywords. *Hidden Markov models; Initial conditions; Finite mixtures; Conditional models.*

1 The embedding approach

Let us introduce the random vector $\mathbf{Y}_{it}$, for unit $i = 1,\ldots,I$ at time $t = 1,\ldots,T$, following a d-dimensional Normal distribution, with mean $\boldsymbol{\mu}_{it}$ and covariance matrix $\boldsymbol{\Sigma}$, i.e. $\mathbf{Y}_{it} \sim N_d(\boldsymbol{\mu}_{it}, \boldsymbol{\Sigma})$. The random unit vector

$$\mathbf{U}_{it} = \frac{\mathbf{Y}_{it}}{||\mathbf{Y}_{it}||}$$

is said to follow a projected Normal distribution, i.e. $\mathbf{U}_{it} \sim PN_d(\boldsymbol{\mu}_{it}, \boldsymbol{\Sigma})$; see Wang and Gelfand (2013). The general version of the projected normal distribution allows asymmetry and bimodality, i.e different shapes can be modelled. However, the general projected normal distribution is not identified and substantially increases the computational burden required in the estimation step. The distribution of $\mathbf{U}_{it}$ is unchanged if $(\boldsymbol{\mu}_{it}, \boldsymbol{\Sigma})$ is replaced by $(c\boldsymbol{\mu}_{it}, c^2\boldsymbol{\Sigma})$ for any $c > 0$, but this lack of identifiability can be addressed imposing constraints on $\boldsymbol{\Sigma}$. Wang and Gelfand (2013) suggest to set one of the variances in $\boldsymbol{\Sigma}$ to 1 to provide identifiability, resulting in a four-parameter distribution. Other constraints could be also considered as e.g. restricting the determinant of $\boldsymbol{\Sigma}$ to equal 1.

The $\mathbf{U}_{it}$ variable can be converted to an angular random variable, say Θ_{it}, relative to some direction treated as 0. Indeed, any Θ_{it} can be obtained from the radial projection of the bivariate normal distribution by using the *arctan** function defined by Jammalamadaka and SenGupta (2001; p. 13), i.e.

$\Theta_{it} = arctan^*\left(\frac{Y_{it2}}{Y_{it1}}\right) = arctan^*\left(\frac{U_{it2}}{U_{it1}}\right)$. The following explicit relation exists between $\mathbf{Y}_{it}$ and the circular variable Θ_{it}

$$\mathbf{Y}_{it} = \begin{bmatrix} Y_{it1} \\ Y_{it2} \end{bmatrix} = \begin{bmatrix} R_{it}\cos\theta_{it} \\ R_{it}\sin\theta_{it} \end{bmatrix} = R_{it}\mathbf{U}_{it},$$

where $R_{it} = ||\mathbf{Y}_{it}||$.

In the following, we will focus exclusively on the case $\boldsymbol{\Sigma} = \mathbf{I}$ and $d = 2$ (i.e. on circular data). If in addition, $\boldsymbol{\mu}_{it} = \mathbf{0}$, then $\mathbf{U}_{it}$ is uniformly distributed on the circle; otherwise the distribution of $\mathbf{U}_{it}$ is unimodal and rotationally symmetric about its mean direction $\boldsymbol{\mu}_{it}/||\boldsymbol{\mu}_{it}||$. Indeed, departure from zero for the two means, in the case of an identity covariance matrix, creates one mode in the trigonometric quadrant with the same sign of the means, e.g. if $\mu_{it1} > 0$ and $\mu_{it2} < 0$, where $\boldsymbol{\mu}_{it} = (\mu_{it1}, \mu_{it2})$, then the mode is in the quadrant with positive cosine and negative sine.

The joint density $f(\theta_{it}, r_{it} \mid \boldsymbol{\mu}_{it}, \mathbf{I})$ can be easily obtained by transforming the bivariate normal distribution of $\mathbf{y}_{it}$ to polar coordinates, i.e. $f(\theta_{it}, r_{it} \mid \boldsymbol{\mu}_{it}, \mathbf{I}) = f(r_{it}\mathbf{u}_{it} \mid \boldsymbol{\mu}_{it}, \mathbf{I})r_{it}$ and, thus, $f(\theta_{it} \mid \boldsymbol{\mu}_{it}, \mathbf{I}) = \int f(\theta_{it}, r_{it} \mid \boldsymbol{\mu}_{it}, \mathbf{I})dr_{it} = \phi(\mu_{it1}, \mu_{it2}; \mathbf{0}, \mathbf{I}) + \mu_{it1}\cos\theta_{it} + \mu_{it2}\sin\theta_{it}$, i.e. $\theta_{it} \sim PN_2(\boldsymbol{\mu}_{it}, \mathbf{I})$, with $\phi(\cdot)$ denoting the density function of the bivariate normal distribution.

In empirical applications, the angle θ_{it} is usually collected. However, as discussed above, we prefer to work with its radial projections as the resulting model can be easily dealt with by using standard regression modelling strategies. In other words, we model $\mathbf{Y}_{it}$ and focus our interest upon the parameter vector $\boldsymbol{\mu}_{it}$, which is modelled, in a regression framework, by defining a multivariate linear mixed model, as defined in the following sections.

2 The random effects model

The temporal evolution of the random effects can be conveniently described by including a vector of time-varying random effects, say $\mathbf{b}_{it} = (b_{it1}, b_{it2})$. Regarding $\mathbf{b}_{it}$'s distribution, we assume a (hidden) Markov chain with states $\mathbf{b}_k = (b_{k1}, b_{k2}), k = 1, \ldots, K$, initial probabilities $\pi_{ik} = \Pr(\mathbf{b}_{i1} = \mathbf{b}_k) = \pi_k$ and transition probability matrix $\boldsymbol{\Pi} = \{\pi_{it,k|h}\}$ with $\pi_{it,k|h} = \Pr(\mathbf{b}_{it} = \mathbf{b}_k \mid \mathbf{b}_{it-1} = \mathbf{b}_h) = \pi_{k|h}, t > 1$, i.e. Markov chain's parameters will be assumed independent on any covariates and shared among subjects.

The modelling framework is completed by defining the regression model (see also Maruotti et al., 2016)

$$\mu_{itj} = \mathbf{x}'_{it}\boldsymbol{\beta}_j + b_{itj}, \quad j = 1, 2,$$

where $\mathbf{x}_{it} = (1, x_{it1}, \ldots, x_{itp}, \theta_{it-1})$, $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{1j}, \ldots, \beta_{pj}, \beta_{p+1,j})$ represents the $(p+2)$-dimensional vector of regression parameters referred to the j-th projection and $\mathbf{b}_i = (b_{i1}, b_{i2})$ denotes a set of subject- and projection-specific random effects. However, in order to easily implement the estimation steps, the following multilevel specification can be considered

$$\mu_{itj} = \sum_{j=1}^{2} d_{itj}(\mathbf{x}'_{it}\boldsymbol{\beta}_j + b_{itj})$$

where we use a set of indicator variables d_{itj}, with $d_{itj} = 1, \forall i = 1, \ldots, I; t = 1, \ldots, T, j = 1, 2$ iff the j-th projection is to be modelled and 0 otherwise. Using a matrix notation

$$\boldsymbol{\mu}_{it} = \mathbf{x}^*_{it}\boldsymbol{\beta}^* + \mathbf{b}^*_{it}$$

where

$$\boldsymbol{\mu}_{it} = \begin{bmatrix} \mu_{it1} \\ \mu_{it2} \end{bmatrix}, \quad \mathbf{x}_{it}^{*} = \mathbf{I}_2 \otimes \mathbf{x}_{it}, \quad \boldsymbol{\beta}^{*} = vec(\boldsymbol{\beta}) = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}, \quad \mathbf{b}_{it}^{*} = vec(\mathbf{b}_{it}) = \begin{bmatrix} b_{it1} \\ b_{it2} \end{bmatrix}.$$

If the covariates are not the same for both projections, some of the elements of $\boldsymbol{\beta}_j$ would be set equal to zero.

We would remark that the circular mean direction and concentration, i.e., the circular counterpart of the mean and precision of a linear random variable, are respectively $\bar{\mu}_{it} = arctan^{*}\left(\frac{\mu_{it2}}{\mu_{it1}}\right)$ and $c_{it} = (\pi\gamma_{it}/2)^{1/2}\exp(-\gamma_{it})(I_0(\gamma_{it}) + I_1(\gamma_{it}))$, where $\gamma_{it} = ||\boldsymbol{\mu}_{it}||^2/4$ and $I_v(\gamma)$ is the modified Bessel function of the first kind of order v, see Wang and Gelfand (2013). Both $\bar{\mu}_{it}$ and c_{it} depend on the means of the projections hence the regression type specification of $\mu_{it\,j}$ can adjust for change in mean direction and concentration due to different levels of covariates.

2.1 Likelihood inference

Inference for the proposed model is based on the log-likelihood

$$\ell(\boldsymbol{\lambda}) = \sum_{i=1}^{I} \log \left\{ \sum_{\mathbf{b}_{i1}} \cdots \sum_{\mathbf{b}_{iT}} \left[\pi_{\mathbf{b}_{i1}} \prod_{t>1} \pi_{\mathbf{b}_{it}|\mathbf{b}_{it-1}} \prod_{t} f(\theta_{it} \mid \mathbf{x}_{it}, \mathbf{b}_{it}) f(\theta_{i0} \mid \mathbf{x}_{i0}, \mathbf{b}_{i0}) \right] \right\}$$

with the sum $\sum_{\mathbf{b}_{it}}$ extended to all possible configurations of $\mathbf{b}_{it}$ and where $\boldsymbol{\lambda}$ is a short-hand notation for all non-redundant parameters. However, inferences can be highly sensitive to misspecification of $f(\theta_{i0} \mid \mathbf{x}_{i0}, \mathbf{b}_{i0})$. Thus, we rewrite the previous expression as

$$\ell(\boldsymbol{\lambda}) = \sum_{i=1}^{I} \log \left\{ \sum_{\mathbf{b}_{i1}} \cdots \sum_{\mathbf{b}_{iT}} \left[\pi_{\mathbf{b}_{i1}}(\theta_{i0}) \prod_{t>1} \pi_{\mathbf{b}_{it}|\mathbf{b}_{it-1}}(\theta_{i0}) \prod_{t} f(\theta_{it} \mid \mathbf{x}_{it}, \mathbf{b}_{it}) f(\theta_{i0} \mid \mathbf{x}_{i0}) \right] \right\}$$

or equivalently

$$\ell(\boldsymbol{\lambda} \mid \theta_{i0}) = \sum_{i=1}^{I} \log \left\{ \sum_{\mathbf{b}_{i1}} \cdots \sum_{\mathbf{b}_{iT}} \left[\pi_{\mathbf{b}_{i1}}(\theta_{i0}) \prod_{t>1} \pi_{\mathbf{b}_{it}|\mathbf{b}_{it-1}}(\theta_{i0}) \prod_{t} f(\theta_{it} \mid \mathbf{x}_{it}, \mathbf{b}_{it}) \right] \right\}$$

resulting in a conditional likelihood, where $\pi_{\mathbf{b}_{i1}}(\theta_{i0})$ and $\pi_{\mathbf{b}_{it}|\mathbf{b}_{it-1}}(\theta_{i0})$ allows the random effects distribution to dependent on θ_{i0}.

References

[1] Jammalamadaka RA, SenGupta A (2001). *Topics in Circular Statistics*. World Scientific

[2] Maruotti A, Punzo A, Mastrantonio G and Lagona F (2016). A time-dependent extension of the projected normal regression model for longitudinal circular data based on a hidden Markov heterogeneity structure. *Stochastic Environmental Research and Risk Assessment*, 30: 1725-1740.

[3] Wang F, Gelfand A (2013). Directional data analysis under the general projected normal distribution. *Statistical Methodology* **10**: 113–127.

Multivariate change-point analysis for climate time series

Gianluca Mastrantonio[1*], Giovanna Jona Lasinio[2], Alessio Pollice[3], Giulia Capotorti[4], Lorenzo Teodonio[5], Carlo Blasi[4]

[1]*Department of Mathematical Sciences, Politecnico di Torino ; gianluca.mastrantonio@polito.it*
[2]*Department of Statistical Sciences, Sapienza Università di Roma*
[3]*Department of Economics and Finance, Università di Bari Aldo Moro*
[4]*Department of Environmental Biology, Sapienza Università di Roma*
[5]*ICRCPAL, Ministry of Cultural Heritage and Activities and Tourism, Roma*
[*]*Corresponding author*

Abstract.
The aim of this work is to find individual and joint change-points in a large multivariate database of climate data. We model monthly values of precipitation, minimum and maximum temperature recorded in 360 stations covering all Italy for 60 years (12 $\times$ 60 months). The proposed three variate Gaussian change-point model exploits the Hierarchical Dirichlet process, allowing for a formalization that lets us estimate a different change-point model for each station. As stations possibly share some of the parameters of the trivariate normal emission distribution, this model framework provides an original definition of the change-points corresponding to changes in any subset of the 9 model parameters.

Keywords. *Change-point model; Hierarchical Dirichlet process; Climate data*

Climate elements and regimes, such as temperature, precipitation and their annual cycles, primarily affect the type and distribution of plants, animals, and soils as well as their combination in complex ecosystems. The ecological classification of climate represents one of the basic steps for the definition and mapping of ecoregions, i.e. of broad ecosystems occurring in discrete geographical areas. As a matter of fact, the large temporal scale of the dataset implies that the time span may subtend several evolving patterns of the underlying series. From a botanical perspective, bioclimatic time regimes correspond to abrupt changes in the climatic behavior and support inferences on the potential effects of these changes on ecosystem composition, functionality, distribution, and dynamics at different time scales ([2]).

To investigate the presence of change-points in thermo-pluviometric historical data over the Italian peninsula, we consider monthly records of precipitation and min/max temperature at 360 monitoring stations over 60 years (1951-2010). The data were mostly provided by national institutions (ISPRA, CRA/CREA, Meteomont and ENEA) and local authorities. Monthly records were obtained considering monthly cumulative precipitations and monthly averages of daily minimum and maximum air temperatures. The full database has $3 \times 360 \times 60 \times 12$ entries, though almost all time series are affected by variable amounts of missing data. Series observed over a very large time span are usually subject to changes of their structure and features, concerning both the first order (means) and the second order (variances and correlations) properties. Moreover, seasonality is present and cannot be disregarded.

We consider mixture modeling for model-based clustering, where mixture components act as cluster generators and classification is equivalent to the model fitting process. In particular, we view the multi-

variate data as realizations of a Gaussian process with a finite or countably infinite set of vector-valued parameters. In order to simplify the joint distribution modeling of the climate variables, we standardize the temperatures and rescale the precipitation with its standard deviation; the latter is then seen as the realization of a latent variable belonging to the real line ($\mathbb{R}$), where the negative values are associated to the event "no precipitation" ([3]). Two time-points are clustered (i.e. they belong to the same regime) if they correspond to realizations from a common member of this set. The stochastic features of the set of vector-valued parameters are specified in a further level of the hierarchical model by Dirichlet process modeling ideas ([5]).

In this work we present some preliminary results.The model was implemented on the TeraStat cluster [1]. The code is written in R/C++, and uses the openMP library [4] to perform parallel computing. Our proposal allows for a very rich inference on the joint change-point detection. Posterior estimates are obtained in 6 days, with 40000 iterations per day and 10 GB of ram usage. Initial results are very encouraging and are partially limited only by computational issues, that we plan to solve in the near future.

Acknowledgments. The work of the first three authors is partially developed under the PRIN2015 supported-project Environmental processes and human activities: capturing their interactions via statistical methods (EPHAStat) funded by MIUR (Italian Ministry of Education, University and Scientific Research) (20154X8K23-SH3).

References

[1] Umbero Ferraro Petrillo and Guerriro Raimato. Terastat computer cluster for high performance computing. "http://www.dss.uniroma1.it/en/node/6554" Department of Statistical Science Sapienza university of Rome, 2014.

[2] Y. Liu and H. Lei. Responses of natural vegetation dynamics to climate drivers in china from 1982 to 2011. *Remote Sensing*, 7:10243–10268, 2015.

[3] Gianluca Mastrantonio, Giovanna Jona Lasinio, Alessio Pollice, Giulia Capotorti, Lorenzo Teodonio, Giulio Genova, and Carlo Blasi. A hierarchical multivariate spatio-temporal model for clustered climate data with annual cycles. *Annals of Applied Statistics*, in press.

[4] OpenMP Architecture Review Board. OpenMP application program interface version 3.0, May 2008.

[5] Yee Whye Teh and Michael I. Jordan. Hierarchical Bayesian nonparametric models with applications. In N. Hjort, C. Holmes, P. Müller, and S. Walker, editors, *Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press, 2010.

Monitoring the spatial interactions among data streams generated by spatio-temporal processes.

A. Balzanella[1], A. Irpino[1] and R.Verde[1]

[1] *Università della Campania Luigi Vanvitelli, Dep. Mathematics and Physics; antonio.balzanella@unicampania.it, antonio.irpino@unicampania.it, rosanna.verde@unicampania.it*
[*] *Antonio Balzanella*

Abstract. *Massive datasets having the form of potentially unbounded data streams are becoming very common due to the availability of sensor networks which record data at very high rate. We can think of monitoring environmental variables, electricity consumptions, pollutions. In these real world applications dataset size grows very quickly and data is expected to evolve over time. Often, data collected by sensors still depend on the geographic location of each sensing device.*
This paper introduces a strategy for measuring the spatial dependence among data streams. Traditional data mining tools fail to cope with data streams since they require that data is stored for being processed. Data stream learning should satisfy the following constraints: 1) Time required for processing the incoming observations has to be small and constant; 2) Allowed memory resources are orders of magnitude smaller than the total size of input data; 3) Algorithms cannot perform more than one scan of the data; 4) Knowledge about data should be available at any point in time or on user demand.
Thus, data stream mining algorithms should update incrementally and continuously over time the knowledge about the monitored phenomenon, keeping into account its evolution.
Consistently with the aforementioned setting, we introduce an approach which analyses data streams arriving from each sensor in parallel, assuming that they are generated by an unknown random process and that the covariance between observations depends only on the spatial distance among sensors.
We propose to represent each data stream as an unbounded histogram time series that is, a collection of histograms ordered over time. Such approach allows monitoring data distribution and supports fast computing as well as low memory occupation. Each histogram time series is the input of a local computing unit which gets a summarization of the stream. The results of this summarization are sent to a central computing unit which measures the spatial dependence among the streams in a time window and provides data predictions. The strategy is based on comparing histograms by the L^2-Wasserstein distance for histogram data. Wasserstein distances are metrics between probability distributions which have been used successfully for the comparison of complex objects since they consider information about the characteristics of compared distributions, such as location, variability and shape. The special case of L^2-Wasserstein distance provides also an elegant and intuitive notion of Fréchet mean and a useful measure of variance for distribution data. This paper uses the L^2-Wasserstein distance for detecting groups of similar histograms over time and for defining a variogram for histogram data, as a tool for measuring the spatial dependence among data streams

Keywords. *Data stream mining; Wasserstein distance; Spatial data.*

Radon time series analysis of Italian monitoring network

Siino Marianna[1*], Salvatore Scudero[1] and Antonino D'Alessandro[1]

[1] *Istituto Nazionale di Geofisica e Vulcanologia;*
marianna.siino@ingv.it, salvatore.scudero@ingv.it, antonino.dalessandro@ingv.it
[*] *Corresponding author*

Abstract. *The environmental radioactivity fluctuations of radon have been of recent interest. In the analysis, the relevant aspects concern identifying the main time series properties in terms of seasonality and autocorrelation structure and disentangling the effects of environmental factors that might control the radon concentrations. In this work, complementary methods are applied for detecting multi-seasonality, for evaluating the presence of long-range correlation and for describing the effect of weather variables on radon time series. We analyse radon measurements at different frequencies recorded in some Italian monitoring sites. The results indicate that there are sub-daily, daily and yearly persistent periodicities that are common for all the observed time series. However, the influence of the weather variables is strictly site-specific.*

Keywords. *radon; seasonality; time series; weather variables*

1 Introduction

There are several applications in geosciences based on the environmental radioactivity measurements, such as the radon (Rn^{222}) that is a gas with a short half-life of 3.8 days. It is often used as a potential earthquake precursor, as environmental tracer in hydrological settings and to study rock stress in volcanic and active zones.
The main difficulty of the use of radon as earthquake precursor is that the earthquake-related anomaly cannot be easily discriminated. As a matter of fact, radon variations are influenced by several factors such as flux of carrier gases, environmental and climatic variables, characteristics of the ground soil, tide, solar effect, etc (see [5] and references therein).
Following the explanation in [2], the analysis of radon time series presents several challenges because these measurements have a complex dynamic structure. Radon time series might present a non-stationary behaviour, multiple-seasonality, heteroscedasticity and long-range memory.
The soil radon emission is a topic of great concern in Italy where the mean annual concentration has been estimated to $70 Bq/m^3$, higher with respect the global annual mean of $40 Bq/m^3$. In this work, we investigate the main properties of 2-hours and daily radon time series of some Italian monitoring sites [1]. The identification of seasonality variations, both in the long-term and short-term, which are usually related to the environmental and climatic variables can allow filtering the radon signals and enhance the anomalies related to the geological processes.

2 Methods

Complementary methods are used to identify the main properties of the radon time series. First of all, the different time series are characterised according their underlying long-memory correlation structure. The presence of long-range memory is assessed estimating the Hurst exponent [4], if it holds the statistical dependence decays more slowly than an exponential decay. The long-range memory is quantify properly fitting autoregressive fractionally integrated moving average models, ARFIMA(p,d,q), [3]. This family of models allows to handle explicitly both the short-term and the long-term correlation structure.
Moreover, we perform a power spectral analysis in time-frequency domains using the continuous wavelet transformation (CWT) [6] identifying the (local) wavelet power spectrum (WPS). This quantity can be interpret as the local variance of the time series. The cross-wavelet analysis provides appealing information such as the similarity between the wavelet power spectrum of two series (computing the cross-wavelet power) and the series synchronicity (estimating the phase differences at certain periods). These quantities are computed for comparing the radon time series with the available weather variables (such as the temperature).

3 Main results

The analysed radon time series exhibit transient dynamics and the magnitude of the WPS is not constant over time fixing a specific frequency. The power spectral density calculated from the sub-daily series shows a main peak indicating a marked period at 1-day for all the selected sites. In the literature, this periodicity is mainly ascribed to the effect of the diurnal pressure and temperature cycle. The daily series shows a main cycle at about 1 year at each station and a subordinate cycle is present at about 180 days.
Depending on the site, the radon concentrations are positive or negative correlated with the temperature Furthermore, the phase differences in the band between 360 and 370 days show different values and also the overall in-phase/out-of-phase relationships is site-dependent. This is mainly due to the conditions resulting from local geological and environmental settings and the installation types of the monitoring station. For describing the auto-correlation structure, the ARFIMA models are estimated. The results suggest that there is a statistically significant evidence of long-range memory.

References

[1] Valentina Cannelli, Antonio Piersanti, Gianfranco Galli, and Daniele Melini. Italian radon monitoring network (iron): A permanent network for near real-time monitoring of soil radon emission in italy. *Annals of Geophysics*, 61(4):444, 2018.

[2] Reik V Donner, Stelios M Potirakis, Susana M Barbosa, José AO Matos, Alcides JSC Pereira, and Luis JPF Neves. Intrinsic vs. spurious long-range memory in high-frequency records of environmental radioactivity. *The European Physical Journal Special Topics*, 224(4):741–762, 2015.

[3] J. R. M. Hosking. Fractional differencing. *Biometrika*, 68(1):165–176, 04 1981.

[4] HE Hurst. Long-term storage capacity of reservoirs. *American Society of Civil Engineering*, 76, 1950.

[5] Marianna Siino, Salvatore Scudero, , Valentina Cannelli, Antonio Piersanti, and Antonino D'Alessandro. Multiple seasonality in radon time series: insights from continuous wavelet analysis. *submitted*, 2019.

[6] Christopher Torrence and Gilbert P Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological society*, 79(1):61–78, 1998.

Spatio-temporal earthquake clustering: insights and outlooks from Network Analysis

E. Varini[1,*], A. Peresan[2] and J. Zhuang[3]

[1] *Institute of Applied Mathematics and Information Technologies E. Magenes, National Research Council, Milano, Italy; elisa@mi.imati.cnr.it*
[2] *National Institute of Oceanography and Experimental Geophysics. CRS-OGS, Udine, Italy; aperesan@inogs.it*
[3] *Institute of Statistical Mathematics, Tokyo, Japan; zhuangjc@ism.ac.jp*
[*] *Corresponding author*

Abstract. *The seismic history of a region is characterized by its earthquake clusters, namely periods when the occurrence rate of earthquakes is higher than usual. Clustering in space and time is an essential key to understanding earthquake source mechanisms (fault geometry, rupture dynamics, status of the stress field, etc.), and several methodologies for cluster analysis have been proposed so far. However the definition of clusters is not univocal. Thus, for the identification of earthqauke clusters we consider two recent data-driven declustering algorithms, one based on nearest-neighbor distance and the other on a self-exciting point process.*
Since different classifications of earthquakes into main and secondary events can be obtained from different methods, we compare their performance by exploiting tools from Network theory. In particular, in order to highlight possible classification similarities/dissimilarities, the earthquake clusters obtained from both algorithms are represented as rooted trees, and their complexity is evaluated and compared through suitable centrality measures.

Keywords. *Earthquake clustering; Centrality measures; Nearest-neighbor distance; Stochastic declustering.*

The main purpose of this study is identifying suitable tools for the identification and quantitative robust characterization of earthquake clusters in a catalog. The data used for this analysis are extracted from the bulletins compiled at the National Institute of Oceanography and Experimental Geophysics; the catalog includes all events occurred in North-Eastern Italy and Western Slovenia, in the period between 1994 and 2018 and magnitude above 2.0.
The identification of earthquake clusters is performed by two declustering methods: the nearest-neighbor (NN) algorithm [2] and the stochastic declustering (SD) algorithm [3]. NN-method is based on the NN-distance η_{ij} between two earthquakes, which is a combination of the inter-occurrence time, the fractal dimension of the hypocenters distribution, and the Gutenberg-Richter law. Each event i is connected to its nearest-neighbor $j = argmin_k \eta_{ik}$; then earthquake clusters are clearly identified by removing all connections η_{ij} such that $\eta_{ij} > \eta_0$, for a selected threshold η_0 (more details in [1]). SD-method is based on the space-time epidemic-type aftershock sequence model, a branching point process defined by its intensity function $\lambda(t,x,y \mid \mathcal{H}_t) = \mu(x,y) + \sum_{i:t_i<t} \nu(t-t_i, x-x_i, y-y_i \mid \mathcal{H}_t)$ conditional on past history $\mathcal{H}_t$, where the rate $\mu(x,y)$ is due to spontaneous events and $\nu(t-t_i, x-x_i, y-y_i \mid \mathcal{H}_t)$ is the contribution to the hazard function due to events triggered by earthquake i. Connections between triggering and triggered events are established by thinning the point process according to probabilities proportional to $\mu(x,y)$ and

$\nu(t-t_i, x-x_i, y-y_i \mid \mathcal{H}_t)$ for all $t_i < t$; so earthquake clusters are immediately identified. Clearly, thinning simulation provides many possible cluster scenarios.
The earthquake clusters obtained from the two declustering algorithms are then compared, so as to identify classification similarities and differences. The analysis highlights a good agreement in the overall clusters identification. As an example, we illustrate the case of the cluster dominated by the 1998/04/12 earthquake, magnitude M5.6. Both methods identify the 1998 cluster: NN-cluster includes 720 events, while SD-cluster has 697 events. We observe that 677 events are consistently assigned to both clusters. Still, despite the large number of events identified by both methods, the hierarchical structure of the SD-cluster is more complex than that obtained from NN method. That is apparent from the representation as rooted trees of NN- and SD-clusters, but also according to some measures which express the way earthquakes (tree nodes) get organized within the cluster (tree). These measures are known as centrality measures in Network theory and, hereafter, only closeness centrality is considered. Given a rooted tree of n earthquakes, closeness centrality of earthquake x_i is defined by

$$close(x_i) = \frac{n-1}{\sum_{x_j} d(x_i, x_j)} \tag{1}$$

where $d(x_i, x_j)$ is the geodesic (shortest path) distance from x_i to x_j along the tree; if node x_j is not reachable from x_i, geodesic distance $d(x_i, x_j)$ is set equal to n. Closeness centrality ranges in $[0,1]$. According to closeness centrality, the most central node x^* is, on average, the closest node to the others. A global index, named closeness centralization, is defined by

$$C = \sum_x \left[close(x^*) - close(x)\right] / (n-1) \,, \tag{2}$$

which also ranges in $[0,1]$. High C values indicate simple structures inside the cluster, in which few nodes dominate others; on the contrary, small C values denote more complex hierarchical structures. As for the 1998 cluster, closeness centralization is 0.63 for the NN-cluster and 0.19 for the SD-cluster, confirming that SD-cluster has higher internal complexity than NN-cluster.

Acknowledgments. This work is supported by National grant MIUR, PRIN-2015 program, Prot. 20157 PRZC4: Complex space-time modeling and functional analysis for probabilistic forecast of seismic events. We also acknowledge financial support from Civil Defence of the Friuli Venezia Giulia Region.

References

[1] Peresan, G. & Gentili, S. (2018). Seismic clusters analysis in Northeastern Italy by the nearest-neighbor approach. *Physics of the Earth and Planetary Interiors* **274**, 88–104.

[2] Zaliapin, I. & Ben-Zion, Y. (2016). A global classification and characterization of earthquake clusters. *Geophysical Journal International* **207**, 608–634.

[3] Zhuang, J. (2006). Second-order residual analysis of spatiotemporal point processes and applications in model evaluation. *Journal of the Royal Statistical Society: Series B* **68**, 635–653.

Spatial Bayesian Hierarchical models to study the bacterium *Xylella fastidiosa*

M. Cendoya[1], J. Martínez-Minaya[2,*], V. Dalmau[3], A. Ferrer[3], D. Conesa[2], A. López-Quílez[2] and A. Vicent[1]

[1] *Centre de Protecció Vegetal i Biotecnología, Institut Valencià d'Investigacions Agràries. CV-315, Km 10, 7, 46113, Valencia. Spain; cendoya@alumni.uv.es, vicent_antciv@gva.es*
[2] *Departament d'Estadística i Investigació Operativa. Universitat de València. C/ Dr. Moliner 50. Burjassot. 46500. Valencia. Spain; Joaquin.Martinez-Minaya@uv.es, David.V.Conesa@uv.es, Antonio.Lopez@uv.es*
[3] *Servei de Sanitat Vegetal, Conselleria d'Agricultura, Medi Ambient, Canvi Climàtic i Desenvolupament Rural, Silla; dalmau_vic@gva.es, ferrer_ampmat@gva.es*
[*] *Corresponding author*

Abstract. *In this work, we present an analysis of the prevalence of X. fastidiosa in southern Italy and mainland Spain using hierarchical Bayesian models. The aim is to present different spatial models to better understand the epidemiological factors driving disease spread. The integrated nested Laplace approximation (INLA) was employed to obtain posterior and predictive distributions.*

Keywords. *INLA; spatial; Xylella fastidiosa*

1 Bayesian hierarchical models

In the last years, the use of complex statistical models has increased to improve our knowledge on the spread of diseases and the distribution of species. The complexity of these models makes the inferential and predictive processes challenging to perform. Bayesian statistics, which is based on the premise that both information and uncertainty can be expressed in terms of probability distributions, represents a good alternative in this context. But moreover, because complexity can be handled via hierarchical Bayesian models not difficultly. Usually, a Hierarchical Bayesian model can be expressed in three different levels: the likelihood, which represents the information given by the data; the prior distributions for the parameters and random effects; and the hyperpriors. However, obtaining the posterior distribution of the parameters governing the models is not straightforward, but, the Integrated Nested Laplace Approximation methodology (INLA) [2] is a powerful tool which facilitate the posterior distribution computation for a particular case of Bayesian Hierarchical models: the latent Gaussian models.

