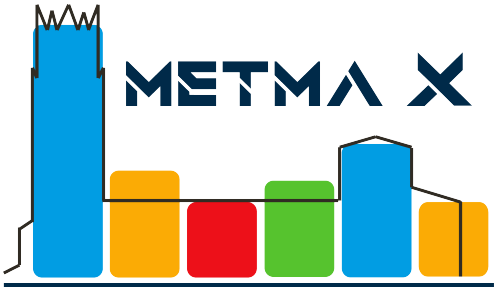
METMA X

Proceedings of the 10th International Workshop on Spatio-Temporal Modelling

Lleida (Spain) 1-3 June 2022

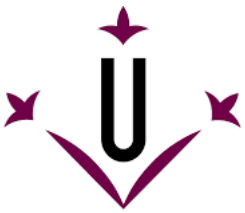


Edited by

Carles Comas and Jorge Mateu

Sponsored by

Universitat de Lleida
Departament de Matemàtica



Universitat de Lleida

Diputació de Lleida

Carles Comas
Department of Mathematics, University of Lleida, Lleida, Spain

Jorge Mateu
Department of Mathematics, University Jaume I, Castellón, Spain

Logo desing: Blanca Capdevila

Layout and cover desing: Carles Comas and Jorge Mateu

ISBN 978-84-9144-364-3

DOI 10.21001/METMA_X

Organizing Committee

Carles Comas
University of Lleida, Spain

Carles Capdevila
University of Lleida, Spain

Cristina Dalfó
University of Lleida, Spain

Nacho López
University of Lleida, Spain

Jorge Mateu
University Jaume I, Spain

Josep Conde
University of Lleida, Spain

Pol Llagostera
University of Lleida, Spain

Sílvia Miquel
University of Lleida, Spain

Scientific Committee

Carles Comas
University of Lleida, Spain

José Miguel Angulo
University of Granada, Spain

Noel Cressie
Univ. Wollongong, NIASRA, Australia

Matthias Eckardt
Humboldt-Universitt zu Berlin, Germany

Alan Gelfand
Duke University, US

Rosalba Ignaccolo
Univ. degli Studi di Torino, Italy

Tomas Mrkvička
University of South Bohemia, Czech Republic

María Dolores Ugarte
Public University of Navarra, Spain

Aila Särkkä
Chalmers University of Technology
and University of Gothenburg, Sweden

Jorge Mateu
University Jaume I, Spain

Adrian Baddeley
Curtin University, Australia

Peter J. Diggle
Lancaster University, UK

María del Pilar Frías
University of Jaen, Spain

Wenceslao Gonzalez-Manteiga
Univ. of Santiago de Compostela, Spain

Raquel Menezes
Universidade do Minho, Portugal

Frederic Schoenberg
Univ. of California at Los Angeles, US

Cristina Vega-García
University of Lleida, Spain

Radu Stoica
Université de Lorraine, France

Technical Secretary

Marta Iglesias
Fundació Universitat de Lleida

Sara Bastien
Fundació Universitat de Lleida

Òscar Clavel
Fundació Universitat de Lleida

Foreword

METMA workshop comes to its 10th edition meaning 20 years of history since its first version held in Castellon (Spain) back in 2001. Since then, its location has been moving along through Spain, Portugal, France and Italy, and has become an international reference for space-time statistics. This edition was aimed to be held in 2020 but was postponed to 2022 due to covid pandemic, and thus it somehow will be remembered as the covid-time edition. The purpose of this conference is to promote the development and application of spatial, temporal, and mainly spatio-temporal statistical methods to different fields related to the environment. The general aim is to bring together practitioners and researchers of different areas and countries all over the world. Cross-disciplinary actions to solve environmental problems are very welcome. The scientific program (http://www.metma-x.udl.cat/) features sessions covering topics on the latest advancements in theory, methods and applications, and presentations from keynotes, invited and a number of oral contributed and posters completes the program.

Spatial statistics has developed rapidly during the last thirty years. We have seen an interesting progress both in theoretical developments and in practical studies. It seems to be honest to remark that the increasing availability of computer power and skilful computer software has stimulated the ability to solve increasingly complex problems. Clearly, these problems have some common elements: they were all of a spatial nature. Some far-reaching theories have developed: image reconstruction, Markov random fields, point process statistics, Geostatistics, and random sets, to mention just a few. As a next stage, these theories were applied successfully to new disciplinary problems leading to modifications and extensions of mathematical and statistical procedures. We therefore notice a general scientific process that has occurred in the field of spatial statistics: well-defined problems with a common character were suddenly on the agenda, and data availability and intensive discussion with practical and disciplinary researchers resulted in new theoretical developments. In this way, spatial statistics has become a refreshing wind in statistics. We do not need to do well much longer on difficult equations, long lists of data and tables with simulated controlled scenarios. But, to be clear, on the back of all these nice pictures a sound science, with sometimes difficult and tedious derivations, and deep thoughts are still required to make serious progress. Spatial statistics has hence emerged as an important new field of

science.

Spatial statistics recognises and exploits the spatial locations of data when designing for, collecting, managing, analysing, and displaying such data. Spatial data are typically dependent, for which there are classes of spatial models available that allow process prediction and parameter estimation. Spatially arranged measurements and spatial patterns occur in a surprisingly wide variety of scientific disciplines. Geology, soil science, image processing, epidemiology, crop science, ecology, forestry, astronomy, atmospheric science, or simply any discipline that works with data collected from different spatial locations, need to develop models that indicate when there is dependence between measurements at different locations. Spatio-temporal variability is a relatively new area within Spatial Statistics, which explains the scarcity of spacetime statistical tools 20 years ago.

Spatial statistics is one of the major methodologies of environmental statistics. Its applications include producing spatially smoothed or interpolated representations of air pollution fields, calculating regional average means or regional average trends based on data at a finite number of monitoring stations, and performing regression analyses with spatially correlated errors to assess the agreement between observed data and the predictions of some numerical model. The notion of proximity in space is implicitly or explicitly present in the environmental sciences. Proximity is a relative notion, relative to the spatial scale of the scientific investigation. When a spatial dimension is present in an environmental study, the statisticians job is to create a statistical framework within which one carries out defensible inferences on processes and parameters of interest. These modelling and inference strategies are not always easy to do, but are never impossible. If Statistics is to continue to be the broker of variability, it must address difficult questions such as those found in the environmental sciences, otherwise it will become marginalised as a discipline. Problems in the environmental sciences are inherently spatial (and temporal), observational in nature, and have experimental units that are highly variable.

In the last decade, spatial statistics has undergone enormous development in the area of statistical modelling. It started slowly, building from models that were purely descriptive of spatial dependence. Then, it became apparent that the process of interest was usually hidden by measurement error, and that the principal goal should be inference on the hidden process from the noisy data. It has only been in the last few decades that the full potential for hierarchical spatial statistical modelling has been glimpsed. There is an enormous amount

of flexibility in hierarchical statistical models, such as the opportunity to account for nonlinearities. Their attractive feature is that at each level of the hierarchy the model specification is simple, yet globally the model can be quite complex. This approach could be summarised as: model locally, analyse globally.

Lately, there has been a rich and growing literature on space-time modelling. Fundamentally, it is clear that in the absence of a temporal component, second-order geostatistical models can be used to represent spatial variability. These are descriptive in the sense that, although they model spatial correlation, there is no causative interpretation associated with them. Thus, for space-time modelling, the geostatistical paradigm assumes a descriptive structure for both space and time (i.e., covariance structures are directly specified). For example, one can extend the geostatistical kriging methodology for spatial processes by assuming that time is just another spatial dimension. Alternatively, one can treat time slices of a spatial field as variables and apply a multivariate or cokriging approach. Although these approaches have been successful in many applications, there are fundamental differences between space and time, and it is not likely that realistic covariance structures can be specified that accurately capture the complicated dynamical processes as found in geophysical applications.

This volume reflects recent contributions that develop theoretical spatial and spatio- temporal statistics to mimic real space-time phenomena. These contributions have been presented at the Tenth International Workshop on Spatio-Temporal Modelling (METMA), which is made possible only thanks to all participants and their continuous support to this type of workshop.

The editors June 2022

Table of Contents

Keynotes

Design and Analysis of Prevalence Surveys for Neglected Tropical Diseases

P.J. Diggle[1,*] , B. Amoah , C. Fronterre , E. Giorgi and O. Johnson .

[1] *CHICAS, Lancaster University. p.diggle@lancaster.ac.uk*
[*] *Corresponding author*

In low-resource settings, where disease registries do not exist, prevalence mapping relies on data collected from surveys of disease prevalence taken in a sample of the communities at risk within the region of interest, possibly supplemented by remotely sensed images that can act as proxies for environmental risk factors. A standard geostatistical model for data of this kind is a generalized linear mixed model,

$$Y_i \sim \text{Binomial}[m_i; P(x_i)],\ \log[P(x_i)/\{1-P(x_i)\}] = d(x_i)\beta + S(x_i) + U_i,\ i = 1,\ldots,n.$$

where Y_i is the number of positives in a sample of m_i individuals at location x_i, $d(x)$ is a vector of spatially referenced explanatory variables available at any location x within the region of interest, $S(x)$ is a Gaussian process and the U_i are iid Gaussian.

In this talk, I will first show how the application of statistical methods associated with this standard model to some Africa-wide control programmes for Neglected Tropical Diseases (NTDs) can bring very substantial gains in efficiency by comparison with the classical survey sampling methods that are currently used in this context. I will then briefly describe some methodological extensions of the standard model to incorporate information from multiple data-sources.

References

[1] Diggle, P.J. and Giorgi, E. (2019). Model-based Geostatistics: Methods and Applications in Global Public Health. Boca Raton: CRC Press.

[2] Fronterre, C., Amoah, B., Giorgi, E., Stanton, M.C. and Diggle, P.J. (2020). Design and analysis of elimination surveys for neglected tropical diseases. *Journal of Infectious Diseases*. doi: 10.1093/infdis/jiz554.

[3] Diggle, P.J., Amoah, B., Fronterre, C., Giorgi, E. and Johnson, O. (2021). Rethinking NTD prevalence survey design and analysis: a geospatial paradigm. textitTransactions of the Royal Society of Tropical Medicine and Hygiene, **115**, 208-210. doi:10.1093/trstmh/trab020.

[4] Johnson, O., Giorgi, E., Fronterre, C., Amoah, B., Atsame, J., Ella, S.N., Biamonte, M., Ogoussan, K., Hundley, L., Gass, K. and Diggle, P.J. (2022). Geostatistical modelling enables effcient safety assessment for mass drug administration with ivermectin in Loa loa endemic areas through a combined antibody and LoaScope testing strategy for elimination of onchocerciasis. PLOS Neglected Tropical Diseases, **16**, e0010189.

Framing a spatio-temporal digital earth concept, around data science, analytics and Statistics

E.M. Scott

School of Mathematics and Statistics, University of Glasgow, Glasgow, UK. Marian.scott@glasgow.ac.uk.

Abstract. *Our understanding of the environment, its connections to biodiversity, health and well being as well as how it is changing are informed by data. The mechanisms and procedures generating those data are continually evolving, which means that the complexity of environmental systems can be studied in greater depth, and hidden connections discovered. Statistical methods also need to evolve to deal with these new data streams. It is in this landscape, that we often see mention of the digital environment agenda, or sometimes digital twin, and more recently digital earth initiatives. These terms all capture the concept that we are studying a temporally evolving system over space, and that monitoring and measurement are essential. Using examples of freshwater quality and biodiversity connectivity, I will illustrate some of the challenges and potential solutions to statistical thinking about a digital earth.*

Fitting and simulating Neyman-Scott cluster process models

A. Baddeley[1,*] , Y.-M. Chang[2] , T.M. Davies[3] , M.L. Hazelton[3] , S. Rakshit[1] and T.R. Turner[4]

[1]*School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, GPO Box U1987, Perth WA 6845, Australia; adrian.baddeley@curtin.edu.au, suman.rakshit@curtin.edu.au*
[2] *Department of Statistics, Tamkang University, 151 Yingzhuan Rd, Tamsui District, New Taipei City 251301, Taiwan (R.O.C.); yamei628@gmail.com*
[3] *Department of Mathematics & Statistics, University of Otago, 730 Cumberland Street, Dunedin North, Dunedin 9016, New Zealand; tilman.davies@otago.ac.nz, martin.hazelton@otago.ac.nz*
[4] *Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand; r.turner@auckland.ac.nz*
[*]*Corresponding author*

Abstract. *When a model for spatial clustering is fitted to a spatial point pattern which is not strongly clustered, algorithms may fail, and parameter estimates may be nonsensical. These failures are not caused by the choice of model-fitting technique, but by weaknesses of the model itself. In particular the model does not include the Poisson process. We propose a new parametrisation of the model involving an index of clustering strength; the Poisson process is included in the model by setting the clustering strength to zero. Close attention to the new parameter space leads to improved performance of fitting algorithms, comprehensible results for the fitted models, and improved algorithms for predicting and simulating cluster process models.*

Keywords. *Brix-Kendall algorithm; Cluster strength; Composite likelihood; Siblings; Total variation.*

---

1. Introduction

Neyman-Scott cluster processes are popular models for spatially clustered patterns of points. Techniques for fitting these models to data are well-established. However, if the evidence for clustering is weak, these fitting techniques often fail: the fitting algorithm fails to converge, or is numerically unstable, or the parameter estimates are extreme, physically implausible values.

In applications, it is reasonable to interpret such failures to mean that a cluster process model was not appropriate. However, a technique which fails *sometimes* does not inspire confidence. Indeed, the technique can also be unreliable for moderately clustered data, which may undermine scientific findings.

It is widely believed that these failures are caused by weaknesses of the fitting method. Consequently, researchers have devoted substantial effort to developing new fitting techniques. Unfortunately, this has not resolved the problem: the new methods all seem to suffer from similar failures.

Here we show that these failures are not caused by the choice of model-fitting technique, but by weaknesses of the model itself. We propose solutions to the problems [1, 2].

2. Motivating example

Figure 1 shows two standard examples of spatial point patterns, both giving the locations of tree saplings: California Giant Redwoods in the left panel, and Japanese Black Pines in the right panel.

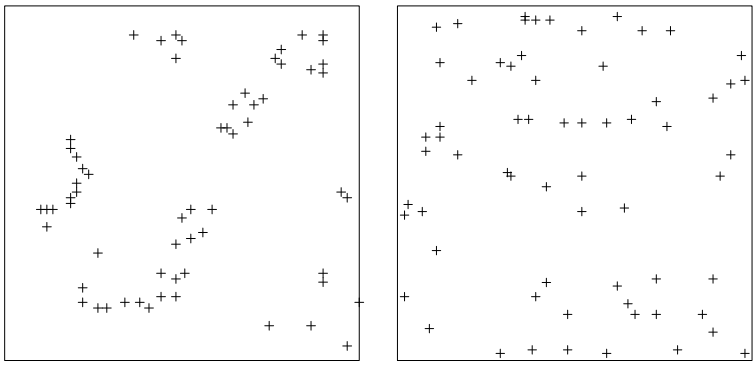


Figure 1: Clustered and non-clustered point patterns. *Left:* California Giant Redwood seedlings and saplings in a 19.2-metre square. *Right:* Japanese Black Pine seedlings and saplings in a 5.7-metre square.

We model each dataset as a realisation of a modified Thomas process, defined as follows [4, 5]. 'Parent' points come from a Poisson point process with constant intensity κ. Each parent is replaced by a random number of 'offspring', according to a Poisson distribution with mean μ, and the offspring are independently displaced from the parent according to an isotropic Gaussian density with standard deviation σ in each coordinate. The model parameters are $\theta = (\kappa, \mu, \sigma)$.

Method	κ	μ	σ	φ	p
Minimum contrast K	0.0641	2.61	0.84	1.75	0.64
Minimum contrast g	0.0635	2.64	0.72	2.40	0.71
Composite likelihood	0.1080	1.55	0.67	1.64	0.62
Palm likelihood	0.0573	2.92	0.79	2.23	0.70

Table 1: Parameter estimates for the Thomas model fitted to the redwoods data.

Table 1 shows parameter estimates for the redwoods data, fitted by minimum contrast to the empirical K-function or pair correlation function $g(r)$, Guan's second order composite likelihood, and Palm likelihood [4, 5]. The estimates broadly agree, and are plausible. Spatial coordinates are in metres, so that κ is the expected number of parents per square metre, and σ is the standard deviation of the offspring displacement in metres. The derived parameters φ, p are explained in the next section.

Table 2 shows the corresponding estimates for the Japanese Pines. There are wide discrepancies between

Method	κ	μ	σ	φ	p
Minimum contrast K	47740	0.00042	50.2	6.6×10^{-10}	6.6×10^{-10}
Minimum contrast g	47740	0.00042	50.2	6.6×10^{-10}	6.6×10^{-10}
Composite likelihood	30	0.06800	6.6	6.2×10^{-5}	6.2×10^{-5}
Palm likelihood	3280	0.00061	30.4	2.6×10^{-8}	2.6×10^{-8}

Table 2: Parameter estimates for the Thomas model fitted to the Japanese Pines data.

the results of different fitting methods, and the parameter estimates are physically implausible.

Extreme values of the fitted parameters cause difficulties with inference, prediction and simulation. Existing simulation algorithms for Neyman-Scott processes fail when the parameters are extreme, because of excessive memory requirements or unacceptably long computation time [1, 3].

The results in Table 2 do have a simple interpretation. Since κ is large and μ is small, the fitted model is close to a Poisson process ("complete spatial randomness"). A Poisson process is an appropriate description of the Japanese Pines data, but does not correspond to any point $\boldsymbol{\theta}$ in the parameter space of the Neyman-Scott model, except in a limiting sense. Despite the unusual behaviour, the algorithm has selected an appropriate model. The fundamental problem with the Neyman-Scott model is that it is not closed under convergence in distribution, and in particular, does not include the Poisson process.

3. Cluster strength

Define the *cluster strength* parameter $\varphi = g(0) - 1$. Then we have the following results [2], which are stated for the Thomas process, but which generalise to any correlation-stationary Neyman-Scott Cox process $\boldsymbol{X}$ with isotropic kernel h. First

$$\varphi = \frac{c}{\kappa\sigma^2} \tag{1}$$

where $c = \|h\|_2^2$ is a constant, equal to $1/(4\pi)$ for the Thomas model. The pair correlation function is

$$g(r) = 1 + \varphi\, a(r/\sigma), \qquad r \geq 0, \tag{2}$$

where $a(r)$ is a nonincreasing function with $a(0) = 1$, determined by h. The model is also a Cox process with driving random intensity function $\Lambda(u)$, $u \in \mathbb{R}^2$, such that for any fixed location u,

$$\mathsf{var}\left[\frac{\Lambda(u)}{\lambda}\right] = \varphi. \tag{3}$$

Indeed the probability distribution of $\Lambda(u)/\lambda$ for fixed u depends only on φ, for the model in question.

Given that there are two points of $\boldsymbol{X}$ at locations separated by a distance $r \geq 0$, the conditional probability (two-point Palm probability) that these two points are *siblings* (offspring of the same parent point) is $p(r) =$

$(g(r)-1)/g(r)$, so that

$$p = p(0) = \frac{\varphi}{1+\varphi} \tag{4}$$

is the sibling probability at distance zero, and conversely $\varphi = p/(1-p)$ is the odds associated with the sibling probability p. For a spatial domain $W \subset \mathbb{R}^2$ of area $|W|$, the total variation distance between the distribution of $\boldsymbol{X}$ and a Poisson process $\boldsymbol{\Pi}$ of the same intensity inside W, satisfies

$$d_{\mathrm{TV}}(\boldsymbol{X}, \boldsymbol{\Pi}) \leq \lambda |W| \sqrt{\varphi}. \tag{5}$$

Consequently, whenever φ is small, the model is close to a Poisson process.

Estimates of φ and p are shown in Tables 1 and 2. Estimated cluster strength for the redwoods is substantial, while for the Japanese pines the fitted model is effectively a Poisson process.

Many of the numerical failures of the fitting algorithms can be avoided by adopting φ as a parameter of the model, and optimising over the space of (φ, σ) pairs. The Poisson process can be included in the model by allowing $\varphi = 0$.

If the parameter vector $\boldsymbol{\theta}$ diverges, the Neyman-Scott model converges to a Poisson process, a mixed Poisson process, a Poisson process with duplicated points, or an improper "explosive" limit [2, Sec. 8–9]. This makes it possible to interpret the fitted model in a comprehensible way in all cases.

4. Cluster scale

The other important parameter of the model is the spatial scale of the clusters. In the Thomas model, cluster scale is controlled by the standard deviation σ. The cluster scale is undefined or unidentifiable when $\varphi = 0$, and is "poorly identified" when the evidence for clustering is weak [4]. This has been another major source of difficulty in fitting Neyman-Scott models.

A standard remedy for unidentifiability is to add a shrinkage penalty to the objective function which is maximised when the model is fitted. The simplest approach is to penalise values of the cluster scale which are physically implausible. We have demonstrated [2] that this strategy improves statistical performance.

For simulation of the fitted model, extreme values of the cluster scale also cause difficulty. If cluster scale is large, the naive or direct simulation algorithm [5] becomes prohibitively slow, while if cluster scale is small, the Brix-Kendall [3] simulation algorithm is prohibitively slow. The insights above lead to improved algorithms for simulating cluster process models, which are 'robust' in the sense that, for a given expected sample size, computation time is uniformly bounded as a function of the model parameters [1]. This makes Monte Carlo inference practical for these models.

References

[1] Baddeley, A. and Chang, Y.-M. (2022). Robust algorithms for simulating cluster processes. Submitted for publication.

[2] Baddeley, A., Davies, T.M., Hazelton, M.L., Rakshit, S. and Turner, T.R. (2022). Fitting spatial cluster process models. Submitted for publication.

[3] Brix, A. and Kendall, W.S. (2002) Simulation of cluster point processes without edge effects. *Advances in Applied Probability* **34**, 267–280.

[4] Diggle, P. (2014). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns* (Third ed.). Chapman and Hall/CRC. Boca Raton, FL.

[5] Møller, J. and Waagepetersen, R. (2004). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman and Hall/CRC. Boca Raton, FL.

Estimation of spatial-temporal point process models using a Stoyan-Grabarnik statistic

C. Kresin[1] and F. Schoenberg[2,*]

[1] *8142 Math-Science Building, Department of Statistics, UCLA, Los Angeles, CA 90095-1554, USA; conorkresin@ucla.edu*
[2] *8142 Math-Science Building, Department of Statistics, UCLA, Los Angeles, CA 90095-1554, USA. frederic@stat.UCLA.edu.*
[*] *Corresponding author*

Abstract. *Parameters in spatial-temporal point process models are typically fit by maximum likelihood estimation (MLE), or some close variant. Here, we show that such parameters can instead be estimated consistently, under general conditions, by instead minimizing the Stoyan-Grabarnik (SG) statistic. More specifically, the spatial-temporal region is divided up into cells, and the sum of squares of the SG statistic is minimized. The resulting estimator has desirable properties, is extremely easy and quick to compute, and does not require approximation of the pesky integral in the log-likelihood formula. Examples and applications to crimes and earthquakes are presented.*

Keywords. *Crimes; Earthquakes; Goodness-of-fit; Hawkes models; Seismology.*

The Stoyan Grabarnik (SG) statistic

$$\bar{m} = \frac{1}{\lambda} \tag{1}$$

was introduced as the exponential "mean mark" in the context of the Palm distribution of marked Gibbs processes [10]. As a primary property of Equation (1), it is noted in [10] that the sum of exponential marks at the points in Borel set $\mathcal{B}$ is equal to the Lebesgue measure of $\mathcal{B}$. The SG statistic has since been proposed as a goodness-of-fit model diagnostic for point processes [1]. More recently, the SG statistic has been used in the context of kernel bandwidth optimization for intensity estimation of point processes [2].

We propose using SG more generally as a tool for estimating the parameters governing the conditional intensity of a space-time point process. Currently, such parameters are typically estimated via maximum likelihood estimation (MLE), as under general conditions, the asymptotic properties of such estimates are consistent and efficient, with standard errors readily constructed using the diagonal elements of the inverse of the Hessian [4, 5].

For a simple stationary process N observed on the time interval $[0,T]$ and on the observed space S, the log

likelihood is given by

$$\begin{aligned} \log \ell(\theta) &= \int_0^T \int_S \log \lambda_\theta(s,t) dN - \int_0^T \int_S \lambda_\theta(s,t) ds dt \\ &= \sum_t \log \lambda_\theta(s,t) - \int_0^T \int_S \lambda_\theta(s,t) ds dt. \end{aligned} \tag{2}$$

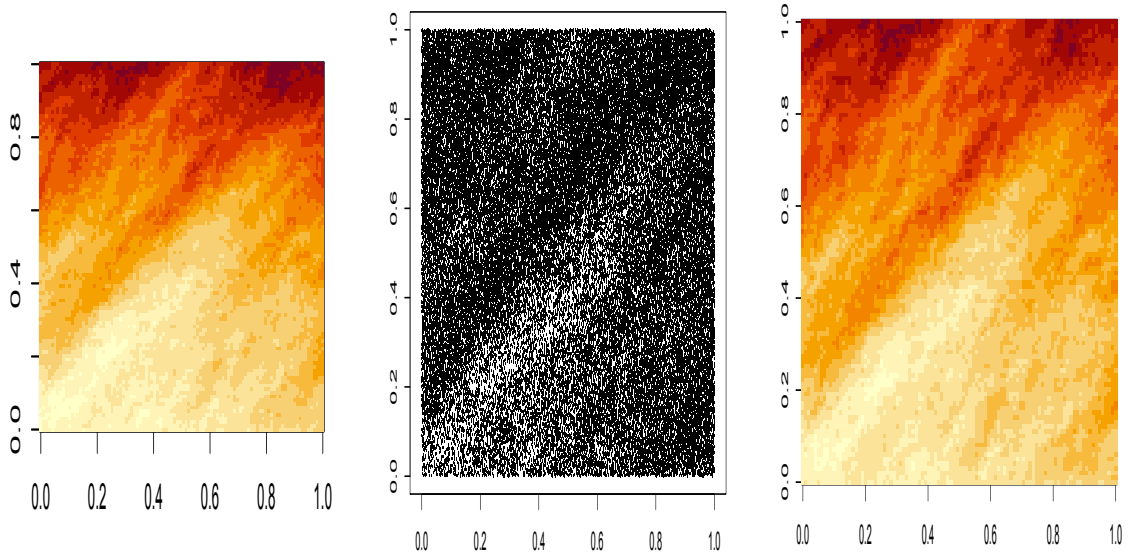


Figure 2: (a) True (simulated) conditional intensity $\lambda(t,x,y) = e^{\alpha x} + \beta e^y + \gamma xy + \delta x^2 + \eta y^2 + w(x,y)/2$, where $(\alpha,\beta,\gamma,\delta,\eta) = (-2,3,4,5,-6)$, and w is a Brownian Sheet on the unit square with variance 100 for t in $[0,1]$; (b) Resulting points from simulation; (c) Conditional intensity with parameters $(\alpha,\beta,\gamma,\delta,\eta)$ estimated by proposed SG method.

Due to the second term, the integral term, in Equation (2), MLE estimates are often computationally intractable for large point processes, or point processes with complex parametric structures, and even approximation of this integral term can be extremely difficult and frought with problems [3, 6, 8]. Further, computing MLEs for intensities with large parameter spaces is often similarly intractable in practice [7]. Complexity of the intensity function is $O(n^2)$ for a point process of size n, and further, flatness in the log-likelihood as a function of the parameters can lead to slow rate of convergence to the optimum [11]. Various remedies have been proposed, including non parametric estimation [6, 8], MLE by way of EM [11], and stochastic de-clustering [12]. Despite computational limitations, maximum likelihood remains the most common method for estimating the parameters of conditional intensities for space-time point processes.

We propose to use the SG statistic as a sort of method of moments type estimator for the parameters θ governing a conditional intensity $\lambda(s,t)$ of a space-time point process. Specifically, for a realization $\{(s_1,t_1),\ldots,(s_n,t_n)\}$ of N on the observation region $\mathcal{S}$ partitioned into p subdivisions, $\{I\}_{j=1}^{p}$, we define the SG estimator as the value of θ minimizing

$$\sum_{j=1}^{p}[\sum_{i:(s_i,t_i)\in I_j}\frac{1}{\lambda_\theta(s_i,t_i)}-\mu(I_j)]^2 \tag{3}$$

We show that, under quite general conditions, the resulting estimator is consistent. The result is verified using simulations. An example is presented in Figure 1, which shows the result of estimating a point process with intensity given by a Brownian Sheet plus exponential plus quadratic in x and y. The benefits of the SG estimator seem to be substantial in terms of programming and computation time. The computationally intensive integral term necessary for the MLE is replaced with a term that is trivial to program and with complexity $O(1)$.

Acknowledgments

Thanks to Jorge Mateu, Carles Comas and the rest of the Organizing Committee, the Department of Mathematics of the University of Lleida, the University of Lleida as well as the SEIO for supporting the METMA X workshop. Thanks also to James Molyneux, Jeff Brantingham, and the Southern California Earthquake Center.

References

[1] Baddeley A., Turner R, Møller J, Hazelton M. (2005). Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(5), 617–666.

[2] Cronie O., van Lieshout M.N.M. (2018). A non-model-based approach to bandwidth selection for kernel estimators of spatial intensity functions. *Biometrika* **105**(2), 455-462.

[3] Harte, D. (2013). Bias in fitting the ETAS model: a case study based on New Zealand seismicity. *Geophys. J. Int.* **92**(1), 390-412.

[4] Krickeberg K. (1982). Processus ponctuels en statistique. in *Ecole d'Eté de Probabilités de Saint-Flour X-1980*, pp. 205–313, Springer, Berlin.

[5] Ogata Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics* **30**(1), 243–261.

[6] Ogata Y., Katsura K. (1988). Likelihood analysis of spatial inhomogeneity for marked point patterns. *Annals of the Institute of Statistical Mathematics* **40**(1), 29–39.

[7] Reinhart A. (2018). A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science* **33**(3), 299-318.

[8] Schoenberg F. (2005). Consistent parametric estimation of the intensity of a spatial–temporal point process. *Journal of Statistical Planning and Inference* **128**(1), 79–93.

[9] Schoenberg, F.P. (2013).Facilitated estimation of ETAS. *Bulletin of the Seismological Society of America* **103**(1), 601–605.

[10] Stoyan D, Grabarnik P (1991). Second-order characteristics for stochastic structures connected with Gibbs point processes. *Mathematische Nachrichten* **151**(1), 95–100.

[11] Veen A, Schoenberg F (2008). Estimation of space–time branching process models in seismology using an EM-type algorithm. *Journal of the American Statistical Association* **103**(482), 614–624.

[12] Zhuang J, Ogata Y, Vere-Jones D (2002). Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association* **97**(458), 369–380.

Spatio-temporal prediction of global carbon-dioxide fluxes at Earth's surface using the fully Bayesian WOMBAT framework

N. Cressie[1,*] , M. Bertolacci and A. Zammit-Mangion

[1]*School of Mathematics and Applied Statistics, University of Wollongong, Wollongong NSW 2522. Australia; ncressie@uow.edu.au*
[*]*Corresponding author*

Abstract. *Locations across Earths surface where the leading greenhouse gas carbon dioxide (CO2) is added to or removed from the atmosphere are known as CO2 sources and sinks. CO2 flux is the rate at which this happens, and a critical goal of carbon-cycle science is to characterise the pattern and scale of sources and sinks in both space and time. There is considerable variability in CO2 fluxes: For example, temperate forests occupy large parts of the terrestrial biosphere and transition from sinks to sources during the year, while volcanoes are local sources with sporadic and unpredictable outgassing of CO2. Human activity has also caused changes to the natural processes that cause these sources and sinks. In this talk, I shall present a framework for predicting fluxes globally and locally in space and time, which is called WOMBAT (the WOllongong Methodology for Bayesian Assimilation of Trace-gases); see [1]. It is fully Bayesian and produces both spatio-temporal predictions and quantifications of their uncertainties. The framework allows scientists and policy makers to take into account uncertainty in CO2flux predictions and, consequently, to produce better mitigation/adaptation strategies for climate change.*

It is now well accepted by the scientific community that greenhouse-gas emissions, unless mitigated, will raise global temperatures, irrevocably alter our climate, and adversely affect Earths ecosystems and human wellbeing. All countries are accountable for their emissions through the 2015 COP21 Paris Agreement, which was adopted in an effort to limit global warming to no more than two degrees Celsius. However, the Paris Agreement framework requires countries to report their own emission estimates, which are based on data and statistics that are sometimes unreliable and incomplete. Validation and corroboration of these estimates is vitally important, to ensure countries commitments to the Agreement, and to ensure that Agreement targets will be reached.

Satellite measurements of CO2 are taken globally and regularly in time, and they are independent of different countries reporting of their inventory targets. WOMBAT, short for WOllongong Methodology for Bayesian Assimilation of Trace-gases, is a fully Bayesian hierarchical modeling framework that produces spatio-temporal predictions of the surface sources and sinks (i.e., fluxes) of greenhouse gases [1]. Prediction is based on (satellite and in situ) measurements of the gass mole fraction after atmospheric transport of its fluxes; that is, the fluxes are not observed directly, only the consequences of that outgassing (sources) and absorption (sinks). The geophysical problem is commonly referred to as flux inversion, and WOMBAT performs this inversion over the entire surface of Earth, resulting in spatio-temporal prediction of the greenhouse-gas fluxes and their uncertainties, at a spatial resolution of 2×2.5 deg. lat-lon and a temporal resolution of one hour.

WOMBAT is the first system to use a fully Bayesian approach to capture uncertainties in the global flux

field, the global mole-fraction field and its measurements, and the parameters used in modelling both the spatio-temporal process errors and the measurement errors. In this talk, we give CO2 flux inversion using mole-fraction data (in ppm) from NASAs Orbiting Carbon Observatory-2 (OCO-2) satellite and from in situ mole-fraction observations collated by the National Oceanic and Atmospheric Administration (NOAA) from across the globe. The full posterior distribution given by WOMBAT yields spatio-temporal predictions of the underlying CO2 fluxes and their uncertainties, as well as posterior inferences on model parameters.

Flux inversion is a highly ill-posed problem that requires one to take into account how the gas moves in the (three-dimensional) atmosphere via a computationally intensive chemical transport model. Flux-inversion frameworks that pre-date WOMBAT are based on quite straightforward, non-hierarchical (empirical) Bayesian statistical models that typically return just flux predictions (and very rarely uncertainty quantifications). WOMBAT is the first to jointly model and account for

- Spatio-temporal correlated error in the chemical transport model,
- Uncertainty in the prior fluxes,
- Uncertainty in the prior fluxes,
- Biases in the satellite retrievals,
- Uncertainties on the reported error statistics of the satellite retrievals.

Statistical computing features extensively in WOMBATs inferential pipeline. Chemical transport models require high-performance computer infrastructure, and dimension-reduced fully Bayesian inference was implemented using Gibbs sampling while employing the use of GPUs to facilitate targeted intensive matrix operations. The WOMBAT framework also employs likelihood approximations that induce matrix sparsity and allow for spatio-temporal correlations. In this talk, WOMBAT is implemented on more than two years of CO2 data from NASAs OCO-2 satellite. Then its flux predictions are compared to those (for the same time period) from a model intercomparison project (MIP), which involved nine flux-inversion groups across the world. This independent validation showed that WOMBATs results always performed favourably against, and sometimes outperformed, those from the other flux-inversion groups.

WOMBAT will be a part of NASAs next OCO-2 MIP to be held in 2022, and it will also form a part of the NASA OCO Science Team report in a contribution to the United Nations 2023 Global Stocktake (an initiative stemming from the 2015 Paris Agreement to assess the global state of emissions that will directly lead to national and international policy recommendations).

In version 2.0 of WOMBAT (Bertolacci et al., in preparation), we build on the foundations laid in version 1.0 to capture changes in the climatology of CO2 fluxes. The talk will end with a brief discussion of this innovation.

References

[1] Zammit-Mangion, A., Bertolacci, M., Fisher, J., Stavert, A., Rigby, M., Cao, Y., and Cressie, N. (2022).WOMBAT v1.0: a fully Bayesian global flux-inversion framework.*Geoscientific Model Development* **15**, 45-73 (doi:10.5194/gmd-15-45-2022)

The role of Preferential Sampling in Spatial and Spatio-temporal Geostatistical Modeling

A.E. Gelfand

Duke University.223A Old Chem Bldg Durham, NC 27708. alan@stat.duke.edu

The notion of preferential sampling was introduced into the literature in the seminal paper of [1]. Subsequently, there has been considerable follow up research. A standard illustration arises in geostatistical modeling. Consider the objective of inferring about environmental exposures. If environmental monitors are only placed in locations where environmental levels tend to be high, then interpolation based upon observations from these locations will necessarily produce only high predictions. A remedy lies in suitable spatial design of the locations, e.g., a random or space-filling design for locations over the region of interest is expected to preclude such bias. However, in practice, sampling may be designed in order to learn about areas of high exposure.

While the set of sampling locations may not have been developed randomly, we study it as if it was a realization of a spatial point process. That is, it may be designed/specified in some fashion but not necessarily with the intention of being roughly uniformly distributed over D. Then, the question becomes a stochastic one: is the realization of the responses independent of the realization of the locations? If no, then we have what is called preferential sampling. Importantly, the dependence here is stochastic dependence. Notationally/functionally, the responses are associated with the locations.

Another setting is the case of species distribution modeling with a binary response, presence or absence, recorded at locations. Here, bias can arise when sampling is designed such that ecologists will tend to sample where they expect to find individuals. This setting can be extended to data fusion where we have both presence/absence data and presence-only data. Other potential applications include missing data settings and hedonic modeling for price with property sales. Very recent work explores preferential sampling in the context of multivariate geostatistical modeling.

Fundamental issues are: (i) can we identify the occurrence of a preferential sampling effect, (ii) can we adjust inference in the presence of preferential sampling, and (iii) when can such adjustment improve predictive performance over a customary geostatistical model? We consider these issues in a modeling context and illustrate with application to presence/absence data, to property sales, and to tree data where we observe mean diameter at breast height (MDBH) and trees per hectare (TPH).

References

[1] Diggle, P.J., Menezes, R. and Su, T. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **59** (2), 191232.

Invited

Spatial analysis of epidermal nerve fiber patterns

K. Konstantinou, , U. Picchini , and A. Särkkä*

Department of Mathematical Sciences, Chalmers University of Technology and the University of Gothenburg, Gothenburg, Sweden; konkons@chalmers.se, picchini@chalmers.se, aila@chalmers.se.
**Corresponding author*

Abstract. *Epidermal nerve fibers (ENFs) are thin sensory nerve fibers in the epidermis, the outermost cell layers in the skin. After they have entered the epidermis, they grow and branch and finally, terminate. Small fiber neuropathies, such as diabetic neuropathy, can cause damage to the ENF structure. For example, it has been established that the ENF density and summed length of ENFs per epidermal surface area are reduced, and ENFs may appear more clustered within the epidermis in subjects suffering from diabetic neuropathy compared to healthy subjects [2]. We have data from healthy subjects and subjects suffering from mild or moderate diabetic neuropathy. We regard the nerve patterns as spatial point patterns consisting of entry points and end points. We concentrate on 2D projections of the 3D data since the main interest is on how the skin is covered by the nerve endings that are responsible for transferring signals, such as heat and pain, to the central nervous system. We will present some point process models for the nerve patterns and compare the healthy and mild patterns in terms of some summary statistics and model parameters [1, 3, 4, 5]. In addition, we would like to understand how a healthy pattern changes as neuropathy advances and present some thinning mechanisms that may explain the change of the ENF structure from healthy to mild neuropathy and further from mild to moderate neuropathy.*

Keywords. *Clustering; Hierarchy; Spatial point pattern; Thinning.*

References

[1] Andersson, C., Guttorp, P., and Särkkä, A. (2016). Discovering early diabetic neuropathy from epidermal nerve fiber patterns. *Statistcs in Medicine* **35**(24), 4427–4442.

[2] Kennedy, W.R., Wendelschafer-Crabb, G., and Johnson, T. (1996). Quantitation of epidermal nerves in diabetic neuropathy. *Neurology* **47**(4), 1042–1048.

[3] Konstantinou, K. and Särkkä, A. (2021). Spatial modeling of epidermal nerve fiber patterns. *Statistics in Medicine* **40**(29), 6479–6500.

[4] Olsbo, V., Myllymäki, M., Waller, L.A., and Särkkä, A. (2013). Development and evaluation of spatial point process models for epidermal nerve fibers. *Mathematical Biosciences* **243**, 178–189.

[5] Ghorbanpour, F., Särkkä, A., and Pourtaheri, R. (2021). Marked point process analysis of epidermal nerve fibers. *Journal of Microscopy* **283**(1), 41–50.

Crop Yield Prediction Using Bayesian Spatially Varying Coefficient Models with Functional Predictors

B. Li

Departament of Statistics, University of Illinois at Urbana-Champaign, USA. libo@illinois.edu.

Abstract. *Reliable prediction for crop yield is crucial for economic planning, food security monitoring, and agricultural risk management. This study aims to develop a crop yield forecasting model at large spatial scales using meteorological variables closely related to crop growth. The influence of climate patterns on agricultural productivity can be spatially inhomogeneous due to local soil and environmental conditions. We propose a Bayesian spatially varying functional model (BSVFM) to predict county-level corn yield for five Midwestern states, based on annual precipitation and daily maximum and minimum temperature trajectories modeled as multivariate functional predictors. The proposed model accommodates spatial correlation and measurement errors of functional predictors, and respects the spatially heterogeneous relationship between the response and associated predictors by allowing the functional coefficients to vary over space. The model also incorporates a Bayesian variable selection device to further expand its capacity to accommodate spatial heterogeneity. The proposed method is demonstrated to outperform other highly competitive methods in corn yield prediction, owing to the flexibility of allowing spatial heterogeneity with spatially varying coefficients in our model. Our study provides further insights into understanding the impact of climate change on crop yield.*

Big problems in spatio-temporal disease mapping, pragmatic solutions

E. Orozco-Acosta[1,2] , A. Adin[1,2] and M. D. Ugarte[1,2,*]

[1] *Department of Statistics, Computer Science and Mathematics, Public University of Navarre, 31006 Pamplona, Spain; erick.orozco@unavarra.es, aritz.adin@unavarra.es, lola@unavarra.es*
[2] *INAMAT2, Public University of Navarre, 31006 Pamplona, Spain*
[*] *Corresponding author*

Abstract. *Much of the research in disease mapping is based on Bayesian hierarchical spatio-temporal models that borrow strength from space and time to smooth the risks and reduce their variability. However, when the number of areas is very large, model fitting is generally time-consuming or even unfeasible. In this talk we will discuss a pragmatic solution based on the idea of "divide and conquer". This is a simple idea that works very well in this context as models are defined to borrow strength locally in space and time, providing reliable risk estimates. We evaluate the new proposal in a simulation study with a twofold objective: to estimate risks accurately and to detect extreme risk areas while avoiding false positives. An analysis of real data will also be discussed.*

Keywords. *Areal data; INLA; Small areas; Scalable models*

Information-based structural complexity analysis of subordinated spatiotemporal random fields

J.M. Angulo[1,*] and M.D. Ruiz-Medina[1]

[1] *University of Granada, Granada, Spain; jmangulo@ugr.es, mruiz@ugr.es.*
[*] *Corresponding author*

Abstract. *In the framework of structural complexity analysis, this work analyzes uncertainty measures, applying information theory, involving the bivariate probability distributions of spatial or spatiotemporal subordinated random fields. Assumptions of homogeneity and isotropy in the spatial case, and also stationarity in time in the spatiotemporal case, are considered. The properties of the formulated spatial and spatiotemporal structural complexity measures are investigated in some cases of interest, including the special case of Minkowski functionals subordinated to Gaussian and Gamma-correlated random fields. The results are illustrated in the context of geometrical analysis of sample paths.*

Keywords. *Gamma-correlated subordinated random fields; Gaussian subordinated random fields; information measures; spatial functional models; structural complexity*

1. Introduction

There is a growing interest on structural complexity analysis based on sojourn measures of spatiotemporal Gaussian and Gamma-correlated random fields. Indeed, there exists a vast literature in the context of stochastic geometrical analysis of the sample paths of random fields based on these measures (see, e.g., Bulinski *et al.* [3]; Ivanov and Leonenko [4], among others). A parallel literature has also been developed in the context of long-range dependent random fields (see Leonenko [6]; Leonenko and Olenko [7]; Makogin and Spodarev [10]). Recently, Leonenko and Ruiz-Medina [8] derive reduction theorems and central and non-central limit results for the asymptotic analysis of functionals of spatiotemporal Gaussian subordinated random fields involving these measures. As a motivation, the special case of Minkowski functionals (see, e.g., [11]) allows, for instance in 2D, the geometrical interpretation of the total area of all 'hot' regions, the total length of the boundary between 'hot' and 'cold' regions, and the Euler characteristic, which counts the number of isolated 'hot' regions minus the number of isolated 'cold' regions within the 'hot' regions.

Information measures have played a fundamental role in probabilistic-statistical theoretical and applied research, with a vast related literature disseminated in a wide variety of knowledge areas. In particular, entropy and divergence measures are often used for informational characterization and comparative assessment of probability distributions describing structural aspects of stochastic systems. Divergence measures constitute

the basis for definition of certain forms of mutual information useful for dependence quantification. Furthermore, product-type complexity measures are also constructed based on entropy and divergence measures. In some cases, a natural interpretation can be given in terms of diversity; a review and discussion, particularly in reference to Rényi-information related measures, is given in [1].

2. Subordinated random fields

Let $(\Omega, \mathcal{A}, P)$ be the basic complete probability space, and denote by $\mathcal{L}^2(\Omega, \mathcal{A}, P)$ be the Hilbert space of zero-mean second-order random variables on $(\Omega, \mathcal{A}, P)$. Consider $X = \{X(\mathbf{z}),\ \mathbf{z} \in D \subseteq \mathbb{R}^d\}$ to be a zero-mean spatial homogeneous and isotropic mean-square continuous second-order random field, with correlation function $\gamma(\|\mathbf{x}-\mathbf{y}\|) = \mathrm{Corr}\,(X(\mathbf{x}), X(\mathbf{y}))$. Assume that the marginal probability distributions are absolutely continuous, having probability density $p(u)$. Let now $L^2((a,b), p(u)du)$, $-\infty \leq a < b \leq \infty$, be the Hilbert space of equivalence classes of measurable real-valued functions on the interval (a,b) which are square-integrable with respect to the measure $\mu(du) = p(u)du$.

Assume that there exists a complete orthonormal basis $\{e_k,\ k \geq 0\}$, with $e_0 = 1$, of the space $L^2((a,b), p(u)du)$ such that

$$\frac{\partial^2}{\partial u \partial v} P\left[X(\mathbf{x}) \leq u, X(\mathbf{y}) \leq v\right] =: p(u, v, \|\mathbf{x}-\mathbf{y}\|) = p(u)\ p(v)\ \left[1 + \sum_{k=1}^{\infty} \gamma^k(\|\mathbf{x}-\mathbf{y}\|)\ e_k(u)\ e_k(v)\right]. \qquad (1)$$

The family of random fields X satisfying the above conditions is known as the *Lancaster-Sarmanov* random field class (see, e.g., Lancaster [5], Leonenko *et al.* [9], and Sarmanov [13]). Note that this class is not empty since, for instance, both the Gaussian and Gamma-correlated random field classes satisfy the above introduced conditions. In particular, for the Gaussian random field case, $\{e_k,\ k \geq 0\}$ coincides with the (normalized) Hermite polynomial system (see, for example, [12]), and $\{e_k,\ k \geq 0\}$ are the Laguerre polynomials in the case of Gamma-correlated random fields. An interesting special case of the latter is defined by the chi-square random field family.

It is well-known that non-linear transformations of these random fields can be approximated in terms of the above series expansions, since, for every $g \in L^2((a,b), p(u)du)$,

$$g(x) = C_0^g + \sum_{k=m}^{\infty} C_k^g e_k(x), \quad C_k^g = \int_{(a,b)} g(u) e_k(u) p(u) du,\ k \geq 0, \qquad (2)$$

where m denotes the rank of function g in the orthonormal basis $\{e_k,\ k \geq 1\}$. In the particular case of Gaussian and Gamma-correlated subordinated random fields we will refer to the Hermite and Laguerre ranks, respectively, of function g.

As mentioned in Section 1, an interesting example is Minkowski functional $M_0(\nu; X, D) = \int_D 1_\nu(X(\mathbf{y}))d\mathbf{y} = \lambda(S_{X,D}(\nu))$, which is defined from $g(x) = 1_\nu(x)$, the indicator function based on thresh-

old ν, with $S_{X,D}(\nu) = \{\mathbf{z} \in D;\ X(\mathbf{z}) \geq \nu\} = \{\mathbf{z} \in D;\ g(X(\mathbf{z})) = 1\}$. Thus, in the Gaussian case, with $p(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)$, equation (2) leads to $1_\nu(x) = \sum_{k=0}^{\infty} \frac{G_k(\nu)}{k!} \mathcal{H}_k(x)$, where $\{\mathcal{H}_k,\ k \geq 0\}$ denotes the basis of Hermite polynomials, and $G_k(\nu) = \left\langle 1_\nu, \mathcal{H}_k \right\rangle_{L^2((a,b),p(u)du)}$, with $G_0(\nu) = \frac{1}{\sqrt{2\pi}} \int_\nu^\infty \exp(-u^2/2)du$, and $G_k(\nu) = \frac{1}{\sqrt{2\pi}} \exp(-\nu^2/2)\mathcal{H}_{k-1}(\nu)$, $k \geq 1$. Hence, $M_0(\nu; X, D) = \sum_{k=0}^{\infty} \frac{G_k(\nu)}{k!} \int_D \mathcal{H}_k(X(\mathbf{y}))d\mathbf{y}$ (see [8]).

The following assumption on the correlation function γ is considered:

Assumption I.

$$\begin{aligned} 1-\gamma(\|\mathbf{x}-\mathbf{y}\|) &= O(\|\mathbf{x}-\mathbf{y}\|^{\alpha}), \quad \|\mathbf{x}-\mathbf{y}\| \to 0, \quad \alpha \in (0,d) \\ \gamma(\|\mathbf{x}-\mathbf{y}\|) &= O\left(\|\mathbf{x}-\mathbf{y}\|^{-\rho}\right), \quad \|\mathbf{x}-\mathbf{y}\| \to \infty, \quad \rho \in (0,d). \end{aligned} \tag{3}$$

In the following section we apply equations (1) and (3) in the derivation of asymptotic orders at spatial microscale and macroscale of mutual information between the marginal components of Lancaster-Sarmanov subordinated random fields. Under **Assumption I**, these asymptotic orders are related to the fractality and long-range dependence parameters of the underlying Lancaster-Sarmanov random field. The results can be interpreted as a limit (infinitesimal, infinite) analysis in the structural complexity framework based on Rényi entropy.

3. Mutual information and spatial structural complexity

Let $\{X(\mathbf{x}),\ \mathbf{x} \in D\}$ be an element of the Lancaster–Sarmanov random field class. From equation (1), mutual information between component r.v.'s $X(\mathbf{x})$ and $X(\mathbf{y})$ can be computed as follows:

$$\mathcal{S}_{\alpha,\rho}(\|\mathbf{x}-\mathbf{y}\|) := I(X(\mathbf{x}), X(\mathbf{y})) = \int_a^b \int_a^b p(u,v,\|\mathbf{x}-\mathbf{y}\|) \ln\left(\frac{p(u,v,\|\mathbf{x}-\mathbf{y}\|)}{p(u)p(v)}\right) dudv$$

$$= \int_a^b \int_a^b p(u)\ p(v) \left[1 + \sum_{k=1}^{\infty} \gamma^k(\|\mathbf{x}-\mathbf{y}\|)\ e_k(u)\ e_k(v)\right] \ln\left(1 + \sum_{k=1}^{\infty} \gamma^k(\|\mathbf{x}-\mathbf{y}\|)\ e_k(u)\ e_k(v)\right) dudv. \tag{4}$$

Under **Assumption I**, the asymptotic behavior of $\mathcal{S}_{\alpha,\rho}(\|\mathbf{x}-\mathbf{y}\|)$ when $\|\mathbf{x}-\mathbf{y}\| \to 0$ involves the fractality parameter α, providing an indicator of spatial structural complexity at microscale, whose maximum is attained for α values close to 0. While when $\|\mathbf{x}-\mathbf{y}\| \to \infty$, the asymptotic behaviour of $\mathcal{S}_{\alpha,\rho}(\|\mathbf{x}-\mathbf{y}\|)$ involves the long-range dependence (LRD) parameter ρ, providing an indicator of spatial structural complexity at macroscale, whose minimum is also attained for ρ values close to 0. For $g \in L^2((a,b), p(u)du)$ a similar asymptotic behavior is displayed by mutual information $I(g(X(\mathbf{x})), g(X(\mathbf{y})))$ when $\|\mathbf{x}-\mathbf{y}\| \to \infty$, involving the LRD parameter ρ scaled by the rank m of function g in the orthonormal basis $\{e_k,\ k \geq 1\}$ (Hermite and Laguerre ranks in the Gaussian and Gamma-correlated cases, respectively). However, the spatial microscale behavior of $I(g(X(\mathbf{x})), g(X(\mathbf{y})))$ is not affected by the rank m of function g.

4. Final comments on the spatiotemporal case

The formulation of mutual information as a measure for spatiotemporal structural complexity analysis can be addressed, for the general class of Lancaster-Sarmanov random fields (see [9]), adopting the infinite-dimensional spatial framework introduced in [2].

Let $X = \{X_{\mathbf{x}}(\cdot),\ \mathbf{x} \in D\}$ be a zero-mean homogeneous and isotropic spatial functional random field on the separable Hilbert space $(H, <\cdot,\cdot>_H)$, mean-square-continuous w.r.t. the H norm. In the following, we will assume that $H = L^2(\mathcal{T})$, with $\mathcal{T} \subseteq \mathbb{R}_+$. For every $\mathbf{x},\mathbf{y} \in D$, $(X_{\mathbf{x}}(\cdot), X_{\mathbf{y}}(\cdot))^T$ is a random element in the separable Hilbert space $\left(H^2, \langle\cdot,\cdot\rangle_{H^2}\right)$ of vector functions $\mathbf{f} = (f_1, f_2)^T$, with the inner product given by $\langle \mathbf{f}, \mathbf{g}\rangle_{H^2} = \sum_{i=1}^2 \langle f_i, g_i\rangle_H$, $\forall \mathbf{f},\mathbf{g} \in H^2$. Thus, for every $\mathbf{x},\mathbf{y} \in D \subseteq \mathbb{R}^d$, we consider the measurable function $(X_{\mathbf{x}}(\cdot), X_{\mathbf{y}}(\cdot))^T$: $(\Omega, \mathcal{A}, P) \longrightarrow \left(H^2, \mathcal{B}(H^2), P_{X_{\mathbf{x}}(\cdot),X_{\mathbf{y}}(\cdot)}(dh_1, dh_2)\right)$. Let us denote by $\{P_{X_{\mathbf{x}}(\cdot)}(dh),\ \mathbf{x} \in D\}$ the marginal infinite-dimensional probability distributions, with $P_{X_{\mathbf{x}}(\cdot)}(dh) = P(dh)$, for every $\mathbf{x} \in D$. Let $L^2(H, P(dh))$ be the space of measurable functions $g : H \longrightarrow H$ such that $\int_H \|g(h)\|_H^2 P(dh) < \infty$. Assume that there exists an orthonormal basis $\{\mathcal{B}_k,\ k \geq 1\}$ of $L^2(H, P(dh))$ such that

$$P_{X_{\mathbf{x}}(\cdot),X_{\mathbf{y}}(\cdot)}(dh_1, dh_2) = p_{X_{\mathbf{x}}(\cdot)}(h_1) p_{X_{\mathbf{y}}(\cdot)}(h_2) \left[1 + \sum_{k\geq 1} \gamma^k_{\|\mathbf{x}-\mathbf{y}\|}(\cdot,\cdot)\mathcal{B}_k(h_1)\mathcal{B}_k(h_2)\right] dh_1 dh_2,$$

with $\gamma_{\|\mathbf{x}-\mathbf{y}\|}(\cdot,\cdot) = \mathrm{Corr}(X_{\mathbf{x}}(\cdot), X_{\mathbf{y}}(\cdot))$ being the kernel of the spatial correlation operator $\gamma_{\|\mathbf{x}-\mathbf{y}\|}$, for every $\mathbf{x},\mathbf{y} \in \mathcal{D}$.

We refer to the class of spatial functional random fields satisfying the above conditions as the functional version of Lancaster-Sarmanov random fields.

In a similar way to Section 3, under **Assumption I**, the asymptotic behavior at spatial microscale and macroscale can be studied from the infinite-dimensional formulation of Kullback-Leibler divergence given in [2], as well as its extension based on Rényi divergence.

Acknowledgments

This work has been supported in part by grants MCIU/AEI/ERDF, UE PGC2018-098860-B-I00 and PGC2018-099549-B-I00, grant A-FQM-345-UGR18 cofinanced by ERDF Operational Programme 2014-2020 and the Economy and Knowledge Council of the Regional Government of Andalusia, Spain, and grant CEX2020-001105-M MCIN/AEI/10.13039/501100011033.

References

[1] Angulo, J.M., Esquivel, F.J, Madrid, A.E., and Alonso F.J. (2021) Information and complexity analysis of spatial data. *Spatial Statistics* **42**, 100462.

[2] Angulo, J.M., and Ruiz-Medina, M.D. (2022) Infinite-Dimensional Divergence Information Analysis. In: *Trends in Mathematical, Information and Data Sciences*, N. Balakrishnan, M.A. Gil, N. Martín, D. Morales and M.C. Pardo (eds.). Springer (in press)

[3] Bulinski, A., Spodarev, E., and Timmermann, F. (2012) Central limit theorems for the excursion volumes of weakly dependent random fields. *Bernoulli* **18**, 100–118.

[4] Ivanov, A.V., and Leonenko, N.N. (1989) *Statistical Analysis of Random Fields*. Dordrecht: Kluwer Academic.

[5] Lancaster, H.O. (1958) The structure of bivariate distributions. *The Annals of Mathematical Statistics* **29**, 719–736.

[6] Leonenko, N.N. (1999) *Limit Theorems for Random Fields with Singular Spectrum*. Mathematics and its Applications **465**. Dordrecht: Kluwer Academic.

[7] Leonenko, N.N., and Olenko, A. (2014) Sojourn measures of Student and Fisher-Snedecor random fields. *Bernoulli* **20**, 1454–1483.

[8] Leonenko, N.N., and Ruiz-Medina, M.D. (2022) Sojourn functionals for spatiotemporal Gaussian random fields with long-memory. *Advances in Applied Probability* (in press).

[9] Leonenko, N.N., Ruiz-Medina, M.D., and Taqqu, M.S. (2017) Non-central limit theorems for random fields subordinated to Gamma-correlated random fields. *Bernoulli* **23**, 3469–3507.

[10] Makogin, V., and Spodarev, E. (2022) Limit theorems for excursion sets of subordinated Gaussian random fields with long-range dependence. *Stochastics* **94**, 111–142.

[11] Marinucci, D., and Peccati, G. (2011) *Random Fields on the Sphere. Representation, Limit Theorems and Cosmological Applications*. London Mathematical Society Lecture Note Series **389**. Cambridge: Cambridge University Press.

[12] Peccati, G., and Taqqu, M.S. (2011) *Wiener Chaos: Moments, Cumulants and Diagrams*. New York: Springer.

[13] Sarmanov, O.V. (1963) Investigation of stationary Markov processes by the method of eigenfunction expansion. *Selected Translations in Mathematical Statistics and Probability* **4**, 245–269.

Modeling complex-valued random fields in environmental sciences

S. De Iaco

DES- Section of Mathematics and Statistics, Universitá del Salento, Italy. sandra.deiaco@unisalento.it

Abstract. *In geostatistical literature, the study of the evolution of vector data with two components in space and space-time is often developed in the framework of the theory of complex-valued random fields. This formalism can reflect the specific characteristics of the components of a vectorial random field, which are associated with a physical phenomenon described by homogeneous quantities, expressed in the same unit of measure, such as wind velocity, force, electric or magnetic field, and available at the same spatial points over the domain. In this case, the corresponding realization of a vectorial field in two dimensions is an expression of a single entity, i.e., a complex number, where the decomposition in modulus and angle is natural and has a physical interpretation. Thus, a compact and unified formalism has to be suitably adopted for this kind of data. In other terms, the use of a complex formalism and consequently of complex modeling for prediction purposes is not an arbitrary choice, but it is strictly motivated by the nature of the phenomenon.*
The aim of this work is to introduce the theoretical background regarding the complex formalism of a spatial and spatio-temporal random field and to present some families of complex-valued covariance models.

Bayesian MCMC inference for complex cluster models

T. Mrkvička

Department of Applied Mathematics and Informatics, Faculty of Economics, University of South Bohemia, Studentská 13, 37005 České Budějovice, Czech Republic; mrkvicka.toma@gmail.com.

Abstract. *The stationary Neyman-Scott point process can be extended for inhomogeneity in many ways. The center points, cluster sizes or cluster spread can be inhomogeneous. Also a combination of these types can be of interest. Further, the distribution of cluster sizes can be non Poisson, usually the Poisson distribution is assumed. We consider all these models and propose the Bayesian MCMC algorithms to estimate parameters of these models. The Bayesian MCMC approach is tractable for all these models and in cases where the faster methods can be applied it gives more precise results. We developed an R package binspp avaylable at* `https://github.com/tomasmrkvicka/binspp` *which contains these algorithms.*

Keywords. *Double inhomogeneity; Generalised Poisson distribution; Inference for covariate effect; Inhomogeneity; Neymann-Scott point process*

The Neyman-Scott point process model is widely used cluster point process model due to its straightforward interpretability in biology, astronomy, forestry, medicine, etc.. This model is built as doubly stochastic process: firstly, cluster centers are randomly drawn under a given spatial point process; secondly, daughter points are randomly spread around cluster centers. Thus, the Neyman-Scott point process is specified by the distribution of cluster centers, the distribution of the number of points per cluster (thereafter called cluster size), and the distribution of daughter points around their cluster center (thereafter called cluster spread). This construction ensure that the user can access directly three quantities, i.e. κ - intensity of process of cluster centers, α - mean number of points in a cluster and ω - parameter determining the cluster spread.

Introducing inhomogeneity into the Neyman-Scott process is important due to the expressing the dependence of the point process on a set of covariates and checking its significance. But the inhomogeneity can be modelled in various ways. If the intensity of cluster centers κ is made to be inhomogeneous, then the clusters remains same but the number of clusters vary in the space. Such a model is called Neyman-Scott point process with inhomogeneous cluster centers and the inference for this process was investigated in [8]. This model is not second order inhomogeneity reweighting stationary (SOIRS) [2], but if instead the mean number of points in cluster α is made inhomogeneous, the resulting process is SOIRS. The inference for such process can be done by two step methods based on the contrast or composite likelihood in spatstat package [10]. Nevertheless the Bayesian MCMC inference method proposed here is more precise. If the inhomogeneity is introduced in ω the localy scaled Neyman-Scott point process is built [4]. It is also possible to introduce inhomogeneity simultaneously in α and ω, then we talk about Neyman-Scott point process with growing clusters [7]. The last three models keep the number of clusters same but vary the shape of the clusters, therefore we talk about cluster

inhomogeneity, while for the Neyman-Scott point process with inhomogeneous cluster centers we talk about inhomogeneity of centers.

Combining together cluster inhomogeneity and inhomogeneity of centers we talk about doubly inhomogeneous Neyman-Scott point process. The Bayesian MCMC inference for such a process was studied in [9]. Other kinds of inference were found to be useless for such complex models. Therefore, we build up an R package binspp which contains the Bayesian MCMC estimation procedure for all mentioned kinds of inhomogeneous Neyman-Scott point process but for the homogeneous one as well.

As it was mentioned above, the minimum contrast approach and maximum composite likelihood approach is available for SOIRS Neyman-Scott point process in R, it is also described in [8] for Neyman-Scott point process with inhomogeneous cluster centers. It is not available for other kinds of inhomogeneity up to our best knowledge.

The binspp package contains also Bayesian MCMC estimation procedure for homogeneous generalised Neyman-Scott process. Which uses generalised Poisson distribution (GPD) as a distribution for number of points in a cluster. The GPD allows for modelling of under- or over-dispersion of number of points in a cluster. The method were described in [1]. This process is useful for determining the under- or over-dispersion of number of points in a cluster or more precise modelling of the distribution of number of points in a cluster.