2 Xylella fastidiosa

In the last years, numerous epidemiological studies have been carried out to prevent, eradicate or contain plant disease spread under different scenarios. Here, we focus on diseases caused by the bacterium *Xylella fastidiosa* which was recently detected in the Mediterranean Basin. The olive quick decline,

caused by *X. fastidiosa* subsp. *pauca*, has devastated extensive areas in the south east Italy. An outbreak of almond leaf scorch, caused by *X. Fastidiosa* was detected in 2017 in Alicante province, eastern Spain. The introduction and spread of *X. fastidiosa* in other regions could cause yield losses and costly control measures not only in olive or almond but also in other economically important crops such as grapes, citrus, or stone fruits. Presence/absence data of the pathogen was avaliable in Alicante and in Apulia. Climatic and topographic variables, and geographical coordinates were also avaliable and employed for the analysis.

3 Spatial hierarchical Bayesian models for *Xylella fastidiosa*

As the data design was completely different, two different ways were employed to deal with them. With respect to Alicante, a spatial latent Gaussian model to deal with lattice data was used, meanwhile, in Apulia, we needed a geostatistical term to find a model that suited the data.

Likelihood

$$\begin{aligned} Y_i &\sim \text{Binomial}(n, \pi_i), \\ \text{logit}(\pi_i) &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{V}_i + \boldsymbol{U}_i, \end{aligned}$$

Latent Gaussian field

$$\begin{aligned} \beta_0, \ldots, \beta_M &\sim \text{N}(0, 10^4), \\ \boldsymbol{V_i} \sim \text{N}\left(\frac{1}{n}\sum_{i\sim j} V_i, \frac{1}{n_i \tau_{\boldsymbol{V}}}\right) &\quad , \quad \boldsymbol{U_i} \sim \text{N}\left(0, \tau_{\boldsymbol{U}}^{-1}\right), \end{aligned}$$

Hyperparameters

$$\tau_{\boldsymbol{V}}, \tau_{\boldsymbol{U}} \sim \text{LogGamma}(1, 5 \cdot 10^{-5}).$$

- $\boldsymbol{V}$ is a structured Besag spatial effect. Two locations are neighbors if the distance between them is 2.5Km. $\boldsymbol{U}$ is an unstructured random effect.

Likelihood

$$\begin{aligned} Y_i &\sim \text{Ber}(\pi_i), \\ \text{logit}(\pi_i) &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{W}_i, \end{aligned}$$

Latent Gaussian field

$$\begin{aligned} \beta_0, \ldots, \beta_M &\sim \text{N}(0, 10^4), \\ \boldsymbol{W} &\sim \text{N}(0, \boldsymbol{Q}^{-1}(r, \sigma_{\boldsymbol{W}})), \end{aligned}$$

Hyperparameters

$$\begin{aligned} \log(r) &\sim \text{N}(\mu_r, 10) \\ \log(\sigma_{\boldsymbol{W}}) &\sim \text{N}(\mu_{\sigma_{\boldsymbol{W}}}, 10). \end{aligned}$$

- $\boldsymbol{W}$ is a spatial effect with Matérn covariance function, r is referred to the range of the spatial effect and $\sigma_{\boldsymbol{W}}$ to the standard deviation of the spatial effect [1].

Results showed that it is as important to have in mind the presence of the spatial autocorrelation as the covariates, because it can change completely the modelization of a phenomenon. In addition, it was demonstrated that INLA is an efficient tool to deal with spatial data.

References

[1] Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73(4)**, 423-498.

[2] Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71(2)**, 319-392.

Improving the environmental impact of statistics - the need for multidisplinary collaboration.

K. Meissner[1*]

[1] *Finnish Environment Institute, SYKE, Programme for Environmental information, Survontie 9 A FI-40500 Jyväskylä; Kristian.Meissner@ymparisto.fi*
**Corresponding author*

Abstract. *The field of environmental statistics is fast evolving and is exploring continuously more sophisticated techniques for a more precise analysis of environmental data. Often however these novel solutions are either sought to very specific cases or generalized for a restricted set of similar cases. To the disappointment of statisticians striving for the wider uptake of their research, such solutions very seldomly are adopted by decision makers for routine use. Rather, environmental managers seem to "stubbornly" continue the use of outdated and less efficient approaches instead. While this can be a function of a different approach of managers towards knowledge production, it need not always be. In my talk I stress the importance of cooperation and dialogue with environmental managers for statisticians that seek to have a larger environmental impact. I particularly stress the need for statisticians to understand the nature and goals of the current managerial system, its history and limitations, as well as the need and willingness to relate theoretical concepts also to audiences of non-experts of their field.*

Keywords. *Novel statistical methods; Methods uptake; Environmental management.*

National-scale spatial modelling of landscape connectivity and stressor interactions on aquatic biodiversity

C. Wilkie[1], C. Miller[1], M. Scott[1,*], J. Belmont Osuna[1]

[1] *School of Mathematics and Statistics, University of Glasgow, UK; Craig.Wilkie@glasgow.ac.uk, Claire.Miller@glasgow.ac.uk, Marian.Scott@glasgow.ac.uk, Jafet.BelomontOsuna@glasgow.ac.uk*
[*] *Corresponding author*

Abstract. *Biodiversity in aquatic ecosystems is affected by stressors including pollution, and urban and agricultural land cover in their catchments, but also by landscape and hydrological connectivity. Connectivity here relates to the number or size of hydrological features near each water body, such as the length of rivers in the catchment. The interaction between connectivity and stressors is not well understood, though it is hypothesised that greater connectivity leads to healthier water bodies in the absence of stressors, but that this relationship can reverse for higher levels of stressors. Improved understanding of the effects on biodiversity of connectivity and stressors and their interactions is vital for conservationists and ecologists to better understand the distributions of aquatic species. Conservationists and ecologists have long collected data on species occurrence using routine surveys, citizen science, or more opportunistic visits to sites. Using rich existing data sources, and including the species groups of interest of dragonflies as well as aquatic plants (macrophytes) observed across the UK at more than 7,000 water bodies, or on a regular 1km grid, we present a statistical modelling framework to examine the ecological hypotheses.*

We propose a spatial modelling approach adapted to account for species detectability (where appropriate) and highly correlated connectivity and stressor covariates. The approach links ecological response data with landscape connectivity and stressor data on a national scale for the UK to address questions related to how multiple stressors (linked to urban or agricultural land use, acidification,) interact with connectivity to affect biodiversity, and how stress-response relationships are affected by differing measures of landscape connectivity.

We have adopted a Bayesian hierarchical approach, accounting for detection probability for hard-to-identify species that are likely to be under-recorded, with the resulting estimated species occupancy/richness modelled in response to connectivity and stressor metrics. Random forests are used to select from an extensive set of possible covariates, many of which are highly correlated, with the spatial data structure accounted for implicitly through including catchment covariates (such as altitude and spatial location).

Data and expert ecological advice were provided by University of Stirling, Centre for Ecology and Hydrology and British Trust for Ornithology. This work was funded as part of the NERC Hydroscape project (NE/N005740/1) (https://hydroscapeblog.wordpress.com/about/).

Keywords. *Biodiversity; Occupancy modelling; Hierarchical spatial models; Connectivity.*

An R-based widget for Bayesian disease mapping

Gardini[1], Greco[1,*] and Trivisano[1]

[1] *Department of Statistical Sciences "P. Fortunati", University of Bologna; aldo.gardini2@unibo.it, fedele.greco@unibo.it, carlo.trivisano@unibo.it*
[*] *Corresponding author*

Abstract. *Disease mapping encompasses a set of methodologies employed to describe the disease risk distribution over a study region. When the disease under study is rare, counts are heavily affected by random variability, and the estimates of the relative risk at the small-area level are unstable. The main aim of disease mapping studies is the identification of the underlying distribution of the risk.*
Several approaches have been proposed for modelling unstructured and spatially structured components and the Bayesian inferential framework is usually adopted. Among the different proposed models, the Besag-York and Mollié (BYM) model is one of the most employed in the literature. In the last years, several works focused on the specification of priors for the random effects variances: in particular, priors that can account for the structure of the effects are desirable.
We present a novel approach based on the solution of an integral equation that allows the user to have full control on prior specification, evaluating the prior marginal variance of the effects and their ratio. From a technical point of view, the equation is solved by means of the Mellin transform of the distribution of a quadratic form.
Despite the mathematical complexity of the prior specification procedure, to encourage the use of our methodology by practitioners, we developed an easily interpretable R interface. It allows to estimate the chosen BYM model through MCMC methods, after an aware prior specification for the variance components. Moreover, the interface provides tools for studying prior sensitivity and for performing posterior analyses.

Keywords. *Mellin transform, Bayesian hierarchical models, Spatial epidemiology*

A Hierarchical Bayesian Spatio-Temporal Model to Estimate the Short-term Effects of Air Pollution on Human Health

C. Grazian[1,*], L. Fontanella[1], L. Ippoliti[1] and P. Valentini[1]

[1] *Università degli Studi "Gabriele d'Annunzio"*
clara.grazian@unich.it, lara.fontanella@unich.it, luigi.ippoliti@unich.it, pvalent@unich.it
[*] *Corresponding author*

Abstract. *We introduce a hierarchical spatio-temporal regression model to study the spatial and temporal association existing between health data and air pollution. The model is developed for handling measurements belonging to the exponential family of distributions and allows the spatial and temporal components to be modelled conditionally independently via random variables for the (canonical) transformation of the measurements mean function.*

Keywords. *distributed lag models; spatio-temporal models; hierarchical models; air pollution*

There exists a huge statistical literature about the effect of air pollution on human health. In particular, The temporal relationship has been studied through time series models, developed in multi-sites frameworks, for example, in [Peng et al.(2009)].

The health data, which often come as mortality or morbidity rates or counts of hospital admissions for particular (e.g. respiratory) diseases, are, in general, collected as time series at different locations and estimating the health risks of air pollution involved considering a spatial relationship. However, the data analysed are often measured at different resolutions because they come from independent sources: environmental variables are generally registered at specific locations corresponding to environmental registration stations, while the health data correspond to areal units (the area associated with the hospital or the health district).

A simple approach to deal with this problem, often named change of support problem, is to average the environmental measurements recorded in the same health district; however, this procedure does not take into account any variability.

In this work, we propose a hierarchical spatio-temporal regression model, which is able to deal with the change of support problem, by changing the support of air pollution data (regressors) to achieve alignment with the health data measured at area level, following [**?**], in such a way that takes into account for spatio-temporal variation using a temporal autoregressive variable with spatially correlated innovations

We focus the interest on the effect of air pollution on the number of hospital admissions for respiratory diseases in the short period through a regression model that includes lagged exposure variables as covariates. The hierarchical model is a lag-distributed model, which assumes a general effect which is observed with variability at each location.

Acknowledgments. This work has been developed under the PRIN2015 supported-project: Environmental processes and human activities: capturing their interactions via statistical methods (EPHAStat), funded by MIUR (Italian Ministry of Education, University and Scientific Research) (20154X8K23-SH3).

References

[Peng et al.(2009)] R.D. Peng, F. Dominici, and L.J Welty. A Bayesian hierarchical distributed lag model for estimating the time course of risk of hospitalization associated with particulate matter air pollution. *Journal of the Royal Statistical Society: Series C*, 58:3–24, 2009.

[Sigrist et al.(2012)] Fabio Sigrist, Hans R Künsch, Werner A Stahel, et al. A dynamic nonstationary spatio-temporal model for short term prediction of precipitation. *The Annals of Applied Statistics*, 6(4):1452–1477, 2012.

Nonparametric Bayesian Approaches to Covariance Functions on Spheres

Porcu E.*[1], Bissiri P.[2], Tagle F.[3], Quintana F.[4]

[1]*School of Mathematics, Statistics and Physics, Newcastle University, E-mail: emilio.porcu@newcastle.ac.uk*

[2]*School of Mathematics & Statistics, Newcastle University*

[3]*Department of Applied and Computational Mathematics and Statistics, University of Notre Dame*

[4]*Pontificia Universidad Catòlica de Chile, Santiago, Chile*

[*]*Corresponding author*

Abstract. We provide a non-parametric spectral approach to the modeling of correlation functions on spheres. The sequence of Schoenberg coecients and their associated covariance functions are treated as random rather than assuming a parametric form. We propose a stick-breacking representation for the spectrum, and show that such a choicespans the support of the class of geodesically-isotropic covariance functions under uniform convergence. Further, we examine the rst order properties of such representation, from which geometric properties can be inferred, in terms of H older continuity, of the associated Gaussian random eld. The properties of the posterior, in terms of existence, uniqueness, and Lipschitz continuity, are then inspected. Our ndings are validated with MCMC simulations and illustrated using a global data set on surface temperatures.

Efficient estimation of nonstationary spatial covariance functions with application to high-resolution climate model emulation

Ying Sun

[1] *King Abdullah University of Science and Technology (KAUST), Saudi Arabia;E-mail: ying.sun@kaust.edu.sa*

Abstract. Spatial processes exhibit nonstationarity in many climate and environmental applications. Convolution-based approaches are often used to construct nonstationary covariance functions in Gaussian processes. Although convolution-based models are flexible, their computation is extremely expensive when the data set is large. Most existing methods rely on fitting an anisotropic, but stationary model locally, and then reconstructing the spatially varying parameters. In this study, we propose a new estimation procedure to approximate a class of non-stationary Matern covariance functions by local-polynomial fitting the covariance parameters. The proposed method allows for efficient estimation of a richer class of nonstationary covariance functions, with the local stationary model as a special case. We also develop an approach for a fast high-resolution simulation with non-stationary features on a small scale and apply it to precipitation data in climate model outputs.

The impact of earthquakes on demographic changes in Italy. A comparison between L'Aquila and the Emilia Romagna's cases

E. Ambrosetti[1], F. Licari[2], S. Miccoli[1] and C. Reynaud[3,*]

[1] *Dipartimento Metodi e Modelli per l'Economia, il Territorio e la Finanza –Sapienza Università di Roma ; elena.ambrosetti@uniroma1.it , sara.miccoli@uniroma.it*
[2] *Istat; licari@istat.it*
[3] *Dipartimento di Scienze Politiche – Università degli studi Roma Tre; cecilia.reynaud@uniroma3.it*

[*] *Corresponding author*

Abstract. *This paper analyses the environmentally-induced migration and displacement resulting from two earthquakes: Abruzzo (2009) and Emilia Romagna (2012). After a general critical overview of the social science literature on this topic, the main changes in the two migration systems are analysed looking at the different roots and trajectories of the forced human displacement that followed the two earthquakes. Additionally, we look at the long-run effects of earthquakes on population density growth across Italian municipalities affected by earthquakes during the period of 2002–2017.*

Moving from the fact that similar events may occurring in different contexts, may have different outcomes according to the specific vulnerability experienced by the territory, we assess the pre-disaster context and recovery period with the aim to offer a comparative analysis of the challenges related to post earthquake demographic movements and post-disaster resettlement.

The goal of our paper is twofold: first, we aim to understand how the two migration systems have been influenced by the pre-existent vulnerabilities and pre-quake social and institutional backgrounds before and after the hazard. Second, we investigate the long-run effects of

earthquakes on population density growth in Italy applying spatial regression models. In the analyses we will take into account the main economic trends in the earthquake's area. Relying on ISTAT data on the internal migration in Italy, we finally offer a general model of how environmental disaster might affect displacement and suggest the main challenges related to the post-disaster governance.

Keywords*: Environmental Disaster, Migration, Displacement, Population density, Spatial analysis*

1 Introduction

The primary focus in migration research has traditionally been labor migration. However, the relevance of

environmentally-induced migration and displacement, including those resulting from disasters and natural hazards, have increasingly drawn attention from academia and policy makers [1]. The reasons behind this increasing attention are included and not limited to the growing 'environmentalization' that has characterized both social science and the public debate in the past three decades, a period marked by a huge number of environmental global disasters that can be narrow down in three main fields:

-natural disasters (such as the 2004 South-east Asian tsunami, the 2009 L'Aquila earthquake, the 2012 Hurricane Sandy in the United States);

-human induced disasters (such as the 1986 Chernobyl nuclear disaster in the former USSR and the 1984 Bhopal disaster in India);

-and mixed disasters (such as the 2005 Hurricane Katrina that resulted in flooding when the levee collapsed in New Orleans).

Events like these have entailed a wide range of social, economic and demographic consequences, especially as concerns the rising phenomenon of the Environmentally-Induced Displacement, namely, the rapid, unforeseen option of last resort for those affected by an environmental hazard.

Over the last decade, at least ten key disasters had a significant long-term impact on the dynamics of long-lasting displacement. According to the estimates of international organizations, more than 1,7 million people were forced to displace following the Asiatic tsunami of December 2004. In August 2005, as a result of the Hurricane "Katrina" over the Gulf of Mexico, over 300,000 people were resettled, while the disaster caused losses estimated at over 86 billion dollars. In February 2010 more than 1,5 million people have been displaced in the aftermath of destructive 8.8 magnitude earthquake in Chile. 2011's earthquake in Haiti has deprived more than 1 million residents of homes. Furthermore, Japan's March 2011 earthquake, with its 9 magnitude and accompanying tsunami wave, had a significant impact on the dynamics of internal migration for Japanese nationals. According to the United Nations, a total of 590,000 were evacuated or displaced as a result of the quake and tsunami disaster, including more than 100,000 children [2].

In the last decade, geographical research into the causation of the disaster-related displacement, began to involve multi-scalar analysis with an emphasis on interaction across multiple spatial-temporal scales. This particular approach called for a rethinking of disasters from a political economic perspective, based on the high correlation between disaster predisposition, low local income and under-development, and leads to the conclusion that the root causes of disasters lay more in society than in nature.

In this theoretical approach the concept of "vulnerability" is crucial because it allows to go in the depth in the understanding of disasters, recognizing that disasters are not caused by a single agent but by the complex interaction of both environmental and social features and forces.

Although part of the research has been focused mainly on disaster and displacement, it is important to keep in mind that disasters do not affect all individuals, households and communities equally, and environmental hazard is not faced in the same way everywhere and by everyone [3]. Events that are rooted in nature such as earthquakes or tsunami, if they are of identical intensity, can produce diverse outcomes according to the characteristics of the communities and of the territory where they take place.

Differently from the wide literature that analyzes the role played by environmental disasters in shaping population movements in under-developed countries, this paper analyses the environmentally-induced migration and displacement resulting from two earthquakes occurred in the context of a developed country: more specifically, in the Italian regions of Abruzzo (2009) and Emilia Romagna (2012). The analysis will be lead at municipalities' level. After a general critical overview of the social science literature on this topic, the main changes in the two migration systems are analysed looking at the different roots and trajectories of the forced human displacement that followed the two earthquakes. Additionally, we look at the long-run effects of earthquakes on population density growth across Italian municipalities affected by earthquakes during the period of 2002–2017.

Moving from the fact that similar events may occurring in different contexts, may have different outcomes according to the specific vulnerability experienced by the territory, we assess the pre-disaster context and recovery period with the aim to offer a comparative analysis of the challenges related to post earthquake demographic movements and post-disaster resettlement.

The goal of our paper is twofold: first, we aim to understand how the two migration systems have been influenced by the pre-existent vulnerabilities and pre-quake social and institutional backgrounds before and after the hazard. Second, we investigate the long-run effects of earthquakes on population density growth in Italy applying spatial regression models. In the analyses we will take into account the main

economic trends in the earthquake's area. Relying on ISTAT data on the internal migration in Italy, we finally offer a general model of how environmental disaster might affect displacement and suggest the main challenges related to the post-disaster governance. It reviews the main socio-demographic and economic tendencies, with the aim to understand how the natural disaster shaped them and their impact on population growth and density.

2 Data and methods

We use Istat data referred to all the Municipalities which were hit by the earthquakes in Abruzzo and Emilia Romagna, respectively in 2009 and 2012. These municipalities are 57 in Abruzzo and 54 in Emilia Romagna. The L'Aquila earthquake of April 6 2009 killed 309 people. The earthquake of May 2012 hit an extensive area of Emilia Romagna.
For the demographic indicators we use data of Population registers for the period 2002-2017. For every year, we consider population at 1.1.t and some structure indicators, such as the proportion of 65 old people. Furthermore, we consider the population evolution in each municipality: natural increase and international and internal net migration, by distinguishing between Italians and foreigners' migration. We include also Istat data about the economic conditions.
A descriptive analysis is conducted to observe and describe the trends of demographic dynamics, in particular migration, in the territories which experienced the earthquake.
At a later stage, we exploit a spatial model based on the formulation proposed by Wang [4]. Through this model we aim to investigate about the effects of earthquakes on population density growth in these territories, by looking at the difference between the periods before and after the earthquake. The application of a spatial model permits to take into account the influence that space can have on some not observable variables which influence demographic events. Demographic and economic variables are inserted in the model as other factors which can affect population distribution beyond a natural disaster as the earthquake. Indeed, we hypothesis that demographic, social, economic factors are interrelated with change in population distribution before and after the earthquake. Population structure affect demographic and social trends. Also pre-existing in-flow and out-flow towards specific territories can shape migration movements also after the earthquakes. Regarding migration, we expect also that internal migration of Italians and foreigners follow different trajectories and evolution, connected also with reconstruction.

3 Preliminary results

Italy is one of the most earthquake-prone countries in the world. The L'Aquila earthquake destroyed a large part of the built environment, as well as essential infrastructure networks. The earthquake and the relief and recovery operations have changed the territories. In this area, the earthquake exhausted populations which were already experiencing demographic and economic challenges [6]. The Emilia Romagna earthquake hit a densely and wealthy populated area. These municipalities represented one of the most productive areas in Italy, contributing in a significant way to regional and national economy.
The L'Aquila earthquake has changed the demographic distribution across the territories which were hit by damages and losses, but these changes are complex and multifaceted. After the L'Aquila earthquake, no massive movements occurred across the municipalities. Part of L'Aquila population was resettled in other crater' municipalities, which indeed recorded a positive net migration (both internal and international) in the period after the earthquake. However, as outlined by Petrei & Petrei [5], people may have chosen to keep their administrative residence in L'Aquila even if they moved to another municipality, in order to receive benefits connected with the recovery measures prepared by the Italian government. Indeed, aids are allocated to the people on the basis of their administrative residence. The recovery period is characterized by a strong increase of out-flows from the analyzed municipalities to municipality of other provinces, within the Abruzzo region and outside the Abruzzo region. In the period after the disaster, a slight increase of in-flows in some municipalities occurred.
After the earthquake in Emilia Romagna, movements from hit municipalities toward Lombardy and

Veneto regions increase. Movements from municipalities located in Modena province towards municipalities located in other provinces (outside and inside Emilia Romagna region) decrease. The earthquake has partially affected the population of these municipalities determining few changes and challenges connected with change in population distribution.

The analysis conducted at municipality level show different evolutions according both to the population dimensions and the impact of the earthquake. The spatial analysis at this level permits to summarize different evolutions and aspects of these changes, pointing out homogenous territorial areas, by controlling for the spatial autocorrelation effect.

Preliminary results outlined that similar events may have different consequences across various spatial context. Pre-existing characteristics of the context can affect the consequences of a natural disaster such as an earthquake. Indeed, demographic and economic vulnerability of a territorial context may contribute to shape different trajectories of displacement following natural disaster.

References

[1] Ambrosetti, E., & Petrillo, E. R. (2016). Environmental disasters, migration and displacement. Insights and developments from L'Aquila's case. *Environmental science & policy* **56**, 80-88.

[2] Terminski, B. (2012). *Environmentally Induced Displacement: Theoretical Frameworks and Current Challenges*. University of Liège.

[3] Piguet, E. (2010). Linking climate change, environmental degradation, and migration: a methodological overview. Wiley Interdiscip. Rev.: Clim. Change 1 (4), 517–524.

[4] Wang, C. (2019). Did natural disasters affect population density growth in US counties?. *The Annals of Regional Science* **62(1)**, 21-46.

[5] Petrei, F. & Petrei, F. (2010). Ad un anno dal terremoto a L'Aquila: dinamiche migratorie e sociali nel post-sisma. *Rivista Italiana di Economia, Demografia e Statistica* **64(4)**, 239-246.

[6] Pesaresi, C. (2017). I comuni del cratere sismico, prima e dopo il terremoto del 2009. Considerazioni sui movimenti demografici in atto. *Semestrale di Studi e Ricerche di Geografia*, **24(1)**, 69-84.

An INLA spatio-temporal model for zero-inflated marine plastic litter abundance

C. Calculli[1,*], A. Pollice[1], I. Paradinas[2], L. Sion[3] and P. Maiorano[3]

[1] *Department of Economics and Finance, University of Bari, Largo Abbazia S. Scolastica 53, 70124 Bari, Italy; crescenza.calculli@uniba,it, alessio.pollice@uniba.it*
[2] *Asociación Ipar Perspective, C/Karabiondo, 48600 Sopela, Spain; paradinas.iosu@gmail.com*
[3] *Department of Biology, LRU CoNISMa, University of Bari, via E. Orabona 4, 70125 Bari, Italy; letizia.sion@uniba.it, porzia.maiorano@uniba.it*
[*] *Corresponding author*

Abstract. *The marine plastic litter pollution is a worldwide growing environmental concern. Despite its negative effects on marine ecosystems, the phenomenon is still not well-known at global and local scale. This work aims at assessing the spatio-temporal distribution of plastic litter amounts found at the sea-floor in a region of the central Mediterranean (Ionian sea). Inspired by species distribution models, we propose a two-parts model to accommodate the excess of zeros and the spatio-temporal correlation characterizing abundance monitoring data. A common spatial effect that links the plastic abundances and the probabilities of occurrences is implemented with the Stochastic Partial Differential Equation approach extended to a non-stationary barrier model. The INLA methodology allows to efficiently perform Bayesian inference to fit complex spatio-temporal models including effects of environmental covariates and enables to investigate the assemblages of plastic litter over the study region.*

Keywords. *Hurdle model; Integrated nested Laplace approximation (INLA); Marine ecology*

1 Introduction

The extensive use of disposable plastic items with a cultural propensity of increasingly over-consuming, discarding and littering, has become a lethal combination for marine ecosystems. While it is true that not all marine garbage is plastic, recent literature clearly indicates that plastic is the dominant material littering seas [3]. Despite the known negative effects of plastic accumulation on habitats and communities [7], the magnitude of the marine litter pollution has yet to be deeply investigated at global and local scale. In the Mediterranean basin, the majority of plastic comes from terrestrial inputs (coastal human populations and rivers) as well as from discarded fishing gears and shipping traffic that greatly contribute to the overall litter of sea bottoms [2]. Even though devoted to the study of benthic and demersal fish stocks, experimental bottom trawl surveys regularly carried out in the Mediterranean (MEDITS program), represent a valuable source of information about wastes. Litter categories might be seen as special abiotic items or additional "species", caught by trawl nets together with real marine species. Therefore, the analysis of litter abundances is not far different from species distribution modeling (SDM). In this spirit, we propose to analyze the spatio-temporal distribution of plastic abundances at the sea-floor in a region of the central Mediterranean (Ionian sea), investigating environmental factors that might affect litter assemblage

dynamics at local scale. To this end, a suitable modeling approach is proposed in order to accommodate the zero-inflation and the spatio-temporal dependence characterizing abundance data. Hurdle models assume that zero and non-zero data are generated by two independent processes, one for the probabilities of occurrences and the other for the intensities of the non-zero responses. For semi-continuous data, the commonly assumed independence between the two processes might be considered unsuitable since it neglects the relation between abundances and probabilities of occurrences. It is far more realistic to expect low abundance intensities associated with low probabilities of occurrences and vice versa. According to [5, 6], a framework with a common latent gaussian random field (GRF) that allows the two processes to be related, is considered. To model the spatio-temporal structure, we refer to the stochastic partial differential equations (SPDE) approach that consists in defining the continuously indexed Matérn Gaussian field (GF) as a discretely indexed spatial random process (GMRF) using piece-wise linear basis functions defined on a triangulation of the domain of interest. SPDE provides a representation of the whole spatial process that varies continuously in the considered domain [4]. The SPDE approximation is currently implemented using the Integrated Nested Laplace approximation (INLA) [8] via the R-INLA package (`http://www.r-inla.org`) designed to make Bayesian inference accessible for a large class of latent Gaussian models. Using this approach, accurate approximations of the posterior marginals are obtained with computationally efficient tools alternative to cumbersome MCMC simulations. The INLA method is an efficient approach for modeling spatio-temporally correlated data with excessive zeros considering the effects of environmental covariates affecting the dynamics dynamics of plastic litter assemblages in the central Mediterranean.

2 Data description

Monitoring data are collected during experimental trawl surveys conducted from 2013 to 2016 in the North-Western Ionian Sea as part of MEDITS project (MEDiterranean International Trawl Surveys) activity. The same 70 depth-stratified hauls are carried out between 10 and 800 m in depth every year, summing to 280 hauls in 4 years. For each survey at every haul location, plastic density indices (N/km^2) are obtained scaling the number of collected items to the swept surface unit (1 km^2). Data concerning *sea currents* and *fishing activities*, characterized by different spatio-temporal support, were first aligned and then considered to investigate environmental factors affecting the distribution of plastic litter densities over the study region. In particular, we investigate the effects of: 1) the superficial eastward and northward sea water velocities (U and V) retrieved from the Copernicus Marine Environment Monitoring Service (`http://marine.copernicus.eu/`); 2) the daily average transit (MVH) and fishing time (MFH) for 3 vessel types (Drifted Longlines, Fixed Gears and Trawlers) provided by the International Global Fishing Watch organization (`https://globalfishingwatch.org/`).

3 The Bayesian spatio-temporal hurdle model

Density-based spatio-temporal abundance processes are commonly measured in R_+, resulting in semi-continuous non-negative datasets. A convenient representation of such datasets is obtained by Hurdle models that consider two independent sub-processes: an occurrence process and a conditional-to-presence continuous process. Let y_{st} and z_{st} being the occurrence and the conditional-to-presence abundance sub-processes at time t ($t = 1,\dots,T$) and location s ($s = 1,\dots,n_t$), then for the plastic litter densities we get:

$$z_{st} \sim Ber(\pi_{st})$$
$$logit(\pi_{st}) = \beta_0^{(1)} + \sum_{i=1}^{p} \beta_i^{(1)}(x_{ist}) + V_{st}(1) \quad \text{with} \quad s = 1,\dots,70; \quad t = 1,\dots,4 \tag{1}$$
$$y_{st} \sim Gamma(a_{st}, b_{st})$$
$$log(\mu_{st}) = \beta_0^{(2)} + \sum_{i=1}^{p} \beta_i^{(2)}(x_{ist}) + V_{st}(2) \quad \text{with} \quad s = 1,\dots,70; \quad t = 1,\dots,4 \tag{2}$$

where π_{st} and $\mu_{st} = a_{st}/b_{st}$ are modeled through the logit and logarithm links, respectively. In linear predictors, the $\beta_0^{(1)}$ and $\beta_0^{(2)}$ represent the intercepts, $\beta^{(1)}$ and $\beta^{(2)}$ are fixed effects of spatio-temporally varying covariates x_i (U, V, MVH and MFH). In order to account for information shared by related occurrence and abundance sub-processes, the V_{st} components in Eq.(1)-(2) are assumed common and modeled by a Gaussian field through the SPDE approach [4] extended to the non-stationarity case to ensure that the spatial correlation seeps around coastlines [1], thus $\boldsymbol{V} \sim N(0, Q(\kappa,\tau))$ and $(log(\kappa), log(\tau)) \sim MVN(\mu,\rho)$ where the covariance function of the spatial effect Q depends on a range effect (κ) and a total variance parameter (τ). Due to the short time series of 4 years available, a first order autoregressive (AR1) model is preliminary adopted to capture the temporal effect. The lack of information leads to assign vague prior distributions to all model parameters, as implemented by default in INLA.

4 Main results

Figure 1 reports estimated fixed effects for the positive plastic densities of the Hurdle model in Eq.(2). Although none of the estimated effects is relevant in affecting the intensity of plastic densities, some insights come from these results. In particular, the left panel shows estimated higher densities with respect to current towards north-western direction. The right panel highlights, instead, higher densities associated with increasing transit time and the opposite negative effect for the fishing time covariate (the more fishing activity, the lesser sampling litter items observed in the area). The shared spatio-temporal field, represented in Figure 2, shows changes concerning densities and presence of plastic hot-spots from year to year. The spatial variation clearly distinguishes hauls with higher densities of plastic in the same areas of Sicily, Calabria and Apulia over the years. On the other hand, the available time series is too short for the identification of a temporal trend.

This work represents a starting point for the analysis of spatio-temporal structured monitoring data for multiple litter categories. To combine environmental information from multiple sources and different temporal sampling frequencies, further development includes the identification of a suitable modeling framework to face the change of support problem.

Acknowledgments. C. Calculli and A. Pollice were supported by the PRIN2015 project EphaStat - *Environmental processes and human activities: capturing their interactions via Statistical methods*, funded by the Italian Ministry of University and Research (MIUR). The MEDITS surveys have been carried out within the Data Collection Framework. For fishing activities data availability, we thank W. Zupa, COISPA Tecnologia & Ricerca, Bari, Italy.

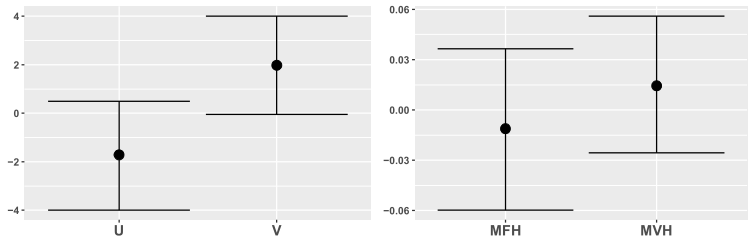

Figure 1: Estimated covariates effects for positive plastic densities (bars represent 95%CI)

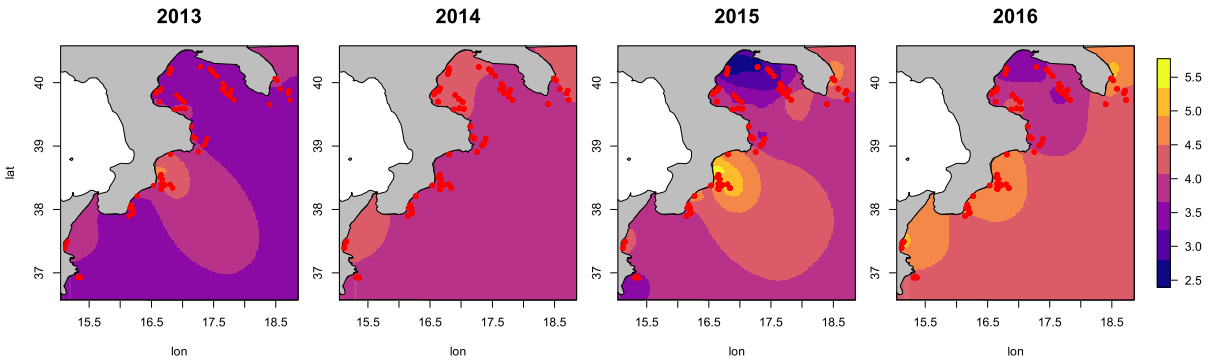

Figure 2: Yearly posterior means of spatial effect

References

[1] Bakka, H., Rue, H., Fuglstad, G-A., Riebler, A,, Bolin, D., Illian, J., Krainski, E., Simpson, D., Lindgren, F. (2018) Spatial modeling with R-INLA: A review. *WIREs Comput Stat*, 10:e1443

[2] Consoli, P., Romeo, T., Angiolillo, M., Canese, S., Esposito, V., Salvati, E., Scotti, G., Andaloro, F., Tunesi, L.(2019). Marine litter from fishery activities in the Western Mediterranean sea: The impact of entanglement on marine animal forests. *Environmental Pollution* **249**, 472–481

[3] Law, K. L.(2017). Plastics in the marine environment. *Annual review of marine science* **9**, 205–229

[4] Lindgren, F., Rue, H., Lindstrom, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach (with discussion). *Journal of the Royal Statistical Society*, **73**, 423–498

[5] Paradinas, I., Conesa, D., López-Quílez, A., Bellido, J. M. (2017). Spatio-Temporal model structures with shared components for semi-continuous species distribution modelling. *Spatial Statistics*, **22**, 434–450

[6] Quiroz, Z. C., Prates, M. O., Rue, H. (2014) A Bayesian approach to estimate the biomass of anchovies off the coast of Perù. *Biometrics*, **71**, 208–217

[7] Rochman, C. M., Browne, M. A., Underwood, A. J., Van Franeker, J. A., Thompson, R. C., Amaral-Zettler, L. A. (2016). The ecological impacts of marine debris: unraveling the demonstrated evidence from what is perceived. *Ecology*, **97**, 302–312

[8] Rue, H., Martino, S., Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by usingintegrated nested Laplace approximations. *Journal of the Royal Statisitcal Society, Series B*, **71**, 319–392

Semiautomatic dictionary-based tweet classification for measuring well-being

M. Cameletti[1], S. Fabris[1,*], S. Schlosser[2] and D. Toninelli[1]

[1] *University of Bergamo (IT);* michela.cameletti@unibg.it, silvia.fabris@unibg.it, daniele.toninelli@unibg.it.
[2] *University of Göttingen (D);* stephan.schlosser@sowi.uni-goettingen.de.
[*] *Corresponding author*

Abstract.

In this paper we describe a semiautomatic dictionary-based approach to filter tweets talking about specific topics. In particular, we are interested in studying the citizen well-being (WB) and, for this aim, we select tweets pertaining two WB dimensions such as environment and health. For this purpose, we use dictionaries containing keywords selected by analyzing tweets published by some Official Social Accounts linked with the two topics. The selected tweets are then processed in order to estimate the sentiment of the population with respect to such specific subjects. In this paper, we present some preliminary results for Great Britain (GB) using tweet collected on the whole country for the six-weeks period from 2019/01/14 to 2019/02/24. The results show that, on the one hand, our dictionary-based classification approach reaches good levels of accuracy, sensitivity and specificity; on the other hand, we assess the spatial variability across GB of the two dimensions we are studying by means of the tweets sentiment analysis.

Keywords. *Twitter; sentiment analysis; health; environment; spatial analysis*

1 Introduction

Measuring individual well-being (WB) is extremely challenging due to the multidimensional, country-specific and latent nature of this concept. Standard approaches for WB evaluation are mainly based on large-scale surveys and rely on several multivariate statistical methods. For example, in [2] the Structural Equation Modelling (SEM) approach is applied to data collected through the European Social Survey (ESS) in order to estimate WB in 16 European countries. In particular, by means of the SEM, the paper identifies seven latent dimensions linked to WB: social involvement, country attachment and trust, discrimination, income perception, environment, health and work status. These dimensions are then used to estimate the WB level of the considered European countries.

Nowadays, in the era of social networks, a huge quantity of data is available that can potentially be used to estimate WB. The collection and the analysis of such data is still an evolving research field that can lead to some advantages: data obtained from the Internet are available at lower costs, in shorter times and are easier to collect than traditional survey-based data. Nevertheless, the collection of this new kind of data is also challenging, from the methodological point of view. Social networks, for example, are used for many different purposes and shared posts can be about personal opinions, ideas, goals and events, but they can also include a huge amount of advertisements and news. For this reason, the identification and

the selection of truly informative data can be a difficult task.

The purpose of this research is to test a reliable semiautomatic dictionary-based method to filter, by topic, posts shared on Twitter, in order to retrieve tweets related to the seven WB dimensions defined in [2]. In particular, in this paper, we focus on two dominions: "health" (HEA) and "environment" (ENV). For this purpose, we define one specific dictionary for each dimension by using a list of keywords chosen by analyzing tweets published by a selected list of Official Social Accounts (OSA).

In this paper we aim to evaluate the reliability of our semiautomatic method using tweets posted in Great Britain (GB). Moreover, since the selected tweets are geolocalized, we are able to study the spatial variability across GB of tweets sentiment, which is a proxy of the two selected WB dimensions.

2 Data and methods

Our data include tweets posted in GB from 2019/01/14 to 2019/02/24 and collected through the "circle approach" described in [1]. Just 1% of these tweets provides GPS coordinates; nevertheless, the "circle approach" allows us to geo-localize all tweets, making it possible to associate each text to one of the circles covering GB (see Figure 1). After having preliminary removed bots (i.e. accounts which post more than 3 times a day, on average), we analyzed 22,193,719 text messages, an average of 26,233 tweets for each circle. In cleaning the corpus of the selected tweets we try to keep as much information as possible by replacing, with equivalent-meaning expressions, htmls, emoji, slangs, word elongations and money symbols; moreover, we keep hashtags and quotations in the tweet text.

The first objective of our analysis was to define two dictionaries to filter among all the available tweets the ones pertaining ENV and HEA. For this purpose, we analyzed several Twitter OSA linked to each of the two dimensions and belonging to no-profit associations, news media, research institutes and intergovernmental organizations[1]. In particular, we collected all the available tweets posted up to 2019/04/04, obtaining 38,604 tweets about ENV and 38,651 about HEA. Our analysis relies on the four following steps.

(1) OSA tweets cleaning. All tweets are cleaned, removing url links, html code, non-ascii and special characters. **(2)** Setting up dictionaries. We select the top trending hashtags used by the selected OSA. These hashtags are keywords used in the OSA description (e.g. #UseLessPlastic for the @LessPlasticUK account) or created by OSA for particular international events (e.g. #PlasticFreeFriday). Among the top trending ones, we selected the most used hashtags: 60 about ENV and 11 for HEA. These thresholds are set, by topic, in order to avoid the selection of acronyms and of too general words (such as for example #women, #plastic, #brexit, #ue, etc.). The selected hashtags constitute the basis of the dictionaries; we then further enrich the list of keywords by analyzing the corpus of the OSA tweets. In particular, we include in the dictionaries the most common bigrams and trigrams (excluding the ones containing stop words). In order to choose combinations of words widely used, we take into account, for each dictionary, bigrams occurred at least 65 times and trigrams occurred at least 35 times. Finally, we manually review the selected hashtags, bigrams and trigrams in order to exclude expressions too generic and not related to the studied WB dimensions (e.g. "facebook live", "fake account", "million people"). The obtained ENV dictionary contains 61 hashtags and 53 bigrams/trigrams; the HEA dictionary includes 11 hashtags and 62 bigrams/trigrams. **(3)** Tweets selection. Using the dictionaries obtained at step (2), among the 22+ millions of tweets collected for GB, we select the ones containing at least one keyword included in the dictionaries. We obtained 35,250 tweets about ENV and 50,610 about HEA. **(4)** Sentiment analysis. These selected tweets are processed by using the AFINN and of the BING lexicon-based approaches. In

[1] List of selected OSA. For ENV: *@climateprogress, @ClimateReality, @friends_earth, @Greenpeace, @GreenpeaceUK, @LessPlasticUK, @PlasticPollutes, @UNEnvironment, @UNFCCC, @World_Wildlife, @WWF, @WWFScotland*. For HEA: *@bbchealth, @CDCgov, @goodhealth, @NBCNewsHealth, @NYTHealth, @EverydayHealth, @NIHClinicalCntr, @theNCI, @CDC_HIVAIDS, @CDCSTD, @CDC_Cancer, @cdchep.*

particular, the first method ranks each word with a score included between -5 and +5 (where negative and positive scores indicate negative and positive sentiment, respectively). The BING lexicon associates -1, 0 and +1 to negative, neutral and positive words, respectively. The total sentiment score of each tweet is computed as the sum of the scores linked to all the words included in the tweet. Thus, for each tweet we obtain two scores (coming from the AFINN and the BING lexicon, respectively).

3 Results

In order to evaluate the performance of our semiautomatic filtering, we selected randomly 100 tweets for each dimension and we manually classify them into two categories: "related" and "non-related" to the topic. To compare the dictionary-based classifications with the manual one (our benchmark), we compute the following performance indexes: *accuracy* (**A**, i.e. the percentage of correctly classified tweets), *sensitivity* (**SE**, i.e. the percentage of topic related tweets correctly identified by the classifier) and *specificity* (**SP**, i.e. the percentage of topic non-related tweets correctly not identified by the classifier). For ENV we obtained the following values for the performance indexes: A=98, SE=97, SP=99. For HEA the observed performance indexes were equal to: A=97.5, SE=95, SP=100. All values denote a very good performance of our semiautomatic dictionary-based approach in filtering tweets related to a given topic.

The sentiment analyses based on AFINN and BING lexicon do not show any significant difference in mean; for this reason (and for space concerns) we present here just the results obtained using BING lexicon. Figure 1 shows the spatial distribution by quartiles of the standardized average sentiment score for ENV (left) and HEA (right). The spatial correlation Moran's index is equal to 0.06 (p=0.04) for ENV and to -0.04 (p=0.90) for HEA, showing absence of relevant spatial correlation between circles. This could be caused by the fact that we are averaging sentiment values across the whole time period and some circles, located in remote zones, may contain a very small number of tweets. Moreover, we implemented a correlation test between sentiment results concerning ENV and HEA, in order to check if the two dimensions influence each other: the correlation is very low and nonsignificant (corr. coeff.=0.09; p=.013).

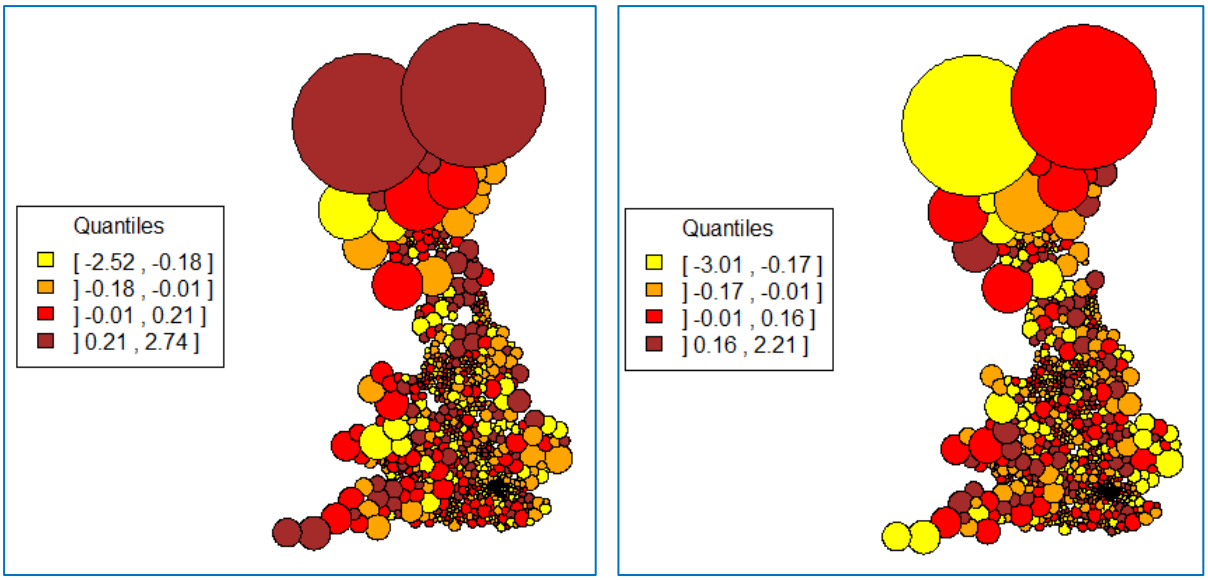

Figure 1: Standardized average tweets sentiment for ENV (left) and HEA (right) (BING lexicon; spatial distribution by quartiles).

4 Conclusions

The aim of this paper is twofold. On the one hand, we want to check if the dictionaries set up by means of our prototypal methodology (see sect. 2) are able to select tweets linked to two WB dimensions, i.e. ENV and HEA. On the other hand, our target is to obtain estimates of the level of two WB dominions over the GB. This second step was based on the sentiment analysis of tweets selected by means of our dictionaries and on the use of the AFINN/BING lexicon-based approaches (the two methods did not show any significant difference in the obtained results).

For what pertains the first objective, all the classification performance indexes (A, SE and SP) show that the both dictionaries perform very well. In fact, they are able to identify posts that have a content actually linked with the dimensions of interest and to exclude the ones which have a content linked to different topics. Our unsupervised topic-classification method is not still fully automatic because we have to select the thresholds, separately for each topic, for the number of hashtags and bigram/trigrams that have to be considered (see Sect. 2) in order to remove keywords not concerning the dimensions of interest. As future research, we intend to include in the dictionary acronyms and to gradually and continuously increment the number of included keywords by periodic analysis of the OSA accounts.

With respect to the second objective, we are able to predict separately for each GB circle the sentiment of the population using tweets about two topics related to WB (ENV and HEA). We expected to find some correlation between neighboring circles and between the two topics within circles. Our findings did not confirm our initial expectations. Moreover, the time lag we took into consideration is probably too short to get a clear picture, that would probably emerge by studying a longer period of analysis (some months of tweets).

Future work will extend the current framework to a different spatial resolution (we will aggregate circles at the area level, by using the so-called NUTS statistical regions of UK); moreover, we will take into account the distribution of sentiment across time, since tweets distribution can be susceptible to daily/weekly events. The final aim of this research is to compare the tweet-based estimation of WB with some benchmark estimates provided by large-scale survey projects, such as the ESS. To this regard, we will consider not only a longer time interval for tweets collection, but also the full set of seven WB dimensions.

References

[1] Schlosser, S., Cameletti, M., Toninelli, D. (2019). Optimized strategies for enhancing the territorial coverage in Twitter data collection. in Keusch, F., Struminskaya, B., Hellwig, O., Stützer, C. M., Thielsch, M., Wachenfeld-Schell, A. (Eds.): 21st General Online Research Conference. Proceedings. Cologne 2019 (link: https://www.gor.de/gor19/index.php?page=browseSessions&form_session=48&presentations=show).

[2] Toninelli, D., Cameletti, M. (2018). Is Structural Equation Modelling Able to Predict Well-being?. In Abbruzzo, A., Brentari, E., Chiodi, M., Piacentino, D. (Eds.). Book of short Papers SIS 2018, 1529-1534, Pearson (link: https://it.pearson.com/content/dam/region-core/italy/pearson-italy/pdf/Docenti/ISTITUZIONI%20-%20HE%20-%20PDF%20-%20SIS%20V2.pdf).

On the reliability of some tests on type of non-separability and type of class of covariance models

C. Cappello[1], S. De Iaco[1,*], M. Palma[1] and D. Posa[1]

[1] *University of Salento, Via per Monteroni, Complesso Ecotekne, Lecce, Italy; claudia.cappello@unisalento.it, sandra.deiaco@unisalento.it, monica.palma@unisalento.it, donato.posa@unisalento.it*
[*] *Corresponding author*

Abstract. *In the literature, various tests for evaluating some characteristics of space-time covariance functions, such as symmetry and separability, are widely used. Recently, in case of rejection of the separability hypothesis, innovative tests have been proposed for evaluating the type of non-separability of space-time covariance functions and testing some well known classes of non-separable positive or negative covariance function models.*
In this paper a study on simulated data is proposed in order to assess the performance of the tests on the type of non-separability and on the classes of covariance functions.

Keywords. *Space-time covariance; Non-separability; Type of non-separability test; Test on class of covariance function models.*

1 Introduction

Apart from various tests for checking some second order properties such as symmetry and separability (Mitchell et al., 2005, 2006; Li et al., 2007, 2008), a test for the type of non-separability, as well as a statistical test for some classes of space-time covariance models were proposed in Cappello et al. (2018) and implemented in the R package `covatest`, which is available on CRAN (De Iaco et al., 2017). These tests help researchers in choosing the appropriate class of spatio-temporal covariance function model, for the spatio-temporal data analyzed.
In this paper a study on simulated data is proposed in order to assess the performance of the test on the type of non-separability and on some well known classes of covariance functions (i.e., the Gneiting and product-sum class of covariance functions).

2 Simulation study

In the following simulation study, the reliability of the test statistics defined in Cappello et al. (2018) and implemented in the R package `covatest` (De Iaco et al., 2017) is discussed.
Zero-mean simulated space-time realizations have been used to test the null hypotheses formulated on different types of non-separability and types of class of models. In particular the product-sum model

$C(\mathbf{h},u) = k_1 C_s(\mathbf{h})C_t(u) + k_2 C_s(\mathbf{h}) + k_3 C_t(u)$, $k_1 > 0, k_2 \geq 0, k_3 \geq 0$, and the Gneiting model $C(\mathbf{h},u) = \sigma^2 \left(\frac{1}{(b|u|^{2\alpha}+1)^{\tau}}\right) \cdot exp\left(-\frac{a||\mathbf{h}||^{2\gamma}}{(b|u|^{2\alpha}+1)^{\gamma\beta}}\right)$, $\beta \in [0,1], \tau \geq \beta d/2$, have been used to generate simulated space-time data regularly distributed over a range of grid sizes (spatial grids of dimensions 9×9 and 15×15), with temporal lengths $|T_n|$=600 and $|T_n|$=1000. The product-sum model has exponential marginals with spatial and temporal effective ranges equal, respectively, to 3 and 20 and parameters $(k_1, k_2, k_3) = (0.5, 0.3, 0.2)$. On the other hand the Gneiting model has marginals with linear behaviour near the origin (with smoothness parameters γ and α equal to 0.5) and $(a, b, \beta, \tau, \sigma^2) = (1, 0.75, 1, 1, 1)$ (which correspond to spatial and temporal marginals that decay approximately at 3 and 20, respectively).
These two classes of covariance function models have been considered to produce alternative simulations since they present two different types of non-separability, i.e., the product-sum class is negative non-separable and the Gneiting class is positive non-separable.
The goodness of the tests have been evaluated through the study of 900 simulations, obtained through a Gaussian-related program, that is the sequential simulation algorithm, based on the above mentioned classes of covariance function models.
The simulation study focused on the analysis of the empirical size and power of the tests for different grid sizes, temporal lengths and classes of models. In particular, for the test on the type of non-separability

- data sets simulated through the product-sum model, which is uniform negative non-separable, have been considered to compute (a) the empirical size through the frequency of rejecting the uniform negative non-separability ($Fr\{R_{H_0^{(-)}}|H_0^{(-)}\}$), and (b) the empirical power through the frequency of rejecting the uniform positive non-separability ($Fr\{R_{H_0^{(+)}}|H_1^{(-)}\}$);
- data sets simulated through the Gneiting model, which is uniform positive non-separable, have been used to compute (a) the empirical size through the frequency of rejecting the uniform positive non-separability ($Fr\{R_{H_0^{(+)}}|H_0^{(+)}\}$), and (b) the empirical power through the frequency of rejecting the uniform negative non-separability ($Fr\{R_{H_0^{(-)}}|H_1^{(+)}\}$).

Moreover, an indirect way of approximating the power of the test has been also proposed. It has been evaluated how large is the p-value for the decision of non-rejection (when the null hypothesis is true), therefore the frequencies of non-rejecting the null hypotheses with large p-values (greater than 0.9), denoted with $Fr\{\bar{R}_{H_0^{(+)}}|H_0^{(+)}; \text{p-values} > 0.9\}$ and $Fr\{\bar{R}_{H_0^{(-)}}|H_0^{(-)}; \text{p-values} > 0.9\}$, have been computed. For the tests on the type of class of models the size and power have been also determined. In particular

- Gneiting model-based data have been used to compute (a) the empirical size, through the frequency of rejecting the same Gneiting model ($Fr\{R_{H_0^{\text{Gn}}}|H_0^{\text{Gn}}\}$) and (b) the empirical power through the frequency of rejecting the null hypotheses formulated on two different classes, such as the product-sum model ($Fr\{R_{H_0^{\text{PS}}}|H_1^{\text{Gn}}\}$) and the integrated product model ($Fr\{R_{H_0^{\text{IP}}}|H_1^{\text{Gn}}\}$). Note that the power of the test on the Gneiting class has been analyzed with respect to the product-sum model, which is negative non-separable and the integrated product model, which is positive non-separable;
- product-sum model-based data have been used to determine (a) the empirical size through the frequency of rejecting the product-sum model, ($Fr\{R_{H_0^{\text{PS}}}|H_0^{\text{PS}}\}$) and (b) the empirical powers, the frequency of rejecting the null hypotheses formulated on the Gneiting class ($Fr\{R_{H_0^{\text{Gn}}}|H_1^{\text{PS}}\}$) and the integrated product class ($Fr\{R_{H_0^{\text{IP}}}|H_1^{\text{PS}}\}$), which are positive non-separable.

In addition, the frequencies of non-rejecting the null hypotheses (when it is true) with large p-values (greater than 0.9) have been computed as an indirect way to approximate the power of the test. These

frequencies are denoted with $Fr\{\bar{R}_{H_0^{\text{Gn}}}|H_0^{\text{Gn}};\text{p-values} > 0.9\}$ and $Fr\{\bar{R}_{H_0^{\text{PS}}}|H_0^{\text{PS}};\text{p-values} > 0.9\}$.
As stated above, the testing procedure has been applied to the zero-mean simulated data sets, obtained for different alternatives in terms of grid size, temporal length and class of models; spatial couples and temporal lags at distances 1 and 2 have been considered for the tests. The results of the test on the type of non-separability, i.e., the empirical size with respect to the nominal level 0.05 and power are given in Tab. 1. Looking at the results, it is clear that the size of the test (p_1 and p_1') is close to the nominal level and the power (p_3 and p_3') approaches 1 as the grid size and temporal length increase; similarly for the approximated powers (p_2 and p_2'), measured in terms of frequencies of non-rejecting the null hypotheses (when it is true) with large p-values (greater than 0.90). These results confirm the reliability of the test and that there is strong confidence in rejecting the null hypothesis of negative/positive non-separability when the alternative hypothesis is valid, as well as in failing to reject the null hypothesis when the null hypothesis is valid.

		Negative non-separable model-based simulations			Positive non-separable model-based simulations		
		p_1	p_2	p_3	p_1'	p_2'	p_3'
9×9	$\lvert T_n\rvert = 600$	0.080	0.093	0.747	0.080	0.093	0.693
	$\lvert T_n\rvert = 1000$	0.053	0.107	0.933	0.040	0.107	0.920
15×15	$\lvert T_n\rvert = 600$	0.067	0.107	0.893	0.067	0.093	0.813
	$\lvert T_n\rvert = 1000$	0.040	0.120	0.987	0.053	0.120	0.973

Table 1: Values of the empirical size and power for the tests on type of non-separability for data simulated through a uniform negative non-separable model ($p_1 = Fr\{R_{H_0^{(-)}}|H_0^{(-)}\}$, $p_2 = Fr\{\bar{R}_{H_0^{(-)}}|H_0^{(-)};\text{p-values} > 0.9\}$ and $p_3 = Fr\{R_{H_0^{(+)}}|H_1^{(-)}\}$) and through a uniform positive non-separable model ($p_1' = Fr\{R_{H_0^{(+)}}|H_0^{(+)}\}$, $p_2' = Fr\{\bar{R}_{H_0^{(+)}}|H_0^{(+)};\text{p-values} > 0.9\}$ and $p_3' = Fr\{R_{H_0^{(-)}}|H_1^{(+)}\}$).

The results for the test on the type of class of models are show in Tab. 2. The size (p_1 and p_1') is close to the nominal level for each option, while the empirical power (p_3, p_4 and p_3', p_4') supports the rejection decision of the null hypothesis when it is false. The approximated powers (p_2 and p_2') are consistent with respect to the nominal frequency of the non-rejection decision of the null hypothesis (when it is valid) with p-value greater than 0.9. Note also that the powers and the approximated powers of all alternatives are nearly equivalent when the temporal length is equal to 1000. Moreover, the tests have greater power when the underlining data are generated by a covariance model characterized by a different type of non-separability with respect to the class of model under the null hypothesis (i.e., p_3 is greater than p_4).

		Gneiting model -based simulations				Product-sum -based simulations			
		p_1	p_2	p_3	p_4	p_1'	p_2'	p_3'	p_4'
9×9	$\lvert T_n\rvert = 600$	0.080	0.080	0.827	0.707	0.067	0.080	0.893	0.853
	$\lvert T_n\rvert = 1000$	0.040	0.107	0.973	0.773	0.040	0.107	0.987	0.973
15×15	$\lvert T_n\rvert = 600$	0.067	0.093	0.947	0.720	0.053	0.107	0.907	0.880
	$\lvert T_n\rvert = 1000$	0.053	0.133	0.987	0.813	0.040	0.120	1.000	0.987

Table 2: Values of the empirical size and power for the tests on type of class of covariance function models for data simulated through the Gneiting model ($p_1 = Fr\{R_{H_0^{\text{Gn}}}|H_0^{\text{Gn}}\}$, $p_2 = Fr\{\bar{R}_{H_0^{\text{Gn}}}|H_0^{\text{Gn}};\text{p-values} > 0.9\}$, $p_3 = Fr\{R_{H_0^{\text{PS}}}|H_1^{\text{Gn}}\}$ and $p_4 = Fr\{R_{H_0^{\text{IP}}}|H_1^{\text{Gn}}\}$) and through the product-sum model ($p_1' = Fr\{R_{H_0^{\text{PS}}}|H_0^{\text{PS}}\}$, $p_2' = Fr\{\bar{R}_{H_0^{\text{PS}}}|H_0^{\text{PS}};\text{p-values} > 0.9\}$, $p_3' = Fr\{R_{H_0^{\text{Gn}}}|H_1^{\text{PS}}\}$ and $p_4' = Fr\{R_{H_0^{\text{IP}}}|H_1^{\text{PS}}\}$).

From the results in Tab. 1 and 2 it is evident that (a) the grid size does not significantly affect the size of the test, which is around the nominal level even if the series length is equal to 600 and (b) the power

increases as temporal length increases.
Finally, the product-sum and the Gneiting model-based simulations have been also used to evaluate how rapidly the empirical distribution function of the test statistic on the type of non-separability and on the type of class of models converges in distribution to a normal and a Chi-square, respectively, according to the results of the multivariate delta theorem of Mardia et al. (1979) and Li et al. (2007). In particular, the temporal length of simulated data increases from 400 up to 1000, with increments of 200 time points for each step and the Kolomogorov-Smirnov tests have been applied for comparing the observed cumulative distribution functions of the test statistics with the corresponding theoretical distributions. The empirical distribution function of the test statistic on the type of non-separability rapidly converges to a normal distribution, even when the temporal length is greater than 400. The p-values for the Kolomogorov-Smirnov tests support the non-rejection of the null hypothesis for all options. The same goes for the empirical distribution function of the test statistic on the class of models. In this case, the p-values, which support the non-rejection of the null hypothesis for all options, are greater than 0.8 when the temporal length is greater than 800 and approach 1 when the temporal length is equal to 1000.