The R package binspp is avaylable at `https://github.com/tomasmrkvicka/binspp`.

The most important statistical problem is to address the dependence of the data on the given covariates. Our package handle the spatial covariates attached in any combination to the intensity of process of cluster centers κ, cluster size α and parameter determining the cluster spread ω. The Bayesian MCMC procedure is time consuming, but on the other hand the significance of all covariates is automatically provided from the posterior distribution. E.g. in case of ihomogeneity of centers and when a faster estimation method is used, it is necessary to perform parametrical bootstrap in order to obtain the significance of the covariates, which is as time consuming as Bayesian MCMC procedure [8]. Only in the case of inhogeneous cluster sizes, which produce the SOIRS process, it is possible to use fast estimation method and the significance of covariates can be obtained by asymptotic normality derived in [10].

Bayesian estimation for homogeneous Neyman-Scott point processes was carried out with a MCMC algorithm in [3, 6, 7, 5], for example. In this approach, the cluster centers and the model parameters are updated in each step of the MCMC algorithm. After reaching the equilibrium, posterior distributions of the parameters can be provided. The cluster centers are generally viewed as nuisance parameters when they do not correspond to an interpretable element of the phenomenon under study.

Considering the inhomogeneous clusters, the MCMC algorithm proceeds in the same way as in the homogeneous case, except that the likelihood is influenced by parameters connected with cluster inhomogeneity.

Considering the inhomogeneous centers, the proposed Bayesian MCMC procedure is performed in two steps, similarly like in [10]. First, the inhomogeneity of centers is estimated from the Poisson likelihood,

then in the second step the estimated inhomogeneity of centers is plug in to the Bayesian MCMC procedure which estimates the cluster inhomogeneity and the rest of the parameters. The inference about the covariates attached to the inhomogeneity of the centers is then performed from the posterior distribution of the cluster centers obtained in the second step. It is also possible to perform full Bayesian approach by estimating the inhomogeneity of centers in the Bayesian MCMC procedure, but this approach was found to be less precise than twe two step approach, probably by unidentifiability issues in the full likelihood.

Considering the generalised Neyman-Scott process, the MCMC algorithm consist of one more step in addition to the traditional Metropolis-Hastings update of parameters and Birdth-death-move update of cluster centers. It is the update of connection between points and cluster centers, since by using the non Poisson distribution the connection takes its part in the likelihood of the process.

Acknowledgments

The research has been supported financially by the Grant Agency of Czech Republic (Project No. 19-04412S).

References

[1] Andersson C., Mrkvička T. (2020). Inference for cluster point processes with over- or under-dispersed cluster sizes, *submitted to Statistics and Computing*.

[2] Baddeley, A.J., Møller, J., Waagepetersen, R. (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica* **54**, 329 - 350.

[3] Guttorp, P., Thorarinsdottir, T.L. (2012). Bayesian inference for non-Markovian point processes, in: Porcu, E., Montero, J.M., Schlather, M. (Eds.), *Advances and Challenges in Space-time Modelling of Natural Events.* Springer, pp. 79 - 102.

[4] Hahn, U., Jensen, E.B.V., Lieshout, M.C.V., Nielsen, L.S. (2003). Inhomogeneous spatial point processes by location-dependent scaling. *Advances in Applied Probability* **35**, 319 - 336.

[5] Kopecký, J., Mrkvička, T. (2016). On Bayesian estimation for Neyman-Scott point processes. *Applications of Mathematics* **61**, 503 - 514.

[6] Møller, J., Waagepetersen, R.P. (2007). Modern statistics for spatial point processes. *Scandinavian Journal of Statistics* **34**, 643 - 684.

[7] Mrkvička T. (2014). Distinguishing different types of inhomogeneity in Neyman-Scott point processes, *Methodology and Computing in Applied Probability* **16/2**, 385 - 395.

[8] Mrkvička T., Muška M., Kubečka J. (2014). Two step estimation for Neyman-Scott point process with inhomogeneous cluster centers, *Statistics and Computing* **24/1**, 91-100.

[9] Mrkvička T., Soubeyrand S. (2017). On Parameter Estimation for Doubly Inhomogeneous Cluster Point Processes, *Spatial statistics* **20**, 191-205.

[10] Waagepetersen, R., Guan, Y. (2009). Two-step estimation for inhomogeneous spatial point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**, 685 - 702.

Research needs in wildfire risk assessment spatiotemporal modelling

C. Vega-García[1,2,*] , M. Rodrigues[3] , F.J. Alcasena[1] , C. Comas[4]

[1]*Department of Agricultural and Forest Engineering, University of Lleida, Alcalde Rovira Roure 191, 25198 Lleida, Catalonia, Spain, cristina.vega@udl.cat, fermin.alcasena@udl.cat*
[2]*Joint Research Unit CTFC-Agrotecnio, Ctra. Sant Lloren Km.2 25280, Solsona, Spain*
[3]*Department of Geography, University of Zaragoza, Pedro Cerbuna 12, 50009 Zaragoza, Spain, rmarcos@unizar.es*
[4]*Department of Mathematics, University of Lleida, Spain; carles.comas@udl.cat*
[*]*Corresponding author*

Keywords. *Fire behaviour; Probability of occurrence, Risk assessment; Wildfire simulations*

Due to a total fire exclusion policy, and the accompanying development of fire management systems, wildfires have decreased in number and burned area in Southern European countries during the last decade. Extreme events, however, have escaped this trend. A few events are responsible every year for most of the burned area, damages and threats to lives and properties, usually under drought and heat wave conditions than place fires beyond suppression capacity, and under multiple-fire occurrences that overload firefighting resources.

The realization that total protection of valued resources and assets, properties and lives is virtually impossible under current conditions has come to be recognized by fire managers only recently, if at all, demanding that new policies and strategies are implemented in Mediterranean areas beyond the usual prevention and suppression solutions applied.

Planning for better civil protection, emergency and resources management though safe and efficient suppression, resilient landscapes and fire-adapted communities within a comprehensive strategy for Mediterranean landscapes [1] requires wildfire risk analysis. Risk is often portrayed as a combination of probability or likelihood of a fire start, exposure caused by propagation or fire intensity, and vulnerability of exposed assets [4]. While all these components of risk have been modelled before, and often separately, all of them admit improvements over current methodologies.

The analysis of fire occurrence by intensity level is often called exposure analysis [3]. Fire occurrence, the probability of a fire starting, or probability of ignition, has been modelled using different techniques for many spatial and temporal scales, ranging from high-resolution observations (small pixels) to forest districts, or provinces, and from daily models to models encompassing several years. A review of work done in this area can be accessed in [2]. Developments in the field of fire occurrence prediction, for instance, achieved success at dealing with the problem that fires are rare events, but most previous work has avoided the issue of how best

to stratify probability of ignition models in time and space; since probabilities of ignition cannot be considered stationary, we need to evaluate how different models should be applied in different areas or seasons, but this spatiotemporal issue has not been solved by research.

Fire propagation is an extremely complex process depending on fuel characteristics, topographic features, and weather (wind) parameters, all of them difficult to measure and model in a realistic manner. Simulations providing probabilistic outputs like conditional flame length or burn probability are used as the basis for fire management actions, but the optimization of the simulation areas and temporal windows remain also spatiotemporal issues that need further research.

References

[1] Alcasena F.J., Ager A.A., Bailey J.D., Pineda N. and Vega-Garca C. (2019) Towards a comprehensive wildfire management strategy for Mediterranean areas: Framework development and implementation in Catalonia, Spain. *Journal of Environmental Management* **231**: 303 320.

[2] Costafreda-Aumedes S., Comas C. and Vega-Garcia C. (2017) Human-caused fire occurrence modelling in perspective: A review. *International Journal of Wildland Fire* **26**(12):983-998.

[3] Fairbrother A. and Turnley J.G. (2005) Predicting risks of uncharacteristic wildfires: application of the risk assessment process. *Forest Ecology and Management* **211**:28-35.

[4] Finney M.A. (2005) The challenge of quantitative risk analysis for wildland fire. *Forest Ecology and Management* **211**:97-108.

Shadow Simulated Annealing a new algorithm for point processes parameter estimation

R.S. Stoica[1,*] , M. Deaconu[1] , A. Philippe[2] and L. Hurtado-Gil[3]

[1]*Université de Lorraine, CNRS, Inria, IECL, F-54000, Nancy, France ; radu-stefan.stoica@univ-lorraine.fr, madalina.deaconu@inria.fr*
[2]*Université de Nantes, Laboratoire de Mathḿatiques Jean Leray, Nantes, France ; anne.philippe@univ-nantes.fr*
[3]*eDreams ODIGEO, C/ Bailn 67-69, 08009, Barcelona ; lluis.hurtado@edreamsodigeo.com*
[*]*Corresponding author*

Abstract. *This talk introduces a global optimisation procedure based on the ABC Shadow simulation dynamics. First the general method is explained, and then results are presented. The method is general, in the sense that it applies for probability densities that are continuously differentiable with respect to their parameters.*

Keywords. *Approximate Bayesian computation, Computational methods in Markov chains, Maximum likelihood estimation, Point processes, Spatial pattern analysis.*

1. Set-up of the problem

Let us assume that an object pattern $\mathbf{y}$ is observed in a compact window $W \subset \mathbb{R}^d$. The observed pattern is supposed to be the realisation of a spatial process. Such a process is given by the probability density

$$p(\mathbf{y}|\theta) = \frac{f(\mathbf{x}|\theta)}{c(\theta)} = \frac{\exp[-U(\mathbf{y}|\theta)]}{c(\theta)} \tag{1}$$

with $f(\mathbf{x}|\theta)$ the unormalised probaility density, $U(\mathbf{y}|\theta)$ the energy function and $c(\theta)$ the normalising constant. The model given by (1) may be considered as a Gibbs process, and it may represent a Markov random field or a marked point process. Let $p(\theta|\mathbf{y})$ be the conditional distribution of the model parameters or the posterior law

$$p(\theta|\mathbf{y}) = \frac{\exp[-U(\mathbf{y}|\theta)]p(\theta)}{Z(\mathbf{y})c(\theta)}, \tag{2}$$

where $p(\theta)$ is the prior density for the model parameters and $Z(\mathbf{y})$ the normalising constant. The posterior law is defined on the parameter space Θ. For simplicity, the parameter space is considered to be a compact region in $\mathbb{R}^r$ with r the size of the parameter vector.

In the following, it is assumed that the probability density $p(\mathbf{y}|\theta)$ is strictly positive and continuous differen-

tiable with respect to θ. This hypothesis is strong but realistic, since it is often required by practical applications.

This paper proposes a global optimisation method based on the Shadow Simulated Annealing (SSA) process to compute :

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} p(\theta|\mathbf{y}). \tag{3}$$

.

2. SSA algorithm : general description

The SSA algorithm is a global optimisation method that uses the ABC Shadow dynamics [8] that can be used to maximise posterior densities.

Classical Simulated Annealing algorithms are based on the following principle. Assume that the probability density p is to be maximised. This is achieved by sampling $p^{1/T}$ while $T \to 0$. If the temperature parameter T goes to 0 in an appropriate way, then the SA algorithm converges asymptotically towards the global optimum. This method is rather general. Under smooth assumptions, the algorithm can be generalised to minimise any criteria U that can be written as $p \propto \exp(-U)$.

SA algorithms for maximising probability densities for random fields and marked point process such as (1) are presented in [1, 4, 7]. The obtained cooling schedules for the temperature parameter are of the form

$$T = \frac{T_0}{1+\log n}$$

with $n > 0$. The solution guaranteed by the method converges towards the uniform distribution over the subspace of configurations that maximises (1).

The difficulty of solving (3) is due to the fact that the normalising constant $c(\theta)$ is not available in analytic closed form. Hence, special strategies are required to sample from the posterior distribution (2). The present paper use for this purpose, the ABC Shadow simulation dynamics [8].

The main steps of the SSA algorithm are presented below :

Algorithm 1 Shadow Simulated Annealing (SSA) algorithm : *fix* $\delta = \delta_0$, $T = T_0$, *n and* $k_\delta, k_T : \mathbb{R}^+ \to \mathbb{R}^+$ *two positive functions. Assume the observed pattern is* $\mathbf{y}$ *and the current state is* θ_0.

1. *Generate* $\mathbf{x}$ *according to* $p(\mathbf{x}|\theta_0)$.
2. *For* $k = 1$ *to* m *do*
   - *Generate a new candidate* ψ *following the density* $U_\delta(\theta_{k-1} \to \psi)$ *defined by*
   $$U_\delta(\theta \to \psi) = \frac{1}{V_\delta}\mathbf{1}_{b(\theta,\delta/2)}\{\psi\},$$
   with V_δ *the volume of the ball* $b(\theta, \delta/2)$.
   - *The new state* $\theta_k = \psi$ *is accepted with probability* $\alpha_s(\theta_{k-1} \to \psi)$ *given by*
   $$\begin{aligned}\alpha_s(\theta_{k-1} \to \theta_k) &= \\ &= \min\left\{1, \left[\frac{p(\theta_k|\mathbf{y})}{p(\theta_{k-1}|\mathbf{y})} \times \frac{f(\mathbf{x}|\theta_{k-1})}{f(\mathbf{x}|\theta_k}\right]^{1/T} \times \frac{\mathbf{1}_{b(\theta_k,\delta/2)}\{\theta_{k-1}\}}{\mathbf{1}_{b(\theta_{k-1},\delta/2)}\{\theta_k\}}\right\} \\ &= \min\left\{1, \left[\frac{f(\mathbf{y}|\theta_k)p(\theta_k)}{f(\mathbf{y}|\theta_{k-1})p(\theta_{k-1})} \times \frac{f(\mathbf{x}|\theta_{k-1})}{f(\mathbf{x}|\theta_k)}\right]^{1/T}\right\} \qquad (4)\end{aligned}$$
   otherwise $\theta_k = \theta_{k-1}$.
3. *Return* θ_m.
4. *Stop the algorithm or go to step* 1 *with* $\theta_0 = \theta_n$, $\delta_0 = k_\delta(\delta)$ *and* $T_0 = k_T(T)$.

It is easy to see that the SSA algorithm is identical to the ABC Shadow dynamics whenever δ and T remain unchanged [8].

3. Results

The SSA lgorithm is applied here to the statistical analysis of patterns which are simulated from a Strauss model [3, 10]. This model describes random patterns made of points exhibiting repulsion. Its probability density is

$$\begin{aligned}p(\mathbf{y}|\theta) \propto \beta^{n(\mathbf{y})}\gamma^{s_r(\mathbf{y})} &= \\ &= \exp\left[n(\mathbf{y})\log\beta + s_r(\mathbf{y})\log\gamma\right]. \qquad (5)\end{aligned}$$

Here $\mathbf{y}$ is a point pattern in the window W, while $t(\mathbf{y}) = (n(\mathbf{y}), s_r(\mathbf{y}))$ and $\theta = (\log\beta, \log\gamma)$ are the sufficient statistic and the model parameter vectors, respectively. The sufficient statistics components $n(\mathbf{y})$ and $s_r(\mathbf{y})$ represent respectively, the number of points in W and the number of pairs of points at a distance closer than r.

The Strauss model on the unit square $W = [0,1]^2$ and with density parameters $\beta = 100$, $\gamma = 0.8$ and $r = 0.1$, was considered. This gives for the parameter vector of the exponential model $\theta = (4.60, -0.22)$. The CFTP algorithm (see Chapter 11 in [6]) was used to get 1000 samples from the model and to compute the empirical means of the sufficient statistics $\bar{t}(\mathbf{y}) = (\bar{n}(\mathbf{y}), \bar{s_r}(\mathbf{y})) = (65.23, 51.51)$. The SSA algorithm was run using $\bar{t}(\mathbf{y})$ as observed data.

The prior density $p(\theta)$ was the uniform distribution on the interval $[3, 5.5] \times [-5, 0]$. Each time, the auxiliary variable was sampled using 100 steps of a MH dynamics [5, 6]. The Δ and m parameters were set to $(0.01, 0.01)$ and 100, respectively. The algorithm was run for 10^6 iterations. The intial temperature was set to $T_0 = 10^4$. For the cooling schedule a slow polynomial scheme was chosen

$$T_n = k_T \cdot T_{n-1}$$

with $k_T = 0.9999$. A similar scheme was chosen for the Δ parameters, with $k_\Delta = 0.99999$. Samples were kept every 10^3 steps. This gave a total of 1000 samples.

Figure 1 shows the results obtained after running the SSA algorithm. The final values for $\log\beta$ and $\log\gamma$ were 4.60 and -0.22, respectively. These values are almost identical to the true model parameters.

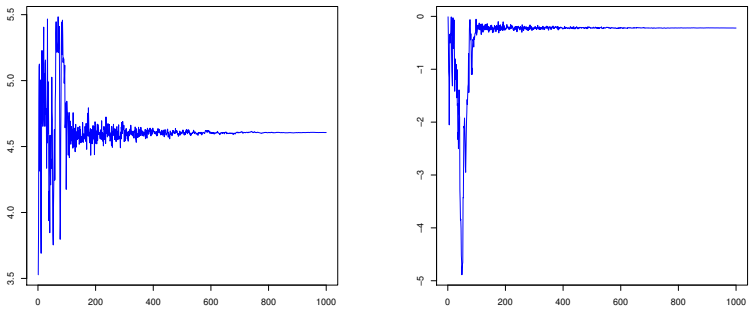


Figure 1: SSA results for computing the MAP estimates for the Strauss model parameters.

4. Conclusions and perspectives

The algorithm was also applied on real astronomical data, and the obtained models were tested and validated. Since the ABC Shadow is an approximate algorithm, the theoretical convergence of the SA procedure based on it was also established [9].

References

[1] Geman, S. and Geman, S (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

[2] Haario, H. and Saksman, E. (1991). Simulated annealing process in general state space. *Advances in Applied Probability* **23**, 866–893.

[3] Kelly, F. P. and Ripley, B. D. (1976). A note on Strauss's model for clustering. *Biometrika* **63**, 357–360.

[4] van Lieshout, M. N. M. (1994). Stochastic annealing for nearest neighbour point processes with application to object recognition. *Advances in Applied Probability* **26**, 281–300.

[5] van Lieshout, M. N. M. (2000). *Markov point processes and their Applications*. Imperial College Press. London.

[6] Møller, J. and Waagepetersen R. P. (2004). *Statistical inference and simulation for spatial point processes*. Chapman and Hall/CRC. Boca Raton.

[7] Stoica, R. S., Gregori, P. and Mateu, J. (2005). Simulated annealing and object point processes : tools for analysis of spatial patterns. *Stochastic Processes and their Applications* **115**, 1860–1882.

[8] Stoica, R. S., Philippe, A., Gregori, P. and Mateu, J. (2017). ABC Shadow algorithm: a tool for statistical analysis of spatial patterns. *Statistics and Computing* **27**, 1225–1238.

[9] Stoica, R. S., Deaconu, M., Philippe, A., and Hurtado-Gil, L. (2021). Shadow Simulated Annealing: a new algorithm for approximate Bayesian inference of Gibbs point processes. *Spatial Statisitcs*, **23**, 2021.

[10] Strauss, D. J. (1975). A model for clustering. *Biometrika* **62**, 467–475.

Extending planar point with scalar marks to more complex mark scenarios

M. Eckardt

Humboldt-Universität zu Berlin, Chair of Statistics, Unter den Linden 6 (UL6), D-10099 Berlin, Germany; m.eckardt@hu-berlin.de

Abstract. *The analysis of (marked) spatial point processes remains to be the subject of interest in various fields. While the points have been extended to more complex domains, i.e. the network space or the sphere, in recent years, the marks most commonly remain scalar-valued quantities. Leaving simple mark scenarios behind, this talk focusses on extensions to the case where the attributes themselves are objects rather than scalar-valued quantities.*

Keywords. *Complicated mark structures; Mark Characteristics; Non-Euclidean; Non-standard mark spaces*

The analysis of random point configurations $\{\mathbf{s}_i, m(\mathbf{s}_i)\}$ with points in $\mathbf{S} \subset \mathbb{R}^2$ and marks in $\mathbb{M}$ has become a vivid field of research. Apart from different summary characteristics which help to decide on the structural properties of the points, i.e. the ground process, and deviations from complete spatial randomness, various tools for integer-valued (multitype) and real-valued point attributes or combinations of integer- and real-valued marks have been derived. While recent years have witnessed extensions for marked patterns with points on e.g. network structures or the sphere, the marks themselves remain most commonly scalar-valued quantities. Addressing this gap, we consider the analysis of points with object-valued point attributes $\{\mathbf{s}_i, o(\mathbf{s}_i)\}$ and introduce extensions of classical mark characteristics to the present context. As special cases, non-simple (multitype) point patterns with multiple coincident points and points with vector-valued marks are included in the class of object-valued marks.

Contributed

Nonparametric tests of dependence between a spatial point process and a covariate

J. Dvořák[1,*], T. Mrkvička[2] , J. Mateu[3] and J.A. González[3]

[1]*Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 186 75 Prague, Czech Republic; dvorak@karlin.mff.cuni.cz*
[2]*Department of Applied Mathematics and Informatics, Faculty of Economics, University of South Bohemia, Studentská 13, 370 05 České Budějovice, Czech Republic; mrkvicka.toma@gmail.com*
[3]*Department of Mathematics, University Jaume I, Campus Riu Sec, 12071, Castellón de la Plana, Castellón, Spain; mateu@mat.uji.es, jmonsalv@mat.uji.es*
[*]*Corresponding author*

Abstract. *In this contribution we consider the problem of testing possible dependence between a point process and a spatial covariate. We recall the available methods which are based on the assumption of a Poisson process. Taking advantage of the recent development of the random shift methods we study the nonparametric tests using random shifts which are suitable also for non-Poisson processes. We study the performance of the different tests in a simulation study and conclude that the random shift test, either with the toroidal correction or the variance correction, depending on the strength of interactions in the point process, performs well for non-Poisson processes.*

Keywords. *Spatial point process; Covariates; Independence; Random shifts; Preferential sampling.*

1. Introduction and background

Let X be a spatial point process observed in a compact observation window $W \subset \mathbb{R}^2$, and let $Y(u), u \in W$, be a spatial random field (covariate). In the classical geostatistical setting the random field values are observed only at the points of X. In this case the independence between the random field Y and the process of sampling locations X is referred to as non-preferential sampling, as opposed to the preferential sampling which refers to situations where Y and X are stochastically dependent [3]. For testing the null hypothesis of non-preferential sampling the paper [3] suggests to use the test by [6], based on fitting a Gaussian random field model to the observed data and comparing the data to the simulations from the fitted model in the Monte Carlo fashion [6].

In the following we step outside the geostatistical context and consider the random field Y to be observed at every location $u \in W$, at least on a fine pixel grid. This is usual if the values of Y are obtained by remote sensing (e.g. altitude and slope of a terrain), interpolated from a given set of measurements by kriging or other methods (mineral content in soil) or constructed analytically from a given set of observations (distance from a geological fault). In such a case Y corresponds to a spatial covariate that may influence the distribution of X.

The problem of testing whether Y affects the distribution of X has been tackled under the assumption of Poisson process e.g. in [2]. For non-Poisson processes parametric may be employed as e.g. in [7]. We focus here on simple methods that do not require fitting a complex model to the data.

As a natural starting point it is possible to compare the empirical cumulative distribution function $\widehat{F}(y)$ of the sampled values $Y(x_i)$ with the cumulative distribution function $F_0(y) = \frac{1}{|W|}\int_W \mathbf{1}\{Y(u) \leq y\}\,\mathrm{d}y$ of the covariate values across whole W. This can be done using the Kolmogorov-Smirnov statistic, Cramér-von Mises statistic or Anderson-Darling statistic [1, Sec. 10.5.2]. Under Poisson assumption these are in fact tests of the constant intensity hypothesis (Complete Spatial Randomness, CSR).

Another approach is suggested by [2]. Let $\{x_1,\ldots,x_n\}$ denote the observed points of X in W. The Berman's Z_1 and Z_2 tests assume X is a Poisson process with intensity function $\lambda(u;\theta) = b(u)\exp\{\theta_0 + \theta_1 Y(u)\}$ where $\theta = (\theta_0,\theta_1)$ are parameters and $b(u)$ is a known baseline function [1, Sec. 10.3.5]. The Z_1 test is in fact the score test of the null hypothesis that $\theta_1 = 0$, based on the sum of observed values $\sum_i Y(x_i)$. The Z_2 test is based on the transformed values $u_i = F_0(Y(x_i))$ where $F_0(y)$ was defined in the previous paragraph.

In order to relax the Poisson assumption [2] suggests, for rectangular windows, to perform instead a Monte Carlo test, with a suitable test statistic, based on the random shifts with torus correction (see below for details). The author comments that "hopefully, the edge effects introduced by the wrapping procedure will have a minimal effect on the statistic" [2, p.60]. However, experience shows that this may not be the case, depending on the choice of test statistic and the range of interactions in the point process, resulting in liberality of the test. For this reason we consider here also the random shift test with variance correction proposed in [5] which can be used also for irregular observation windows.

2. Random shift tests

In the classical Monte Carlo test one computes a test statistic value T_0 from the observed data, obtains in a certain way M replications of the data under the null hypothesis and computes the values of the test statistic $T_1,\ldots,T_M$ from the replications. The p-value of the test is then determined by assessing how typical or extreme the value T_0 is with respect to the population of $(T_1,\ldots,T_M)$. The random shift tests are based on a specific strategy for producing the Monte Carlo replications [4]. In order to break possible dependence structure between a pair of spatial processes (such as the point process X and the random field Y in this paper), one of the processes is kept fixed while the other one is shifted by a random vector. Note that the random shift approach in general requires at least one of the spatial processes to be stationary, but not necessarily both. Different versions of the random shift test are available, using different ways to deal with the part of data that is shifted outside the observation window W. Here we focus on two of them, the well-established torus correction and the variance correction from [5].

The torus correction approach [4, 2] makes the shifts respecting the toroidal geometry induced by identifying the opposite edges of the observation window. Wrapping the data onto the torus introduces cracks in the correlation structure of the shifted data which in turn introduces liberality of the test [5]. To compensate for the liberality [5] proposed a variance correction strategy which is based on dropping out the part of the data shifted outside W. In this way no cracks in the correlation structure are created. On the other hand the amount of

data used for computing the test statistic values $T_1, \ldots, T_M$ decreases and their means and variances are possibly different, resulting in the need for standardization to get closer to exchangeability. Details are given in [5, Sec. 2.1.3].

In this contribution we investigate the performance of the random shift tests using the sample mean as the test statistic, i.e. $T = \frac{1}{n}\sum_{i=1}^{n} Y(x_i)$, and compare it with the tests discussed in Section 1. For the variance correction we use the correction factor $1/\sqrt{n}$, motivated by the fact that for the sample mean computed from n i.i.d. observations the order of variance is $1/n$. It can be shown that the same correction factor is appropriate also in the current context where the observed values $Y(x_i)$ are not independent.

3. Simulation study

In the following we consider the random shift tests with the torus correction and the variance correction (RS_T, RS_V, based on 999 random shifts), the Schlather et al. [6] simulation-based test (*SIM*, based on 99 simulations, the test statistic $E(h)$ and l_2-norm with constant weights), the tests based on the cumulative distribution function and Kolmogorov-Smirnov, Cramér-von Mises and Anderson-Darling statistic (F_{KS}, F_{CvM} and F_{AD}, respectively) and the two Berman's tests (Z_1, Z_2). The nominal significance level of the tests is chosen to be 0.05. The observation window is the unit square $W = [0,1]^2$ and the random shift vectors are generated uniformly on a disc with radius 1/2. We perform the tests for 1000 independent realizations from each model and report the rejection rates.

Let Y_1, Y_2 be independent zero-mean unit-variance isotropic Gaussian random fields with correlation function $r(s) = \exp\{-5s\}, s \geq 0$. In both experiments considered here the point process X is a stationary log-Gaussian Cox process with the driving intensity function $\Lambda(u) = \exp\{5 + \alpha Y_1(u) - \alpha^2/2\}, u \in W$, where $\alpha \in \mathbb{R}$ is a parameter governing the strength of interactions in X. Intensity of X is approx. 148.

In experiment 1 the covariate Y is taken to be the random field affecting the distribution of X, i.e. $Y = Y_1$. This experiment hence corresponds to X being a Poisson process with intensity function depending on the covariate. Therefore the use of the tests based on cumulative distribution function and the Berman's tests is justified. Setting $\alpha = 0$ implies independence between X and Y and corresponds to the null hypothesis. Setting $\alpha \neq 0$ allows assessing the power of the tests against selected alternatives. In experiment 2 the covariate Y is taken to be the random field independent of X, i.e. $Y = Y_2$. In this case X is a clustered point process independent of the covariate Y (null hypothesis holds) and we can investigate robustness of the tests to departures from the Poisson assumption. Note that the random shift tests and the Schlather *et al.* [6] test do not rely on this assumption and hence are expected to perform well.

The top part of Table 1 gives the rejection rates in experiment 1. In the case with $\alpha = 0$ we observe that some of the tests match the nominal significance level rather well ($RS_T, F_{KS}, F_{CvM}, Z_2$) while others are slightly conservative (RS_V, Z_1) or slightly liberal (*SIM*). The F_{AD} test is extremely liberal and will not be discussed further. With increasing absolute value of α the dependence of X on Y gets stronger and the rejection rates indicate increasing power of the tests, RS_T, F_{CvM}, Z_2 showing the highest power. The bottom part of Table 1 gives rejection rates in experiment 2. All choices of α now correspond to the null hypothesis so the rejection rates should be close to 0.05. For $\alpha = 0$ the point process X follows CSR and all the test are theoretically

α	RS_T	RS_V	*SIM*	F_{KS}	F_{CvM}	F_{AD}	Z_1	Z_2
0.6	0.999	0.961	0.828	0.996	0.999	0.999	0.994	1.000
0.4	0.975	0.669	0.593	0.945	0.964	0.976	0.963	0.977
0.2	0.561	0.175	0.214	0.463	0.548	0.654	0.533	0.573
0.0	0.043	0.026	0.082	0.046	0.051	0.211	0.030	0.046
-0.2	0.542	0.179	0.216	0.456	0.528	0.625	0.516	0.556
-0.4	0.968	0.648	0.572	0.939	0.964	0.977	0.959	0.972
-0.6	0.998	0.961	0.838	1.000	1.000	1.000	1.000	1.000
α	RS_T	RS_V	*SIM*	F_{KS}	F_{CvM}	F_{AD}	Z_1	Z_2
0.6	0.078	0.034	0.069	0.214	0.236	0.376	0.184	0.220
0.4	0.057	0.035	0.058	0.100	0.118	0.275	0.093	0.124
0.2	0.046	0.026	0.061	0.055	0.070	0.249	0.050	0.072
0.0	0.055	0.018	0.060	0.036	0.048	0.196	0.041	0.050
-0.2	0.051	0.023	0.055	0.066	0.065	0.252	0.053	0.062
-0.4	0.065	0.028	0.076	0.118	0.133	0.294	0.108	0.133
-0.6	0.066	0.025	0.070	0.193	0.216	0.363	0.180	0.210

Table 1: Fractions of rejections of the null hypothesis by the given tests. Top part: experiment 1; bottom part: experiment 2. Nominal significance level 0.05.

justified, only the *SIM* test showing slight liberality. With increasing absolute value of α the point process X has more and more prominent interactions (clustering) and the tests based on the Poisson assumption break down, showing high degree of liberality, as noted in [1, Sec. 10.5.4]. Also the RS_T and *SIM* tests show increasing liberality. The RS_V test is able to preserve the interaction structure of X, its rejection rates not exceeding the nominal significance level.

Note that the observations made for the random shift methods (increasing liberality of RS_T with increasing strength of interactions in the process, conservativeness of RS_V) are in line with the findings reported in [5]. To conclude, the random shift tests provide a simple, fully nonparametric way of testing the given hypothesis. For rectangular observation windows and point processes with not very strong interactions the torus correction is appropriate. If the interactions are strong and the user is not willing to accept the degree of liberality of the torus correction indicated by simulation studies, the variance correction should be used instead.

Acknowledgments

This work was supported by Grant Agency of Czech Republic (Project 19-04412S).

References

[1] Baddeley, A.J., Rubak, E. and Turner, R. (2016). *Spatial Point Patterns*. Chapman and Hall/CRC. New York.

[2] Berman, M. (1986). Testing for spatial association between a point process and another stochastic process. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **35** (1), 54–62.

[3] Diggle, P.J., Menezes, R. and Su, T. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **59** (2), 191–232.

[4] Lotwick, H.W. and Silverman, B.W. (1982). Methods for analysing spatial processes of several types of points. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **44** (3), 406–413.

[5] Mrkvička, T., Dvořák, J., Mateu, J. and González, J.A. (2020). Revisiting the random shift approach for testing in spatial statistics. To appear in *Spatial Statistics*. https://doi.org/10.1016/j.spasta.2020.100430

[6] Schlather, M., Ribeiro, P.J. and Diggle, P.J. (2004). Detecting dependence between marks and locations of marked point processes. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **66** (1), 79–93.

[7] Waagepetersen, R.P. and Guan, Y. (2009). Two-step estimation for inhomogeneous spatial point processes. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **71** (3), 685–702.

Presence-only for Marked Point Process under Preferential Sampling

G.A. Moreira[1,*] and R. Menezes[2]

[1]*Centro de Biologia Molecular e Ambiental, Universidade do Minho, Campus de Gualtar, 4710-057 Braga - Portugal; d12582@bio.uminho.pt*
[2]*Centro de Matemática, Universidade do Minho, Campus de Gualtar, 4710-057 Braga - Portugal; rmenezes@math.uminho.pt*
[*]*Corresponding author*

Abstract. *Preferential Sampling models have received much attention in the last few years. Although its original model is applied to geostatistics, it can be recognized in other types of data, such as point processes, in the form of presence-only data. This has been already identified in the Statistics literature. It is valuable to draw advantages from both presence-only and preferential sampling specific literatures. In particular, we propose a way to deal with biased sampling of a continuous variable collected by opportunistic sampling. For our particular case, we employ the idea on sardine biomass collected during professional fishing expeditions. The data, although intuitively understood, presents complications such as two types of preferential sampling. One is about the fish presence locations, and the preferentiability happens due to the travel pattern of fishing boats not being representative of the region. The other happens with respect to the biomass itself, as the fishermen prefer visiting regions with larger biomass. These and other theoretical and practical aspects of the problem are discussed. A probabilistically well defined approach is discussed. Its results may be an incentive to apply data collection in fishing expeditions commercial fisheries as a means for decision making aimed at benefiting both ecological and economical aspects.*

Keywords. *Inhomogeneous Poisson Process; Bayesian Analysis; Preferential sampling; Data augmentation; Spatial Statistics*

1. Introduction

A major challenge of quantitative ecology is a class of problems known as Species Distribution Models. It consists of methodologies whose aim is twofold. Firstly it explains the occurrence of species in relation to geological, ecological and climactic drivers. Secondly, it proposes to predict the occurrence over a specified region. Its applications range from conservation and reserve planning, evolution, epidemiology, invasive-species management and other fields [11].

Scientific collection of ecological data is often expensive. It requires planning, careful consideration of the study objectives and usage of specially suited equipment and personnel. Consequently, other sources of information are considered, particularly data that has not been randomly or systematically collected. These cases are often called opportunistic sampling and can contain biased information. However, if the model accounts for the bias, then it can adequately estimate scientifically relevant quantities.

Presence-only data is the result of opportunistic sampling when the collected data is the observed locations of the object of study. In this case a point process [1] is adequate and the bias can result in a model predicting the studied object less often where observers don't tend to go. For example, a group of biologists may record the locations of a certain species they are studying, but the data only exists close to locations easily accessible to people. This work benefits from [10] who dealt with this problem using exact inference on an Inhomogenous Poisson Process (IPP), dealing with identifiability issues otherwise mentioned in [5] and [4].

The problem becomes more intricate when there is a measured variable in the observed locations. For example in fishing data, the fish biomass may be recorded in addition to the location where they were found. In this case, the point process extends to the marked point process case. Additionally, the fishing expeditions will likely favor locations with known higher fish biomass, causing a biased sample. In this case, the preferential sample description of [3] is adequate, albeit still being in the presence-only field. This work joins these ideas to deal with the opportunistically sampled marked point process. In order to achieve scalability, the theory of Nearest Neighbor Gaussian Process is employed to sample the latent processes, using similar ideas as [12].

There has been a case which deals with presence-only data in the context of preferential sampling. Namely [7] use the latter approach to model the sampling bias of presence-only data. The authors have not been able to find methods that consider the case where preferential sampling happens in addition to presence-only sampling. This can happen with fishery data collected during fishing expeditions.

It is common that fishermen prefer going to locations with the most fish biomass. It is also assumed that fish biomass is a variable that has spatial smoothness, which is modeled using an approximated Gaussian Process. Therefore it is reasonable that the this process can also be used to measure the sampling bias due to biomass.

Another aspect of current problems is the inclusion of information from multiple data sources, some of which may have no sampling bias. For this reason some discussion is proposed about good practices of modeling species distributions when sampling bias may be present.

The mixing of a Gaussian Process in the intensity function extends the presence-only model of [10]. The idea is based on the doubly stochastic process of [9]. To make the procedure more computationally efficient, the recent nearest neighbor approximation of [2] has been discussed in the point process context in [12].

2. Proposal

The proposed model uses a data augmentation scheme to achieve exact computation of the Poisson Process likelihood.

2.1 Motivating data

The data which motivates this development comes from fishing expeditions, which constitutes presence-only data. The biomass of sardines is recorded as well. Since it is recorded from the fishing expeditions which favor higher biomass, it constitutes preferential sampling as well.

2.2 Model and notation

The available data is composed of an unordered set of paired variables $(X,Z) = \{(x_1,z_1),\ldots,(x_{n_x},z_{n_x})\}$ observed in a closed region $\mathcal{D}$. The component x_i represents the i-th location of sardines detection and z_i represents its recorded biomass.

The data are modeled as a marked point process based on the Inhomogeneous Point Process (*IPP*). In addition, a data augmentation scheme is used to avoid performing approximations of the likelihood function.

$$\begin{aligned} X &\sim IPP(q(\cdot)p(\cdot)\lambda^*) \\ X' &\sim IPP(q(\cdot)(1-p(\cdot))\lambda^*) \\ U &\sim IPP((1-q(\cdot))\lambda^*) \\ Z(s) \mid s \in x \cup x' &\sim Gamma(a, a/\eta(s)) \\ log\ \eta(s) &= W_z(s)\beta_z + S(s),\ s \in \mathcal{D} \\ logit\ q(s) &= W_{int}(s)\beta_{int},\ s \in \mathcal{D} \\ logit\ p(s) &= W_{obs}(s)\beta_{obs} + \gamma S(s),\ s \in \mathcal{D} \\ S(\cdot) &\sim NNGP\left(0, \sigma^2\rho(\cdot)\right)), \end{aligned} \tag{1}$$

where $W_z(\cdot), W_{int}(\cdot)$ and $W_{obs}(\cdot)$ are sets of covariates.

Parameter γ measures the preferentiability of the sampling procedure. The complete infinite-dimensional vector of unknown quantities is $\Theta = (\beta_z, \beta_{int}, \beta_{obs}, \lambda^*, X', U, a, \gamma, S(\cdot), \theta)$.

2.3 Inference

The inference is done under the Bayesian paradigm on the posterior distribution $\pi(\Theta \mid x,z) \propto L_x(\Theta)\pi(\Theta)$ ([8]) where $L_x(\Theta)$ is the likelihood function and $\pi(\Theta)$ is the prior distribution. The posterior is not known in closed form. An MCMC sampling scheme allows inference to be made. A Metropolis-within-Gibbs ([6]) sampling procedure using concepts from [10] and [12] is employed to achieve the sampling.

References

[1] Cressie, N. A. C. (1993). *Spatial Point Patterns*. John Wiley and Sons, Inc.

[2] Datta A., Banerjee S., Finley A. O. Gelfand A. E. (2016). Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets, *Journal of the American Statistical Association*, **111:514**, 800–812.

[3] Diggle, P. J., Menezes, R. and Su, T.- L . (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **59** 191–232.

[4] Dorazio, R. M. (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography* **23** 1472–1484.

[5] Fithian, W. and Hastie, T. (2013). Finite-sample equivalence in statistical models for presence-only data. *Ann. Appl. Stat.* **7** 1917–1939.

[6] Gamerman, D. and Lopes, H. (2006). *Markov Chain Monte Carlo - Stochastic Simulation for Bayesian Inference*, 2nd ed. CRC Press.

[7] Gelfand, A. E. and Schliep, E. M. (2018). Bayesian Inference and Computing for Spatial Point Patterns. *NSF-CBMS Regional Conference Series in Probability and Statistics* **10** i–125.

[8] Gelman, A. and Carlin, J.B. and Stern, H.S. and Dunson, D.B. and Vehtari, A. and Rubin, D.B. (2013). *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science.

[9] Gonalves, F. B. and Gamerman, D. (2018). Exact Bayesian inference in spatiotemporal Cox processes driven by multivariate Gaussian processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80** 157–175.

[10] Moreira, G. A., Gamerman, D. (2022). Analysis of presence-only data via exact Bayes, with model and effects identification. *Ann. Appl. Stat.*, To appear.

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

[12] Shirota, S., Gelfand, A. and Banerjee, S. (2019). Spatial joint species distribution modeling using Dirichlet processes. *Statistica Sinica* **29** 1127–154.

Disease mapping method comparing the spatial distribution of a disease with a control disease

O. Petrof[1,*], T. Neyens[1,2] , M. Vranckx[1] , V. Nuyts[3] , K. Nackaerts[4] , B. Nemery[3] and C. Faes[1]

[1]*Hasselt University, Data Science Institute (DSI), The Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Hasselt, Belgium; oana.petrof@uhasselt.be, thomas.neyens@uhasselt.be, maren.vranckx@uhasselt.be, christel.faes@uhasselt.be*

[2]*KU Leuven, Leuven Biostatistics and Statistical Bioinformatics Centre (L-BioStat), Department of Public Health and Primary Care, Leuven, Belgium; thomas.neyens@uhasselt.be*

[3]*KU Leuven, Centre for Environment and Health, Department of Public Health and Primary Care, Leuven, Belgium; valerie.nuyts@kuleuven.be, ben.nemery@kuleuven.be*

[4] *KU Leuven, Department of Pneumology, University Hospital Leuven, Leuven, Belgium; kristiaan.nackaerts@uzleuven.be*

[*]*Corresponding author*

Abstract. *Traditional disease mapping models are based on relating the observed number of disease cases per spatially discrete area to an expected number of cases for that area. Expected numbers are often calculated by internal standardisation, which requires both accurate population numbers and disease rates per age group. However, confidentiality issues or the absence of high-quality information about the characteristics of a population-at-risk can hamper those calculations. Based on methods in point process analysis for situations without accurate population data, we propose the use of a case-control approach in the context of lattice data, in which an unrelated spatially unstructured disease is used as a control disease. We correct for the uncertainty in the estimation of the expected values, which arises by using the control disease's observed number of cases as a representation of a fraction of the total population. We apply our methods to a Belgian study of mesothelioma risk, where pancreatic cancer serves as the control disease. The analysis results are in close agreement with those coming from traditional disease mapping models based on internally standardised expected counts. We show that the proposed method can adequately address the problem of inaccurate population data in disease mapping analysis.*

Keywords. *BYM model; Case-control study; Disease mapping; Mesothelioma; Standardization.*

---

1. Introduction

The classical hierarchical models for disease mapping make use of data including the population at risk or a local number of cases "expected" under some null model of disease transmission. Due to medical confidentiality, it is often difficult to obtain accurate and detailed population data [2]. Census data can be used to reflect the population data of a specific region. However, countries' census areas can be large or population data are not available for some countries. Census data are collected for a single snapshot in time, every decade, meaning that any changes in populations between census counts will add to the uncertainty of these data [2], while no data are available for the intercensal years. The objective of this paper is to propose a disease mapping method,

where a control disease is used as a proxy for the population at risk, extending the case-control methods for point-pattern data towards lattice data. In this study, interest is in mesothelioma cancer, while pancreatic cancer is used as control disease.

2. Methodology

2.1 Classical Disease Mapping Method

The response $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ represents the observed number of disease cases per areal unit throughout the study period. A Poisson model is commonly assumed to estimate the disease risk per area:

$$Y_i \sim Poisson(E_i\theta_i), \quad i = 1, ..., n, \tag{1}$$

where E_i represents the expected number of disease cases in area i and θ_i expresses the disease risk for the i^{th} area. The expected number of cases is defined as [6]:

$$E_i^I = \sum_g \frac{Y_g}{N_g} N_{i,g} = \sum_g r_g N_{i,g} \tag{2}$$

where r_g is the age-specific incidence rate in the standard population calculated as the observed number of cases in age group g and the age-specific population number. This ratio is multiplied by $N_{i,g}$ representing the population size of municipality i in age group g.

2.2 Disease Mapping with Control Disease

An approach commonly used in the context of point-pattern data is to compare the location of disease cases with that of a set of carefully selected controls for the population at risk [5]. In the context of disease mapping, it is assumed that only the aggregated number of cases for the disease of interest ($\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$) and the number of cases for the control disease ($\mathbf{Z} = (Z_1, Z_2, ..., Z_n)$) are available. The expected number of cases (under the assumption of no excess risk) for the disease of interest can be represented by

$$E_i^C = \frac{Z_i}{\sum_{j=1}^N Z_j} \left(\sum_{j=1}^N Y_j \right) = r_i^Z Y_\cdot \tag{3}$$

where r_i^Z is the rate of the control disease in area i and $Y_\cdot$ is the total number of cases of the disease of interest.

Any disease utilized as a control disease will introduce uncertainty in the model, as it represents a sample from the population data. The calculated expected values will have a lot of uncertainty if only a small number (or no) cases of the control disease are present. To account for the uncertainty in the estimation of the expected number E_i^C, we will assume that this is not a fixed known quantity. Since the expected number is based on the proportion of the total number of controls in the areas r_i^Z, we assume that the control disease follows a

multinomial distribution

$$(Z_1,...,Z_N) \sim Multinomial(Z_\cdot,(r_1^Z,...,r_N^Z)),$$

where $Z_\cdot$ represents the total number of controls. Given that the Poisson counts can be considered jointly multinomially distributed, we make use of the multinomial-Poisson transformation developed by [1]:

$$\begin{aligned} Z_i &\sim Poisson(\lambda_i) \\ \lambda_i &\sim Gamma(0.5,0.05), \\ r_i^Z &= \frac{\lambda_i}{\sum_{j=1}^N \lambda_j}, \end{aligned} \tag{4}$$

where the number of control cases Z_i in municipality i follows a Poisson distribution with mean λ_i. The resulting expected value is denoted as E_i^{C2}.

A conditional autoregressive convolution model [3] was used to analyse and compare the three methods presented above.

3. Data analysis

3.1 Data description

Residential information about all mesothelioma and pancreatic cancer patients diagnosed between 2004 and 2015 is available (Belgian Cancer Registry) as well as information about the population distribution in all areas during the period 2009-2015.

3.2 Results

Figure 1 presents the results of the classical method using indirect standardized number (upper panel), our proposed control-disease's standardized number (middle panel) and control-disease's standardized number accounting for uncertainty (lower panel). All methods show a cluster of municipalities in the Central Northern part of Flanders, and in the Central Eastern part of the country. However, on the middle panel areas with increased risk are more dispersed over the country, as compared to the classical method. The lower panel results show less variability as compared to the model in which the expected number is considered to be a fixed value (middle panel). By incorporating more variability for the expected values, a smoothed map is observed for the new method (lower panel), leading to a more accurate approximation of the Poisson convolution model results.

4. Conclusion

In this paper, we have proposed a method similar to methods used in point-pattern data [4], in which the in-

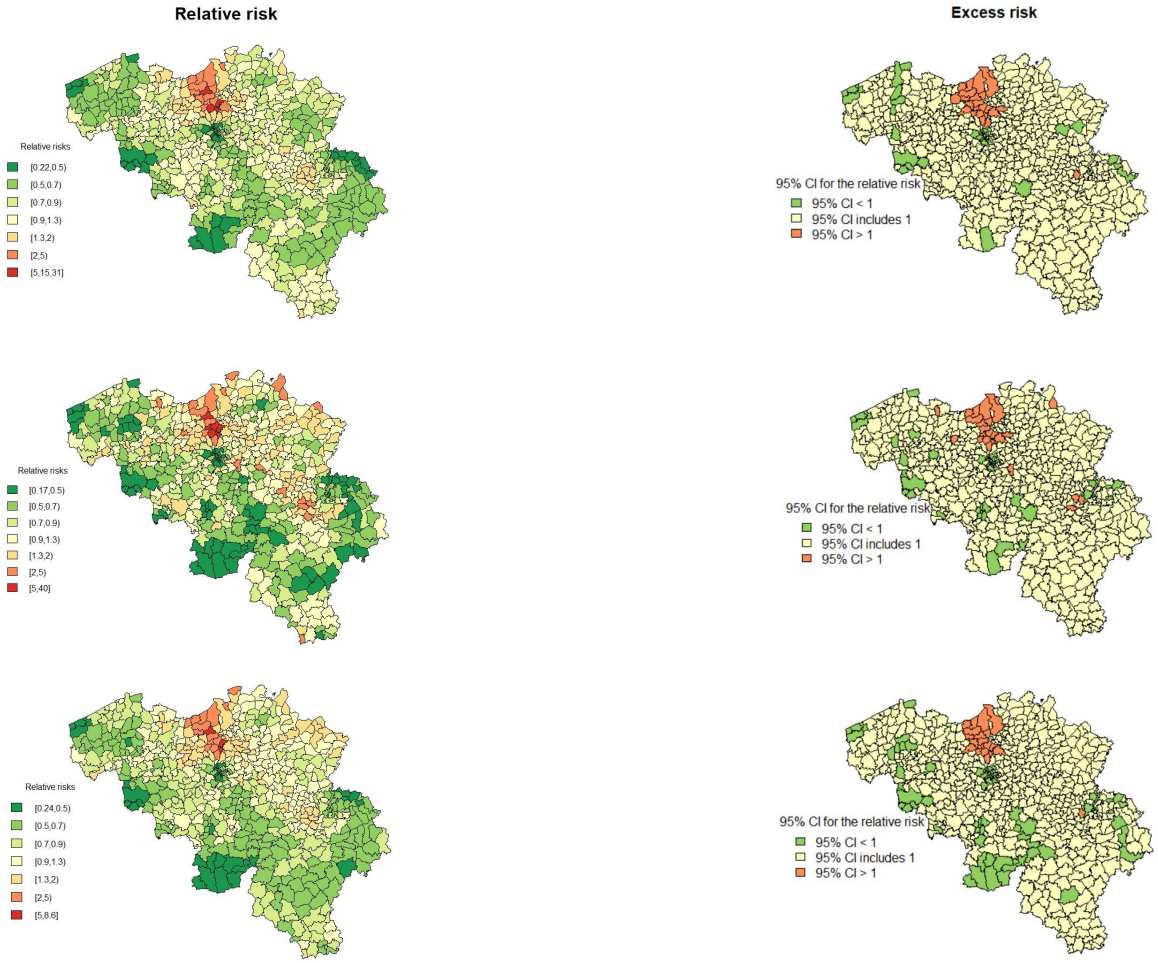


Figure 1: Map of the relative risks for the Poisson Convolution model. Upper panel: using indirect standardized number; Middle panel: using control-disease's standardized number; Lower panel: using control-disease's standardized number accounting for uncertainty.

cidence of the disease of interest is compared to the incidence of a control disease, in the context of lattice data. Allowing for extra variability through the use of a distribution for the expected values, leads to a control-disease approach used for a Poisson convolution model which had similar results with the classical methodology, where the expected values were calculated based on a standard population. Our proposed method can be used either when a standard population including patients' characteristics factors (including age, gender strata) is missing, or when a standard population is not available at all.

References

[1] Baker, S. G. (1994). The multinomial-Poisson transformation. *The Statistician* **43**, 495-504.

[2] Beale, L., Abellan, J. J., Hodgson, S., & Jarup, L. (2008). Methodologic Issues and Approaches to Spatial Epidemiology. *Environmental Health Perspectives* **116**, 1105–1110.

[3] Besag, J., York, J., & Molli, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics* **43**, 1–20.

[4] Diggle, P. J., Morris, S. E., & Wakefield, J. C. (2000). Point-source modelling using matched case-control data. *Biostatistics* **1**, 89–105.

[5] Kelsall, J. E., & Diggle, P. J. (1998). Spatial variation in risk of disease: a nonparametric binary regression approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **47**, 559–573.

[6] Waller, L. A., & Gotway, C. A. (2004). *Applied spatial statistics for public health data.* John Wiley & Sons. New Jersey.

Spatial return level surfaces for non-stationary spatio-temporal processes

L. Bel[1,*] , J. Gomez-Garcia[1] and B. Sawadogo[1]

[1] *UMR MIA-Paris-Sacaly, AgroParisTech, INRAE, Université Paris-Saclay;*
liliane.bel@agroparitech.fr, jose.gomez-garcia@agroparistech.fr, bewentaore.sawadogo@agroparistech.fr
[*] *Corresponding author*

Abstract. *We extend the return level concept, usually defined point-wise, to spatial surfaces for spatio-temporal processes. By analogy with the univariate theory of excesses above a threshold we use the recent results on convergence to l-Pareto processes for spatial exceedances. We investigate the case of non-stationary spatio-temporal processes by modeling the temporal tail distribution as the product of a stationary tail distribtion and a skedasis temporal function, and expressing the return level as the expected number of exceedances during the return period. The methodology is experienced on simulation of max-stable processes and climate data. In a climate change context it provides spatial scenarios of potential future extreme temperature or precipitation surfaces.*

Keywords. *Spatio-temporal processes; Non-stationarity; ℓ-Pareto processes ; Return level.*

1. Introduction

We are interested in this work in investigating future extreme values of some climate variables. When dealing with climate data, especially in the framework of climate change a major issue is the non stationarity in space and time. An usual way to deal with non-stationarities in space is to transform margins via a function of the coordinates and working pointwise in the stationary framework. Similarly non stationarity in time may be handled through modelling the marginal parameters as a function of time, polynomials or splines for instance. Working pointwise does not allow us to take into account the spatial structure of the data and spatial models are needed to describe to spatial dependence. To deal with the spatial nature of climate data when focusing on extreme values we use the ℓ-Pareto models introduced by [5] that are an extension of the Generalized Pareto Distribution for spatial exceedances given by the ℓ function. Non-stationarities in space and time are handled as in [7] via a skedasis function that establishes a relationship between the tail distribution of the non-stationnary spatio-temporal process and a spatial latent process. A by-product of the Extreme Value Theory is to provide indicators such as return levels associated to return periods which are widely used by decision makers. Extension of the original definition of return levels to the non stationary case have been provided by several authors [8], [3] writing the return level as value that has a probability to be exceeded once during the period τ according to the non-stationary distribution of the margins. This gives point-wise return levels that can be spatialized through kriging or other spatial extrapolation methods. We define return level surfaces similarly as a surface which is exceeded by the spatio-temporal process in average once during the period τ. The surface is calculated

by means of simulations according to the estimated ℓ-Pareto model and the marginal non-stationarities. The method is illustrated on simulation and temperature data which are likely fulfilling the max-stable assumption necessary for the ℓ-Pareto modelling.

2. Methodology

Let $X = \{X_t(s), s \in S, t \in T\}$ be a continuous non-stationary space-time stochastic process, $F_{t,s}$ the continuous univariate marginal distribution with a common right endpoint x_F. X is observed at locations $s_1, \cdots, s_m$ at times $t = 1, \cdots, n$. Let $Z = \{Z(s), s \in S\}$ be an unobserved latent spatial stationary process satisfying the proportional tail condition

$$\lim_{x \to x_F} \frac{P(X_t(s_j) > x)}{P(Z(s_j) > x)} = c_\theta\left(\frac{t}{n}, s_j\right), \text{with} \frac{1}{m}\sum_{j=1}^{m} \int_0^1 c_\theta(u, s_j)\, du = 1, \tag{1}$$

with $c_\theta : [0,1] \times S \longrightarrow (0, \infty)$ a continuous and positive function depending on a parameter vector θ, called tail trend or skedasis function [7]. The skedasis function is the density of the point process made by the time occurrences of exeedances of x at location s.

Assuming that there are normalization functions $a_n(.) > 0$, $b_n(.) \in \mathbb{R}$ such that $F_{t,s}$ are in the maximum domain of attraction of common index $\gamma \in \mathbb{R}$ then peudo observations of Z may be recovered from observations of X writing [2] :

$$Z_t(s_j) = \left\{c_\theta\left(\frac{t}{n}, s_j\right)\right\}^{-\gamma} \left[X_t(s_j) - \frac{\left\{c_\theta\left(\frac{t}{n}, s_j\right)\right\}^{\gamma} - 1}{\gamma}\left(a_n(s_j) - \gamma b_n(s_j)\right)\right], \tag{2}$$
$$j = 1, \cdots, m\,;\, t = 1, \cdots, n,$$

If it is assumed a parametric form for the skedasis function $c_\theta(., s_j)$ the parameters θ may be estimated by maximum likelihood. Similarly the marginal parameters $a_n(s_j)$, $b_n(s_j)$ and γ are estimated by independent maximum likelihood.

The extremal behaviour of the process Z is then modeled as an ℓ-Pareto process [5]. If Z is regularly varying with exponent γ and spectral measure σ and $\ell : \mathcal{C}(S) \longrightarrow \mathbb{R}^+$ is a continuous and homogeneous non-negative function then

$$P\left\{u^{-1}Z \in . \mid \ell(Z) \geq u\right\} \longrightarrow P\left\{W^{\ell}_{\gamma,\sigma} \in .\right\}, u \to \infty, \tag{3}$$

$W^{\ell}_{\gamma,\sigma}$ is a ℓ-Pareto process, it can be represented $W^{\ell}_{\gamma,\sigma} = P_\gamma Y$, P_γ is a γ-Pareto r.v. and Y is a continuous process with distribution σ. Selecting the process Y with a distribution that lies in a parametric family which parameters may be inferred leads to an estimation of the spatial structure. A standard choice is a log-Gaussian process with a parametric variogram function, for which the corresponding max-stable process is the well-known Brown-Resnick process. Algorithms for estimating and simulating ℓ-Pareto process are developed by [4], [1], and software are available.

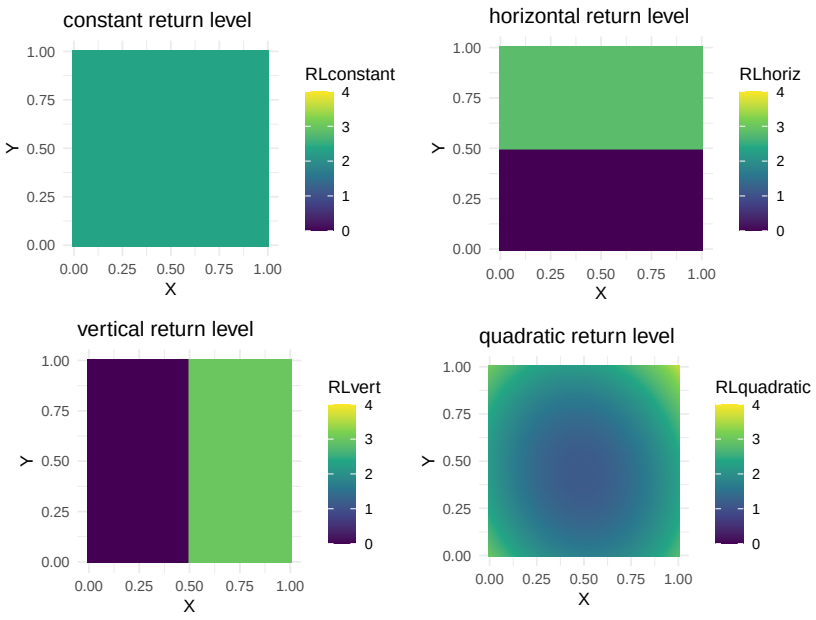


Figure 1: Different shapes for the 500 period return level.

Several definition have been proposed to extend the notion of return levels to the non-stationary case. The Expected Number of Exceedances(ENE) [8] $x_\tau(s)$ is the value for which the expected number of events exceeding $x_\tau(s)$ during the period $1,\dots,\tau$ equals to one, it is the solution of the equation:

$$1 = \sum_{t=1}^{\tau} \mathbb{P}\left(X_t(s) > x_\tau(s)\right) = \sum_{t=1}^{\tau} 1 - F_{t,s}\left(x_\tau(s)\right) \tag{4}$$

We define the surface level return the same way, as a function $f(s)$ minimizing the following criterion: Using the transformation from the non-stationary process $X_t(s)$ to the stationary process $Z(s)$ and using the convergence result to a ℓ-Pareto process of the ℓ excesses of Z we can write :

$$\mathbb{P}\left(X_{t,.} > f(.)\right) \approx \mathbb{P}(W^{\ell}_{\gamma,\sigma} > L_t(f))\mathbb{P}\left(\ell\left(Z\right) > u\right)$$

with $L_t(f) = \left(c_\theta^{-\gamma}(t,.)\left[f - \dfrac{c_\theta^{-\gamma}(t,.)-1}{\gamma}(a_n(.) - \gamma b_n(.)\right] - b_n(.)\right)/\ell(a_n)$.
$\mathbb{P}(W^{\ell}_{\gamma,\sigma} > L_t(f))$ is estimated thanks to simulations of the ℓ-Pareto process $W^{\ell}_{\gamma,\sigma}$ that has been estimated previously, and $\mathbb{P}\left(\ell(Z) > u\right)$ is estimated empirically on the pseudo-observations.
Giving a parametric form to the function $f(.)$, for instance polynomial, allows us to retrieve $f(.)$ using a standard optimization algorithm.

3. Simulation and application to climate data

We simulate a data set of 1000 replications of a spatial process according to a Brown-Resnick model. Then we simulate 10000 replications of the associated ℓ-Pareto processes for the risk $\ell(X) = max(X)$ and we check the matching of some basic statistics on the marginals of the ℓ exceedances. With the ℓ-Pareto simulations we calculate the probabilities necessary to derive the return level surfaces according to several shapes. Figure 1 shows for constant, horizontal vertical and quadratic shapes, the 500 period return level resulting surfaces. The methodology is used to derive return surfaces for daily maximum temperatures in France according to different evolution scenarios.

References

[1] Belzile, L. (2022). mev: Modelling Extreme Values. *R package version 1.14* `https://CRAN.R-project.org/package=mev`

[2] Cabral, R. and Ferreira, A. and Friederichs, P. (2020). Space-time trends and dependence of precipitation extremes in North-Western Germany. *Environmetrics* **31** e2605.

[3] Cooley, D. (2013). Return periods and return levels under climate change, *Extremes in a changing climate*, 97–114, Springer

[4] De Fondeville, R. and Davison, A.C. (2018). High-dimensional peaks-over-threshold inference. *Biometrika* **105** 575–592.

[5] Dombry, C. and Ribatet, M. (2015). Functional regular variations, Pareto processes and peaks over threshold *Statistics and its Interface*, **8**, 1, 9–17.

[6] Ferreira, A. and De Haan, L. (2014). The generalized Pareto process; with a view towards application and simulation. *Bernoulli* **20** 1717–1737.

[7] Mefleh, A. and Biard, R. and Dombry, C. and Khraibani, Z. (2020). Trend detection for heteroscedastic extremes, *Extremes*, **=**23, 1, 85–115.

[8] Parey, S., Hoang, T. and Dacunha-Castelle, D. (2010). Different ways to compute temperature return levels in the climate change context, *Environmetrics*, **21**, 7-8, 698–718.

Assessing spatio-temporal point process intensities using adaptive kernel estimators

J.A. González[1,*] and P. Moraga[1]

[1]*Computer, Electrical and Mathematical Science and Engineering Division, King Abdullah University of Science and Technology; jonathan.gonzalez@kaust.edu.sa, paula.moraga@kaust.edu.sa*
[*]*Corresponding author*

Abstract. *We extend the adaptive methodology for estimating the first-order intensity function of a point process from the planar case to the spatio-temporal case. In this context, two bandwidths are considered for each point in a point pattern, one for space and another for time, and a non-separable estimator is considered by summing the contributions of kernel weights. We formulate several statistical approaches for facing the issues of adaptive estimation in the spatio-temporal case. In particular, we extend some algorithms such as bandwidth binning and the fast Fourier approach to accelerate the computing of the adaptive estimator.*

Keywords. *Bandwidth selection; Bandwidth binning; Fourier transform; Intensity function; Spatio-temporal point process.*

1. Introduction

When we talk about spatio-temporal point processes, one of the essential characteristics of a given observation, that is, of a point pattern, is the first-order intensity, which corresponds to the expected number of points per unit area in the observation window [8]. Kernel smoothing is a non-parametric technique classically used to estimate some types of functions such as probability density functions. This technique in spatial and spatio-temporal statistics is increasingly common, especially in those cases where additional information is not available to understand the distribution of points in space or space-time [7].