3 Conclusions

In this paper the reliability of the statistical tests for checking different forms of non-separability and some classes of space-time covariance function models was analyzed. The empirical results obtained through the simulated data confirm the goodness of these tests and can stimulate their use in the applications.

References

Cappello, C., De Iaco, S., Posa, D. (2018). Testing the type of non-separability and some classes of space-time covariance function models. *Stoch. Environ. Res. and Risk Assess.* **32**, 17–35.

De Iaco, S. and Posa, D. (2013). Positive and negative non-separability for space-time covariance models. *J. Stat. Plan. Infer.* **143**(2), 378–391.

De Iaco, S., Cappello, C., Posa, D., Maggio, S. (2017). Covatest: Tests on properties of space-time covariance functions. The Comprehensive R Archive Network, https://CRAN.R-project.org/package=covatest, 1–18.

Li, B., Genton, M. G. and Sherman, M. (2007). A Nonparametric Assessment of Properties of Space-Time Covariance Functions. *Journal of the American Statistical Association*. **102**(478), 736–744.

Li, B., Genton, M. G. and Sherman, M. (2008). On the asymptotic joint distribution of sample space-time covariance estimators. *Bernoulli*. **14**(1), 228–248.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). Multivariate Analysis, New York: Academic Press. 521p.

Mitchell, M., Genton, M. G. and Gumpertz, M. (2005). Testing for separability of space-time covariances. *Environmetrics*. **16**, 819–831.

Mitchell, M. W., Genton, M. G., and Gumpertz, M. L. (2006). A Likelihood Ratio Test for Separability of Covariances. *Journal of Multivariate Analysis*. **97**, 1025–1043.

Rodrigues, A., Diggle, P. J. (2010). A Class of Convolution-Based Models for Spatio-Temporal Processes with Non-Separable Covariance Structure. *Scand. J. Stat.* **37**(4), 553–567.

Modeling hydrologic data by means of re-parametrization of Beta-Singh-Maddala distribution

F. Condino[1,*], F. Domma[1] and S. Franceschi[1]

[1] *Department of Economics, Statistics and Finance "Giovanni Anania", University of Calabria; francesca.condino@unical.it, filippo.domma@unical.it, sara.franceschi@unical.it*
[*] *Corresponding author*

Abstract. *In this paper we propose a new parametrization of the four-parameters Beta-Singh-Maddala distribution suitable for the context of hydrologic studies. With this aim, we reparameterize the Beta-Singh-Maddala distribution to make its parameters directly interpretable in terms of measures much more relevant for their practical use than the classical shape, location and scale parameters of the parametric families generally used for modeling hydrologic events. Moreover, in order to evaluate how climatic or physic characteristics could affect these measures, we will express them as functions of a set of covariates that could have an effect separately and/or simultaneously.*

Keywords. *Regional flood frequency; Extreme events; Regression.*

1 Introduction

Nowadays, the occurrence and impact of hydrologic extreme events and their possible relationship with climate change represents a crucial theme for human life. In this context, the statistics of extremes plays a fundamental role and represents a strategic tools for the assessment of current and future exposure to risks. The improvement of models for better exploring observed extremes, with an emphasis on flood quantiles, are strategic activities for the assessment of current and future exposure to risks and the development of some appropriate tools for accurately describing some particular phenomena are crucial. With this aim, Domma and Condino [1] propose the use of two new four-parameters distribution functions, namely the Beta-Dagum and Beta-Singh-Maddala distributions which seem to possess the main suitable features to be used for the analysis of extreme events. Furthermore, following Domma et al. [2], in this paper we consider the reparameterization of the four-parameters Beta-Singh-Maddala (Beta-SM4) distribution in order to make its parameters directly interpretable in terms of median, return level and return period. So, the new reformulation allows of making the parameters of the distribution directly interpretable in terms of measures much more relevant for their practical use than the classical shape, location and scale parameters of the parametric families used as in modeling hydrologic events. Moreover, in order to evaluate how climatic or physic characteristics could affect these measures, we will express them as functions of a set of covariates that could have an effect separately and/or simultaneously.

2 Reparameterization of the Beta-SM4 distribution

The Beta-SM4 distribution, in its original parameterization, has the following distribution function (*df*):

$$F_{Beta-SM4}(x;\boldsymbol{\xi}) = \left[1-(1+\gamma_2 x^{\gamma_3})^{-\gamma_1}\right]^a \tag{1}$$

where $\boldsymbol{\xi}' = (\gamma_1,\gamma_2,\gamma_3,a)$, with $a>0$ and $\gamma_i>0$ for $i=1,2,3$. The probability density function (*pdf*) is given by $f_{Beta-SM4}(x;\boldsymbol{\xi}) = a\left[F_{SM}(x;\boldsymbol{\gamma})\right]^{a-1} f_{SM}(x;\boldsymbol{\gamma})$. where $F_{SM}(x;\boldsymbol{\gamma})$ and $f_{SM}(x;\boldsymbol{\gamma})$ are, respectively, the *df* and *pdf* of SM distribution.

Following the proposal of [2], we consider the possibility of reformulate the Beta-SM4$(\gamma_1,\gamma_2,\gamma_3,a)$ in terms of new parameters, $I_j, j=1,...,4$, that are indicators describing some peculiarities of hydrologic data distribution and such that there exist a one-to-one transformation of the kind $I_j = g_j(\gamma_1,\gamma_2,\gamma_3,a), j=1,...,4$, in order to have a unique solution in terms of $\gamma_1,\gamma_2,\gamma_3$ and a:

$$\begin{cases} \gamma_1 = \gamma_1(I_1,I_2,I_3,I_4) \\ \gamma_2 = \gamma_2(I_1,I_2,I_3,I_4) \\ \gamma_3 = \gamma_3(I_1,I_2,I_3,I_4) \\ a = a(I_1,I_2,I_3,I_4) \end{cases}. \tag{2}$$

Substituting the solution (2) in (1), it is possible to obtain the expressions of the *cdf* in terms of the chosen indicators. Analogously to the generalized linear models, the measures I_j are related to the set of covariates, $\mathbf{x}_{j,i}$, by $I_{j,i} = h_j(\mathbf{x}_{j,i},\boldsymbol{\gamma}_j)$, where $h_j(.)$ are suitable link function.

2.1 Formulation in terms of median and return level

In this paper, the original parameters are substituted by the following one-to-one transformation $(\gamma_1,\gamma_2,\gamma_3,a) \mapsto (\tau, me, x_0, a)$ where $\tau = \frac{1}{\gamma_1}$, me is the median of distribution, given by

$$me_{Beta-SM4}(p) = \gamma_2^{-\frac{1}{\gamma_3}}\left[(1-0.5^{\frac{1}{a}})^{-\frac{1}{\gamma_1}}-1\right]^{\frac{1}{\gamma_3}}$$

and x_0 is the return level, corresponding to a pre-fixed return period π_{x_0}, i.e.

$$x_0(\pi_{x_0}) = \gamma_2^{-\frac{1}{\gamma_3}}\left\{\left[1-\left(1-\frac{1}{\pi_{x_0}}\right)^{\frac{1}{a}}\right]^{-\frac{1}{\gamma_1}}-1\right\}^{\frac{1}{\gamma_3}}.$$

After simple algebra, we obtain

$$\begin{cases} \gamma_1 = \frac{1}{\tau} \\ \gamma_2 = [(1-0.5^{1/a})^{-\tau}-1]\cdot me^{-\frac{\log\left\{\left[1-\left(1-\frac{1}{\pi_{x_0}}\right)^{1/a}\right]^{-\tau}-1\right\}-\log\left\{[1-0.5^{1/a}]^{-\tau}-1\right\}}{\log x_0-\log me}} \\ \gamma_3 = \frac{\log\left\{\left[1-\left(1-\frac{1}{\pi_{x_0}}\right)^{1/a}\right]^{-\tau}-1\right\}-\log\left\{[1-0.5^{1/a}]^{-\tau}-1\right\}}{\log x_0-\log me} \\ a = a \end{cases}. \tag{3}$$

It is immediate, from (1), to obtain the new expression of *cdf* of Beta-SM4 *r.v.* in terms of median and return level.

3 Application

In order to show the usefulness of the proposed model, we consider the data from Hydroclimatic Data Network of U.S. Geological Survey (USGS). In particular, we focuses our attention on annual peak flows considered in [3] for basins located in Texas Region. We consider the area of drainage (A, in km^2), the slope of main channel (S, in m/km), the mean elevation of drainage basin over MSL (E, in m) and the length of main channel from divide to gauge (L, in km), as covariates and a return period of 50 years, as in the cited paper. Therefore, in this example, we investigate the direct effects of the covariates on the median and the 50-years return level, using the reparametrization in (3) and choosing $\exp(.)$ as the log-link function. We consider no covariate effects on the remaining parameters. Table 1 reports maximum likelihood estimates (MLEs) of the parameters, the corresponding standard errors, t-tests and p-values, related to the four indicators $I_1 = \tau$, $I_2 = me$, $I_3 = x_0$ and $I_4 = a$. As we expected, many of the considered variables seem to have a significant influence on annual peak flows, in particular with regards to its median value and 50-years return level. Finally, Figure 1 shows empirical and fitted distribution obtained inserting the sample mean for covariates in the expressions of me and x_0, to simulate the case of a representative basin which well summarizes the Texas Region peak flows data.

Covariate	Estimate	SE	t	p-value
	$\tau = \exp(\boldsymbol{x}_{1,i}, \gamma_1)$			
Intercept	-2.118	1.412	-1.500	0.1336
	$me = \exp(\boldsymbol{x}_{2,i}, \gamma_2)$			
Intercept	8.252	5.260×10^{-2}	156.875	< 0.001
A	1.833×10^{-5}	3.597×10^{-6}	5.097	< 0.001
E	-4.111×10^{-4}	3.412×10^{-5}	-12.049	< 0.001
L	4.869×10^{-3}	2.040×10^{-4}	23.865	< 0.001
	$x_0 = \exp(\boldsymbol{x}_{3,i}, \gamma_3)$			
Intercept	10.224	8.468×10^{-2}	120.737	< 0.001
S	1.368×10^{-2}	5.655×10^{-3}	2.419	0.0156
L	3.090×10^{-3}	2.690×10^{-4}	11.486	< 0.001
	$a = \exp(\boldsymbol{x}_{4,i}, \gamma_4)$			
Intercept	5.00×10^{-1}	4.347×10^{-1}	1.150	0.250

Table 1: MLEs of the parameters (log-likelihood: -29103.63)

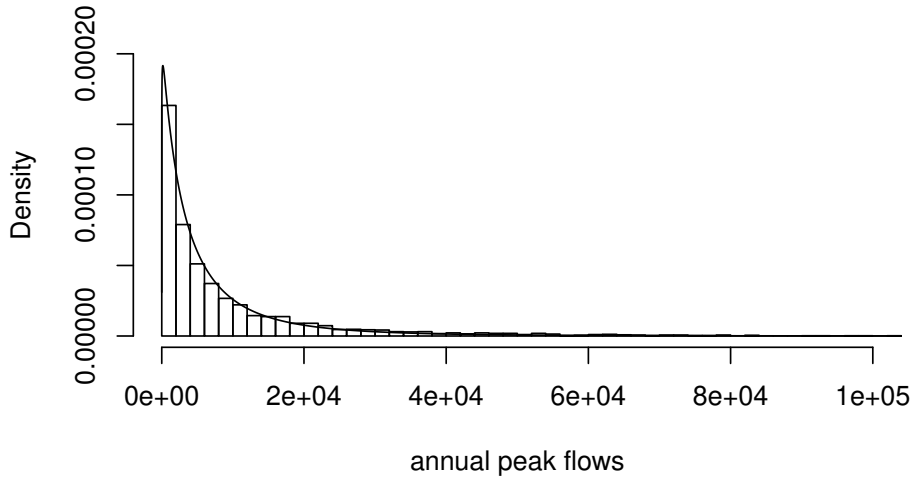

Figure 1: Empirical and fitted Beta-SM4 distribution

References

[1] Domma, F. and Condino F. (2017) Use of the Beta-Dagum and Beta-Singh-Maddala distributions for modeling hydrologic data. *Stochastic Environmental Research and Risk Assessment* **31**, 799–813.

[2] Domma, F., Condino, F., Giordano S. (2018). A New Formulation of the Dagum Distribution in terms of Income Inequality and Poverty Measures. *Physica A: Statistical Mechanics and its Applications* **511**, 104–126.

[3] Latraverse, M., Rasmussen, P.F., Bobe, B. (2002). Regional estimation of flood quantiles: Parametric versus nonparametric regression models. *Water Resources Research* **38**, 1-1-1-11.

Multivariate geostatistical tools for time series modeling and prediction

S. De Iaco[1], S. Maggio[1], M. Palma[1,*] and D. Pellegrino[1]

[1] *University of Salento, Via per Monteroni, Complesso Ecotekne, Lecce, Italy; sandra.deiaco@unisalento.it, sabrina.maggio@unisalento.it, monica.palma@unisalento.it, daniela.pellegrino@unisalento.it*
[*] *Corresponding author*

Abstract. *Modeling and prediction multivariate geostatistical techniques can be successfully applied to study the temporal behaviour of several correlated time series. In particular, in the time domain, by using variogram-based tools the analyst can easily a) identify trend and periodicity which characterize each time series, b) fit a properly Multivariate Linear Temporal (MLT) model to multiple correlated time series, c) predict the variable of interest (primary variable) at some time points after the last available observation, by taking into account the fitted model as well as the auxiliary information coming from the secondary variables. In this paper the convenience of performing a complete analysis of multiple correlated time series on the basis of geostatistical tools is illustrated through a case study concerning three environmental variables. As regards the computational aspects, a new version of the GSLib Cokb3d routine has been implemented for prediction purposes.*

Keywords. *Multivariate linear temporal model; Temporal cross-variogram; Temporal cokriging.*

1 Introduction

In time series analysis, the methodology developed by Box and Jenkins (1976) is commonly applied to detect the most suitable model which reasonable might describe the temporal evolution of the analyzed process. Then, the model is used in the prediction stage. On the basis of the Box-Jenkins approach, the auto-correlation and the partial auto-correlation functions (ACF and PACF, respectively), as well as the cross-correlation function (CCF) have a crucial role in the modeling selection, indeed through the visual inspection of the sample ACF, PACF and CCF, the most appropriate model for the process under study can be identified. In the multivariate context, several approaches have been proposed in order to model the joint relationships between multiple time series. Among the different types of models (Reinsel, 2003), the most common are Vector AutoRegressive, AutoRegressive-MovingAverage or AutoRegressive-Integreted-MovingAverage models in the presence of exogenous variables, the models based on a transferring function and the co-integrated models (mainly used in the economic field). However, for the analysis of multiple correlated time series, multivariate geostatistics could also be a very useful approach, nevertheless it is widely applied to investigate, through the matrix variogram, the spatial direct and cross-correlation which characterize the variables of interest and make predictions at unsampled locations of the spatial phenomena.
In this paper the use of variogram-based multivariate geostatistical techniques have been enlarged to analyze multiple time series, in order to identify trends and periodicity exhibited by the data, model the temporal evolution of the variables and make temporal predictions for the primary variable using the auxiliary information coming from the secondary available variables. The computational aspects have

been tackled by implemented a new version of the GSLib *Cokb3d* routine (Deutsch and Journel, 1998) which allows the analyst to use the fitted model in the cokriging system and define appropriate temporal search neighborhoods for prediction purposes.

2 Variogram-based modeling and prediction for multiple time series

In time series analysis, the measurements of $p \geq 2$ correlated variables, at different time points or intervals, can be considered as a finite realization of a real-valued Multivariate Random Process (MRP) $\{Z(t), t \in T \subseteq \mathbb{R}\}$, with $\mathbf{Z}(t) = [Z_1(t), Z_2(t), \ldots, Z_p(t)]^T$. Under second-order stationarity, the mean vector of $\mathbf{Z}$ exists and does not depend on t, and the $(p \times p)$ variogram matrix $\boldsymbol{\Gamma}$ defined for two MRP, $\mathbf{Z}(t)$ and $\mathbf{Z}(t')$, exists and depends on the temporal separation h, i.e.:

$$\boldsymbol{\Gamma}[\mathbf{Z}(t), \mathbf{Z}(t')] = E\left\{[\mathbf{Z}(t) - \mathbf{Z}(t')][\mathbf{Z}(t) - \mathbf{Z}(t')]^T\right\} = \boldsymbol{\Gamma}(h) = [\gamma_{ij}(h)],$$

where $h = (t - t')$ and $\gamma_{ij}(h),\ i, j = 1, \ldots, p,$ are the cross-variogram (if $i \neq j$) between the random variables $Z_i(t)$ and $Z_j(t+h)$ and the direct variogram (if $i = j$) of the i-th random variable. In the multivariate context, the empirical temporal variogram matrix can be modelled through the most used model in the spatial multivariate analysis, namely the *Linear Coregionalization Model* (Wackernagel, 2003). In this case, a Multivariate Linear Temporal (MLT) model $\boldsymbol{\Gamma}(h) = \sum_{l=1}^{L} \mathbf{B}_l\, g_l(h)$, can be developed, where $\mathbf{B}_l = [b_{ij}^l], i, j = 1, \ldots, p$, are $(p \times p)$ positive-definite matrices and $g_l(h)$, $l = 1, \ldots, L$, are basic temporal variograms identified at $L \geq 2$ temporal variability scales. Before modeling the temporal direct and cross-correlation among the variables, the direct and cross-variograms are estimated as follows:

$$\widehat{\gamma}_{ii}(r) = \frac{1}{2\,|N_i(r)|} \sum_{N_i(r)} [Z(t+h) - Z(t)]^2; \quad \widehat{\gamma}_{ij}(r) = \frac{1}{2|N_{ij}(r)|} \sum_{N_{ij}(r)} \left[(Z_i(t+h) - Z_i(t)) \cdot (Z_j(t+h) - Z_j(t))\right],$$

where $N_i(r) = \{t, t+h \in H_i, i = 1, \ldots, p$, such that $|r-h| < \delta\}$, $|N_i(r)|$ is the cardinality of this last set, $N_{ij}(r) = \{t, t+h \in (H_i \cap H_j), i, j = 1, \ldots, p,\ i \neq j$ such that $|r-h| < \delta\}$, and $|N_{ij}(h)|$ is its cardinality, r is the temporal lag, δ is the tolerance and H_i is the set of the measurements for the i-th time series, $i = 1, \ldots, p$. As pointed out in De Iaco et al. (2013), the variogram could be efficiently applied in time series analysis (Haslett, 1997), since it can describe a wide class of stochastic processes (the class of intrinsic stochastic processes), and also its estimation does not require the knowledge of the expected value of the associated stochastic process. Moreover, the variogram is a useful tool to identify trend and periodicity exhibited by data and to make temporal predictions for the variable of interest. For a second-order stationary MRP $\mathbf{Z}$, a linear prediction of the time series under study at an unsampled time point $t \in T$, can be obtained by using the well-known cokriging predictor (Wackernagel, 2003). In this case, the temporal cokriging predictor is expressed as: $\widehat{\mathbf{Z}}(t) = \sum_{\alpha=1}^{N} \boldsymbol{\Lambda}_\alpha(t)\mathbf{Z}(t_\alpha)$, where $t_\alpha \in T$, $\alpha = 1, \ldots, N$, are the sampled points and $\boldsymbol{\Lambda}_\alpha(t)$, $\alpha = 1, \ldots, N$, are the $(p \times p)$ matrices of the weights which are determined so that the above temporal predictor is unbiased and efficient (Journel and Huijbregts, 1981). The ordinary cokriging requires only the knowledge of the model for the matrix variogram and it is used when the expected value of the process is constant and unknown.

3 A case study

Two atmospheric variables, i.e. daily Temperature ($^\circ C$) and daily Wind Speed (m/sec), as well as PM_{10} daily concentrations ($\mu g/m^3$), measured from 2010 to 2013, at one survey station belonging to the environmental network of the Apulian Protection Agency and located in Brindisi district (South of Italy),

have been analyzed by multivariate geostatistical tools. The survey station, called "Torchiarolo", is very close to the thermoelectric power station "Enel-Federico II", and all the surrounding area is considered being at high risk of air pollution, especially during the winter and during long period of low rainfall. PM_{10} is strongly influenced by meteorological conditions. In particular, the horizontal transport, dispersion and resuspension of PM_{10} are mainly determined by Wind speed: low values of this meteorological variable are related to high PM_{10} concentrations (Harrison et al., 1997; Sayegh et al., 2014). Moreover, temperature is considered as one of the strongest predictors of PM_{10} concentrations. High values of this air pollutant are measured in winter, specially when the difference between maximum and minimum daily temperature is large (Perez et al., 2002). In the following sections, the advantages and the flexibility of the multivariate geostatistical techniques to analyze the times series under study will be pointed out.

3.1 Exploratory analysis and modeling

Exploratory data analysis has clearly highlighted that: a) PM_{10} daily concentrations present an annual periodicity at 12 months, b) Temperature and Wind Speed are characterized by opposite seasonal behaviors: in winter time, Temperature decreases, while Wind Speed increases; on the other hand, in summer time Temperature increases and Wind Speed decreases, c) over the four-year span (from 2010 to 2013), the PM_{10} daily values have exceeded 243 times the threshold value (50 $\mu g/m^3$) fixed by the national law for the human health protection; in particular, during the summer, PM_{10} does not exceed this limit value, instead, in winter time changes in the lower layer of the troposphere determine PM_{10} stagnation and consequently high concentrations of PM_{10}. The sample direct temporal variograms for the analyzed time series highlight the presence of periodicity for all variables (Fig. 1). These periodic components

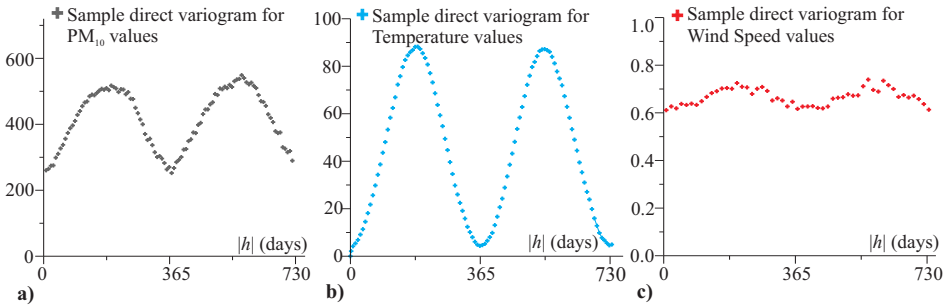

Figure 1: Sample temporal direct variograms of a) PM_{10}, b) Temperature and c) Wind Speed daily averages.

have been factored out from the observed data through moving average and monthly averages techniques. Hence, the residuals have been considered as a realization of a second-order stationary MRP $\mathbf{Z}(t) = [Z_1(t), Z_2(t), Z_3(t)]^T$, with $t \in T \subseteq \mathbb{R}$, and have been used in the following steps of the analysis. After computed the sample direct and cross temporal variograms of the residuals, two different scales of temporal variability have been detected through the visual inspection of the sample variograms. Hence, the following MLT model has been fitted to the sample matrix variogram:

$$\boldsymbol{\Gamma}(h) = \mathbf{B}_1\, g_1(h) + \mathbf{B}_2\, g_2(h), \tag{1}$$

where g_1 is the short-scale temporal component described by an exponential model (Cressie, 1993) with unit sill and range equal to 30 days, g_2 is the long-scale temporal component described by an exponential model with unit sill and range equal to 365 days and the positive-definite matrices $\mathbf{B}_l, l = 1,2$, are:

$$\mathbf{B}_1 = \begin{bmatrix} 230 & 3.07 & -3.4 \\ 3.07 & 4.9 & -0.025 \\ -3.4 & -0.025 & 0.58 \end{bmatrix}, \qquad \mathbf{B}_2 = \begin{bmatrix} 30 & 1.2 & -1.1 \\ 1.2 & 0.37 & -0.016 \\ -1.1 & -0.016 & 0.073 \end{bmatrix}. \tag{2}$$

At this point, it is convenient to check if the fitted model (1) can be considered suitable to make predictions of the primary variable, thus a validation procedure has been properly performed.

3.2 Model validation and temporal prediction

The goodness of model (1) has been checked through the cross-validation technique. In this stage of the analysis a modified version of the GSLib program *Cokb3D* (Deutsch and Journel, 1998), named *T-Cok*, has been implemented and used to compute temporal predictions of PM_{10} on the basis of a) the auxiliary variables, b) the fitted MLT model and c) a properly defined neighborhood, i.e. a subset of time data which can be considered in the cokriging system. Hence the cross-validation has been performed and the correlation between PM_{10} residuals and estimated ones has been measured. The high values of the linear correlation coefficient (0.780) has confirmed the goodness of the fitted MLT model, which can be used to predict PM_{10} daily concentrations in time points after the last available data. In particular, PM_{10} residuals have been predicted for six time points (1-6 January 2014), by using the new GSLib routine *T-Cok*. The deseasonalized PM_{10} observations, the residuals of the auxiliary variables and the model (1) are the input information for the *T-Cok* routine. Then, the diurnal component has been added to the predicted PM_{10} residuals in order to obtain predictions of PM_{10} daily concentrations. By comparing PM_{10} daily concentrations measured from the 1st to the 6th of January 2014 and the predicted ones, it is worth highlighting that the behavior of the predicted values is quite similar to the true PM_{10} daily concentrations; moreover, as it is for the true values recorded in the period 1-6 January 2014, some predicted values are greater than the limit value of 50 $\mu g/m^3$ and it can represent a hazardous condition for air quality and human health.

4 Conclusions

In this paper, the time series of PM_{10} daily concentrations and two meteorological variables (Temperature and Wind Speed), correlated with the pollutant under study, were analyzed through multivariate geostatistical techniques. The importance and the advantages of using variogram-based procedures were pointed out during both modeling and prediction stages. The scientific community should consider the flexibility of the geostatistical tools for the analysis of time series and more theoretical and computational efforts should be made in order to extend the variogram-based techniques in the time domain.

References

Box, G. E. P., Jenkins, G. M. (1976). *Time series analysis: forecasting and control*. Holden Day. San Francisco.

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley Series in Probability and Mathematical Statistics. New York.

De Iaco, S., Palma, M., Posa, D. (2013). *Geostatistics and the Role of Variogram in Time Series Analysis: A Critical Review.* In: Montrone S., Perchinunno P. (eds) Statistical Methods for Spatial Planning and Monitoring. Contributions to Statistics. Springer, 47–75.

Deutsch, C. V., Journel, A. G. (1998). *GSLib: Geostatistical Software Library and User's Guide*. Oxford University Press. New York.

Haslett, J. (1997). On the sample variogram and sample autocovariance for non-stationary time series. *Statistician* **46**(4) 475–485.

Harrison, R. M., Deacon, A. R., Jones, M. R., Appleby, R. S. (1997). Sources and processes affecting concentrations of PM10 and PM2.5 particulate matter in Birmingham (U.K.). *Atmospheric Environment* **31** 4103–4117.

Journel, A. G., Huijbregts C.J. (1981). *Mining Geostatistics*. Academic Press. London.

Perez, P., Reyes, J. (2002). Prediction of maximum of 24-h average of PM10 concentrations 30 h in advance in Santiago, Chile. *Atmospheric Environment* **36**(28) 4555–4561.

Reinsel, G. C. (2003). *Elements of Multivariate Time Series Analysis*. Springer Science and Business Media.

Sayegh, A. S., Munir, S., Habeebullah, T. M. (2014). Comparing the performance of statistical models for predicting PM10 concentrations. *Aerosol Air Quality Research* **14**(3) 653–665.

Wackernagel, H. (2003). *Multivariate geostatistics-an introduction with applications*. 3rd edn. Springer. Berlin.

Sensitization to allergens and environmental features: a preliminary analysis to study their relation.

Silvia Fabris[1*], Giovanna Jona Lasinio[2] and Mario Santoro[3]

[1*] *University of Bergamo (IT); silvia.fabris@unibg.it*
[2] *DSS, University of Rome "Sapienza"; giovanna.jonalasinio@uniroma1.it*
[3] IAC-CNR, Rome; m.santoro@iac.cnr.it
[*] *Corresponding author*

Abstract. *Prevalence of allergic disease, in the last decades, increases in all countries, as much that The World Health Organization consider allergy a non-transmittable disease which is out of control. The most common allergies are determined by production of Immunoglobulin E (IgE), that can cause different disorders. Diagnosis started at the end of the nineteenth century by in vivo test and during the seventies IgE detection in blood had been introduced and it allowed to identify allergenic molecules; FABER test combines these two sources of information. The purpose of this study is to explore 16 allergens behavior on the province of Rome and to find the most appropriate model to define a possible relationship between sensitizations' occurrences and environmental features. In this case of study, we will take into account rainfall and minimum, maximum and average temperature recorded by ARPA Lazio.*

Keywords. *Log-Gaussian processes; Epidemiology; Spatio-temporal models, Bayesian methods.*

1 Introduction

Prevalence of allergic disease, in the last decades, increases in all countries, as much that The World Health Organization consider allergy a non-transmittable disease which is out of control. The most common allergies are determined by production of Immunoglobulin E (IgE), that can cause different disorders. Allergens are protein contained in allergenic sources; sensitization occurs when specific IgE are produced by atopic individuals and bind the trigger molecules [1]. Diagnosis started at the end of the nineteenth century, with the introduction in medicine of the first clinical allergy basic test: the *Skin Test* (*ST*), which have some limitation and it is not riskless. During the seventies IgE detection in blood had been introduced and it allowed to identify allergenic molecules. FABER test combines these two sources of information; first patients are tested with skin prick test that indirectly shows the presence of specific IgE, then direct IgE detection is made on serum sample by several *in vitro* method [2]. The purpose of this study is to explore 16 allergens behavior on the province of Rome and to find the most appropriate model to define a possible relationship between sensitizations' occurrences and environmental features. We consider eleven allergens belonging to plants, Ambrosia (Amb a 1),

Artemisia vulgaris (Art v 1), Betula pendula (Bet v 1), Birch, Hazel and Oak species (Cor a 1), Arizona Cypress (Cup a 1), Olea europaea (Ole e 1), Parietaria (Par j 2), Grasses (Phl p 1, Phl p 2 and Phl p 5), American Sycamore (Pla a 1), two due to cat (Fel d 1) and dog (Can f 1), two regarding house dust mite (Der p 1, Der p 2) and one coming from the Alternaria fungi (Alt a 1).

2 Data

We focus on 5523 clinical tests, collected between 2012 and 2017, from patients who live in the province of Rome. IgE values on the continuous scale, as they are recorded by the test, are not comparable between different allergens, for this reason, in order to evaluate the impact of each allergen on the population, data has been recoded into presence and absence of sensitization. Blood analysis reveal that 2032 patients do not show any positive response, 751 subjects recorded one sensitization and just one reveals to be affected by every allergic source. The most observed molecules are Cup a 1, Phl p 1, Der p 2, Fel d 1, the remaining allergens occur in less than 20% of cases, Art v 1, Pla a 1 and Amb a 1 are detected in less than 5% of the tests. The phenomenon act in a similar way about male and female separately, since the number of sensitizations is proportionally similar distributed between sexes.
Weather data has been collected from ARPA Lazio web portal: we considered available data, recorded by 26 meteorological stations, spread all over the province, collected from Spring 2012 to 31th December 2017, about minimum, maximum and average temperature and rainfall. Daily time series of those variables suggest constant trend and annual seasonality for all temperature's measurements. Precipitations, analyzed on the log-scale, do not show the same seasonal behavior, but again we have a constant trend. Moreover, average, minimum and maximum temperatures and rainfall show a stationary trend that does not differ much between stations too. Observations have been quarterly summarized by seasons, not considering periods where data are partially missing. For each season of each year of observation, mean, minimum and maximum value have been displayed, except for rainfall data, for which only the average amount has been considered. Dividing the new data into quartiles, the spatial behavior of temperature variables can be discussed: despite moderate changes between years of observation, the area of Rome and the cost present the higher temperature values, on the other hand, Castelli Romani, lake of Bracciano and the Nord - Est area are the coldest one.

3 Methods

Once sensitization and weather have been explored, we can go on studying the relation between them, modelling the phenomenon and estimating the parameters of interest with an MCMC algorithm. Having the exact place of residence of each patient, the finite pattern of point of positive response is easily represented on the entire province of interest: Figure 1 maps positive sensitizations of Fel d 1, Ole e 1 and Pla a 1, these allergens have been chosen because of their different impact on the observed sample.

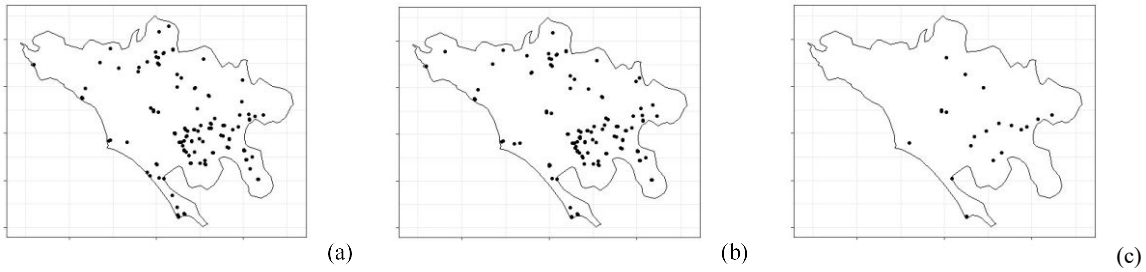

Fig. 1: Point Pattern of (a) Fel d 1, (b) Ole e 1, (c) Pla a 1 over the province of Rome.

Considering these locations as a random realization of a point pattern over a bounded window, fixed

and known, represented by the province of Rome, allergens phenomenon can be treated as a Poisson Process [3]. The phenomenon is clearly changing from one municipality to the other and many environmental phenomena may influence such occurrences; for this reason, we adopted an Inhomogeneous Poisson Process with intensity function varying in space. The class of models we are interested in is the Cox Process, in particular we used the log-Gaussian Cox Process [4]. In order to interpolate weather information all over the area of the Rome province, the time series needed to be predicted on every centroid of each municipality; we choose to model available spots using a Generalized Additive Model (GAM) [5]:

$$E(temperature) = \beta_0 + f(longitude, latitude) + f(time)$$
$$E(rain) = \beta_0 + f(longitude, latitude) + time\beta_1$$

For the purpose of the study, only prediction of the minimum of the minimum temperatures, the mean of the average temperatures and the maximum of the maximum temperatures and annually average rainfall have been used. Those have been summarized into one value for each centroid: first the median of average seasonal temperatures, minimum of minimum seasonal temperatures and maximum of maximum seasonal temperatures have been calculated, then these information have been transformed into their principal components. The first two components have been passed to the Cox process as covariates. Moreover, rainfall information have been summarized as the mean of the annual means. A grid, with square 2.3 X 2.3 km cells, overlaids the observation windows and covariates have been interpolated all over the grid by areal weight sum.

4 Results

Diagnostic results of the MCMC algorithm, concerning the considered allergenic, gave very good results, furthermore, inferential analysis of the estimated parameters finds out that meteorological features influence sensitization depending on molecules. Rain coefficient β is significantly different from 0 at 90% confidence interval just for Fel d 1: indeed the mean value of this coefficient shows a negative relation between allergens occurrence and rainfall, it means that each unit increase in rainfall lead to a reduction in relative risk with a mean of 3.373. Furthermore, the model can not explain any possible relation between those allergens that afflict less the population.

5 Further studies

The chosen protocol has several limitations, the main being the overly smoothed covariates added to the model. Furthermore, as far as vegetational allergens are concerned it would be of great interest to add information on the vegetation present in each municipality. Future developments will include a different interpolation of covariates and an ad hoc implementation of the Bayesian log-Gaussian Cox process.

References

[1] Alessandri C., Ferrara R., Bernardi M., Zennaro D., Tuppo L., Gianrieco I., Tamburrini M., Mari A., and Ciardiello M. (2017). Diagnosing allergic sensitization in the third millennium: why clinicians should know allergen molecule structures. *Clinical and translational allergy*, 7(1):21.

[2] Mari A. (2008). When does a protein become an allergen? Searching for a dynamic definition based on most advanced technology tools. *Clinical & Experimental Allergy*, 38(7):1089-1094

[3] Carlin, B. P., Gelfand, A. E., & Banerjee, S. (2014). *Hierarchical modeling and analysis for spatial data.*

Chapman and Hall/CRC.

[4] Taylor, B., Davies, T., Rowlingson, B., & Diggle, P. (2015). Bayesian inference and data augmentation schemes for spatial, spatiotemporal and multivariate log-Gaussian Cox processes in R. *Journal of Statistical Software*, 63, 1-48.

[5] Zuur, A. F. (2012). *A beginner's guide to generalized additive models with R*. Newburgh, NY, USA: Highland Statistics Limited.

Multiresolution Decomposition of Areal Count Data

R. Flury[1] and R. Furrer[2]

[1] *Department of Mathematics, University of Zurich, Switzerland; roman.flury@math.uzh.ch*
[2] *Department of Mathematics & Department of Computational Science, University of Zurich, Switzerland; reinhard.furrer@math.uzh.ch*

Abstract. *Multiresolution decomposition is commonly understood as a procedure to capture scale-dependent features in random signals. Such methods were first established for image processing and typically rely on raster or regularly gridded data. In this article, we extend a particular multiresolution decomposition procedure to areal count data, i.e. discrete irregularly gridded data. More specifically, we incorporate in a new model concept and distributions from the so-called Besag–York–Mollié model to include a priori demographical knowledge. These adaptions and subsequent changes in the computation schemes are carefully outlined below, whereas the main idea of the original multiresolution decomposition remains. Finally, we show the extension's feasibility by applying it to oral cavity cancer counts in Germany.*

Keywords. *Spatial scales; Lattice data; Intrinsic GMRF; Besag–York–Mollié model; MCMC.*

1 Introduction

Decomposing an observed signal or spatial field into scale-dependent components allows recognizing its inherent and prominent features. Those features give insight to where local or global phenomena manifest themselves and assist in understanding the structure of hierarchical information. Holmström et al. (2011) proposed a procedure in the tradition of image processing that hence is applicable to Gaussian data distributed on regular grids [7]. We extend this method to count data which is potentially observed on an irregular grid, often termed 'areal count data' [3]. The original multiresolution decomposition approach can be divided into three individual steps: 1) spatial field resampling based on a Bayesian hierarchical model, 2) smoothing on multiple scales, then calculating differences between these smooths to specify details for each resampled field separately, and 3) posterior credibility analysis. In the following paragraphs we summarize a) the Bayesian hierarchical model for step 1) and b) how to calculate differences between smooths in step 2). Those are the relevant parts in the procedure for the proposed extension, outlined in Section 2. The original multiresolution decomposition assumes that an observed field $\boldsymbol{y}$ consists of the true field $\boldsymbol{x}$ and additive white noise. Based on these flexible model assumptions the hierarchical model is constructed.

a) Bayesian hierarchical model: the true field $\boldsymbol{x}$ is presumed to follow a Gaussian distribution, which implies a selfsame likelihood function. Its positive valued variance is modeled with a scaled–inv–χ^2 prior and the spatial component of the field $\boldsymbol{x}$ is captured with an intrinsic Gaussian Markov random

field (IGMRF) using a precision matrix $\boldsymbol{Q}$ [10]. With those choices, the resulting marginal posterior is of closed form and corresponds to a multivariate t-distribution [4].

b) Calculate differences between smooths: the proposed penalty smoother is defined as $\boldsymbol{S}_\lambda = (\mathbf{I} + \lambda\boldsymbol{Q})^{-1}$, where λ is the scale or smoothing parameter, such that $0 = \lambda_1 < \lambda_2 < \ldots < \lambda_L = \infty$. The spatial field $\boldsymbol{x}$ is interpreted as random vector, $\boldsymbol{S}_{\lambda_1}\boldsymbol{x} = \boldsymbol{x}$ defines the identity mapping and $\boldsymbol{S}_{\lambda_L}\boldsymbol{x} = \boldsymbol{S}_\infty\boldsymbol{x}$ the mean field. On the ground of those preliminaries, $\boldsymbol{x}$ can be decomposed as differences of consecutive smooths: $\boldsymbol{x} = \sum_{l=1}^{L-1}\left(\boldsymbol{S}_{\lambda_l} - \boldsymbol{S}_{\lambda_{l+1}}\right)\boldsymbol{x} + \boldsymbol{S}_\infty\boldsymbol{x}$. Scale-dependent details are then formalized as $\boldsymbol{z}_l = \left(\boldsymbol{S}_{\lambda_l} - \boldsymbol{S}_{\lambda_{l+1}}\right)\boldsymbol{x}$ for $l = 1, \ldots, L-1$ and $\boldsymbol{z}_L = \boldsymbol{S}_\infty\boldsymbol{x}$.

Pivotal for a) and b) is the definition of the precision matrix $\boldsymbol{Q}$:

$$\boldsymbol{x}^\top\boldsymbol{Q}\boldsymbol{x} = \sum_j \left(\sum_{i\sim j} x_i - 4x_j\right)^2, \tag{1}$$

where $i\sim j$ denotes neighboring grid locations. To ensure four neighbors at every grid location i, the boundary values of $\boldsymbol{x}$ are extended across the initial grid. This definition inherently demands the data allocated to a regular grid but bears the advantage that individual computational steps can be optimized based on $\boldsymbol{Q}$'s fast eigendecomposition, such that large dimensional problems can be solved efficiently.

2 Extension

To decompose areal count data, first the resampling pattern described in a) needs modification. Assuming the n observed counts $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ are realizations from a conditionally independent Poisson distribution and the expected counts $\boldsymbol{e} = (e_1, \ldots, e_n)^\top$ are known for every location in the spatial field. The Poisson's rate for a location i, is defined as the product of the expected count e_i and the respective relative risk, denoted as $\exp(\eta_i)$. We construct the hierarchical model, to resample the spatial field, with the likelihood function

$$\pi(\boldsymbol{y}|\eta_1, \ldots, \eta_n) \propto \prod_{i=1}^{n} \exp\left(y_i\eta_i - e_i\exp(\eta_i)\right), \tag{2}$$

which corresponds to the classical Besag–York–Mollié (BYM) model [1]. Whereat $\boldsymbol{\eta}$ is modeled as the composition of the true log-relative risk $\boldsymbol{u}$ and a normal zero-mean noise term $\boldsymbol{v}$, with unknown precision parameter $\kappa_{\boldsymbol{v}}$. Analogous to the original model, we use a first order IGMRF process to model the spatial component with accompanying precision parameter $\kappa_{\boldsymbol{u}}$, such that

$$\pi(\boldsymbol{u}|\kappa_{\boldsymbol{u}}) \propto \kappa_{\boldsymbol{u}}^{\frac{n-1}{2}} \exp\left(-\frac{\kappa_{\boldsymbol{u}}}{2}\sum_{i\sim j}(u_i - u_j)^2\right) = \kappa_{\boldsymbol{u}}^{\frac{n-1}{2}} \exp\left(-\frac{\kappa_{\boldsymbol{u}}}{2}\boldsymbol{u}^\top\boldsymbol{R}\boldsymbol{u}\right). \tag{3}$$

Again $i\sim j$ denotes neighboring lattice locations but here in terms of regions sharing a common border. Assigning Gamma priors for both precision parameters implies a posterior distribution of non-closed form. Hence, we use a Gibbs sampler with a Metropolis-Hastings (MH) step to resample the log-relative risks $\boldsymbol{u}$, the noise components $\boldsymbol{v}$ and parameters [6]. Finally, we exploit that the mean of a Poisson distribution is equivalent to its rate and reconstruct the spatial field with $\boldsymbol{e}\cdot\exp(\boldsymbol{u} + \boldsymbol{v})$, for every sampled field $\boldsymbol{u}$ and $\boldsymbol{v}$.

We form the scale-dependent details still relying on a penalty smoother. Instead of using the matrix $\boldsymbol{Q}$ from the original model, we include the precision matrix $\boldsymbol{R}$ of the first order IGMRF [10]. The

definition of $\boldsymbol{R}$ does not limit the data to be associated with a regular grid and can be constructed based on adjacency relations of the respective observations. Since we use a different precision matrix, the optimized implementation relying on $\boldsymbol{Q}$ cannot be employed but we alternatively take advantage of the precision's sparse structure and apply tailored algorithms [5].

3 Application

The extension's feasibility is demonstrated on the German oral cavity cancer dataset [8]. This data includes cancer counts for 544 districts of Germany over 1986–1990, as well as the expected number of cases derived demographically. The main bulk of the oral cavity counts range between one and hundred counts per district but single highly populated districts have up to 500. The data including additional relevant information is available via the R package **spam** [5]. Following the multiresolution decomposition steps, we first resample the areal counts using suitable sampler specifications [6] and verify the convergence of the MH sampler with common diagnostic tools [2]. Figure 1 shows how well the reconstructed field corresponds to the original data. Only in northeast Germany, where the field is less smooth, the differences are larger. Since the BYM model was designed not to be oversensitive to extreme counts, part of the resampling difference can be explained through its damping effect [11].

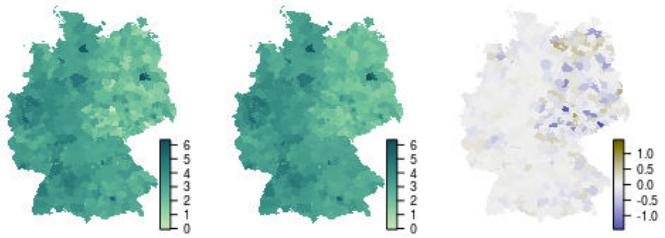

Figure 1: Oral cavity cancer data on logarithmic scale. Left: the observed number of cases; middle: the mean of the reconstructed fields; right: the difference between the left and the middle panels.

In the second step, we choose suitable scales ([9]) $\lambda_1 = 0$, $\lambda_2 = 1$ and $\lambda_3 = 25$ and form scale-dependent details (Figure 2). Completing the decomposition, we calculate pointwise probability maps [7] (Figure 3). The detail $\boldsymbol{z}_1$ reflects spatial noise as well as the relatively low or high counts in the data. This is also supported by its pointwise probability map, where no large red or blue clusters are visible. $\boldsymbol{z}_2$ catches larger patches of districts and shows local peculiarities. Detail $\boldsymbol{z}_3$ consists of the largest scale range and shows the east-west or nationwide trend but this trend is less distinct compared to the more local ones, indicated by the legends of each panel.

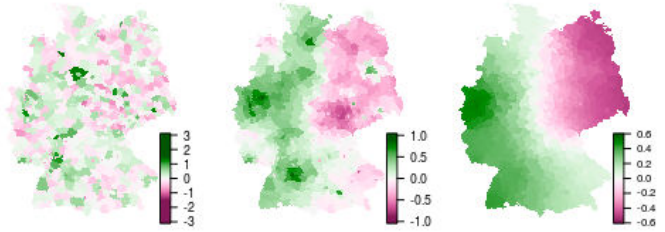

Figure 2: Scale dependent details $\boldsymbol{z}_l = \boldsymbol{S}_{\lambda_l}\log(\boldsymbol{e}\cdot\exp(\boldsymbol{u}+\boldsymbol{v})) - \boldsymbol{S}_{\lambda_{l+1}}\log(\boldsymbol{e}\cdot\exp(\boldsymbol{u}+\boldsymbol{v}))$, summarized by their posterior means. Left: $\mathrm{E}(\boldsymbol{z}_1|\boldsymbol{y})$; middle: $\mathrm{E}(\boldsymbol{z}_2|\boldsymbol{y})$; right: $\mathrm{E}(\boldsymbol{z}_3|\boldsymbol{y})$.

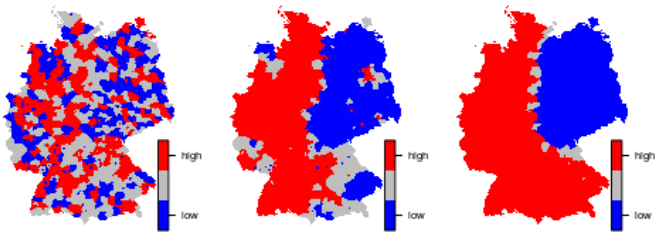

Figure 3: Pointwise probability maps. Left: z_1; middle: z_2; right: z_3. The map indicates which features are jointly credible: blue and red areas indicate jointly credibly negative and positive areas, respectively.

4 Discussion

We extended the multiresolution decomposition approach from Holmström et al. (2011), which originally processes data coming from a Gaussian distribution on a regular grid, to areal count data. Establishing an MH sampling model makes it possible to resample count data and use an arbitrary precision matrix. Employing the BYM model to include prior demographical knowledge, in the form of the known expected counts, enables us to model the data without being oversensitive to possible outliers. The R code to reproduce this example is available at https://git.math.uzh.ch/roflur/bymresa.

References

[1] Besag J. and York J. and Mollié A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43**, 1–20.

[2] Brooks S. P. and Gelman A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7**, 434–455.

[3] Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Wiley. New York.

[4] Erästö P. and Holmström L. (2005). Bayesian multiscale smoothing for making inferences about features in scatterplots. *Journal of Computational and Graphical Statistics* **14**, 569–589.

[5] Furrer R. and Sain. S. R. (2010). A sparse matrix R package with emphasis on MCMC methods for Gaussian Markov random fields. *Journal of Statistical Software* **36**, 1–25.

[6] Gerber F. and Furrer R. (2015). Pitfalls in the implementation of Bayesian hierarchical modeling of areal count data: An illustration using BYM and Leroux models. *Journal of Statistical Software* **63**, 1–32.

[7] Holmström L. and Pasanen L. and Furrer R. and Sain S. R. (2011). Scale space multiresolution analysis of random signals. *Computational Statistics & Data Analysis* **55**, 2840–2855.

[8] Knorr-Held L. and Raßer G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics* **56**, 13–21.

[9] Pasanen L. and Launonen I. and Holmström L. (2013). A scale space multiresolution method for extraction of time series features. *Stat* **2**, 273–291.

[10] Rue H. and Held L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC. London.

[11] Waller L. A. and Carlin B. P. (2010). Disease mapping. *Chapman & Hall/CRC Handbooks of Modern Statistical Methods* **2010**, 217–243.

Spatio-temporal analysis of extreme river flow

M. Franco-Villoria[1,*], M. Scott[2] and T. Hoey[2]

[1] *University of Torino, Italy; maria.francovilloria@unito.it,*
[2] *University of Glasgow, UK; marian.scott@glasgow.ac.uk, trevor.hoey@glasgow.ac.uk*
[*] *Corresponding author*

Abstract. *A quantile regression model is proposed to assess spatio-temporal trends and seasonality in extreme river flow in Scotland over the period 1st January 1996 to 31st December 2013. The model is built in a generalized additive model framework that allows inclusion of three-variate smooth functions to account for space-time interaction effects. The results suggest a clear East/West gradient in the 95th quantile of river flow that is in agreement with previous studies.*

Keywords. *Quantile regression; P-splines; PIRLS.*

1 Introduction

Recent studies [1, 5] report increases in both frequency and intensity of extreme events such as flooding. Climate change impacts are expected to vary spatially and to result in changes in river flows, the extremes of which are critical for flood risk estimation. Identification of patterns in extreme river flow behaviour, mainly in the form of seasonality and long term trends, is essential for planning purposes so that changes can be identified and decisions appropriately made to avoid or alleviate any negative impacts.

We introduce a new framework for spatio-temporal quantile regression [3], exploiting the flexibility of P-splines. The regression model is built as an additive model that includes smooth functions of time and space, as well as space-time interaction effects, and can be easily extended to incorporate potential covariates. Model parameters are estimated using a penalized iterative reweighed least squares approach instead of linear programming methods, classically used in quantile parameter estimation. The model is illustrated on a data set of daily river flows in 98 rivers across Scotland over the period 1st January 1996 to 31st December 2013.

2 Data

Daily river flow data for 98 gauging stations, shown in Figure 1, were provided by the Scottish Environment Protection Agency (SEPA) and the National River Flow Archive (NRFA) over the period 1st January 1996- 31st December 2013.

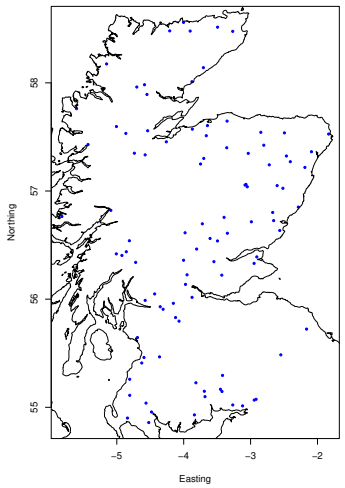

Figure 1: Location of the 98 selected gauging stations.

3 The model

Quantile regression [4] allows estimation of the relationship between response and explanatory variables at any percentile of the distribution of the response (conditioned on the explanatory variables). As a result, rates of change in the response variable can be estimated for the whole distribution and not only in the mean. The conditional quantile can be expressed as $Q_Y(\tau|X=x) = F_Y^{-1}(\tau|X=x)$, where $\tau \in (0,1)$, Y is the response variable with cumulative distribution function F_Y and $X = (X_1,\ldots,X_p)$ is a vector of explanatory variables [4, 2].

We propose the following model:

$$Q_{y_i}(\tau|t_i,d_i,z_i) = s_1(t_i) + s_2(d_i) + s_3(z_i) + s_4(t_i,d_i) + s_5(t_i,z_i) + s_6(d_i,z_i), \tag{1}$$

where $s_1(t)$, $s_2(d)$ are smooth functions of time and day of the year and $s_3(z)$ is a bivariate smooth function of easting and northing coordinates, accounting for the temporal, seasonal and spatial trends in river flow. The terms $s_4(t,d)$, $s_5(t,z)$ and $s_6(d,z)$ represent the time-season, space-time and space-season interactions, respectively. Estimating Model (1) involves minimizing a sum of weighted absolute deviations, where the weights are asymmetric functions of τ. In the classic quantile regression literature, linear programming methods are used for doing so [4]. We introduce an alternative approach by approximating the absolute residuals with the squared residuals; this way, model estimation can be done using the penalized iterative reweighed least squares (PIRLS) approach; see [3] for details.

4 Results and Discussion

We estimate Model (1) with $\tau = 0.95$ and log(daily flow) as the response variable, t=time (1996 to 2013), d=day within the year (1 to 365) and z=(easting, northing). Each univariate smooth term is re-expressed as a linear combination of B-spline basis functions, while interaction terms can be built using the tensor product of the marginal basis functions. We add a penalty term to control the amount of smoothness, and impose a periodicity constrain on the seasonal component.

The estimated temporal trend is fairly flat, while the seasonal effect shows lower values during the summer, as expected. The estimated spatial pattern suggests a slight East-West gradient, with greater values on the Western side. Regarding the interaction effects, the seasonal effect varies considerably over space, and in some years is very different from the rest. Overall, the results suggest that trends in the 95^{th} quantile of river flow are not homogeneous across Scotland; this information might prove useful in decision making, for example, to provide more accurate flood warnings.

References

[1] Black, A. and C. Burns (2002). Re-assessing the flood risk in Scotland. *The Science of the Total Environment* **294**, 169–184.

[2] Cade, B. and B. Noon (2003). A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment* **1(8)**, 412–420.

[3] Franco-Villoria M., M. Scott and T. Hoey (2018). Spatio-temporal modelling of hydrological return levels. A quantile regression approach. *Environmetrics* DOI: 10.1002/env.2522.

[4] Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs.

[5] Werritty, A. (2002). Living with uncertainty: climate change, river flows and water resource management in Scotland. *The Science of the Total Environment* **294**, 29–40.

Data fusion for air quality mapping using sensor data: feasibility and added-value through an application in Nantes

Alicia Gressent[1], Laure Malherbe[2*] and Augustin Colette[1]

[1] *alicia.gressent@ineris.fr, augustin.colette@ineris.fr*
[2] *laure.malherbe@ineris.fr*
[*] *Corresponding author*

Abstract. *The recent technological developments and the increased interest for public information lead to a fast-growing use of microsensors for air quality monitoring. Measurement campaigns are conducted to assess the potential of these low-cost instruments by deploying fixed sensors (e.g. on top of buildings, street lights or reference stations) and/or mobile sensors (e.g. on top of cars, bikes, or carried by citizens). These experiments allow to measure pollutant concentrations at high resolution in space and time. The large amount of collected information offers new opportunities of developments in air quality modelling and mapping. This work aims to take the best of these sensors despite the related measurement uncertainty to produce urban air pollution maps at fine spatial and temporal resolution. A geostatistical methodology (data fusion) is presented, which uses sensor observations as well as dispersion model outputs. It is applied to PM_{10} data in the French city of Nantes. It involves new challenges such as the consideration of the quick change of the sensor location if it is mobile, the temporal variability of the measurements, the analysis of numerous and heterogeneous data, the spatial representativeness of the measurements and the measurement uncertainties. Also, efforts still need to be done on the sampling design to ensure appropriate spatial coverage of the considered domain and get more accurate estimates.*

Keywords. *Air Quality Mapping; Microsensor; Data Fusion; PM_{10}*

1 Introduction

Air quality monitoring is conventionally based on a network of stations which allows a continuous report of pollutant concentrations. The related measurement uncertainty is constrained by the European existing legislation [1, 2] ensuring observation accuracy. Nevertheless, the installation and maintenance of such a network are expensive and so the number of stations in each region is limited. The use of numerical modelling on various scales (regional, urban, local) has thus increased during the last 15 years to supplement station observations and support air quality assessment.
In parallel, the technological progress allowed the development of miniaturized and low-cost instruments to measure pollutant concentrations [3]. Many projects of crowdsourcing and citizen science are emerging. In addition, field measurement campaigns are conducted to assess the potential of these low-cost devices by deploying fixed sensors (on top of buildings, street lights, reference stations) and/or mobile sensors (on top of cars, bikes, or carried by citizens) offering higher spatial coverage than

reference stations. Because microsensors suffer from metrological weaknesses, a calibration is generally applied to the raw data [4, 5].
The large amount of collected information offers new opportunities of development in air quality modelling and mapping at urban scale that are the scope of recent studies. Statistical methodologies are broadly used to derive air quality maps from sensor data, in particular the Land Use Regression models (LUR), but they generally do not take spatial dependence into account. Geostatistical approaches have been less frequently applied to such type of data but provide significant advantages to combine sensor measurements and auxiliary information such as dispersion model outputs [6].

In this paper, data collected from fixed and mobile micro-sensors are used together with urban-scale modelling data to map PM_{10} in the city of Nantes (France).

2 PM_{10} data

PM_{10} sensor data were provided by AtmoTrack (https://atmotrack.fr), a French company created in 2015 in Nantes. PM_{10} data measured at reference monitoring stations (quarter-hourly mean concentrations) and simulation data (ADMS-Urban model) on the city of Nantes were provided by the French air quality monitoring association Air Pays de la Loire (http://www.airpl.org/).

2.1 Sampling routes and frequency of measurements

In November 2018, PM_{10} sensor measurements were collected in Nantes. During this sampling period, the company deployed 16 fixed sensors including 3 sensors at the Victor Hugo station (reference station for traffic typology) and 3 other sensors at the Bouteillerie station (reference station for urban background typology). In addition, 19 mobile sensors were installed on-board of driving school cars to measure PM_{10} concentrations over numerous routes each day of the sampling period. The vehicles routes ensure a satisfactory spatial coverage over the entire urban area even if they are totally dependent on the driving school car itineraries and on the lesson time (only daytime).

2.2 Measurement accuracy

Considering the measurement uncertainty, the three available datasets (data from the reference stations, the fixed sensors and the mobile sensors) can be related to three monitoring networks of respectively low (up to 25%), medium (up to 50%) and high (up to 125%) uncertainty (Figure 1).

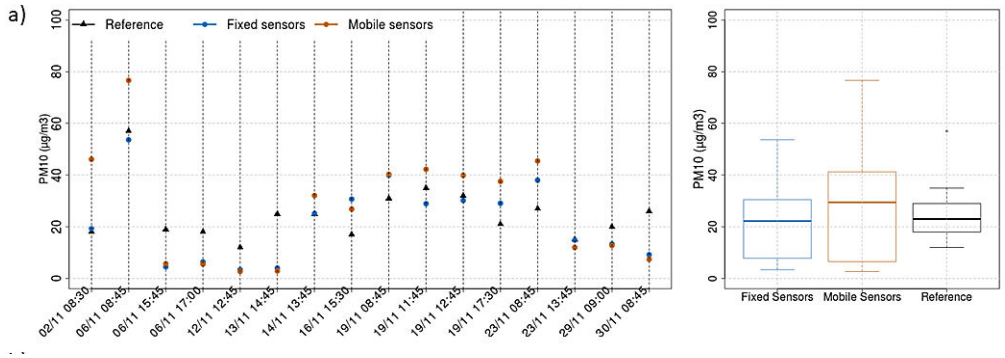

Figure 1: Comparison of the three networks (fixed sensors in blue, mobile sensors in orange and reference station in black) at Victor Hugo station for November 2018. Example of PM_{10}.

Microsensors offer a unique spatial and temporal coverage of pollutant concentrations. However, the accuracy of the measurements and their meaning, in case of mobile sensors, are real challenges to include them in air quality maps. In the following sections, a methodology of data fusion is detailed and a first test using fixed and mobile sensor data in Nantes (France) is presented.

3 Data fusion

Kriging [7] involves deriving linear combination of the data which ensures minimal estimation variance under a non-bias condition. Its strength is to give an information about the uncertainty of the estimated map.
Among the kriging methods, the universal or external drift kriging makes it possible to consider auxiliary information to increase the estimation accuracy. The main hypothesis is that the global mean is not constant through the domain and relies on explanatory variables, entailing an additional condition on the kriging weights. This approach has long been applied to air quality mapping [8, 9, 10, 11] and was used in this work to perform data fusion between:

- the hourly average concentrations measured by the fixed and mobile microsensors (after bias correction) as main variable;
- the 2016 annual average concentrations of the pollutant simulated by the ADMS-Urban dispersion model (https://www.cerc.co.uk/environmental-software/ADMS-Urban-model.html) as drift of the mean.

In addition, the measurement uncertainty of the sensors was taken into account by defining the variance of measurement errors (hereafter VME) as an input of the calculation.

4 Results

4.1 Estimation of PM_{10} concentration fields

Data fusion was performed for 27/11/2018, the day for which the amount of data was the largest. At every measurement position, the hourly mean of the observations is calculated, and external drift kriging is applied. The mobile and fixed sensor observations at 5pm and the annual modelled concentration field are presented for PM_{10} in figure 2a). Figure 2b) presents the VME for the same sampling routes. In this case, the measurement uncertainty is set to 25%, i.e. to the maximum uncertainty of the reference station observations. The uncertainty definition is totally arbitrary here and could be considered between 25% to 125%. Note that the fixed station measurements are not included in this estimation because they were used to correct and prepare the sensor data before kriging.