One of the main disadvantages of Kernel estimation is the prior knowledge of the bandwidth. Regardless of the data dimensionality, this smoothing parameter is fundamental for the adequate estimation of the intensity and a wrong choice may have unfortunate consequences [2]. It is widespread in the statistical literature to use fixed bandwidth kernels, mainly due to the simplicity of the implementation and the steps involved. However, this classical approach's lack of spatial and temporal adaptability often results in poor estimation, especially in highly heterogeneous point processes with very complicated underlying features that affect the intensity structure within the study region [4, 5]. The fixed kernel density estimator will struggle to capture important finer details in crowded areas when much smoothing is applied to control noise where data is sparse. Conversely, if less smoothing is applied (to retain more of this detail), we can expect to see spurious bumps generated by isolated points in low-density regions.

A more intuitive approach consists of variable smoothing. in this technique, the amount of smoothing is

inversely related to the density of the points. Known as *adaptive*, this smoothing has shown better levels of bias [1, 9] and lower integrated square error [6]. We can solve this issue by the rescaling the bandwidth of the kernels associated with the data points. This technique is called the *point fitting, sample smoothing* or *sample point* method [11]. Following [4], in this work, we focus on adaptive point estimators with the *square root* methodology, i.e., the variable bandwidths are inversely proportional to the square root of the underlying intensity function [1]. To illustrate the idea of point adaptation, Figure 1 provides an example of a spatio-temporal point pattern. The ellipsoids correspond to a variable bandwidth in space and time for each point.

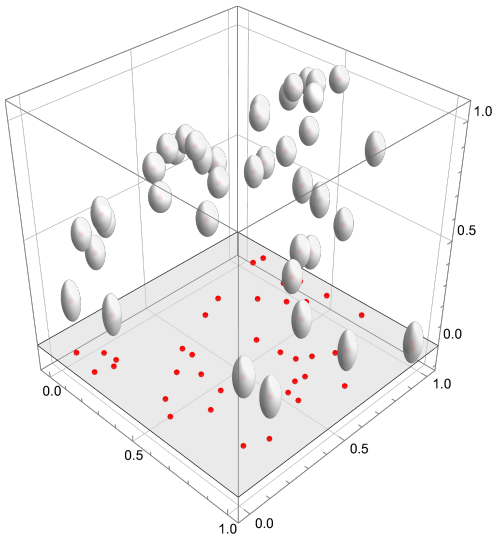


Figure 1: Spatio-temporal point pattern with variable kernel bandwidth in space and time for each point (as translucent isosurfaces). The equatorial radii of the ellipsoids represent the spatial bandwidths and the polar radii the temporal ones.

2. Statistical methodology

2.1 First-order intensity function

We consider a spatio-temporal point pattern X as a countable set of random points with spatial and temporal coordinates; X has a generating stochastic underlying mechanism known as a point process. The number of points in every subset of a spatio-temporal observation window $W \times T$ governs the univariate distributions of the points of X. If the *intensity function* $\lambda(\cdot)$ exists,

$$\mathbb{E}[N(A \times B)] = \int_{A \times B} \lambda(\mathbf{u}, v) \mathrm{d}\mathbf{u}\mathrm{d}v.$$

When $\lambda()$ is a constant, X is called *homogeneous*.

2.2 Adaptive estimation

An adaptive kernel estimator for the first-order intensity function in the spatio-temporal case can be written as

$$\hat{\lambda}_{\varepsilon,\delta}(\mathbf{u},v) = \frac{1}{e_{\varepsilon,\delta}(\mathbf{u},v)} \sum_{i=1}^{n} K^{s}_{\varepsilon(\mathbf{u}_i)}(\mathbf{u}-\mathbf{u}_i)\, K^{t}_{\delta(v_i)}(v-v_i), \qquad (\mathbf{u},v) \in W \times T,$$

where

$$e_{\varepsilon,\delta}(\mathbf{u},v) = \int_W \int_T K^{s}_{\varepsilon(\mathbf{u}')}\left(\mathbf{u}-\mathbf{u}'\right) K^{t}_{\delta(v')}\left(v-v'\right) \mathrm{d}\mathbf{u}'\mathrm{d}v',$$

is an edge correction extended from the proposed by [10] , and the bandwidth functions are defined as

$$\varepsilon(\mathbf{u}) = \frac{\varepsilon^{\star}}{\gamma^{s}} \sqrt{\frac{n}{\lambda^{s}(\mathbf{u})}}, \quad \text{and} \quad \delta(v) = \frac{\delta^{\star}}{\gamma^{t}} \sqrt{\frac{n}{\lambda^{t}(v)}},$$

where $\varepsilon^{\star}, \delta^{\star}$ are *global bandwidths*, $\lambda^{s}(\mathbf{u}), \lambda^{t}(v)$ are marginal intensity functions in space and time, and γ^{s}, γ^{t} are the geometric mean terms for the marginal intensities evaluated in the points of the point pattern.

3. Efficient computation

3.1 Partitioning algorithm

We follow the methodology proposed by [4] and discretise the bandwidths chosen for each point through the empirical quartiles of its sampling distribution. Given the set of spatial and temporal bandwidths $\{\varepsilon_1,\ldots,\varepsilon_n\}$ and $\{\delta_1,\ldots,\delta_n\}$, we consider the empirical pth quantiles, $\hat{\varepsilon}^{(p)}$ and $\hat{\delta}^{(p)}$ of the bandwidths together with two *quantile steps*, $\xi_1, \xi_2 \in (0,1]$ such that $C_1 = \xi_1^{-1}$ and $C_2 = \xi_2^{-1}$ are integers. We define the bandwidth bins through the values $\{\hat{\varepsilon}^{(0)}, \hat{\varepsilon}^{(\xi_1)}, \hat{\varepsilon}^{(2\xi_1)}, \ldots, \hat{\varepsilon}^{(1)}\}$, and $\{\hat{\delta}^{(0)}, \hat{\delta}^{(\xi_2)}, \hat{\delta}^{(2\xi_2)}, \ldots, \hat{\delta}^{(1)}\}$, and we place each observation $(\mathbf{u}_i, v_i)$ in one of the bins.

The sets of bins generate a disjoint partition of the original point pattern X into $C_1 \times C_2$ sets Y_{ij}, and

$$X = \bigcup_{ij} Y_{ij}.$$

If in each subset Y_{ij} the intensity is estimated using a bandwidth defined as the midpoint of the respective bin where it belongs, then the intensity can be approximated as,

$$\hat{\lambda}_{\varepsilon,\delta}(\mathbf{u},v) \approx \sum_{i=1}^{C_1} \sum_{j=1}^{C_2} \hat{\lambda}^{*}_{\bar{\varepsilon}_i,\bar{\delta}_j}(\mathbf{u},v|Y_{ij}),$$

where $\bar{\varepsilon}_i$ and $\bar{\delta}_j$ represent the midpoints of the ith spatial and jth temporal bins and $\hat{\lambda}^*_{\bar{\varepsilon}_i,\bar{\delta}_j}(\mathbf{u},v|Y_{ij})$ is a fixed-bandwidth estimate based on the sub-pattern Y_{ij}.

3.2 Estimation via 5D FFT

The main idea for this technique is to introduce two extra dimensions that represent the logarithms of the spatial and temporal bandwidths ans see the problem using a *scale space* approach [3]. We then define a five-dimensional kernel in scale space as

$$\mathscr{K}(x,y,\varepsilon,t,\delta) = K^2_{\exp(-\varepsilon)}(x,y)K^1_{\exp(-\delta)}(t),$$

where K^2 and K^1 represent the two- and one-dimensional Gaussian kernel functions. Consider a counting measure $\mathcal{N}$ that assigns 1 to each of the points $\{x'_i,y'_i,\log(\varepsilon'_i),t'_i,\log(\delta'_i)\}_{i=1}^n$, and then, consider the convolution of the kernel with the measure,

$$\begin{aligned}(\mathscr{K}*\mathcal{N})(x,y,\varepsilon,t,\delta) &= \int_{\mathbb{R}^5}\mathscr{K}(x-x_0,y-y_0,\varepsilon-\varepsilon_0,t-t_0,\delta-\delta_0)\mathrm{d}\mathcal{N}(x_0,y_0,\varepsilon_0,t_0,\delta_0)\\ &= \sum_{i=1}^n \mathscr{K}(x-x'_i,y-y'_i,\varepsilon-\varepsilon'_i,t-t'_i,\delta-\delta'_i)\\ &= \sum_{i=1}^n K^2_{\varepsilon'_i\exp(-\varepsilon)}(x-x'_i,y-y'_i)K^1_{\delta'_i\exp(-\delta)}(t-t'_i).\end{aligned}$$

For the edge correction can be written analogously as a convolution of the kernel in scale space and the Lebesgue measure. It follows that the convolution evaluated in the hyperplane where $\varepsilon=\delta=0$ yelds $\hat{\lambda}_{\varepsilon,\delta}(\mathbf{u},v)$. One of the fundamental properties of the convolution is that the Fourier transform of a convolution of two integrable functions is the point-wise product of their Fourier transforms. It implies that we can compute the adaptive intensity using the Fourier transform, which is always faster than the traditional algorithms.

References

[1] Abramson, I. S. (1982). On bandwidth variation in kernel estimates-a square root law, *The Annals of Statistics,* **10(4),** 1217-1223.

[2] Baddeley, A., Rubak, E. and Turner, R. (2015). Spatial Point Patterns: Methodology and Applications with R, Chapman & Hall Interdisciplinary Statistics Series, CRC Press, Boca Raton, Florida.

[3] Chaudhuri, P. and Marron, J. S. (2000). Scale space view of curve estimation, *The Annals of Statistics,* **28(2),** 408-428.

[4] Davies, T. M. and Baddeley, A. (2018). Fast computation of spatially adaptive kernel estimates, *Statistics and Computing,* **28(4),** 937-956.

[5] Davies, T. M., Marshall, J. C. and Hazelton, M. L. (2018). Tutorial on kernel estimation of continuous spatial and spatiotemporal relative risk, *Statistics in Medicine,* **37(7),** 1191-1221.

[6] Diggle, P. J., Rowlingson, B. and Su, T.-l. (2005). Point process methodology for on-line spatio-temporal disease surveillance, *Environmetrics,* **16(5),** 423-434.

[7] Fernando, W. S. and Hazelton, M. L. (2014). Generalizing the spatial relative risk function, *Spatial and Spatio-temporal Epidemiology,* **8**, 1-10.

[8] Jonatan A. González and Francisco J. Rodríguez-Cortés and Ottmar Cronie and Jorge Mateu. (2016). Spatio-temporal point process statistics: A review. Spatial Statistics, **18(B),** 505-544.

[9] Hall, P. and Marron, J. S. (1988). Variable window width kernel estimates of probability densities, *Probability Theory and Related Fields,* **80(1),** 37-49.

[10] Marshall, J. C. and Hazelton, M. L. (2010). Boundary kernels for adaptive density estimators on regions with irregular boundaries, *Journal of Multivariate Analysis,* **101(4),** 949-963.

[11] Scott, D. W. (2015). Multivariate density estimation: theory, practice, and visualization, Wiley series in probability and statistics, second edn, John Wiley & Sons.

Spatial heterogeneity of Covid-19 cases in Italy

M. Franco-Villoria[1,*] , M. Ventrucci[2] and H. Rue[3]

[1] *University of Modena and Reggio Emilia; maria.francovilloria@unimore.it*
[2] *University of Bologna; massimo.ventrucci@unibo.it*
[3] *King Abdullah University of Science and Technology; haavard.rue@kaust.edu.sa*
[*] *Corresponding author*

Abstract. *We study weekly Covid-19 incidence rates in Italy from the 24^{th} February 2020 to late July 2021 using space-time disease mapping models. These models often include an interaction term to account for the complexity in the data. In a Bayesian hierarchical framework, prior elicition of precision parameters, responsible for the smoothness of the process, remains a difficult task. We use an intuitive reparametrization of space-time interaction models by means of mixing parameters that control the proportion of variability explained by each term.*

Keywords. *Variance partitioning; Kronecker product IGMRF; Penalized complexity prior.*

1. Introduction

Italy was the first European country to report Covid-19 cases in the early beginning of 2020. The outbreak of the pandemic was followed by a national lockdown to try and control the expansion of the virus. Data on new Covid-19 cases have been collected since then at province level. Knowledge of the space-time evolution of the disease, that can be studied using disease mapping models [1, 2, 3, 4], can be useful to evaluate the effectiveness of the measures put into place by local and/or national authorities. We study variations in Covid-19 incidence rates in Italy using a recently proposed reparametrization [5] of the space-time interaction models introduced by Knorr-Held [6].

2. The data

Weekly data on new Covid-19 cases are available for all of the 107 Italian provinces for 70 weeks, going from 24^{th} February 2020 to late July 2021. We want to investigate whether the disease has spread differently across geographical macro-regions and time windows. In particular, we are interested in assessing the contribution of the space-time interaction to the total variability in incidence rates, as this can be considered as a proxy for local heterogeneity [7]. We divide the dataset in 3 subsets corresponding to the northern (N), central (C) and southern (S) regions (Figure 1). For each of these, we consider two time periods: W1, corresponding to the first 18 weeks (national lockdown period) and W2, that covers the rest of the time window. The time series plots of the data can be seen in Figure 2.

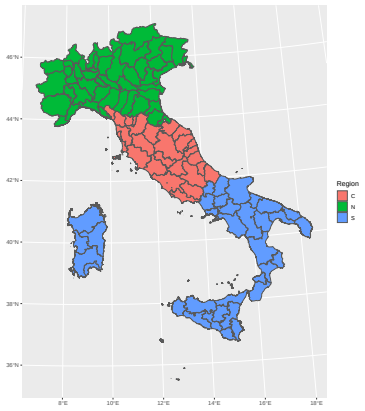


Figure 1: Northern (N), central (C) and souther (S) Italian regions.

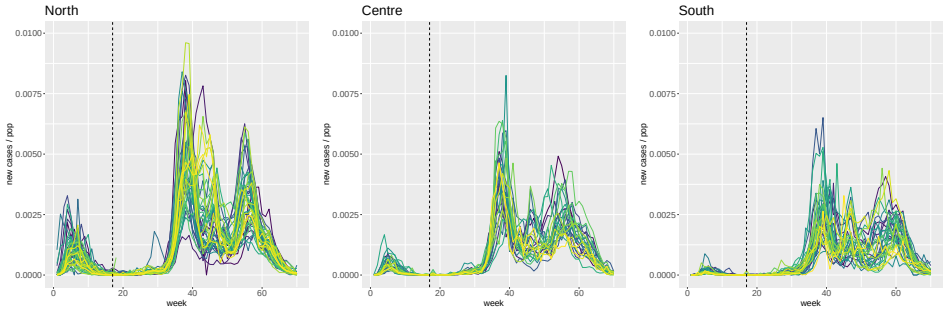


Figure 2: Weekly Covid-19 incidence rates in the North (left), Centre (central) and South (right) of Italy. The vertical dashed line marks the separation between the first (W1) and second (W2) time period.

3. The model

Let y_{ij} be the number of new Covid-19 cases at week $i = 1,\ldots,n_1$ and province $j = 1,\ldots,n_2$ and pop_j be the population at risk in province j. We consider the variance partitioning model proposed by Franco-Villoria et al. [5]:

$$
\begin{aligned}
y_{ij} &\sim \text{Bin}(\text{pop}_j, \exp(\eta_{ij})/\exp(1+\eta_{ij})), \\
\eta_{ij} &= \alpha + \sqrt{\tau^{-1}}\Bigg(\sqrt{1-\gamma}\Bigg(\underbrace{\sqrt{1-\phi}\left(\sqrt{1-\psi_1}\beta_{1_i} + \sqrt{\psi_1}\varepsilon_{1_i}\right)}_{\texttt{main temporal effect}} + \underbrace{\sqrt{\phi}\left(\sqrt{1-\psi_2}\beta_{2_j} + \sqrt{\psi_2}\varepsilon_{2_j}\right)}_{\texttt{main spatial effect}}\Bigg) + \\
&\qquad \underbrace{\sqrt{\gamma}\delta_{ij}}_{\texttt{interaction effect}}\Bigg) \qquad (1) \\
&\boldsymbol{\beta}_1 \sim N\left(\mathbf{0}, \tilde{\boldsymbol{R}}_1^{-}\right), \quad \boldsymbol{\beta}_2 \sim N\left(\mathbf{0}, \tilde{\boldsymbol{R}}_2^{-}\right), \quad \boldsymbol{\delta} \sim N\left(\mathbf{0}, (\tilde{\boldsymbol{R}}_2 \otimes \tilde{\boldsymbol{R}}_1)^{-}\right), \\
&\boldsymbol{\epsilon}_1 \sim N\left(\mathbf{0}, \boldsymbol{I}_{n_1}\right), \quad \boldsymbol{\epsilon}_2 \sim N\left(\mathbf{0}, \boldsymbol{I}_{n_2}\right),
\end{aligned}
$$

where $\tilde{\boldsymbol{R}}_1$ and $\tilde{\boldsymbol{R}}_2$ are the scaled [8, 9] structure matrices of a RW1 and an ICAR models on the temporal and spatial main effects, respectively, $\tau > 0$ is an overall precision parameter and $0 < \gamma < 1, 0 < \phi < 1, 0 < \psi_1 < 1$, $0 < \psi_2 < 1$ are mixing parameters. We assume a type IV space-time interaction on $\boldsymbol{\delta}$, modelled as a Kronecker product IGMRF following Knorr-Held [6]. Model (1) allows to investigate the different sources of variation in Covid-19 incidence rates in a convenient scale. Of particular interest is the mixing parameter γ that represents the contribution of the interaction effect to the total variance. Regarding prior choices, we use the Gumbel type II PC prior for τ proposed by Simpson et al. [10] and the PC prior in Franco-Villoria et al. [5] for γ. The remaining mixing parameters are assigned a uniform prior on (0,1).

4. Results

Table 1 reports posterior estimates for γ for the three geographical areas (N, C, S) and the two time periods (W1, W2). As it can be seen, the impact of the interation term is greater during the second time period, suggesting greater local hetereogeneity from the end of the national lockdown onwards. In both time periods, the interaction explains a slightly greater proportion of variability in the South compared to the other two regions.

	W1			W2		
Region	quant2.5	mean	quant97.5	quant2.5	mean	quant97.5
North	0.23	0.39	0.53	0.43	0.55	0.66
Centre	0.17	0.28	0.41	0.45	0.57	0.67
South	0.27	0.42	0.57	0.55	0.69	0.8

Table 1: Posterior estimates for the mixing parameter γ (contribution of the interaction term), for both time periods W1 and W2 in North, Centre and South regions.

5. Discussion

The reparametrization of the interaction model in Eq. (1) as a weighted sum of main and interaction effects can prove useful in practical applications. The intuitive interpretation of the mixing parameter γ makes prior elicitation simpler, particularly in disease mapping, where the nature of the disease can provide useful information on the importance of the interaction term.

References

[1] Martinez-Beneito, M. and Botella-Rocamora, P. (2019). *Disease Mapping: From Foundations to Multidimensional Modeling (1st ed.)*. Chapman and Hall/CRC.

[2] Lawson, A. (2018). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology (3rd ed.)*. Chapman and Hall/CRC.

[3] MacNab, Y.C. (2022). Bayesian disease mapping: Past, present, and future. *Spatial Statistics*, https://doi.org/10.1016/j.spasta.2022.100593

[4] Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, **8(2)**, 158–183.

[5] Franco-Villoria, M. and Ventrucci, M. and Rue, H. (2021). Variance partitioning in spatio-temporal disease mapping models. *ArXiv:2109.13374*

[6] Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, **19(17-18)**, 2555–2567.

[7] Picado, A. and Guitian, J. and Pfeiffer, D. (2007), Space-time interaction as an indicator of local spread during the 2001 FMD outbreak in the UK. *Preventive Veterinary Medicine*, **79**, 3–19.

[8] Sørbye, S.H. and Rue, H. (2013). Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, **8**, 39–51.

[9] Freni-Sterrantino, A. and Ventrucci, M. and Rue, H. (2018) A note on intrinsic conditional autoregressive models for disconnected graphs. *Spatial and Spatio-Temporal Epidemiology*, **26**, 25–34.

[10] Simpson, D. and Rue, H. and Riebler, A. and Martins, T. and Sørbye, S. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, **32(1)**, 1–28.

Point process learning

O. Cronie[1,*] , M. Moradi[2] and C. Biscio[3]

[1]*Department of Mathematical Sciences, Chalmers University of Technology & University of Gothenburg, 412 96 Gothenburg, Sweden; ottmar@chalmers.se*
[2] *Universidad Publica de Navarra, Campus de Arrosadia, 31006 Pamplona, Pamplona, Spain; mehdi.moradi@unavarra.es*
[3] *Department of Mathematical Sciences, Skjernvej 4A, 9220 Aalborg, Denmark; christophe@math.aau.dk*
[*]*Corresponding author*

Abstract. *Point processes are random sets which generalise the classical notion of a random (iid) sample by allowing i) the sample size to be random and/or ii) the sample points to be dependent. Therefore, point processes have become ubiquitous in the modelling of spatial and/or temporal event data, e.g. earthquakes and disease cases. We here present a recent approach by the authors, point process learning, which is the first statistical learning framework for general point processes. It is based on a subtle combination of two new concepts: prediction errors and cross-validation for point processes. The general idea is to split a point process in two, through thinning, and estimate parameters by predicting one part using the other. By repeating this procedure, we implicitly induce a conditional repeated sampling scheme. Having discussed its properties, we illustrate how it may be applied in a spatial statistical setting and, numerically, show that it outperforms the state of the art.*

Keywords. *(Non-)parametric intensity estimation; Papangelou conditional intensity modelling; Point process cross-validation; Point process prediction; Statistical learning*

Classical statistical learning [5] may compactly be summarised as fitting a family of functions f^* by minimising the risk functional $f^* \mapsto \mathbb{E}[f^*(x_i)]$, under the assumption that $X = \{x_i\}_{i=1}^N \subseteq S$ is a random (iid) fixed-size sample. In practice, however, one minimises the empirical risk, $f^* \mapsto \frac{1}{N}\sum_{i=1}^N f^*(x_i)$, which is motivated by the classical law of large numbers. The typical supervised learning form is obtained when each $x_i = (x_{i1}, x_{i2})$ is a pair of input-output variables and $f^*(x_i) = \mathcal{E}(f(x_{i1}), x_{i2})$ for some family of functions f and some context-specific discrepancy measure $\mathcal{E}$.

One may here ask the relevant question how this should be handled when we deal with the generalised random sampling framework of a point process, i.e. when the elements of X are allowed to be dependent and the sample size is allowed to be random. Such a setting would typically be of interest when one is dealing with predictive modelling of different kinds of point pattern data, e.g. forestry data, earthquakes and disease case data.

In this talk we present the work in [2], which provides a way of doing statistical learning for general point processes. It is based on two new concepts: point process cross-validation and point process prediction errors. The former allows us to consider a form of conditional repeated sampling of the underlying point process and the latter provides a way of measuring how a well parametrised model/characteristic manages to predict one

point process/pattern by means of another process/pattern.

1. Point process preliminaries

Given a (simple) point process $X = \{x_i\}_{i=1}^N$ in a space[1] S, e.g. a d-dimensional Euclidean domain, its (Papangelou) conditional intensity, λ, may be defined through the GNZ formula, which states that

$$\mathbb{E}\left[\sum_{x\in X} h(x, X\setminus\{x\})\right] = \int_S \mathbb{E}[h(u,X)\lambda(u;X)]\mathrm{d}u \tag{1}$$

for any non-negative h on $S\times\mathcal{X}$, where $\mathcal{X}$ is the space of point configurations/patterns in S. Given an infinitesimal neighbourhood $du \ni u \in S$ with measure/volume du, it may be interpreted as $\mathbb{P}(X(du)=1|X\cap du^c = \mathbf{x}\cap du^c) = \lambda(u;\mathbf{x})\mathrm{d}u$, where $X(A) = \#X\cap A$, $A\subseteq S$, and # denotes cardinality. Moreover, the intensity function of X satisfies

$$\rho(u)\mathrm{d}u = \mathbb{E}[X(du)] = \mathbb{E}[\lambda(u;X)]\mathrm{d}u,$$

which has the interpretation that $\mathbb{P}(X(du)=1) = \rho(u)\mathrm{d}u$; we say that X is (in)homogeneous if the intensity function is (non-)constant. By replacing X by $X^n_{\neq} = \{(x_1,\ldots,x_n)\in X^n : x_i\neq x_j \text{ if } i\neq j\}\subseteq S^n$, we obtain the nth order conditional intensity and product density of X.

Conditional intensities have become one of the main tools for full model description of a point process, largely because likelihood functions for point processes contain intractable normalising constants. An example of a model conveniently specified by its conditional intensity is the Strauss hard-core model, for which

$$\lambda_\theta(u;\mathbf{x}) = \beta_{\theta'}(u)\mathbf{1}\left\{u\notin\bigcup_{x\in\mathbf{x}} b(x,R)\right\}, \quad \theta=(\theta',R), \tag{2}$$

where $b(x,R)$ denotes a closed R-ball around x. If we have $\beta_{\theta'}(\cdot)\equiv\theta'>0$, we obtain a homogeneous version of this model.

1.1 Point process statistics

In practice, we observe a point pattern $\mathbf{x} = \{x_i\}_{i=1}^n \subseteq S$, which we assume is a realisations of X. Most estimators we consider can be characterised by a *general parametrised estimator family*

$$\Xi_\Theta = \{\xi_\theta(u;\mathbf{y}) : u\in S, \mathbf{y}\in\mathcal{X}, \theta\in\Theta\}, \quad \Theta\subseteq\mathbb{R}^l, l\geq 1,$$

e.g. a non-parametric intensity estimator, $\xi_\theta = \widehat{\rho}_\theta$, or a parametric family of conditional intensities, $\xi_\theta = \lambda_\theta$,

[1] Polish

and in some cases it is constant in the sense that

$$\xi_\theta(\cdot;\mathbf{y}) = \xi_\theta(\cdot), \quad \mathbf{y} \in \mathcal{X},$$

which e.g. is the case for a parametric intensity estimator, $\xi_\theta = \rho_\theta$. In addition, intensity estimation is typically carried out using i) $\rho_\theta(u) \equiv \theta > 0$ if X is assumed to be homogeneous, ii) a kernel intensity estimator $\widehat{\rho}_\theta(u;\mathbf{x})$, where $\theta > 0$ is the smoothing bandwidth, and iii) $\rho_\theta(u) = \exp\{z(u)^T\theta\}$ when the intensity is modelled parametrically by means of covariates $z(u) = (z_1(u),\dots,z_l(u))^T$ on S.

To obtain an estimate $\widehat{\theta} = \widehat{\theta}(\mathbf{x}) \in \Theta$ of the true parameter $\theta_0 \in \Theta$, e.g. $\lambda = \lambda_{\theta_0}$, one typically specifies some loss function $\mathcal{L}(\theta) = \mathcal{L}(\xi_\theta(\cdot;\mathbf{x})) \geq 0$, $\theta \in \Theta$, to be minimised. In other words, this is the definition of the statistical method used to fit the model.

2. Point process learning

As previously mentioned, our approach is based on two new concepts, which we present below.

2.1 Cross-validation

Cross-validation (CV) partitioning is simply about splitting the dataset into two parts, where typically one of the parts is referred to as a training set and the other part as a validation set. Hence, in point process terminology, this means carrying out thinning and calling the retained points the validation set.

Definition 1 *Split the point pattern* $\mathbf{x}$ *(point process* X*) into two parts, a training set* $\mathbf{x}^T$ *(*X^T*) and a validation set* $\mathbf{x}^V = \mathbf{x} \setminus \mathbf{x}^T$ *(*$X^V = X \setminus X^T$*), using some partitioning mechanism (thinning); repeat this procedure* $k \geq 1$ *times to generate the pairs* $(\mathbf{x}_i^T, \mathbf{x}_i^V)$*,* $i = 1,\dots,k$*.*

Training sets are typically used for model fitting whereas validation sets are used for model validation. As we will see, here they are used somewhat differently. Because independent thinnings are particularly tractable, we propose to carry out CV by means of p-thinning, i.e. each $\mathbf{x}_i^V$ is a p-thinning with retention probability $p \in (0,1)$. Note e.g. that

$$\rho_{X^V}(u) = p\rho_X(u),$$
$$\lambda_{X^V}(u;X^V) = p\mathbb{E}[\lambda_X(u;X)|X^V].$$

A version of k-fold CV here would be to independently assign a label $i \in \{1,\dots,k\}$, $k \geq 2$, to each point, where each label has assignment probability $1/k$, and let $\mathbf{x}_i^V$ consist of all points with label i. Another approach is Monte-Carlo CV, where we independently assign each point $x \in \mathbf{x}$ to $\mathbf{x}_i^V$ with assignment probability p.

2.2 Prediction errors

Consider general parametrised estimator families $\Xi_\Theta = \{\xi_\theta : \theta \in \Theta\}$ and $\mathcal{H}_\Theta = \{h_\theta : \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^l$, $l \geq 1$, and refer to $\mathcal{H}_\Theta$ as *test/weight functions*.

Definition 2 *For point processes Z and Y, and any $A \subseteq S$, the* ($\mathcal{H}$-weighted) bivariate prediction errors *are defined as*

$$I^{h_\theta}_{\xi_\theta}(A;Z,Y) = \begin{cases} \sum_{x\in Z\cap A} h_\theta(x) - \int_A h_\theta(u)\xi_\theta(u)\mathrm{d}u & \text{if } \xi_\theta(\cdot;\mathbf{y}) = \xi_\theta(\cdot), \\ \sum_{x\in Z\cap A} h_\theta(x;Y\setminus\{x\}) - \int_A h_\theta(u;Y)\xi_\theta(u;Y)\mathrm{d}u & \text{otherwise.} \end{cases}$$

Here, the the random sum collects h_θ-generated "predictions" of the points of $Z\cap A$ (based on $Y\cap A$) and the integral represents a ξ_θ-governed "expected counterpart"/"compensator". This is made precise by the following result:

Theorem 1 *Given a point process X in S, let X^V be a p-thinning of X, $p \in (0,1)$, and let $X^T = X \setminus X^V$. For a parametric family of intensity functions, ρ_θ, $\theta \in \Theta$, if $\xi_\theta(\cdot) = p\rho_\theta(\cdot)$ then $\mathbb{E}[I^{h_\theta}_{\xi_\theta}(A;X^V,X^T)] = 0$ for any A and any test function if and only if $\rho_\theta = \rho_{\theta_0}$ a.e.. For a parametric family of conditional intensity functions, λ_θ, $\theta \in \Theta$, if $\xi_\theta(\cdot) = p\lambda_\theta(\cdot)/(1-p)$ then $\mathbb{E}[I^{h_\theta}_{\xi_\theta}(A;X^V,X^T)] = 0$ for any A and any test function if and only if $\lambda_\theta = \lambda_{\theta_0}$ a.e..*

Note here that i) when $\xi_\theta = \rho_\theta$ is a parametric intensity estimator, then $\mathcal{L}(\theta) = I^{h_\theta}_{\rho_\theta}(W;\mathbf{x},\mathbf{x})$ essentially yields quasi-likelihood estimation [4], when $\xi_\theta = \lambda_\theta$ is a parametric conditional intensity estimator then $\mathcal{L}(\theta) = I^{h_\theta}_{\lambda_\theta}(W;\mathbf{x},\mathbf{x})$ is an innovation [1] which yields Takacs-Fiksel estimation, and iii) when $\xi_\theta = \widehat{\rho}_\theta$ is a non-parametric intensity estimator then $\mathcal{L}(\theta) = I^{h_\theta}_{\widehat{\rho}_\theta}(W;\mathbf{x},\mathbf{x})$ yields the loss functions in [3].

2.3 Point process learning

Given training-validation set pairs $(\mathbf{x}_i^V, \mathbf{x}_i^T)$, $i = 1,\ldots,k$, consider

$$I_i(\theta) = I^{h_\theta}_{\xi_\theta}(W;\mathbf{x}_i^V,\mathbf{x}_i^T) = \sum_{x\in\mathbf{x}_i^V\cap W} h_\theta(x) - p\int_W h_\theta(u)\rho_\theta(u)\mathrm{d}u$$

when doing parametric intensity estimation and

$$I_i(\theta) = I^{h_\theta}_{\xi_\theta}(W;\mathbf{x}_i^V,\mathbf{x}_i^T) = \sum_{x\in\mathbf{x}_i^V\cap W} h_\theta(x;\mathbf{x}_i^T) - \frac{p}{1-p}\int_W h_\theta(u;\mathbf{x}_i^T)\lambda_\theta(u;\mathbf{x}_i^T)\mathrm{d}u$$

when doing parametric conditional intensity estimation (and non-parametric intensity estimation). Point process learning is based on the idea of minimising these prediction errors in some suitable sense.

Definition 3 *Any method based on exploting* $I_i(\theta)$, $\theta \in \Theta$, $i = 1, \ldots, k$, *to carry out estimation is called a* point process learning *method.*

An example here would be to minimise an associated loss function, e.g.

$$\mathcal{L}_j(\theta) = \frac{1}{k} \sum_{i=1}^{k} |I_i(\theta)|^j, \quad \theta \in \Theta, j = 1, 2,$$

where, in many cases, we let $I_i(\theta) = 0$ if any of $\mathbf{x}_i^V, \mathbf{x}_i^T$ is empty. Note that the things we need to specify here include good choices for the CV parameters (k and p) and the test functions $\mathcal{H}_\Theta = \{h_\theta : \theta \in \Theta\}$ (which influence the variances of the prediction errors).

In this talk we will introduce point process learning and illustrate it by looking closer at how it can be used to fit the hard-core model in (2). In particular, we will show that it outperforms pseudolikelihood estimation in terms of MSE.

References

[1] Baddeley, A., Turner, R., Moller, J. and Hazelton, M. (2005). Residual analysis for spatial point pro- cesses (with discussion). *Journal of the Royal Statistical Society: Series B,* **67**(5):617666.

[2] Cronie, O., Moradi, M., and Biscio C, AN. (2021). Statistical learning and cross-validation for point processes. arXiv preprint arXiv:2103.01356,

[3] Cronie, O. and van Lieshout, M.N.M. (2018). A non-model-based approach to bandwidth selection for kernel estimators of spatial intensity functions. *Biometrika*, **105**(2):455462.

[4] Guan, Y., Jalilian, A. and Waagepetersen, R. (2015). Quasi-likelihood for spatial point processes. *Journal of the Royal Statistical Society: Series B*, **77**(3):677697.

[5] Vladimir, V.N. (1990). An overview of statistical learning theory. I*EEE Transactions on Neural Networks*, **10**(5):988999.

Estimation of the intensity of a point process and its support through point process learning

M. Pereira[1,2,*] and O. Cronie[1,2]

[1]*School of Public Health and Community Medicine, University of Gothenburg, Sweden. mike.pereira@chalmers.se*
[2]*Dept. of Mathematical Sciences, Chalmers University of Technology & University of Gothenburg, Sweden; ottmar.cronie@gu.se*
[*]*Corresponding author*

Abstract. *In this work, we propose an approach for the joint estimation of the intensity function of a point process and its support based on a single realization of the point process. This method relies on the minimization of the expectation of loss functions related to point process innovations. This minimization is performed by combining a stochastic gradient descent algorithm and the statistical learning framework for point processes introduced by one of the authors in a previous work.*

Keywords. *Point process learning; Support; Stochastic gradient descent.*

1. Introduction

Let X be a (finite) point process with state space $S = \mathbb{R}^d$, $d \in \mathbb{N}$. We assume in the following that X admits an intensity function ρ_X and aim at estimating jointly ρ_X and its support W_X, as defined by

$$W_X = \{u \in S : \rho_X(u) > 0\},$$

from a realization of X. To that end, let $\Xi = \{\xi_\theta : \theta \in \Theta\}$ be a family of functions parametrized by some set $\Theta \subset \mathbb{R}^{m_1}$ ($m_1 \in \mathbb{N}$), and such that $\xi_\theta : S \to (0, +\infty)$ for any $\theta \in \Theta$. Let $\mathcal{W} = \{W_r : r \in \mathcal{R}\}$ be a family of subsets of S parametrized by some set $\mathcal{R} \subset \mathbb{R}^{m_2}$ ($m_2 \in \mathbb{N}$). In particular, we assume that for any $\theta, \theta' \in \Theta$, $\theta \neq \theta' \Rightarrow \xi_\theta \neq \xi_{\theta'}$, and similarly for any $r, r' \in \mathcal{R}$, $r \neq r' \Rightarrow W_r \neq W_{r'}$. Finally we assume that there exists there exits some $r_0 \in \mathcal{R}$ and some $\theta_0 \in \Theta$ such that ρ_X can be written as

$$\rho_X(u) = \xi_{\theta_0}(u) \mathbb{1}_{W_{r_0}}(u), \quad u \in S.$$

Note that this is in fact a very general setting, as any intensity ρ_X can be written as $\rho_X = \rho_X \mathbb{1}_{W_X}$ where W_X is the support of ρ_X. Then, estimating ρ_X and W_X comes down to estimating the values $r_0 \in \mathcal{R}$ and $\theta_0 \in \Theta$.

In the remainder of the text, and for $p \in (0,1)$, we denote by X_p an independent p-thinning of X. Recall then that X_p has an intensity function ρ_{X_p} given by

$$\rho_{X_p}(u) = p\, \rho_X(u) = p\, \xi_{\theta_0}(u) \mathbb{1}_{W_{r_0}}(u), \quad u \in S.$$

2. Prediction errors

Following the Point Process Learning (PPL) approach of [2], in order to estimate the intensity and support parameters of the point process, we rely on the definition of so-called "prediction errors" which are closely related to the notion of point process innovations [1].

Let $g : S \to \mathbb{R}$ be some (square-integrable) test function, and let $\theta \in \Theta$, $r \in \mathcal{R}$. We introduce the prediction errors $I_g(\theta, r; X_p)$ defined by

$$I_g(\theta, r; X_p) = p \int_{W_r} g(u)\xi_\theta(u)du - \sum_{x \in X_p} g(x). \tag{1}$$

Following the Campbell formula [3], we have that

$$\mathbb{E}[I_g(\theta, r; X_p)] = p \int_S g(u)\big(\xi_\theta(u)\mathbb{1}_{W_r}(u) - \rho_X(u)\big)du = p \int_S g(u)\big(\xi_\theta(u)\mathbb{1}_{W_r}(u) - \xi_{\theta_0}(u)\mathbb{1}_{W_{r_0}}(u)\big)du.$$

Note that, if for any test function g, we have $\mathbb{E}[I_g(\theta, r; X_p)] = 0$, then we must have that for (almost every) $u \in S$,

$$\xi_{\theta_0}(u)\mathbb{1}_{W_{r_0}}(u) = \xi_\theta(u)\mathbb{1}_{W_r}(u). \tag{2}$$

This in turns yields that $W_{r_0} = W_r$ (since any point u in $W_{\theta_0} \setminus W_\theta$ or $W_\theta \setminus W_{\theta_0}$) would contradict (2)), and therefore that $r = r_0$ and $\theta = \theta_0$. The problem of finding the optimal parameters $(\theta_0, r_0) \in \Theta \times \mathcal{R}$ can therefore be restated as follows:

Problem 1 *Find $(\theta, r) \in \Theta \times \mathcal{R}$ such that, for any test function g, $\mathbb{E}[I_g(\theta, r; X_p)] = 0$.*

We now propose an alternative definition of the prediction errors. Let $h : S \to (0, \infty)$ be a test function, and consider the prediction errors defined by

$$\mathcal{J}_h(\theta, r; X_p) = p \int_{W_r} h(u)\xi_\theta(u)^2 du - \sum_{x \in X_p} 2h(x)\xi_\theta(x). \tag{3}$$

Using once again the Campbell formula gives

$$\begin{aligned}\mathbb{E}[\mathcal{J}_h(\theta, r; X)] &= p \int_S h(u)\big(\xi_\theta(u)^2\mathbb{1}_{W_r}(u) - 2\rho_X(u)\xi_\theta(u)\mathbb{1}_{W_r}(u)\big)du \\ &= p \int_S h(u)\big(\xi_\theta(u)\mathbb{1}_{W_r}(u) - \rho_X(u)\big)^2 du - p \int_S h(u)\rho_X(u)^2 du.\end{aligned}$$

Hence, for any choice of test function $h : S \to (0, \infty)$, the function

$$J_h(\theta, r) = \mathbb{E}[\mathcal{J}_h(\theta, r; X)], \quad \theta \in \Theta, \quad r \in \mathcal{R} \tag{4}$$

is minimized for (θ^*, r^*) such that $\xi_{\theta^*}\mathbb{1}_{W_{r^*}} = \rho_X$ over S, which in turn means that $\theta^* = \theta_0$ and $r^* = r_0$. The problem of finding the optimal parameters $(\theta_0, r_0) \in \Theta \times \mathcal{R}$ can therefore be restated as follows:

Problem 2 *Given a test function $h : S \to (0,\infty)$, find the minimizers $(\theta^*, r^*) \in \Theta \times \mathcal{R}$ of the function J_h defined in* (4)*.*

3. Estimation by point process learning

We assume that we observe a realization χ of the point process X, and aim at estimating parameters (θ_0, r_0) defining the intensity of the process and the support of that intensity. Since this is equivalent to solving either Problem 1 or Problem 2, we propose to solve these problems instead, using the PPL approach of [2]. Hence, let $p \in (0,1)$ and let $\chi_p^{(1)}, \dots, \chi_p^{(K)}$ be a large number K of independent p-thinnings of χ.

3.1 Solving Problem 1

Following [2], instead of finding parameters $(\theta, r) \in \Theta \times \mathcal{R}$ such that the expectation of the prediction errors (1) is (exactly) zero, we simplify the problem and try instead to make them as small as possible for a large class of test functions. To do so, we first note that since the map $g \mapsto \mathbb{E}[I_g(\theta, r; X_p)]$ is linear, it is enough to show that $\mathbb{E}[I_g(\theta, r; X_p)] \approx 0$ only for some test functions g taken in some basis of functions of S (eg. polynomial functions). Besides, for any test function g, we have $|\mathbb{E}[I_g(\theta, r; X_p)]| \leq \mathbb{E}[I_g(\theta, r; X_p)^2]^{1/2}$. Hence, instead of minimizing $|\mathbb{E}[I_g(\theta, r; X_p)]|$, we can minimize $\mathbb{E}[I_g(\theta, r; X_p)^2]$.

Following these remarks, let then $\mathcal{G}$ be a finite subset of a basis of test functions of S. We consider the following simplified version of Problem 1:

Problem 1bis *Find $(\theta^*, r^*) \in \Theta \times \mathcal{R}$ that minimize the cost function*

$$L(\theta, r) = \sum_{g \in \mathcal{G}} \mathbb{E}[I_g(\theta, r; X_p)^2] = \mathbb{E}\left[\sum_{g \in \mathcal{G}} I_g(\theta, r; X_p)^2 \right], \quad (\theta, r) \in \Theta \times \mathcal{R}. \tag{5}$$

Note that the cost function (5) can be seen as the expectation, over the distribution of thinned point processes X_p, of the loss function given by

$$y \mapsto \sum_{g \in \mathcal{G}} I_g(\theta, r; y)^2, \tag{6}$$

where y denotes here a point pattern in S. Under the assumption that independent samples of X_p are available, the minimization of this cost function can be handled using a (batch) stochastic gradient algorithm where at each iteration, the gradient of the loss function is computed using a different sample (or a different batch of samples) [5]. Since the only data at hand is the realization χ of X, we propose instead to replace these samples by the thinned patterns $\chi_p^{(1)}, \dots, \chi_p^{(K)}$ computed from χ. This yields the Batch stochastic gradient algorithm presented in Algorithm 1, and that we use to solve Problem 1bis.

Algorithm 1 Batch stochastic gradient algorithm

Input: Thinned patterns $\chi_p^{(1)}, \ldots, \chi_p^{(K)}$, Initial estimates $(\theta_{\text{init}}, r_{\text{init}}) \in \Theta \times \mathcal{R}$, Step size $\alpha \in (0,1)$, Batch size N_{batch}.

Output: Estimates $(\hat{\theta}, \hat{r})$ of the minimizers of the cost function (5) of Problem 1bis.

1: $\hat{\theta} = \theta_{\text{init}}$, $\hat{r} = r_{\text{init}}$

2: **for** $q = 0, \ldots, \lfloor K/N_{\text{batch}} \rfloor$ **do**

3: $$(\hat{\theta}, \hat{r})^T \leftarrow (\hat{\theta}, \hat{r})^T - \alpha \sum_{k=1+qN_{\text{batch}}}^{(q+1)N_{\text{batch}}} \nabla \left(\sum_{g \in \mathcal{G}} I_g(\hat{\theta}, \hat{r}; \chi_p^{(k)})^2 \right)$$

4: **end for**

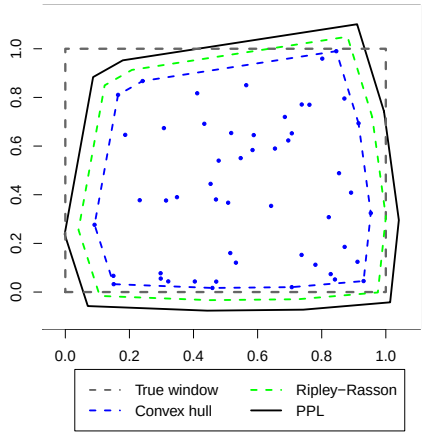
5: **return** $(\hat{\theta}, \hat{r})$.

3.2 Solving Problem 2

Similarly to Problem 1bis, the cost function defining to be minimized in Problem (2) can also be written as the expectation of a non-negative loss function. Hence, Algorithm 1 can be used to solve Problem (2), after substituting the loss function (6) in Step 3 of the algorithm by the loss function $y \mapsto \mathcal{J}_h(\theta, r; y)$ defined in (3).

	Intensity	Area of the support
True	50	1
True support only	52	1
Convex hull	73.2	0.71
Ripley–Rasson	59.13	0.88
PPL	**49.38**	**1.04**

(a) Results of the parameter estimation.



(b) Plot of the point pattern and estimated windows.

Figure 1: Results of the numerical experiment: "True" corresponds to the true process, "Convex hull" (resp. "Ripley–Rasson", "True support only") corresponds to the case where we estimate the support of the intensity by the convex hull of the points (resp. the Ripley–Rasson estimate, the true support of the intensity), "PPL" corresponds to our approach.

4. Example of application

In the following examples, we aim at estimating jointly the intensity of a homogeneous Poisson point process, as well as the support of that intensity, given a single realization χ of the process. We consider a Poisson process with intensity 50, defined on the unit square $[0,1]^2$. To estimate these parameters, we solve Problem 1bis using the Batch stochastic gradient algorithm. Following the approach of Ripley and Rasson [4], we take the family $\mathcal{W}$ of possible supports for the intensity function to be given by scalings of the convex hull of χ. As for the family of test functions $\mathcal{G}$, we choose the polynomial basis functions up to order 2. We choose a thinning parameter $p = 0.25$, a number of thinned patterns $K = 200$ and a batch size $N_{\text{batch}} = 1$.

The results of this numerical experiment are presented in Figure 1. They show that our approach consisting in carrying out a joint estimation of the intensity and its support yields better estimates for both the intensity and the support, compared to more classical approaches consisting in estimating the support first (through the convex hull of the points, their Ripley-Rasson method [4] or even when using the actual support of the intensity), and then estimating the intensity using this support.

References

[1] Baddeley, A., Turner, R., Moller, J. and Hazelton, M. (2005). Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society: Series B,* **67**(5):617666.

[2] Cronie, O., Moradi, M., and Biscio C, AN. (2021). Statistical learning and cross-validation for point processes. arXiv preprint arXiv:2103.01356,

[3] Daley, D. J. and D. Vere-Jones, D.(2008). An introduction to the theory of point processes - Volume II. Springer, New York, 2nd edition, 2008.

[4] Ripley, B.D. and Rasson, J.P. (1977). Finding the edge of a poisson forest. *Journal of Applied probability*, **14**(3):483 491.

[5] Shalev-Shwartz, S. and S. Ben-David, S. (2014) Understanding machine learning: From theory to algorithms. Cambridge university press, 2014.

Classification of intensity functions of inhomogeneous point processes

I. Fuentes-Santos[1,*] , M.I. Borrajo[2] and W. González-Manteiga[2]

[1]*Instituto de Investigaciones Marinas, Spanish National Research Council, Vigo (Spain); isabel.fuentes@usc.es*
[2]*Department of Statistics, Mathematical Analysis and Optimization. University of Santiago de Compostela*
[*]*Corresponding author*

Abstract. *A common question when a given point process is observed in more than one population is whether those patterns share the same structure or they can be partitioned in a ceitain number of groups. A k-means algorithm could be used to classify the densities of event locations. However, the space of density functions does not fulfill Hilbert conditions, so specific measures should be adopted. In this work we propose some possibilities, and we compare their performance through a simulation study. Real data problems, such as the classification of COVID-19 infection curves can be addressed.*

Keywords. *Density of event locations, functional data, k-means algorithm, synthetic data, Wasserstein metric.*

1. Motivation

Point processes are mathematical models generating a random number of events randomly located on a mathematical space, $\mathcal{S}$, such as any d-dimensional Euclidean space, the spatiotemporal domain, or linear graphs. In many real-world problems we observe a point process in two or more populations defined in a common observation domain. This is for example the case of spatiotemporal patterns of wildfire with different cause, different types of crime, or the temporal patterns of COVID-19 infections in differnt regions. In these cases, we may be interested on testing whether these patterns have the same distribution. This issue has already been addressed for two spatial point processs by measuring the distance between their corresponding densities of event locations, see [2] for more details. We could use defining statistical procedures to compare the structure of a larger number of point processes and classify them into groups.

Classification is the problem of identifying which of a set of categories or groups, an observation (or observations) belongs to. Focussing on point processes observed in the one-dimensional Euclidean space for simplicity, our aim is to classify a set of Poisson point processes $\{X_i\}_{i=1}^n$ observed in a given interval $S=[a,b]\subset\mathbb{R}$ into K groups, conditioning on point processes in each group share a same distribution. To address this problem, we estimate the corresponding first-order intensity function $\{\hat{\lambda}_i(x)\}_{i=1}^n$, moving from the point process framework to the intensity space, Ω. As argued by [3], this space can be seen as a product metric space $\Omega=\mathcal{D}\times\Omega_S$, where $\mathcal{D}\subset\{f:S\to\mathbb{R}^+;\int_S f(x)dx=1\}$ denotes the spaces of density functions in S, and $\Omega_S=\mathbb{R}^+$ the space of intensity factors, which determine the shape and expected number of events (size) of the point process, respectively. Therefore, we can use a L^2 product metric, d, between a given pair of intensity functions $\lambda_1=(m_1,f_2)$

and $\lambda_2 = (m_2, f_2)$ given by

$$d(\lambda_1, \lambda_2) = \left(d_{\mathcal{D}}^2(f_1, f_2) + d_E^2(m_1, m_2)\right)^{1/2}, \quad (1)$$

where d_E is the one-dimensional Euclidean metric and $d_{\mathcal{D}}$ is a metric in the density space. Considering this decomposition, the structure of point processes relies on the density of event locations and, consequently, our problem can be reduced to that of classifying density estimates in groups.

2. Classification algorithm

Let $\{X_i\}_{i=1}^n$ be a set of point patterns observed in a bonded interval S, and $\{\hat{f}_i(x), x \in S\}_{i=1}^n$ the kernel estimators of their densities of event locations. Let assume that these point processes belong to K categories characterized by the densities of events locations $\{f_k\}_{k)1}^K$, referred as centers. Classification of the n density estimates into the K groups can be conducted by a k-means algorithm in the space of density functions, $\mathcal{D}$, with a certain metric, d, which is implemented as follows.

Step 1 Estimate the initial centers, $\{f_k^0\}_{k=1}^K$, as the K density estimates maximizing $\sum_{i,j \in p_j} d\left(\hat{f}_i, \hat{f}_j\right)$, where $p_j \in C_{n,K}$, all possible combinations of $\{1, \ldots, n\}$ in groups of K elements.

Step 2 Once obtained the initial centers, compute pairwise distances between the remaining $n-K$ densities and the K centers, $d\left(\hat{f}_i, f_k^0\right)$, and assign each density to the group with the closest center.

Step 3 Once partitioned the n functions into the K clusters, estimate the mean of the density curves in each group to be the density of event locations that characterizes that group,

Intutively, the density estimates can be considered as functional data, and the k-means algortihm for functional data could be used to proceed with classification. However, $\mathcal{D}$ is not Hilbert, and consequently, statistlcal procedures for functional data can not be directly applied in the density space. In particular, we cannot use the L^2 distance as discrepancy measure in the k-means algorithm. Following a common practice for statistical modeling and computing of densities, we should conduct the classification in a representative space. These spaces have been mainly defined under two perspectives, the functional and object-oriented approaches, see details in [4]. In this work we use both of them. We propose a transformation approach for functional data representation, as well as two object-oriented metrics to determine the discrepancy between density functions.

- **Transformation approach (L^2-LQD)**, density curves can be treated as functional data after transformation into the Hilbert space. Here we use the log-quantile density (LQD) transformation, and the L^2 distance in the transformed space:

$$d_{LQD}(f_1, f_2) = \left(\int_0^1 \left(\psi_{LQD(f_1)}(x) - \psi_{LQD(f_2)}(x)\right)^2 dx\right)^{1/2},$$

where $\psi_{LQD(f)}(\cdot)$ denotes the LQD-transformed density.

- **L^2-Wasserstein distance (L^2-WS)**, is an optimal transport distance that measures the cost of transporting one distribution to another in the object-oriented framework and can be defined in quite general spaces. For absolutely continuous distributions it can be defined as the L^2- distance between their respective quantile functions, $Q_j, j = 1,2$:

$$d_W(f_1, f_2) = \left(\int_0^1 (Q_{f_1}(r) - Q_{f_2}(r))^2 \, dr \right)^{1/2}.$$

- **Fisher-Rao distance (FR)**, first used as a Riemmanian structure for parametric models, is the spherical geodesic distance between square root densities:

$$d_{FR}(f_1, f_2) = \arccos \left(\int_a^b \sqrt{f_1(x) f_2(x)} dx) \right),$$

  where, arccos denotes the arcocosine function. The square root of a density lies on the Hilbert unit sphere, so d_{FR} measures the length of an arch connecting $\sqrt{f_1}$ and $\sqrt{f_2}$ along this sphere.

3. Simulation study

We have conducted a simulation study to test the performance of the k-means algorithm for density functions with the different distance measures outlined above, We use the L^2 and Kullback-Leibler (KL) metrics as benchmark criteria. The latter is defined as follows

$$d_{KL}(f_1, f_2) = \frac{1}{2} \left(\int_a^b f_1(x) \log \left(\frac{f_1(x)}{f_2(x)} \right) dx + \int_a^b f_2(x) \log \left(\frac{(f_2(x)}{f_1(x)} \right) dx \right).$$

We simulated $(n_j = 20)_{j=3}^3$ point patterns with densities of event locations given by Model A and Model B, see Figure 1 (thick lines in the first column) with intensity factor $(m_j = 200)_{j=3}^3$, and estimated the density of event locations by kernel smoothing with plug-in bandwidth [1]. Once obtained the density estimates, we applied the k-means algorithm detailed in Section 2. for functional data with the different distance measures under comparison.

Table 1 shows the correct classification rates obtained with the k-means algorithm for functional data with the 5 distance measures under study. The L^2 and Kullback-Leibler correct classification rates are below the 80%. The LQD transformation approach varies between models reported a poor performance for Model A, as the log-quantile densities of the three groups are similar, but a perfect classification for Model B. The object-oriented approaches with the Fisher-Rao and Wasserstein metrics provide perfect classifications for both models.

In real data applications we way not know the number of clusters in the population. Table 1 shows the classification matrix obtained applying the k-means algorithm with $K = 2$ using the FR and l^2-WS metrics. Both approaches identify correctly the point processes in the most distant groups, but distributes those in the remaining group between them. These results highlight the need of some mechanism to estimate the number of

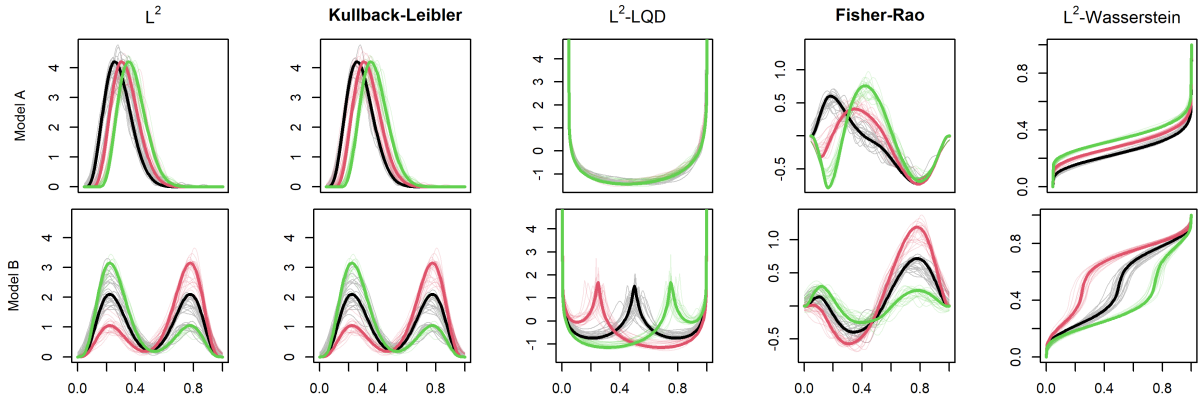


Figure 1: Density of event locations (thick lines) for Models A and B, and density estimates assigned to each cluster (thin lines) in the different representation spaces.

	L^2	KL	L^2-LQD	FR	L^2-WS
Model A	0.78	0.68	0.42	1	1
Model B	0.77	0.75	1.	1	1

Table 1: Right classification rates

groups prior to apply the k-means algorithm.

	Model A				Model B			
	FR		L^2-WS		FR		L^2-WS	
	$\hat{C}_1$	$\hat{C}_2$	$\hat{C}_1$	$\hat{C}_2$	$\hat{C}_1$	$\hat{C}_2$	$\hat{C}_1$	$\hat{C}_2$
C_1	20	0	20	0	20	0	20	0
C_2	5	15	11	9	12	8	1	19
C_3	0	20	0	20	0	20	0	20

Table 2: Classification matrix provided by the k-means algorithm with $K = 2$, $\{C_j\}_{j=1}^3$ target clusters, and $\{\hat{C}_j\}_{j=1}^2$ clusters assigned by the k-means algorithm.

Acknowledgments

This work has been supported by Project PID2020-116587GB-I00 (AEI/FEDER, UE).

References

[1] Chacón, J. E., and Duong, T. (2010). Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *Test*, **19(2)**, 375-398. DOI: 10.1007/s11749-009-0168-4.

[2] Fuentes-Santos, I., González-Manteiga, W., and Mateu, J. (2021). Testing similarity between first order intensities of spatial point processes. a comparative study. *Communications in Statistics- Simulation and Computation*. DOI: 10.1080/03610918.2021.1901118.

[3] Gajardo, A., and Möller, H. G. (2021). Point process models for COVID-19 cases and deaths. *Journal of Applied Statistics*, DOI: 10.1080/02664763.2021.1907839.

[4] Petersen, A., Zhang, C., and Kokoszka, P. (2022). Modeling probability density functions as data objects. *Econometrics and Statistics*, **21**, 159-178. DOI; 10.1016/j.ecosta.2021.04.004.

Spatial statistical calibration on linear networks: an application to the analysis of traffic volumes

A. Gilardi[1,*] , R. Borgoni[1] and J. Mateu[2]

[1] *Universit degli Studi di Milano-Bicocca, Italy; andrea.gilardi@unimib.it, riccardo.borgoni@unimib.it*
[2] *Universitat Jaume I, Castellon, Spain; mateu@uji.es*
[*] *Corresponding author*

Abstract. *The estimation of traffic volumes on a street network represents a fundamental step to im- prove transport planning protocols and develop effective road safety interventions. The traditional ways to derive traffic figures involve manual counts or fine-tuned automatic tools (e.g. cameras or inductive loops). Unfortunately, the manual counts are extremely time consuming, whereas the fixed instruments are typically very expensive and geographically sparse. However, given the increasing availability of mobile sensors (e.g. smartphones and GPS sat-nav), in the last years we observed a surge of methods to infer traffic counts from geo-referenced mobile devices. This paper proposes a spatial statistical calibration technique to combine accurate fixed counts and extensive GPS mobile data for the estimation of traffic flows, re-adapting the statistical methods to the spatial network con- text. The suggested methodology is exemplified using data collected in the City of Leeds (UK).*

Keywords. *Geographical weighted regression; Spatial networks; Statistical Calibration; Traffic flows*

1. Introduction

The estimation of traffic volumes on a street network represents a critical issue to improve transport planning protocols and develop effective road safety interventions. In fact, the traditional ways to produce traffic figures typically involve manual counts with ad-hoc cameras or automatic counts with road-fixed sensors (e.g. inductive loops and spirals). Unfortunately, both techniques have several limitations linked to their limited spatial coverage and high economical costs of installation and maintenance. More recently, traffic information have been collected by geo-referenced mobile sensors (e.g. smartphones and sat-navs) using ad-hoc models. These mobile sensors have several advantages, such as extremely detailed spatial resolution and extensive spatial coverage. However, since not all vehicles are equipped with GPS devices, traffic counts from mobile sensors typically underestimate the real flows.

The precision of the mobile information can be improved by integrating mobile figures using more traditional road-fixed sensors. This allows one to calibrate extensive mobile measurements using more accurate data. We propose a statistical technique to spatially calibrate mobile sensor data using geographical weighted regression (GWR). Being traffic flows a classical example of a phenomenon occurring in a spatial network, the usual GWR was modified to take into account the spatial domain.

2. Spatial calibration by geographical weighted regression

The term *statistical calibration* represent a series of techniques adopted in several research fields to adjust the values of one measurement, say X, using some other measurements, say Y. The need for calibration arises when X is more expensive or more difficult to measure than Y or when the values of X are not recorded and cannot be retrieved [3].

We consider an absolute calibration problem where X is assumed to be measured without error. In particular, X represents the traffic flows measured by the fixed traffic cameras installed on a restricted number of segments of the city network, whereas Y represents the traffic counts derived from mobile sensors installed on cars and available on each street segment. The next two sections will briefly introduce the regression calibration problem and present its spatial re-adaptation in this context.

2.1 Regression Calibration

In a linear regression calibration context, it is assumed that, given a sample of n pairs of observations, the relationship between an imprecise measure Y and a gold-standard or reference value X has the following functional form: $Y_i = f(X_i) = \beta_0 + \beta_1 X_i + \varepsilon_i,\ i = 1,\dots,n$ where ε_i represents the measurement error of the less precise instrument. Two main approaches have been developed when Y is calibrated on X, the so called *classical* and *inverse calibration*.

The classical calibration technique is articulated in two steps. In the first step, Y is regressed on X to obtain the estimated model $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$. Then, the estimated regression model is inverted to obtained predictions of X, i.e. $\hat{X} = (Y - \hat{\beta}_0)\hat{\beta}_1^{-1}$, using known experimental values of Y for each unit where X is not available. On the other hand, in the inverse calibration approach, X is regressed on Y and the values of X are predicted from the estimated model, i.e. $\hat{X} = \hat{\alpha}_0 + \hat{\alpha}_1 Y$, for the units where the gold-standard is not available.

2.2 Spatial calibration via GWR

The analysis of spatial data typically requires ad-hoc adjustments to take into account the nature of the spatial domain. In fact, the relationship between the traffic flows measured by fixed cameras and mobile devices can change according to the spatial location. Hence, a global calibration as described in the previous section is not appropriate and a more local approach would seem preferable.