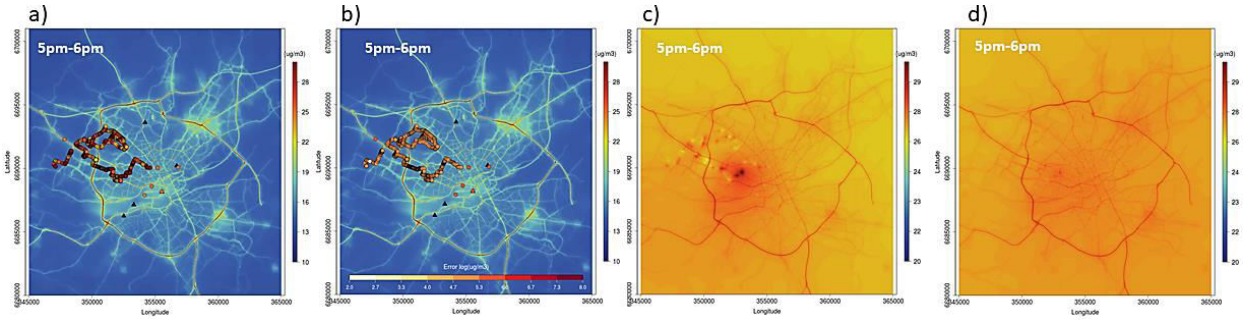

Figure 2: Data fusion of the sensor data on 27/11/2018 at 5pm: the 2016 annual average concentrations simulated by ADMS-Urban and the hourly-averaged sensor data (a), the variance of the measurement errors (b), the fused map with 25% uncertainty on measurements (c), and the fused map with 75% uncertainty on measurements (d).

As shown by the fused maps (Figure 2c and 2d), the modelled annual average allows to define the general patterns of the pollutant fields. Then the sensor observations which are associated with higher concentrations (by a factor of two) increase the concentration levels in the estimation domain, with some PM_{10} hotspots where data were collected (Figure 2c). When data fusion is performed with higher VME (75%, figure 2d), the hotspots are not represented anymore and the local effects of the sensor data is minimized.

5 Conclusion

The recent technological developments for miniaturizing the instruments that measure outdoor ambient air offer new possibilities for air quality modelling and mapping. The new portable and low-cost devices could provide observations of pollutants with higher spatial coverage than the reference monitoring networks. As long as several challenges can be dealt with (measurement uncertainty, representativeness of the sampling…), they could help to produce more accurate pollution maps. In this work, we investigate the potential added value of these data for air quality mapping by applying a data fusion technique. The dataset refers to PM pollution in the French city of Nantes. Hourly averaged sensor data and the annual mean concentration field simulated by ADMS-Urban model are combined by external drift kriging to estimate hourly PM_{10} concentrations, taking the variance of the measurement error into account. Those calculations were performed for one day but the next step of this work will be to consider each hour of the whole sampling period (November 2018). Further investigations will be carried out to estimate the influence of the amount of data, their position and their related uncertainties on the interpolation results. In addition, several ways of improvement have been identified such as the consideration of the spatial anisotropy in kriging and the application of spatiotemporal kriging. Besides geostatistical methods, machine learning techniques will be tested allowing to learn about historical data to improve the current estimate.

Acknowledgments

We thank AtmoTrack company for providing their sensor data and Air Pays de La Loire for providing the reference station observations and the results from their simulations with the ADMS-Urban model.

References

[1]. Directive 2008/50/EC of the European Parliament and the Council of 21 May 2008 on ambient air quality and cleaner air for Europe.

[2]. Directive 2004/107/EC of the European Parliament and the Council of 15 December 2004 relating to arsenic, cadmium, mercury, nickel and polycyclic aromatic hydrocarbons in ambient air.

[3]. Kumar, P., et al. (2015). The rise of low-cost sensing for managing air pollution cities. Environment International, 75, 199-205.

[4]. Spinelle, L., Gerboles, M., Villani, M. G., Aleixandre, M., Bonavitacola, F., (2015). Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: Ozone and nitrogen dioxide. Sensors and Actuators B: Chemical, 215, 249–257, http://dx.doi.org/10.1016/j.snb.2015.03.031.

[5]. Maag, B., Zhou Z., Thiele L. (2015). A survey on sensor calibration in air pollution monitoring deployments. IEEE Internet of Things journal, 5 (6), 4857 – 4870.

[6]. Schneider, P., Castell, N., Vogt, M., Dauge, F. R. and Lahoz, W. A., (2017). Mapping urban air quality in near real-time using observations from low cost sensors and model information. Environment International, 106, 234-247.

[7]. Chilès, J.-P., Delfiner, P.P. (2012). Geostatistics: Modeling Spatial Uncertainty. John Wiley & Sons, 726 pp.

[8]. Wackernagel H., Lajaunie C., Blond N., Roth C., and Vautard R. (2004). Geostatistical risk mapping with chemical transport model output and ozone station data. Ecological Modelling, 179, 177–185.

[9]. Malherbe L., Ung A., Colette A., Debry E. (2011). Formulation and quantification of uncertainties in air quality mapping. ETC/ACM Technical Paper 2011/9.

[10]. Lahoz, W.A., Schneider, P.P. (2014). Data assimilation: making sense of earth observation, Frontiers in Environmental Science, 2 (16), 1–28, http://dx.doi.org/10.3389/fenvs.2014.00016.

[11]. Beauchamp M., de Fouquet Ch., Malherbe L. (2017). Dealing with non-stationarity through explanatory variables in kriging-based air quality maps. Spatial Statistics, 22 (1), 18-46.

CircSpaceTime: an R package for spatial and spatio-temporal modeling of Circular data

Giovanna Jona Lasinio[1*], Mario Santotoro[1,2] and Gianluca Mastrantonio[3]

[1*] *DSS, Sapienza University of Rome, giovanna.jonalasinio@uniroma1.it*
[2] *IAC CNR,m.santoro@iac.cnr.it*
[3] *Department of Mathematical Sciences, Polytechnic of Turin, gianluca.mastrantonio@polito.it*

Abstract. *CircSpaceTime is an R package that implements Bayesian models, recently developed, for spatial and spatio-temporal interpolation of circular data. Such data are often found in applications where, among the many, wind directions, animal movement directions, and wave directions are involved. To analyze such data we need models for observations at locations* $\mathbf{s}$ *and times* t*, so-called geostatistical models, providing structured dependence which is assumed to decay in distance and time. For example, for wave directions in a body of water, we conceptualize a wave direction at every location and every time and introduce structured dependence into these angular data. The approach we take begins with Gaussian processes defined for linear variables over space and time. Then, we use either wrapping or projection to obtain processes for circular data. The models are cast as hierarchical, with fitting and inference within a Bayesian framework. Altogether, this package implements work developed by a series of papers; the most relevant being [9, 25, 16]. All procedures are written using Rcpp. Estimates are obtained by MCMC allowing parallelized multiple chains runs. As running example, for the spatial setting, we use a wave direction dataset while simulated data are used to illustrate the spatio-temporal models.*

Keywords. *Directional data; Spatial model; Spatio-temporal model; Rcpp*

1 Summary of existing circular packages and what CircSpaceTime is adding

In the last ten years the interest in circular data has increased, with new theoretical results and models (for an extended review of both theory and applications see [11] or [12]). There exist several R packages dealing with circular data. The best known are **circular** [2] and **CircStats** [13], both implementing inference for univariate data as described in [8]. Another recent set of functions specifically devoted to wrapped distributions is **Wrapped** [18]. The package computes the probability density function, cumulative distribution function, quantile function and many more features for several (about fifty) univariate wrapped distributions. A very interesting set of functions is implemented in **CircSizer** [20] where a non-parametric approach is adopted. Based on scale-space ideas, **CircSiZer** presents a graphical device to assess which observed features are statistically significant, both for density and regression analysis. Also a book on circular data in R has been published [21] with many nice examples and a narrative of the topic that makes easy to learn how to run inferences on univariate data. In 2013 the first version of the package **isocirc** was presented [3], making available functions to perform constrained inference using isotonic regression for circular data [22, 6]. The **CircOutlier** [7] collects functions to detect outliers

in circular-circular regression as proposed in [1]. Bayesian estimation for univariate regression models is implemented in **bpnreg** [4] that presents models developed in [19] and [5]. Again in the Bayesian framework the work in [17] is implemented in **circglmbayes**. More recent is the **Directional** [24] package, mostly linked to the textbook by [14]. A series of wrapper functions to implement the 10 maximum likelihood models of animal orientation, described by [23], are included in the **CircMLE**. The proposals in [26] are presented in **circumplex**.

Dependent and multivariate circular data are often found in applications (see [12] for recent developments). To handle them in a likelihood framework we can refer to the package **CircNNTSR**, that implements functions to plot, fit by maximum likelihood, and simulate models based on non-negative trigonometric sums for circular, multivariate circular, and spherical data.

None of the above packages deal with spatial or spatio-temporal interpolation of circular data, that is the main objective of **CircSpaceTime**, the package we are proposing. In what follows we are going to present models that have been developed, starting from 2012 [9]; a summary of these models can also be found in [10]. **CircSpaceTime** is available at `https://github.com/santoroma/CircSpaceTime`.

There are different approaches to specify valid circular distributions, see for example [8], here we focus on the two methods that allow to built a circular distribution starting from a linear one, namely the wrapping, and the projection. Both revealed to be useful in the definition of spatial and spatio-temporal data modeling, see for example [15] and [25]. Under both methods, the resulting distribution has a complex functional form but introducing suitable latent variables, the joint distribution of observed and latent variables, in a fully Bayesian framework, are really easy to handle. We are going to show some examples of implementation to illustrate the package features.

Acknowledgments. This work has been partially developed under the PRIN2015 supported-project *Environmental processes and human activities: capturing their interactions via statistical methods* (EPHAStat) funded by MIUR (Italian Ministry of Education, University and Scientific Research) (20154X8K23-SH3). Gianluca Mastrantonio research has been partially supported by MIUR grant Dipartimenti di Eccellenza 2018 - 2022 (E11G18000350001).

References

[1] A. H. Abuzaid, A. G. Hussin, and I. B. Mohamed. Detection of outliers in simple circular regression models using the mean circular error statistic. *Journal of Statistical Computation and Simulation*, 83(2):269–277, 2013.

[2] Claudio Agostinelli and Ulric Lund. *R package `circular`: Circular Statistics (version 0.4-93)*. CA: Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University, Venice, Italy. UL: Department of Statistics, California Polytechnic State University, San Luis Obispo, California, USA, 2017.

[3] Sandra Barragán, Miguel A. Fernández, Cristina Rueda, and Shyamal Das Peddada. isocir: An R package for constrained inference using isotonic regression for circular data, with an application to cell biology. *Journal of Statistical Software*, 54(4):1–17, 2013.

[4] Jolien Cremers. *bpnreg: Bayesian Projected Normal Regression Models for Circular Data*, 2018. R package version 1.0.0.

[5] Jolien Cremers, Kees Tim Mulder, and Irene Klugkist. Circular interpretation of regression coefficients. *British Journal of Mathematical and Statistical Psychology*, 71:5–95, 2017.

[6] Miguel A. Fernàndez, Cristina Rueda, and Shyamal D. Peddada. Identification of a core set of signature cell cycle genes whose relative order of time to peak expression is conserved across species. *Nucleic Acids Research*, 40(7):2823–2832, 11 2011.

[7] Azade Ghazanfarihesari and Majid Sarmad-Ferdowsi University Of Mashhad. *CircOutlier: Detection of Outliers in Circular-Circular Regression*, 2016. R package version 3.2.3.

[8] S. Rao Jammalamadaka and A. SenGupta. *Topics in Circular Statistics*. World Scientific, Singapore, 2001.

[9] G. Jona Lasinio, A. E. Gelfand, and M. Jona Lasinio. Spatial analysis of wave direction data using wrapped Gaussian processes. *Annals of Applied Statistics*, 6(4):1478–1498, 2012.

[10] Giovanna Jona Lasinio, Gianluca Mastrantonio, and Alan E. Gelfand. *Applied Directional Statistics: Modern Methods and Case Studies*, chapter Spatial and Spatio-Temporal Circular Processes with application to Wave Directions, pages 129–162. Interdisciplinary statistics. Chapman and Hall/CRC, 2019.

[11] Christophe Ley and Thomas Verdebout. *Modern Directional Statistics*. Interdisciplinary statistics. Chapman and Hall/CRC, 2017.

[12] Christophe Ley and Thomas Verdebout, editors. *Applied Directional Statistics: Modern Methods and Case Studies*. Interdisciplinary statistics. Chapman and Hall/CRC, 2019.

[13] Ulric Lund and Claudio Agostinelli. *CircStats: Circular Statistics, from "Topics in Circular Statistics" (2001)*, 2018. R package version 0.2-6.

[14] K. V. Mardia and P. E. Jupp. *Directional Statistics*. John Wiley and Sons, Chichster, 1999.

[15] G. Mastrantonio, A. E. Gelfand, and G. Jona Lasinio. The wrapped skew Gaussian process for analyzing spatio-temporal data. *Stochastic Environmental Research and Risk Assessment*, 30(8):2231–2242, 2016.

[16] G. Mastrantonio, G. Jona Lasinio, and A. E. Gelfand. Spatio-temporal circular models with non-separable covariance structure. *TEST*, 25:331–350, 2016.

[17] Kees Mulder and Irene Klugkist. Bayesian estimation and hypothesis tests for a circular generalized linear model. *Journal of Mathematical Psychology*, 80:4 – 14, 2017.

[18] Saralees Nadarajah and Yuanyuan Zhang. *Wrapped: Computes Pdf, Cdf, Quantile, Random Numbers and Provides Estimation for any Univariate Wrapped Distributions*, 2017. R package version 2.0.

[19] G. Nuñez-Antonio and E. Gutiérrez-Peña. A Bayesian model for longitudinal circular data based on the projected normal distribution. *Computational Statistics and Data Analysis*, 71(C):506–519, 2014.

[20] María Oliveira, Rosa M. Crujeiras, and Alberto Rodríguez-Casal. Circsizer: an exploratory tool for circular data. *Environmental and Ecological Statistics*, 21(1):143–159, Mar 2014.

[21] Arthur Pewsey, Markus Neuh auser, and Graeme D Ruxton. *Circular Statistics in R*. Oxford University Press, 2013.

[22] Cristina Rueda, Miguel A. Fernández, and Shyamal Das Peddada. Estimation of parameters subject to order restrictions on a circle with application to estimation of phase angles of cell cycle genes. *Journal of the American Statistical Association*, 104(485):338–347, 2009.

[23] Jon T. Schnute and Kees Groot. Statistical analysis of animal orientation data. *Animal Behaviour*, 43(1):15 – 33, 1992.

[24] Michail T Tsagris, Giorgos Athineou, Anamul Sajib, Eli Amson, and Micah J. Waldstein. *Directional: Directional Statistics*, 2018. R package version 3.3.

[25] F. Wang and A. E. Gelfand. Modeling space and space-time directional data using projected Gaussian processes. *Journal of the American Statistical Association*, 109(508):1565–1580, 2014.

[26] Johannes Zimmermann and Aidan G. C. Wright. Beyond description in interpersonal construct validation: methodological advances in the circumplex structural summary approach. *Assessment*, 24(1):3–23, 2017. PMID: 26685192.

Goodness of Fit Test For Wrapped Normal Distribution

Anahita Nodehi

Department of Statistics, Computer Science, Applications (DiSIA), Florence University, Florence, Italy.

Abstract. *One of the main difficulties in any statistical method is whether the data could have actually been drawn from that fitted distribution or not. To extend in circular data, it is necessary to consider nature feature of this data. The Wrapped Normal and Von Mises are two most important and famous distributions in circular. Based on the auther's knowledge, there is no Goodness-of-Fit test for Wrapped Normal. To enhance this issue, in this paper, we present an appropriate test and compare its performance based on simulation study.*

Keywords. *Goodness-of-Fit Test; Wrapped Normal; Von Mises; Circular Data*

1 Introduction

Directional data is a type of data that has a wide range of applications in applied sciences. For consideration of value to directions, it is common to specify an angle on a unit circle since an initial direction and the orientation of the circle have been chosen. Therefore, having such periodic feature makes one to consider the topological feature of the non-Euclidean space. Accordingly, many methods and statistical techniques have been developed to analyze and understand this type of data. The popular approaches have been embedding, wrapping and intrinsic approaches. Based on every approach, great number of distributions are proposed which are in non-Euclidean space. The Wrapped Normal and the Von Mises are two important distributions on the circle, which resemble on circle the Normal distribution on Euclidean space. To provide a better sense of this phenomenon, certain overview of the embedding and intrinsic approaches can be found in Jammalamadaka and SenGupta (2001), Mardia (1972) and Mardia and Jupp (2000).

One of the fundamental questions that arise in every statistical application is whether the data could have actually been drawn from that fitted distribution. This is the so-called Goodness-of-Fit problem. The importance of this problem arises especially in some estimation methods as the first step is to check if the assumption is hold or not. For example, Nodehi et al (2018) proposed two algorithms which estimate the parameters of Wrapped Normal distribution. With regard to that, it is necessary to check whether the data is Wrapped Normal or not.

To do so, in circular data, one should consider the periodic feature of data. The Goodness-of-Fit testing for a Von Mises distribution fitted using maximum likelihood estimation (without bias correction for

the estimation of κ), were obtained by Lockhart and Stephens (1985). Based on the auther's knowledge, there are no contribution to find the same test for Wrapped Normal distribution. In that sense, the main goal in this paper is to propose a Goodness-of-Fit test based on Wrapped Normal distribution.

The reminder of this paper is organized as follows. In section 2, the review of circular densities is presented. Afterwards, Section 3 is based on Goodness-of-Fit procedure. Section 4 provides simulation study while Section 5 gives final comments and remarks.

2 Statistical Modeling

As mentioned in Introduction, there are three approaches to modeling circular data: embedding, wrapping and intrinsic approaches. In the embedding approach the sample space is considered as part of larger space and the distributions on the S^{P-1} (the circular sample space) can be obtained by radial projection of the in line distributions on R^P. In general, most of the literature is focused on developing statistical methods for the projected Normal distribution, which is, the only a significant limitation of the embedding approach.

In the intrinsic approach, the circle is used as the sample space. The directions are represented as points on the circle and probability distributions are defined on the circle directly. The main probability distributions obtained from this approach are the Uniform, Cardioid and Von Mises distributions.

The wrapping approach consists to wrap a known distribution in the real line around a circumference of a circle with a unit radius. In that sense, the main characteristic of this approach is flexibility. Elaborating on, it is a rich class of distributions on the circle that can be obtained using the wrapping technique, as it is possible to wrap any known distribution in the real line onto the circle. Therefore, the most famous probability distribution based on this approach is Wrapped Normal which resembles Normal distribution in Euclidean space. Since the main contribution of this paper is to propose a Goodness-of-Fit for Wrapped Normal, it is necessary to reviwe certain features of this distribution.

Any linear random variable X may be transformed to a circular random variable by reducing its modulo 2π. i.e.

$$\theta = X(mod\ 2\pi)$$

This operation is equal to taking a line random variable and wrapping around circle of unit radius, accumulating probability over all points $X = (\theta + 2K\pi)$ where $K \in \mathbb{Z}$. If F represents the circular distribution function and G distribution function of line random variable, we have

$$F(\theta) = \sum_{K=-\infty}^{+\infty} \{G(\theta + 2K\pi) - G(2K\pi)\}\ , \qquad 0 \leq \theta \leq 2\pi\ .$$

In particular, if θ has a circular density function f and g is density function of X then

$$f(\theta) = \sum_{K=-\infty}^{+\infty} g(\theta + 2K\pi)\ .$$

A Wrapped Normal distribution is obtained by wrapping a $N(\mu, \sigma^2)$ distribution around the circle. Its

pdf is given by

$$f(\theta) = \sum_{K=-\infty}^{+\infty} g(\theta - \mu + 2K\pi) = \frac{1}{\sqrt{2\pi\sigma^2}} \sum_{K=-\infty}^{+\infty} \exp\{\frac{-(\theta - \mu + 2K\pi)^2}{2\sigma^2}\}$$

An alternate and more useful representation of this density using Fourier expansion and properties of characteristic function can be shown to be

$$f(\theta) = \frac{1}{2\pi}(1 + 2\sum_{p=1}^{+\infty} \rho^{p^2} \cos p(\theta - \mu))$$

where $\rho = \exp(\frac{-\sigma^2}{2})$ (Jammalamadaka and SenGupta, 2001). In this regard, some properties of Wrapped Normal are as follows: it is unimodal and symmetric about the value μ, the mean resultant length is $\rho = \exp(\frac{-\sigma^2}{2})$, as $\rho \to 0$, the distribution converges to the uniform distribution, as $\rho \to 1$, it tends to a point distribution at μ, it appears in the central limit theorem and Brownian Motion, the convolution of two Wrapped Normal variables is also Wrapped Normal, unlike the Von Mises distribution

3 Goodness-of-fit test

According to Pewsay et al (2013), considering the circular analogue of the probability integral transformation, it follows implicitly that the Goodness-of-Fit of a posited distribution with distribution function $F(\theta)$ that can be tested by calculating the values of $2\pi F(\theta_1), \ldots, 2\pi F(\theta_n)$ and applying any test of circular uniformity to them. If the data do come from the postulated distribution, then we would expect circular uniformity not to be rejected. The problem with this approach is that the usual critical values of the tests for circular uniformity do not apply if the parameters of the distribution have been estimated from the data. The difference between the correct critical values and those for the usual tests of circular uniformity should not be great, however, for large sample sizes. Lockhart and Stephens (1985) proposed a Goodness-of-Fit testing for a Von Mises distribution fitted using maximum likelihood estimation based on the critical values of Watsons U^2 test which is implemented within the function watson.test available in Rs circular package (Agostinelli and Lund, 2017) if its argument dist is specified as vonmises. Since, Wrapped Normal and Von Mises have close relationships, it is possible to use the same procedure by calculating the values of $2\pi F(\theta_1), \ldots, 2\pi F(\theta_n)$ and applying any test of circular uniformity to them. In this regard, it is expected that under some conditions ($\sigma \to 0$), the two distributions have the same behavior.

4 Simulation study

To compare the performance of the proposed method we consider simulation study. To do so, we consider sample size $n = 50, 100$, $\mu_0 = 0$, $\sigma_0 = (\pi/8, \pi/4, \pi/2, \pi, 3/2\pi, 2\pi)$, and the number of Monte Carlo replications 100. As canbe seen in Table 1, the values (within 100 replications) are based on number of times, the test has been accepted. In other words, we generate the data of Wrapped Normal and Von Mises distribution and see whether the data could have actually been drawn from that fitted distribution or not. According to Kent (1978), any Von Mises distribution can be approximated by a Wrapped Normal distribution when σ is small or $\kappa \to \infty$; i.e.

$$f_{VM}(\theta, \mu, \kappa) - f_{WN}(\theta, \mu, A_1(\kappa)) = O(\kappa^{\frac{-1}{2}})$$

		Fitted distribution							
		WN				VM			
		Kuper		Watson		Kuper		Watson	
Real Distribution	σ	$n=50$	$n=100$	$n=50$	$n=100$	$n=50$	$n=100$	$n=50$	$n=100$
WN	$\frac{\pi}{8}$	100	92	99	93	100	92	99	93
	$\frac{\pi}{4}$	97	96	96	96	98	94	96	92
	$\frac{\pi}{2}$	95	88	98	92	97	95	95	92
	π	85	77	78	65	90	96	92	96
	$\frac{3\pi}{2}$	86	74	85	76	96	92	96	92
	2π	81	78	80	69	91	100	94	98
VM	$\frac{\pi}{8}$	85	93	83	91	92	99	91	99
	$\frac{\pi}{4}$	89	86	91	81	94	97	96	96
	$\frac{\pi}{2}$	95	92	96	94	96	92	98	96
	π	94	86	93	79	95	97	94	96
	$\frac{3\pi}{2}$	93	83	92	76	97	96	96	96
	2π	94	78	93	73	96	96	97	95

Table 1: Results of the Monte Carlo simulation based on 100 replications.

where f_{VM} and f_{WN} are the densities of the Von Mises (μ,κ) and the Wrapped Normal $(\mu, A_1(\kappa))$ distribution, respectively. Therefore, when σ is sufficiently small, the test cannot distinguish between these two distribution, as expected. Moreover, by $\sigma \to \infty$, the both tend to circular Uniform distribution. Thus, it is better to increase the number of replications and sample sizes and add another distribution with different features in simulation study to see more distinctions obtained by the test.

5 Conclusion

Based on the auther's knowledge, there is yet no test for Goodness-of-Fit to Wrapped Normal distribution. According to periodic feature of circular data and close relationship between Wrapped Normal and Von Mises distribution, in simulation study, we show the performance of this test.

References

[1] Agostinelli, C. and Lund, U. (2017). R package `circular`: Circular Statistics (version 0.4-93), https://r-forge.r-project.org/projects/circular/.

[2] Jammalamadaka, S. R. and SenGupta, A. (2001). *Topics in Circular Statistics*, World Scientific, Singapore.

[3] Kent, J. T. (1978). Limiting Behaviour of the Von Mises-Fisher Distribution. *Math. Proc. Cambridge Phil. Soc.*, **84**, 531-536.

[4] Lockhart, R. A. and Stephens, M. A. (1985). Tests of Fit for the Von Mises Distribution. *Biometrika*, **72**, 647–52.

[5] Mardia, K. V. (1972). *Statistics of Directional Data*, Academic Press, London.

[6] Mardia, K. V. and Jupp, P. (2000). *Directional Statistics*, John Wiley, Chichester.

[7] Nodehi, A., Golalizadeh, M., Maadooliat, M. and Agostinelli, C. (2018). Estimation of Multivariate Wrapped Models for Data in Torus, *arXiv preprint arXiv:1811.06007*.

[8] Pewsey, A., Neuhauser, M., and Ruxton, G. D. (2013). *Circular Statistics in R*. Oxford University Press, England.

Graphical model selection for air quality time series

L. Paci[1,*] and G. Consonni[1]

[1] *Department of Statistical Sciences, Università Cattolica del Sacro Cuore; lucia.paci@unicatt.it, guido.consonni@unicatt.it;*
[*] *Corresponding author*

Abstract. *We propose an objective Bayes approach based on graphical models for learning dependencies among multiple air quality time series within the framework of Vector Autoregressive (VAR) models. Using a fractional Bayes factor approach, we obtain the marginal likelihood in closed form and construct an MCMC algorithm for Bayesian graphical model determination with limited computational burden. We apply our method to study the interactions between four air pollutants over the municipality of Milan (Italy).*

Keywords. *Decomposable graphical model; Fractional Bayes factor; Multiple pollutants.*

1 Introduction

Air pollution is a major global environmental risk to human health. Because humans are simultaneously exposed to a complex mixture of air pollutants, many organizations are moving toward a multi-pollutant approach to air quality [4]. Key aspects of such approach are the estimation of the health risk of multiple pollutants, the setting of regulatory standards and the design of compliance strategies for multiple pollutants. For example, a strategy to reduce levels of one pollutant, say particulate matter, may also affect the levels of other pollutants, say ozone. To take on these challenges, a better understanding of the interactions between air pollutants is required.

Pollutant measurements or numerical model estimates usually arise a multivariate time series collected at fixed locations or aggregated over a given spatial domain. Vector Autoregressive (VAR) models offer a suitable framework for analyzing multiple time series, such as air quality data. VAR models can be naturally represented by graphs, with directed edges reflecting the autoregressive structure over time while undirected edges describe the contemporaneous interactions among variables.

In this paper we describe an objective Bayes methodology to learn dynamic and contemporaneous dependencies among multiple pollutants modeled through a graphical VAR. Using a fractional Bayes factor approach, we are able to obtain the marginal likelihood in closed form and perform Bayes graphical model selection with limited computational burden because we focus on marginal likelihood, and disregard inference on model parameters. We apply our method to analyze the time series of four pollutants over the municipality of Milan (Italy). Results offer helpful insights about the relationship between these pollutants.

1.1 VAR model

Let $\mathbf{y}_t$ be a $(q \times 1)$ vector of observations collected at time t, $t = 1, \ldots, T$. The reduced form of a stable VAR of order k, VAR(k), is given by

$$\mathbf{y}_t = \sum_{i=1}^{k} \mathbf{B}_i \mathbf{y}_{t-i} + \boldsymbol{\epsilon}_t, \qquad t = 1, \ldots, T, \tag{1}$$

where $\mathbf{B}_i$ are $(q \times q)$ matrices of coefficients or lag matrices, determining the dynamics of the system and $\boldsymbol{\epsilon}_t$ is a $(q \times 1)$ dimensional white noise process, that is $\boldsymbol{\epsilon}_t \mid \boldsymbol{\Sigma} \sim N_q(\mathbf{0}, \boldsymbol{\Sigma})$, independently over time. The observation vector at time t depends linearly on the previous k observations, where k is assumed to be known. The intercept and exogenous variables can be added to the model, leading to straightforward modifications of the results shown here; for simplicity we omit details. Let $\mathbf{z}_t = (\mathbf{y}'_{t-1}, \ldots, \mathbf{y}'_{t-k})'$ denote the $(kq \times 1)$ vector of lagged observations at time t and $\mathbf{B}' = (\mathbf{B}_1, \ldots, \mathbf{B}_k)$ be the $(q \times kq)$ matrix obtained by collecting together the corresponding coefficient matrices. Let $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_T)'$ be the $(T \times q)$ matrix collecting all observations and $\mathbf{Z}$ be the $(T \times kq)$ matrix containing all the lagged variables, i.e., $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_T)'$. Equation (1) can be rewritten in matrix form as

$$\mathbf{Y} = \mathbf{Z}\mathbf{B} + \mathbf{E}, \tag{2}$$

where $\mathbf{E} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_T)'$ is the $(T \times q)$ matrix of errors following a Matrix Normal distribution with zero mean, cross-covariance matrix between vector column j and vector column j' of $\mathbf{Y}$ equal to $\sigma_{jj'}\mathbf{I}_T$ and covariance matrix of vector row i equal to $\boldsymbol{\Sigma}$, which we write $\mathbf{E} \mid \boldsymbol{\Sigma} \sim \mathcal{N}_{T,q}(\mathbf{0}, \mathbf{I}_T, \boldsymbol{\Sigma})$.

2 Model selection for graphical VAR

Let $G = (V_{TS}, E)$, be a graph with node set $V_{TS} = V \times \mathbb{Z}$, $V = \{1, 2, \ldots, q\}$, and edge set E, whose edges have at most k lags and which is invariant under translation. If $(\mathbf{B}_i)_{vw}$ is the (v,w)-element of matrix $\mathbf{B}_i$ in (1) and $(\boldsymbol{\Omega})_{vw}$ is the (v,w)-entry of precision matrix $\boldsymbol{\Sigma}^{-1}$, then the VAR model with the following constraints on the parameters

$$\begin{aligned} &i) \quad (v, t-i) \rightarrow (w,t) \in E \Leftrightarrow (\mathbf{B}_i)_{vw} \neq 0 \qquad i = 1, \ldots, k \\ &ii) \quad (v,t) \text{ — } (w,t) \in E \Leftrightarrow (\boldsymbol{\Sigma}^{-1})_{vw} \neq 0 \qquad t = 1, \ldots, T \end{aligned} \tag{3}$$

represents a VAR(k, G) model [6]. It follows from (3) that nonzero elements in $\mathbf{B}$ correspond to directed edges in the graph reflecting the recursive structure of the time series, while nonzero elements in $\boldsymbol{\Sigma}^{-1}$ correspond to undirected edges that specify conditional independencies at any given time t. In other words, learning the dynamic structure of a graphical VAR translates into a variable selection problem while learning the interactions among variables translates into a covariance selection problem.