We thus propose a spatial calibration approach based on geographical weighted regression. GWR is a local form of spatial analysis that allows the estimation of relationships between a dependent variable and a set of predictors that vary over space [2]. More precisely, given a sample of n units in a region S observed at locations $\mathbf{s}_i,\ i = 1,\dots,n$, the GWR model reads as

$$Y(\mathbf{s}_i) = \beta(\mathbf{s}_i)'\mathbf{X}(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i), \quad i = 1,\dots,n, \tag{1}$$

where $Y(\mathbf{s}_i)$ denotes the response variable, $\mathbf{X}(\mathbf{s}_i)$ a column vector of explanatory covariates with a column of constant unitary values that represent the intercept, $\beta(\mathbf{s}_i)$ the corresponding spatially-varying coefficients, and $\varepsilon(\mathbf{s}_i)$ is a zero mean random error.

Parameter estimation at a selected location $\mathbf{s}_j \in S$ is carried out using locally weighted least squares

$$\hat{\beta}(\mathbf{s}_j) = \left[\mathbf{X}'(\mathbf{s})W(\mathbf{s}_j)\mathbf{X}(\mathbf{s})\right]^{-1}\mathbf{X}(\mathbf{s})'W(\mathbf{s}_j)Y(\mathbf{s}_j), \tag{2}$$

where $W(\mathbf{s}_j) = diag(w_{1j},\ldots,w_{nj})$ is a local weighting square matrix with entry w_{ij} giving the weight associated to unit i when the regression is estimated at location $\mathbf{s}_j$, and $\mathbf{X}(\mathbf{s})$ represents the design matrix. The weights are defined in terms of a kernel function K that decays gradually with d_{ij}, i.e. the distance between the ith observation and the point $\mathbf{s}_j$. In particular, a Gaussian kernel function is adopted in this paper: $K(d_{ij}) = \exp\{-d_{ij}^2/2h\}$, where the bandwidth parameter h determines the spatial range of the kernel. In the case study presented in the next section, the value of h is selected using cross-validation by minimising the mean square error of traffic flows predictions.

Re-adapting the calibration equations described before to the GWR framework is actually straightforward. In fact, since the regression coefficients $\hat{\beta}(\mathbf{s}_j)$ depend upon the spatial locations, the GWR permits one to map the variation in the regression parameters and, more importantly, to calibrate the variable of interest taking into account the spatial pattern of the two measures. More precisely, the inclusion of a GWR into a classical calibration approach can be performed as follows. First, we need to estimate a local model that written as

$$\hat{Y}(\mathbf{s}_i) = \hat{\beta}_0(\mathbf{s}_i) + \hat{\beta}_1(\mathbf{s}_i)X(\mathbf{s}_i), \tag{3}$$

and then we calculate the calibrated value of $X(\mathbf{s})$ by inverting the local equation at any desired location $\mathbf{s}$ i.e. $\hat{X}(\mathbf{s}) = (Y(\mathbf{s}) - \hat{\beta}_0(\mathbf{s}))(\hat{\beta}_1(\mathbf{s}))^{-1}$.

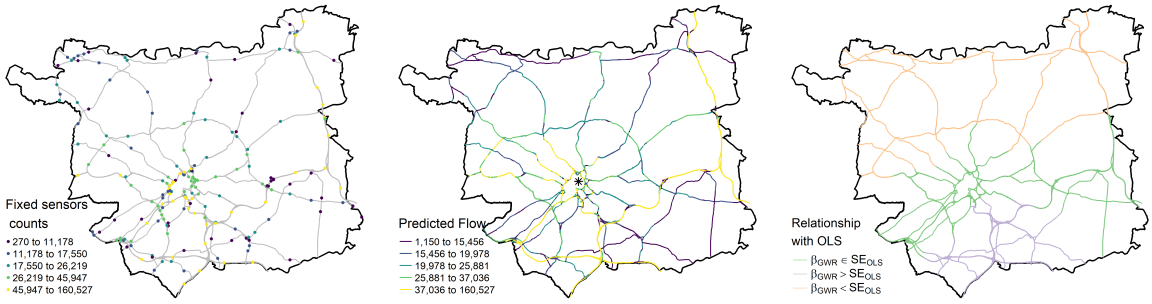
Similarly, the spatial inverse calibration can be performed by estimating a local regression

$$\hat{X}(\mathbf{s}_i) = \hat{\alpha}_0(\mathbf{s}_i) + \hat{\alpha}_1(\mathbf{s}_i)Y(\mathbf{s}_i) \tag{4}$$

which can be used to predict $X(\mathbf{s})$ at any desired location $\mathbf{s}$.

Finally, we note that the usual distance metric adopted in a spatial regression context is the Euclidean distance e.g. $d_{ij} = ||\mathbf{s}_j - \mathbf{s}_i||$. However, the problem that motivated the analysis presented in this paper develops on a one-dimensional linear domain and the sample unit considered below is a road segment represented by its centroid $\mathbf{s}_i$. Therefore, we argue that the distances d_{ij} should be calculated preserving the graph structure of the network. More precisely, indicating by $L = (V,E)$ the one-dimension graph object representing a street network (where V and E denote the sets of vertices and edges, respectively), a path ρ_{ij} connecting any two generic locations $\mathbf{s}_i$ and $\mathbf{s}_j$ on the network is defined as a finite sequence $\{\mathbf{p}_m\}_{m=1}^{M}$ of adjacent vertices in V such that the edges with endpoints $[\mathbf{s}_i,\mathbf{p}_1]$ and $[\mathbf{p}_\mathbf{M},\mathbf{s}_j]$ belong to E. The length of ρ_{ij} can be computed as

$$||\mathbf{s}_i - \mathbf{p}_1|| + \sum_{m=1}^{M-1} ||\mathbf{p}_{m+1} - \mathbf{p}_m|| + ||\mathbf{p}_M - \mathbf{s}_j||,$$



(a) (b) (c)

Figure 1: (a) The segments represent the road network of Leeds, whereas the coloured dots represent the location of the fixed cameras and the corresponding traffic figures; (b) Predicted traffic flows for all segments in the network using an inverse spatial calibration. The black star denotes the city centre; (c) Comparison between classical and spatial calibration in terms of estimated slope coefficient.

and we define d_{ij} as the minimum length of all paths connecting $\mathbf{s}_i$ and $\mathbf{s}_j$ [1].

3. Results and Conclusions

The statistical methods presented in this paper were exemplified considering fixed and mobile traffic data recorded in the street network of Leeds (UK) from January to December 2019. The road network and the mobile traffic counts were obtained from TomTom Move provider (*https://move.tomtom.com/*). In particular, the network is composed by 8959 geo-referenced segments that are associated to traffic volumes estimated using mobile devices connected to cars and anonymous GPS-equipped smartphones. On the other hand, the fixed camera counts were derived using data shared by the Department for Transport (*https://roadtraffic.dft.gov.uk/downloads*). The linear network, the fixed cameras and the corresponding traffic estimates are displayed in Figure 1(a).

Figure 1(b) displays the predicted traffic flows obtained using the inverse spatial calibration technique. The figure clearly highlights several roads corresponding to a motorway (i.e. the yellow segments connecting the south area with the north/north-east) and the most important arterial thoroughfares. Similar results were obtained in the classical spatial calibration framework. Finally, we explored the stationarity of the relationship between the two available measurements. Figure 1(c) compares the slope estimates given by a non-local inverse calibration obtained using OLS regression with the suggested extension. The results highlight a non-stationarity in the relationship between mobile and fixed counts. The Pearson correlation coefficient between observed and calibrated counts (obtained by a leave-one-out approach) were found equal to 0.956 (classical calibration) and 0.965 (inverse calibration).

We plan to extend the analysis presented in this paper in several directions. In fact, there exists a vast literature on classical and inverse statistical calibration problems (see [3]) that can be extended to the problem at hand. In particular, we will focus on a) deriving estimates of the standard errors of the regression coefficients; b) comparing classical and inverse spatial calibrations using a variety of criteria (e.g. MSE or consistency);

c) exploring more sophisticated statistical methods such as multivariate calibration (joining information from different providers), robust GWR and truncated calibration.

References

[1] Baddeley, A., Nair, G., Rakshit, S., McSwiggan, G. and Davies, T.M. (2021). Analysing point patterns on networks - A review. *Spatial Statistics* **42**, p. 100435

[2] Fotheringham, A.S., Brunsdon, C. and Charlton, M. (2003) *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons

[3] Osborne, C., 1991. Statistical calibration: a review. *International Statistical Review/Revue Internationale de Statistique*, pp.309-336.

Random sets on manifolds under an infinite–dimensional Log–Gaussian Cox process approach

M.P. Frías[1,*] , A. Torres[2] and M.D. Ruiz-Medina[3]

[1] *University of Jaén, Jaén, Spain; mpfrias@ujaen.es*
[2] *University of Málaga, Málaga, Spain; atisignes@uma.es*
[3] *University of Granada, Granada, Spain; mruiz@ugr.es*
[*] *Corresponding author*

Abstract. *A new framework is introduced in this paper for modeling and statistical analysis of point sets in a manifold, that randomly arise through time. Specifically, in the characterization of these sets, the random counting measure is assumed to belong to the family of Cox processes driven by a $L^2(\mathcal{M})$–valued Log–Gaussian intensity, where $\mathcal{M}$ denotes here a compact two–point homogeneous space. The associated family of temporal covariance operators on $L^2(\mathcal{M})$ characterizes the n–order product density under stationarity in time. In particular, the pair correlation functional, the reduced second order moment measure or K function can also be constructed from this covariance operator family. Some functional summary statistics of interest are introduced, analyzing their asymptotic properties in the simulation study undertaken.*

Keywords. *Compact two–point homogeneous space; functional summary statistics; Log–Gaussian Cox processes; $\mathcal{M}$–valued Gaussian random fields.*

1. Introduction

In point pattern analysis, different parametric, semiparametric and nonparametric models have been adopted in the estimation of deterministic and random intensities characterizing their counting functions. The nearest neighbor functions, empty space functions, and Ripley's and inhomogeneous K functions arise as classical summary functional statistics in point pattern analysis (see, e.g., [2];[4]). Particularly, a growing interest for point processes in the sphere is observed in recent contributions (see [7]; [8], among others). The goal of the present paper is located in a related more general framework involving point processes driven by log–intensities evaluated in the space $L^2(\mathcal{M})$ of square integrable functions on a compact two–point homogeneous space $\mathcal{M}$. Well–known examples of compact two–point homogeneous spaces are the sphere $\mathbb{S}_d \subset \mathbb{R}^{d+1}$, projective spaces over different algebras (see Section 2 in [5], for more details). Any one of these spaces defines a manifold $\mathbb{M}_d$, where d denotes its topological dimension. We restrict our attention here to random counting functions driven by a temporal Log–Gaussian infinite–dimensional process with values in $L^2(\mathcal{M})$ (see [3] in the Euclidean setting). The n–order joint product density is characterized, in the weak–sense, in terms of test functions lying in the n–fold tensor product $[L^2(\mathcal{M})]^{\otimes_n}$ of the Hilbert space $L^2(\mathcal{M})$. Particularly, we adopt the setting of d–dimensional manifolds $\mathbb{M}_d$ embedded in $\mathbb{R}^{d+1}$. The isometric identification of $(\mathbb{S}_d, d_{\mathbb{S}_d})$ with $(\mathbb{M}_d, d_{\mathbb{M}_d})$ can then be considered via the identity $d_{\mathbb{S}_d}(\mathbf{x}_1, \mathbf{x}_2) = \arccos\left(\mathbf{x}_1^T \mathbf{x}_2\right)$, for $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{S}_d$. In the case of Riemannian manifolds, one can replace the inner product in $\mathbb{R}^{d+1}$ by the family of inner products $\mathcal{G}_{\mathbf{x}}(\cdot,\cdot)$ defined on the tangent

space, smoothly varying over $\mathbf{x} \in \mathbb{M}_d$.

2. Log–Gaussian intensity with values in $L^2(\mathbb{M}_d)$

Let $\{X_t(\cdot),\ t \in \mathcal{T} \subseteq \mathbb{R}\}$ be an infinite–dimensional random process such that, for each $t \in \mathcal{T} \subseteq \mathbb{R}$, almost surely $\log(X_t) \in L^2(\mathbb{M}_d)$, $E[\log(X_t)] \underset{L^2(\mathbb{M}_d)}{=} 0$, with $\log(X_t)$ having characteristic functional

$$\begin{aligned} f_{\log(X_t)}(h) &= \int_{L^2(\mathbb{M}_d)} \exp\left(i\langle h, \log(x_t)\rangle_{L^2(\mathbb{M}_d)}\right) \mu_{\log(X_t)}(d\log(x_t)) \\ &= \exp\left(-\frac{\langle \mathcal{R}_0(h), h\rangle_{L^2(\mathbb{M}_d)}}{2}\right), \quad h \in L^2(\mathbb{M}_d), \end{aligned} \tag{1}$$

where $\mathcal{R}_0 = E\left[\log(X_t) \otimes \log(X_t)\right] \in \mathcal{L}^1(L^2(\mathbb{M}_d))$ denotes the covariance operator of $\log(X_t)$, and $\mathcal{L}^1(L^2(\mathbb{M}_d))$ denotes the space of trace or nuclear operators on $L^2(\mathbb{M}_d)$. Here, $\mu_{\log(X_t)}$ is the induced Gaussian measure by $\log(X_t)$ on $\left(L^2\left(\mathbb{M}_d\right), \mathcal{B}(L^2(\mathbb{M}_d))\right)$, with $\mathcal{B}(L^2(\mathbb{M}_d))$ being the σ–algebra generated by all cylindrical subsets of $L^2(\mathbb{M}_d)$. In the subsequent development, we will also assume that, for any $t, s \in \mathcal{T}$,

$$E\left[\log(X_t)(\mathbf{z})\log(X_s)(\mathbf{y})\right] = r_{t-s}\left(d_{\mathbb{M}_d}(\mathbf{z},\mathbf{y})\right) = \widetilde{r}\left(d_{\mathbb{M}_d}(\mathbf{z},\mathbf{y}), t-s\right), \quad \mathbf{z},\mathbf{y} \in \mathbb{M}_d, \tag{2}$$

i.e., stationarity in time and isotropy over $\mathbb{M}_d$ in the weak sense are assumed. Note that the covariance operator $\mathcal{R}_{t-s}$ with kernel $r_{t-s}(\cdot,\cdot)$ is a nuclear operator, and kernel $\widetilde{r}(d_{\mathbb{M}_d}(\mathbf{z},\mathbf{y}), t-s) = r_{t-s}\left(d_{\mathbb{M}_d}(\mathbf{z},\mathbf{y})\right)$ is assumed to be continuous.

For the special case $r_{t-s}(\cdot,\cdot) = r_{s-t}(\cdot,\cdot)$, the following series expansion is obtained from Theorems 4 and 5 in [5]:

$$\log(X_t)(\mathbf{z}) = \sum_{n=0}^{\infty} V_n(t) P_n^{(\alpha,\beta)}\left(\cos\left(d_{\mathbb{M}_d}(\mathbf{z},\mathbf{U})\right)\right),\ \mathbf{z} \in \mathbb{M}_d,\ t \in \mathbb{R}, \tag{3}$$

where $\{V_n(t),\ n \in \mathbb{N}_0\}$ is a sequence of independent stationary random processes on $\mathcal{T} \subseteq \mathbb{R}$, satisfying $E[V_n(t)] = 0$ and $E[V_n(t_1)V_n(t_2)] = a_n^2 b_n(t_1 - t_2)$, $n \in \mathbb{N}_0$. The random variable $\mathbf{U}$ is uniformly distributed on $\mathbb{M}_d$, and is independent of $\{V_n(t),\ n \in \mathbb{N}_0\}$, and $\sum_{n=0}^{\infty} b_n(0) P_n^{(\alpha,\beta)}(1)$ converges. Also, $\operatorname{cov}\left(V_n(t)P_n^{(\alpha,\beta)}\left(\cos\left(d_{\mathbb{M}_d}(\mathbf{z},\mathbf{U})\right)\right), V_m(t)P_m^{(\alpha,\beta)}\left(\cos\left(d_{\mathbb{M}_d}(\mathbf{z},\mathbf{U})\right)\right)\right) = 0$, for $m \neq n$, and $\mathbf{z} \in \mathbb{M}_d$, and $t \in \mathcal{T}$.

3. Cox processes family

Let now consider the measure $d\nu(\mathbf{z})$ induced on the homogeneous space $\mathbb{M}_d = G/K$, by the probabilistic invariant measure on G, with G being the connected component of the group of isometries of $\mathbb{M}_d$, and K be the stationary subgroup of a fixed point $\mathbf{o} \in \mathbb{M}_d$. As before, $H = L^2\left(\mathbb{M}_d, d\nu(\mathbf{x})\right)$. Consider $\mathbf{Y} = \{\mathbf{Y}_t,\ t \in \mathcal{T} \subseteq \mathbb{R}\}$ to be a family of finite random subsets of $\mathbb{M}_d$, arising at the random times in the interval family $\{[0,t],\ t \in \mathcal{T}\}$.

Let $\{N_t(\cdot),\, t \in \mathcal{T}\}$ be the family of counting measures associated with $\mathbf{Y} = \{\mathbf{Y}_t,\, t \in \mathcal{T} \subseteq \mathbb{R}\}$. For every $t \in \mathcal{T}$, and any Borel set $A \subseteq \mathbb{M}_d$, $N_t(A)$ denotes the number of points in $\mathbf{Y}_t$ falling in a region $A \subseteq \mathbb{M}_d$, at the random times specified by $\mathbf{Y}_t$ in the interval $[0,t]$. The state space for $\mathbf{Y}_t$ is then the set of all possible combinations of finite subsets of $\mathbb{M}_d$ with finite time sets of $[0,t]$, equipped with the σ–algebra $\mathcal{F}$ generated by the events $\{N_t(A) = n\}$, indicating that n points in $\mathbf{Y}_t$ falling in a region $A \subseteq \mathbb{M}_d$, at some specific times in $[0,t]$, for any Borel set $A \subseteq \mathbb{M}_d$, interval $[0,t]$, and integer $n \in \mathbb{N}$. Assume that, for each $t \in \mathcal{T}$, given a realization $\{x_t(\mathbf{z}),\ \mathbf{z} \in \mathbb{M}_d\}$, of X_t, satisfying (1)–(3), the conditional distribution of $N_t(A)/\{x_t(\mathbf{z}),\ \mathbf{z} \in \mathbb{M}_d\}$ is a Poisson distribution with parameter $\lambda_t = \int_0^t \int_A x_s(\mathbf{z}) d\nu(\mathbf{z}) ds$. The n–order product density $\rho^{(n)}_{t_1,\ldots,t_n}(\mathbf{z}_1,\ldots,\mathbf{z}_n)$ is such that $\rho^{(n)}_{t_1,\ldots,t_n}(\mathbf{z}_1,\ldots,\mathbf{z}_n) d\nu^{(n)}(\mathbf{z}_1,\ldots,\mathbf{z}_n) dt_1,\ldots,dt_n$ indicates the probability that $\mathbf{Y}_t$ has a point in each of n infinitesimally small regions on $\mathbb{M}_d$ around $\mathbf{z}_1,\ldots,\mathbf{z}_n$, of surface measure $d\nu(\mathbf{z}_1)\cdots d\nu(\mathbf{z}_n)$, over the infinitesimal time intervals around $t_1,\ldots,t_n$, of length $dt_1,\ldots,dt_n$. From equation (3), for any $t_1,\ldots,t_n \in \mathbb{R}$, one can compute $\rho^{(n)}_{t_1,\ldots,t_n}$ as follows:

$$\rho^{(n)}_{t_1,\ldots,t_n}(\mathbf{z}_1,\ldots,\mathbf{z}_n) = E\left[\prod_{i=1}^{n} \exp\left(X_{t_i}(\mathbf{z}_i)\right)\right] = E\left[\exp\left(\sum_{i=1}^{n} X_{t_i}(\mathbf{z}_i)\right)\right]$$

$$= [\rho]^n \exp\left(\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=0}^{\infty} b_k(t_i - t_j) P_k^{(\alpha,\beta)}\left(\cos\left(d_{\mathbb{M}_d}(\mathbf{z}_i,\mathbf{z}_j)\right)\right)\right),\ \forall z_i \in \mathbb{M}_d, i = 1,\ldots,n.$$

In particular, for any $t \in \mathcal{T}$, and, for any $t_1, t_2 \in \mathcal{T}$, the intensity function $\rho_t = \rho_0 = \rho^{(1)}(t)$, and the pair correlation function $g_{t_1-t_2}\left(\cos\left(d_{\mathbb{M}_d}(\mathbf{z}_1,\mathbf{z}_2)\right)\right)$, $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{M}_d$, respectively admit the following expressions:

$$\begin{aligned} \rho = \rho_0(\mathbf{z}) &= \exp\left(\frac{1}{2}\sum_{n=0}^{\infty} b_n(0) P_n^{(\alpha,\beta)}(1)\right), \quad \forall \mathbf{z} \in \mathbb{M}_d, \\ g_{t_1-t_2}\left(\cos\left(d_{\mathbb{M}_d}(\mathbf{z}_1,\mathbf{z}_2)\right)\right) &= \frac{\rho^{(2)}_{t_1-t_2}\left(\cos\left(d_{\mathbb{M}_d}(\mathbf{z}_1,\mathbf{z}_2)\right)\right)}{\rho^2} = \exp\left(\sum_{n=0}^{\infty} b_n(t_1-t_2) P_n^{(\alpha,\beta)}\left(\cos\left(d_{\mathbb{M}_d}(\mathbf{z}_1,\mathbf{z}_2)\right)\right)\right). \end{aligned}$$

4. Functional summary statistics and simulation

A simulation study is undertaken for an asymptotic analysis of the usual functional summary statistics. We will consider the empirical counterparts of the nearest neighborhood function, given by $\widehat{G}_t(s) = \frac{1}{N_t(\mathbb{M}_d)} \sum_{(u,\mathbf{y}) \in \mathbf{Y}_t} 1_{\{\inf_{(v,\mathbf{z}) \in \mathbf{Y}_t \setminus \{(u,\mathbf{y})\}} d_{\mathbb{M}_d}(\mathbf{z},\mathbf{y}) | u-v | \leq s\}}$, and of the empty space function defined as $\widehat{F}_t(s) = \frac{1}{m(t)} \sum_{(u,\mathbf{x}) \in \mathcal{Q}_t} 1_{\{\inf_{(v,\mathbf{z}) \in Y_t} d_{\mathbb{M}_d}(\mathbf{z},\mathbf{x}) | u-v | \leq s\}}$, for $s \in [0,\pi]$, and $t \in \mathcal{T}$, with $\mathcal{Q}_t$ being a finite grid on $\mathbb{M}_d \times [0,t]$, of $m(t) > 0$ points. For the K_t function, its empirical version $\widehat{K}_t(s) = \frac{1}{\nu(\mathbb{M}_d)\widehat{\rho}^2} \sum_{(u,\mathbf{x}) \neq (v,\mathbf{y}) \in \mathbf{Y}_t} 1_{\{d_{\mathbb{M}_d}(\mathbf{x},\mathbf{y}) | u-v | \leq s\}}$ will also be asymptotically analyzed, with $\widehat{\rho}$ being an unbiased estimator of ρ, and assuming $\mathbf{Y}_t$ is fully observed on $\mathbb{M}_d$, provided $N_t(\mathbb{M}_d) > 0$, for any $t \in \mathcal{T}$. This asymptotic analysis is achieved from simulations, under the particular model $r_{t-s}(\langle \mathbf{x},\mathbf{y} \rangle) = \sum_{l=0}^{\infty} C_l(t-s) \frac{2l+1}{4\pi} P_l(\langle \mathbf{x},\mathbf{y} \rangle)$, $t,s \in \mathcal{T}$, introduced in [1];[6] on the sphere $\mathbb{S}_2$, in terms of Legendre polynomials. Figure 1 displays the short–memory case at the first four rows, for $\tau = t - s$,

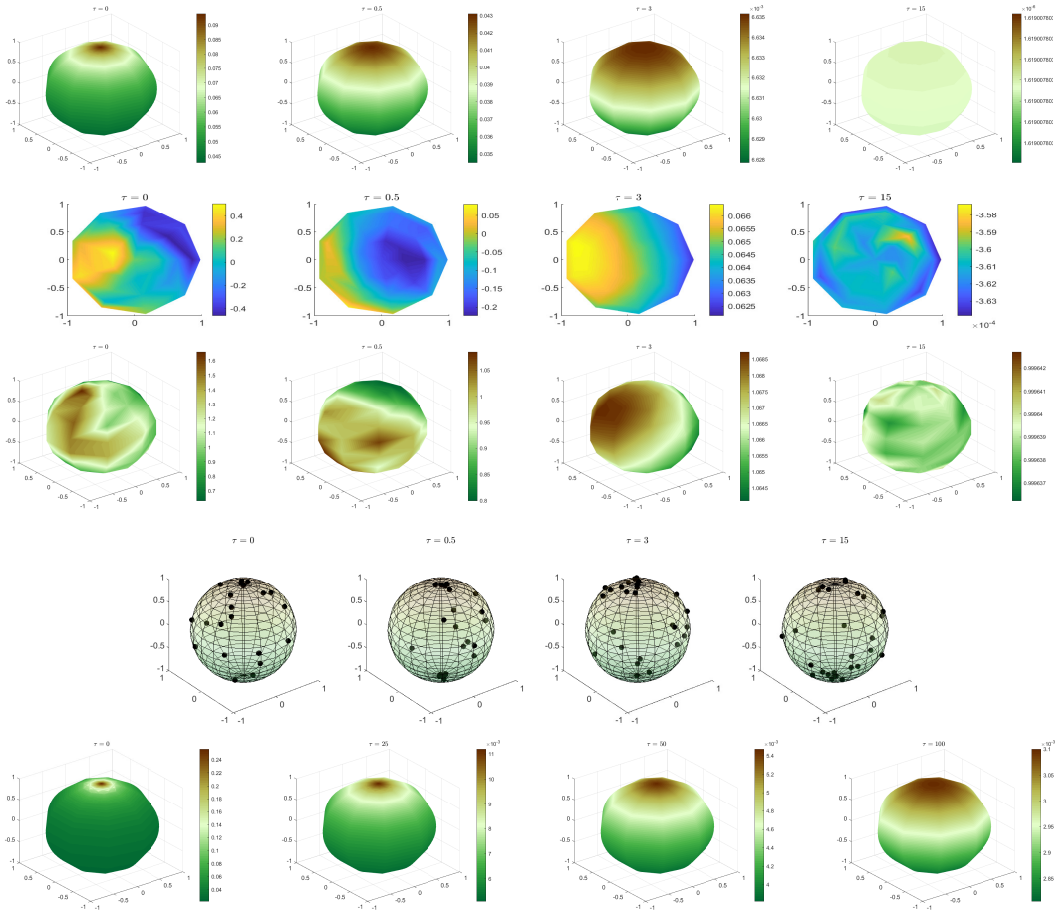


Figure 1: At the top, *short–memory* covariance operator model is displayed for different temporal lags, simulated realizations of the log-intensity and the intensity random processes are given at the second and third rows, respectively. At the fourth row, point patterns generated from the corresponding log–Gaussian Cox process are drawn for $\tau = t - s = 0, 0.5, 3, 15$. Finally, *long–memory* covariance operator model for temporal lags $\tau = t - s = 0, 25, 50, 100$ is displayed at the bottom row.

$C_l(\tau) = \frac{\phi_l^{|\tau|} C_{l;Z}}{1-\phi_l^2}$ with $\phi_l = G(l+1)^{-\alpha_\phi}$, $\alpha_\phi = 3$, $C_{l;Z} = G_Z(1+l)^{-\alpha_Z}$, $\alpha_Z = 3$, $G = G_Z = 0.5$. At the bottom of Figure 1, the long–memory case is showed, for $C_l(\tau) = G_l(\tau) g_l(\tau)$, $G_l(\tau) = G(l+1)^{-2-k_\tau \tau}$, $g_l(\tau) = (1+|l|)^{-\beta_l}$, $\beta_l = \frac{k_\beta (l+1)}{\sqrt{(l+1)^2+1}}$, $k_\beta = 0.8$, $G = 0.5$ and $k_\tau = 0.03$. In both cases Legendre series is truncated at $L_{max} = 30$ considering a spherical regular grid.

Acknowledgments

This work has been supported in part by projects MCIN/ AEI/PGC2018-099549-B-I00, and the Economy and Knowledge Council of the Regional Government of Andalusia, Spain (A-FQM-345-UGR18, and CEX2020-001105-M MCIN/ AEI/10.13039/501100011033).

References

[1] Caponera, A. and Marinucci,D. (2021). Asymptotics for spherical functional autoregressions. *The Annals of Statistics* **49**, 346 – 369.

[2] Diggle, P. J. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. Taylor & Francis. Boca Raton.

[3] Frías, M. P., Torres–Signes, A., Ruiz–Medina, M. D. and Mateu, J. (2022). Spatial Cox processes in an infinite–dimensional framework. *Test*. **31**, 175–203.

[4] Illian, J., Penttinen, A., Stoyan, H. and Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. John Wiley & Sons. New York.

[5] Ma, C. and Malyarenko, A. (2020). Time varying isotropic vector random fields on compact two points homogeneous spaces. *Journal of Theoretical Probability* **33**, 319–339.

[6] Marinucci, D., Rossi, M. and Vidotto, A. (2020). Non-universal fluctuations of the empirical measure for isotropic stationary fields on $\mathbb{S}^2 \times \mathbb{R}$. *Annals of Applied Probability* **31**, 2311–2349.

[7] Møller, J. and Rubak, E. (2016). Functional summary statistics for point processes on the sphere with an application to determinantal point processes. arXiv: 1601.03448v2.

[8] Robeson, S. M., Li, A. and Huang, C. (2014). *Point-pattern analysis on the sphere. Spatial Statistics* **10**, 76–86.

Spatiotemporal point processes with moderate and extreme marks: application to wildfires

T. Opitz[1,*] , J. Koh[2] , F. Pimont[3] and J.-L. Dupuy[3]

[1]*BioSP, INRAE, Avignon, France; thomas.opitz@inrae.fr*
[2]*University of Bern, Switzerland; jonathan.koh@stat.unibe.ch*
[3]*URFM, INRAE, Avignon, France; francois.pimont@inrae.fr,jean-luc.dupuy@inrae.fr* *Corresponding author*

Abstract. *Accurate spatiotemporal stochastic modeling of conditions leading to moderate and large wildfires provides better understanding of mechanisms driving fire-prone ecosystems and improves risk management. Typically, the distribution of burnt areas is very heavy-tailed, such that the few largest wildfires have dominant contribution to the aggregated burnt area of all wildfires in a region. We here propose a novel joint model, called* Firelihood*, for the occurrence intensity and the wildfire size distribution that combines extreme-value theory and point processes within a Bayesian hierarchical regression framework. The model is used to study daily summer wildfire data for the French Mediterranean basin during the 1995–2018 period. The occurrence component models wildfire ignitions as a spatiotemporal log-Gaussian Cox process. Burnt areas are numerical marks attached to points and are considered as extreme if they exceed a high threshold. Therefore, the model's size component is a two-component mixture varying in space and time that jointly models moderate and extreme fires. We capture potentially non-linear influence of covariates (Fire Weather Index for weather drivers, forest cover for exposure to wildfire risk) through component-specific smooth functions, which may further vary with season. We also propose estimating shared random effects between model components to reveal and interpret common drivers of different aspects of wildfire activity. For instance, this mechanism could incorporate behavior where larger occurrence numbers are associated with either larger or smaller wildfires, depending on the subregion. This sharing of random effects leads to increased parsimony and reduced estimation uncertainty with better predictions. We fit various models using the integrated nested Laplace approximation, and we compare and validate them through predictive scores and visual diagnostics. Our methodology provides a holistic approach to explaining and predicting the drivers of wildfire activity and associated uncertainties.*

Keywords. *Bayesian hierarchical model; Cox process; Extreme-value theory; Forest fires; Common-component model*

1. Wildfires: a challenging global problem

Wildfires represent major environmental and ecological risks worldwide. They provoke many human casualties and substantial economic costs, and can trigger extreme air pollution episodes and important losses of biomass and biodiversity. While climate change is expected to exacerbate their frequency and extent (see [1]), wildfires themselves contribute an important fraction of global greenhouse gases that can accelerate climate change. To aid in wildfire prevention and risk mitigation, one must identify the factors contributing to wildfires and predict their spatiotemporal distribution. Prediction maps of various components of wildfire risk

are relevant for the study of historical periods, for short-term forecasting and for long-term projections. The joint modeling of wildfire occurrences and sizes is highly challenging since wildfire activity depends on many factors and their complex interactions, such as weather, season, vegetation type and socioeconomic variables. We here discuss modeling extensions in the framework of *Firelihood* introduced in [3], a Bayesian hierarchical model for spatiotemporal modeling, prediction and long-term projection of wildfire activity.

2. Available data for Southeastern France

The French *Prométhé* database, filled since the 1970s, provides data of wildfire occurrences in Southeastern France. For each wildfire, it contains the date of ignition, the location of wildfire ignition with 2km precision, and the burnt area, among other information. We here focus on the period 1995–2018, for which Figure 1 shows a map of wildfire occurrences, including contour lines of a kernel-based intensity estimate of the spatial point pattern, highlighting strong spatial heterogeneity. For this study period, we also have weather reanalysis data from the SAFRAN model of Météo rance at daily 8km pixel resolution for France. Weather data have been preprocessed to obtain the Fire Weather Index, a scalar metric of weather-induced fire danger, at day-pixel level. Another important covariate is the forested area (FA) per pixel and year, obtained from the Corine Land Cover database, which provides a metric for the exposure to wildfire risk 8km pixel for each year.

3. Modeling approach

Wildfire size distributions are known to be heavy-tailed, and for our dataset we find that approximately 1% of the larges fires contribute around 99% of the aggregated burnt area. Therefore, appropriate modeling of large wildfires is crucial for making reliable predictions of aggregated burnt areas, which are a key indicator for sanitary, ecological and economic damages of wildfires. In this work, we propose using split modeling where the distribution of wildfire size is modeled as a mixture of two components: a generalized Pareto distribution (GPD) for sizes exceeding a fixed high threshold $u > 0$, as suggested by extreme-value theory, and an appropriately rescaled Beta distribution in $[0, u]$ to model the size of small and moderate fires. The mixing probability is determined by the probability of fires to exceed the threshold u. Here, exploratory analyses have led us to fix $u = 79$ha.

For explanatory and predictive modeling, Bayesian hierarchical models are useful; they can include latent Gaussian components to allow for observation and estimation uncertainty and to capture nonlinear influences of covariates. In this framework, we construct four spatiotemporal regression equations at day-pixel resolution using Gaussian random effects, where the response distributions are of the following type: Poisson for day-pixel wildfire counts, GPD for positive excesses of wildfire sizes above the threshold u, Bernoulli for the binary indicator of a wildfire exceeding the threshold, and rescaled Beta for wildfire sizes below the threshold u.

Conceptually, our model defines a marked log-Gaussian Cox process with a mixture distribution for the marks. We here use the integrated nested Laplace approximation for Bayesian inference, and we combine it with the Stochastic Partial Differential Equation (SPDE) approach to leverage Gauss–Markov representations of Gaussian fields with Matérn covariance for flexible inference with many observation locations; see [2] for

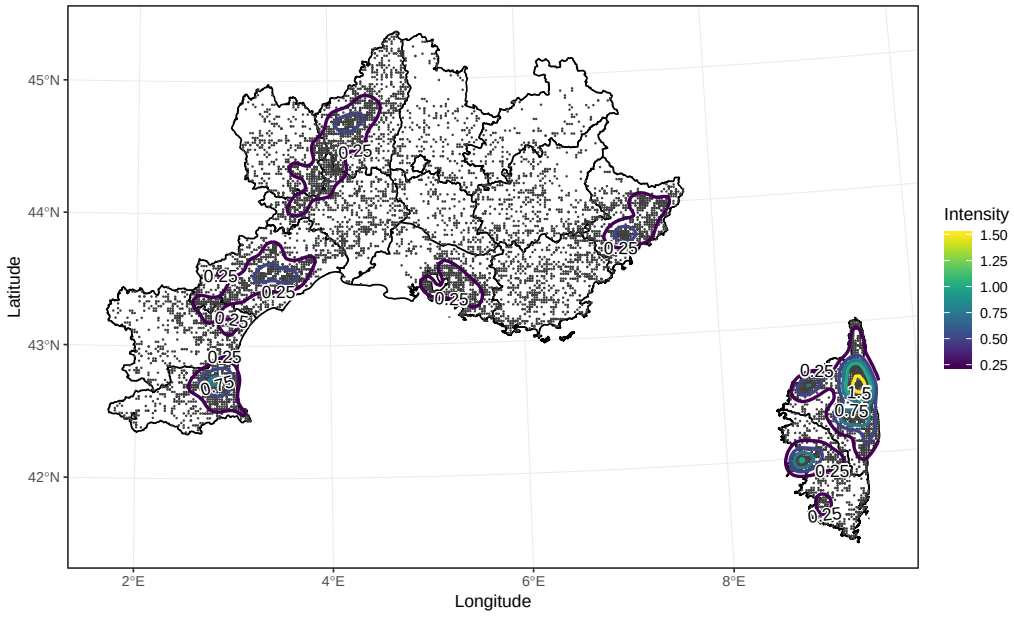


Figure 1: Spatial point pattern of wildfire data in Southern France. Each point refers to a single or multiple occurrence of wildfires during the period 1995–2018. Overlaid contour lines have been obtained through a kernel estimate of the point process intensity function with spatial unit of 1 km^2.

details. The SPDE approach is used to define the Gaussian priors of spatial effects and of nonlinear effects of the two covariates FWI and FA, where the FWI contribution is further allowed to vary across months.

We further study the benefits of sharing spatial random effects across several of the four components, where a spatial random effect $W(s)$ included in one component can also be shared towards another component as $\beta W(s)$ with an additional scaling factor β to be estimated. For instance, having positive β generates positive correlation between the two components. This sharing mechanism can reduce estimation uncertainties and can allow for useful interpretations of the interactions among the four model components.

4. Results and discussion

Our model diagnostics show that using sophisticated techniques such as split modeling of moderate and extreme marks or sharing of random effects provides improvements over alternative modeling approaches and leads to a realistic stochastic representation of spatiotemporal wildfire activity.

Our findings improve decision support in wildfire management. Spatial and temporal random effects quantify the spatiotemporal variation in wildfire activity not explained by the available explanatory variables, i.e., FWI and Forest Area. Our shared spatial effects explain how residual spatial variability is correlated across wildfire numbers and extreme sizes, and allow us to provide maps of the significant disparities between regions. Weather forecasts, and the derived FWI forecasts typically delivered at regional levels, are currently the main components guiding fire detection and suppression resources as well as the temporary shutdown of forest areas to the public. However, FWI maps used for fire danger rating must be interpreted with care because of the strongly nonlinear and seasonal FWI effect on wildfire risk highlighted by our model. Moreover, strong residual spatial effects estimated in our model could also hint at weather effects not captured by FWI. The precise regional forecasting of fire activity that our model can provide, especially of the expected number of fires and of the expected number and sizes of extreme fires, equips wildfire managers with additional objective criteria to aid decision-making.

References

[1] Jones, M.W., Smith, A., Betts, R., Canadell, J.G., Prentice, I.C., Le Quéré, C. (2020). ScienceBrief Review: Climate change increases the risk of wildfires. *Critical Issues in Climate Change Science*.

[2] Krainski, E.T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., Rue, Håvard, (2018). *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. Chapman and Hall/CRC.

[3] Pimont, F., Fargeon, H., Opitz, T., Ruffault, J., Barbero, R., Martin-StPaul, N., Rigolot, E., Rivière, M., Dupuy, J.-L. (2021). Prediction of regional wildfire activity in the probabilistic Bayesian framework of Firelihood. *Ecological Applications*, 31(5), e02316.

Implementing a class of non-stationary non-separable spacetime models

E.T. Krainski[1] , F. Lindgren[2] and H. Rue[1]

[1]*KAUST, Thuwal - Saudi Arabia; Elias.Krainski@kaust.edu.sa, Haavard.Rue@kaust.edu.sa*
[2]*The University of Edinburgh, Scotland; Finn.Lindgren@ed.ac.uk*
[*]*Corresponding author*

Abstract. *When modeling data collected in different locations (space) and at different times one may specify a model that contains a stochastic model accounting for the correlation over the spacetime domain. Although there are some theorectical properties to be accounted for, it is also of great practical importance to consider computational aspects. In this talk we will highligh some important details of a flexible class of spacetime models. We can achieve computing time similar to somewhat much simpler models. An illustration is shown for a real dataset on particulate matter concentration.*

Keywords. *spacetime; stochastic partial differential equation; non-separable; non-stationary.*

1. A framework: the SPDE approach

There are a number of approaches to be considered when specifying a stochastic model. One famous model was specified considering a (discrete) lattice system in [16], and in Eq. (54) of this work the correlation function was given as a function of (continous) distance. This work was further extended in [9], [17] and [7]. The latter stablished an (explicit) link between models over discrete basis representations and continuous domain fields. The framework introduced in [7], so called Stochastic Partial Differential Equations (SPDE) approach, can be schematically represented in Figure 1.

The SPDE has parameters that directly relate to the local properties of the model, giving interpretation related to the dynamics of the defined process, as well as spectral properties for stationary models. These parameters are translated into the parameters of the conditional distributions implied by a Gaussian Markov Random Field (GMRF) representation. The parameters in GMRFs are common in statistics, starting with autoregressions in time series (spectral methods are very common in time serie methods). It is usual to consider the marginal (covariance) property parameters, that can be derived from the former ones, even for intrinsic models, see [11].

A key point in the SPDE approach is that it does not require covariances to be computed. This is appealing when working with spatio-temporal models. The commonly considered problem is the complexity of numerical factorizations for the working matrices. Even evaluating covariances can be already challenging when dealing with good spacetime models as shown in [12] and [13]. In these two works, the covariance was derived from the spectral density, as in Figure 1, but only for integer smoothness parameters.

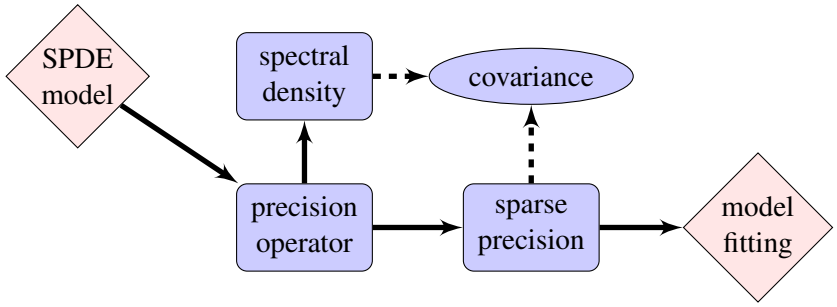


Figure 1: Work flow with the SPDE model formulation.

The SPDE framework was further developed still enabling computations with sparse precision matrices, [6]. This alleviates the double hard work of computing with covariances: no need of evaluating covariances and no need of solving dense matrices. When needed, an efficient algorithm for computing covariances of selected elements of the precision matrix exists and is available in the **INLA** package. In addition, the computation of conditional expectations, or kriging, can take advantage of this formulation.

In [13], three desired properties for spacetime models were stated. The third one is regarding to computation: covariance "computed accurately and efficiently". We can improve on this by having the whole fitting approach to be computationally efficient, without the need to explicitly evaluate covariances.

2. An SPDE based spatio-temporal model class

Consider the operator $L_s = \gamma_s^2 - \Delta$ on a spatial domain $\mathcal{D}$ and introduce the precision operator for the Matérn covariance as $Q(\gamma_s, \gamma_e, \alpha) = \gamma_e^2 L_s^\alpha$, corresponding to the stationary solutions $v(\boldsymbol{s})$ to

$$\gamma_e L_s^{\alpha/2} v(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \qquad \boldsymbol{s} \in \mathcal{D} \tag{1}$$

where $\mathcal{W}$ is a spatial white noise process, as in [17] and [7].

We replace the damping coefficient κ with a fractional dampened diffusion operator L_s defining a model family of time-stationary solutions to a iterated diffusion-like processes

$$\left(-\gamma_t^2 \frac{d^2}{dt^2} + L_s^{\alpha_s}\right)^{\alpha_t/2} u(\boldsymbol{s},t) = \mathrm{d}\mathbb{E}[Q](\boldsymbol{s},t), \qquad (\boldsymbol{s},t) \in \mathcal{D} \times \mathbb{R}. \tag{2}$$

When $\mathcal{D} = \mathbb{R}^d$, the space-stationary solutions are used. For compact manifolds with boundary, the operators L_s and Q are equipped with suitable boundary conditions on $\partial\mathcal{D}$. In total, the model has three non-negative smoothness parameters $(\alpha_t, \alpha_s, \alpha_e)$ and three positive scale parameters $(\gamma_t, \gamma_s, \gamma_e)$.

The smoothness parameters can be chosen so that we can have models such as separable, fully nonseparable, diffusion and iterated diffusion. The scale parameters can be written as function of intepretable parameters:

temporal range, spatial range and marginal variance. For more details see [1] and [15].

3. Practical implementation and example

In order to work with this model in practice, a discretization has to be done. The considered discretization combines discretizations over space and time following [7]. The resulting precision matrix is a linear combination of kronecker product between finite element matrices from temporal and spatial discretizations. [6] discusses the basis function options, showing that the SPDE models do not require the triangle-FEM approach. For example, one can mix temporal B-splines with spherical harmonics, which can be used for smooth global models.

An appealing property of separable models is that the factorization of the spatio-temporal precision matrix simplifies into the factorization of a purelly temporal matrix and a purely spatial one. However, when considering a general hierarchical model formulation as in **INLA**, the factorization is done for the joint problem precision matrix, which already takes the form of a *sum* of sparse matrix products, and direct sparse linear solversare used. Therefore, because the derived precision matrices for the nonseparable case has similar sparsity order as the separable one, the computational complexity is also similar.

The particular case of $\alpha_t = 1$, $\alpha_s = 2$ and $\alpha_e = 1$ was considered in [5]. The implementation illustrated there was a particular way for building the precion matrix to be considered in the **INLA** package. This package was further developed including a new way for building the precision matrix and a series of methodological improvements that further improved the computations, see [14].

We considered the data in [3], where a separable spacetime model was present in the fitted model. The authors of that paper reported a computation time of 240 seconds. With the current **INLA** the same analysis took a bit over a minute. This factor of four is mainly due to all the improvements cited in [14].

The recent development of the `cgeneric` interface in **INLA** has enabled a flexible and efficient way for building general precision matrices. This also takes full advantage of performing the matrix factorizations in parallel, which is particularly appealing for huge spacetime models. We fitted the nonseparable model in a bit more than two minutes. The fitted model, as in [3], consider a triangulated domain with *142* nodes as well and the *182* time points.

4. Conclusion and further developments

The basis of working with models for spacetime domains through the SPDE approach is set in the literature, see [6] and [1]. The adequate numerical representations of the model make possible the use of efficient algorithms within a general modeling framework as implemented in the **INLA**, [10]. This combination provides the capability for chosing between separable and nonseparable models, without restricting this choice due to computational limitations.

The extension for the non-stationary case is a follow up of the spatial case as proposed in [4]. Under this

approach, the parameters can be specified considering covariates or it can be considered basis functions to provide smoothing over desired support. It applies either to the parameters in the SPDE or the derived ones related to the marginal properties, as detailed in [8].

The theoretical work done under the SPDE framework is not restricted to Euclidean domains. One can specify the model on a sphere, useful for modeling global data. Furthermore, physical barriers can also be considered as in [2]. This implementation is a work in progress, and we have started a new R package dedicated to this work. The aforementioned extensions for non-stationary models are under develompment.

References

[1] Bakka, H., Krainski, E. T., Bolin, D., Rue, H. and Lindgren. F. (2020) The diffusion-based extension of the matrn field to space-time. arXiv. https://arxiv.org/abs/2006.04917.

[2] Bakka, H., Vanhatalo, J., Illian, J.B., Simpson, D. and Rue. H. (2019) Non-stationary Gaussian models with physical barriers. *Spatial Statistics*, **29**:268 288.

[3] Cameletti, M., Lindgren, F., Simpson, D. and Rue. H. (2013) Spatio-temporal modeling of particulate matter concentration through the spde approach. *AStA Advances in Statistical Analysis*, **97**(2):109131,

[4] Ingebrigtsen, R., Lindgren, F. and Steinsland, I. (2014) Spatial models with explanatory variables in the dependence structure. *Spatial Statistics*, **8**:2038.

[5] Krainski, E.T. (2018). Statistical Analysis of Space-time Data: New Models and Applications. PhD thesis, Norwegian University of Science and Technology. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2500188.

[6] Lindgren, F., Bolin, D. and Rue, H. (2022) The SPDE approach for Gaussian and non-Gaussian fields: 10 years and still running. *Spatial Statistics* (accepted).

[7] 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):423498.

[8] Lindgren, F. Rue, H. et al. (2015) Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, **63**(19):125.

[9] Matérn, B. (1960) Spatial variation-stochastic models and their application to some problems in forest surveys and other sampling investigations. Meddelanden fran statens skogsforskningsintitut, almaenna foerlaget, stock- holm. (1986), 49 (5).

[10] Rue, H., Martino, S. and Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**(2):319392.

[11] Sørbye, S.H. and Rue. H. (2014) Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, **8**:3951.

[12] Stein, M.L. (2005) Spacetime covariance functions. *Journal of the American Statistical Association*,**100**(469):310321.

[13] Stein, M.L. (2013). On a class of space-time intrinsic random functions. *Bernoulli*, **19**(2):287408, 2013.

[14] van Niekerk, J., Krainski, E.T. Rustand, D. Rue, H. (2022). A new avenue for Bayesian inference with INLA. submitted. https://arxiv.org/abs/2204.06797.

[15] Carrizo-Vergara, R., Allard, D. and Desassis, N. (2022) A general framework for spde-based stationary random fields. *Bernoulli*, **28**.

[16] Whittle, P. (1954) On stationary processes in the plane. Biometrika, 41(3/4):pp. 434449.

[17] Whittle, P. (1963) Stochastic-processes in several dimensions. *Bulletin of the International Statistical Institute,* **40**(2):974994.

Detecting climate change in daily temperatures with a space-time quantile autoregressive model

J. Castillo-Mateo[1,*] , A.E. Gelfand[2] , J. Asín[1] and A.C. Cebrián[1]

[1]*Department of Statistical Methods, University of Zaragoza, Zaragoza, Spain; jorgecm@unizar.es, jasin@unizar.es, acebrian@unizar.es*
[2]*Department of Statistical Science, Duke University, Durham NC, USA; alan@duke.edu*
[*]*Corresponding author*

Abstract. *We propose a flexible mixed effects autoregressive model in time using four spatial processes to detect space-time quantile changes in a point-referenced collection of daily maximum temperature series with a wide climatic variability in the Ebro Basin (Spain). We consider regression through asymmetric Laplace (AL) errors. Using the AL specification, we propose a method to extract marginal quantiles from the conditional quantiles in the autoregression.*

Keywords. *Asymmetric Laplace; Bayesian hierarchical model; climate change; daily temperature; quantile autoregression.*

1. Introduction

Climate change can lead to changes in various aspects of the distribution of climatic variables, in particular, for the series of daily maximum temperatures (Tmax), a different evolution could occur in the central and extreme quantiles. Quantile regression (QR) makes it possible to detect this type of heterogeneous pattern in evolution. The modeling approach for QR called *multiple* QR, follows the original ideas by Koenker and Bassett [1], offers a separate regression model for each of the quantiles of interest, and inference proceeds by minimizing a check loss function or assuming asymmetric Laplace (AL) errors. An alternative approach, called *joint* QR, specifies an appropriate joint model for all quantiles [2]. Frequently, spatial QR models assume that the data have no temporal dependence [3]. However, to study Tmax in daily scale, the statistical modeling must include components that represents the strong serial correlation. Reich [4] works with spatial time series but its model is not autoregressive in time.

We develop a flexible spatio-temporal model for QR that specifies temporal dependence through autoregression and that introduces spatial dependence through Gaussian processes (GPs). Additionally, an approach to obtain marginal quantiles from the conditional quantiles is proposed.

The quantiles of Tmax in the Ebro Basin are analyzed. This Spanish region is a challenge because it includes mountains and aride subregions with a wide variety of climate conditions in a relatively small area. An exploratory analysis shows that the elevation is a factor, but not the only one, having an influence on the mean and standard deviation of the Tmax distribution, and that trend and serial correlation are spatially varying.

These characteristics motivate the spatial random effects introduced in the model.

2. Methodology

2.1 The space-time model

We propose a spatio-temporal quantile autoregression (QAR) model for Tmax, where each quantile is modeled separately. Denote by $Y_{t\ell}(\mathbf{s})$ the daily maximum temperature for day ℓ ($\ell = 2,\ldots,L$) of year t ($t = 1,\ldots,T$) at location $\mathbf{s}$ ($\mathbf{s} \in \mathcal{D}$ our study region), $\tau \in (0,1)$ the quantile order and $Q_{Y_{t\ell}(\mathbf{s})}(\tau \mid Y_{t,\ell-1}(\mathbf{s}))$ the τ conditional quantile of $Y_{t\ell}(\mathbf{s})$ given $Y_{t,\ell-1}(\mathbf{s})$. The model expresses separately fixed and random effects in $q^{\tau}_{t\ell}(\mathbf{s})$ and the autoregressive term,

$$Y_{t\ell}(\mathbf{s}) = Q_{Y_{t\ell}(\mathbf{s})}(\tau \mid Y_{t,\ell-1}(\mathbf{s})) + \varepsilon^{\tau}_{t\ell}(\mathbf{s}) = q^{\tau}_{t\ell}(\mathbf{s}) + \rho^{\tau}(\mathbf{s})\left(Y_{t,\ell-1}(\mathbf{s}) - q^{\tau}_{t,\ell-1}(\mathbf{s})\right) + \varepsilon^{\tau}_{t\ell}(\mathbf{s}). \tag{1}$$

In particular,

$$\begin{aligned} q^{\tau}_{t\ell}(\mathbf{s}) &= \beta^{\tau}_0 + \alpha^{\tau} t + \beta^{\tau}_1 \sin(2\pi\ell/365) + \beta^{\tau}_2 \cos(2\pi\ell/365) + \beta^{\tau}_3 \,\mathrm{elev}(\mathbf{s}) + \gamma^{\tau}_t(\mathbf{s}), \\ \gamma^{\tau}_t(\mathbf{s}) &= \beta^{\tau}_0(\mathbf{s}) + \alpha^{\tau}(\mathbf{s}) t + \psi^{\tau}_t + \eta^{\tau}_t(\mathbf{s}). \end{aligned}$$

The *fixed effects* are given by β^{τ}_0, a global intercept, $\alpha^{\tau} t$, a global long-term linear trend, sin and cos terms that capture the annual seasonal component, and $\mathrm{elev}(\mathbf{s})$, the elevation at $\mathbf{s}$. The *random effects* given by $\gamma^{\tau}_t(\mathbf{s})$ capture space-time dependence through GPs. In particular, $\beta^{\tau}_0(\mathbf{s}) \sim GP(0, C(\cdot;\sigma^{2,\tau}_{\beta_0}, \phi^{\tau}_{\beta_0}))$ and $\alpha^{\tau}(\mathbf{s}) \sim GP(0, C(\cdot;\sigma^{2,\tau}_{\alpha}, \phi^{\tau}_{\alpha}))$ provide local adjustments to the intercept and the long-term linear trend, where $C(\cdot;\sigma^2,\phi)$ is the exponential covariance function. In addition, $\psi^{\tau}_t \sim$ i.i.d. $N(0, \sigma^{2,\tau}_{\psi})$ provides annual intercepts and $\eta^{\tau}_t(\mathbf{s}) \sim$ i.i.d. $N(0, \sigma^{2,\tau}_{\eta})$ provides local annual intercepts. We also specify $\rho^{\tau}(\mathbf{s})$ spatially varying to capture spatial autoregession dependence through $Z^{\tau}_{\rho}(\mathbf{s}) = \log\{(1+\rho^{\tau}(\mathbf{s}))/(1-\rho^{\tau}(\mathbf{s}))\} \sim GP(Z^{\tau}_{\rho}, C(\cdot;\sigma^{2,\tau}_{\rho}, \phi^{\tau}_{\rho}))$.

The error term is $\varepsilon^{\tau}_{t\ell}(\mathbf{s}) \sim$ ind. $AL(0, \sigma^{\tau}(\mathbf{s}), \tau)$. The AL distribution is characterized by location, scale, and asymmetry parameters, μ, σ, τ; by setting $\mu = 0$ to ensure $P(\varepsilon \le 0) = \tau$, the density of $\varepsilon \sim AL(0,\sigma,\tau)$ is written as $f(\varepsilon) = \tau(1-\tau)\sigma \exp\{-\sigma\varepsilon[\tau - \mathbf{1}(\varepsilon < 0)]\}$. A convenient strategy for generating ε's is to use the following representation, $\varepsilon = \sqrt{\frac{2U}{\sigma^2\tau(1-\tau)}} Z + \frac{1-2\tau}{\sigma\tau(1-\tau)} U$, where $Z \sim N(0,1)$ and $U \sim Exp(1)$. So, $\varepsilon \mid \sigma, U$ is normally distributed enabling us to use all the familiar Gaussian theory. In the same way as above, we specify $\sigma^{\tau}(\mathbf{s})$ to capture spatial scale dependence through $Z^{\tau}_{\sigma}(\mathbf{s}) = \log\{\sigma^{\tau}(\mathbf{s})\} \sim GP(Z^{\tau}_{\sigma}, C(\cdot;\sigma^{2,\tau}_{\sigma}, \phi^{\tau}_{\sigma}))$.

Model inference is implemented in a Bayesian framework. The conditional AL distribution for all $Y_{t\ell}(\mathbf{s})$ can be expressed as normal when it is conditioned on $U^{\tau}_{t\ell}(\mathbf{s}) \sim Exp(1)$. To complete the model we specify diffuse and, when available, conjugate priors such as normal and inverse gamma for all model parameters. We develop a Metropolis-within-Gibbs algorithm to obtain Markov chain Monte Carlo samples from the joint posterior distribution. Full conditional distributions for each of the parameters are derived, including the $n \times T \times (L-1)$ reparameterized latent exponential variables $\xi^{\tau}_{t\ell}(\mathbf{s}) = U^{\tau}_{t\ell}(\mathbf{s})/\sigma^{\tau}(\mathbf{s})$.

Adequacy of the model is studied considering the performance across the L days within year, T years and n locations. We apply a leave-one-out cross-validation where the conditional quantiles are obtained using one-step ahead prediction. A version of the $R^1(\tau)$ by Koenker and Machado [5], and the probability $p(\tau)$ that an observation is less than the conditional quantile are calculated. In-sample, $R^1(\tau)$ takes values between 0 and 1, where a value close to 1 indicates a better fit. The target for $p(\tau)$ is proximity to τ. Analogous versions of these measures without averaging over days, years, or sites have also been considered.

2.2 Marginal quantiles

The conditional quantile model is used to extract a marginal quantile from the conditional quantile. The first idea is to add an adjustment term to $q^{\tau}_{t\ell}(\mathbf{s})$ in (1). Note that $q^{\tau}_{t\ell}(\mathbf{s})$ is not immediately a marginal quantile for $Y_{t\ell}(\mathbf{s})$ because $P(Y_{t\ell}(\mathbf{s}) \leq q^{\tau}_{t\ell}(\mathbf{s})) \neq \tau$. The proposed adjustment to $q^{\tau}_{t\ell}(\mathbf{s})$ will adjust this probability to τ.

For sake of simplicity, space and years and the superscript τ in the parameters are suppressed. We have $Y_\ell = q_\ell + \rho(Y_{\ell-1} - q_{\ell-1}) + \varepsilon_\ell$, where $\varepsilon_\ell \sim$ i.i.d. $AL(0,\sigma,\tau)$. Using this notation, $Q_{Y_\ell}(\tau \mid Y_{\ell-1}) = q_\ell + \rho(Y_{\ell-1} - q_{\ell-1})$ is the τ quantile of the QAR. For convenience, write this model as $W_\ell = \rho W_{\ell-1} + \varepsilon_\ell$ with $W_\ell = Y_\ell - q_\ell$. Upon substitution, we have $W_\ell = \rho^\ell W_0 + \sum_{j=0}^{\ell-1} \rho^j \varepsilon_{\ell-j}$. We consider the τ quantile of W_ℓ, call it $d^{\tau}_{\ell}(\rho,\sigma)$, so that the τ quantile of $W_\ell - d^{\tau}_{\ell}(\rho,\sigma)$ is 0, and therefore the τ marginal quantile of Y_ℓ is $q_\ell + d^{\tau}_{\ell}(\rho,\sigma)$. Using the conditional normal form for ε_ℓ and defining $\tilde{\varepsilon}_\ell \equiv \sum_{j=0}^{\ell-1} \rho^j \varepsilon_{\ell-j}$, we have

$$\tilde{\varepsilon}_\ell \mid \rho,\sigma,U_\ell,U_{\ell-1},\ldots,U_1 \sim N\left(\frac{1-2\tau}{\sigma\tau(1-\tau)}\sum_{j=0}^{\ell-1}\rho^j U_{\ell-j}, \frac{2}{\sigma^2\tau(1-\tau)}\sum_{j=0}^{\ell-1}\rho^{2j}U_{\ell-j}\right). \tag{2}$$

Though $\tilde{\varepsilon}_\ell$ does not have an AL distribution we can find its τ quantile. For any d, we seek

$$P(\tilde{\varepsilon}_\ell < d \mid \rho,\sigma) = \int\int\cdots\int P(\tilde{\varepsilon}_\ell < d \mid \rho,\sigma,\{U_j : j = 1,2,\ldots,\ell\})[\{U_j\}]dU_1dU_2\cdots dU_\ell.$$

Given $\{U_j : j = 1,2,\ldots,\ell\}$, we have the distribution for $\tilde{\varepsilon}_\ell$ in (2). We can do a Monte Carlo integration to calculate $P(\tilde{\varepsilon}_\ell < d \mid \rho,\sigma)$ by generating many sets $\{U_j : j = 1,2,\ldots,\ell\}$, all i.i.d., all distributed as $Exp(1)$. Then, using a simple search, we can find $d^{\tau}_{\ell}(\rho,\sigma)$. In our modeling setting we can create the posterior distribution of the τ marginal quantile for any year, day, and site. In the sequel, we denote this as $\tilde{q}_{Y_{t\ell}(\mathbf{s})}(\tau) \equiv q^{\tau}_{t\ell}(\mathbf{s}) + d^{\tau}_{t\ell}(\rho^\tau(\mathbf{s}),\sigma^\tau(\mathbf{s}))$.

These marginal quantiles can be kriged over a spatial region for any τ, year, and day within year to reveal the temperature *quantile surface*; i.e., we can obtain the posterior distribution of $\tilde{q}_{Y_{t\ell}(\mathbf{s}_0)}(\tau)$ at any new site $\mathbf{s}_0$. We obtain this for a sufficiently spatially resolved grid, and we can obtain the posterior mean at each point and represent the posterior τ quantile surface for the given day within year.

3. Results

The data include the time series from $n = 18$ sites at a daily scale from 1956 to 2015, provided by the

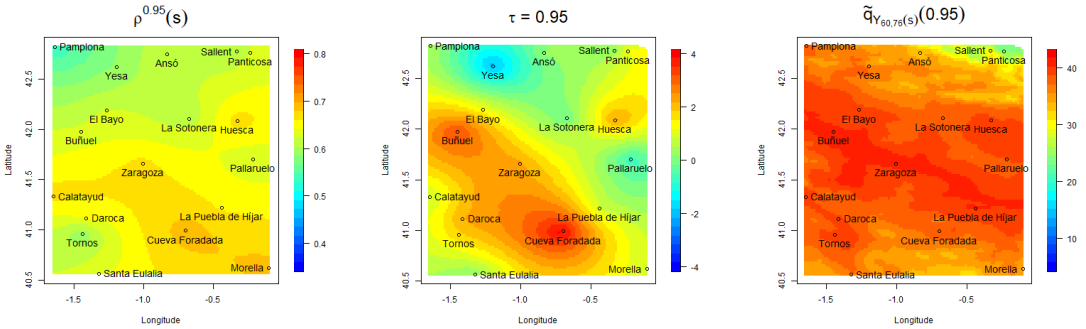


Figure 1: Left: Spatially varying autorregression coefficients. Center: Difference in °C between marginal quantiles of the last and the first decade. Right: Marginal 0.95 quantile on July 15, 2015.

Spanish Meteorological Office, but we focus on the warm months from May 1 to September 30. We show some results for the $\tau = 0.95$ quantile as an illustration.

The model assessment yields $p(0.95) = 0.944$ and $R^1(0.95) = 0.442$. The daily, annual and local measurements also show the good assessment of the model. Figure 1 shows maps of $E(\rho^{0.95}(\mathbf{s}) \mid data)$, $E\left(\sum_{t\in D6} \tilde{q}_{Y_{t\ell}(\mathbf{s})}(0.95) - \sum_{t'\in D1} \tilde{q}_{Y_{t'\ell}(\mathbf{s})}(0.95) \mid data\right)/10$ where $D1$ is the first decade (1956–1965) and $D6$ the last (2006–2015), and $E(\tilde{q}_{Y_{60,75}(\mathbf{s})}(0.95) \mid data)$. It is observed that $\rho^{0.95}(\mathbf{s})$ shows a strong serial correlation that varies spatially from 0.53 to 0.69. A general warming between decades is observed, it exceeds 3°C in the southwest, but a cooling pattern appears in the northwest. Finally, marginal quantiles enjoy direct interpretation, the range for this marginal quantile goes from 23.3°C to 41.1°C. Other quantiles have been fitted, obtaining remarkable differences in components and trends.

Summary and future work

A modeling approach to predict a specific quantile in a spatio-temporal framework is proposed. We have specified a spatial autoregressive model on a daily scale using the AL distribution for the errors, that captures serial correlation. An attractive approach to obtain marginal quantiles at daily scale from the conditional quantiles fitted by the model is also developed. This allows posterior inference to evaluate distributional changes between marginal quantiles.

Future work focuses on characterizing a joint QAR model. This will allow the comparison of persistence and long-term trends between quantiles jointly.

Acknowledgments

This work was partially supported by the Spanish Ministe of Science under Grant PID2020-116873GB-I00; Gobierno de Aragón under Research Group E46_20R: Modelos Estocásticos; and JC-M was supported by Go-

bierno de Aragón under Doctoral Scholarship ORDEN CUS/581/2020.

References

[1] Koenker R., and Bassett G. (1978). Regression quantiles. *Econometrica*, **46**: 33–50.

[2] Yang Y., and Tokdar S. T. (2017). Joint estimation of quantile planes over arbitrary predictor spaces. *Journal of the American Statistical Association*, **112**: 1107–1120.

[3] Lum K., and Gelfand A. E. (2012). Spatial quantile multiple regression using the asymmetric Laplace process. *Bayesian Analysis*, **7**: 235–258.

[4] Reich B. J. (2012). Spatiotemporal quantile regression for detecting distributional changes in environmental processes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **61**: 535–553.

[5] Koenker R., and Machado J. A. F. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, **94**: 1296–1310.

Estimation of the Spatial Weighting Matrix

P. Otto[1,*] , M.S. Merk[2,*] and R. Steinert[3,*]

[1]*Leibniz University Hannover, Germany; otto@ikg.uni-hannover.de*
[2]*University of Göttingen, Germany*
[3] *European University Viadrina, Frankfurt (Oder), Germany*
[*]*Corresponding author*

Abstract. *Spatial econometric research typically relies on the assumption that the spatial dependence structure is known in advance and is represented by a deterministic spatial weights matrix. This matrix, like an adjacency matrix in graphical models, defines how the locations are connected, meaning which locations are possibly dependent and by which extend. The spatial autoregressive term is then classically modelled as multiple of the product of this predefined weighting matrix and the vector of observations. Of course, estimated spatial autoregressive coefficients, therefore, depend on the choice of this matrix. Thus, all coefficients as well as inference on these parameters should always be done conditional on the definition of the weighting scheme. From this perspective, the current practice is quite unsatisfactory. In this talk, I will present the results of two papers. First, we propose a two-stage lasso estimation approach for the estimation of a full spatial weights matrix of spatiotemporal autoregressive models. In addition, we allow for an unknown number of structural breaks in the local means of each spatial location. These locally varying mean levels, however, can easily be mistaken as spatial dependence and vice versa. Thus, the proposed approach jointly estimates the spatial dependence, all structural breaks, and the local mean levels. For selection of the penalty parameter, we propose a completely new selection criterion based on the distance between the empirical spatial autocorrelation and the spatial dependence estimated in the model. Secondly, we investigate the estimation of sparse spatial dependence structures for regular lattice data. In particular, an adaptive least absolute shrinkage and selection operator (LASSO) is used to select and estimate the individual connections of the spatial weights matrix. To recover the spatial dependence structure, we propose cross-sectional resampling, assuming that the random process is exchangeable. The estimation procedure is based on a two-step approach to circumvent simultaneity issues that typically arise from endogenous spatial autoregressive dependencies*

Keywords. *Adaptive LASSO; Change Points; Cross-Sectional Resampling; Spatial Weights Matrix.*

References

[1] Merk, M. S., Otto, P. (2020) Estimation of the spatial weighting matrix for regular lattice data – An adaptive lasso approach with cross-sectional resampling. *arXiv* arXiv:2001.01532.

[2] Otto, P., Steinert, R. (2018) Estimation of the Spatial Weighting Matrix for Spatiotemporal Data under the Presence of Structural Breaks. *arXiv* arXiv:1810.06940

Using a constructed covariate that accounts for preferential sampling

A. Monteiro[1,*] , M.L. Carvalho[2] , I. Figueiredo[2,3] , P. Simões[1,4] and I. Natário[1,5]

[1]*NOVA MATH - Center for Mathematics and Applications (CMA), NOVA University of Lisbon, Portugal; andreiaforte50@gmail.com*
[2] *Centre of Statistics and its Applications (CEAUL), Faculty of Sciences of the University of Lisbon, Portugal; mlucilia.carvalho@gmail.com*
[3] *Portuguese Institute for Sea and Atmosphere (IPMA), Portugal; ifigueiredo@ipma.pt*
[4] *Military Academy Research Center - Military University Institute (CINAMIL), Portugal; pc.simoes@campus.fct.unl.pt*
[5] *Department of Mathematics, NOVA School of Science and Technology, Portugal; icn@fct.unl.pt*
[*] *Corresponding author*

Abstract. *In Geostatistics, the common assumption is that the selection of the sampling locations does not depend on the values of the spatial variable of interest. However, dependence can be observed for example in fishery data, where catches are certainly associated with the locations where the fisheries take place, in order to optimize capture effort. Thus, the process under study determines the data-locations and the above mentioned assumption is violated. This phenomenon is coined as preferential sampling and ignoring the preferential nature of the sampling can lead to biased estimates and misleading inferences. We plan to investigate the use of constructed covariates, based on an average value of the nearest neighbors observations distances, that are able to mitigate preferential sampling. The objective is that the inclusion of this covariate is able to explain the stochastic dependence of sampling locations on the spatial variable under study. If this dependence is no longer detected after this adjustment, then we can use standard statistical techniques. This approach is assessed in a simulation study and we also discuss issues specific to this approach that arise when several study configurations are accounted in a model. The methodology is illustrated using a real data set provided by the Instituto Portuguls do Mar e da Atmosfera*

Keywords. *Preferential sampling; Constructed covariates; Nearest neighbour distances.*

1. Geostatistical model for preferential sampling

Diggle and colleagues [2] developed a model for geostatistical data collected in a preferential way, where sampling locations and observations are jointly modelled depending on a common unobserved random field. According to authors, the model for point locations is a log Gaussian Cox process with intensity

$$\Lambda(x_i) = \exp\{\alpha + \beta S(x_i)\} \tag{1}$$

where S is a stationary Gaussian Process with mean μ_s and variance σ_s^2 and β controls the degree of preferentiality, for example, when $\beta > 0$ the sample points are concentrated, predominantly, near the maximum of the observed values and when $\beta = 0$ it corresponds to the situation of a non-preferential sampling, corresponds to

a homogeneous Poisson process with intensity $\exp(\alpha)$.
The model for the data takes the form

$$Y(x_i) = S(x_i) + W_i \tag{2}$$

$Y(x_i)$ denotes the measured value at the location x_i and W_i is a Gaussian random error with mean 0, variance τ^2 and $i = 1 \cdots n$, where n is the number of locations.
The modeling approach suggested [2], accounts for preferential sampling using likelihood-based inference with Monte Carlo methods, but Bayesian inference based on a SPDE-INLA approach has more recently been used [3]. Geostatistical model, can be regarded from the perspective of a marked point process, modelling the marks the observed quantities and the points the sampling locations.

2. Constructed Covariates

Due to the computational challenges of fitting joint models, detecting preferential sampling or dependence between marks and points is therefore an important issue. When there are covariates available, it is possible that when they are explicitly included in the model they are sufficient to account for this relationship between points and marks. The discovery of these covariates may justify the continued use of standard methodologies, [5].
We consider an approach to deal with preferential sampling using a constructed covariate. Constructed covariates are summary characteristics defined for any location in the observation window reflecting inter-individual spatial behavior such as local interaction or competition, [4]. The constructed covariate considered in this paper is based on the nearest point distance, which is simple and fast to compute.
The constructed covariate considered is based on the averaged K nearest neighbour distances. We consider a grid for the region and for each point of the grid, s, we find the distance to the k nearest point in the observation pattern $X = (x_1, \cdots, x_n)$ as

$$d(s,k) = \frac{1}{k} \sum_{j=1}^{k} ||s - x_j||$$

where $||\cdot||$ denotes the Euclidean distance.

3. Numerical Studies

In this Section, we document the performance of including the constructed covariate to mitigate the effect of preferential sampling. This is demonstrated across a range of simulated data settings. We consider three degrees of preferentiality, with $\beta = -2$, $\beta = 0.3$ and $\beta = 2$ and data with three sample sizes, $n = 50$, $n = 100$ and $n = 250$. Each experimental setting is repeated 100 times. To illustrate these sampling schemes, we represent in Figure 1, 100 sampling points, assuming β equals 2, -2 and 0.3.
Considering the average distance to 5% of the nearest neighbors as the construct covariate and performing for the different degrees of preferentiality and sample sizes, a total of 100 independent samples, Table 1 summarizes the percentage of replicas in which the inclusion of the constructed covariate failed to mitigate the effect of preferential sampling, considering 10% confidence level. By analysing Table 1, the results are quite satisfac-

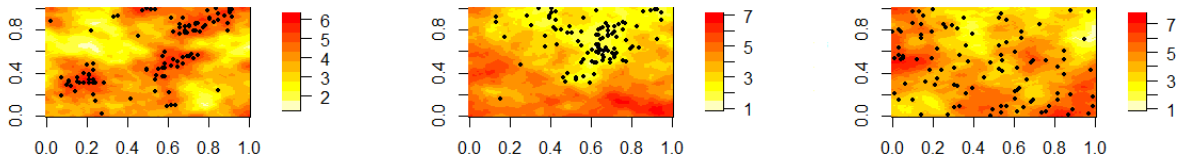


Figure 1: Sampling schemes assuming $\beta = 2$; $\beta = -2$; $\beta = 0.3$.

tory and we believe that the inclusion of the construct covariate is able to explain the stochastic dependence of sampling locations on the spatial variable under study. This allows proceeding with data analysis using standard statistical techniques.
Further studies with different combinations of the parameters, namely for β and different percentages of nearest neighbors lead to similar conclusions.

	$n = 50$	$n = 100$	$n = 250$
$\beta = 2$	1%	0%	2%
$\beta = -2$	2%	2%	7%
$\beta = 0.3$	0%	0%	2%

Table 1: Percentage of replicas in which the inclusion of the constructed covariate failed to mitigate the effect of preferential sampling.

4. Data example

We illustrate the previously described methodology on a real data provided by the Instituto Portuguls do Mar e da Atmosfera (IPMA) on black scabbardfish catches. A subset of the original data described in [1] was taken for this data analysis: the fishing area with latitude minor than 39.3^o, captures that have occurred from September to February for the years between 2009 to 2013, resulting in a total set of 733 observations. The data, include the Black Scabbardfish (BSF) catches (in kg) by fishing haul of the longline fishing fleet but also include the location of each fishing haul, Figure 2.

The existence of preferential sampling was detected in BSF data with a degree of preferentiality equal to

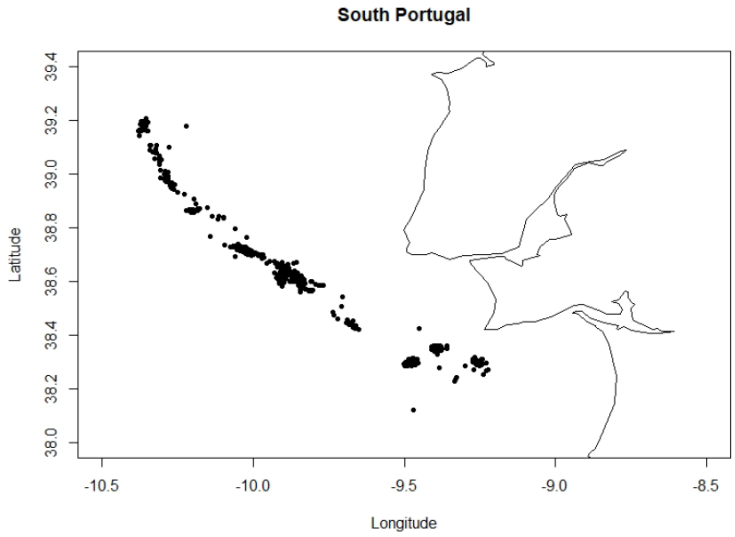


Figure 2: Locations of the BSF catches.

0.5. With the inclusion of the constructed covariate (average distance to 5% of nearest neighbors), we mitigate the preferential sampling effect and we can now proceed the analysis using standard statistical techniques.

5. Concluding remarks and further work

We present a methodology that allows to deal with sampling designs that depend on the observed values of the spatial variable and the suggested approach exhibited, in the numerical studies, a quite satisfactory performance. For future investigation we intend to investigate more general covariates, certainly also suitable, such as the local intensity or the number of points within a fixed interaction radius from a location $s \in \mathbb{R}^2$.

Acknowledgments

This work is funded by national funds through the FCT - Fundação para a Ciência e Tecnologia, I.P., under the scope of the projects PREFERENTIAL, PTDC/MAT-STA/28243/2017; UIDB/00297/2020 and UIDP/00297/2020 (Center for Mathematics and Applications); and UIDB/00006/2020 (CEAUL).

References

[1] André, L. M., Figueiredo, I., Carvalho, M. L., Simões, P. and Natário, I. (2020). Spatial modelling of black scabbardfish fishery off the portuguese coast. In *International Conference on Computational Science and Its Applications*, 332–344. Springer.

[2] Diggle, P., Menezes, R. and Su, T. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):191–232.

[3] Dinsdale, D. and Salibian-Barrera, M. (2019). Modelling ocean temperatures from bio-probes under preferential sampling. *The Annals of Applied Statistics*, 13(2):713–745.

[4] Illian, J. B., Sørbye, S. H. and Rue, H.(2012). A toolbox for fitting complex spatial point process models using integrated nested laplace approximation (INLA). *The Annals of Applied Statistics*, pages 1499–1530.

[5] Watson, J. (2021) A perceptron for detecting the preferential sampling of locations and times chosen to monitor a spatio-temporal process. *Spatial Statistics*, 43:100–500.

Mitigating Spatial Confounding by Explicitly Correlating Gaussian Random Fields

I. Marques[1,*] , T. Kneib[1] and N. Klein[2]

[1]*Georg-August-University of Göttingen, Chairs of Statistics, Humboldtallee 3, 37073 Göttingen, Germany; imarques@uni-goettingen.de, tkneib@uni-goettingen.de*
[2]*Humboldt-Universität zu Berlin, School of Business and Economics, Unter den Linden 6, 10099 Berlin, Germany; nadja.klein@hu-berlin.de*
[*]*Corresponding author*

Abstract. *Spatial models are used in a variety of research areas, such as environmental sciences, epidemiology, or physics. A common phenomenon in such spatial regression models is spatial confounding. This phenomenon is observed when spatially indexed covariates modeling the mean of the response are correlated with a spatial random effect included in the model, for example, as a proxy of unobserved spatial confounders. As a result, estimates for regression coefficients of the covariates can be severely biased and interpretation of these is no longer valid. Recent literature has shown that typical solutions for reducing spatial confounding can lead to misleading and counterintuitive results. In this paper, we develop a computationally efficient spatial model that explicitly correlates a Gaussian random field for the covariate of interest with the Gaussian random field in the main model equation and integrates novel prior structures to reduce spatial confounding. Starting from the univariate case, we extend our prior structure also to the case of multiple spatially confounded covariates. In simulation studies, we evaluate the performance of our model. Finally, as a real data illustration, we study the effect of elevation and temperature on the mean of monthly precipitation in Germany.*

Keywords. *Bayesian inference; Penalized complexity prior; Spatial statistics; Stochastic partial differential equation.*

1. Introduction

Spatial regression data are regression data $\{(y(\boldsymbol{s}), \boldsymbol{z}(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{S}\}$, where both the response variable $y(\boldsymbol{s})$ and the explanatory variables $\boldsymbol{z}(\boldsymbol{s})$ are indexed by a spatial variable $\boldsymbol{s}$ representing the location of the corresponding observational unit in a spatial domain $\mathcal{S}$. Based on observations collected at locations $\boldsymbol{s}_i \in \mathcal{S}$, $i = 1, \ldots, n$, a standard regression model would then be of the form

$$y(\boldsymbol{s}_i) = \beta_0 + \boldsymbol{z}(\boldsymbol{s}_i)^T \boldsymbol{\beta} + \varepsilon_i, \tag{1}$$

with i.i.d. error terms $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. However, this model could only be applied if – after adjusting for the covariates $\boldsymbol{z}(\boldsymbol{s}_i)$ – no spatial dependence remains such that the residuals can indeed be assumed to be independent. This assumption is questionable for typical, observational spatial data where spatial dependence is likely to arise, for example, due to direct interaction between observational units of interest with spatial proximity

determining the intensity of mutual interactions or due to omitted variables that themselves obey spatial dependence. In this paper, we are considering situations of the second type, i.e. model (1) is assumed to represent the true data generating mechanism, yet we only have access to a subset of covariates instead of the complete vector.

More formally, we assume that $\boldsymbol{z}(\boldsymbol{s}) = (\boldsymbol{z}_{\text{obs}}(\boldsymbol{s})^T, \boldsymbol{z}_{\text{unobs}}(\boldsymbol{s})^T)^T$ and $\boldsymbol{z}_{\text{obs}}(\boldsymbol{s})$ is a subset of covariates that are observed, while the remainder $\boldsymbol{z}_{\text{unobs}}(\boldsymbol{s})$ is unobserved. This refers to the classical setup of unobserved confounders where: (1) naive estimation of the model relying exclusively on observed data will be biased unless observed and unobserved covariates are uncorrelated, and (2) estimation uncertainty will usually be underestimated in the naive model based on observed data only.

In spatial data, it is particularly unlikely that $\boldsymbol{z}_{\text{obs}}(\boldsymbol{s})$ and $\boldsymbol{z}_{\text{unobs}}(\boldsymbol{s})$ are uncorrelated when both obey spatial dependence themselves. A common approach that is considered to adjust for unobserved spatial covariates is to approximate their overall effect as $\gamma(\boldsymbol{s}) \approx \boldsymbol{z}_{\text{unobs}}(\boldsymbol{s})^T \boldsymbol{\beta}_{\text{unobs}}$ and to model $\gamma(\boldsymbol{s})$ as a spatially correlated random effect in the model

$$y(\boldsymbol{s}_i) = \beta_0 + \boldsymbol{z}_{\text{obs}}(\boldsymbol{s}_i)^T \boldsymbol{\beta}_{\text{obs}} + \gamma(\boldsymbol{s}_i) + \varepsilon_i. \tag{2}$$

Unfortunately, this modeling strategy does not completely solve the problem since the correlation between the observed and unobserved part of the covariates carries over to dependence between $\boldsymbol{z}_{\text{obs}}(\boldsymbol{s})$ and $\gamma(\boldsymbol{s})$, which is then usually referred to as spatial confounding. Whether or not spatial confounding is a relevant concern, depends both on the specific properties of the omitted variables $\boldsymbol{z}_{\text{unobs}}(\boldsymbol{s})$ and the purpose of the analysis. For the former, the scale at which spatial variability is observed and how this scale relates to the spatial variability in $\boldsymbol{z}_{\text{obs}}(\boldsymbol{s})$ is of major relevance (e.g. [3]).

In this paper, we pragmatically consider dealing with spatial confounding as a way of getting a more realistic approximation of the underlying data generating process that enables meaningful interpretation of $\boldsymbol{\beta}_{\text{obs}}$, positioning our contribution close to the evaluation of spatial risk factors [1]. Thus, we distance ourselves from the causal inference literature. We develop a Bayesian framework that allows to deal with spatial confounding in continuously indexed spatial models using a novel prior structure. For this, we model $\gamma(\boldsymbol{s})$ and $\boldsymbol{z}_{\text{obs}}(\boldsymbol{s})$ jointly using a multivariate Gaussian random field (MGRF) distribution. We do estimation with the stochastic partial differential equation (SPDE) approach [2] which utilizes a Gaussian Markov random field (GMRF) for computations and use Markov Monte Carlo (MCMC) for inference. Moreover, we explore a penalized complexity (PC) prior for the correlation parameter that controls the shrinkage towards a base model, i.e. the case of no spatial confounding [5]. Finally, we extend our model to the case of multiple spatially confounded covariates in a spatial model, while explicitly accounting for spatial scale and computational complexity.

2. Spatial confounding in one covariate

In the SPDE-approach [2], the Gaussian random field (GRF) $\gamma(\boldsymbol{s})$ (similarly for $\boldsymbol{z}(\boldsymbol{s})$) that solves a given SPDE is expanded into a piecewise linear basis through

$$\gamma(\boldsymbol{s}) = \sum_{m=1}^{M} \psi_m(\boldsymbol{s}) \gamma_m, \tag{3}$$

where the joint distribution of the GMRF weight vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_M)^T$ are normally distributed weights $\boldsymbol{\gamma}$ with mean zero and sparse precision matrix $\boldsymbol{Q}_{\boldsymbol{\gamma}}$ such that $\boldsymbol{\gamma} \sim N(\mathbf{0}, \boldsymbol{Q}_{\boldsymbol{\gamma}}^{-1})$. Each basis function $\psi_m(\cdot)$ is piecewise linear. Thus, the GRF and GRMF are empirically equivalent and connected via a SPDE and we benefit from both the computational benefits of the GMRF and a well-defined continuous GRF counterpart

Consider the model in (2). Let $\boldsymbol{z}$ be a covariate having a spatial structure that can be represented by a GMRF with mean $\boldsymbol{\mu_Z}$ and precision matrix $\boldsymbol{Q}_z$, i.e., $\boldsymbol{z} \sim \mathcal{N}_p(\boldsymbol{\mu_z}, \boldsymbol{Q}_z)$. Let $\boldsymbol{\gamma} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{Q}_\gamma)$. For positive definite $\boldsymbol{Q}_{\boldsymbol{\gamma}}$, there is a unique Cholesky triangle $\boldsymbol{L}_{\boldsymbol{\gamma}}$ such that $\boldsymbol{L}_{\boldsymbol{\gamma}}$ is a lower triangular matrix with $L_{\gamma(ii)} > 0 \ \forall i$ and $\boldsymbol{Q}_{\boldsymbol{\gamma}} = \boldsymbol{L}_{\boldsymbol{\gamma}} \boldsymbol{L}_{\boldsymbol{\gamma}}^T$. By exploiting the Markov graph structure of the GMRF, the Cholesky factor of the precision matrix can be guaranteed to be sparse [4]. The same logic applies to $\boldsymbol{z}$. We assume that $\boldsymbol{\gamma}$ and $\boldsymbol{z}$ are jointly Gaussian distributed and we can write the conditional mean and precisions as

$$\boldsymbol{\mu}_{\boldsymbol{\gamma}|\boldsymbol{z}} = \mathbf{0} + \rho \, \boldsymbol{L}_{\boldsymbol{\gamma}}^{-T} \boldsymbol{L}_{\boldsymbol{z}}^T (\boldsymbol{z} - \boldsymbol{\mu_Z}) \tag{4}$$

$$\boldsymbol{Q}_{\boldsymbol{\gamma}|\boldsymbol{z}} = \boldsymbol{\Sigma}_{\boldsymbol{\gamma}|\boldsymbol{z}}^{-1} = \frac{1}{1-\rho^2} \boldsymbol{Q}_{\boldsymbol{\gamma}} \tag{5}$$

where $\rho \in [-1, 1]$ such that $\text{Corr}(\gamma_m, Z_m) = \rho$, for all $m = 1, \ldots, M$. We refer to a prior that assumes $\rho = 0$ as the base prior, and the model associated with it as the base model. The more general case with $\rho \in [-1, 1]$ is denoted MGRF prior, and the model associated with it the MGRF model. We use a PC-prior [5] on ρ that shrinks the model towards the less complex base model.

We need to avoid boundary cases $|\rho| \approx 1$. To address this, one can consider the unconstrained GMRF, $\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}_{\boldsymbol{\gamma}}^{-1})$, and then compute

$$\boldsymbol{\gamma}^* = \boldsymbol{\gamma} + \rho \, \boldsymbol{L}_{\boldsymbol{\gamma}}^{-T} \boldsymbol{L}_{\boldsymbol{z}}^T (\boldsymbol{z} - \boldsymbol{\gamma}_{\boldsymbol{z}}), \text{ where } \boldsymbol{\gamma}_{\boldsymbol{z}} \sim \mathcal{N}(\boldsymbol{\mu_z}, \boldsymbol{\Sigma_z}). \tag{6}$$

The resulting $\boldsymbol{\gamma}^*$ has the correct conditional distribution. This structure essentially removes from the linear predictor the part of the spatial effect that is correlated with the covariate of interest.

3. Spatial confounding in multiple covariates

Spatial confounding typically biases regression coefficients when $\boldsymbol{z}$ has a larger spatial scale than $\boldsymbol{\gamma}$, since when the spatial effect operates on a smaller scale than the covariate, it is likely to explain the data better than the covariate [3, 1]. With this in mind, in the multiple covariates case we use principal component analyzes (PCA) to reduce the dimensionality of the design matrix $\boldsymbol{z}$ when deriving the prior for $\boldsymbol{\gamma}$. We choose the first B^* (largest) eigenvalues $\lambda_1, \ldots, \lambda_{B^*}$ and their respective eigenvectors, as these should capture large scale behavior. The vector of variables resulting from PCA is $\boldsymbol{w} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{b^*}, \ldots, \boldsymbol{w}_{B^*})^T$. The model follows

$$y(\boldsymbol{s}_i) = \beta_0 + \boldsymbol{z}(\boldsymbol{s}_i)^T \boldsymbol{\beta} + \boldsymbol{\gamma}(\boldsymbol{s}_i) + \varepsilon_i$$

but now the spatial effect has the prior $\boldsymbol{\gamma}|\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\gamma}|\boldsymbol{w}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}|\boldsymbol{w}})$. Either (4) and (5) or (2) can be used.

4. Simulation study

In the simulation study, we were able to confirm that spatial confounding typically biases regression coefficients when $\boldsymbol{z}$ has a larger spatial scale than $\boldsymbol{\gamma}$. Our prior structure behaved quite well in these scenarios, in both the single and multiple covariate cases, and successfully shrinked the model towards the base model in the remainder.

	β_1^{obs} (elevation)		β_2^{obs} (temperature)	
Model	Mean	95% CI	Mean	95% CI
NS	0.024	[-0.024, 0.071]	-0.381	[-0.429, -0.332]
Base model	0.641	[0.493, 0.802]	-0.035	[-0.140, 0.066]
RSR	0.023	[-0.04, 0.088]	-0.170	[-0.228, 0.113]
MGRF	0.165	[0.068, 0.258]	-0.143	[-0.202, -0.086]

Table 1: Mean and equal-tailed 95% credible interval for the posterior of β_1^{obs} and β_2^{obs} in the five models.

5. Application

We analyze average monthly precipitation in Germany in October 2015 using open-access data from the German Meteorological Institute We consider model (2). Variable $y(\boldsymbol{s}_i)$ is the *standardized* amount of precipitation in milliliters at weather station $\boldsymbol{s}_i \in \mathcal{S}$, where $\mathcal{S}$ represents Germany. Covariate $z_1(\boldsymbol{s}_i)$ is the standardized elevation in meters and $z_2(\boldsymbol{s}_i)$ is the standardized average monthly minimum temperature in degrees Celsius. We test four models with different specifications for $\gamma(\boldsymbol{s})$: (i) $\gamma(\boldsymbol{s})$ is excluded from the model, i.e., non-spatial (NS) model, (ii) $\gamma(\boldsymbol{s})$ has a base prior that does not account for spatial confounding, (iii) $\gamma(\boldsymbol{s})$ has our novel MGRF prior, (iv) $\gamma(\boldsymbol{s})$ is the spatial effect not in the span of the fixed effects, i.e., restricted spatial regression (RSR), which is the classic solution for spatial confounding.

Although linearity is a strong assumption, one would generally expect a positive association between elevation and monthly precipitation ($\beta_1^{\text{obs}} > 0$) and a negative association between minimum temperature and monthly precipitation ($\beta_2^{\text{obs}} < 0$). Table 1 shows the posterior summaries for β_1^{obs} and β_2^{obs} and it clearly demonstrates the discrepancies in the posterior distribution of β_1^{obs} and β_2^{obs} for different models.

By using our prior we go further into the expected direction of the association between the two covariates and monthly precipitation: none of the credibility intervals cover zero and both coefficients show the expected association. Thus, the MGRF seems to pull the coefficients in the direction of the expected association, while RSR and NS behave similarly, as previously reported in the literature (e.g. [1]).

References

[1] Adin, A., Goicoa, T., Hodges, J. S., Schnell, P. M. and Ugarte, M. D. (2021). Alleviating confounding in spatio-temporal areal models with an application on crimes against women in India. *Statistical Modelling* 1471082X211015452.

[2] 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* **4**, 423–498.

[3] Paciorek, C. J. (2010). The importance of scale for spatial-confounding bias and precision of spatial regression estimators.*Statistical Science* **25**, 107–125.

[4] Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. Chapman & Hall/CRC. Boca Raton.

[5] Simpson, D., Rue, H., Riebler, A., Martins, T.G. and Sørbye, S.H. (2017). Penalising model component complexity: a principled, practical approach to constructing priors. *Statistical Science* **32**, 1–28.

Data fusion in a Bayesian spatio-temporal model using the INLA-SPDE

S. J. Villejo[1]

[1] *University of Glasgow, Stephen.Villejo@glasgow.ac.uk; University of the Philippines, svvillejo@up.edu.ph*

Abstract. *A two-stage model is proposed motivated by an epidemiological problem which involves data with different spatial supports. The first-stage model performs data fusion to combine measurements from pollution monitors and from a high-resolution data of air quality from numerical models or from satellites. The second-stage model fits a Poisson GLMM to link exposures and the health outcomes. The proposed method is applied on NO2 measurements and respiratory hospitalizations for year 2007 in England. The results show that an increase in NO2 levels is significantly associated with an increase in the relative risks of the health outcome. Also, there is a strong spatial structure in the risks, a strong temporal autocorrelation, and a significant spatio-temporal interaction effect.*

Keywords. *Integrated nested laplace approximation (INLA); data fusion; spatial misalignment*

1. Introduction

Dealing with spatially misaligned data is a common problem in spatial modelling. [1] [2] [5] This paper focuses on a particular case of spatial misalignment relevant to spatial epidemiology. Data on health outcomes such as the incidence of certain diseases are available as aggregated counts on irregular areal units. Measurements on exposures are typically collected from a sparse network of monitoring stations. Two additional sources of information, which has wider spatial coverage, are also used: satellite images and computer simulation using numerical or deterministic models. These data are also called *proxy data*. This problem of combining data from monitoring stations and proxy data is referred to as *data fusion*. [4]

This paper proposes the use of a two-stage model to link exposures and the health outcomes. The first-stage model performs data fusion using an approach similar to the *Bayesian melding* approach. [3] The Bayesian melding approach assumes that the point-referenced outcomes from monitors and the high-resolution proxy data have a common latent spatial process. The former is assumed to be observed with measurement error and the latter as a spatial average of point outcomes of the latent process within a grid. The point outcome inside a grid is a linear function of the true latent process, which has an additive and multiplicative calibration parameters and possibly an additional noise term. The second-stage model fits a Poisson GLMM in addition to spatial effects, temporal effects, and their interactions. Both models are fitted using the integrated nested Laplace approximation(INLA) method. [7] In addition, the exposures model are fitted using the stochastic partial differential equation (SPDE) approach. The use of the INLA approach is motivated by its computational benefits. The INLA method is a deterministic approach for doing Bayesian inference, as opposed to MCMC

which is a simulation-based approach. The SPDE approach also provides a means to speed up computation [6].

2. Methodology

Suppose we denote by $\boldsymbol{w}_t = \begin{pmatrix} w(\boldsymbol{s}_1,t) & w(\boldsymbol{s}_2,t) & \dots & w(\boldsymbol{s}_M,t) \end{pmatrix}^\intercal$ the measurements from M monitors at time $t, t = 1,\dots,T$. Also, we denote by $\tilde{\boldsymbol{x}}_t = \begin{pmatrix} \tilde{x}(\mathbf{g}_1,t) & \tilde{x}(\mathbf{g}_2,t) & \dots & \tilde{x}(\mathbf{g}_G,t) \end{pmatrix}^\intercal$ the data from the single proxy data at time t, where $\tilde{x}(\mathbf{g}_i,t)$ is the observed value at the grid whose centroid is $\mathbf{g}_i$ at time $t, g = 1,\dots,G$ and $t = 1,\dots,T$. The true latent exposures field is denoted by $\boldsymbol{x}_t$, and both $\boldsymbol{w}_t$ and $\tilde{\boldsymbol{x}}_t$ are error-prone realizations of the true exposure values $\boldsymbol{x}_t$. Suppose the vector of true exposures combined for both the monitors and the centroids of the grids of the proxy data at time t is given by $\boldsymbol{x}_t = \begin{pmatrix} \boldsymbol{x}_{t,M} & \boldsymbol{x}_{t,G} \end{pmatrix}^\intercal$. The spatial dependence are assumed to be induced by $\boldsymbol{\xi}_t = \begin{pmatrix} \boldsymbol{\xi}_{t,M} & \boldsymbol{\xi}_{t,G} \end{pmatrix}$, with $\boldsymbol{\xi}_{t,M}$ as the vector of spatial random effects at the monitors for time t and $\boldsymbol{\xi}_{t,G}$ as the vector of spatial random effects at the centroids of the grids of the proxy data. The first-stage estimation procedure fits the following joint model:

$$\boldsymbol{w}_t = \boldsymbol{x}_{t,M} + \boldsymbol{e}_t, \quad \boldsymbol{e}_t \sim N(\mathbf{0}, \sigma_e^2 \mathbb{I}_M), \ t = 1,\dots,T \tag{1}$$

$$\tilde{\boldsymbol{x}}_t = \alpha_0 \mathbf{1}_G + \alpha_1 \boldsymbol{x}_{t,G} + \boldsymbol{\delta}_t, \quad \boldsymbol{\delta}_t \sim N(\mathbf{0}, \sigma_\delta^2 \mathbb{I}_G), \ t = 1,\dots,T \tag{2}$$

$$\begin{pmatrix} \boldsymbol{x}_{t,M} \\ \boldsymbol{x}_{t,G} \end{pmatrix} = \beta_0 \mathbf{1}_{G+M} + \beta_1 \begin{pmatrix} \boldsymbol{z}_{t,M} \\ \boldsymbol{z}_{t,G} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\xi}_{t,M} \\ \boldsymbol{\xi}_{t,G} \end{pmatrix} \tag{3}$$

$$\begin{pmatrix} \boldsymbol{\xi}_{t,M} \\ \boldsymbol{\xi}_{t,G} \end{pmatrix} = \varsigma \begin{pmatrix} \boldsymbol{\xi}_{t-1,M} \\ \boldsymbol{\xi}_{t-1,G} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\omega}_{t,M} \\ \boldsymbol{\omega}_{t,G} \end{pmatrix}, \ \begin{pmatrix} \boldsymbol{\omega}_{t,M} \\ \boldsymbol{\omega}_{t,G} \end{pmatrix} \sim N(\mathbf{0}, \boldsymbol{\Sigma}), \ t = 1,\dots,T \tag{4}$$

where $\boldsymbol{\omega}_t = \begin{pmatrix} \boldsymbol{\omega}_{t,M} & \boldsymbol{\omega}_{t,G} \end{pmatrix}^\intercal$ is a temporally-independent Gaussian vector with mean zero and covariance matrix $\boldsymbol{\Sigma}$ whose elements are computed using the Matern covariance function with parameters σ_ω^2 and κ, and $\boldsymbol{z}_t = \begin{pmatrix} \boldsymbol{z}_{t,M} & \boldsymbol{z}_{t,P} \end{pmatrix}$ is a vector of covariates.

Suppose $\boldsymbol{\chi}_t$ is the extended latent (exposures) field, and $\boldsymbol{\theta} = \begin{pmatrix} \sigma_e^2 & \sigma_\delta^2 & \alpha_0 & \alpha_1 & \beta_0 & \beta_1 & \varsigma & \sigma_\omega^2 & \kappa \end{pmatrix}^\intercal$, the posterior distribution of interest for the stage model is $\pi(\boldsymbol{\chi}, \boldsymbol{\xi}, \boldsymbol{\theta} | \boldsymbol{w}, \tilde{\boldsymbol{x}}, \mathbf{0})$ which is estimated using the INLA-SPDE approach. Suppose we denote by $\hat{x}(G_{\boldsymbol{s}_j}, t)$ the predicted value at the prediction grid whose centroid is $\boldsymbol{s}_j$ at time t. With $N(\mathbf{s}_j \in B_i)$ as the number of prediction grids with centroids inside B_i, the block-level estimate of exposures at B_i and time t, denoted by $\hat{x}(B_i, t)$, is

computed as

$$\hat{x}(B_i,t) = \frac{1}{N(\mathbf{s}_j \in B_i)} \sum_{\mathbf{s}_j \in B_i} \hat{x}(G_{\boldsymbol{s}_j},t). \tag{5}$$

The second model is specified as follows:

$$Y(B_i,t) \overset{iid}{\sim} \text{Poisson}\Big(\mu(B_i,t) = P(B_i,t)\lambda(B_i,t)\Big) \tag{6}$$

$$\log(\lambda(B_i,t)) = \gamma_0 + \gamma_1 \hat{x}(B_i,t) + \phi_i + \psi_i + \nu_t + \zeta_t + \upsilon_{it} \tag{7}$$

where $Y(B_i,t)$ is the observed count at block B_i at time t, $P(B_i,t)$ is the expected count, and $\lambda(B_i,t)$ is the relative risk. The relative risk is modelled as a function of $\hat{x}(B_i,t)$, a block-specific *iid* effect ϕ_i, a spatial random effect ψ_i which follows the intrinsic conditional autoregressive (AR) process, a time-specific *iid* effect ν_t, a time effect ζ_t which follows the AR process of order 1, and a spatio-temporal interaction effect υ_{it}. The interaction term υ_{it} can take several forms. The first case is when the unstructured effects ϕ_i and ν_t interact, which is called Type I interaction. When the structured temporal effect ζ_t and the unstructured spatial effect ϕ_i interact, this is Type II interaction. Type III and Type IV interaction are defined as the case when ν_t and ψ_i interact, and when both the structured effects ψ_i and ζ_t interact, respectively. The criteria for model selection are the following: marginal likelihood, widely applicable Bayesian information criterion (WAIC), deviance information criterion (DIC), predictive integral transform (PIT), and the conditional predictive ordinate (CPO). Smaller values for the CPO indicates better model fit. Moreover, if the model fits the data well, the values of the PIT should be close to a uniform distribution. To properly account for the uncertainty in the block-level exposure estimates when fitting the second-stage model, several samples from the estimated posterior predictive distribution of the latent field $\hat{\pi}(\boldsymbol{x}_t|\cdot)$ will be generated. For each sample, $\hat{x}(B_i,t)$ will be computed and will be used to fit the second-stage model. All posterior estimates from all samples will be combined to come up with the final posterior estimates.

The method is applied on monthly respiratory hospitalizations at the level of local authorities in England for year 2007, NO2 measurements from the the Automatic Urban and Rural Network AURN), and the pollutant estimates from the Air Quality Unified Model (AQUM) on a $12km^2$ grid.

3. Results and Discussion

Figures 1 and 2 shows a side-by-side plot of the block-level NO2 estimates and the high-resolution AQUM dataset for two time points. The block-level estimates appear to correspond well to the AQUM dataset. Table 1 shows the marginal likelihoods, WAIC, DIC, PIT and CPO for the five second-stage models considered. The model with Type II interaction has the highest marginal likelihood, but the model with Type IV interaction has the lowest WAIC and DIC, although the WAIC and DIC for the Type II model is not too different from the Type IV model. All the mean PIT are close to 0.5 which is the mean of a uniform distribution from 0 to 1, and all the mean CPO are also close to 0. Based on the values, Type II seems to provide the best fit among the five models considered.

Figure 3 shows the estimated posterior distribution of γ_1, $\hat{\pi}(\gamma_1|\cdot)$, for the five models considered. The γ_1

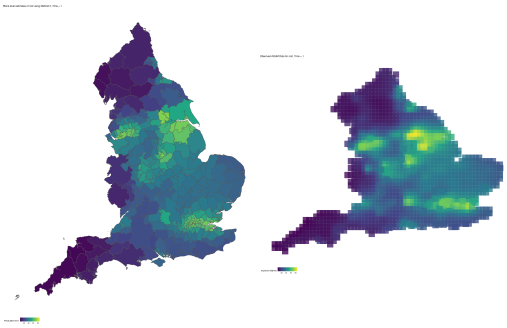


Figure 1: Estimated block exposures (left), AQUM data (right), Time = 1

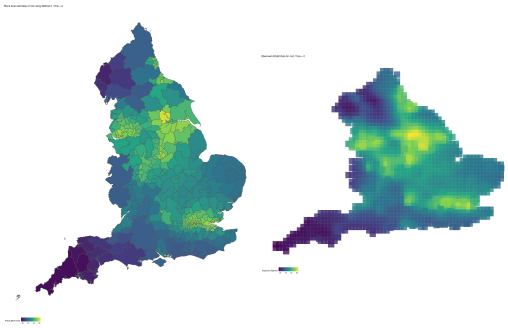


Figure 2: Estimated block exposures (left), AQUM data (right), Time = 2

estimates are quite close to each other except for the model with Type III interaction. As shown in table 1, the model with Type III interaction has 11.42% of the observations with predictive measures which are not reliable due to some numerical problems. Figure 4 shows the parameter estimates and the 95% credible intervals for the model with Type II interaction. Since the block-level estimates of exposures were log-transformed, for a 10% increase in the NO2 levels for a block, we expect the relative risk of respiratory illness to increase by 2.67%. There seems to be a high spatial correlation in the risks since the estimate for the variance of the spatially structured effect σ^2_ψ (0.3050) is higher than the unstructured effect σ^2_ϕ (0.0608). Also, the temporal correlations is very evident as seen by the estimated AR coefficient ρ_{ζ_t} of 0.7970 and the estimate of the variance of the structured time effect $\sigma^2_{\zeta_t}$ (0.0625) which is bigger than the variance of the unstructured time effect (0.0006). The estimated AR coefficient in the interaction term is 0.8105 which indicates a strong interaction effect. The Type II interaction says that *i*th area/block has its own autoregressive structure which is independent from the other areas.

Model	Mlik.Integ	Mlik.Gauss	WAIC	DIC	Failure	PIT	CPO
No Interaction	-17208.3	-17209.0	33646.12	32872.71	0%	0.5093	0.0230
Type I	-16684.5	-16685.1	32404.37	30966.22	0.9%	0.5046	0.0204
Type II	-15831.3	-15831.7	29366.97	29284.8	2.51%	0.4972	0.0206
Type III	-17460.0	-17460.9	29555.34	29516.32	11.42%	0.4873	0.0195
Type IV	-18120.7	-18121.1	29308.01	29203.12	2.65%	0.4999	0.0212

Table 1: Model choice criteria values

4. Conclusions

The proposed method worked in the actual data with the expected result that NO2 is significantly associated with respiratory diseases and additional insights about the spatial and temporal structure of the risks. The

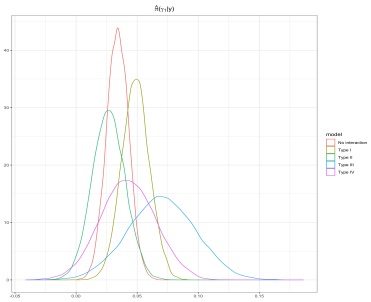


Figure 3: $\hat{\pi}(\gamma_1|\cdot)$ of the five models

Parameter	Mean	P2.5%	P50%	P97.5%
γ_0	-0.5034	-0.8283	-0.5192	-0.0690
γ_1	0.0267	0.0006	0.0266	0.0533
σ^2_ϕ	0.0608	0.0295	0.0592	0.1026
σ^2_ψ	0.3050	0.1800	0.2931	0.4914
$\sigma^2_{\zeta_t}$	0.0625	0.0173	0.0476	0.1954
ρ_{ζ_t}	0.7970	0.5054	0.8189	0.9592
$\sigma^2_{\nu_t}$	0.0006	0.0000	0.0001	0.0048
$\sigma^2_{\upsilon_{it}}$	0.0225	0.0181	0.0223	0.0281
$\rho_{\upsilon_{it}}$	0.8105	0.7446	0.8023	0.8540

Figure 4: Estimates of the final second-stage model

SPDE worked well to efficiently estimate the spatial field, while the INLA method sped up the estimation of the posterior marginals of all parameters. The next step is the perform simulation studies to look at the performance of the proposed method under different scenarios and its sensitivity to priors.

References

[1] Blangiardo, M., Finazzi, F., and Cameletti, M. (2016). Two-stage Bayesian model to evaluate the effect of air pollution on chronic respiratory diseases using drug prescriptions. *Spatial and spatio-temporal epidemiology* **18**, 1–12.

[2] Cameletti, M., Gómez-Rubio, V., and Blangiardo, M. (2019). Bayesian modelling for spatially misaligned health and air pollution data through the INLA-SPDE approach. *Spatial Statistics* **31**, 100353.

[3] Fuentes, M. and Raftery, A. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics* **61**, 36–45.

[4] Lawson, A., Banerjee, S., Haining, R., and Ugarte, M. (2016). Handbook of spatial epidemiology. *CRC Press*

[5] Lee, D., Mukhopadhyay, S., Rushworth, A., and Sahu, S. (2017). A rigorous statistical framework for spatio-temporal pollution prediction and estimation of its long-term impact on health. *Biostatistics* **18**, 370–385.

[6] 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**, 423–498.

[7] 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**, 319–392.

Infectious Diseases Spatio-temporal Modeling with integrated Compartment and Point Process Models

A.V. Ribeiro-Amaral [1,*] , J.A. González[1] and P. Moraga[1]

[1]*King Abdullah University of Science and Technology, CEMSE Division. Thuwal 23955-6900, Saudi Arabia; andre.ribeiroamaral@kaust.edu.sa, jonathan.gonzalez@kaust.edu.sa, and paula.moraga@kaust.edu.sa*
[*]*Corresponding author*

Abstract. *Infectious disease modeling plays an important role in understanding and preventing diseases from spreading, and different approaches can be taken to describe them. In this context, the well-known SIR (Susceptible, Infected and Recovered) compartment model is a common choice for modeling problems of this kind. In this work, we will use the SIR model machinery to describe the disease-spreading evolution in time, and a Cox process to model the spatial correlation in each of the discretized time windows. By means of simulation, we verified that adding the SIR model output in the mean component of the Cox process may drastically improve the quality of the obtained intensity function, especially when making prediction. In summary, our work proposes a framework for a common problem in epidemiology; in particular, we integrate two well-known modeling approaches for the distribution of infectious individuals in space and time in such a way that the predictions in space are more accurate as long as we can correctly characterize the epidemic dynamics in time.*

Keywords. *Compartment Model; Spatio-temporal Modeling; Point Process; Cox Process.*

1. Introduction

Infectious diseases may have a huge impact on individuals' lives and put enormous pressure on healthcare systems globally. In this sense, one common approach to describe such diseases dynamics in time is the `SIR` compartment model, where individuals are assigned to one of the following three groups: Susceptible, Infected, or Recovered. However, we may also be interested in studying how these infectious individuals are distributed in space (and time), and that is our focus with this work.

2. Methodology

Aiming to integrate a compartment model with a log-Gaussian Cox process modeling approach, we will divide our work into two steps, namely (1) temporal modeling, and (2) spatio-temporal modeling. In the first stage, we will use a compartment model to describe the dynamics of the infectious individuals in the population, and in the second stage, using information acquired in the previous step, we will study the intensity of the infectious group over space for each of the previously discretized time points. In this section, we will detail these two steps and state the corresponding models.

2.1 Temporal Modeling

Firstly, for $t \in \mathcal{T} \subset \mathbb{R}_{\geq 0} = \{x \in \mathbb{R} : x \geq 0\}$, let $\mathtt{S}(t)$, $\mathtt{I}(t)$, and $\mathtt{R}(t)$ denote the number of Susceptible, Infected, and Recovered individuals, respectively, in a region $\mathcal{U}$ at time t. Also, let $\mathtt{S}(t) + \mathtt{I}(t) + \mathtt{R}(t) = \mathtt{N}(t) = \mathtt{N}$, $\forall t$; that is, the population size is kept constant over time. Under this setting, we will use the `SIR` model [3] to describe the infectious disease dynamics over time, in particular, we will set

$$d\mathtt{S}(t)/dt = -\beta \mathtt{S}(t)\mathtt{I}(t) \quad d\mathtt{I}(t)/dt = \beta \mathtt{S}(t)\mathtt{I}(t) - \gamma \mathtt{I}(t) \quad d\mathtt{R}(t)/dt = \gamma \mathtt{I}(t), \tag{1}$$

such that $\beta > 0$ and $\gamma > 0$. Under initial conditions ($\mathtt{S}(0)$, $\mathtt{I}(0)$, $\mathtt{R}(0)$), such system can be numerically solved, and the β and γ parameters can be estimated in different forms; for instance, we can introduce a sampling distribution for the observed number of infectious individuals in such a way that

$$\mathtt{I}_{\mathrm{OBS}}(t) \sim \text{Negative Binomial}(\mathtt{I}_{\mathrm{ODE}}(t), \phi),$$

where $\mathtt{I}_{\mathrm{ODE}}(t)$ is the solution of the system of Ordinary Differential Equations (ODEs) in (1), and ϕ is an overdispersion parameter. Thus, given some data $\mathcal{Y}$, estimating $\theta = \{\beta, \gamma, \phi\}$ can be done by maximizing the log-likelihood $\log(\mathcal{L}(\theta; \mathcal{Y}))$ function; alternatively, under the Bayesian framework, we can estimate θ by sampling from the posterior distribution $p(\theta|\mathcal{Y})$, as in [2].

2.2 Spatio-temporal Modeling

Secondly, for a partition of $\mathcal{T} = [0, \mathrm{T}]$ given by $\{t_k : k = 0, 1, \cdots, n\}$, and $\mathbf{u} \in \mathcal{U} \subset \mathbb{R}^2$, let $\xi(t_k)$ be a *spatial point process* defined in $\mathcal{U}$ at $t = t_k$. In particular, let $\xi(t_k)$ denote the locations of infectious individuals in the region of interest for a point in time. For such spatial point process, we may define an *intensity function* $\lambda(\mathbf{u}; t_k)$, such that $\int_A \lambda(\mathbf{u}; t_k) d\mathbf{u} < +\infty$, for all bounded $A \subset \mathcal{U}$, in the following way

$$\mathbb{E}[\mathcal{N}(A; t_k)] = \int_A \lambda(\mathbf{u}; t_k) d\mathbf{u},$$

where $\mathcal{N}(A; t_k)$ is the random variable that counts the number of events in A at $t = t_k$. Also, $\xi(t_k)$ is a *Poisson point process* if $\mathcal{N}(A; t_k) \sim \text{Poisson}\,(\int_A \lambda(\mathbf{u}; t_k) d\mathbf{u})$ and, conditional on $\mathcal{N}(A; t_k) = n$, the points in $\xi(t_k) \cap A$ are independent and identically distributed (i.i.d.) with density proportional to $\lambda(\mathbf{u}; t_k)$ [1].

However, since the Poisson point process may not correctly describe our phenomena of interest, we will model the events related to the infectious individuals as a log-Gaussian Cox process; that is, for a non-negative valued stochastic process $\Lambda(\mathbf{u}; t_k)$, $\xi(t_k)$ is said to be a Cox process if, conditional on $\Lambda(\mathbf{u}; t_k) = \lambda(\mathbf{u}; t_k)$, $\xi(t_k)$ is a Poisson process with intensity $\lambda(\mathbf{u}; t_k)$. In particular, for a log-Gaussian Cox process, we will have

$$\Lambda(\mathbf{u}; t_k) = \exp\{\mu(\mathbf{u}; t_k) + \zeta(\mathbf{u}; t_k)\}, \tag{2}$$

where $\zeta(\mathbf{u};t_k)$ is a Gaussian Process with variance σ^2, correlation function $\text{Corr}(\zeta(\mathbf{u}_1;t_k),\zeta(\mathbf{u}_2;t_k)) = \rho(h;t_k)$, such that $h = ||\mathbf{u}_1 - \mathbf{u}_2||$, and constant mean function given by $-\sigma^2/2$.

From Equation (2), we will set the mean term $\mathbb{E}(\Lambda(\mathbf{u};t_k)) = \exp\{\mu(\mathbf{u};t_k)\}$ as $\lambda_0(\mathbf{u};t_k)\cdot \mathtt{I}(t_k)/|\mathcal{U}|$, where $\lambda_0(\mathbf{u};t_k)$ represents the population at risk (in our case, it will be constant over time), such that $\int_{\mathcal{U}}\lambda_0(\mathbf{u};t_k)d\mathbf{u} = 1$, and $\mathtt{I}(t_k)$ is the estimated (or predicted) number of infectious individuals obtained as described in Subsection 2.1.

In that way, and under a Bayesian framework, we can estimate the intensity function for the infectious individuals in $\mathcal{U}$ and all $\{t_k\}$ by fitting the following model

$$\begin{aligned}
\xi(t_k)|\Lambda(\mathbf{u};t_k) = \lambda(\mathbf{u};t_k) &\sim \text{Poisson}\left(\int_{\mathcal{U}}\lambda(\mathbf{u};t_k)d\mathbf{u}\right), \text{ for } k = 0,1,\cdots,n\\
\Lambda(\mathbf{u};t_k) &= \lambda_0(\mathbf{u};t_k)\cdot \mathtt{I}(t_k)/|\mathcal{U}|\cdot\exp\{\zeta(\mathbf{u};t_k)\}\\
\zeta(\mathbf{u};t_k|\sigma^2,\eta) &\sim \text{Gaussian Process}(-\sigma^2/2, \sigma^2\rho(h;t_k|\eta))\\
\sigma^2,\eta &\sim \text{priors.}
\end{aligned}$$

3. Results

As a way to validate our model, we will use a mix of real and synthetic data. In particular, for a region of approximately 3 km^2 in São Paulo city, Brazil (Figure 1), we will use the estimated values [5] for the population size defined in each of the (approx.) 100 $\times$ 100 m cells (with 39,040 individuals in total) as a way to mimic the real intensity function that describes how infectious individuals are distributed over space. For a simulated epidemic dynamics sampled in $\{t_k : k = 0,1,\cdots,100\}$ and obtained from Model (1) with a set of chosen parameters, the true intensity function will be drawn from the following scheme

$$\begin{aligned}
\Lambda(\mathbf{u};t_k) &= \mathtt{pop}(\mathbf{u})\cdot \mathtt{I}(t_k)/|\mathcal{U}|\cdot\exp\{\zeta(\mathbf{u};t_k)\}, \text{ for } k = 0,1,\cdots,n \qquad (3)\\
\zeta(\mathbf{u};t_k) &= -\sigma^2/2 + \vartheta(\mathbf{u};t_k) + u(t_k),
\end{aligned}$$

such that $\mathtt{pop}(\mathbf{u}) = \mathtt{pop}(\mathbf{u};t_k)$, $\forall t_k$, is given by the normalized populational grid, $u(t_k) = u(\mathbf{u};t_k)$, $\forall \mathbf{u}\in\mathcal{U}$, is a zero-mean temporally independent Gaussian process with variance σ_u^2, and $\vartheta(\mathbf{u};t_k) = \delta(\mathbf{u};t_k)$, if $k = 0$, and $\vartheta(\mathbf{u};t_k) = a\vartheta(\mathbf{u};t_{k-1}) + \delta(\mathbf{u};t_k)$, if $k\in\{1,\cdots,n\}$, where $\delta(\mathbf{u};t_k)\sim$ Gaussian Process$(0,\sigma_\delta^2\rho(h;t_k))$ and $\rho(h;t_k)$ is given by the Matérn model. Also, as before, σ^2 is the variance of $\zeta(\mathbf{u};t_k)$. Note that, in this case, we are modifying the populational grid by incorporating a noise spatio-temporal structure into it.

3.1 Model Fitting

For the previously described data set, we first fitted the temporal model as described in Subsection 2.1. To

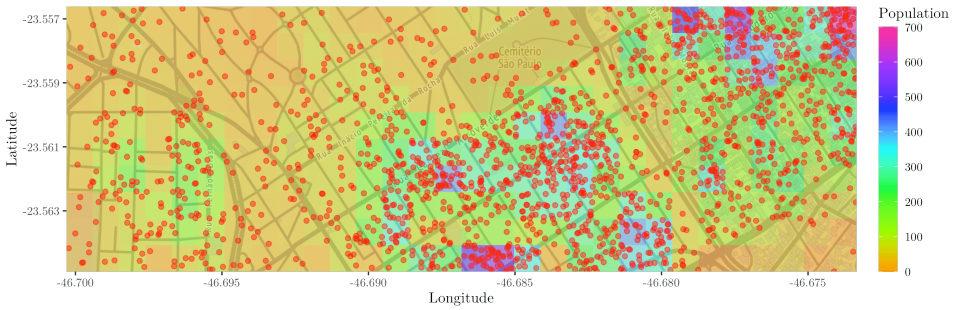


Figure 1: $\mathcal{U}$ with an overlapped grid for the estimated population. Red points are the infectious locations.

do so, we used `RStan`, and the obtained $\texttt{I}(t_k)$ curve can be seen in Figure 2 (LEFT PANEL). Notice that, in that case, we only observed data up to $t_k = 50$, and to make prediction for the remaining points, all we had to do is solving Equations (1) for $\texttt{I}(t_k)$ using the posterior sample of the estimated parameters. Then, based on the obtained $\texttt{I}(t_k)$, we can continue with the spatio-temporal modeling procedure.

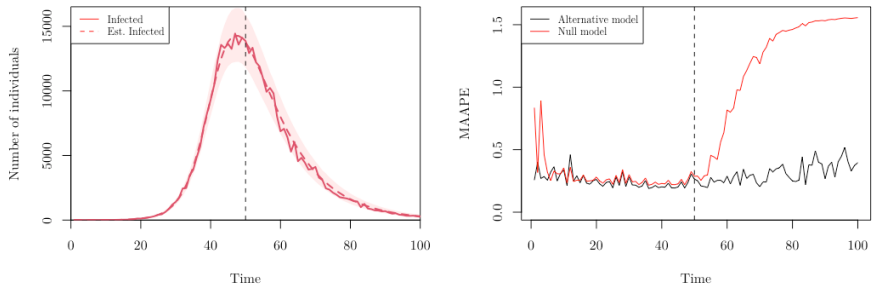


Figure 2: LEFT PANEL: Estimated $\texttt{I}(t_k)$ curve with its 95% prediction interval (shaded area). RIGHT PANEL: MAAPE for the true and estimated intensity function using both `NULL` and `ALTERNATIVE` models. For both figures, the corresponding models were fitted with data up to $t_k = 50$.

Now, under the same setting as in Equations (3), we can fit the model introduced in Subsection 2.2. To do this, we used `R-INLA`. Recall that we are using $\texttt{I}(t_k)$ in the mean component of the log-Gaussian Cox process, and therefore, we should expect better results when comparing it with a model in which

$$\Lambda(\mathbf{u};t_k) = \beta_0 \cdot \exp\{\zeta(\mathbf{u};t_k)\},$$

where β_0 is a parameter to be estimated (call it "`NULL` model", as oposed to our "`ALTERNATIVE` model").

Now, to assess the results, we can compute an error measure for the difference between the true and estimated intensity functions. To do this, we will use the Mean Arctangent Absolute Percentage Error (MAAPE) [4]. In

particular, we want to compute

$$\text{MAAPE} = \frac{1}{n}\sum_{i=1}^{n} \arctan\left(\left|\frac{f_i - \hat{f}_i}{f_i}\right|\right),$$

such that f and $\hat{f}$ are the true and estimated functions, respectively. Figure 2 (RIGHT PANEL) shows the results for the two fitted models, namely `NULL` and `ALTERNATIVE`.

From this figure, we can see that our model performs much better (with respect to the proposed error measure) than the `NULL` model, especially when making predictions. This means that including the `SIR` model output into the mean component of the log-Gaussian Cox process may drastically improve the quality of the obtained intensity function for the spatial distribution of the infectious individuals for all t.

References

[1] Diggle, P. J. (2013). Statistical Analysis of Spatial and Spatio-Temporal Point Patterns, Chapman & Hall Monographs on Statistics & Applied Probability, third edn, CRC Press, Boca Raton, Florida.

[2] Grinsztajn, L., Semenova, E., Margossian, C. C. and Riou, J. (2021). Bayesian workflow for disease transmission modeling in Stan. Statistics in medicine *40*, 6209–6234.

[3] Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character *115*, 700–721.

[4] Kim, S. and Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. International Journal of Forecasting *32*, 669–679.

[5] WorldPop (2020). Population Counts. `https://www.worldpop.org/geodata/listing?id=78`.

Local test of random labelling for functional marked point processes

N. D'Angelo[1,*] , G. Adelfio[1] , J. Mateu[2] , and O. Cronie[3]

[1]*Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy;* [*]*nicoletta.dangelo@unipa.it, giada.adelfio@unipa.it*
[2]*Department of Mathematics, University Jaume I, Castellon, Spain; mateu@uji.es;*
[3]*Department of Mathematical Sciences, Chalmers University of Technology, University of Gothenburg, Gothenburg, Sweden; ottmar.cronie@gu.se.*
[*]*Corresponding author*

Abstract. *We introduce the local t-weighted marked nth-order inhomogeneous K-function, in a Functional Marked Point Processes framework. We employ the proposed summary statistics to run a local test of random labelling, useful to identify points, and consequently regions, where this assumption does not hold, i.e. the functional marks are spatially dependent.*

Keywords. *Spatio-temporal point process; Local features; K-function; Random labelling; Envelopes*

1. Introduction

Despite of the relatively long history of point process theory, few approaches have been performed to analyse spatial point patterns where the features of interest are functions (i.e. curves) instead of qualitative or quantitative variables. Examples of point patterns with associated functional data include: forest patterns where for each tree we have a growth function, curves representing the incidence of an epidemic over a period of time, and the evolution of distinct economic parameters such as unemployment and price rates, all for distinct spatial locations. [2] introduced a very broad framework for the analysis of Functional Marked Point Processes (FMPPs), indicating how they connect the point process theory with both Functional Data Analysis (FDA) and geostatistics. In their work, they defined a new family of summary statistics, so-called *t-weighted nth-order marked inhomogeneous K-function*, together with their nonparametric estimators, to analyse Spanish population structures, such as demographic evolution and sex ratio over time. This summary statistic can be used to run a test of random labelling by means of the global envelopes test (GET, [3]), to assess whether the functional marks of the analysed pattern are spatially dependent. However, this procedure is essentially global, since it does not provides information on the points which mostly contributed to the rejection of the random labelling hypothesis. Therefore, motivated by the will of detecting those points, and therefore regions, where the functional marks really do depend on the spatial locations, in this paper, we introduce the *local t-weighted marked nth-order inhomogeneous K-functions*, and use them for proposing a *local test of random labelling*.
The structure of the paper is as follows. Section 2. gives notation of functional marked point processes. The local nth-order summary statistics and the local test of random labelling are proposed in Section 3. In Section

4., by an application to simulated data, we show that our proposal succeeds in identifying points (and thus regions) where the functional marks are spatially dependent. The conclusions are drawn in Section 5.