Let $\boldsymbol{\Gamma}$ the binary connectivity matrix such that $(\boldsymbol{\Gamma})_{vw} = 1 \Leftrightarrow (\mathbf{B})_{vw} \neq 0$. Let $G^u = (V_{TS}, E^u)$ denote the undirected graph corresponding to the contemporaneous dependencies. We assume that $\boldsymbol{\Sigma}$ is Markov with respect to G^u, i.e., condition ii) is satisfied. We confine our analysis to the class of decomposable graphs for all time points, although we provide posterior graph summaries that go beyond this assumption. Given $\boldsymbol{\Gamma}$ and G^u, we denote $\mathbf{B}^{(\boldsymbol{\Gamma})}$ the associated coefficient matrix and $\boldsymbol{\Sigma}^{(G^u)}$ the associated covariance matrix. Then the likelihood of a graphical VAR(k, G) factorizes as

$$f\left(\mathbf{Y} \mid \mathbf{B}^{(\boldsymbol{\Gamma})}, \boldsymbol{\Sigma}^{(G^u)}\right) = \frac{\prod_{C \in \mathcal{C}} f\left(\mathbf{Y}_C \mid \mathbf{B}_C^{(\boldsymbol{\Gamma})}, \boldsymbol{\Sigma}_{CC}^{(G^u)}\right)}{\prod_{S \in \mathcal{S}} f\left(\mathbf{Y}_S \mid \mathbf{B}_S^{(\boldsymbol{\Gamma})}, \boldsymbol{\Sigma}_{SS}^{(G^u)}\right)}, \tag{4}$$

where $\mathcal{C}$ and $\mathcal{S}$ denote the set of cliques and separators of the undirected decomposable graph, while $\mathbf{B}_C^{(\mathbf{\Gamma})}$ and $\mathbf{B}_S^{(\mathbf{\Gamma})}$ are the matrices whose columns contain the nonzeros coefficients of the selected responses $\mathbf{Y}_C$ and $\mathbf{Y}_S$, respectively. Notice that the set of cliques C and separators S depend on G^u, which is omitted for simplicity.

In this work, we employ an objective approach for model selection based on the Fractional Bayes Factor (FBF), originally presented in [2]. The idea of the FBF is to train a noninformative, typically improper, prior using a small fractional power b of the likelihood, thus converting the noninformative prior into a proper prior. The latter is then used to compute the marginal likelihood based on the complementary fractional power $(1-b)$ of the likelihood.

We start with a prior for $\left(\mathbf{B}^{(\mathbf{\Gamma})}, \mathbf{\Sigma}^{(G^u)}\right)$ that is a limiting form of a Matrix Normal Hyper-Inverse Wishart distribution. Combining such prior with a fraction $b = T_0/T$ of likelihood (4), we obtain the fractional prior of a VAR(k,G) that is a Matrix Normal Hyper-Inverse Wishart distribution, $\mathcal{MNHIW}\left(\hat{\mathbf{B}}^{(\mathbf{\Gamma})}, \underline{\mathbf{C}}, d, \underline{\mathbf{R}}\right)$, where $\hat{\mathbf{B}}^{(\mathbf{\Gamma})}$ is the ordinary least square estimate of nonzero coefficients, $\underline{\mathbf{C}} = T/T_0\,(\mathbf{Z}'\mathbf{Z})^{-1}$, $d = T_0 - kq$ and $\underline{\mathbf{R}} = T_0/T\,\hat{\mathbf{E}}'\hat{\mathbf{E}}$ with $\hat{\mathbf{E}} = \mathbf{Y} - \mathbf{Z}\hat{\mathbf{B}}^{(\mathbf{\Gamma})}$. Hence, the fractional prior factorizes as

$$p^F\left(\mathbf{B}^{(\mathbf{\Gamma})}, \mathbf{\Sigma}^{(G^u)}\right) = \frac{\prod_{C\in\mathcal{C}} \mathcal{N}_{kq,|C|}\left(\mathbf{B}_C^{(\mathbf{\Gamma})}, \underline{\mathbf{C}}, \mathbf{\Sigma}_{CC}^{(G^u)}\right) I\mathcal{W}_{|C|}\left(d+|C|-1, \mathbf{R}_{CC}\right)}{\prod_{S\in\mathcal{S}} \mathcal{N}_{kq,|S|}\left(\mathbf{B}_S^{(\mathbf{\Gamma})}, \underline{\mathbf{C}}, \mathbf{\Sigma}_{SS}^{(G^u)}\right) I\mathcal{W}_{|S|}\left(d+|S|-1, \mathbf{R}_{SS}\right)}. \tag{5}$$

Because of conjugacy of prior (5), we can write

$$m^F\left(\mathbf{Y} \mid \mathbf{\Gamma}, G^u\right) = \frac{\prod_{C\in\mathcal{C}} m^F\left(\mathbf{Y}_C \mid \mathbf{\Gamma}\right)}{\prod_{S\in\mathcal{S}} m^F\left(\mathbf{Y}_S \mid \mathbf{\Gamma}\right)}, \tag{6}$$

where, $m^F\left(\mathbf{Y}_C \mid \mathbf{\Gamma}\right)$ and $m^F\left(\mathbf{Y}_S \mid \mathbf{\Gamma}\right)$ can be obtained in closed form using the results in [5].

2.1 Computational details

Posterior inference on the space of decomposable graphs is carried out through Markov Chain Monte Carlo (MCMC) methods. In particular, at each step of our collapsed Gibbs sampling, we locally modify $\mathbf{\Gamma}$ and G^u and then update through the following Metropolis-Hasting steps:

- we move from $\mathbf{\Gamma}$ to $\mathbf{\Gamma}_*$ with acceptance probability $r(\mathbf{\Gamma}, \mathbf{\Gamma}_*) = min\left\{1, \frac{m^F(\mathbf{Y}|\mathbf{\Gamma}_*, G^u)p(\mathbf{\Gamma}_*)q(\mathbf{\Gamma}|\mathbf{\Gamma}_*)}{m^F(\mathbf{Y}|\mathbf{\Gamma}, G^u)p(\mathbf{\Gamma})q(\mathbf{\Gamma}_*|\mathbf{\Gamma})}\right\}$;
- we move from G^u to G^u_* with acceptance probability $r(G^u, G^u_*) = min\left\{1, \frac{m^F(\mathbf{Y}|\mathbf{\Gamma}, G^u_*)p(G^u_*)q(G^u|G^u_*)}{m^F(\mathbf{Y}|\mathbf{\Gamma}, G^u)p(G^u)q(G^u_*|G^u)}\right\}$.

We compute $m^F\left(\mathbf{Y} \mid \mathbf{\Gamma}_*, G^u\right)$ using (6) while a multiplicity-correction prior for both the directed dynamic graph and the undirected contemporaneous graph is assumed [5]. Finally, $q(\mathbf{\Gamma}_* \mid \mathbf{\Gamma}) = \alpha$ when adding an edge, and $q(\mathbf{\Gamma}_* \mid \mathbf{\Gamma}) = 1-\alpha$ when deleting an edge; same proposal is employed for G^u, see [1].

Given the MCMC output we can approximate the posterior inclusion probability of edge (v,w) as the proportion of MCMC iterations, after the burn-in, wherein the edge (v,w) appears. A variety of summaries of the MCMC output can be adopted to estimate the data generating graph. Here, we employ a Bayesian version of the (approximate) expected false discovery rate (FDR; [3]), i.e., we estimate the graph considering those edges whose posterior probability of inclusion is greater than $1-r$, where r is determined so that the FDR is at most 5%.

3 Analysis of air quality data

We analyze a 4-dimensional time series of air pollutants from January 2016 to December 2018 over the municipality of Milan (Italy). Data are provided and aggregated by the air quality system of ARPA (the regional environmental protection agency of Lombardia). In particular, we study the time series of daily average of nitrogen dioxide (NO_2), daily 8-hour maximum ozone (O_3), daily particulate matter (PM_{10}) and daily fine particulate matter ($PM_{2.5}$). Left panel of Figure 1 displays such time series, highlighting the cyclical pattern of the pollutants.

A VAR$(1, G)$ is fitted using the approach described in Section 2. Daily average of temperature and precipitation over the city are employed as covariates. Preliminary results of the variable selection algorithm applied to these variables show that temperature is a relevant predictor for both NO_2 and O_3 while precipitation is relevant for particulate matter and nitrogen dioxide. Right panel of Figure 1 shows

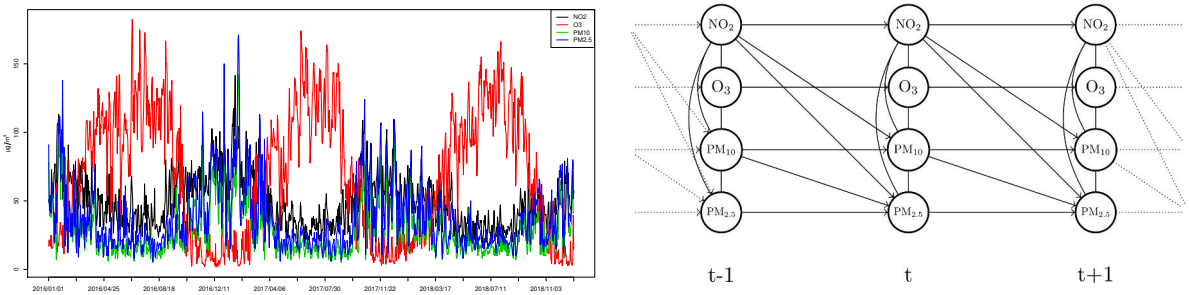

Figure 1: Left panel: daily averages of the pollutants. Right panel: estimated VAR graph.

the estimated VAR graph obtained using the FDR criterion described in Section 2. Some of the links of the graph can be explained by the chemical and physical transformation of the pollutants, e.g., NO_2 is a precursor to ozone while particulate matter and NO_2 are both indicators of urban pollution. However, interactions between air pollutants are very complex and require further investigations.

Acknowledgments. The research of the authors has been partially supported by a grant from UCSC (track D1) and by the EU COSTNET project (CA15109).

References

[1] Bhadra A. and Mallick B. K. Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis. *Biometrics*, 69:447–457, 2013.

[2] O'Hagan A. Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B*, 57:99–138, 1995.

[3] Peterson C., Stingo F. C., and Vannucci M. Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association*, 110:159–174, 2015.

[4] Dominici F., Peng R., Barr C. D., and Bell M. Protecting human health from air pollution: Shifting from a single-pollutant to a multi-pollutant approach. *Epidemiology*, 21:187–194, 2010.

[5] Consonni G., La Rocca L., and Peluso S. Objective Bayes covariate-adjusted sparse graphical model selection. *Scandinavian Journal of Statistics*, 44:741–764, 2017.

[6] Eichler M. Graphical modelling of multivariate time series. *Probability Theory and Related Fields*, 153:233–268, 2012.

A nonparametric spatio-temporal approach for multiple CUSUM charts of evapotranspiration

D. Pellegrino[1,*], G. Giungato[1] and S. De Iaco[1]

[1] *University of Salento, Via per Monteroni, Complesso Ecotekne, Lecce, Italy; daniela.pellegrino@unisalento.it, giuseppina.giungato@unisalento.it, sandra.deiaco@unisalento.it*
[*] *Corresponding author*

Abstract. *In recent years, the interest in natural resources management is increased. In this context, control charts, developed to monitor and maintain quality in industrial processes, are a useful monitoring and decision tool.*
In this paper, the behavior of an agro-metereological variable, named evapotranspiration, in an area of the northern part of Italy during a 25-year span (from 1992 to 2016) is studied through a nonparametric spatio-temporal geostatistical analysis and multiple CUSUM control charts. In particular, the probability that the variable registers an "out of control" is estimated over the area of interest, for three decades from the 10th to the 30th of January 2017.

Keywords. *Spatio-temporal geostatistics; Indicator kriging; Control charts.*

1 Introduction

The sustainable management of natural resources is an increasingly complex issue for environmental sciences. Hence, monitoring represents an important activity for decision-making procedures. In this context, the control charts might be useful for natural resources management, although they were developed as a tool in the Statistical Process Control (SPC) for improving industrial processes. On the basis of the classical approach, these SPC techniques are a representation of the quality characteristic measured in a sample or in several samples of an industrial process and allow pointing out if the process is "out of control" and it should be stopped (Montgomery, 2009).
The convenience of using the control charts approach in different fields such as environmental, economics, financial, social and healthcare sciences was discussed in several studies. In particular, the interest in SPC techniques to analyze environmental phenomena is increasing (Paroissin et al., 2016; Garthoff and Otto, 2016). On the other hand, few attempts to integrate the control charts with Geostatistics have been made, such as in Grimshaw et al. (2013). However, the geostatistical methods applied in the above mentioned papers were not used in a joint way in the space and in the time.
In this paper, the Cumulative Sum (CUSUM) charts, introduced by Page (1961), are used to study an agro-meteorological variable, i.e. evapotranspiration (ET_0), in 26 stations of Veneto region, in the period 1992-2016. In particular, the CUSUM charts technique has been integrated with nonparametric spatio-temporal geostatistical methods in order to predict the probability that the CUSUM chart signals that the variable is "out of control". These results could be useful to plan adequate water management strategies, since the ET_0 monitoring plays an important role in irrigation scheduling, watershed level budgeting, as well as climate and weather models.

2 CUSUM charts and geostatistical framework

In environmental field, the detection of changes in a phenomenon could represent a useful tool to define management and controlling plans. Hence, the CUSUM charts, based on the cumulative sums of deviations of the analyzed values from a target value (μ_0), might be a convenient technique for monitoring this variability. Neverthless, it is worth noting that the variables under study are usually nonstationary, so the residuals must be considered (Montgomery and Mastrangelo, 1991).
The residual values of a spatio-temporal environmental phenomenon, recorded at different time points and spatial locations, can be considered as a realization of a second-order stationary spatio-temporal random function (*STRF*), $\{Y(\mathbf{u}), \mathbf{u} = (\mathbf{s},t) \in D \times T\}$, where $D \subseteq \mathbb{R}^d$ and $T \subseteq \mathbb{R}$.
In particular, for each spatial location, the chart is obtained by plotting over the time the cumulative values $CS(\mathbf{s},t) = \sum_{j=1}^{t} [Y(\mathbf{s},j) - \mu_0] = [Y(\mathbf{s},t) - \mu_0] + CS(\mathbf{s},t-1)$, where $CS(\mathbf{s},0) = 0$. On the other hand, the CUSUM could be expressed in the form of decision-interval, based on the cumulative sums of positive and negative deviations from the target value μ_0 that are greater than a reference value indicated with K, respectively:

$$CS^{+}(\mathbf{s},t) = max[0, Y(\mathbf{s},t) - (\mu_0 + K) + CS^{+}(\mathbf{s},t-1)],$$
$$CS^{-}(\mathbf{s},t) = max[0, (\mu_0 - K) - Y(\mathbf{s},t) + CS^{-}(\mathbf{s},t-1)],$$

with starting values $CS^{+}(\mathbf{s},0) = CS^{-}(\mathbf{s},0) = 0$. The CUSUM chart is obtained by plotting these statistics over the time. In particular, if measurements are above the reference value, the upper CUSUM CS^{+} shows an upward trend; likewise, the lower CUSUM CS^{-} exhibits a downward trend if the phenomenon is consistently below the reference value.
Finally, the parameters K and H must be fixed. K is related to the size of the smallest shift in the level of the reference value that can be detected; while H is the threshold that CS^{+} and CS^{-} should not exceed in order to consider the phenomenon "in-control".
In a nonparametric context, given the fixed threshold $z = H$ and the CS^{+} and CS^{-} computed from residuals, a spatio-temporal indicator random field (*STIRF*),

$$\{I(\mathbf{u},z), \mathbf{u} = (\mathbf{s},t) \in D \times T\}$$

can be defined as follows:

$$I(\mathbf{u},z) = \begin{cases} 1 & \text{if } CS^{+}(\mathbf{u}) \geq z \text{ or } CS^{-}(\mathbf{u}) \geq z\,, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Under the second-order stationarity, the spatio-temporal indicator variogram, which describes the correlation, depends on the threshold z and the lag vector $\mathbf{h}$, i.e. $2\gamma_I(\mathbf{h};z) = Var[I(\mathbf{u}+\mathbf{h};z) - I(\mathbf{u};z)]$, where $\mathbf{h} = (\mathbf{h}_s, h_t)$.
In this context, the empirical spatio-temporal indicator variogram can be modelled through the following generalized product-sum model (De Iaco et al., 2001), selected among different spatio-temporal models proposed in literature:

$$\gamma_I(\mathbf{h}_s, h_t; z) = \gamma_I(\mathbf{h}_s, 0; z) + \gamma_I(\mathbf{0}, h_t; z) - k\gamma_I(\mathbf{h}_s, 0; z)\gamma_I(\mathbf{0}, h_t; z), \tag{2}$$

where $\gamma_I(\mathbf{h}_s, 0; z)$ and $\gamma_I(\mathbf{0}, h_t; z)$ are, respectively, spatial and temporal valid bounded marginal variograms and $k \in]0, 1/max\{sill\gamma_I(\mathbf{h}_s, 0; z), sill\gamma_I(\mathbf{0}, h_t; z)\}]$ is the parameter of spatio-temporal interaction.
For a second-order stationary *STIRF* I, a linear prediction of the probability that the CS^{+} or CS^{-} is greater than the threshold z, that means that the phenomenon is "out of control", can be obtained by using a linear combination of neighbouring indicator variables, expressed by the spatio-temporal indicator kriging predictor $\widehat{I}(\mathbf{u};z) = \sum_{\alpha=1}^{n} \lambda_\alpha(\mathbf{u}_\alpha;z) I(\mathbf{u}_\alpha;z)$, where $I(\mathbf{u}_\alpha;z)$, $\alpha = 1, 2, \ldots, n$ represent the indicator random variables at the sampled points $\mathbf{u}_\alpha \in D \times T$ and $\lambda_\alpha(\mathbf{u}_\alpha;z)$ are the kriging weights, determined by solving the indicator kriging system (Journel, 1983).

3 Case study

The control of ET_0 levels in a geographic area is an important tool for water management and planning, since this variable is a very crucial factor in river discharge, irrigation water requirement and soil moisture contents (Mohan and Arumugam, 1996).
In the present case study, the ET_0 levels (expressed in *mm*) provided by an Italian web system, named SCIA (Desiato et al., 2007), for 26 agro-meteorological stations located in the northeastern part of Italy (Veneto Region) have been analyzed. Note that selected data are averaged every ten days and refer to a 25-year span (from 1992 to 2016).
ET_0 is characterized by a periodic behavior: high levels registered in autumn and in winter are in contrast to low measurements in the other seasons. Hence, in order to remove the periodic component exhibited by the data, the FORTRAN program REMOVEMULT described in De Iaco et al. (2010) has been used; consequently the residual data have been used in the steps of the analysis.
From residuals, for each spatial location positive (CS^+) and negative (CS^-) CUSUM have been computed by fixed the parameters $K = 2\sigma$ and $H = 3\sigma$, where σ is the global standard deviation equals to 0.359. Hence, by considering the parameter $H = 1.077$ as the threshold z, a nonparametric analysis has been conducted on the indicator variable $I(\mathbf{u};z)$, in order to estimate the probability that the CUSUM CS^- exceeds the threshold z and to predict the probability that the ET_0 is "out of control" for three future time points, that is the 10th, 20th and 30th of January 2017.
After computing the sample spatio-temporal indicator variogram (Fig.1), the space-time correlation of the indicator variable has been modeled through the following product-sum model:

$$\gamma_I(\mathbf{h}_s,h_t;z) = N_s + c_s\,Exp(\|\mathbf{h}_s\|;a_s) + N_t + c_t\,Exp(h_t;a_t) - k\{[N_s + c_s\,Exp(\|\mathbf{h}_s\|;a_s)]\cdot[N_t + c_t\,Exp(h_t;a_t)]\}$$

where N_s and c_s are, respectively, the nugget and the sill contribution of the spatial marginal indicator variogram model which is $\gamma_I(\mathbf{h}_s,0;z) = 0.015 + 0.019Exp(\|\mathbf{h}_s\|;a_s)$, while N_t and c_t are, respectively, the nugget and the sill contribution of the temporal indicator variogram model which is $\gamma_I(\mathbf{0},h_t;z) = 0.012 + 0.097Exp(h_t;a_t)$, with spatial range a_s equals to 80 km and temporal range a_t equals to 55 days. Note that the parameter k, which is equals to 7.104, is such that the admissibility condition is satisfied and the global sill, equals to 0.117, is fitted.

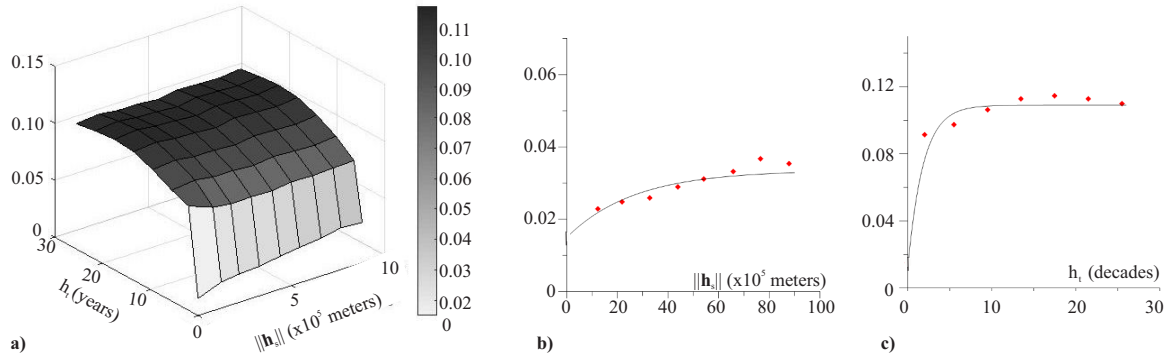

Figure 1: a) Sample indicator spatio-temporal variogram surface, b) marginal spatial variogram and fitted model, c) marginal temporal variogram and fitted model.

The reliability of the fitted spatio-temporal model is evaluated through cross-validation technique and some fitting indexes. The linear correlation coefficient between the observed values and the estimates

from cross-validation, equals to 0.8, confirms the goodness of the fitted model. Moreover, the Mean Error (ME) and the Root Mean Square Error (RMSE) computed on the fitting errors between the empirical surface and the model, equal to 0.007 and 0.005, respectively, confirm the accuracy of the fitted spatio-temporal model.
Finally, these models have been applied in order to obtain spatio-temporal indicator kriging predictions over the area of interest for three decades, from the 10th to the 30th of January 2017, through a modified *GsLib* routine (De Iaco et al., 2011). Then, the probability maps of the negative CUSUM CS^{-} exceeding the fixed threshold have been obtained. The results highlight that, in the analyzed area, there are low probabilities that the CUSUM exceeds the fixed threshold. Hence, the ET_0 behavior will be "in-control" in these three decades.

References

[1] De Iaco, S., Myers, D.E., Posa, D. (2001). Space-time analysis using a general product-sum model. *Statistics and Probability Letters* **52**(1), 21–28.

[2] De Iaco, S., Myers, D.E., Palma, M., Posa, D. (2010). FORTRAN programs for space-time multivariate modeling and prediction. *Computer & Geosciences* **36**(5), 636–646.

[3] De Iaco, S., Posa, D. (2011). Predicting spatio-temporal random fields: some computational aspects. Computer & Geosciences, doi: 10.1016/j.cageo.2011.11.014

[4] Desiato, F., Lena, F., Toreti, A. (2007). SCIA: a system for a better knowledge of the Italian climate. *Bollettino di Geofisica Teorica ed Applicata* **48**(3), 351–358.

[5] Garthoff, R., Otto, P., (2017). Control charts for multivariate spatial autoregressive models. *AStA Adv Stat Anal* **101**, 67–94.

[6] Grimshaw, S. D., Blades, N. J., Miles, M. P. (2013). Spatial Control Charts for the Mean. *Journal of Quality Technology* **45**(2), 130–148.

[7] Journel, A.G. (1983). Nonparametric estimation of spatial distributions. *Mathematical Geology* **15**(3), 445–468.

[8] Mohan, S., Arumugam, N., 1996. Relative importance of meteorological variables in evapotranspiration: Factor analysis approach. *Water Resource Management* **10**(1), 1–20.

[9] Montgomery, D. C. (2009). *Introduction to Statistical Quality Control*. Sixth Edition. John Wiley & Sons, Inc.

[10] Montgomery, D. C., Mastrangelo, C. M. (1991). Some Statistical Process Control Methods for Autocorrelated Data. *Journal of Quality Technology* **23**(3), 179–193.

[11] Page, E. S. (1961). Cumulative sum charts. *Technometrics* **3**(1), 1–9.

[12] Paroissin, C., Penalva, L., Pétrau, A., Verdier, G. (2016). New control chart for monitoring and classification of environmental data. *Environmetrics* **27**, 182–193.

Estimation of Spatial Deformation for Non-stationary Processes via Variogram Alignment

Ghulam A. Qadir[1], Ying Sun[1*] and Sebastian Kurtek[2]

[1] *CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia; ghulam.qadir@kaust.edu.sa; ying.sun@kaust.edu.sa*
[2] *Department of Statistics, The Ohio State University, Columbus, OH 43210, USA; kurtek.1@stat.osu.edu*
[*] *Corresponding author*

Abstract. *In modeling spatial processes, a second-order stationarity assumption is often made. However, for spatial data observed on a vast domain, the covariance function often varies over space, leading to a heterogeneous spatial dependence structure, therefore requiring non-stationary modeling. Spatial deformation is one of the main methods for modeling non-stationary processes, assuming the non-stationary process has a stationary counterpart in the deformed space. The estimation of the deformation function poses severe challenges. Here, we introduce a novel approach for non-stationary geostatistical modeling, using space deformation, when a single realization of the spatial process is observed. Our method is based, at a fundamental level, on aligning local variograms, where warping variability of the distance from each subregion explains the spatial non-stationarity. We propose to use multi-dimensional scaling to map the warped distances to spatial locations. We asses the performance of our new method using multiple simulation studies. Additionally, we illustrate our methodology on soil moisture data to estimate the heterogeneous spatial dependence and to perform spatial predictions.*

Keywords. *Functional data registration; distance warping; spatial statistics.*

Anisotropic attenuation of the macroseismic intensity

R. Rotondi[1*], E. Varini[1]

[1] *CNR - Istituto di Matematica Applicata e Tecnologie Informatiche, Via Bassini 15, Milano (I)*
renata.rotondi@mi.imati.cnr.it, elisa.varini@mi.imati.cnr.it
[*] *Corresponding author*

Abstract. *Macroseismic intensity is a measure of the size of an earthquake in terms of the damages caused to the anthropic and natural environment; it is an ordinal quantity expressed through the twelve degrees of the macroseismic scale. The set of intensity values recorded in the sites around the epicenter constitutes the macroseismic field of the event, that is, the damage scenario produced by the earthquake. Knowing the expected spatial distribution of the effects of a future quake would allow to carry out prevention actions and to intervene more promptly in the case of disastrous event. At first we studied a probabilistic model for the intensity at site under the assumption that the attenuation trend was circular; actually, drawing the isoseismal lines (lines of equal felt seismic intensity) of many earthquakes we noted that the trend is quite complex and it is influenced by the ground conditions and the orographic configuration. Therefore, to generalize the shape of the isoseismal lines, we considered an elliptical trend where the major axis of the first isoseismal line corresponds to the fault. In this work we extend the previous results by proposing a method to determine location and extremes of the fault when they are unknown. Examples of the damage scenario estimated for some volcanic and recent tectonic earthquakes are given.*

Keywords. *Anisotropy; Beta-binomial probability model; Ellipse; Macroseismic intensity; Bayesian inference*

1 Beta-binomial model and ellipsoid hull

Conditioned on the epicentral intensity I_0, and on a fixed epicentral distance, the intensity at a given site I_s is assumed to have a binomial distribution with parameter p. First we assume that the decay is isotropic, i.e., we assume to have a point source and circular isoseismal lines bounding the points of equal intensity. In this case we draw J circular bins around the epicentre and suppose that in all of the sites within each j-th bin, I_s - so as ΔI - has the same binomial distribution with parameter p_j, i.e.:

$$Pr(I_s = i \,|\, I_0 = i_0, p_j) = Pr(\Delta I = I_0 - i \,|\, I_0 = i_0, p_j) = \binom{i_0}{i} p_j^i (1-p_j)^{i_0 - i}. \tag{1}$$

In its turn, each p_j has a beta distribution with hyperparameters α_j and β_j that, according to the Bayesian approach, we assign by exploiting the information drawn from previous databases; then the posterior

mean of each p_j provides the estimate of these paramaters. To extend the value of p at any epicentral-site distance we approximate the estimates $\hat{p_j}$ by the smoothing inverse power function $g(d) = [c_1/(c_1 + d)]^{c_2}$, whose coefficients c_1, c_2 are estimated by the method of least squares. In this way we are able to forecast in terms of macroseismic intensity I_s at site the damage scenario that a future earthquake of given intensity I_0 could cause by the *smoothed* binomial probability distribution obtained by replacing p_j with $g(d)$ in Eq. (1) and by using the mode i_{smooth} of this distribution as forecast value of the intensity I_s at any site distant d from the epicentre [2]. Three criteria were used to validate the results: the logarithmic scoring rule, the ratio between the probability that the fitted model assigns to an observation and the probability of the forecast value, and the absolute discrepancy between observed and estimated intensities at site.

Since it has been observed that more rapid decay can be visibly recognizable along the direction perpendicular to that of the fault, it can be appropriate to use an elliptical shape for the isoseismal lines when we have information on the fault rupture that caused an earthquake, in particular on the direction and length of the rupture [1]. The solution we have found to do that, consists in a plane transformation that turns the ellipse of major axis equal to the fault rupture into the circle of radius equal to the width of the first bin; we repeat the estimation procedure in the transformed plane and then we associate the estimated probability distribution of the intensity I_s that will be felt at a site to the original position of that site [3].

The problem arises when we do not have information on the causative fault, e.g. when the fault does not appear on the surface, being completely hidden underneath surface rock layers (blind fault). Taking into account that the shape of the area of highest intensity is generally elongate along the direction of the active fault plane, we propose to deduce the fault dimensions from those of the ellipsoid hull that includes all the sites with $I_0 - I_s \leq 1$, i.e. the ellipsoid of minimal area such that all given points lie just inside or on the boundary of the ellipsoid. The method has been tested on some volcanic earthquakes of Etna area for which the fault is known and on L'Aquila earthquake with a blind fault.

Acknowledgments. This work was partly financed by the Italian Ministry of Education, University and Research (MIUR) in the framework of the PRIN-2015 project 'Complex space-time modeling and functional analysis for probabilistic forecast of seismic events'.