2. Functional marked point processes

We consider marked point processes $\Psi = \{(x_i, m_i)\}_{i=1}^{N}$ in the sense of [1] (Definition 6.4.1), with ground points x_i in $\mathbb{R}^d$, which is equipped with the Euclidean metric and the Lebesgue measure $l(A) = \int_A \mathrm{d}z$ for Borel sets $A \in \mathcal{B}(\mathbb{R}^d)$; a closed r-ball around $x \in \mathbb{R}^d$ will be denoted by $b[x,r]$. The ground process Ψ_g, obtained from the marginal Ψ w.r.t. the marks, is by definition a well-defined point process on $\mathbb{R}^d$ in its own right. We shall assume that Ψ is simple, that is, almost surely (a.s.) does not contain multiple points. We assume that the mark space $\mathcal{M}$ is Polish and equipped with a finite reference measure ν on the Borel σ-algebra $\mathcal{B}(\mathcal{M})$. The Borel σ-algebra $\mathcal{B}(\mathbb{R}^d \times \mathcal{M}) = \mathcal{B}(\mathbb{R}^d) \otimes \mathcal{B}(\mathcal{M})$ is endowed with the product measure $A \times E \mapsto l(A)\nu(E)$, $A \times E \in \mathcal{B}(\mathbb{R}^d \times \mathcal{M})$. Given this general setup, one may obtain various forms of marked point processes, most notably multivariate/multitype point processes with $\mathcal{M} = \{1, \ldots, k\}$ and FMPP with $\mathcal{M}$ given by a suitable function space.
In FDA, one analyses collections of functions $\{f_1(t), \ldots, f_n(t)\}$, $t \in \mathcal{T} \subset [0, \infty)$, $n \geq 1$, in some Euclidean space $\mathbb{R}^k$, $k \geq 1$, which belong to some suitable family of functions (usually an L_2-space); note that the argument t does not need to represent time, but it could represent spatial distance. Classically, it has been assumed that these functions are realisations/sample paths of some collection of independent and identically distributed (iid) random functions/stochastic processes $\{F_1(t), \ldots, F_n(t)\}$, $t \in \mathcal{T}$. However, it is often the case that these functions have some sort of spatial dependence. For example, two functions f_i and f_j, with starting points $f_i(0)$ and $f_j(0)$ which are spatially close to each other in $\mathbb{R}^k$, either gain or lose from being close to each other. Accordingly, it seems natural to generate $F_1, \ldots, F_N$ conditionally on some collection of (dependent) random spatial locations. Moreover, these random functions may not be iid. More specifically, it makes sense to describe the collection of locations and functions as a FMPP, which is defined as a marked point process where the marks are random elements in some (Polish) function space, most notably the L_2 space of functions $f : \mathcal{T} \to \mathbb{R}^k$.

3. Local nth-order summary statistics and test for random labelling

Assume we observe a FMPP Ψ within a bounded spatial domain $W \in \mathcal{B}(\mathbb{R}^d), l(W) > 0$, i.e. $\Psi \cap W \times \mathcal{M}$. We define the *local t-weighted marked nth-order inhomogeneous K-function* for the ith point (x,m) as

$$
\begin{aligned}
L^{(i)} = \hat{\mathcal{K}}_{\mathfrak{t}}^{(x,m)\times\prod_{i=1}^{n-1}E_i}(r) = & \sum_{(x_1,m_1),\ldots,(x_{n-1},m_{n-1})\in\Psi\setminus\{(x,m)\}} \frac{w(x,x_1,\ldots,x_{n-1})}{\nu_{\mathcal{M}}(E)\prod_{i=1}^{n-1}\nu_{\mathcal{M}}(E_i)} \times \\
& \times t(m,m_1,\ldots,m_{n-1}) \frac{I\{(x,m)\in W\times E\}}{\hat{\rho}_g(x)} \prod_{i=1}^{n-1} \frac{I\{x_i \in (W\cap b[x,r]\}I\{m_i\in E_i\}}{\hat{\rho}_g(x_i)}.
\end{aligned}
\tag{1}
$$

with $r \geq 0$, $w(\cdot)$ an edge correction term, and $\hat{\rho}_g(x)$ an estimator of the ground intensity $\rho_g(x)$.

It can be prove that its expectation is

$$\mathcal{K}_{\mathfrak{t}}^{(x,m)\times\prod_{i=1}^{n-1}E_i}(r) = \frac{1}{l(W)\nu_{\mathcal{M}}(E)\prod_{i=1}^{n-1}\nu_{\mathcal{M}}(E_i)}\times$$
$$\times\mathbb{E}\left[\sum_{(x_1,m_1),\ldots,(x_{n-1},m_{n-1})\in\Psi\setminus\{(x,m)\}}\frac{t(m,m_1,\ldots,m_{n-1})}{\rho(x,m)}\prod_{i=1}^{n-1}\frac{I\{x_i\in b[x,r]\}I\{m_i\in E_i\}}{\rho(x_i,m_i)}\right].$$

Simple hypotheses for spatial point patterns (such as CSR) are commonly tested using an estimator of a global summary statistic, e.g., the Ripley's K-function. In this context, one typically resorts to the Monte Carlo simulation. The first step is then to generate Q simulations under the null hypothesis, and to calculate the chosen summary statistics for both the observed pattern and the simulations. In order to study whether there is *random labelling* in a FMPP Ψ, the simulations are obtained by permuting the functional marks, that is, randomly assigning them to the spatial points of the ground pattern Ψ_g, which are kept fixed. Then, the chosen summary statistic, is computed for each of these permutations, and global envelopes at a given nominal level are generated based on them. The result of the test can be assessed graphically: if the summary statistic of the observed pattern goes outside the envelopes, we proceed with the assumption that the functional marks of the analysed pattern are indeed not randomly labelled. Furthermore, it is possible to calculate a point estimate for the p-value based on the position of the observed summary statistic within the qth envelopes, following [3]. We know however, that the conclusion drawn from the application of the above-mentioned global test is referred to the whole analysed process, indicating whether all the functional marks of the analysed pattern are randomly labelled or not. Motivated by the will of further detecting the specific points, and regions, where the functional marks really do depend on the spatial locations and dependencies, we propose a *local test of random labelling*. The main idea is to run a global envelope test on each point of the analysed pattern by means of the previously proposed *local t-weighted marked inhomogeneous K-functions*, to obtain p-values and to draw different conclusions on the individual points.
We next outline the proposed local test:

- Set a number of simulations Q;
- For each $q = 1,\ldots,Q$:
  - Randomly sample n functional marks, from the original n ones;
  - Compute the local summary statistic, say $L_q^{(i)}$ in (1) ;
- For each point x_i of ground pattern Ψ_g, run the global test of random labelling, where the qth envelopes are given by the all the local summary statistics of the ith point, computed over the Q permutations $\{L_q^{(i)}\}_{q=1}^{Q}$.

The testing procedure ends with providing a p-value p_i for each point of the analysed FMPP Ψ.
The null hypothesis is rejected for the i^{th} point if $p_i \leq \alpha$, with α the fixed nominal value of type I error.

4. Application to simulated spatially dependent functional marks

We simulate a homogeneous spatial point pattern with 250 points on the unit square. This represents the ground pattern Ψ_g. For each point x_i, we simulate a functional mark from $y_t = \mu_t + \varepsilon_t$, with $\varepsilon_t \sim \mathcal{N}(0,\sigma_t^2)$, where

σ_t^2 is a variance function approximated by a piecewise constant regression function with K_0+1 segments. We set $K_0=2$, $\mathcal{T}=500$ equispaced observations $z_t=t\in[0;1]$ and $\sigma_t^2=0.2+7.5I(t>0.4)-5I(t>0.6)$. The mean signal μ is just taken equal to zero. To make the functional marks spatially dependent, we then superimpose a homogeneous spatial point pattern with 50 points, generated in the $[0,0.5]\times[0,0.5]$ square, i.e. the bottom left region of the entire analysed window. For these additional points, we generate different functional marks than before, namely with underlying trend $\mu_t=10+6\sin(3\pi z_t)$. Therefore, we have simulated a FMPP Ψ with spatially dependent functional marks, i.e. not random labelled. We therefore expect a global test of random labelling to confirm it.

We first run a global test of random labelling, by means of the t-weighted nth-order marked inhomogeneous K-function of [2], with $n=2$, making it a second-order summary statistic. As test function $t(\cdot)$, we consider the functional marked counterpart of the test function for the classical variogram $t(f_1,f_2)=t_v(f_1,f_2)=\int_a^b(f_1(t)-\bar{F}(t))(f_2(t)-\bar{F}(t))\mathrm{d}t$, with $\bar{F}(t)=(1/n)\sum_{i=1}^n f_i(t)$, that is the average functional mark at time t for the observed functional part of the point pattern. We run $Q=39$ permutations, and obtain a global p-value of 0.025. This, together with the observed K-function lying outside the envelopes (left panel of Figure 1), indicates the ability of the global test to correctly detect the spatial dependence of the functional marks. We know, however, that this conclusion show not be drawn for each point of the pattern, but specifically for those in the $[0,0.5]\times[0,0.5]$ square.

We therefore proceed by running our proposed local test, based on the proposed local K-functions (1) in their second-order version, and with the same choice for the test function $t(\cdot)$ as the global one. The right panel of Figure 1 depicts the points of the simulated point pattern, and displays in pink those for which the local test resulted significant. From this, we know that the proposed local test is able to correctly identify some of the points, and then the region, where the hypothesis of random labelling does not hold.

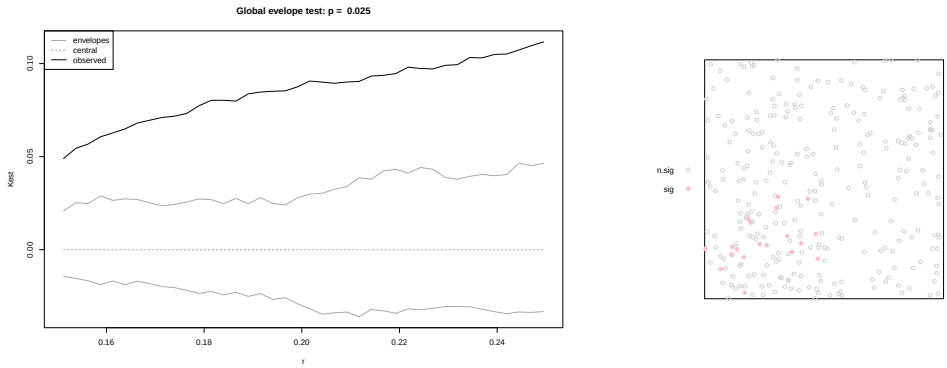


Figure 1: *Left panel*: Result of the global envelopes test; *Right panel*: The simulated point pattern. Significant points for which the hypothesis of random labelling is rejected are in pink.

5. Conclusions and future directions

In this work, we have proposed the local t-weighted marked nth-order inhomogeneous K-functions for spatial point processes with functional marks. We have employed them to construct a local test for random

labelling, to identify points, as well as regions, where this hypothesis does not hold.
In future, we plan to run an extensive simulation study to assess the performance of the test under different scenarios in reference to both the ground pattern (random, clustered, regular) and the functional marks (which could differ in mean, variance, and correlation structure).
Finally, we aim at analysing seismic data. Indeed, while spatial (and temporal) location of the epicenter of the earthquake is typically studied within the theory of point processes, the seismic waveforms are commonly investigated in separate analyses through FDA. Applying the local test would allow to identify where one would expect waveforms (i.e. functional marks) to be similar to those of close points or not.

References

[1] Daley, D. J. and Vere-Jones, D. (2007). An Introduction to the Theory of Point Processes. Volume II: General Theory and Structure. Springer-Verlag, New York, second edition.

[2] Ghorbani, M., Cronie, O., Mateu, J., and Yu, J. (2021). Functional marked point processes: a natural structure to unify spatio-temporal frameworks and to analyse dependent functional data. *Test*, **30**(3):529568.

[3] Myllymki, M., Mrkvika, T., Grabarnik, P., Seijo, H., and Hahn, U. (2017). Global envelope tests for spatial processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **79**(2):381404.

The influence of gas production on seismicity in the Groningen field

Z. Baki[1,2,*] and M.N.M. van Lieshout[2,1]

[1]*Department of Applied Mathematics, University of Twente, P.O. Box 217, NL-7500 AE Enschede, The Netherlands; Z.Baki@utwente.nl*
[2]*CWI, P.O. Box 94079, NL-1090 GB, Amsterdam, The Netherlands; M.N.M.van.Lieshout@cwi.nl*
[*]*Corresponding author*

Abstract. *In this paper we investigate the effect of gas production volumes on seismicity in the Groningen field by means of a log-linear Poisson model. First, we consider the annual counts and then refine to the temporal point pattern of earthquake occurrence times.*

Keywords. *Count data; Gas extraction; Induced seismicity; Log-linear Poisson model; Point pattern.*

1. Introduction

Discovered in 1959, the Groningen gas field, the largest gas field in Europe with an estimated recoverable gas volume of around $2,900$ billion cubic meters (bcm), has been a massive boost to the Dutch economy. Production in Groningen started in 1963, initially only to accommodate the high demand for gas during the winter months [8]. However, the closure of smaller fields in the country led to an increase in production. By 2012, annual production volumes had climbed to over 40 bcm per year [6].

Increasing production volumes and the resulting depletion of the gas field have led to induced earthquakes in the previously tectonically inactive Northern Netherlands. Depletion causes a decrease of the gas pressure, which causes compaction of the gas reservoir, noticeable by subsidence. Additionally, a drop in gas pressure increases stress in the faults of the region. Due to the increased stress, faults slip and cause seismicity [10]. The most significant event to date, in August 2012 near Huizinge with a magnitude of 3.6, attracted massive public attention, prompting the Ministry of Economic Affairs to reduce production volumes.

Numerous studies have been conducted, of which we mention a few. For example, [5] models the times in between earthquakes in terms of the cumulative and annual production rates, pressure, subsidence and fault zones. A more recent example of such a study is [12]. Van Hove *et al.* [7] propose a Poisson auto-regression model for the annual hazard maps in terms of subsidence, fault lines and gas extraction in previous years. Sijacic *et al.* [14] focus on the detection of changes in the rate of a temporal Poisson point process by Bayesian and frequentist methods. Moreover, [1] modify Ogata's space-time model [11] to include changes in stress level and estimate the probability of fault failures. Other papers [2, 3, 13] discuss the modelling of seismicity in relation to stress changes based on a differential equation. Both [6] and [15] explore the temporal development

of seismicity in Groningen by proposing a linear model for the relation between the number of earthquakes over specific periods and gas production volumes. In this paper, we take a similar approach towards the temporal development of seismicity in Groningen including data up to 2020. Details on the methodology used can be found, e.g., in [9].

2. Data

An earthquake catalogue for The Netherlands is being maintained by the Royal Dutch Meteorological Office (KNMI) at `www.knmi.nl/kennis-en-datacentrum/dataset/aardbevingscatalogus`. Data on the period before 1995 are not reliable due to the inaccuracy of the equipment used. Moreover, a threshold on the magnitude is necessary to guarantee data quality. According to [4], earthquakes with a magnitude of 1.5 or larger can be reliably recorded. We therefore restrict ourselves to induced earthquakes with a magnitude of 1.5 or higher between January 1st, 1995, and December 31st, 2020. The resulting pattern of occurrence times consists of 322 earthquakes.

Monthly production figures are available at the site of the Dutch Oil Company (NAM) at `https://www.nam.nl/feiten-en-cijfers/gaswinning.html`, both for the gas field as a whole and per individual well. The figures are published in cubic meters, which we re-scale to bcm. Since we focus on the temporal aspects, we use only the cumulative numbers over the entire field.

3. Annual counts

To explore the relationship between seismicity and gas production, Figure 1 plots the total annual production in bcm for the years $1994, \ldots, 2019$ and the annual number of earthquakes in the next year against time. A dip in the produced volumes is observed during the late 1990s, followed by a steep increase. After 2014 the volumes decrease following government regulations. As for the number of earthquakes, there seems to be a general upwards trend up to about 2013. From 2014 onward, the frequency of earthquakes also tends to decrease.

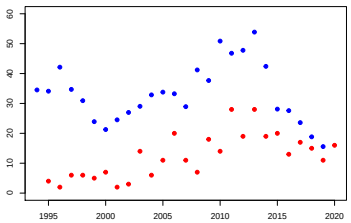


Figure 1: Plots of the annual number of earthquakes (1995–2020) in the Groningen gas field (red dots) and annual gas production volumes (1994–2019, blue dots, in bcm) against time.

A statistical model that captures these aspects is the following. Assume that the number of earthquakes $N(t)$ in year $t \in \{1995,\ldots,2020\}$ is Poisson distributed with intensity parameter $\lambda(t)$ such that

$$\log\lambda(t) = \alpha_1 + \alpha_2 C(t-1) + \alpha_3 \log(\tilde{C}(t-1)). \tag{1}$$

Here $C(t-1)$ denotes the gas produced in year $t-1$, $\tilde{C}(t-1)$ is the cumulative gas production up to year $t-1$. By maximizing the likelihood function we obtained the parameter estimates $\hat{\alpha}_1 = -2.23$, $\hat{\alpha}_2 = 0.025$ and $\hat{\alpha}_3 = 0.64$. We are especially interested in α_2 as it quantifies the effect of the gas production in each consecutive year. Its asymptotic approximate 95% confidence interval is $(0.015, 0.035)$, from which we conclude that an increase in production leads to increased seismicity.

To validate the model, Figure 2 shows the Pearson residuals (left-most panel) and the empirical inhomogeneous auto-correlation function (central panel). We conclude that the model fits reasonably well. The estimated and predicted number of earthquakes are shown in the right-most panel. For $\lambda(2021)$, an approximate 95% confidence interval is $(7.16, 13.69)$. The actual number was 12. For comparison [6] predicted a 16 ± 8 events in 2016. Our prediction for 2016 is tighter, 15.67 ± 2.61. In reality, there were 13 earthquakes.

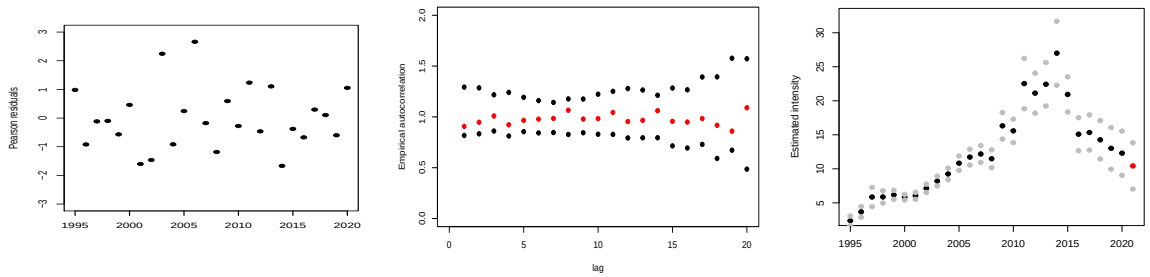


Figure 2: Left: Pearson residuals plotted against time. Central panel: empirical inhomogeneous auto-correlation function for lags $h = 1,\ldots,15$ and local envelopes based on 19 simulations from the fitted model. Right: estimated number of earthquakes (black dots) and predicted number (red dot) for 2021 with associated approximate 95% confidence intervals (grey dots).

4. Temporal point pattern

So far, we used aggregated count data. Since the earthquake times are being recorded, we may also consider a temporal Poisson point process model. Taking days as our unit of time, suppose that the intensity function λ satisfies

$$\log\lambda(t) = \alpha_1 + \alpha_2 C(t,12) + \alpha_3 \log(\tilde{C}(t)). \tag{2}$$

for $t \in (0, 9497]$. Here $C(t,12)$ denotes the amount of gas produced over the twelve months preceding time $t \in (0, 9497]$ and $\tilde{C}(t)$ is the cumulative amount produced from 1994 and preceding time t. The maximum likelihood estimates are $\hat{\alpha}_1 = -8.56$, $\hat{\alpha}_2 = 0.023$ and $\hat{\alpha}_3 = 0.72$.

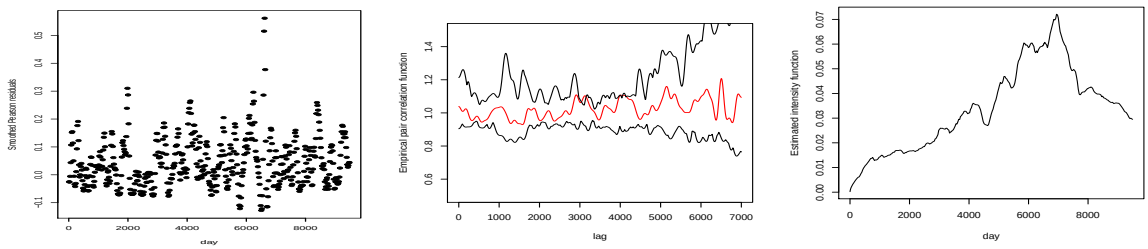


Figure 3: Left: smoothed Pearson residuals using a Gaussian kernel with $\sigma = 30$. Central panel: empirical inhomogeneous pair correlation function for lags $h = 1, \ldots, 7000$ and local envelopes based on 19 simulations from the fitted model. Right: estimated intensity function.

Figure 2 shows the smoothed Pearson residuals (left-most panel) and the empirical inhomogeneous pair correlation function (central panel). We conclude that the model fits reasonably well. The estimated intensity function is shown in Figure 3. Compared to Figure 2, it is more detailed, but the general interpretation is similar. The approximate 95% confidence interval for α_2 is $(0.010, 0.036)$. Since it does not contain 0, $C(t, 12)$ is significant. The earthquake hazard $\lambda(t)$ at the dawn of the new year, January 1st, 2021, based on the gas production $C(9498, 12) = 7.95$ in 2020, has an approximate confidence interval $(0.018, 0.0413)$. The first induced earthquake with a magnitude of at least 1.5 happened on January 24, 2021, near Tjuchem.

Acknowledgments

This research was funded by the Dutch Research Council NWO (Deep.NL.2018.033). We thank R.L. Markwitz and H. Paulssen for useful comments.

References

[1] Bourne, S.J., Oates, S.J. & Van Elk, J., 2018. The exponential rise of induced seismicity with increasing stress levels in the Groningen gas field and its implications for controlling seismic risk. *Geophysical Journal International* **213**:1693–1700.

[2] Candela T. et al., 2019. Depletion-induced seismicity at the Groningen gas field: Coulomb rate-and-state models including differential compaction effect. *Journal of Geophysical Research: Solid Earth* **124**:7081–7104.

[3] Dempsey, D. & Suckale, J., 2017. Physics-based forecasting of induced seismicity at Groningen gas field, the Netherlands. *Geophysical Research Letters* **44**:7773-7782.

[4] Dost, B. et al., 2012. Monitoring induced seismicity in the North of the Netherlands: status report 2010. Scientific Report KNMI, WR 2012–03.

[5] Geerdink, E., 2014. Modeling the induced earthquakes in Groningen as a Poisson process using GLM and GAM. BSc thesis, University of Groningen.

[6] Hettema, M.H.H. et al., 2017. An empirical relationship for the seismic activity rate of the Groningen gas field. *Netherlands Journal of Geosciences* **96**:149–151.

[7] Hove, E. van, Van Lingen, R. & Riemens, S., 2015. Geïnduceerde aardbevingen in gasveld Groningen. Een statistische analyse (in Dutch). BSc thesis, University of Twente.

[8] Jager, J. de & Visser, C., 2017. Geology of the Groningen field – an overview. *Netherlands Journal of Geosciences* **96**:3–15.

[9] Lieshout, M.N.M. van, 2019. Theory of spatial statistics. A concise introduction. Chapman and Hall/CRC Press (Boca Raton).

[10] Nepveu, M., Van Thienen–Visser, K. & Sijacic, D., 2016. Statistics of seismic events at the Groningen field. *Bulletin of Earthquake Engineering* **14**:3343–3362.

[11] Ogata, Y., 1988. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association* **83–401**:9–27.

[12] Post, R.A.J. et al., 2021. Interevent-time distribution and aftershock frequency in non-stationary induced seismicity. *Scientific Reports* **11**:3540.

[13] Richter, G. et al., 2020. Stress-based, statistical modeling of the induced seismicity at the Groningen gas field, The Netherlands. *Environmental Earth Sciences* **79**:252.

[14] Sijacic, D. et al., 2017. Statistical evidence on the effect of production changes on induced seismicity. *Netherlands Journal of Geosciences* **96**:27–38.

[15] Vlek, C., 2019. Rise and reduction of induced earthquakes in the Groningen gas field, 1991–2018: statistical trends, social impacts, and policy change. *Environmental Earth Sciences* **78**:59.

A spatial analysis of sex differences in chess expertise across 24 countries in Europe

A. Blanch[1,*] and C. Comas[2]

[1]*Department of Psychology, University of Lleida, Spain; angel.blanch@udl.cat*
[2]*Department of Mathematics, University of Lleida, Spain; carles.comas@udl.cat*
[*]*Corresponding author*

Abstract. *Sex differences in chess expertise tend to be large, with males scoring higher than females in the Elo chess rating. Past reports attribute these differences either to the differential participation of males and females in chess, or to biological and cultural factors. For instance, a previous study comparing the top-hundred male and female chess players in* 24 *Eurasian countries highlights that differential sex ratios explain only in part the sex difference in chess expertise [1]. There is in addition considerable variability across countries that might depend on geographical aspects. Countries with narrower sex gaps in chess expertise predominate in Eastern Europe or have smaller territories (Georgia, Czech Republic, Romania, Hungary, Slovakia or Netherlands). On the other hand, countries with wider sex gaps in chess expertise predominate in southern Europe or have larger territories (Greece, Russia, Bulgaria, Croatia, Portugal, or France). In this study, we reanalyze the data in [1] in order to address the degree by which raw and expected (RED) sex differences followed such geographical patterns. More specifically, we evaluate whether the occurrence of RED at distinct countries can depend on adjacent countries or country size, or else this spatial structure is independent of the spatial configuration of neighboring countries (country location location or size). The findings from this study can shed additional light about the geographical distribution of sex differences in chess expertise.*

Keywords. *Chess expertise; LISA; Sex differences; Spatial analysis.*

1. Introduction

Sex differences in chess expertise tend to be large, with males scoring higher than females in the Elo chess rating. Past reports attribute these differences either to the differential participation of males and females in chess, or to biological and cultural factors. For instance, a previous study comparing the top-hundred ranked male and female chess players in 24 Eurasian countries highlights that differential sex ratios explain only in part the sex difference in chess expertise [1].

In addition, there is considerable variability across countries that might depend in geographical aspects. Countries with narrower sex gaps in chess expertise either are Eastern European countries, or have small territories (Georgia, Czech Republic, Romania, Hungary, Slovakia or Netherlands). On the other hand, countries with wider sex gaps in chess expertise either are Southern European countries, or have large territories (Greece, Russia, Bulgaria, Croatia, Portugal, or France).

In this study, we reanalyze the data in [1] in order to address the degree by which raw and expected (RED)

sex differences follow such geographical patterns. More specifically, we evaluate the spatial autocorrelation structure of these variables in terms of three distinct distance matrices and whether the occurrence of RED at adjacent countries or depending on country size can be accounted for by randomness alone. This analysis is considered bearing in mind several geographical factors: 1) the country size, 2) the proximity between countries, and 3) the proximity to the Mediterranean sea. The findings from this study can shed additional light about sex differences in chess expertise across when considering the geographical distribution across 24 countries in Europe.

2. Input data

There are two kinds of data for this study. First, there are three variables or attributes for each country: a) the raw sex differences in chess expertise (R), b) the expected sex differences considering differential Male:Female ratios by country (E), and c) the difference between the observed and estimated sex difference in chess expertise (D). Second, there are two matrices according with each of the three geographical factors of interest. Each of these matrices describes some sort of distance across the studied countries: a) country size differences and b) nearest neighbor distances.

To evaluate the spatial structure of these RED measures assuming distinct distances matrices, we consider global measures of autocorrelation (such as the Moran's I measure) and their counterpart local version (LISA functions) [2]. Our intention is to investigate the spatial configuration of these RED variables, and evaluate if their expected spatial structures, in terms of the distinct distance matrices, have a global scale of interaction, or else, they act at more local scales. Given the social an cultural differences between the analyzed countries, it is expected the presence of some local spatial dependencies affecting the spatial structure of these RED variables.

3. Expected outcomes

According with the findings in [1], non-random spatial structures are to be expected regarding Eastern Europe countries, the proximity to the Mediterranean sea, and the country size. Therefore, and according with these geographical factors, the countries meeting these conditions should be more similar and display a more consistent spatial structure regarding RED.

References

[1] Blanch, A. (2016) Expert performance of men and women: A cross-cultural study in the chess domain. *Personality and Individual Differences* **101**, 90-97.

[2] Anselin, L. (1995). Local indicators of spatial association - LISA. *Geographical Analysis* **27**, 93-115.

Exploring spatial point pattern interactions at different scales - a glimpse into Portugal active fire data

I.J.F. Correia[1,*] , S.A. Pereira[1] , T.A. Marques[1,2] , and J.M. Pereira[3]

[1]*Centro de Estatística e Aplicações Universidade de Lisboa, Faculdade de Ciências da Universidade de Lisboa, Portugal*
[2]*Centre for Research into Ecological and Environmental Modelling, University of St. Andrews, Scotland, tam2@st-andrews.ac.uk*
[3]*Instituto Superior de Agronomia, Universidade de Lisboa, Portugal*
[*]*Corresponding author*

Abstract. *This work is part of a PhD project whose main goal is to study and model point patterns that exhibit two different types of interactions at different scales. Specifically, how observations of active fires in mainland Portugal interact (clustering or regularity) in space (and/or time) for small and large distances (or intervals). The data belong to the MODIS Collection 6 Active Fire, from the Fire Information for Resource Management System (FIRMS, NASA, US). They consist in daily point detections of fire hot spots represented by the centroid of their respective* $1km^2$ *pixel from a grid for mainland Portugal in 2001. Thus, within the spatial statistics framework, these data are target to spatial point process statistical methodologies. As a starting point of this work, we are presenting a brief spatial exploratory analysis of the pattern aforementioned. Exploratory analysis usually includes functions like the inhomogeneous K-function and the inhomogeneous pair correlation function which are presented and plotted. Their estimates showed presence of both aggregation and repulsion behaviour at different scales. Some spatial point pattern models like empirical models (e.g., Geyer or Matrn-thinned Cox processes) or mechanistic models (e.g., self-exciting processes) that allow capturing this behaviour are certainly the following steps once a better understanding is achieved concerning the dynamic within the analysed data.*

Keywords. *Active Fires; Exploratory Spatial Analysis; Spatial Interactions; Spatial Point Patterns.*

References

[1] Andersen, I. T. and Hahn, U. (2016). Matrn-thinned Cox processes, *Spatial Statistics* **15**, 1–21.

[2] Baddeley, A., Rubak, E., and Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R*. London: Chapman and Hall/CRC Press.

[3] Baddeley, A., Mateu,, J., and Bevan, A. (2013). Hybrids of Gibbs point process models and their implementation, *Journal of Statistical Software* **55**, 1–43.

[4] Giglio, L., Schroeder, W., and Justice, C. O. (2016). The collection 6 MODIS active fire detection algorithm and fire products, *Remote Sensing of Envirenment* **178**, 31.41.

[5] Møller, J. and Waagepetersen, R. P. (2003). *Statistical inferernce and simulation for spatial point processes*. Chapman and Hall.

[6] Raeisi, M., Bonneu, F. and Gabriel, E. (2021). A spatio-temporal multi-scale model for Geyer saturation point processes: Application to forest fire occurences, *Spatial Statistics* **41**, 100492.

[7] González, J. A., Rodrígues-Corts, F. J., Cronie, O., and Mateu, J. (2016). Spatio-temporal point process statistics: A review, *Spatial Statistics* **18**, 505-544.

Locally weighted spatio-temporal minimum contrast for Log-Gaussian Cox Processes

N. D'Angelo[1,*] , G. Adelfio[1] and J. Mateu[2]

[1]*Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo, Italy; nicoletta.dangelo@unipa.it, giada.adelfio@unipa.it*
[2]*Department of Mathematics, University Jaume I, Castellon, Spain; mateu@uji.es*
[*]*Corresponding author*

Abstract. *We propose a local version of the spatio-temporal log-Gaussian Cox processes (LGCPs) employing the Local Indicators of Spatio-Temporal Association (LISTA) functions into the minimum contrast procedure to obtain space as well as time-varying parameters. We resort to the joint minimum contrast method fitting method to estimate the set of second-order parameters for the class of spatio-temporal LGCPs. This approach has the advantage of being usable in the case of both separable and non-separable parametric specifications of the correlation function of the underlying Gaussian Random Field (GRF).*

Keywords. *Spatio-temporal point processes; Local models; Log-Gaussian Cox Processes; Minimum contrast; Second-order characteristics.*

1. Introduction

Local extensions of spatio-temporal point process models are very welcome in many fields of study, such as epidemiology and seismology. Indeed, one could be interested in identifying the most inhomogeneous locations, both in space and time, to be further examined separately.
For spatial point process, [1] presents a general framework based on the local composite likelihood to detect and model gradual spatial variation in any parameter of a spatial stochastic model, among which the Cox processes. [2] show that their purely local models provide good inferential results by applying them to earthquake data. Motivated by this, we propose a local version of the spatio-temporal log-Gaussian Cox processes (LGCPs) employing the Local Indicators of Spatio-Temporal Association (LISTA) functions in the minimum contrast procedure to obtain space as well as time-varying parameters.

In particular, we extend the joint minimum contrast method [6] to the local context, managing to estimate a set of second-order parameters of the spatio-temporal LGCPs for each point. The joint estimation approach has the advantage of being usable in the case of both separable and non-separable parametric specifications of the correlation function of the underlying Gaussian Random Field (GRF).

The structure of the paper is as follows. In Section 2. the spatio-temporal log-Gaussian Cox Processes are recalled, as well as the joint minimum contrast procedure. Section 3. contains the proposed method to estimate local parameters, whose performance is assessed in 4. Conclusions are drawn in Section 5.

2. Spatio-temporal LGCPs and minimum contrast estimation

Following the inhomogeneous specification in [4], a log-Gaussian Cox process for a generic point with $\mathbf{u}$ and t coordinates in space and time has the intensity

$$\Lambda(\mathbf{u},t) = \lambda(\mathbf{u},t)\exp(S(\mathbf{u},t))$$

where S is a Gaussian process with $\mathbb{E}(S(\mathbf{u},t)) = \mu = -0.5\sigma^2$ and so $\mathbb{E}(\exp S(\mathbf{u},t)) = 1$ and with variance and covariance matrix $\mathbb{C}(S(\mathbf{u}_i,t_i),S(\mathbf{u}_j,t_j)) = \sigma^2\gamma(r,h)$, with $\gamma(\cdot)$ the correlation function of the GRF, and r and h some spatial and temporal distances. Following [5], the first-order product density and the pair correlation function of a log-Gaussian Cox process are $\mathbb{E}(\Lambda(\mathbf{u},t)) = \lambda(\mathbf{u},t)$ and $g(r,h) = \exp(\sigma^2\gamma(r,h))$, respectively. We consider a separable structure for the covariance function of the Gaussian Random Field that has exponential form for both the spatial and the temporal components,

$$\mathbb{C}(r,h) = \sigma^2 \exp\left(\frac{-r}{\alpha}\right)\exp\left(\frac{-h}{\beta}\right), \tag{1}$$

where σ^2 is the variance, α is the scale parameter for the spatial distance and β is the scale parameter for the temporal one. The exponential form is widely used in this context and nicely reflects the decaying correlation structure with distance or time. Moreover, we can consider a non-separable covariance of the GRF useful to describe more general situations.

In general, the Cox model is estimated by a two-step procedure, involving first the intensity and then the cluster or correlation parameters. First, a Poisson model with the same model formula is fitted to the point pattern data, providing the estimates of the coefficients of all the terms in the model formula that characterize the intensity. Second, the estimated intensity is taken as the true one and the cluster or correlation parameters are estimated by one among the method of minimum contrast, Palm likelihood, or composite likelihood. The most common technique is the *minimum contrast*.

Let the function J represent either the pair correlation function g of the K-function, and $\hat{J}$ stands for the corresponding non-parametric estimate. [6] propose a new fitting method to estimate the set of second-order parameters for the class of spatio-temporal log-Gaussian Cox point processes with constant first-order intensity function. Hereafter we will denote by $\boldsymbol{\theta}$ the vector of (first-order) intensity parameters, and by $\boldsymbol{\psi}$ the cluster parameters, also denoted as correlation or interaction parameters by some authors. In the case of a spatio-temporal log-Gaussian Cox process with exponential covariance as the one in Equation (1), the cluster parameters correspond to $\boldsymbol{\psi} = (\sigma,\alpha,\beta)$, that is found by minimizing

$$M_J\{\boldsymbol{\psi}\} = \int_{h_0}^{h_{max}}\int_{r_0}^{r_{max}} \{\hat{J}(r,h) - J(r,h;\boldsymbol{\psi})\}^2 \mathrm{d}r\mathrm{d}h. \tag{2}$$

With simulations, [6] show that the *joint minimum contrast procedure*, based on the spatio-temporal pair correlation function, provides reliable estimates. Its main advantage is that it can be used in the case of both separable and non-separable parametric specifications of the correlation function of the underlying GRF, representing a more flexible method with respect to other available methods.

3. Locally weighted spatio-temporal minimum contrast

Combining the *joint minimum contrast* [6] and the *local minimum contrast* [1] procedures, we can obtain a vector of parameters $\hat{\psi}_i$ for each point i, by minimizing

$$M_{J,i}\{\psi_i\} = \int_{h_0}^{h_{max}} \int_{r_0}^{r_{max}} \{\bar{J}_i(r,h) - J(r,h;\psi)\}^2 \mathrm{d}r\mathrm{d}h, \tag{3}$$

where $\bar{J}_i(r,h)$ is the average of the local functions $\hat{J}_i(r,h)$, weighted by some point-wise kernel estimates. This procedure not only provides individual estimates, but it does also account for the vicinity of the observed points, and therefore the contribution of their displacement on the estimation procedure. Thus, consider the weights $w_i = w_{i,\sigma_s} w_{i,\sigma_t}$ given by some kernel estimates, where w_{σ_s} and w_{σ_t} are weight functions, and $\sigma_s, \sigma_t > 0$ are the smoothing bandwidth. It is not necessary to assume that w_{i,σ_s} and w_{i,σ_t} are probability densities. For simplicity, we shall consider only kernels of fixed bandwidth, even though spatially adaptive kernels could also be used. Then, the averaged weighted local statistics $\bar{J}_i(r,h)$ in Equation (3), for each point i, is

$$\bar{J}_i(r,h) = \frac{\sum_{i=1}^n \hat{J}_i(r,h) w_i}{\sum_{i=1}^n w_i}.$$

In particular, we consider $\hat{J}_i(\cdot)$ as the local spatio-temporal pair correlation function

$$\hat{J}_i(r,h) = \hat{g}_i(r,h) = \frac{1}{4\pi r |W \times T| \hat{\lambda}^2} \sum_{j \neq i} \frac{\kappa_{\varepsilon,\delta}(||\mathbf{u}_i - \mathbf{u}_j|| - r, |t_i - t_j| - h)}{\omega(\mathbf{u}_i, \mathbf{u}_j)\omega(t_i, t_j)} \tag{4}$$

where ω is the edge correction factor. The kernel function κ has a multiplicative form $\kappa_{\varepsilon,\delta}(||\mathbf{u}_i - \mathbf{u}_j|| - r, |t_i - t_j| - h) = \kappa_\varepsilon(||\mathbf{u}_i - \mathbf{u}_j|| - r)\kappa_\delta(|t_i - t_j| - h)$ where κ_ε and κ_δ are kernel functions with bandwidths ε and δ, respectively. Both of them are computed using the Epanechnikov kernel and the bandwidths are estimated with a direct plug-in method.

4. Simulation study

A simulation study is carried out to assess the local estimation method proposed in Section 3. We assume a stationary and isotropic LGCP with a separable structure of the covariance of the underlying GRF, with a purely exponential model both in space and time as in Equation (1) and the vector of parameters given by $\psi = \{\sigma^2, \alpha, \beta\}$. To simulate a point pattern from a homogeneous log-Gaussian Cox process, first, a realisation $S(\mathbf{u},t)$ from a spatio-temporal Gaussian Random Field is generated on a grid with dimension $50 \times 50 \times 50$. Conditioning on the realisation of the Random Field, a point pattern is obtained simulating an inhomogeneous Poisson process with intensity given by $\exp(\lambda + S(\mathbf{u},t))$ where $\lambda = \log(n/|W||T|)$. For each scenario, 200 point patterns are generated with $n = 1000$ expected number of points in the spatio-temporal window $W \times T = [0,1]^2 \times [0,50]$. We consider several degrees of clustering in the process with variance $\sigma^2 = \{5,8\}$ and scale

Table 1: Mean (m) and quartiles of the distributions of the varying parameters estimated for the 200 spatio-temporal LGCPs generated assuming an exponential form in both the spatial and temporal dimensions for the GFR as in Equation (1).

True			$\hat{\sigma}^2$				$\hat{\alpha}$				$\hat{\beta}$			
σ^2	α	β	25%	50%	m	75%	25%	50%	m	75%	25%	50%	m	75%
5	0.05	2	5.27	6.30	6.45	7.60	0.05	0.07	0.14	0.09	1.77	2.26	2.63	2.97
	0.10		4.64	5.51	5.67	6.47	0.09	0.11	0.13	0.15	1.80	2.32	2.61	3.08
	0.25		3.68	4.39	4.63	5.37	0.19	0.24	0.34	0.32	1.69	2.22	2.50	2.92
	0.05	5	4.36	5.43	5.54	6.58	0.05	0.07	0.12	0.09	3.36	4.43	5.03	5.93
	0.10		4.13	4.96	5.14	5.97	0.09	0.11	0.14	0.15	3.40	4.45	5.09	6.17
	0.25		3.29	4.10	4.27	5.02	0.17	0.24	0.40	0.34	3.04	4.21	4.89	5.90
	0.05	10	4.08	5.03	5.20	6.12	0.05	0.06	0.10	0.08	5.69	7.85	8.61	10.54
	0.10		3.66	4.44	4.66	5.50	0.08	0.11	0.16	0.14	5.64	8.00	8.85	11.07
	0.25		3.05	3.73	3.97	4.70	0.16	0.22	0.35	0.30	5.15	7.09	8.15	9.98
8	0.05	2	7.26	8.23	8.29	9.37	0.05	0.06	0.07	0.08	2.27	2.85	3.36	3.87
	0.10		6.32	7.26	7.40	8.36	0.08	0.10	0.12	0.13	2.25	2.84	3.17	3.72
	0.25		5.07	5.97	6.16	7.13	0.17	0.22	0.32	0.29	1.97	2.62	2.83	3.43
	0.05	5	6.72	7.64	7.76	8.83	0.05	0.06	0.08	0.08	3.35	4.43	5.05	6.11
	0.10		5.73	6.74	6.91	7.99	0.08	0.10	0.13	0.13	3.29	4.29	4.82	5.81
	0.25		4.79	5.61	5.81	6.69	0.15	0.19	0.29	0.26	3.01	4.17	4.58	5.59
	0.05	10	6.11	7.06	7.14	8.14	0.05	0.06	0.08	0.08	5.37	7.50	8.30	10.22
	0.10		5.15	6.19	6.33	7.24	0.08	0.10	0.12	0.12	4.93	7.04	7.88	9.84
	0.25		4.19	5.05	5.23	6.19	0.14	0.19	0.28	0.26	4.57	6.51	7.60	9.54

parameters in space and time, $\alpha = \{0.005, 0.10, 0.25\}$ and $\beta = \{2, 5, 10\}$, as done by [6]. The mean of the GRF is fixed $\mu = -0.5\sigma^2$. The results come in Table (1).

The results obtained are quite promising: indeed, even considering fixed bandwidths for the weights in the proposed locally weighted minimum contrast, the procedure manage to provide quite precise estimates. This is particularly evident if compared to the results in [6] and, even before, in [3], where the authors provide a number of simulation studied to assess the overall performance of the minimum contrast procedure under different aspects, concluding that especially the variance σ estimates strongly tend to be underestimated. However, our main goal here is not to provide an alternative to the classical global minimum contrast procedure, but instead to estimate varying parameters. This objective is clearly achieved as we manage to obtain a whole distribution for each parameter (of each analysed point pattern).

5. Conclusions

In this paper, we have introduced a novel local fitting procedure for obtaining space-time varying estimated for a log-Gaussian Cox process fitted to the data. From a methodological point of view, we have resorted to the joint minimum contrast procedure (which is appealing for its flexibility in dealing also with non-separable covariances), extending it to the local context, and therefore allowing to obtain a whole set of covariance parameters for each point of the analysed process.

By simulations, we have shown that the local proposal provides good estimates on average, if compared to the global fitting alternatives. Future work regards the application of the proposed methodology to real spatio-temporal point patterns, where it is of interest to study the characteristics of the underlying process, in relation to the spatial displacement and the temporal occurrence of points. Some examples include seismology, forestry, criminology, epidemiology, and so on.

References

[1] Baddeley, A. (2017). Local composite likelihood for spatial point processes. *Spatial Statistics*, **22**:261295.

[2] DAngelo, N., Siino, M., DAlessandro, A., and Adelfio, G. (2021). Local spatial log-gaussian cox processes for seismic data. Submitted.

[3] Davies, T. M. and Hazelton, M. L. (2013). Assessing minimum contrast parameter estimation for spatial and spatiotemporal log-gaussian cox processes. *Statistica Neerlandica*, **67**(4):355389.

[4] Diggle, P. J. (2013). Statistical Analysis of Spatial and Spatio-Temporal Point Patterns. CRC Press.

[5] Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log gaussian cox processes. *Scandinavian journal of statistics*, **25**(3):451482.

[6] Siino, M., Adelfio, G., and Mateu, J. (2018). Joint second-order parameter estimation for spatio-temporal log- gaussian cox processes. *Stochastic environmental research and risk assessment*, **32**(12):35253539.

A spatially correlated self-exciting spatio-temporal model with conditionally heteroskedastic structure for counts of crimes

I. Escudero[1,2,*] , J.M. Angulo[2] and J. Mateu[3]

[1]*Departamento Estadística, Facultad de Ciencias, Escuela Superior Politcnica de Chimborazo, Riobamba, Ecuador; aescudero@espoch.edu.ec*
[2]*Department of Statistics and Operations Research, University of Granada, Granada, Spain; aisabel@correo.ugr.es; jmangulo@ugr.es*
[3]*Department of Mathematics, University Jaume I, Castellón; mateu@mat.uji.es*
[*]*Corresponding author*

Abstract. *In this paper we introduce a class of spatially correlated self-exciting spatio-temporal models for count data that capture data model dependence, as well as dependence in a latent spatial process involving distance-based covariates that vary naturally in space and time. We considered a B-splines procedure that permits to handle space-time variation and non-linear dependencies. A Bayesian framework is proposed for inference on model parameters. We analyze three distinct crime datasets in city of Riobamba (Ecuador). Our model fits well the data and provides better predictions than more simple alternatives.*

Keywords. *Bayesian inference; B-splines; Crimes; Self-exciting process; Spatio-temporal patterns.*

1. Introduction

Modeling time series of counts has received increasing attention since the 1950s. The conditional distribution of observed counts given past outcomes or a latent process comes from some well-known discrete distributions, such as generalized Poisson and double Poisson distributions, that can treat overdispersion and underdispersion, but they have some shortcomings or limitations. An alternative are integer-valued generalized autoregressive conditionally heteroskedastic (INGARCH) models [4, 2, 7], that show flexibility in representing a wide range of overdispersion and underdispersion cases. The stationary distribution of the INGARCH(1,1) process is also equivalent to a stochastic process given in [4], or self-exciting point process. These processes have shown beneficial to model the dynamics of earthquakes, epidemics, forest fires, traffic accidents, crimes, etc.

This study is motivated by the analysis of crime data in the city of Riobamba (Ecuador) provided by three different governmental agencies with the aim of understanding this crime behavior and its interaction with society to further help public institutions to enhance proper actions. The overall challenge is how to appropriately model the space-time dependence relationship between observations. Following the line of reasoning of [1], we focus on INGARCH models for their relationship between the variance and the mean, and we further consider spatial variation on a small scale by taking a spatial integer-valued generalized autoregressive conditionally

heteroskedastic model due to its flexibility to describe the autocorrelation and variance with the mean proportion of the data. In this framework, and following [1], we formulate a stochastic difference equation for the intensity of the space-time process within a class of spatially correlated self-exciting spatio-temporal models that capture both data model dependence as well as dependence in a latent spatial process. We note that the model in [1] considers a linear regression structure in the covariates assuming these are constant in time. We structure space-time dependency for our count data through a combination of distance-based covariates that vary naturally in both space and time. We develop a B-splines procedure within a generalized additive model that permits to handle space-time variation and non-linear dependencies. Our B-splines strategy is more flexible and adapts better to the our case study.

2. Metodology

We focus on a SPINGARCH(1,1) model that overall allows to define the autocorrelation present in the data and the mean-variance ratio with greater flexibility. We use a conditional Poisson distribution and place spatio-temporal structure on the covariance of the latent Gaussian process. The data model $\mathcal{Y}(s_i,t)$ can be defined conditionally on a process model $X(s_i,t)$. As a result, the process model is a function of both observable spatial and/or temporal covariates and unobservable latent spatial errors. In our case, the spatio-temporal intensity $\lambda(s,t)$ provides the process model, and our full model is a stochastic difference equation operating directly on the intensity function. Thus, crime counts in space and time, $\mathcal{Y}(s_i,t)$, are conditionally distributed Poisson random variables for $i = 1,\ldots,n$, i.e., $\mathcal{Y}(s_i,t)|\lambda(s_i,t) \sim \text{Pois}(\lambda(s_i,t))$, with $\lambda(s_i,t)$ representing the rate s_i in time t. Hence, $\text{E}[\mathcal{Y}(s_i,t)|\lambda(s_i,t)] = \lambda(s_i,t)$. We can assume that a change in crime rate at a specific location and in a specific period is a function of particular spatial features of the location given by $\alpha_t = (\alpha(s_1,t),\alpha(s_2,t),\cdots,\alpha(s_n,t))^T$, together with two other factors, a natural deterioration χ, and repeated victimization η. Thus the final model SPINGARCH(1,1) is defined through the following hierarchical structure:

$$\mathcal{Y}(s_i,t)|\lambda(s_i,t) \sim \text{Pois}(\lambda(s_i,t)), \tag{1}$$

with

$$\begin{aligned}
\text{E}[\mathcal{Y}(s_i,t)|\lambda(s_i,t)] &= \lambda(s_i,t),\\
\lambda_t &= \exp(X_t+\varepsilon_t)+\eta\mathcal{Y}_{t-1}+\kappa\lambda_{t-1},\\
X_t &\sim \text{Gau}\left(\alpha_t,(I_{n,n}-\zeta C)^{-1}\sigma^2\right),\\
\varepsilon_t &\sim \text{Gau}\left(0,I_{n,n}\sigma_\varepsilon^2\right),
\end{aligned}$$

with $\mathcal{Y}(s_i,t)$ being defined conditionally on the intensity $\lambda(s_i,t)$, which can be modeled using observable spatial and temporal covariates $\alpha(s_i,t)$, as well as non-observable latent errors ε_t, where $\lambda_t = (\lambda(s_1,t),\lambda(s_2,t),\ldots,\lambda(s_n,t))^T$ is a Markov chain in $(\mathbb{R}^+)^n$, and the same notation applies for $\mathcal{Y}_t$ and X_t. Note that $I_{n,n}$ is the identity matrix, σ^2 is the conditional variance, σ_ε^2 is the latent conditional variance, and ζ controls the amount of spatial dependence in the model not captured by the covariates in α_t. Large scale spatial structure is accounted for in the latent process X_t by the spatial regression parameter α_t, whereas small scale spatial structure is accounted for by conditionally defining X_t. For the latter, a conditionally autoregressive

(CAR) model is used (through spatially adjacent neighbors), $X(s_i,t)|X(s_j,t),s_j \in N|s_i| \sim N(\mu(s_i,t),\sigma^2)$, with $\mu(s_i,t) = \alpha(s_i,t) + \zeta\sum_{s_j\in N|s_i|} X(s_j,t) - \alpha(s_j,t)$. If locations s_i and s_j are neighbors, the entry (i,j) of C will be one. Note that by adding space-time noise $\varepsilon(s_i,t) \overset{iid}{\sim} N(0,\sigma_\varepsilon^2)$ further variation in the spatio-temporal process is allowed. The hierarchical model defined above depends on a set of parameters in the final level of the hierarchy given by $\theta = (\eta,\kappa,\sigma,\sigma_\varepsilon,\alpha_t,\zeta)$, similarly to a classical Besag-York-Mollié (BYM) model [6, 1] which defines a fully Bayesian spatial model (see [9]). The action of the deterministic covariates depending on space or space-time is highly non-linear onto the responses. Thus, we have used a generalized additive model (GAM) that supports integrated smoothness estimation addressing the lack of linearity and others aspects. The relationship between each predictor x_i and the mean of the response variable, $g(u)$, is indirect because it is calculated using the smooth (usually splines with polynomial bases [8]) function $f(x_i)$, defined as $g(u) = \beta_0 + \sum_{i=1}^{p} f_i(x_i)$, with f_t being a smooth spatial surface in the t-th time. The generalized cross-validation criterion (GCV) was used to estimate the smoothing parameters which provide the degree of smoothness. To define the version of smoothing that best fits the data, we tested the joint interactions of the spatial covariates with crime.

3. Results

The city is divided into $m = 141$ administrative zones, whose centroids are denoted by $\{s_1,s_2,\ldots,s_{141}\}$. We count crimes per zone and month. We take average nearest-neighbor distances from each crime to community police units, to surveillance cameras, to markets, to parks and to hospitals, and the population enters the model as a spatial-only covariate of dimension $m \times 1$. Climatological variables were not significant in our context crime events. Testing all possible combinations for a multivariate GAM we find that the univariate GAM provides more robust predictions (see Figure 1).

Model	η	κ	σ	σ_ε
$\alpha_t^I = \beta_0 + f_2(cam) + f_4(par)$	1.00	1.00	1.00	1.01
$\alpha_t^{II} = \beta_0 + f_3(cc) + f_6(pob)$	1.00	1.01	1.00	1.01
$\alpha_t^{III} = \beta_0 + f_1(upc) + f_2(cam)$	1.00	1.00	1.00	1.02

Table 1: Final models and measure of chain equilibrium Rhat for posterior parameters.

Once the parameter α_t is estimated depending on the covariates, and keeping fixed $\zeta = 0.99$ near the edge of the parameter space [1], the remaining parameters $\theta = (\eta,\kappa,\sigma,\sigma_\varepsilon,\alpha_t,\zeta)$ are estimated using a Bayesian approach. We use informative (prior) Beta distributions for η and κ, and Cauchy for σ and σ_ε, but minimize the impact on the posterior densities. The measures of chain equilibrium (Rhat) are the diagnostic statistics on chain convergence (see Table 1) and for the three models are less than 1.05, so we conclude that the chains have converged. Also we compute mean-square prediction errors (MSPE), and these are small enough, 0.45, 0.21 and 0.61 respectively.

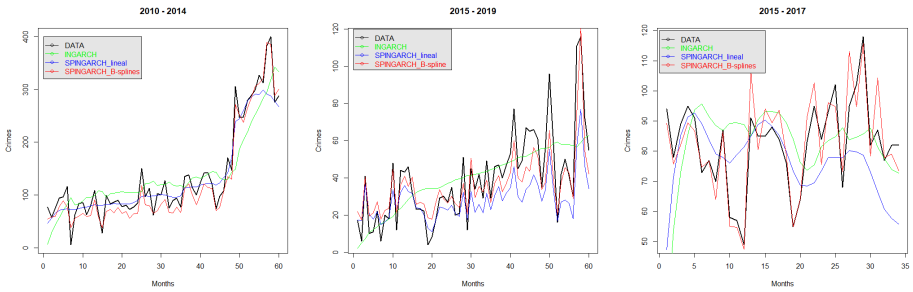


Figure 1: Crimes predictions using INAGRCH, SPINGARCH linear forms on the covariates and constant over time, and SPINGARCH with B-splines on the covariates and evolving in time, for dataset I, II and III, respectively.

4. Conclusions

This study formulates a statistical model that contains both latent spatial dependence and temporal dependence in the form of a stochastic difference equation for the spatio-temporal intensity, i.e., this model is consistent with common beliefs about how violence and crime evolve in space and time. Also our model incorporates the effect of exogenous covariates using non-linear B-splines evolving in time, which provides more robust predictions. Main open ideas in this context include identifying crimes happening on the network of streets in a city as this new support can enhance the modeling task (see [3]). We can also think of using the Next Hit Predictor method (see [5]) in our particular context.

Acknowledgments

J.M. Angulo was partially supported by MCIU/AEI/ERDF, UE grant PGC2018-098860-B-I00, grant A-FQM-345-UGR18 cofinanced by ERDF Operational Programme 2014-2020 and the Economy and Knowledge Council of the Regional Government of Andalusia, Spain, and grant CEX2020-001105-M MCIN/AEI/10.13039/501100011033. J. Mateu was partially supported by grant PID2019-107392RB-I00/AEI/10.13039/501100011033 from the Spanish Ministry of Science and Innovation and grant UJI-B2018-04 from University Jaume I, Spain.

References

[1] Clark, N. J. and Dixon, P. M. (2021). A class of spatially correlated self-exciting statistical models. *Spatial Statistics* **43**, 2211–6753.

[2] Ferland, R. (2006). Integer-valued GARCH Process. *Time Series Analysis* **27**, 923–942.

[3] Gilardi, A. ;Mateu, J.; Borgoni, R. and Lovelace, R. (2022). Multivariate hierarchical analysis of car crashes data considering a spatial network lattice. *Journal of the Royal Statistical Society Series A* **496**, 1–18.

[4] Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika* **58**, 83–90.

[5] Li, Y. and Wang, T. (2018). Next hit predictor self-exciting risk modeling for predicting next locations of serial crimes. *arXiv:1812.05224v1*.

[6] Morris, M.; Wheeler, K.; Simpson, D.; Mooney, S.; Gelman, A. and DiMaggio, C. (2019). Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in stan. *Spatial and Spatio-temporal Epidemiology* **31**.

[7] Reinhart, A. (2018). A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science* **33**, 330–333.

[8] Taylan, P (2010). On the foundations of parameter estimation for generalized partial linear models with B-splines and continuous optimization. *Computers Mathematics with Applications* **60**, 134–143.

[9] Thamrin, S. A. (2019). Geographical mapping of dengue fever incidence 2012-2016 in Makassar, Indonesia. *IOP Conf. Series: Earth and Environmental Science* **279**, 1–8.

A Nonparametric Boostrap Method for Heteroscedastic Functional Data

M. Flores[1,*] , S. Castillo-Páez[2] and R. Fernández-Casal[3]

[1] *Universidad Escuela Politcnica Nacional (Ecuador); miguel.flores@epn.edu.ec*
[2] *Universidad de las Fuerzas Armadas ESPE (Ecuador); sacastillo@espe.edu.ec*
[3] *University of A Coruña (Spain); ruben.fcasal@udc.es*
[*] *Corresponding author*

Abstract. *The aim is to provide a nonparametric bootstrap method for data that consists of independent realizations of a continuous one-dimensional process. The process is assumed to be non stationary, with a functional mean and a functional variance, and dependent. The resampling method is based on nonparametric estimates of the model components. Numerical studies where carried out to check the performance of the proposed procedure, trough the approximation of the bias and the variance of two estimators of the functional mean.*

Keywords. *Resampling methods; Local linear estimation; Conditional variance; Variogram.*

1. Introduction

Assume that $\mathcal{S}_n = \{Y_i(t)\}_{i=1}^n$, for $t \in [a,b] \subset \mathbb{R}$, is a set of n independent observations of a functional variable $Y(t)$ defined over $\mathbb{R}$, verifying:

$$Y_i(t) = \mu(t) + \sigma(t)\varepsilon_i(t), \tag{1}$$

being $\mu(t)$ and $\sigma^2(t)$ deterministic functions, which represent the trend and variance functions, respectively, and $\varepsilon_i(t)$ is a random error process with zero mean, unit variance and correlations

$$Cov\left(\varepsilon_i(t), \varepsilon_{i'}(t')\right) = \delta_{ii'}\rho\left(\left|t-t'\right|\right),$$

for $1 \leq i, i' \leq n$ and $a \leq t, t' \leq b$, where $\delta_{ii'} = 1$ if $i = i'$, $\delta_{ii'} = 0$ if $i \neq i'$ and $\rho(\cdot)$ is the correlogram function.

In practice, each $Y_i(t)$ is observed in a discrete set of points $t_j \in [a,b] \subset \mathbb{R}$, with $j = 1, \ldots, p$. Then, these set of observations can be expresed as a matrix $\mathbf{Y}$ of order $n \times p$, with $\mathbf{Y}_{ij} = Y_i(t_j)$. Furthermore, if $\mathbf{y}_i = (Y_i(t_1), \ldots, Y_i(t_p))^\top$ is the vector corresponding to the i-th row of $\mathbf{Y}$, its covariance matrix $Cov(\mathbf{y}_i) = \mathbf{\Sigma}_0$ (within-curve covariance matrix) has

$$(\mathbf{\Sigma}_0)_{jj'} = \sigma(t_j)\sigma(t_{j'})\rho\left(\left|t_j - t_{j'}\right|\right),$$

for $i = 1, \ldots, n$. Consequently, $\mathbf{\Sigma}_0 = \mathbf{D}\mathbf{\Sigma}_\varepsilon\mathbf{D}$, where $\mathbf{\Sigma}_\varepsilon$ (within-curve correlation matrix) is the covariance matrix of $\varepsilon_i = (\varepsilon_i(t_1), \ldots, \varepsilon_i(t_p))^\top$, for $i = 1, \ldots, n$, and $\mathbf{D} = diag(\sigma(\mathbf{x}_1), \ldots, \sigma(\mathbf{x}_n))$.

2. Nonparametric estimation

The proposed procedure starts with the nonparametric estimation of the trend, the conditional variance and the dependence, following an iterative algorithm similar to that described in [1]. In this case, however, since multiple realizations of the process are available, it was observed that a bias correction in the estimation of the small-scale variability is not typically necessary.

The trend is estimated by linear smoothing of $\{(t_j, Y_i(t_j)) : 1 \leq i \leq n, 1 \leq j \leq p\}$. This estimator can be explicitly written in terms of the sample means $\bar{Y}(t) = \frac{1}{n}\sum_i Y_i(t)$:

$$\hat{\mu}(t) = \mathbf{e}_1^\top \left(\mathbf{X}_t^\top \mathbf{W}_t \mathbf{X}_t\right)^{-1} \mathbf{X}_t^\top \mathbf{W}_t \bar{\mathbf{y}} = s_t^\top \bar{\mathbf{y}} \tag{2}$$

where $\bar{\mathbf{y}} = (\bar{Y}(t_1), \ldots, \bar{Y}(t_p))^\top$, $\mathbf{e}_1 = (1,0)^\top$, $\mathbf{W}_t = diag\{K_h(t_1 - t), \ldots, K_h(t_p - t)\}$, $\mathbf{X}_t$ is a matrix with the j-th row equal to $(1, t_j - t)$, $K_h(u) = \frac{1}{h}K(\frac{u}{h})$, K is a kernel function and h is the bandwidth parameter, which determines the local neighborhood used to estimate the trend. This bandwidth should be selected taking the dependence into account, for instance by using the "bias corrected and estimated generalized cross-validation" criterion (CGCV) proposed in [2], bearing in mind that:

$$Cov\left(\bar{Y}(t_j), \bar{Y}(t_{j'})\right) = \frac{1}{n}\sigma(t_j)\sigma(t_{j'})\rho\left(\left|t_j - t_{j'}\right|\right).$$

The small-scale variability of the process, determined by the conditional variance and the spatial dependence of the error process, is estimated from the residuals $r_{ij} = Y_i(t_j) - \hat{\mu}(t_j)$. Estimates of the conditional variance, $\hat{\boldsymbol{\sigma}}^2 = (\hat{\sigma}^2(t_1), \ldots, \hat{\sigma}^2(t_p))$, are obtained by linear smoothing of $\{(t_j, r_{ij}^2) : 1 \leq i \leq n, 1 \leq j \leq p\}$.

The dependence structure is estimated through the error semivariogram $\gamma_\varepsilon(u) = \frac{1}{2}Var(\varepsilon(t) - \varepsilon(t+u)) = 1 - \rho(u)$. A pilot local linear estimate is obtained by the linear smoothing of the semivariances,

$$\left\{\left(t_j - t_{j'}, \tfrac{1}{2}(\hat{\varepsilon}_{ij} - \hat{\varepsilon}_{ij'})^2\right) : 1 \leq i \leq n, 1 \leq j < j' \leq p\right\},$$

of the standardized residuals $\hat{\varepsilon}_{ij} = r_{ij}/\hat{\sigma}(t_j)$. In fact, as this estimator is not necessarily conditionally negative definite (it cannot be used directly for prediction or simulation), a flexible Shapiro-Botha variogram model [3] is fitted to the pilot estimates to obtain the final variogram estimate $\hat{\gamma}_\varepsilon$.

3. Bootstrap algorithm

The proposed boostrap procedure is as follows:

1. Form the standardized residuals matrix $\hat{\mathbf{E}}$, whose ith row is equal to $\hat{\boldsymbol{\varepsilon}}_i = \hat{\mathbf{D}}^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\mu}})$, where $\hat{\mathbf{D}} = diag(\hat{\sigma}^2(t_1), \ldots, \hat{\sigma}^2(t_p))$ and $\hat{\boldsymbol{\mu}} = (\hat{\mu}(t_1), \ldots, \hat{\mu}(t_p))^\top$.

2. Construct an estimate $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}$ of the within-curve correlation matrix from $\hat{\gamma}_\varepsilon$, and compute its Cholesky

decomposition $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}} = \mathbf{U}^{\top}\mathbf{U}$.

3. Compute the uncorrelated standardized residuals $\mathbf{E} = \hat{\mathbf{E}}\mathbf{U}^{-1}$ and scale them (by subtracting the overall sample mean and dividing by their sample standard deviation).
4. Use the scaled values to derive an independent bootstrap sample $\mathbf{E}^*$ (by resampling the rows and columns of $\mathbf{E}$).
5. Compute the bootstrap errors $\boldsymbol{\varepsilon}^* = \mathbf{E}^*\mathbf{U}$.
6. Obtain the bootstrap sample $\mathbf{Y}^*$, with $\mathbf{y}_i^* = \hat{\boldsymbol{\mu}} + \hat{\mathbf{D}}\boldsymbol{\varepsilon}_i^*$, for $i = 1, \ldots, n$.
7. Repeat B times steps 4-6 to obtain the B bootstrap replicates $\{\mathbf{Y}_1^*, \ldots, \mathbf{Y}_B^*\}$.

The replicates derived from this algorithm can be used to approximate characteristics of the distribution of a statistic under study. For instance, they can be used for approximating the standard error and bias of an estimator.

Acknowledgments

The research of Rubén Fernández-Casal has been supported by MICINN (Grant PID2020-113578RB-I00), and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2020-14 and Centro de Investigación del Sistema Universitario de Galicia ED431G 2019/01), all of them through the ERDF. The research of Sergio Castillo Páez has been supported by the Universidad de las Fuerzas Armadas ESPE, from Ecuador and the research of Miguel Flores has been supported by the Universidad Escuela Politécnica Nacional, from Ecuador.

References

[1] Castillo-Pez, S., Fernández-Casal, R. and García-Soidán, P. (2017). Nonparametric estimation of the small-scale variability of heteroscedastic spatial processes. *Spatial Statistics* **22**, 358–370.

[2] Francisco-Fernández, M. and Opsomer, J.D. (2005). Smoothing parameter selection methods for nonparametric regression with spatially correlated errors. *The Canadian Journal of Statistics* **33**, 279–295.

[3] Shapiro, A. and Botha, J.D. (1991). Variogram fitting with a general class of conditionally non-negative definite functions. *Computational Statistics and Data Analysis* **11**, 87–96.

A nonparametric approach for direct approximation of the spatial quantiles

P. García-Soidán[1,*] and T.R. Cotos-Yáñez[2]

[1]*Department of Statistics and Operations Research & Atlantic Research Center for Information and Communication Technologies (atlantTIC), University of Vigo, Spain; pgarcia@uvigo.es*
[2]*Department of Statistics and Operations Research, University of Vigo, Spain; cotos@uvigo.es*
[*]*Corresponding author*

Abstract. *The approximation of the spatial quantiles has been addressed through different mechanisms. On the one hand, the kriging techniques can be adapted to deal with this issue, by minimizing the corresponding estimation variance. Other proposals are based on minimizing instead a generalized version of the mean absolute deviation or on inverting an estimate of the spatial distribution. However, these approaches suffer from several drawbacks, regarding their lack of optimality or the fact of not providing direct approximations of the spatial quantiles. Thus, the current work introduces an alternative methodology for estimation of the quantiles that tries to overcome the aforementioned issues, by proceeding similarly as done for independent data through the order statistics. With this aim, the available observations are appropriately transformed to yield a sample of the process at each target site, so that the resulting data are then ordered and used to derive the spatial quantile at the corresponding location.*

Keywords. *Distribution function; Kriging; Order statistics; Quantile; Spatial data.*

1. Introduction

The approximation of the quantiles provides a broad information about the random variable under study, as well as allows its comparison with other variables. For independent data, this issue can be easily solved through the order statistics, whose consistency can be checked for continuous distributions. A smoother version is obtained by considering a kernel-type distribution estimator [3] and this procedure has been adapted for dependent observations [4], so that the underlying spatial distribution is first approximated and then inverted to provide the target quantile. The kriging methodology can also be employed for this purpose, by minimizing the resulting estimation variance, although it does not necessarily lead to optimal results, when applied to non-gaussian processes. An alternative has been suggested in [2], based on minimizing a generalized version of the mean absolute deviation, which is not a simple task.

In view of the previous comments, the current work aims to introduce a procedure to approximate the spatial quantiles, which overcomes the aforementioned drawbacks and provides direct quantile estimates. Our proposal tries to mimic the simple mechanism, designed for independent data, which uses the order statistics for this goal, although not applied on the original observations but on transformations of them, which represent a sample of the variable under study.

2. Main results

Let us assume that $\{Z(\mathrm{s}) \in \mathbb{R} : \mathrm{s} \in D \subset \mathbb{R}^d\}$ is a stochastic spatial process, where D denotes the observation region. We require the following model for the spatial process:

$$Z(\mathrm{s}) = \mu(\mathrm{s}) + Y(\mathrm{s})$$

with function μ being the deterministic trend of Z, namely, $\mu(\mathrm{s}) = E[Z(\mathrm{s})]$, for all $\mathrm{s} \in D$, and Y representing a strict stationary random process, with zero mean.

The goal of the current work is to introduce an approach that approximates the α-quantile of $Z(\mathrm{s})$, denoted by $z_{\alpha,\mathrm{s}}$, for any $\alpha \in (0,1)$ and $\mathrm{s} \in D$. Therefore, $F_{\mathrm{s}}(z_{\alpha,\mathrm{s}}) = \alpha$, where F_{s} stands for the unidimensional distribution of $Z(\mathrm{s})$, namely, $F_{\mathrm{s}}(z) = P(Z(\mathrm{s}) \leq z)$, for all $z \in \mathbb{R}$.

Suppose that n data $Z(\mathrm{s}_1), ..., Z(\mathrm{s}_n)$ have been collected at the respective locations $\mathrm{s}_1, ..., \mathrm{s}_n$. Then, we can take $Z_i(\mathrm{s}) = Z(\mathrm{s}_i) - \mu(\mathrm{s}_i) + \mu(\mathrm{s})$, for $i = 1, \ldots, n$. The set of transformed observations $Z_i(\mathrm{s})$ represents a sample of size n of F_{s}, since $P(Z_i(\mathrm{s}) \leq z) = F_{\mathrm{s}}(z)$, for all $z \in \mathbb{R}$ and $i = 1, \ldots, n$.

Then, a natural estimator of $F_{\mathrm{s}}(z)$ is given by the empirical distribution of the transformed data $Z_i(\mathrm{s})$:

$$\hat{F}_{\mathrm{s}}(z) = \frac{1}{n} \sum_{i=1}^{n} I_{\{Z_i(\mathrm{s}) \leq z\}}$$

Under several hypotheses, involving an increasing-domain asymptotics framework, we can prove that $\hat{F}_{\mathrm{s}}(z)$ is an unbiased and consistent estimator of $F_{\mathrm{s}}(z)$, for all $z \in \mathbb{R}$. Thus, this estimator supports the extension to this setting of the typical nonparametric approach employed for independent data and based on considering the approximated α-quantile of the ordered sample.

Following the previous idea, our proposal consists of first computing the order statistics of the transformed data $Z_i(\mathrm{s})$, given by $Z_{(1)}(\mathrm{s}) < Z_{(2)}(\mathrm{s}) < ... < Z_{(n)}(\mathrm{s})$, and then estimating $z_{\alpha,\mathrm{s}}$ through $\hat{z}_{\alpha,\mathrm{s}} = Z_{(k)}(\mathrm{s})$, with k being the smallest integer satisfying that $\alpha \leq \frac{k}{n}$.

By assuming some regularity conditions, we can check that:

$$\hat{z}_{\alpha,\mathrm{s}} \xrightarrow{P} z_{\alpha,\mathrm{s}}$$

for all $\alpha \in (0,1)$.

A smoother alternative for estimation of $z_{\alpha,\mathrm{s}}$ can be derived by considering a weighted distribution estimator $\tilde{F}_{\mathrm{s}}$, instead of $\hat{F}_{\mathrm{s}}$:

$$\tilde{F}_{\mathrm{s}}(z) = \sum_{i=1}^{n} w_i I_{\{Z_{(i)}(\mathrm{s}) \leq z\}}$$

where $w_i = \frac{K\left(\frac{s-s_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{s-s_i}{h}\right)}$, K represents a d-dimensional kernel density and h stands for the bandwidth parameter.

Write $w_{(k)}$ for the weight associated to $Z_{(k)}(s)$ and define $W_{(k)} = \sum_{i=1}^{k} w_{(i)}$. With similar arguments as those that can be applied to $\hat{z}_{\alpha,s}$, we could prove that $\tilde{z}_{\alpha,s} = Z_{(k)}(s)$ provides a consistent estimator of $z_{\alpha,s}$, where k denotes the lowest integer such that $\alpha \leq W_{(k)}$.

To illustrate the behavior of the proposed approaches for approximation of the spatial quantiles, numerical studies have been developed with simulated data. With this aim, 500 samples of size 100 from bivariate gaussian processes, with linear trend, have been drawn on $D = [0,1] \times [0,1]$. The sampling locations were taken on regular grids. In addition, exponential and spherical semivariograms were considered for the dependence structure, with variance 0.25, range 0.25 and nugget 0.1.

For each data set, the theoretical quantiles were estimated through $\hat{z}_{\alpha,s}$ (method 1) and $\tilde{z}_{\alpha,s}$ (method 2) at locations $s^{(1)} = (0.05, 0.05)$, $s^{(2)} = (0.25, 0.25)$ and $s^{(3)} = (0.45, 0.45)$. Figure 1 displays the averages of the squared errors of the quantiles estimators that were achieved at the three target sites.

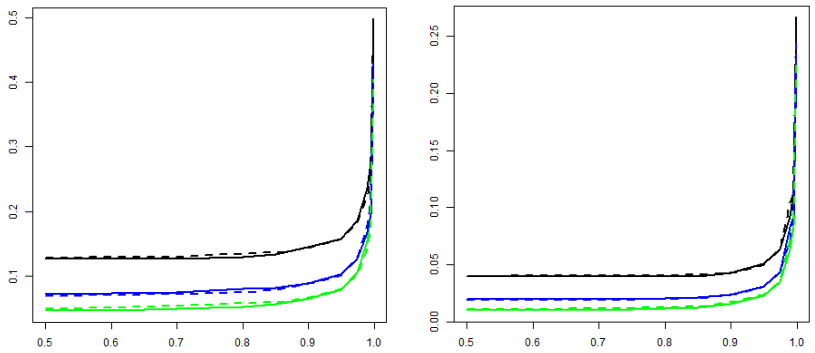


Figure 1: Averages of the squared errors of the quantiles estimators obtained for simulated gaussian data derived from the exponential model (left panel) and the spherical model (right panel), at the three target locations $s^{(1)}$ (black color), $s^{(2)}$ (blue color) and $s^{(3)}$ (green color). The dotted and continuous lines respectively represent the results attained through methods 1 and 2.

The results depicted in Figure 1 provide small values for the averaged squared errors, except for the extreme quantiles. It is also worth noting the similar performance that both approaches show for estimating the spatial quantiles. These preliminary conclusions will be checked by accomplishing an extensive simulation study that covers a greater variety of scenarios.

Alternative quantile estimates could be derived by linear interpolation of the adjacent data $Z_{(k)}(s)$, obtained by either of the two proposals, similarly as done for independent observations [1]. This way of proceeding

yields additional options for approximation of the spatial quantile, such as:

$$\bar{z}_{\alpha,\mathrm{s}} = (1-\delta)Z_{(k-1)}(\mathrm{s}) + \delta Z_{(k)}(\mathrm{s})$$

where $\delta = \alpha n - k + 1$.

Acknowledgments

The first author's research was partially funded by the Spanish Ministry of Science and Innovation project PID2020-113979RB-C22, by the ERDF and by the Xunta de Galicia (Spain), under project ED431C 2019/25 SC7-GRC 2019 and under agreement for funding atlantTIC (Atlantic Research Center for Information and Communication Technologies). The second author's work has been partially supported by project PID2020-118101GB-I00 from the Spanish Ministry of Science and Innovation (AEI/10.13039/501100011033).

References

[1] Hyndman, R. J.; Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician* **50**, 361–365.

[2] Journel, A. G. (1984). mAD and conditional quantile estimators. In: Verly, G.; David, M.; Journel, A. G.; Marechal A. (Eds), *Geostatistics for natural resources characterization* (pp. 261–270). Dordrecht, The Netherlands. Springer.

[3] Nadaraya, E. A. (1964). Some new estimates for distribution functions. *Theory of Probability and its Applications* **15**, 497–500.

[4] Ould Abdi, S. A.; Dabo-Niang, S.; Diop, A., Ould Abdi, A. (2010). Consistency of nonparametric conditional quantile estimator for random fields. *Mathematical Methods of Statistics* **19**, 1–21.

Modeling Spatial Dependencies of Natural Hazards in Island Nations using Barriers

S. Chaudhuri[1] , P. Juan[1,*] , L. Serra-Saurina[2] , D. Varga[2] and M. Saez[2]

[1] *University Jaume I, Castellń, som.rtc@gmail.com, juan@uji.es*
[2] *Research Group on Statistics, Econometrics and Health (GRECS), University of Girona and CIBER of Epidemiology and Public Health (CIBERESP), laura.serra@udg.edu, dievarga@gmail.com, marc.saez@udg.edu.*
[*] *Corresponding author*

Abstract. *Natural hazards like flood, cyclone, landslide, earthquake or, tsunami have deep impacts on environment and society causing damage to both life and property. Computational modeling provides an essential tool to estimate the damage by incorporating spatial uncertainties and local geographic and climatic conditions. The objective of the current study is to explore the application of Integrated Nested Laplace Approximation (INLA) with Stochastic Partial Differential Equation (SPDE) implemented using barrier model. Classical stationary models in spatial statistics inappropriately smooth over features having boundaries, holes, or physical barriers in the study area, e.g. dispersed islands. This leads to the use of non-stationary models like barrier model. In the present study, we have explored the differences between the classical spatial approach and the barrier model using natural hazards data from Maldives. In the broader picture, this research work contributes to the relatively new field of barrier models as well as to initiate and develop scientific research works on the unique island nation of Maldives.*

Keywords. *Barrier mesh, INLA-SPDE, Islands, Maldives, Natural hazards.*

1. Introduction

From an environmental point of view, natural hazards represent a danger to ecosystems, directly affecting geomorphological and hydrological processes, as well as biodiversity. They also endanger human settlements with serious consequences for society [8, 3]. Modeling natural disasters is very important to characterize these phenomena and provide tools to overcome them. Estimating natural hazards, including spatial effects and local conditions, both climatic and geographic, will help in management and even allow anticipation of events [2]. Traditionally, Bayesian approach with Markov Chain Monte Carlo (MCMC) simulation methods can be used to fit for processing generalized linear mixed model (GLMM) [4]. For approximation Bayesian inference a computationally more efficient prediction of the marginal distributions can be achieved by using integrated nested Laplace approximation (INLA) [6]. The analysis of spatial point processes by implementing INLA approach can be explicitly linked between Gaussian function (GF) and Gaussian Markov random fields (GMRF) through SPDE. We are estimating the logarithm of expected number of events (η_k) occurring at k^{th}

sub-region from the entire study area using the following linear predictor:

$$log(\eta_k) = \beta_0 + \sum_{m=1}^{M} \beta_m z_{mk} + \sum_{l=1}^{L} f_l(v_{lk}) \tag{1}$$

where, β_0 is a scalar, which represents the intercept, $\beta = (\beta_1, \ldots, \beta_M)$ are the coefficients of the linear effects of the covariates $z = (z_1, \ldots, z_M)$ on the response, and $f = \{f_1(.), \ldots, f_L(.)\}$ is a collection of functions defined in terms of a set of other covariates represented as $v = v(v_1, \ldots, v_L)$, different from the previous covariates [6]. In the current study, the function used is SPDE that is used to analyze the spatial effect with the Matérn covariance function.

The current study is conducted to explore and model occurrence of natural hazards in the island nation of Maldives. The country consists of 1200 dispersed islands on both sides of the equator. A stationary model can not be aware of the coastline and the island boundaries and will inappropriately smooth over the features. This might result to an unrealistic assumption [1]. In the recent research work by [1] a new non-stationary model was constructed for INLA having syntax very similar to the stationary model. The model, named as barrier model has been designed considering water (Finnish Archipelago Sea) as normal terrain and it is aware of the distinct coastlines and boundaries considered as physical barriers. In the present study, we have explored the barrier model in a converse mode where water body acts as barriers for the dispersed islands and natural hazards are the sample events considered precisely on the land area of the islands.

2. Data sets

Republic of Maldives is located on the south-western region off the coast of India in the Arabian Sea of the Indian Ocean. The Maldivian archipelago consists of about 1190 coral islands grouped into 26 natural atolls. Out of which 188 islands are inhabited [5].

The natural hazards dataset of Maldives for the year 2004, contains 190 records of tsunami affected islands, all being inhabitat islands and provides the number of direct and indirect affected people for individual island. The dataset is published by open data sharing platform, Humanitarian Data Exchange (HDX) managed by the United Nations Office for the Coordination of Humanitarian Affairs (OCHA) under a Creative Commons Attribution 4.0 International license. It is noteworthy to mention that natural hazards like cyclone, typhoon, storm, flood and water shortage can also be accessed from the same open portal. We have used tsunami data as a showcase for the current study.

3. Methodology and Results

Traditionally spatio-temporal modeling using INLA-SPDE is performed by generating a SPDE mesh for the entire study region [7]. In that case, the model result might be unpreventably generalised as it is going to estimate predicted values for the regions where there is no chance of incidents to happen. We construct some spatial polygon covering our study area, then, we define our study area, as a manually constructed polygon,

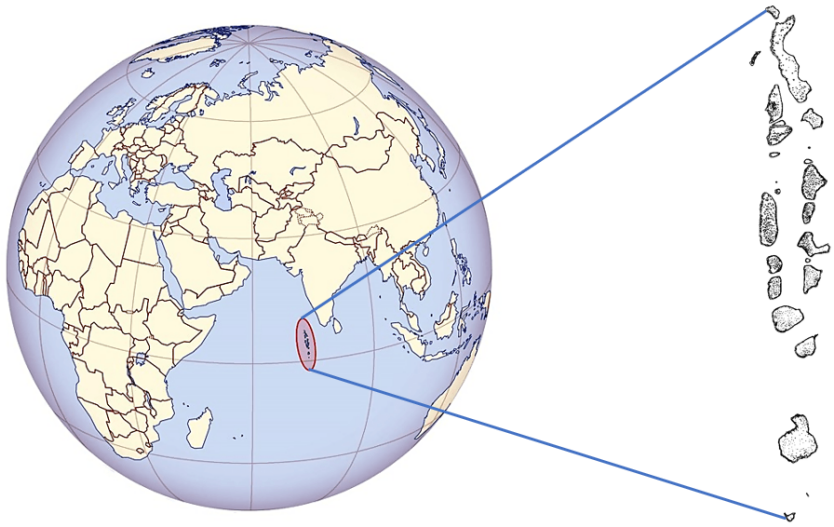


Figure 1: Study area: Republic of Maldives is one of the smallest countries in Asia and the Pacific having the chain of coral islands across an archipelago more than 800 kilometers long and 130 kilometers wide.

Figure 2: Figure on left shows the locations of tsunami affected regions of individual atolls which includes enclosed lagoon or basin, forereef, subtidal reef, pass reef flat and land on reefs. The figure on the right shows the same affected locations precisely on inhabited land on reefs areas.

and intersect this with the coastal area. Since we have a polygon for land, we take the difference instead of an intersection, finally, we construct the mesh we are going to use. The new polygon is where our model assumes there to be land, hence we use this polygon also for plotting the results.

Figure 3: As showcase, the southernmost atoll of Maldives, Addu atoll has been considered for the current study. Both figures display the tsunami affected regions for Addu atoll. Figure on left depicts both the lagoon and reef areas of the atoll, while figure on right shows only land on reefs areas.

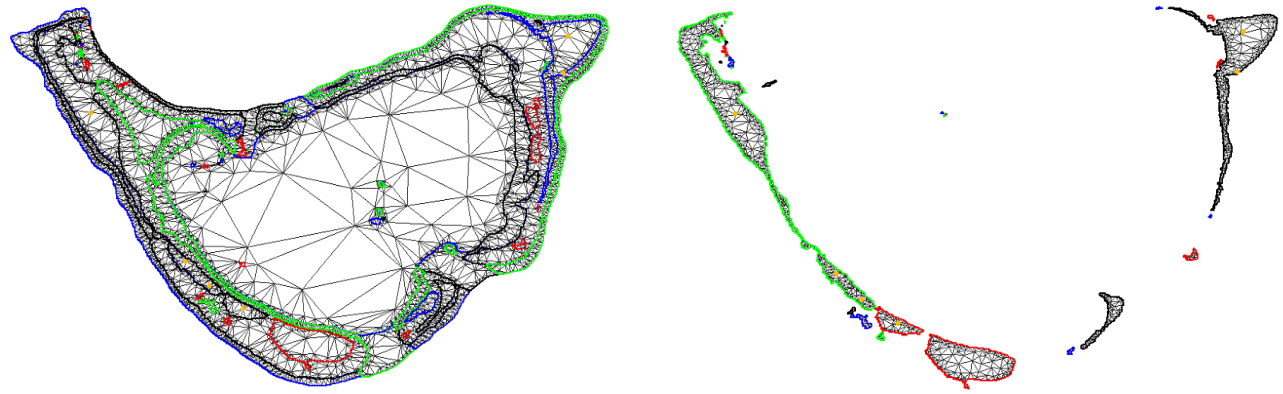


Figure 4: Figure on left represents SPDE triangulation with tsunami affected regions (highlighted in yellow) for the entire region. Figure on right is the same triangulation generated using barrier model where land on reefs are considered as normal terrain and the water bodies (ocean and lagoons) act as physical barriers. In the first case the number of triangulations in the SPDE mesh is 10490 while the barrier model generates 3073 triangulations.

In spatial modelling, classical models are unrealistic when they smooth over holes or physical barriers. Barrier model is more realistic with both sparse data and complex barriers and computational cost is the same as for the stationary models [1]. The current study explore the application of barrier model and relate with spatial dependencies of natural hazards. Barrier models can be appropriate tools to model different types of environmental phenomena that have a clear anisotropic behavior.

References

[1] Bakka, H., Vanhatalo, J., Illian, J.B., Simpson, D. and Rue, H. (2019) Non-stationary Gaussian models with physical barriers. *Spatial Statatistics, 29*, 268-288.

[2] Cutter, S. L., and Finch, C. (2008). Temporal and spatial changes in social vulnerability to natural hazards. *Proceedings of the National Academy of Sciences, 105*(7), 2301-2306.

[3] Emmer, A. (2018). Geographies and scientometrics of research on natural hazards. *Geosciences, 8*(10), 382.

[4] Juan, P., Mateu, J. and Saez, M. (2012). Pinpointing spatio-temporal interactions in wildfire patterns. *Stochastic Environmental Research and Risk Assessment, 26*, 1131-1150.

[5] Maldives. (2015, August 1). Asian Development Bank. Retrieved from *www.ebook.de/de/product/30686580/maldives.html*.

[6] 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.

[7] Verdoy, P. J. (2019). Enhancing the SPDE modeling of spatial point processes with INLA, applied to wildfires. choosing the best mesh for each database. *Communications in Statistics - Simulation and Computation*, 1-34.

[8] Zorn, M., and Komac, B. (2013). Contribution of Ivan Gams to Slovenian physical geography and geography of natural hazards. *Acta Geographica Slovenica, 53*(1), 23-41.

Spatial modeling of epidermal nerve fiber patterns

K. Konstantinou[1,*] and A. Särkkä[1]

[1]*Department of Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden; konkons@chalmers.se, aila@chalmers.se*
[*]*Corresponding author*

Abstract. *Diabetic neuropathy is a disorder characterized by impaired nerve function. The number of epidermal nerve fibers (ENFs) per epidermal surface decreases dramatically, and the two-dimensional spatial arrangement of ENFs becomes more clustered as the neuropathy advances. The entry, branching and termination points of epidermal nerve fibers obtained from the feet of healthy controls and subjects at the earliest stage of the neuropathy are treated as realizations of multitype, marked three-dimensional point processes. The main interest is studying and comparing the complete ENF tree structure between the two groups. For this purpose, three dimensional point process models allowing interactions between the end points within cylindrical regions centered at each point are developed and fitted to the data in the two groups. Due to the anisotropic nature of the data, goodness of fit evaluation of the model is performed using the cylindrical K function.*

Keywords. *Anisotropy; Markov chain Monte Carlo; Markov random field; Point process; Pseudolikelihood*

References

[1] Konstantinou, K., and Särkkä, A. (2021). Spatial modeling of epidermal nerve fiber patterns. Statistics in Medicine, 40(29), 6479-6500.

[2] Konstantinou, K. and Särkkä, A.(2022). Pairwise interaction Markov model for the 3D epidermal nerve fiber endings (Manuscript in preparation).

Optimal path selection for road traffic safety based on wildlife-vehicle collisions

P. Llagostera[1] , C. Comas[1] , C. Dalfó[1] and N. López[1,*]

[1]*Department of Mathematics, University of Lleida, Spain; pol.llagostera@udl.cat, carles.comas@udl.cat, cristina.dalfo@udl.cat, nacho.lopez@udl.cat*
[*]*Corresponding author*

Abstract. *Wildlife-vehicle collisions present an important coexisting problem between human populations and the environment. These type of accidents are a serious problem for the life and safety of car drivers, cause property damage to vehicles, and affect wildlife populations. We present a new approach based on algorithms used to obtain minimum paths between vertices in weighted networks to obtain the optimal (safest) route between two points (departure and destination points) in a road structure based on wildlife-vehicle collision point patterns. We have adapted the road structure into a mathematical linear network and analysed it using some graph theory methodologies. This new approach has been illustrated with a case study in the region of Catalonia, North-East of Spain. This example shows the usefulness of our new approach to identify optimal path between pair of vertices based on weights associated to each edge.*

Keywords. *Point patterns; Road safety; Spatial analysis; Wildlife-vehicle collisions*

1. Introduction

Wildlife-vehicle collisions (WVC) present an important coexisting problem between human populations and the environment. These type of accidents are a serious problem for the life and safety of car drivers, cause property damage to vehicles [4], and are a real peril to wildlife populations [3]. As this type of accidents do not occur randomly neither in space nor in time, in the past decades there has been a growing interest in the analysis and the modelling of such type of events [5], in particular for identifying areas with a high occurrence of accidents (hotspots) [7]. A tentative way to investigate such space events, implies the analysis of point locations (in this case WVC) distributed on linear structures (road configurations). As the occurrence of WVC are affected by several ecological, biological, and meteorological covariates together with some structural road characteristics, one may expected that distinct road sections will have distinct probabilities of having a WVC assuming the covariates surrounding this road area. This suggests the possibility of using weighted graphs to model the risk of WVC, and to define distinct path configurations to optimize this risk. Weighted graphs, based on WVC information, could be considered to find the safest road path between two vertices (a departure and a destination point). our main aim in this paper is to develop a new approach adapting algorithms used to obtain minimum paths between vertices in weighted networks to model road traffic safety based on WVC point patterns.

2. Linear network and point patterns

We represent the road configuration as a linear network L as defined in [1]. Based on this linear configuration and on the point pattern of WVC on L, we obtain an estimator of the rate or intensity function of the point pattern of WVC on L, assuming the diffusion estimator proposed by [6] and implemented in the Spatstat R package [2] as a tentative kernel intensity estimator. Based on this intensity function, we obtain the average intensity value for each edge of L as the integral of this intensity function over a edge. Now each line segment has a weight that represents the average of WVC. This value will be considered as the weight for each line segment to evaluate the risk of having a WVC. For instance, if we consider a constant point intensity, longer line segments are expected to have more accidents than shorter ones, and therefore, the shortest path between two prescribed points on L will result also in the safest one.

3. Algorithms used to obtain minimum paths between vertices in weighted networks

Several algorithms have been proposed to calculate the minimum path between two vertices in a weighted network, and probably the most used are the Dijkstra algorithm, the Bellman-Ford algorithm, the Floyd-Warshall algorithm and the Johnson algorithm. These algorithms are defined to explore weighted networks in search of the path between two points (usually vertices) that has minimum cost. This cost is usually defined as the total sum of the weights associated with the edges of each path. Here a path is defined as a set of unique edges that connects distinct vertices starting from a vertex origin and ending with a vertex destination. In this work, we consider two distinct problem. First, we calculate the total number of possible paths between two points (vertices) and, second, rank the top K-best paths between them, based on a given criterion. To accomplish with the first question we consider an adaptation of the well-known algorithm called Depth First Search (DFS) from the igraph R library, and for the second problem, we use the Yen's algorithm [8].

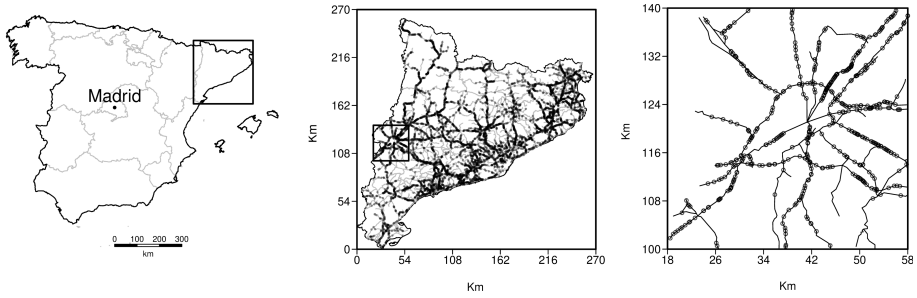


Figure 1: Left and centre: location of the study area together with the location of 6590 roadkills during the period $2010-2014$ and the underlying road network in Catalonia (North-East of Spain), given in km. Right: a magnification of the study region around the city of Lleida (40 km $\times$ 40 km) with the location of 491 roadkills.

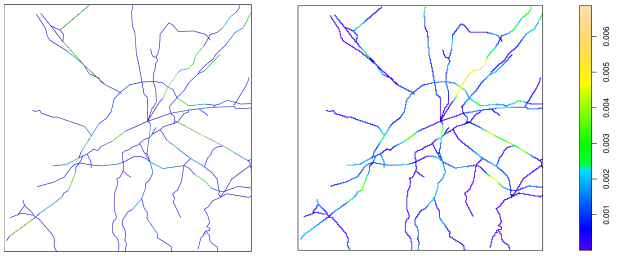


Figure 2: Right: resulting point intensity (roadkills) for the roadkill point pattern based on the diffusion estimator with a bandwidth value around 750 meters. Left: weighted network structure based on the average number of wildlife-vehicle collisions for each linear segment.

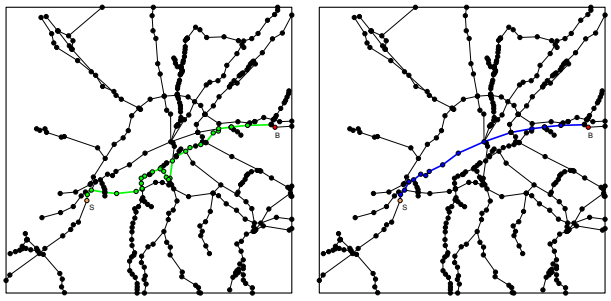


Figure 3: Right: Linear structure of the study region together with the resulting shortest path for the origin/destination points, Soses (S)/Bell-lloch d'Urgell (B) (blue line). Left: resulting safest path for this pair of origin/destination points (green line).

4. Analysing the Wildlife-vehicle collision dataset

We analyse the spatial structure of a dataset containing 491 WVC occurred in a squared area (40 km $\times$ 40 km) around the city of Lleida, North-East of Spain (see Figure 1) during the period 2010-2014. This study area involves 459.050 km of roads for three distinct road categories, namely, highways and paved roads. The Department of Territory and Sustainability of the Autonomic Government of Catalonia (https://web.gencat.cat) provided the roadkill records and the road structure for this study. To illustrate the performance of our new approach, we consider one possible scenario assuming this wildlife-vehicle collision data. Figure 2 shows the resulting point density based on the diffusion estimator with a bandwidth value around 750 meters chosen to provide a good visual fitting to the point pattern, together with the weighted network structure based on the average number of wildlife-vehicle collisions for each edge. We modify the DFS and the Yen's algorithm to find the safest paths between two real town locations Soses and Bell-Lloc d'Urgell (see Figure 3 for town locations). Other pair of vertices on L could have been considered. Figure 3 shows the safest path between this pair of origin/destination points together with the shortest path between these two locations. This highlights that the

safest path is not the shortest one between these two locations. Note that this safest path is the best road route over 607.416 paths obtained by combining the 410 vertices and 437 edges of this linear network configuration.

5. Future work

The next step in our work is to consider traffic flow information in our optimization procedure. Traffic flow information, such as, vehicle traffic volume is crucial to full understand WVC. For instance, two roads with similar intensity of accidents, the road with higher traffic volume is, probably, safer than that with a lower traffic volume. For the same occurrence of accidents the road with a higher traffic volume has less accidents per vehicle than the road with lower traffic volume.

Acknowledgments

Work partially funded by grant PID2020-115442RB-I00 from MCIN/AEI/10.13039/501100011033, the Spanish Ministry of Science and Innovation.

References

[1] Ang, W., Baddeley, A., Nair, G., (2012). Geometrically corrected second order analysis of events on a linear network, with applications to ecology and criminology. *Scandinavian Journal of Statistics* **39**, 591-617.

[2] Baddeley, A., Rubak, E., Turner, R., (2015) Spatial Point Patterns: Method- ology and Applications with R. London: Chapman and Hall/CRC Press.

[3] Coffin, A.W., (2007). From roadkill to road ecology: A review of the ecological effects of roads. *Journal of Transport Geography* **15**, 396-406.

[4] Groot Bruinderink, G.W.T.A., Hazebroek, E., (1996). Ungulate traffic collision in europe. *Conservation Biology* **26**, 1059-1067.

[5] Gunson, K., Mountrakis, G., Quackenbush, L., (2011). Spatial wildlife-vehicle collision models: A review of current work and its application to trans- portation mitigation projects. *Journal of Environmental Management* **92**, 1074-1082.

[6] McSwiggan, G., Baddeley, A., Nair, G., (2016). Kernel density estimation on a linear network. *Scandinavian Journal of Statistics* **44**, 324-345.

[7] Ramp, D., CaldwellS, J., Kathryn, K., Croft, D.W.D., (2005). Modelling of wildlife fatality hotspots along the snowy mountain highway in new south wales, australia. *Biological Conservation* **126**, 474-490.

[8] Yen, J.Y., (1971). Finding the k shortest loopless paths in a network. *Network. Management Science* **17**, 661-786.

Spatio-Temporal Event Studies for Air Quality Assessment

P. Maranzano[1,2]

[1]*Dept. of Economics, Management and Statistics (DEMS), University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126, Milano, Italy; paolo.maranzano@unimib.it*
[2]*Fondazione Eni Enrico Mattei (FEEM), Corso Magenta, 63, Milano, 20123, Milano, Italy*

Abstract. *Event Studies (ES) are statistical tools that assess whether a particular event of interest has caused changes in the level or volatility of one or more relevant time series. We are interested in ES analyses applied to multivariate time series characterized by high spatial dependence. An example of such structures of data are the concentrations of air pollutants observed on geo-referenced monitoring networks located on specific territories. Recent literature showed that cross-sectional dependence is capable of dramatically compromising the performance of ES statistics if not properly adjusted. The objective is to show how the use of spatio-temporal geostatistical models to perform ES applied to georeferenced data, instead of the classical univariate regression models, allows to highlight in a very accurate way the actual effect of shocks due to the events of interest. In particular, we compare ES statistics (both adjusted and non-adjusted with respect to spatial cross-correlation) obtained through three models: 1) Hidden Dynamics Geostatistical Model or HDGM (spatio-temporal model); 2) multiple linear regression (purely temporal model); 3) multiple linear regression with ARIMA errors (purely temporal model). The models are applied to the case study of air quality in Lombardy (Northern Italy), one of the most critical areas for pollution in Europe. The results show that the HDGM is capable of modelling the concentrations and the correlation between stations in a much more precise way than regression models, with the direct consequence of much more reliable and realistic estimates of ES statistics.*

Keywords. *Event Studies; Geostatistics; HDGM; Air quality; Spatial cross-sectional dependence.*

1. Event Studies statistics and intervention analysis

Event studies [3], hereafter ES, are statistical tools used to assess whether a particular event of interest has caused changes in the level or volatility of one or more relevant time series. ES can be directly connected with branches of statistics devoted to the impact assessment of policies or unexpected shocks, as well as intervention analysis. However, while statistical intervention analyses are common tools in studying the air quality and the impact of pollution mitigation policies [4, 6], event studies are only recently receiving attention in pollution-related fields such as energy and oil commodity markets [5, 9, 10].
ES are tools grounded in the *offline hypothesis testing* methods of [1], in which a *without change* scenario (i.e., no abrupt changes occurred in the observed data) is compared to a *with change* scenario (i.e., the data were statistically significantly affected by some shock or event). In ES, the idea is to segment the available observations over time into two windows: a first part is used to estimate the parameters of the regression model (i.e. the *estimation window*), while the second part is used to quantify the effect of the intervention without re-estimating the model (i.e. the *event window*). The estimation window must be unaffected by the intervention

to produce unbiased parameter estimates.

2. Addressing the spatial cross-sectional dependence in ES

We are interested in ES analyses applied to multivariate time series characterized by high spatial (cross-sectional) dependence. An example of such structures of data are the concentrations of air pollutants observed on geo-referenced monitoring networks located on specific territories. As discussed by [7], the presence of high cross-sectional dependence immediately leads to strong biases in classical ES test statistics and misidentifying of the outcomes of the events of interest. Considering the case of airborne pollutants measured through ground monitoring networks, cross-sectional dependence is a direct consequence of the spatial correlation existing the sampling points in space. Indeed, it is reliable to assume that control units located at close distances record similar values under the same environmental conditions.

To address the spatial cross-correlation issue, we adopt a twofold adjustment with respect to classical event studies frameworks: first, we use a linear mixed spatio-temporal regression model called Hidden Dynamics Geostatistical Model (HDMG) [2] to model the relationship between observed concentrations and several exogenous factors, such as meteorology and calendar effects, and at the same time to model the spatio-temporal dynamics between observations. This model will be compared with purely temporal regression models (i.e., not adjusted for spatial dependence); second, we apply and compare a series of sixteen ES test statistics, both parametric and nonparametric, some of which directly adjust for cross-sectional dependence.

3. Hidden Dynamics Geostatistical Model and ES statistics

We consider that the data are driven by a spatio-temporal process $\{Y(s,t) \in \mathbb{R} : s \in D, t = 1,\dots,T\}$, where D is the spatial domain and t represents a discrete point of time. HDGM is composed by a random effects term $w(s,t)$ modelling the spatial and temporal dependence, and by a fixed effects term $v(s,t)$ accounting for all exogenous regressive effects. That is,

$$Y(s,t) = v(s,t) + w(s,t) + \varepsilon(s,t) \tag{1}$$

with $\varepsilon(s,t)$ being the error vector that is assumed to be independent and identically distributed across space and time with mean zero and a constant variance σ^2_ε. The random effects term $w(s,t)$, which accounts for the spatio-temporal dependence in the random process $Y(s,t)$, can be defined using a Markovian process for all the temporal dependencies and interactions and having a spatial Matrn covariance function. The maximum likelihood estimates is computed using the EM algorithm, which is implemented together with the parameter variance-covariance matrix computation in D-STEM software [8].

Let $\Omega_0 = T_0 + 1, T_0 + 2, ..., T_1$ be the set of time indexes included in the *estimation window*, and $\Omega_1 = T_1 + 1, T_1 + 2, ..., T_2$ be the time indexes at which we want to test the presence of abnormal movements on the residuals (i.e. the *event window*). By *abnormal residuals* (AR_{st}), we define the residuals of the regression of the dependent variable $Y(s,t)$ calculated using the event window Ω_0 at each location s, i.e. $AR_{st} = Y(s,t) - \hat{Y}(s,t)$.

The fitted values $\hat{Y}(s,t)$ are the value of $Y(s,t)$ that would be expected if the event did not take place. By the term *cumulated abnormal residuals* ($\text{CAR}_{s,\tau}$) we mean the cumulative sum of abnormal residuals (AR_{st}) in a given time window. The null hypothesis of the ES is that CAR_{s,Ω_1} during the event-window have null mean value (i.e. no abrupt change during the event window), whereas the alternative hypothesis is that the cumulated abnormal residuals in the event-window have negative mean value (i.e. reduction over the event window). The event study is performed considering a set of sixteen test statistics, as in the study of [7], some of them directly adjusted for cross-sectional dependence.

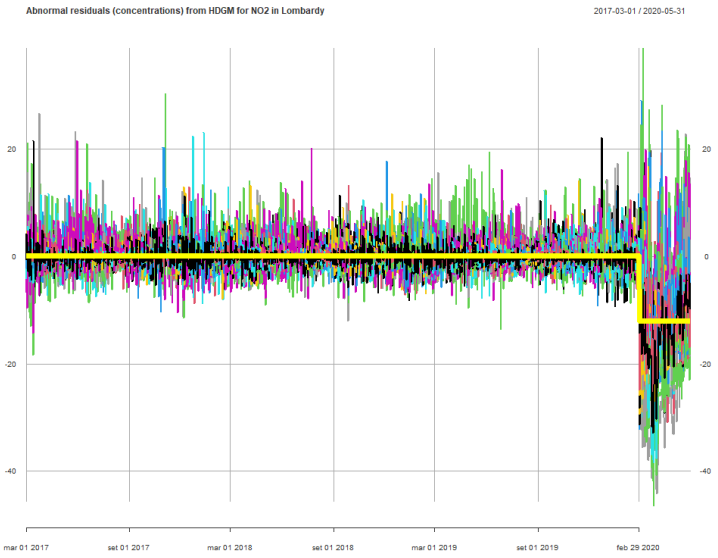


Figure 1: Abnormal residuals for NO_2 from HDGM for the whole sample at the 84 monitoring sites. The solid yellow lines are the pre-event and during-event average NO_2 concentrations, respectively.

4. Application: air quality and COVID-19 in Lombardy

We present an empirical application of ES analysis concerning the effect of the lockdown restrictions imposed on air quality in the Lombardy region (Italy) in response of COVID-19 disease spread in 2020. The aim is to test whether traffic and mobility restrictions have affected the average levels of nitrogen oxide concentrations by exploiting several test statistics applied to daily concentrations of NO_2 at 84 ground monitoring stations from the 9^{th} March to 18^{th} May 2020. Figure 1 shows the estimated abnormal residuals using the HDGM over the whole window. The plot highlights that the HDGM is able to predict very well the spatial and temporal dynamics of the concentrations (i.e. absence of seasonality or patterns in the residuals and cross-correlation on average equal to 0.07), allowing to clearly emphasize the mitigating effect of the lockdown (vertical shift from March 2020). The results support the hypothesis of a generalized reduction of the average nitrogen dioxide

Spatial dep.	Statistic	HDGM		lm		regARIMA	
		Value	Signif.	Value	Signif.	Value	Signif.
Adjusted	P1	-16.90	***	-13.94	***	-1.94	**
Adjusted	P2	-46.25	***	-13.84	***	-1.94	**
Not-adjusted	cross_t_test	-21.40	***	-15.36	***	-9.56	***
Not-adjusted	crude_dep_t_test	-29.03	***	-14.14	***	-1.75	**
Not-adjusted	T_skew	-20.35	***	-21.41	***	-13.38	***
Not-adjusted	Z_patell	-229.84	***	-88.58	***	-10.68	***
Adjusted	Z_patell_adj	-72.21	***	-15.55	***	-1.83	**
Not-adjusted	Z_BMP	-29.49	***	-21.14	***	-10.59	***
Adjusted	Z_BMP_adj	-8.74	***	-2.92	***	-1.41	*
Adjusted	T_grank	-11.26	***	-2.57	***	-1.71	**
Not-adjusted	Z_grank	-14.95	***	-15.06	***	-9.61	***
Adjusted	Z_grank_adj	-5.65	***	-2.57	***	-1.71	**
Adjusted	CumRank	-30.45	***	-12.91	***	-2.11	**
Adjusted	CumRank_mod	-31.40	***	-13.31	***	-2.17	**
Adjusted	CumRank_t	-78.48	***	-14.44	***	-2.17	***
Not-adjusted	CumRank_Z	-108.96	***	-81.36	***	-12.39	***
Adjusted	CumRank_Z_adj	-41.13	***	-13.84	***	-2.18	**
Adjusted	CorradoTuckey	-46.24	***	-13.83	***	-1.94	**

Table 1: Test statistics for NO_2. H_0: $CAR_{\Omega_1} = 0$ (i.e. the average of cumulative abnormal residuals is null during the event period) VS H_1: $CAR_{\Omega_1} < 0$ (i.e. the average of cumulative abnormal residuals reduced during the event period).

concentrations in the region. Indeed, for all the model considered, all the implemented ES statistics unanimously suggest that the lockdown restrictions caused significant reductions of NO_2 concentrations all over the region (see Table 1). Regarding cross-correlation, the adjusted test statistics show smaller estimates, compared to the unadjusted statistics, and are more consistent with the corresponding probability distributions.

References

[1] Basseville, M. and Nikiforov, I. (1993). Detection of Abrupt Change Theory and Application, volume 15.

[2] Calculli, C., Fasso, A., Finazzi, F., Pollice, A., and Turnone, A. (2015). Maximum likelihood estimation of the multivariate hidden dynamic geostatistical model with application to air quality in apulia, italy. *Environmetrics*, **26**(6):406417, 2015.

[3] Campbell, J.Y., Lo, A.W., MacKinlay, AC., and Whitelaw, R.F. (1998). The econometrics of financial markets. *Macroeconomic Dynamics*, **2**(4):559562.

[4] Cujia, A., Agudelo-Castaneda, D., Pacheco-Bustos, C., and Calesso-Teixeira, E. (2019). Forecast of pm10 time-series data: A study case in caribbean cities. *Atmospheric Pollution Research*, **10**(6):20532062.

[5] Demirer, R. and Kutan, A.M. (2010). The behavior of crude oil spot and futures prices around opec and spr announcements: An event study perspective. *Energy Economics*, **32**(6):14671476.

[6] Grange, S.K. and Carslaw, D.C. (2019). Using meteorological normalisation to detect interventions in air quality time series. *Science of The Total Environment*, **653**:578588.

[7] Pelagatti. M. and Maranzano, P. (2021). Nonparametric tests for event studies under cross-sectional dependence. *Quarterly Journal of Finance Accounting*, **59**.

[8] Wang, Y., Finazzi, F., and Fasso, A.,(2021). D-stem v2: A software for modeling functional spatio-temporal data. *Journal of Statistical Software*, **99**(10):1 29.

[9] Zha, D., Zhao, T., Kavuri, A.S., Wu, F., and Wang, Q. (2018) An event study analysis of price ad- justment of refined oil and air quality in china. *Environmental Science and Pollution Research*, **25**(34):34236 34246.

[10] Zhang, X., Yu, L., Wang, S., and Lai, K.K. (2009). Estimating the impact of extreme events on crude oil price: An emd-based event analysis method. *Energy Economics*, **31**(5):768778.

Risk analysis of a log-Gaussian Cox process under scenarios of separability and non-separability

A. Medialdea[1,*] , J.M. Angulo[1] and J. Mateu[2]

[1]*Department of Statistics and Operations Research, University of Granada, Granada, Spain; amedialdea@ugr.es, jmangulo@ugr.es* [2]*Department of Mathematics, University Jaume I, Castellón, Spain; mateu@mat.uji.es*
[*]*Corresponding author*

Abstract. *Log-Gaussian Cox processes define a flexible class of spatio-temporal models which allow the description of a wide variety of dependency effects in point patterns. In this context, the analysis of a spatio-temporal point pattern, corresponding to observed forest fires in Nepal, is performed under two scenarios of separability and non-separability of the spatial and temporal dimensions. The predictive performance of each model is compared graphically using risk maps.*

Keywords. *Conditional simulation; log-Gaussian Cox process; non-separability; risk maps; spatio-temporal point process.*

1. Introduction

Log-Gaussian Cox processes define a class of doubly stochastic Poisson processes useful to model point patterns which are environmentally driven. The clustering structure observed in these patterns can be described by the inclusion of random heterogeneities in an unobservable intensity function. Typically, spatio-temporal point processes have been modeled considering a separable structure of space and time dimensions. Here, we propose a model which combines a non-separable structure for the first-order intensity function and a non-separable correlation structure for the underlying random field, allowing to reflect the interaction of spatial and time dimensions present in the pattern.

2. Statistical model

We consider a two dimensional spatial region S corresponding to Nepal and discrete times $T = \{1,2,...,T\}$ divided in daily intervals from 2012-02-01 to 2016-04-26, retaining the last six days for testing the model. The point process formed by the occurrence of forest fires in $S \times T$ is assumed to follow a log-Gaussian Cox process driven by an intensity function which, according to the specifications in [1], takes the form:

$$\Lambda(\mathbf{u},t) = \lambda(\mathbf{u},t)\exp\{G(\mathbf{u},t)\},$$

where G is a Gaussian process with $E[G(\mathbf{u},t)] = \mu$, $E[\exp\{G(\mathbf{u},t)\}] = 1$ and variance-covariance matrix

$Cov\left(G\left(\mathbf{u},t\right),G\left(\mathbf{v},s\right)\right)=C\left(\left\|\mathbf{u}-\mathbf{v}\right\|,\left|t-s\right|\right)=C\left(r,h\right).$

Separable and non-separable spatio-temporal scenarios have been considered as shown in table 1.

Scenario	Deterministic intensity	Covariance model
Separable	$\lambda\left(\mathbf{u},t\right)=\lambda\left(\mathbf{u}\right)\mu\left(t\right)$	$C\left(r,h\right)=\sigma^2\left(1+\left(\frac{r}{\alpha}\right)^{\gamma_s}+\left(\frac{h}{\beta}\right)^{\gamma_t}\right)$
Non-separable	$\lambda\left(\mathbf{u},t\right)=\lambda_t\left(\mathbf{u}\right)\mu\left(t\right)$	$C\left(r,h\right)=\sigma^2\left(1+\left(\frac{r}{\alpha}\right)^{\gamma_s}+\left(\frac{h}{\beta}\right)^{\gamma_t}\right)^{\delta}$

Table 1: Deterministic intensity functions and covariance models of the Gaussian random fields assumed to model the LGCP for the separable and non-separable scenarios. Iaco-Cesare covariance model has been considered for both cases, with $\delta=1$ for the separable scenario.

The temporal intensity $\mu\left(t\right)$ represents the expected number of forest fires that occurred in the spatial domain S in the time interval t. Under both separable and non-separable assumptions it is estimated through a generalized linear model, using precipitation, wind speed and temperature as linear regressors; we use the month of the year as a dummy variable taking as reference January and include a seasonal component in the model corresponding to a periodicity of a year. Figure 1 shows the observed and predicted number of forest fires. The spatial component $\lambda\left(\mathbf{u}\right)$ describes the spatial variation in the intensity of the forest fires. Its estimation

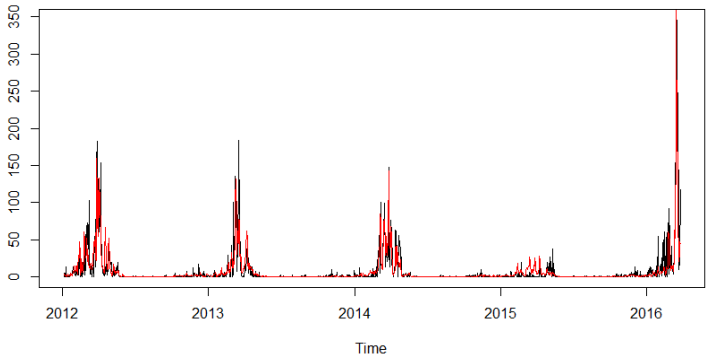


Figure 1: Observed daily forest fires (black line) and fitted (red lines) from adjusted general linear model.

for the separable scenario is performed through a linear point process model using as regressors some climate covariates such as precipitation, wind speed and temperature. The estimation of $\lambda_t\left(\mathbf{u}\right)$ for the non-separable scenario is performed using weighted kernel density estimation, implemented in `spatstat` and based on the method proposed by [5] and [3]; for a time period T, a day t and a location $\mathbf{u}$, we have

$$\lambda_t\left(\mathbf{u}\right)=\frac{\sum_{h\in T}\sum_{i=1}^{y_h}w\left(h,t\right)K_N\left(\mathbf{u},\mathbf{u}_{h,i},s\right)}{\sum_{h\in T}\sum_{i=1}^{y_h}w\left(h,t\right)},$$

with $w(h,t)$ being the product of the temporal and spatial covariates weights for the $\mathbf{u}_{h,i}$ forest fire and $K_N(\mathbf{u},\mathbf{u}_{h,i},s)$ the Jones-Diggle corrected network kernel function with bandwidth s.

The variance and covariance parameters of the covariance models are estimated following the procedure proposed in [4], by minimizing the quadratic distance between the empirical and parametric forms of the pair correlation function $g(r,h)$ with respect of the vector of parameters θ,

$$\int_{r_0}^{r_{max}} \int_{h_0}^{h_{max}} \{\hat{g}(r,h) - g(r,h;\theta)\}^2 \, dhdr,$$

where $r_{max} = 1/4 \cdot \text{lag}_{max_s}$, $h_{max} = 1/4 \cdot \text{lag}_{max_t}$, with lag_{max_s} and lag_{max_t} being the maximum space and time lags, respectively (see [2]). We use an Epanechnikov kernel to compute the spatial component and a box kernel for the temporal component, setting r_0 and h_0 slightly greater that the spatial and temporal bandwidths. The estimates for the separable and non-separable covariance models are shown in table 2.

Scenario	$\hat{\sigma}^2$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}_s$	$\hat{\gamma}_t$	$\hat{\delta}$
Separable	0.04	1.26	11.77	1.99	0.89	
Non-separable	6.78	58.55	158.83	0.42	1.21	6.38

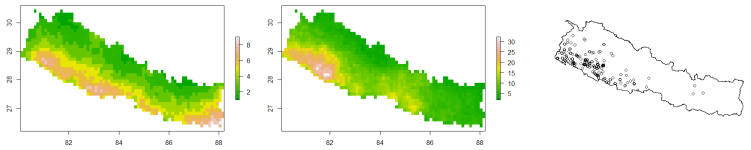
Table 2: Parameter estimates of the covariance families for the separable and non-separable scenarios.

3. Conditional inference and risk maps

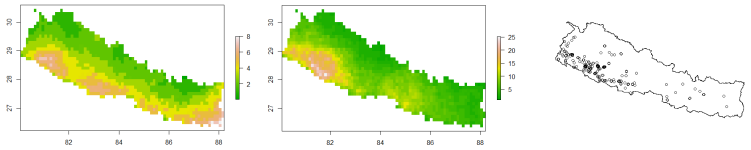
Estimations of the intensity function have been performed using the deterministic models considering six future time periods, whilst for the stochastic component 100 point patterns have been simulated for each time through conditional inference. The predictive accuracy of each model has been compared graphically by means of risk maps, as shown in figure 2, using the Value-at-Risk (VaR) measure computed over the quadrat counts of the simulated points patterns through a sliding window.

4. Conclusions

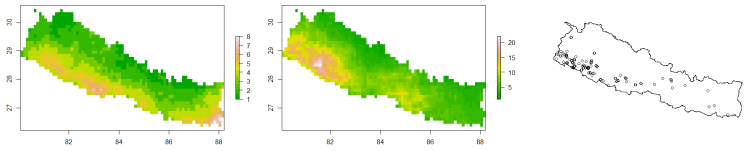
A comparative analysis of separable and non-separable LGCP models is presented for describing and predicting the spatio-temporal structure of forest fires occurred in Nepal from 2012-02-01 to 2016-04-26. The results of the deterministic component of the model, which is responsible of the spatio-temporal structure of the underlying intensity, show an improvement in the spatio-temporal dynamics for non-separable models. As a consequence, for this model, the spatio-temporal interaction allows to observe an increase of the number of forest fires from April to June with more incidence in areas with high wind speed and mild temperatures, and decreasing in areas with high precipitations and extreme temperatures. As for the stochastic component, the



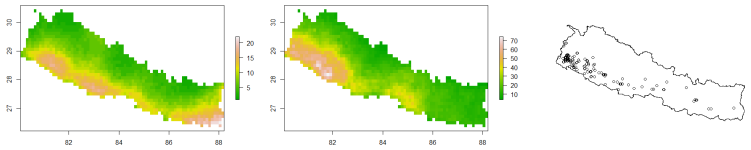
(a) Time: 2016-04-21



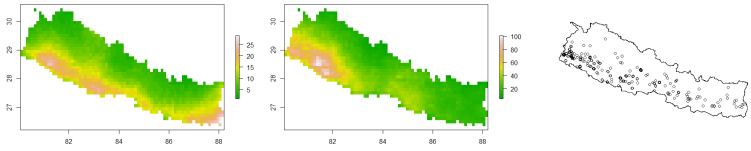
(b) Time: 2016-04-22



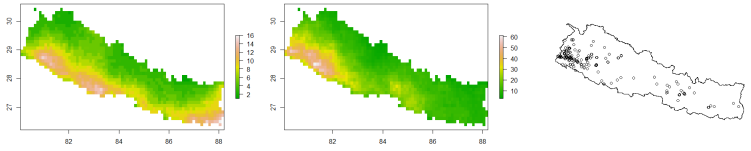
(c) Time: 2016-04-23



(d) Time: 2016-04-24



(e) Time: 2016-04-25



(f) Time: 2016-04-26

Figure 2: Value-at-Risk (VaR) of conditional points patterns for six future time periods. Separable scenario is shown on the left, non-separable scenario at the center and the observed events on the right.

goodness of fit of each covariance model has been tested using and envelope of the non-parametric spatio-temporal K-function showing a better performance for the non-separable scenario.

Acknowledgments

A. Medialdea and J.M. Angulo were partially supported by MCIU/AEI/ERDF, UE grant PGC2018-098860-

B-I00, grant A-FQM-345-UGR18 cofinanced by ERDF Operational Programme 2014-2020 and the Economy and Knowledge Council of the Regional Government of Andalusia, Spain, and grant CEX2020-001105-M MCIN/AEI/10.13039/501100011033. J. Mateu was partially funded by grant PID2019-107392RB-I00/AEI/10.13039/501100011033 from the Spanish Ministry of Science and Innovation and grant UJI-B2018-04 from University Jaume I.

References

[1] Diggle, P.J.; Moraga, P.; Rowlingson, B.; Taylor, B.M. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Statistical Science* **28**, 542–563.

[2] Gabriel, E. (2014). Estimating second-order characteristics of inhomogeneous spatio-temporal point processes. In: *Methodology and Computing in Applied Probability*. **16**, 411–431.

[3] Gilardi, A.; Borgoni, R.; Mateu, J. (2021). A non-separable first-order spatio-temporal intensity for events on linear networks: an application to ambulance interventions. In: *arXiv preprint* **arxiv.2106.00457**.

[4] Siino, M.; Adelfio, G.; Mateu, J. (2018). Joint second-order parameter estimation for spatio-temporal log-Gaussian Cox processes. *Sthochastic Environmental Research and Risk Assessment* **32**, 3525–3539.

[5] Zhou, Z.; Matteson, D.S.; Woodard, D.B.; Henderson, S.G.; Micheas, A.C. (2015). A spatio-temporal point process model for ambulance demand. *Journal of the American Statistical Association* **110.509**, 6–15.

Approximation of the cross-covariance functions of multivariate spatial processes through the direct covariances

P. García-Soidán[1] and R. Menezes[2,*]

[1]*Department of Statistics and Operations Research & Atlantic Research Center for Information and Communication Technologies (atlantTIC), University of Vigo, Spain; pgarcia@uvigo.es*
[2]*Department of Mathematics and Research Centre of Mathematics, University of Minho, Portugal; rmenezes@math.uminho.pt*
[*]*Corresponding author*

Abstract. *The assumption of stationarity simplifies tackling inference problems for spatial data. However, an appropriate characterization of the correlation structure of the underlying process is usually required and this issue is particularly complex in the multivariate scenario. For instance, under second-order stationarity, the main difficulties are due to the number of covariance functions that must be estimated, as well as to the relationships among them, which convey that the characterization of these functions cannot be accomplished in an independent way. Different approaches have been suggested in the statistics literature to overcome the aforementioned drawbacks, although, to our knowledge, none of these proposals aims to solely involve the direct covariances to address the estimation of the whole dependence structure. This is the main goal of the current work, which intends to explore the suggested alternative for approximation of symmetric cross-covariances and, additionally, to quantify the error committed in the estimation, when proceeding in this way. In fact, the resulting error directly depends on the correlation degree between the variables involved.*

Keywords. *Covariance; Multivariate process; Stationarity.*

1. Introduction

Prediction of a spatial variable at unsampled locations can be enhanced by including data from other variables, correlated with the target one. However, the incorporation of auxiliary variables demands taking appropriate decisions, regarding the characteristics of the underlying process and the tools selected to model them. Thus, under stationarity of the involved variables, the prediction goal could be tackled through the cokriging techniques [3], although they demand an adequate characterization of the dependence structure, which is a specially complex task in the multivariate scenario [4]. On one hand, a drawback is the number of functions that must be estimated for the specification of the underlying correlation. For instance, if $p-1$ secondary variables are considered for prediction of the main one and the assumption of second-order stationarity is supposed for the multivariate process, estimation of $p(p+1)/2$ direct and cross-covariance functions is needed. An additional inconvenience is that the aforementioned covariances cannot be obtained in an independent way, because of the relationships among these functions.

To address the above-mentioned problems, different strategies to model the multivariate correlation have

been introduced in the statistics literature. An interesting review of the main proposals is presented in [2]. As far as we know, the proposed tools are not focused on solely involve the direct covariances in the estimation of the whole correlation structure, which is the aim of the current work. The implementation of approaches of this kind is supported by the fact that the approximation of the direct covariances is quite simpler than that of the cross-covariances, particularly for heterotopic data. Taking this into account, we suggest estimating each cross-covariance through an appropriate linear combination of the direct covariances of the involved variables. This way of proceeding requires that the target cross-covariance be symmetric, since this property holds for the direct covariances. A further step of our research is to provide a measure of the error committed in the proposed approximation of each cross-covariance, which turns out to be dependent on the correlation degree between the corresponding variables, as expected.

2. Main results

Denote by $\{Z(\mathrm{s}) = (Z_1(\mathrm{s}), ..., Z_p(\mathrm{s})) : \mathrm{s} \in D \subset \mathbb{R}^d\}$ a p-variate stochastics process that satisfies the second-order stationarity assumption, so that:

(a) $E[Z_i(\mathrm{s})] = \mu_i$, for all $\mathrm{s} \in D$ and some $\mu_i \in \mathbb{R}$, for all $i = 1, ..., p$.