References

[1] Agostinelli, C. and Rotondi, R. (2016). Analysis of macroseismic fields using statistical data depth functions: considerations leading to attenuation probabilistic modelling, *Bull. Earth. Eng.*, 14, 1869–1884, DOI:10.1007/s10518-015-9778-2

[2] Azzaro, R., D'Amico, S., Rotondi, R., Tuvè, T., Zonno G. (2013). Forecasting seismic scenarios on Etna volcano (Italy) through probabilistic intensity attenuation models: A Bayesian approach, *Journal of Volcanology and Geothermal Research*, 251, 149–157.

[3] Rotondi, R., Varini, E. and Brambilla, C. (2016). Probabilistic modelling of macroseismic attenuation and forecast of damage scenarios, *Bull. Earth. Eng.*, 14, 1777–1796, DOI:10.1007/s10518-015-9781-7

The preliminary study of the PM_1 main components and of their seasonal variation using a linear parametric model

A. Speranza[1]*, G. Jona Lasinio[2]and R. Caggiano[1]

[1]*IMAA, Istituto di Metodologie per l'Analisi Ambientale, CNR, 85050 Tito Scalo, PZ, Italy; rosa.caggiano@imaa.cnr.it; antonio.speranza@imaa.cnr.it*

[2]*Department of Statistical, Sciences, University of Rome "la Sapienza"- P.le Aldo Moro 5, 00185 Rome, Italy; giovanna.jonalasinio@uniroma1.it*

**Correspondingauthor*

Abstract. *This study presents the mass concentration PM_1(aerosol particles with an aerodynamic diameter below 1 μm) together with sixteen related trace elements (i.e. Al, Ca, Cd, Cr, Cu, Fe, K, Li, Mg, Mn, Na, Ni, Pb, S, Ti, Zn) measured during Summer and Winter periods in a characteristic anthropized area. Soil, Sulfate, total-metal-oxide (TMO) and the carbonaceous material (CM) masses were evaluated and considered as the main components of PM_1. All variables were log transformed to smooth extreme recording influence. A linear parametric model was estimated to asses components influence on the log PM_1 and to evaluate the possible differential effect of observation periods on the components contribution. Results showed that CM, Sulfate, Soil and TMO explained about 44.32%, 33.56%, 11.4% and 0.2% of the total variance of PM_1, respectively. The contributions of CM, Sulfate and Soil to PM_1 were significative with an error of 5%. The contribution of TMO to PM_1 was significative with an error of 10%. CM, Soil and TMO contributed to PM_1 with a significant difference between Summer and Winter, whereas Sulfate contributed to PM_1 with non significant difference between Summer and Winter. Therefore, the CM, Sulfate and TMO components which were mainly related to anthropogenic origin explained about 78% of the PM_1 total variance, whereas the Soil component which was mainly related to natural origin explained about 11.4% of the PM_1 total variance. CM, Soil and TMO components contributed differently to PM_1 in Summer and Winter. This result suggested possible seasonal sources activities for these components.*

Keywords.PM_1; Linear parametric model; Seasonal sources.

1 Introduction

Human activities and natural processes contribute to the formation and emission in the air of aerosol particles, which are also known as particulate matter (PM). These particles have different sizes, shapes and masses and they are made of many chemical compounds some of which potentially harmful. The size of particle is an important physical parameter because it provides relevant information on particles origin, their formation process and harmful effects. Particles with an aerodynamic diameter smaller that 10 μm (thoracic particles) are of special interest because they can penetrate and be deposited in specific thoracic regions of the lung. It has been reported that PM toxicity increases with the aerodynamic diameter decrease, as particles with smaller aerodynamic diameter can easily reach deep regions of the lung and can vehicle potentially toxic substances. Indeed, physiological and

toxicological considerations have suggested that fine particles (i.e. aerosol particles with an aerodynamic diameter smaller that 2.5 μm) can play the largest role in affecting human health [1]. Fine particles have mainly an anthropogenic origin and they are mostly formed through the processes of combustion and relating condensation/reaction and gas-to-particle conversion of materials containing potentially toxic elements. PM_1 (i.e. aerosol particles with an aerodynamic diameter smaller that 1 μm) is a PM fraction that better represents the contribution of PM anthropogenic sources. The identification of the possible emission sources of PM_1 is a starting point to evaluate and plan actions aimed at mitigating the levels of PM to protect the public health and the environment. This preliminary study aims to evaluate the main components of PM_1 relating to natural and anthropogenic emission sources and to evaluate their seasonal variations using a linear parametric model.

2 Methods

The mass concentration of PM_1 measured during Summer and Winter periods in a characteristic anthropized area was determined and sixteen related trace elements (i.e. Al, Ca, Cd, Cr, Cu, Fe, K, Li, Mg, Mn, Na, Ni, Pb, S, Ti, Zn) were analyzed. The area was characterized from natural emission sources and from diverse anthropogenic emission sources relating to several activities (i.e. industrial, agricultural, domestic heating and traffic).

2.1 PM_1 main components

The main components of PM_1 such as Soil mass, Sulfate mass and total metal oxide (TMO) mass were evaluated as follows:

$$[Soil] = 2.14[Si] + 1.89[Al] + 2..42[Fe] + 1.95[Ca] + 1.67[Ti] + 1.29[Mn] + 1.35[Na] + 1.67[Mg] \quad (1)$$

$$[Sulfate] = 3.063[S] \quad (2)$$

$$[TMO] = 2.15[Li] + 1.14[Cd] + 1.27[Ni] + 1.31[Cr] + 1.25[Cu] + 1.24[Zn] + 1.08[Pb] + 1.21[K] \quad (3)$$

with $[Al]1.89 = Al_2O_3$ and $3Al_2O_3 = SiO_2$ [2-3]. An estimate of the carbonaceous material mass was determined evaluating the PM_1 missing mass (MM) using the reconstructed mass approach as:

$$PM_1 = [Soil] + [Sulfate] + [TMO] + [MM] \quad (4)$$

Thus, it is expected that the missing mass was largely made of carbonaceous material (CM) [4].

2.2 Linear parametric model

It was originally demonstrated that air pollutants follow a lognormal distribution starting from the “law of proportionate effect” [5]. Davies [6] reported that aerosol atmospheric particles have a distribution that can be represented by lognormal distribution. Thus all considered variables were log transformed to smooth extreme recording influence. The following linear parametric model was estimated:

$$Y_i = \beta_{0s} + \Sigma_k \beta_{ks} X_{ki} \quad (5)$$

Where Y_i is the i[th] observation of log-transformed PM_1, $\beta_{ks}(k=0,1,2,4,\ s=1,2)$ are standard regression coefficients changing with the season s and X_{ki} denotes the log-transformed i[th] value of Soil, TMO, MM and Sulfate. Model's components where evaluated with the usual ANOVA table and the percentage of total variation for each component computed.

This model was used to estimate and asses the role of PM_1 main components and to evaluate the possible differential effect of observation periods on the components contribution [7].

3 Results

Figure 1 shows the measured PM_1 mass fraction versus the fitted PM_1 mass fraction using the parametric linear model (see Equation 5). In this model CM, Sulfate, Soil and TMO explained about 44.32%, 33.56%, 11.4% and 0.2% of the total variance of PM_1, respectively (see Table1). The contributions of CM, Sulfate and Soil to PM_1 were significative with an error of 5%.

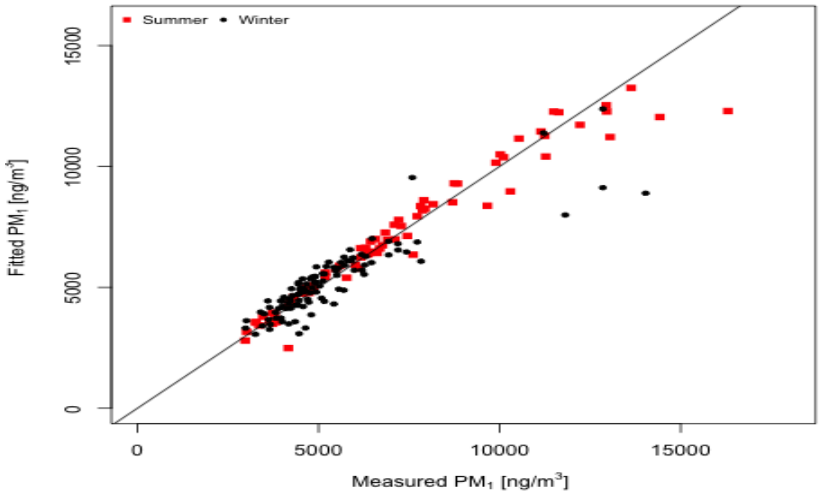

Figure1: Observed PM_1 vs. fitted PM_1 according to model (5). The correlation between fitted and observed is r=0.95, the line is bisector (y=x)

The contribution of TMO to PM_1 was significative with an error of 10%. Carbonaceous material, Soil and TMO contributed to PM_1 with a significant difference between the two considered seasons. Sulfate contributed to PM_1 with non significant difference between Summer and Winter.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	% variance
MM	1	10.2835	10.283	738.228	0.000	40.320
Season	1	0.6292	0.629	45.167	0.000	2.470
Sulfate	1	8.5578	8.558	614.341	0.000	33.560
TMO	1	0.0490	0.049	3.519	0.062	0.200
Soil	1	2.9083	2.908	208.782	0.000	11.400
MM:Season	1	0.4127	0.413	29.623	0.000	1.620
Sulfate:Season	1	0.0003	0.0003	0.019	0.889	0.001
TMO:Season	1	0.1372	0.137	9.847	0.002	0.540
Soil:Season	1	0.1014	0.101	7.279	0.007	0.400
Residuals	174	2.4238	0.0140			
		25.5032				

Table1: ANOVA table from model (5).

The Carbonaceous material component and the Sulfate component which were mainly related to combustion of fossil fuels, industrial activities and inorganic aerosol formation (anthropogenic emission sources) explained about 78% of the PM_1 total variance. The Soil component which was mainly related to possible African dust and to the re-suspension from fields or bare lands of crustal material (natural emission sources) explained about 11.4% of the PM_1 total variance. The TMO which was also related to possible fuels-oils combustion and to traffic related activities explained about 0.2% of the PM_1 total variance. The carbonaceous material component, Soil component and TMO component contributed differently to PM_1 in Summer and Winter (see Table1). In Summer period the processes of resuspension of Soil can be facilitated from dry weather conditions which can involve significative variations of the emissions of crustal material in to the air. Whereas, during Winter the combustion of fossil fuels and firewood due to domestic heating come in to play determining a significative variations of the emissions of carbonaceous material and chemical elements in to the air. As such the observed seasonality for Soil, TMO and CM components. The Sulfate component contributed to PM_1 with non significant difference in the two seasons. This result indicates that the emission sources of Sulfate and the relating anthropogenic activities remain unchanged in the two considered seasons.

Eventually residuals analysis was elaborated showing a heavy tail for larger observations, omoschedastic behavior and negligible autocorrelation

4 Conclusions

The parametric linear model allows for the assessment of the main components of PM_1 and the evaluation of their seasonal variations. In the studied area the anthropogenic emission sources were responsible for about 78% of the PM_1 total variance, while natural emission sources explained about 11.4% of the PM_1 total variance. Significative differential effect of the observation periods on the CM, Soil and TMO components were observed. This suggests that the studied area is characterized from possible anthropogenic as well as natural emission sources with seasonal characteristics.

References

[1]Pope III, C. A., and Dockery, D. W. (2006). Health effects of fine particulate air pollution: lines that connect. Journal of the air & waste management association, 56(6), 709-742.

[2] Trippetta, S., Sabia, S., Caggiano, R. (2016). Fine aerosol particles (PM 1): natural and anthropogenic contributions and health risk assessment. Air Quality, Atmosphere & Health, **9**(6), 621-629.

[3] Malm, W. C., Sisler, J. F., Huffman, D., Eldred, R. A., & Cahill, T. A. (1994). Spatial and seasonal trends in particle concentration and optical extinction in the United States. Journal of Geophysical Research: Atmospheres, 99(D1), 1347-1370.

[4] Prakash, J., Lohia, T., Mandariya, A. K., Habib, G., Gupta, T., Gupta, S. K. (2018). Chemical characterization and quantitativ e assessment of source-specific health risk of trace metals in PM 1.0 at a road site of Delhi, India. Environmental Science and Pollution Research, 25(9), 8747-8764.

[5]Aitchison, J., and Brown, J. (1957). The lognormal distribution with special reference to its uses in economics. New York, NY: Cambridge University Press.

[6]Davies, C. N. (1974). Size distribution of atmospheric particles. Journal of Aerosol Science, 5(3), 293-300.

[7]R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation

for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Induced earthquakes and the ETAS model

Z. Varty[1,*], J. Tawn [1], and S. Bierman[2]

[1] *Lancaster University, Lancaster, UK; z.varty@lancs.ac.uk, j.tawn@lancs.ac.uk*
[2] *Shell Technology Centre, Amsterdam, NL; Stijn.Bierman@shell.com*
[*] *Corresponding author*

Abstract. *The epidemic type aftershock sequence (ETAS) model is widely used in the modelling of earthquake catalogues that include aftershocks. The model has be used successfully in describing tectonic seismicity where the usable catalogue sizes are large. The model is more difficult to apply to induced earthquakes, where catalogue sizes are typically much smaller and the seeding rate of main shocks cannot be assumed to be constant. In both cases, the parameters of the ETAS model are highly correlated under the conventional parameterisation and the resulting log-likelihood function has many flat regions, which can make inference difficult.*
We will introduce issues that arise when modelling induced seismicity caused by gas extraction and put forward an alternative parameterisation for the aftershock component of the ETAS model. The standard ETAS model is nested within our alternative but the correlation of aftershock parameters is greatly reduced. This means that that inference can be made on a broader class of models and more effectively, allowing more model uncertainty to be propagated into earthquake forecasts and simplified parameter interpretation.

Keywords. *Seismic risk; Point processes; Extreme value theory; Spatio-temporal modelling.*

1 Introduction

The Groningen region of the Netherlands does not experience tectonic seismicity. It does, however, contain the largest field of natural gas in Europe. This gas field supplies homes and industries in the Netherlands, Belgium, Germany and France, where gas-powered appliances are specialised to the gas from this field. Despite the Groningen region not being tectonically active, seismic events have been recorded there since the early 1990s. Gas extraction induces these events but there are still questions on the form of the relationship between the two. Figure 1 shows the relationship between one feature of gas extraction and the density of induced events. Understanding these links is critical to informed decision making about future extraction from the Groningen field, based on the associated seismic hazard.

There has been substantial investment into the investigation of this relationship, including improvements to the network of geophones that cover the gas field. It is important to be able to detect and model events with small magnitudes because the gas field is only 3km below surface level and so small magnitude events are still capable of causing damage. The investment has funded a dense monitoring network across the gas field, which can now reliably detect all events down to 1.1 Mw. This value, known as the

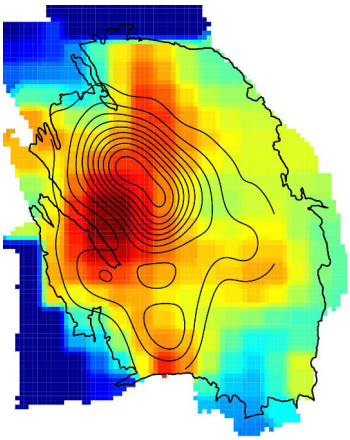

Figure 1: Cumulative compaction caused by gas extraction with event density contours overlaid.

magnitude of completion, has decreased over time but this is not usually accommodated into the model fitting process. The increased ability to detect small magnitude earthquakes is apparent in Figure 2.

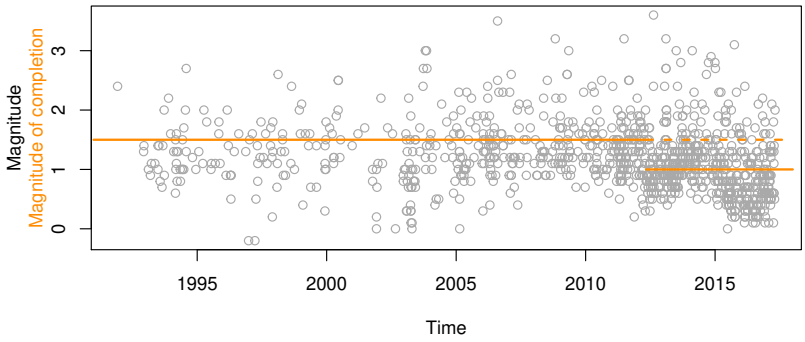

Figure 2: Magnitudes of recorded events and field-wide magnitude of completion.

Modelling seismicity in the Groningen field has additional challenges and opportunities as compared to the usual tectonic setting, these include:

- Covariates such as the cumulative compaction of the gas field in Figure 1 are available but the best way to incorporate these is unclear;
- The variable rate of induced events makes potential aftershock activity difficult to identify;
- The magnitude of completion is decreasing with time but also varies spatially;
- The usable catalogue is small, containing only a few hundred events.

Models that exploit these opportunities and address these challenges could improve our ability to predict

induced seismicity, which is the first step in evaluating future seismic hazard and comparing production scenarios.

2 The ETAS model

The epidemic type aftershock sequence (ETAS) model is currently the standard statistical approach to incorporating aftershock activity. This model is a special case of the Hawkes point process, the class of point process models in which the intensity function λ is dependent on the history of the process, $\mathcal{H}_t$. In particular, the ETAS model locally augments a background intensity function, μ, with an increase in intensity after each earthquake. The size of this increase is determined according to the magnitude of each earthquake and reduces with time and distance from the epicentre. Figure 3 shows an example of such an intensity function for a temporal ETAS point process.

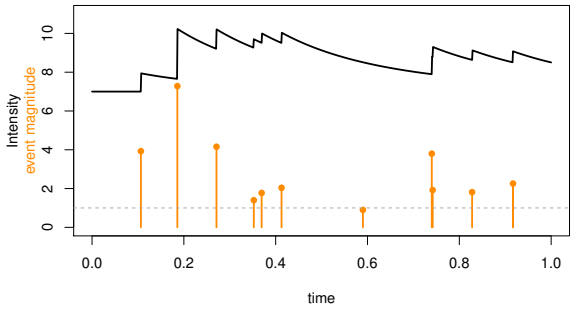

Figure 3: Simulated temporal ETAS catalogue and the associated intensity function.

It is simple to extend the ETAS model to incorporate covariates within the background intensity and to generalise to a spatio-temporal setting, as in Equation (1) . Selecting an appropriate parametric or semi-parametric form for $\mu(x,y,t|X,\theta)$ provides a way of linking gas extraction covariates X and the level of induced seismicity. The functions κ, g and h then describe the aftershock activity by respectively controlling the expected number of aftershocks, their lag and their displacement from the triggering earthquake.

$$\lambda(x,y,t|X,\mathcal{H}_t,\theta) = \mu(x,y,t|X,\theta) + \sum_{i:t_i<t} \kappa(m_i|\theta)g(t-t_i|\theta)h(x-x_i,y-y_i|\theta). \quad (1)$$

3 Reparameterisation

It is well known that the ETAS model is difficult to fit, particularly to small earthquake catalogues like that of Groningen. This is partly because the model was developed in the tectonic setting, where much larger catalogues are available and a temporally constant background rate may be assumed. There are

also issues with the conventional choices for the functions κ, g and h, which are motivated by empirical relationships seen in the tectonic setting. The conventional choice is for κ to be an exponentially increasing function above some threshold M_0. The functions g and h are conventionally described by the modified-Omori law, a heavy tailed power-law distribution, in Δt and $r^2 = \Delta x^2 + \Delta y^2$ respectively. This choice of aftershock functions results in a log-likelihood function that is almost flat in many regions of the parameter space and parameters which are strongly correlated [1]. These issues make both frequentist and Bayesian approaches to inference on the ETAS model difficult.

We suggest alternative forms for the aftershock terms in the ETAS model, within which the current standard choices are nested. The reparameterisation centres the effect of magnitude on aftershock productivity and uses a generalised Pareto distribution rather than the modified-Omori law to describe aftershock lags and displacements. The resulting version of the ETAS model is more flexible than the current approach and is able to describe short-tailed delay and displacement distributions. By using the alternative forms the parameter dependence is greatly reduced, and is negligible within the range of models covered by the conventional parameterisation. The reduction in parameter dependence can be seen in Figure 4.

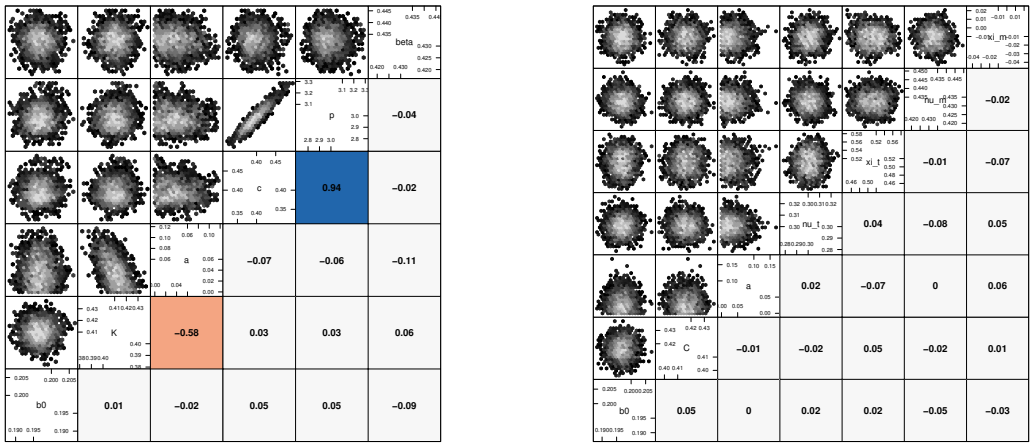

Figure 4: Posterior samples & correlations for a simulated ETAS catalogue using the conventional aftershock parameterisation (left) and the centred generalised Pareto parameterisation (right).

The resulting model allows for more effective inference to be performed, simplifies parameter interpretation and carries uncertainty in the shape of the delay distributions into earthquake forecasts. This is particularly important for application of the model to small catalogues of induced earthquakes such as that of the Groningen gas field.

References

[1] Veen, A. and Schoenberg, F. P. (2008). Estimation of Space-Time Branching Process Models in Seismology Using an EM-Type Algorithm. *Journal of the American Statistical Association* **103** 614–624.

Prior specification in one-factor mixed models applied to community ecology data

Ventrucci M[1,*], Burgazzi G[2], Cocchi D[1] and Laini A [2]

[1] *University of Bologna, Department of Statistical Sciences, massimo.ventrucci@unibo.it* [*]*Corresponding author*
[2] *Department of Chemistry, Life Sciences and Environmental Sustainability, University of Parma*

Abstract. *In community ecology studies the goal is to evaluate the effect of environmental covariates on a response variable while investigating the nature unobserved heterogeneity. We focus on one-factor mixed models in a Bayesian setting and introduce an intuitive Penalized Complexity (PC) prior to balance the variance components of the model. We start with the simple one-way anova and discuss extension to spatially structured residuals, following a Matern exponential covariance.*

Keywords. *Bayesian mixed models; Group model; Intra-class correlation; One-way anova; PC prior.*

1 Mixed models in community ecology

When modelling ecological data several authors report high levels of unexplained variation after considering the effect of environmental covariates [1]. In this cases, the linear regression framework is abandoned in favour of linear mixed models. From a statistician's point of view, accounting for lack of independence in the residuals is required to "adjust" estimates of the regression coefficients. From an ecologist's perspective, investigating the type of residual structure is important in itself to improve understanding of, or generating hypothesis on, the underlying ecological community. For instance, residuals that are correlated within some pre-specified groups/clusters of observational units can be associated to interactions between members of the community, including negative (like competition, predation and parasitism) and positive interactions (like mutualism and commensalism).

We analyze macroinvertebrate community data collected in 6 sampling campaigns carried out in three streams tributaries of the Po River (Northern Italy): Nure Stream, Parma Stream and Enza Stream. For each river a sampling area was sampled twice (in summer and winter), the spatial design including fifty random points aligned along several transects (in total, there are 38 transects). At each point, abundance of macroinvertebrates (response) and environmental covariates such as flow velocity, water depth, substrate composition and benthic organic matter were recorded. The main goals are 1) to investigate the role of the environmental covariates and 2) to assess the presence of small scale interactions within macroinvertebrate communities.

The questions above could be addressed by applying the mixed model framework. In mixed models the effect of the observed covariates and unobserved processes can be neatly separated. Assuming a

Gaussian response $\boldsymbol{Y}$ and covariates $\boldsymbol{X}$ the general formulation of a mixed model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}, \quad ; \quad \boldsymbol{b} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_b) \quad ; \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I})$$

where $\boldsymbol{\beta}$ are the fixed effects and $\boldsymbol{b}$ the random effects. The common interpretation in ecology is that the $\boldsymbol{\beta}$'s account for variability explained by *observed* abiotic factors, while the $\boldsymbol{b}$'s account for variability driven by *unobserved* abiotic or biotic factors [4]. Matriz $\boldsymbol{Z}$ incorporates information about the grouping factors under consideration. In our case study, it is expected that observations tend to be similar within the same sampling campaign or the same transect, thus grouping factors to be considered in the following analysis will be *campaign* (a factors with 6 levels) and *transect* (a factor with 38 levels).

1.1 Exchangeable case: one-way anova

Assume data are grouped according to the levels of a certain *grouping factor*, with y_{ij} being the response at unit $i = 1, \dots, m_j$ within group $j = 1, \dots, n$. The simplest mixed model case is one-way anova,

$$\begin{aligned} y_{ij} &= \alpha + \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\beta} + b_j + \varepsilon_{ij} \qquad i = 1, \dots, m_j \quad j = 1, \dots, n \\ b_j &\sim \mathcal{N}(0, \sigma_b^2) \\ \varepsilon_{ij} &\sim \mathcal{N}(0, \sigma_\varepsilon^2) \end{aligned} \tag{1}$$

where b_j's are random effects quantifying *group-specific* deviations from the intercept α and ε_{ij} are i.i.d. noise terms. It is important to note that introducing the group-specific random effects induces correlation among the residuals $(y_{ij} - (\alpha + \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\beta}))$. For this reason, we refer to model (1) as to the *exchangeable* case.

We note that the b_j's and ε_{ij}'s compete to capture the variance unexplained by environmental covariates. The balance between the two components is regulated by the hyper-parameters σ_b^2 and σ_ε^2. In particular, when $\sigma_b^2 = 0$ model (1) corresponds to the linear regression $y_{ij} = \alpha + \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_{ij}$; the ecological conclusion would be that only environmental covariates matter and the rest is i.i.d. variation. Instead, if $\sigma_b^2 > 0$ there is a certain amount of unexplained variability in the data; the interpretation would be that covariates matters but residuals are not independent, investigating the structure in there can give useful insights on the behaviour of the ecological community.

2 Prior specification in one-factor mixed models

Because the estimates of the variance components σ_b^2, σ_ε^2 drive most of the ecological interpretations on the behaviour of underlying communities, the choice of priors for the hyper-parameters σ_b^2, σ_ε^2 is an important aspect of model specification. [3] address this issue in a general class of one-factor Bayesian mixed models: their proposal is to tackle the choice of priors for the variance components jointly, by specifying a prior on the intraclass correlation (ICC) parameter, $\rho = \sigma_b^2/(\sigma_b^2 + \sigma_\varepsilon^2)$. This prior is derived under the Penalized Complexity (PC) prior framework [2]. By definition, a PC prior is an exponential distribution with rate parameter λ defined on a *distance* scale, d. Such distance d quantifies the increased complexity of the model under consideration w.r.t. to its *base model*, in our case the base model being the linear regression $y_{ij} = \alpha + \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_{ij}$. Thus λ is a scaling parameter controlling the degree of shrinkage to the base model and needs to be specified by the expert user/ecologist. Once the PC prior has been scaled according to a given λ, the prior on the original parameter, ρ, can be computed by the change of

variable rule. For a detailed discussion of the principles underpinning the construction of PC priors and their properties see [2].

In [3] group models assuming different correlation structures for the within group residuals are presented and the prior for the associated correlation parameter (e.g., ρ in the exchangeable case) is always derived under the same principles. There are several practical advantages for the user/ecologist. First, the PC prior ensures proper shrinkage to the base model, thus avoiding overfitting. Second, in the exchangeable case, the scaling parameter λ can be elicited upon a prior statement on the ICC, i.e. the proportion of total variance explained by the grouping factor; for instance, one may compute λ such that $\mathbb{P}(\rho < 0.5) = 0.5$. Third, the prior for ρ is actually defined on an underlying distance scale, which is common to all group models (e.g. exchangeable residuals within transects, serially correlated residuals within transect, they both are extension of the same base model). Thus, the intuitive choice of λ based on eliciting the ICC, is one that can be applied in general for any group model.

2.1 Spatially correlated case

In the present work we extend to the case of spatially correlated residuals, according to a Matern exponential covariance. The *spatially correlated* group model is

$$\begin{aligned} y_{ij} &= \alpha + \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\beta} + \theta_{ij} \qquad i = 1,\dots,m_j \quad j = 1,\dots,n \\ (\theta_{ij},\dots,\theta_{mj})^{\mathsf{T}} &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{R}_j(\phi)) \end{aligned}$$

where the correlation matrix depends on a range parameter $\phi > 0$,

$$\boldsymbol{R}_j(\phi) = \begin{bmatrix} 1 & \exp(-u_{1,2}/\phi) & \cdots & \cdots & \exp(-u_{1,m}/\phi) \\ \exp(-u_{2,1}/\phi) & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 & \exp(-u_{m-1,m}/\phi) \\ \exp(-u_{m,1}/\phi) & \cdots & \cdots & \exp(-u_{m,m-1}/\phi) & 1 \end{bmatrix}. \tag{2}$$

Notation $u_{i,h}$ in matrix (2) indicates the euclidean distance between spatial units i and h. We note that the base model is achieved at $\phi = 0$, in which case we are back to i.i.d. residual case. The PC prior for ϕ can be derived numerically.

3 Concluding remarks

We emphasize that a very intuitive aspect of the proposed PC prior on ϕ is that the scaling parameter λ can be chosen according to a prior statement on the ICC, like in the exchangeable case. We believe this intuitive way to define λ provides an easy-to-elicit prior. The user is then able to balance variance components in an intuitive manner, even in complex models where the variance parameters are difficult to interpret.

The poster presentation will focus in particular n the benefits of using PC priors for residual correlation parameters in a model comparison setting. We will discuss comparison of two different one-factor

mixed models, having different residual structures: *exchangeable* residuals within campaign versus *spatially correlated* residuals within campaign. This comparison would provide insights into the strength of spatial correlation in the residuals, as a preliminary answer to the main questions under study in our motivating example, the one about the presence of possible interactions between members of the ecological community.

Acknowledgments. Daniela Cocchi and Massimo Ventrucci are supported by the PRIN 2015 grant project n. 20154X8K23 (EPHASTAT); Gemma Burgazzi is supported by the PRIN 2015 grant project n. 201572HW8F (NOACQUA). Both projects are funded by the Italian Ministry of Education and University.

References

[1] Lamouroux N., Dolédec S and Gayraud S (2004). Biological traits of stream macroinvertebrate communities: effects of microhabitat, reach, and basin filters. *Journal of the North American Benthological Society* **23(3)**: 449–466.

[2] Simpson D, Rue H, Riebler A, Martins T.G. and Sorbye S.H. (2017). Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. *Statistical Science* **32**: 1–28.

[3] Ventrucci M, Burgazzi G, Cocchi D and Laini A (2019). PC priors for residual correlation parameters in one-factor mixed models. *arXiv:1902.08828*.

[4] Warton et al., (2015) So many variables: joint modeling in community ecology. *Trends in Ecology & Evolution* **30(12)**: 766-779.