(b) $Cov[Z_i(\mathrm{s}), Z_j(\mathrm{s}+\mathrm{t})] = C_{ij}(\mathrm{t})$, for all $\mathrm{s}, \mathrm{s}+\mathrm{t} \in D$ and some function C_{ij}, for all $i, j = 1, ..., p$.

Functions C_{ii} and C_{ij} are referred to as direct covariance and cross-covariance, respectively, for $i, j = 1, ..., p$, with $j \neq i$.

It follows from condition **(b)** that:

$$Cov[Z_i(\mathrm{s}), Z_j(\mathrm{s})] = C_{ij}(0) = \sigma_{ij}$$
$$Var[Z_i(\mathrm{s})] = C_{ii}(0) = \sigma_{ii} = \sigma_i^2$$

for all $\mathrm{s} \in D$ and for all $i, j = 1, ..., p$.

The behavior of the cross-covariances differs from that of the direct covariances [1]. For instance, C_{ii} is a symmetric function, whereas C_{ij} may not satisfy this property, for $i \neq j$. However, $C_{ji}(\mathrm{t}) = C_{ij}(-\mathrm{t})$, for all t, even under asymmetry of function C_{ij}. This yields that the total specification of the dependence structure of the p-variate process requires the estimation of p covariance functions C_{ii} and $(p^2 - p)/2$ cross-covariances C_{ij}.

On the other hand, there are some relationships among the direct and cross-covariances, such as the one given below:

$$C_{ij}(\mathrm{t})^2 \leq \sigma_i^2 \sigma_j^2$$

derived from the Cauchy-Schwarz inequality. Then, the characterization of the different covariance functions cannot be accomplished in an independent way.

Several mechanisms have been proposed to specify the multivariate correlation structure, based on using

nonparametric or parametric techniques. The former methods are typically employed in a first step of the estimation process and then followed by parametric approaches, which encompass the application of the maximum likelihood or the least squares methods on valid models, specifically designed to account for the relationship among the variables. However, we suggest using a different strategy that solely requires the approximation of the direct covariances, so that the remaining functions are derived from them and the resulting error can be quantified. To derive our approach, the cross-covariance functions will be assumed to be symmetric.

For the latter purpose, write:

$$C_{ij}(\mathrm{t}) = \alpha C_{ii}(\mathrm{t}) + \beta C_{jj}(\mathrm{t}) + R_{ij}(\mathrm{t})$$

for some parameters $\alpha, \beta \in \mathbb{R}$. Our aim is to determine optimal values for α and β in the latter relation, thus leading to a negligible error term $R_{ij}(\mathrm{t})$.

By exploring distinct alternatives, we obtain that $C_{ij}(\mathrm{t}) = D_{ij}(\mathrm{t}) + R_{ij}(\mathrm{t})$, with:

$$D_{ij}(\mathrm{t}) = \frac{\rho_{ij}(0)}{1+\rho_{ij}(0)^2}\left(\frac{\sigma_j}{\sigma_i}C_{ii}(\mathrm{t}) + \frac{\sigma_i}{\sigma_j}C_{jj}(\mathrm{t})\right) \tag{1}$$

where $\rho_{ij}(0) = \frac{C_{ij}(0)}{\sigma_i\sigma_j}$, $|R_{ij}(\mathrm{t})| \leq \sigma_i\sigma_j e_{ij}$ and $e_{ij} = \frac{1-\rho_{ij}(0)^2}{1+\rho_{ij}(0)^2}$.

Table 1 displays some values of e_{ij}, for a selection of cross-correlation degrees. According to these results, when the cross-covariance C_{ij} is approximated through D_{ij}, given in (1), the maximum error committed represents less than 22% of the maximum variability of the variables involved, provided that the cross-correlation between the variables (in absolute value) amounts to 80%. The aforementioned error decreases to 10.5% of the maximum variability, for $|\rho_{ij}(0)|$ equaling 90%.

$\lvert\rho_{ij}(0)\rvert$	e_{ij}
0.9	0.1050
0.8	0.2195
0.7	0.3423

Table 1: Values of e_{ij} for different cross-correlation degrees.

In view of the above results, for strongly correlated variables Z_i and Z_j, an accurate approximation of the cross-covariance C_{ij} can be derived from D_{ij} and, therefore, from C_{ii} and C_{jj}. Then, the practical implementation of this approach demands obtaining appropriate specifications of the two direct covariances, together with estimates of $C_{ij}(0)$, σ_i and σ_j.

Acknowledgments

The first author's work has been partially funded by the Spanish Ministry of Science and Innovation project PID2020-113979RB-C22, by the ERDF and by the Xunta de Galicia (Spain), under project ED431C 2019/25 SC7-GRC 2019 and under agreement for funding atlantTIC (Atlantic Research Center for Information and Communication Technologies). The second author acknowledges to FCT Foundation (Fundação para a Ciência e Tecnologia) for partially funding this research through projects PTDC/MAT-STA/28243/2017, UIDB/00013/2020 and UIDP/00013/2020.

References

[1] Chilès J. P., Delfiner P. (2012). *Geostatistics: Modeling spatial uncertainty*. Wiley. New York.

[2] Genton M. G., Kleiber W. (2015). Cross-covariance functions for multivariate geostatistics. *Statistical Science* **30**, 147–163.

[3] Journel A. G., Huijbregts C. J. (1978). *Mining geostatistics*. Academic Press. London.

[4] Wackernagel H. (2003). *Multivariate geostatistics: An introduction with applications*. Springer. Berlin.

Statistical and machine learning models for landscape-level prediction of forest fungal productivity

A. Morera[1,2*] , J. Martínez de Aragón[3] , J.A. Bonet[1,2] , J. Liang[4] and S. de-Miguel[1,2]

[1]*Department of Crop and Forest Sciences, University of Lleida, Av. Alcalde Rovira Roure 191, E-25198 Lleida, Spain; morera.marra@gmail.com; jantonio.bonet@udl.cat; sergio.demiguel@udl.cat*
[2]*Joint Research Unit CTFC-AGROTECNIO-CERCA, 25280 Solsona, Spain*
[3]*Forest Science and Technology Centre of Catalonia, Ctra. Sant Lloren de Morunys km 2, 25280 Solsona, Spain; mtzda@ctfc.cat*
[4]*Forest Advanced Computing and Artificial Intelligence Laboratory, Department of Forestry and Natural Resources, Purdue University, West Lafayette, IN, 47907, USA; jjliang@purdue.edu*
[*]*Corresponding author*

Keywords. *Machine Learning; Statistical Modelling; Biogeography; Climate; Forest; Fungi*

Constant developments in analytical methods and tools including statistical and machine learning models, highlight the need for systematically revising the predictive performance of different spatial-temporal modeling techniques within the field of ecological research. This study compares different statistical and machine learning models for predicting fungal productivity biogeographical patterns from climatic data as a case study for the thorough assessment of the performance of alternative modeling approaches in order to provide accurate and ecologically-consistent predictions [1]. Specifically, we evaluated and compared two statistical modeling techniques, namely, geographically weighted regression and generalized linear mixed-effects models, and four techniques based on different machine learning algorithms, namely, artificial neural networks, random forest, extreme gradient boosting and support vector machine. In addition, this study aims to be a starting point to study fungal productivity from a spatial and temporal perspective; a field little studied to date and, at the same time, of great relevance due to the important role played by fungi in natural ecosystems. We evaluated our models based on a robust, systematic methodology combining random, spatial and environmental blocking together with the assessment of the ecological consistency of spatially-explicit model predictions according to the current scientific knowledge.

We found that fungal productivity predictions were sensitive to the modeling approach and the number of predictors used (Figure 1). In the same vein, the importance assigned to different predictors varied between modeling approaches. Random forest and extreme gradient boosting (both decision tree-based models) performed the best in the prediction of fungal productivity in both in sampling-like environments as well as in extrapolation beyond the spatial and climatic range of the modeling data (increasing prediction accuracy by more than 10% compared to other machine learning approaches, and by more than 20% compared to statistical models). Moreover, the spatial estimates of the decision tree-based models resulted in higher ecological consistence across the landscape. We therefore recommend the use of these models for further research involving the biogeographical patterns and spatial-temporal prediction of fungal productivity. We show that proper variable

selection is crucial to create robust models for extrapolation in biophysically differentiated areas, a task that becomes more evident when trying to predict complex environmental patterns shaped by a large number of drivers. This allows for reducing the dimensions of the model predictors space, resulting in higher similarity between the modeling data and the new environmental conditions. Finally, when dealing with models trained with data sampled annually on the same set of plots (a common case in fungal productivity studies), environmental cross-validation is more suitable than spatial or random cross-validation, resulting in more realistic characterization of the prediction error, also within the context of changing environmental conditions due to global change [2].

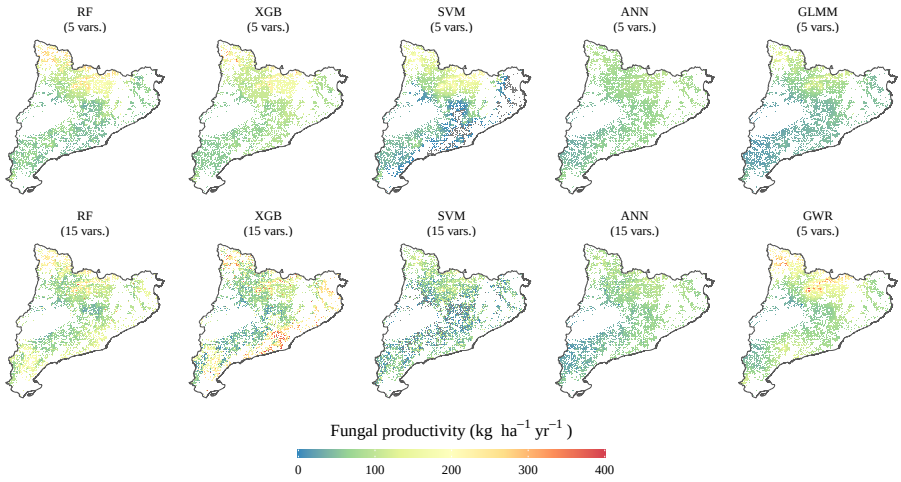


Figure 1: Landscape-level prediction of total annual fungal productivity, using random forest (RF), extreme gradient boosting (XGB), support vector machine (SVM), artificial neural network (ANN), generalized linear mixed models (GLMM) and geographically weighted regression (GWR), trained/fitted with 5 and 15 variables (number of variables is shown between brackets). Modified from [1].

References

[1] Morera A, Martnez de Aragn J, Bonet JA, Liang J, de-Miguel S. 2021. Performance of statistical and machine learning-based methods for predicting biogeographical patterns of fungal productivity in forest ecosystems. *Forest Ecosystems*, **8**, 21.

[2] Morera A, Martnez de Aragn J, De Cceres M, Bonet JA, de-Miguel S. 2022. Historical and future spatially-explicit climate change impacts on mycorrhizal and saprotrophic macrofungal productivity in Mediterranean pine forests. *Agricultural and Forest Meteorology*, **319**, 108918.

Classification in point patterns on linear networks under clutter

J.F. Díaz-Sepúlveda[1] , N. D'Angelo[2] , G. Adelfio[2] , J.A. González[3] and F.J. Rodríguez-Cortés[1,*]

[1]*Escuela de Estadística, Universidad Nacional de Colombia, Medellín, Colombia; jfdiazs0@unal.edu.co, rrodriguezc@unal.edu.co*
[2]*Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo, Palermo, Italy; nicoletta.dangelo@unipa.it, giada.adelfio@unipa.it*
[3]*Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia; jonathan.gonzalez@kaust.edu.sa*
[*]*Corresponding author*

Abstract. *The problem of features detection under present of clutter in point process on linear networks establishes a methodological and computational challenge with multiple kind of applications as traffic accidents among other. Previous works related to the same topical but developed in more simpler geometries tackles the issue of the clutter removal through the distance of nearest-neighbour and show good results with high classification rates. We extend this procedure to the linear networks motivated by the classification of the traffic accidents on the road network of a city. Simulations demonstrate the performance of the method.*

Keywords. *Linear network; Linear Point process; classification; Feature and Clutter; EM-algorithm.*

1. Introduction

Traffic accidents are one of the 10 leading causes of death worldwide and are the first cause of death for people between 15 and 29 years of age. According to the World Health Organization (WHO), 2.5% of deaths were caused by a traffic accident and approximately a third left people injured, these injuries being an important cause of disability worldwide. For this reason, it is necessary to implement new tools for the analysis of this phenomenon, based on which strategies can be formulated to improve road safety and encourage the formulation of public policies to face the high mortality rate in the country due to traffic accidents [7].

The spatial statistics has suffered an extraordinary methodological and computational breakthrough over the past in the last two decades centered on generalization and extension of its own bases to geometric spaces more complex that allow a better statistical analysis of new kinds of spatial data. Identifying features in the presence of clutter is of great interest in spatial point pattern analysis. Unlike the planar case, which is primarily straightforward, identifying features on networks through visual inspection is far from direct since the complexity of this domain [2]. We extend and implement computationally to the linear networks the methodology proposed by [4], which model the Kth nearest neighbor distances of points as a mixture distribution and estimated its parameters using the EM algorithm to classify them as clutter or feature.

2. Procedure for clutter removal on linear networks

The extension of the approach in [4] to the linear networks is consider assuming that the clutter and features are distributed as two homogeneous Poisson point process on a linear network. Features are limited to a sub-network and overlaid to the clutter which is generate on the whole network (see Figure 1), the resulting process is a Poisson point process with piece-wise constant intensity on the linear network.

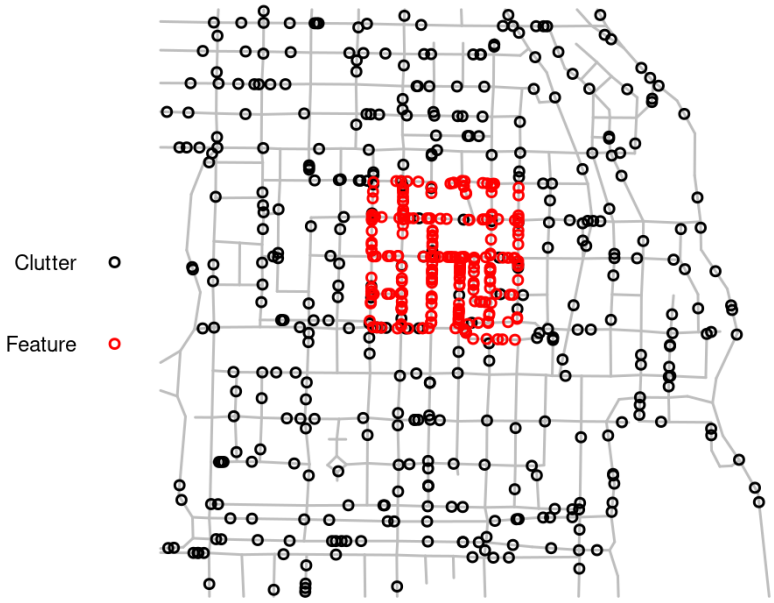


Figure 1: Clutter point pattern (black points) and feature point pattern (red points) are simulated from homogeneous Poisson point process with intensities $\lambda_c = 0.013$ and $\lambda_f = 0.068$, respectively, on linear network `chicago` of the package `spatstat.data`.

2.1 Distances distribution

Let $b_L(u,r)$ the disc with center u and radius r inside the network, and $|b_L(u,r)|$ the length of this disc. For all $u \in L$ and $x \in [0,\infty)$, We propose the following approach for Kth nearest neighbour distribution on the linear network

$$\mathbb{P}(D_K^L \geq x) = \sum_{j=0}^{K-1} \frac{e^{-\lambda x}(\lambda x)^j}{j!} = 1 - F_{D_K^L}(x), \tag{1}$$

where $\mathbb{P}(D_K^L \geq x)$ is the probability that the Kth nearest neighbour point falls out of $b_L(u,x)$ with $|b_L(u,x)| = x$, for more detail and fundamentals about point process on linear network see [2]. Therefore, the density $f_{D_K^L}(x)$

is

$$f_{D_K^L}(x) = \frac{\lambda\,(\lambda x)^{K-1}\,e^{-\lambda x}}{\Gamma(K)}, \tag{2}$$

hence $D_K^L \sim \Gamma(K,\lambda)$. Thus, the maximum likelihood estimation of λ given the observed values of D_K^L is

$$\hat{\lambda} = \frac{K}{\sum_{i=1}^{n} d_i}, \tag{3}$$

where d_i is the ith observed Kth nearest neighbour distance.

2.2 Mixture modeling

We assume that the clutter and feature come from two types of processes to be classified through a mixture of the corresponding Kth nearest neighbour distances. That is, based on Eq. (2), it is assumed that

$$D_K^L \sim p\Gamma(K,\lambda_1) + (1-p)\Gamma(K,\lambda_2), \tag{4}$$

where λ_1 and λ_2 are intensities of two superimposed homogeneous Poisson processes on the linear network and p is the constant related to the D_K^L distribution. The parameters λ_1, λ_2 and p are estimated using an EM algorithm [6], where the Gamma distribution in equation (2) is accommodated in the expectation step and the maximum likelihood estimation of λ given in equation (3) in the maximization step. The previous development involves a proper value of K in advance chosen. However, we implemented this automatic option using an entropy-type measure of separation introduced in [5] through a segmented regression models [8].

3. Simulation study

We explored several simulation scenarios considering different clutter and features shapes as well as combinations of parameters with very effective results of the proposed classification method. In order to save space, we just present one of the cases consider in the simulation study which corresponds to linear network `chicago` of the package `spatstat.data` [3]. The street network has 338 intersections, 503 uninterrupted segments and a total length of 31,150 feet. The sub-network has 39 intersections, 53 uninterrupted segments and a total length of 2,991 feet [1].

We show the results of the classification method for different linear networks in terms of true-positive rate (TPR), false-positive rate (FPR), and accuracy (ACC), averaged over 100 simulated point patterns with $\mathbb{E}[n_c]$ and $\mathbb{E}[n_f]$ expected number of points for clutter and feature, respectively. In addition, the λ_c and λ_f intensities are reported for clutter and feature, respectively. The accuracy is the proportion of correct predictions (both true-positives and true-negatives) among the total number of cases examined. The simulated point pattern on this linear network is shown in Figure 1. We consider the detection with K equal to $\{5,10\}$, and the automatically selected.

						K		
Stage	λ_c	λ_f	$\mathbb{E}[n_c]$	$\mathbb{E}[n_f]$	Rate	5	10	$\hat{K}$
1	0.032	0.100	1000	300	TPR	0.996	0.998	0.997
					FPR	0.001	0.001	0.001
					ACC	0.642	0.638	0.651
2	0.032	0.067	1000	200	TPR	0.986	0.994	0.991
					FPR	0.003	0.001	0.002
					ACC	0.590	0.546	0.568
3	0.032	0.033	1000	100	TPR	0.927	0.979	0.992
					FPR	0.007	0.002	0.001
					ACC	0.516	0.409	0.337
4	0.016	0.017	500	50	TPR	0.898	0.979	0.983
					FPR	0.010	0.002	0.002
					ACC	0.538	0.405	0.325
5	0.064	0.017	2000	50	TPR	0.749	0.865	0.959
					FPR	0.006	0.003	0.001
					ACC	0.445	0.377	0.239
6	0.128	0.017	4000	50	TPR	0.684	0.765	0.846
					FPR	0.004	0.003	0.002
					ACC	0.425	0.379	0.312

Table 1: Results of the classification method for the `chicago` linear network averaged over 100 simulated Poisson point patterns with two different intensities.

From the Table 1 for $\lambda_c < \lambda_f$ in stages 1 and 2, it is clear that the classification method performs well in terms of true- and false-positive rates for $K = 5$, $K = 10$ and $\hat{K}$. Similarly, this happens for $\lambda_c \approx \lambda_f$ in stages 3 and 4. For $\lambda_c > \lambda_f$ in stages 5 and 6, the classification method works well for $\hat{K}$, but in $K = 5$ and $K = 10$ both have a very low performance than $\hat{K}$. The estimated K automatically improve in terms of TPR with respect to $K = 5$ and $K = 10$ as the classification method is reducing its performance, and so $\hat{K}$ became a good estimate of the K to be selected. In any case, $\hat{K}$ is a good indication of the magnitude of the K to be selected.

4. Conclusions

We have extend a simple and intuitive method for estimating in a linear networks differing densities in a point process. It can be applied without input about the shapes of the feature which is a strength when the shape of the feature is not known. The simulation study shows that the performance of the classification procedure is fairly good under the assumption that the feature is a Poisson process overlaid with a clutter Poisson process.

References

[1] Ang, Q. W., Baddeley, A., Nair, G. (2012). Geometrically corrected second order analysis of events on a linear network, with applications to ecology and criminology. *Scandinavian Journal of Statistics* **39(4)**, 591–617.

[2] Baddeley, A., Nair, G., Rakshit, S., McSwiggan, G., Davies, T. M. (2021). Analysing point patterns on networks - a review. *Spatial Statistics* **42**, 100435.

[3] Baddeley, A., Turner, R., Rubak, E. (2021). *spatstat.data: Datasets for "spatstat" Family.* R package version 2.1-2.

[4] Byers, S., Raftery, A. E. (1998). Nearest-neighbor clutter removal for estimating features in spatial point processes. *Journal of the American Statistical Association* **93(442)**, 577–584.

[5] Celeux, G., Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of classification* **13(2)**, 195–212.

[6] Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39(1)**, 1–38.

[7] Moncada, I. (2018). Spatio-temporal analysis of traffic accidents in Bogotá using point process. Master's Thesis. Universidad Nacional de Colombia. URL: `https://repositorio.unal.edu.co/handle/unal/69640?show=full`.

[8] Muggeo, V. M. R. (2003). Estimating regression models with unknown break-points. *Statistics in Medicine* **22(19)**, 3055–3071.

Nonparametric Conditional Risk Mapping under Heteroscedasticity

R. Fernández-Casal[1,*], S. Castillo-Páez[2] and M. Francisco-Fernández[1]

[1]*Departamento de Matemáticas, Facultad de Informática, Universidade da Coruña, 15071, A Coruña, Spain; ruben.fcasal@udc.es, mariofr@udc.es*
[2]*Departamento de Ciencias Exactas, Universidad de las Fuerzas Armadas, Sangolquí, Ecuador; sacastillo@espe.edu.ec*
[*]*Corresponding author*

Abstract. *In this work, a nonparametric procedure to approximate the conditional probability that a non-stationary geostatistical process exceeds a certain threshold value is proposed. On the contrary to traditional methods, the proposed approach does not require the assumption of constant mean and variance. For this, the process is nonparametrically modeled using an iterative algorithm. The local linear estimator is used to estimate the trend. Moreover, the variability is modeled estimating the conditional variance and the variogram from corrected residuals to avoid the bias. From these estimates, conditional bootstrap replicas are generated combining an unconditional bootstrap algorithm with kriging prediction. The performance of the proposed procedure is analyzed by simulation and with its application to a real data set.*

Keywords. *Conditional simulation; Local linear estimation; Heteroscedasticity; Bootstrap.*

1. Introduction

Suppose that the spatial heteroscedastic process $\{Y(\mathbf{x}), \mathbf{x} \in D \subset \mathbb{R}^d\}$, can be modeled as follows:

$$Y(\mathbf{x}) = \mu(\mathbf{x}) + \sigma(\mathbf{x})\varepsilon(\mathbf{x}),$$

where $\mu(\cdot)$ and $\sigma^2(\cdot)$ correspond to the deterministic functions of trend and variance, respectively, and $\varepsilon(\cdot)$ is a second order stationary process with zero mean, unit variance and correlogram $\rho(\mathbf{u}) = Cov(\varepsilon(\mathbf{x}), \varepsilon(\mathbf{x}+\mathbf{u}))$, $\mathbf{x}, \mathbf{x}+\mathbf{u} \in D$. The goal is, from n observations of the process $\mathbf{Y} = (Y(\mathbf{x}_1), ..., Y(\mathbf{x}_n))^t$ at the sample locations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, nonparametrically estimate the conditional probability

$$r_c(\mathbf{x}^e, \mathbf{Y}) = P(Y(\mathbf{x}^e) \geq c | \mathbf{Y}),$$

where $\mathbf{x}^e$ is an (unobserved) estimation location.

It must be taken into account that the spatial dependence of the process Y depends on the variance and the correlogram of ε, since $Cov(Y(\mathbf{x}), Y(\mathbf{x}+\mathbf{u})) = \sigma(\mathbf{x})\sigma(\mathbf{x}+\mathbf{u})\rho(\mathbf{u})$. In addition, the covariance matrix of the observations $\mathbf{Y}$ can be expressed as:

$$\boldsymbol{\Sigma} = \mathbf{DRD},$$

where $\mathbf{R}$ is the covariance matrix of the errors $\boldsymbol{\varepsilon} = (\varepsilon(\mathbf{x}_1), \ldots, \varepsilon(\mathbf{x}_n))^t$ and $\mathbf{D} = \text{diag}(\sigma(\mathbf{x}_1), \ldots, \sigma(\mathbf{x}_n))$.

As usual, instead of estimating the dependence of the error from the covariogram (or from the correlogram), the error semivariogram is used,

$$\gamma_\varepsilon(\mathbf{u}) = \tfrac{1}{2} Var(\varepsilon(\mathbf{x}) - \varepsilon(\mathbf{x}+\mathbf{u})) = 1 - \rho(\mathbf{u}).$$

The first step of the proposed procedure consists of the nonparametric modeling of the process, for which a slight modification of the iterative algorithm proposed in [5] is applied to obtain nonparametric estimates of $\mu(\cdot)$, $\sigma^2(\cdot)$ and $\gamma(\cdot)$. The local linear estimator (e.g. [3]) is used to derive these approximations, in order to avoid possible misspecification problems. The next step consists in using these estimates in the bootstrap algorithm described in [4] to generate unconditional replicates. Finally, in the conditional bootstrap algorithm, the unconditional replicates are combined with the kriging prediction at the estimation locations to obtain the conditional bootstrap samples.

2. Nonparametric heteroscedastic estimation

The local linear estimator of the trend is obtained by the linear smoothing of $\{(\mathbf{x}_i, Y(\mathbf{x}_i)) : i = 1, \ldots, n\}$, and can be expressed as:

$$\hat{\mu}_{\mathbf{H}}(\mathbf{x}) = \mathbf{e}_1^t \left(\mathbf{X}_\mathbf{x}^t \mathbf{W}_\mathbf{x} \mathbf{X}_\mathbf{x}\right)^{-1} \mathbf{X}_\mathbf{x}^t \mathbf{W}_\mathbf{x} \mathbf{Y} = s_\mathbf{x}^t \mathbf{Y},$$

where $\mathbf{e}_1$ is a vector with 1 in the first entry and 0 in the others, $\mathbf{X}_\mathbf{x}$ is a matrix whose i-th row is equal to $(1, (\mathbf{x}_i - \mathbf{x})^t)$, $\mathbf{W}_\mathbf{x} = \text{diag}\{K_\mathbf{H}(\mathbf{x}_1 - \mathbf{x}), \ldots, K_\mathbf{H}(\mathbf{x}_n - \mathbf{x})\}$, $K_\mathbf{H}(\mathbf{u}) = |\mathbf{H}|^{-1} K(\mathbf{H}^{-1}\mathbf{u})$, K is a d-dimensional kernel function and $\mathbf{H}$ is the bandwidth matrix, which controls the shape and size of the local neighborhood used to estimate $\mu(\mathbf{x})$. It is recommended to use the "bias corrected and estimated generalized cross-validation" (CGCV) criterion, proposed in [3], to select this bandwidth.

On the other hand, the estimation of spatial dependence is usually carried out from the residuals $\mathbf{r} = \mathbf{Y} - \mathbf{S}\mathbf{Y}$, where $\mathbf{S}$ is the *smoothing matrix*, whose i-th row is equal to $\mathbf{s}_{\mathbf{x}_i}^t$. However, it is known that estimates based on these residuals underestimate the variability of the spatial process. Indeed,

$$Var(\mathbf{r}) = \boldsymbol{\Sigma} + \mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^t - \boldsymbol{\Sigma}\mathbf{S}^t - \mathbf{S}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_\mathbf{r}.$$

Likewise, the covariance matrix of the standardized residuals $\tilde{\boldsymbol{\varepsilon}} = \mathbf{D}^{-1}\mathbf{r}$ is given by:

$$Var(\tilde{\boldsymbol{\varepsilon}}) = \mathbf{R} + \mathbf{B} = \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\varepsilon}}}, \tag{1}$$

where $\mathbf{B} = \mathbf{D}^{-1}\left(\mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^t - \boldsymbol{\Sigma}\mathbf{S}^t - \mathbf{S}\boldsymbol{\Sigma}\right)\mathbf{D}^{-1}$.

From (1), it can be verified that:

$$Var\left(r(\mathbf{x}_i)/\sqrt{1+b_{ii}}\right) = \sigma^2(\mathbf{x}_i),$$

$$Var\left(\tilde{\varepsilon}(\mathbf{x}_i) - \tilde{\varepsilon}(\mathbf{x}_j)\right) = Var\left(\varepsilon(\mathbf{x}_i) - \varepsilon(\mathbf{x}_j)\right) + b_{ii} + b_{jj} - 2b_{ij},$$

where $r(\mathbf{x}_i)$ is the i-th term of the vector $\mathbf{r}$, b_{ij} is the (i,j)-th element of the matrix $\mathbf{B}$ and $\tilde{\varepsilon}(\mathbf{x}_i) = r(\mathbf{x}_i)/\sigma(\mathbf{x}_i)$ is the i-th component of $\tilde{\boldsymbol{\varepsilon}}$.

From these results, joint nonparametric estimates for the variance and the variogram are obtained using the following iterative algorithm:

1. Obtain $\hat{\mu}_{\mathbf{H}}(\mathbf{x})$ and calculate the corresponding residuals $\mathbf{r}$.
2. Assuming homoscedasticity, $\hat{\mathbf{D}} = \mathbf{I}$ is taken, where $\mathbf{I}$ is the identity matrix, to then obtain a pilot estimator of the error semivariogram $\hat{\gamma}_\varepsilon^0$, using the linear smoothing of $\left(||\mathbf{x}_i - \mathbf{x}_j||, (r(\mathbf{x}_i) - r(\mathbf{x}_j))^2\right)$.
3. Obtain a pilot estimate of $\hat{\mathbf{R}}$ from $\hat{\gamma}_\varepsilon^0$ by fitting a valid Shapiro-Botha variogram model (e.g. [6]).
4. Calculate $\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{D}}\hat{\mathbf{R}}\hat{\mathbf{D}}$ and $\hat{\mathbf{B}} = \hat{\mathbf{D}}^{-1}\left(\mathbf{S}\hat{\boldsymbol{\Sigma}}\mathbf{S}^t - \hat{\boldsymbol{\Sigma}}\mathbf{S}^t - \mathbf{S}\hat{\boldsymbol{\Sigma}}\right)\hat{\mathbf{D}}^{-1}$.
5. Obtain the estimator $\hat{\sigma}^2$ by linear smoothing of $\left(\mathbf{x}_i, r_i^2/(1+\hat{b}_{ii})\right)$ and update $\hat{\mathbf{D}} = \mathrm{diag}(\hat{\sigma}(\mathbf{x}_1), \ldots, \hat{\sigma}(\mathbf{x}_n))$.
6. Compute $\hat{\boldsymbol{\varepsilon}} = \hat{\mathbf{D}}^{-1}\mathbf{r}$ and get an updated version of the variogram estimator $\hat{\gamma}_\varepsilon$ by linear smoothing of $\left(||\mathbf{x}_i - \mathbf{x}_j||, (\hat{\varepsilon}(\mathbf{x}_i) - \hat{\varepsilon}(\mathbf{x}_j))^2 - \hat{b}_{ii} - \hat{b}_{jj} + 2\hat{b}_{ij}\right)$, where $\hat{b}_{ij}$ es el (i,j)-th element of the bias matrix $\hat{\mathbf{B}}$ and $\hat{\varepsilon}(\mathbf{x}_i)$ the i-th component of $\hat{\boldsymbol{\varepsilon}}$.
7. Estimates $\hat{\sigma}^2$ and $\hat{\gamma}_\varepsilon$ are rescaled so that $\widehat{Var}(\varepsilon) = 1$.
8. A new estimate of $\hat{\mathbf{R}}$ is obtained from a Shapiro-Botha model fitted to the rescaled version of $\hat{\gamma}_\varepsilon$ and steps 4-8 are repeated until convergence is obtained.

This procedure is a slight modification of the one proposed in [2], in the sense that valid variogram models are used instead of nonparametric estimates (in steps 3 and 8) and step 7 is added.

3. Heteroscedastic unconditional bootstrap

The following iterative procedure, a modification of the one proposed in [1], allows generating unconditional bootstrap replicates $Y_{NS}^*(\mathbf{x}_\alpha^e)$ for the different estimation locations $\{\mathbf{x}_\alpha^e : \alpha = 1, \ldots, n_0\}$:

1. Using the procedure described in the previous section:
   (a) Obtain $\hat{\mu}_{\mathbf{H}}(\mathbf{x})$, the corresponding residuals $\mathbf{r}$, $\hat{\sigma}^2(\mathbf{x})$ (from the final step) and the pilot and final semivariogram estimates $\hat{\gamma}_\varepsilon^0$ and $\hat{\gamma}_{\hat{\varepsilon}}$, respectively.
   (b) Construct the matrix $\hat{\mathbf{R}}_0$ from the pilot variogram $\hat{\gamma}_\varepsilon^0$ (assuming homoscedasticity), and find the matrix $\mathbf{L}_0$ such that $\hat{\mathbf{R}}_0 = \mathbf{L}_0\mathbf{L}_0^t$, using the Cholesky decomposition.
   (c) Compute $\hat{\mathbf{R}}_\alpha$ corresponding to $\mathbf{x}_\alpha^e$ using $\hat{\gamma}_\varepsilon$, and construct $\mathbf{L}_\alpha$ such that $\hat{\mathbf{R}}_\alpha = \mathbf{L}_\alpha\mathbf{L}_\alpha^t$.
   (d) Construct the "uncorrelated" errors $\mathbf{e} = \mathbf{L}_0^{-1}\mathbf{r}$ and standardize them.

2. Generate the unconditional bootstrap replicas as follows:

   (a) Obtain the independent bootstrap residuals of size n_0 from $\mathbf{e}$, denoted by $\mathbf{e}^*$.

   (b) Compute the unconditional bootstrap residuals $\varepsilon^*_{NC} = \mathbf{L}_\alpha \mathbf{e}^*$.

   (c) Construct the unconditional bootstrap replicas $Y^*_{NC}(\mathbf{x}^e_\alpha) = \hat{\mu}_\mathbf{H}(\mathbf{x}^e_\alpha) + \hat{\sigma}(\mathbf{x}^e_\alpha)\varepsilon^*_{NC}(\mathbf{x}^e_\alpha),\ \alpha = 1,\ldots,n_0$, being $\varepsilon^*_{NC}(\mathbf{x}^e_\alpha)$ the i-th component of the vector ε^*_{NC}.

Note that these unconditional bootstrap replicates do not necessarily coincide with the observed values at the observation locations (e.g. [2], Section 7.3.1). Therefore, it is not recommended to use this algorithm to approximate the conditional probability. However, for the conditional bootstrap algorithm, it would not be necessary to obtain replicas of the entire process (step 2-c), only of the heteroscedastic errors

$$\delta^*_{NC}(\mathbf{x}^e_\alpha) = \hat{\sigma}(\mathbf{x}^e_\alpha)\varepsilon^*_{NC}(\mathbf{x}^e_\alpha)$$

4. Heteroscedastic conditional bootstrap

To generate the conditional bootstrap replicates, a similar procedure to that proposed by [2] was used. This procedure combines unconditional spatial simulation techniques with kriging methods. The proposed approach consists of the following steps:

1. Generate the unconditional bootstrap replicates using the procedure described in the previous section, both at the estimation locations $\delta^*_{NC}(\mathbf{x}^e_\alpha)$, $\alpha = 1,\ldots,n_0$, as well as in the sample locations $\delta^*_{NC}(\mathbf{x}_i)$, $i = 1,\ldots,n$.

2. Using simple kriging, obtain the predictions $\hat{\delta}(\mathbf{x}^e_\alpha)$ and $\hat{\delta}^*_{NC}(\mathbf{x}^e_\alpha)$ from the observed residuals $r(\mathbf{x}_i)$ and the unconditional heteroscedastic errors $\delta^*_{NC}(\mathbf{x}_i)$, respectively.

3. Calculate conditional bootstrap heteroscedastic errors $\delta^*_{CS}(\mathbf{x}^e_\alpha) = \hat{\delta}(\mathbf{x}^e_\alpha) + \left(\delta^*_{NC}(\mathbf{x}^e_\alpha) - \hat{\delta}^*_{NC}(\mathbf{x}^e_\alpha)\right)$.

4. Construct the conditional bootstrap replicates $Y^*_{CS}(\mathbf{x}^e_\alpha) = \hat{\mu}_\mathbf{H}(\mathbf{x}^e_\alpha) + \delta^*_{CS}(\mathbf{x}^e_\alpha)$.

5. Repeat steps 1 to 4 a large number B of times, to get $Y^{*(1)}_{CS}(\mathbf{x}^e_\alpha),\ldots,Y^{*(B)}_{CS}(\mathbf{x}^e_\alpha)$.

6. Finally, the conditional probability is estimated by: $\hat{r}_c(\mathbf{x}^e_\alpha, \mathbf{Y}) = \frac{1}{B}\sum_{j=1}^{B} \mathbb{I}\left(Y^{*(j)}_{CS}(\mathbf{x}^e_\alpha) \geq c\right)$, where $\mathbb{I}(\cdot)$ represents the indicator function.

Acknowledgments

The research of Rubén Fernández-Casal and Mario Francisco-Fernández has been supported by MICINN (Grant PID2020-113578RB-I00), and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2020-14 and Centro de Investigación del Sistema Universitario de Galicia ED431G 2019/01), all of them

through the ERDF. The research of Sergio Castillo Páez has been supported by the Universidad de las Fuerzas Armadas ESPE, from Ecuador.

References

[1] Castillo-Pez, S., Fernndez-Casal, R., and Garca-Soidn, P. (2020). Nonparametric bootstrap approach for unconditional risk mapping under heteroscedasticity. *Spatial Statistics*, 40, 100389.

[2] Chilès, J., and Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*. Wiley & Sons, New York.

[3] Francisco-Fernández, M. and Opsomer, J. D. (2005). Smoothing parameter selection methods for non-parametric regression with spatially correlated errors. *The Canadian Journal of Statistics* **33**, 279–295.

[4] Fernández-Casal, R., Castillo-Páez, S., y Francisco-Fernández, M. (2018). Nonparametric geostatistical risk mapping. *Stochastic Environmental Research and Risk Assessment* **32**, 675–684.

[5] Fernández-Casal, R., Castillo-Páez, S. and García-Soidán, P. (2017). Nonparametric estimation of the small-scale variability of heteroscedastic spatial processes. *Spatial Statistics* **22**, 358–370.

[6] Shapiro, A. and Botha, J.D. (1991). Variogram fitting with a general class of conditionally non-negative definite functions. *Computational Statistics and Data Analysis* **11**, 87–96.

Spatio-temporal variability of the distribution and abundance of sardine in the Portuguese mainland coast and relationship with environmental drivers

D. Silva[1,*] , R. Menezes[2] , A. Moreno[3] , A. Teles-Machado[3] and S. Garrido[3]

[1]*Minho University, Centre of Mathematics (CMAT), Braga, Portugal; danyelasylva2@gmail.com*
[2]*Minho University, Centre of Mathematics (CMAT), Guimarães, Portugal; rmenezes@math.uminho.pt*
[3]*Portuguese Institute for the Sea and Atmosphere (IPMA), Division of Modelling and Management of Fishery Resources, Lisboa, Portugal; amoreno@ipma.pt, ana.machado@ipma.pt, susana.garrido@ipma.pt*
[*]*Corresponding author*

Abstract. *Scientific tools capable of identifying the distribution patterns of species are important as they contribute to improve knowledge of causes of species fluctuations which can contribute to improve the species management, and consequently conserve biodiversity. Species distribution data often implies residual spatial autocorrelation and temporal variability, so both time and space are important components to study the evolution of species distribution from an ecological point of view. This study aims to estimate the spatio-temporal distribution of sardine (Sardina pilchardus) in the western and southern Iberian waters, relating the spatio-temporal variability of the biomass indicator with the environmental conditions. With this objective, a hierarchical two-part model is suggested capable of dealing with data specificities, namely zero-inflated, and with different sources of uncertainty. This work proposes to incorporate environmental covariates with time-lags, not under the usual approach of being fixed, but considering kernel weights.*

Keywords. *Environmental effects; Geostatistics; Hurdle model; Sardina pilchardus; Species Distribution Model.*

1. Introduction

Improving knowledge about biodiversity and species abundance has become an important scientific and societal issue. Scientific tools capable of identifying species distribution patterns are necessary for understanding the causes of these species fluctuations, and for taking decisions on measures that contribute to the conservation of biodiversity.

Species distribution data often implies residual spatial autocorrelation, that is, the observations are not conditionally independent. In this scope, spatial autocorrelation often arises due to the non-consideration of important environmental factors such as climate conditions that influences the species distribution or intrinsic factors such as competition, dispersal, and aggregation. Consequently, the application of spatial and non-spatial methods can lead to different conclusions. Temporal scale could also be an important component to consider in the modelling process, since species abundance varies in both time and space and there is also an ecological interest to study the evolution of species distribution [2].

The purpose of the presented study is to estimate the spatio-temporal distribution of sardine (*Sardina pilchardus*, Walbaum 1792) in western and southern Iberian waters, relating the spatio-temporal variability of the biomass indicator with the environmental conditions. Furthermore, this work will lead to the identification of the main drivers of sardine spatial dynamics and the understanding sardine dynamics over time and space.

Sardine is one of the most relevant pelagic species for Portugal and Spain due to its high socioeconomic importance by representing one of the main targets of the seine fishing. Several studies have been developed in order to understand the sardine distribution, its habitat, and its relationship with the environmental conditions, however the knowledge on sardine distribution in the western and southern Iberian waters is limited. Indeed, environmental and oceanographic conditions differ from marine regions and thus, study the spatio-temporal distribution of species in different regions can improve the knowledge on the species and how the species behaves under different conditions.

2. Material and Methods

2.1 Data

Acoustic data of sardine was obtained during Portuguese spring acoustic surveys (PELAGO) conducted by the Portuguese Institute for the Sea and Atmosphere (IPMA) in the western and southern Iberian waters from 2000 to 2020 (gap in 2012). Over this period, a total of 19920 hauls were carried out. Each haul is identified by a pair of coordinates (longitude and latitude); sector and zone. Our variable of interest is a biomass indicator (name of variable used hereafter), that was obtained from the acoustic energy. Daily environmental data was obtained for the region and time of study, particularly satellite derived sea surface temperature, chlorophyll-a concentration, bathymetry, and intensity and direction of surface ocean currents.

2.2 Spatio-temporal species distribution model

Species Distribution Models are investigated to relate sardine presence/absence and biomass with environmental conditions, aiming at predicting its distribution in unobserved locations and for the unobserved year of 2012. The estimation and prediction of a process and of the parameters that govern the process often define the main objective in the characterization of the phenomena, so a flexible structure is necessary to accommodate complex relationships between the data and process models, incorporating several sources of uncertainty.

In our case, the response variable, species biomass, is indexed in time and space, and in order to incorporate possible temporal, spatio-temporal and smoothing effects we use a two-part model. The two-part model is a hierarchical model where the species biomass distribution is given by the product of the species occurrence distribution and the species biomass distribution under occurrence. Given the hierarchical framework, the problem is decomposed into a series of levels linked by simple rules of probability. The Bayesian paradigm comes suitable for these complex spatial models since it might more easily handle with inference and prediction.

Consequently, numerical approximations are required to do inference, which can be computational challeging when applied to most realistic problems. In this sense, the integrated nested Laplace approximation (INLA) is used to approximate the posterior marginals of latent Gaussian Field (GF) [3]. However, INLA cannot be applied when dealing with continuous GFs, as the parametric covariance function needs to be specified and fitted based on the data. The Stochastic Partial Differential Equation (SPDE) manages to solve this problem since the GF with a Matérn covariance structure is replaced by a Gaussian Markov random field, which is a discretely indexed GF.

Let $Y_{\mathbf{s}t}$ be the spatio-temporal distributed biomass process at year t and location $\mathbf{s} \in D \subset \mathbb{R}^2$, where D represents the region of study, the Portuguese mainland coast. $Z_{\mathbf{s}t}$ denotes the occurrence sub-process, taking the value 0 if no species was observed in location $\mathbf{s}$ at year t, and 1 otherwise. Given the semi-continuous nature of the data, the occurrence process is assumed to come from a Bernoulli distribution, while the biomass process given the occurrence, $Y_{\mathbf{s}t}|(Z_{\mathbf{s}t} = 1)$, requires a semi-continuous distribution as Gamma or log-Normal distributions. In our case, we use the Gamma distribution. Therefore, the model is given by the following:

$$log(\mu_{\mathbf{s}ti}) = \alpha_1 + \sum_{j}^{p} f(K(X_{j\mathbf{s}ti}, c, l)) + \gamma_t + W_{\mathbf{s}t}$$

$$logit(\pi_{\mathbf{s}ti}) = \alpha_2 + \sum_{j}^{p'} f'(K'(X'_{j\mathbf{s}ti}, c, l)) + \gamma'_t + kW_{\mathbf{s}t} \quad (1)$$

The link function used to model the mean of the biomass indicator μ_{sti} under occurrence is the logarithm, while for the probability of occurrence π_{sti} is the logistic function, where i identifies the ith day of the survey in year t. $K(.)$ and $K'(.)$ represent weighted averages of environmental covariates X_{jsti} observed at day i of year t with daily time lags of $c-l, \cdots, c+l$, where the weights were determined by using the gaussian kernel function. The $f(.)$ and $f'(.)$ denote smoother functions such as thin plate and cubic regression splines. The α_1 and α_2 are regression coefficients. The γ_t and γ'_t refer to unstructured yearly effects specified by means of a Gaussian exchangeable prior with mean zero and precision τ_γ and $\tau_{\gamma'}$, respectively. The $W_{\mathbf{s}t}$ represents the spatio-temporal structure of the model. In our case, this latent process changes in time (year) with a first-order autoregressive structure and the spatial covariance is defined based on the Matérn function.

The use of the parameteres c (the mode of the Gaussian kernel) and l (the distance between the mode and the minimum of the Gaussian kernel) is motivated by the fact that the impact of the environmental conditions on sardine biomass and distribution can not be observed at the moment but in the nearest future. Futhermore, the use of various past moments allow a more complex, complete and realistic approach. Various combinations of c and l were tested for each covariate, except for bathymetry (which is considered a static covariate). The evaluation of the goodness of fit of the model by the Deviance Information Criterion (DIC) and the log-conditional predictive ordinates (LCPO) made possible to select the covariates.

3. Results

For the south coast, bathymetry and intensity are shown to be important to explain both sardine occurrence

and biomass, while temperature and chlorophyll-a also help to explain the biomass. Shallow locations favour both sardine occurrence (especially bathymetry between 22m and 46m) and biomass (bathymetry between 22m and 56m). Calmer locations (intensity between 0.08m/s and 0.11m/s) also favour sardine occurrence, while intensity presents different effects on biomass depending on the direction of ocean currents (north, south, or east). Strong currents towards the north influence negatively the biomass, but positively towards the south and east, being even more favourable when they move towards the east. The biomass is also higher for colder temperatures (between 14.4°C and 15.3°C with a maximum at 14.8°C) and where the chlorophyll-a varies between 10mg/m^3 and 25mg/m^3, being harmful for values of chlorophyll-a below 2mg/m^3.

The distribution maps allow to identify unfavourable and recurrent areas for the sardine and evaluate its evolution over two decades. Besides the great changes observed in these areas, in the last years persistent unfavourable areas can be identified.

4. Conclusions

This work provides a deep study of spatial distribution and abundance of sardine over 20 years. These results are relevant to assist in the elaboration of spatially explicit management plans for sardine. Indeed, although great changes were observed during the 20 years, some indicators of habitat selection and the persistence of unfavourable areas are pointed out and should be considered for the species conservation. In short, the study show that spatial modelling can play a key role in ecology, and especially in marine ecology.

Acknowledgments

The authors acknowledge to FCT Foundation (Fundação para a Ciência e Tecnologia) for funding this research through Individual Scholarship PhD. PD/BD/150535/2019, throught projects PTDC/MAT-STA/28243/2017, UIDB/ 00013/2020 and UIDP/00013/2020; to MAR2020 for funding through SARDINHA 2020 project (MAR-01.04.02-FEAMP-0009) and to all colleagues involved in this work. The data was collected under the European Commission's Data Collection Framework - PNAB/EU-DCF Programa Nacional de Amostragem Biológica, (Reg. EC 2008/199).

References

[1] Clark, J. S. and Gelfand, A. E., editors (2006). *Hierarchical modelling for the environmental sciences: statistical methods and applications*. Oxford University Press Inc.

[2] Martínez-Minaya, J., Cameletti, M., Conesa, D. V., and Pennino, M. G. (2018). Species distribution modeling: a statistical review with focus in spatio-temporal issues. *Stochastic Environmental Research and Risk Assessment* **32**, 3227–3244.

[3] 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* **71**, 319–392.

Black Scabbardfish species distribution: Geostatistical Inference under Preferential Sampling

P. Simões[1,2*] , M.L. Carvalho[3] , I. Figueiredo[4] , A. Monteiro[1] and I. Natário[1,5]

[1]*Centro de Matemática e Aplicações (CMA), Faculdade de Ciências e Tecnologia da Universidade NOVA de Lisboa, Portugal; pc.simoes@campus.fct.unl.pt*
[2]*Centro de Investigação, Desenvolvimento e Inovação da Academia Militar (CINAMIL), Portugal*
[3]*Centro de Estatística e Aplicações (CEAUL), Faculdade de Ciências da Universidade de Lisboa, Portugal, mlucilia.carvalho@gmail.com*
[4]*Instituto Português do Mar e da Atmosfera (IPMA); ifigueiredo@ipma.pt*
[5]*Departamento de Matemática, Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa, Portugal; icn@fct.unl.pt*
[*]*Corresponding author*

Abstract. *The spatial distribution and abundance of the black scabbardfish, a deep-water species, in Portugal, is mostly unknown. The available data relies on the commercial fisheries information. It is known that commercial fishing takes place where fishermen expect to find the higher catches of the species, leading to the choice of fishing locations that are not randomly but preferentially selected. The aim of this study is to do a species distribution modeling for the black scabbardfish (BSF) species in Portugal using geostatistical methods that addresses this question. The BSF captures are further analysed, under different scenarios, using a Bayesian approach and INLA methodology, considering stochastic partial differential equations (SPDE) for geostatistical data jointly with a with a Log-Cox point process model. The two best cases are presented, first considering the point process with covariate depth and response with spatial effect, and second one considering the same structure for the point process but incorporating in the response the covariate group of tonnage.*

Keywords. *Preferential sampling; Geostatistics; Point process; SPDE; INLA.*

1. Introduction

Addressing the problem of improving knowledge about the abundance of certain fish species is a determining factor for ensuring the sustainability of commercial fisheries and protecting the biodiversity of species that are of high interest for consumption. The available data on this area is based to on the choice of fishing locations that are not random but preferentially selected, which is referred to as preferential sampling. The sampling sites are deliberately chosen in areas where fishermen tend to look for a specific species in areas where they are believed to find it. As consequence, is a growing need to explore the problem of building mathematical/statistical models that take into account the problem of preferentiality, making it possible to produce maps of abundance that are more consistent and less biased. That will allow the responsible institutions to rely on concrete data to define more precise quotas, a scientific and societal important challenge to fulfill the mission of the fishing organizations. The black scabbardfish (BSF) species, on the portuguese coast, constitutes an important com-

mercial resource. Georefereced data about the location of the fishing hauls and the corresponding captures has been made available by the Portuguese Institute of Sea and Atmosphere (IPMA), for a number of differently sized vessels belonging to the fishing fleet. It is intended to use available information combined with environmental covariates, to predict where a species is likely to be present in unsampled locations for management and conservation purposes. A previous study was conducted considering a classical geostatistical approach trough different regression models with fixed, structured and unstructured random effects under a Bayesian approach [1]. Taking into account the available preferentially sampled data, standard geostatistical methods might have yielded biased results. The information of capture sites of the species under analysis should be accounted for in the modelling process of the [3, 7]. The aim of this study is to perform species distribution modelling to the BSF data using geostatistical methods that takes this preferentiability into account. The BSF captures are further analysed, under different scenarios, using a Bayesian approach and INLA methodology, considering stochastic partial differential equations (SPDE) for geostatistical data jointly with a with a Log-Cox point process model. The two best cases are presented, first considering the point process with covariate depth and response with spatial effect, and second one considering the same structure for the point process but incorporating in the response the covariate group of tonnage.

2. Data

The data considered in this study is a comprehensive data set of geo-referenced captures of black scabbardfish from commercial fisheries, along the Portuguese coast, between the 2002 and 2013. Several other variables have also been registered along with the captures as, for example, the vessel tonnage and identification, the speed and also the depth at which the capture has been made. A subset of the original data was taken for this data analysis: the fishing area with latitude minor than 39.3, captures that have occurred from September to February for the years between 2009 to 2013, resulting in a total set of 732 observations. The locations of the data are displayed in figure 1. Due to a skewed original data, a Box-Cox transformation of BSF data (Y) was carried out, according with the expression $Y^* = \frac{Y^\lambda - 1}{\lambda}$, with $\lambda = \frac{1}{2}$, so that the response follows approximately a Normal distribution.

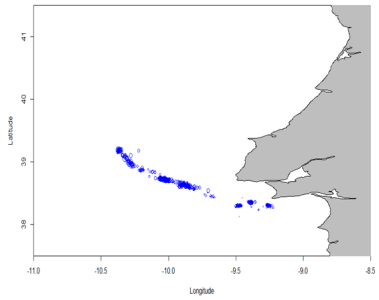


Figure 1: BSF data locations.

3. Geostatistical Inference under preferential sampling

The preferential sampling model is considered as a two part model that share information, the observed species captures and the intensity of the point process, reflecting the sampling intensity through space [6, 7]. The observed locations $(x_1,...,x_n)$, are assumed to come from a non-homogeneous Poisson process, whereby species distribution are described based on a trend function that may depend on covariates with corresponding intensity (the number of points per unit area). Log-Gaussian Cox Process (LGCP) are a specific class of Cox processes in which the logarithm of the intensity surface, λ_i, is a Gaussian random field. The observed captures $Y = (Y_1,...,Y_n)$ is linked to the intensity of the underlying spatial field $S(x)$. $S(x)$ is a stationary Gaussian Process with mean zero, variance σ^2, and a Matérn correlation (correlation shape parameter fixed $k = 1.5$, correlation range ϕ and nugget variance τ^2), $S(x) \sim N(0,\Sigma)$. Using an approximate stochastic weak solution of a Stochastic Partial Differential Equation that is a continuous Gaussian field with a Matérn covariance structure, inference in made considering stochastic partial differential equations (SPDE) models for geostatistical data and INLA (Integrated Nested Laplace Approximation) methodology, mainly through the package R-INLA [5, 8]. The first step is the triangulation of the considered spatial domain by building a mesh that covers the study region, the constrained refined Delaunay triagulation. The SPDE model was defined considering penalized complexity priors (PC Priors) for the model parameters, range r and marginal standard deviation σ, of the spatial effect, ($P[r < 30] = 0.2$, $P[\sigma > 10] = 0.01$) [4]. The two best models (model comparison was performed using the DIC) were:

Model 1: Point process with covariate depth (D) and response with spatial effect,

$$\begin{aligned} &Y_i|S \sim Normal(\mu_i,\tau^2), i = 1,..,n, \\ &Y_i|(S,\boldsymbol{X} = \boldsymbol{x})] = \beta_0^y + \beta^y S(x_i) + e_i, \\ &\boldsymbol{e_i} \sim N(0,\tau^2), \\ &\lambda_i = \exp(\beta_0^{pp} + \beta_1^{pp} D_i + S(x_i)) \end{aligned} \tag{1}$$

Model 2: Point process with covariate depth (D) and response with covariate group of tonnage (PRT) and spatial effect,

$$\begin{aligned} &Y_i|S \sim Normal(\mu_i,\tau^2), i = 1,..,n, \\ &Y_i|(S,\boldsymbol{X} = \boldsymbol{x})] = \beta_0^y + \beta_1^y PRT_i + \beta^y S(x_i) + e_i, \\ &\boldsymbol{e_i} \sim N(0,\tau^2), \\ &\lambda_i = \exp(\beta_0^{pp} + \beta_1^{pp} D_i + S(x_i)) \end{aligned} \tag{2}$$

where β_0^{pp} is the correspondent intercept, $S(x_i)$ spatial effect of the model, where the observed locations are modelled by a LGCP, $X|S$, $i = 1,...,n$ (index of i-location). For $\beta^y > 0$ the response values are higher where

there are more observation. For model 1 the estimated value of β^y was 1.24, and for model 2 was 0.24. Figure 2 shows the posterior predicted mean of the spatial effect for model 1 and 2, respectively.

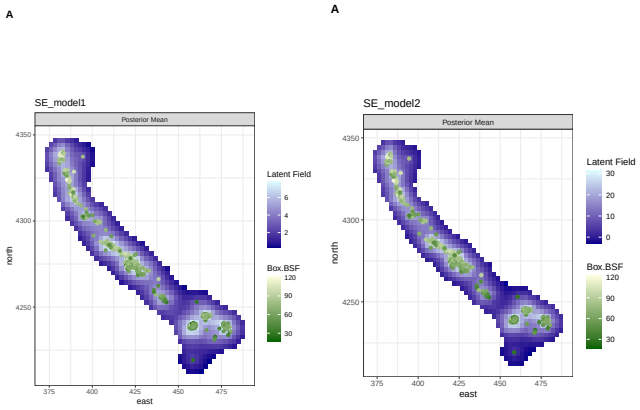


Figure 2: Posterior predicted mean of the spatial effect, model 1 (left) and model 2 (right).

4. Conclusion

The spatial outputs obtained with the preferential model better absorb the variability of BSF captures providing a more realistic pattern of BSF distribution. This approach allows a better knowledge of BSF spatial distribution, assuming that the selection of the sampling locations depends on the values of the observed species. The modelling could be extended in order to include important environment factors, that may be important in the estimation or by incorporating a term for the temporal effect moving on to a spatio-temporal approach [2, 6].

Acknowledgments

This work is funded by national funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects PREFERENTIAL, PTDC/MAT-STA/28243/2017, UIDP/00297/2020 (Center for Mathematics and Applications) and UIDB/00006/2020(CEAUL).

References

[1] André, L. M., Figueiredo, I., Carvalho, M. L., Simões, P., Natário, I. (2020). Spatial Modelling of Black Scabbardfish Fishery Off the Portuguese Coast. *In International Conference on Computational Science and Its Applications* 332–344. Springer.

[2] Blangiardo, M., Cameletti, M., Baio, G., Rue, H. (2013).Spatial and spatio-temporal models with R-INLA. *Spatial and spatio-temporal epidemiology* **4**, 33–49.

[3] Diggle, P., Menezes, R., Su, T. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **59(2)**, 191-232.

[4] Fuglstad, G. A., Simpson, D., Lindgren, F., Rue, H.. (2019). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association* **114(525)**, 445–452.

[5] Lindgren, F., Lindström, J., Rue, H. (2010). *An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach.*. Mathematical Statistics, Centre for Mathematical Sciences, Faculty of Engineering, Lund University.

[6] Martínez-Minaya, J. et al (2018). Species distribution modeling: a statistical review with focus in spatio-temporal issues. *Stochastic environmental research and risk assessment* **32(11)**, 3227–3244.

[7] Pennino, M. et al (2019). Accounting for preferential sampling in species distribution models. *Ecology and evolution* **9(1)**, 653–663.

[8] Rue, H., Martino, S., 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.

Impact of climate and local environment on Dengue and Zika dynamics in Brazil: A joint Bayesian spatio-temporal model

M.H. Suen[1] , F. Lindgren[1] , M. Blangiardo[2] , F. Chiaravalloti-Neto[3] , and M. Pirani[2]

[1]*School of Mathematics, University of Edinburgh; m.h.suen@ed.ac.uk, finn.lindgren@ed.ac.uk*
[2]*MRC Centre for Environment and Health, Department of Epidemiology and Biostatistics, Imperial College London, UK; m.blangiardo@imperial.ac.uk , monica.pirani@imperial.ac.uk*
[3]*Department of Epidemiology, School of Public Health, University of São Paulo; franciscochiara@usp.br*

Abstract. *Arboviral diseases pose a major challenge for public health in Brazil. This project focuses on Dengue and Zika virus, which share the same virus family and vector. Thanks to the multiple data sources, this study aims to deliver a spatio-temporal Bayesian hierarchical model for examining the interactions among climate, diseases distribution and environmental conditions. Hence, the objectives are to (i) to jointly examine the spatial distribution and temporal variability of the arboviral diseases in relationship to climatological factors, in presence of changes in the local environmental conditions, and (ii) to predict the probability of outbreak of the diseases with quantified uncertainty. The data are fitted with fixed or random effects M model. Intrinsic conditional autoregressive prior (iCAR), Leroux prior (LCAR) and proper conditional autoregressive prior (pCAR) are considered to account for the spatial dependence. Random walk of first order (RW1) is applied for the temporal dependence. Here we present preliminary results obtained from monthly data in the state of Bahia in Brazil during Jan 2015 - Jun 2019. The posterior mean of the spatial risk shows different clustering for the diseases. This study provides insights into the discussion of the interactions between global climate changes and arboviral diseases epidemics.*

Keywords. *Arboviral diseases; Bayesian hierarchical model;* `inlabru` *package; Spatio-temporal; Remote sensing*

---

1. Background and Aim

Arboviral diseases and their complications constitute a major threat for public health in Brazil, as the diseases are spreading widely and expanding into geographic regions outside transmission zones. We consider two mosquito-borne diseases: Dengue and Zika virus; the major vector and virus family of which are Aedes aegypti and Flaviviridae respectively. The majority of the current understanding of Dengue and Zika virus spreading in Brazil is obtained by considering them separately within a time frame, rather than modelling the multi-dimension of the spatial and temporal patterns of the two diseases jointly. Here, we will focus on a joint spatio-temporal modelling approach for studying the interrelations between climate, disease distribution and socio-environmental conditions and the synchrony that these systems express.

2. Methods

By taking advantage of multiple data sources, including satellite-derived information on environmental conditions and ecosystem, re-analysis climate data and census-based socioeconomic data, we develop a hierarchical spatio-temporal modelling approach for Dengue and Zika cases obtained from the Brazilian Ministry of Health at municipality level from 2015 to mid-2019. The model, grounded within a Bayesian framework, allows us: (i) to jointly examine the spatial distribution and temporal variability of the arboviral diseases in relationship to climatological factors, in presence of changes in the local environmental conditions, and (ii) to predict the probability of outbreak of the diseases with quantified uncertainty. The Bayesian inference is performed using the integrated nested Laplace approximation (INLA) [5] via the `inlabru` extension package in `R` [1], considering a recent computation-efficient multivariate proposal for areal data, called M models [3, 6].

2.1 Model Framework

We present a Poisson hurdle model to fit the data. This consists first to adopt a Bernoulli distribution if there is an occurrence of disease and then to specify a Poisson distribution for the number of disease cases. For i-th municipality ($i = 1,...,I$), t-th time point ($t = 1,...,T$) and j-th disease ($j = 1,2$), the observed Bernoulli and Poisson responses (Y^{bin}_{itj} and Y^{poi}_{itj}) are modelled as follows using the `inlabru` package:

$$Y^{\text{bin}}_{itj} = \begin{cases} 0, & \text{if there is no disease case,} \\ 1, & \text{otherwise,} \end{cases} \qquad Y^{\text{poi}}_{itj} = \begin{cases} \texttt{NA}, & \text{if } Y^{\text{bin}}_{itj} = 0, \\ N_{itj}, & \text{otherwise,} \end{cases}$$

$$Y^{\text{bin}}_{itj} \sim \text{Bin}(n_{itj} = 1, p_{itj}) \qquad (Y^{\text{poi}}_{itj} | Y^{\text{bin}}_{itj} = 1) \sim \text{Poi}_{>0}(\lambda_{itj})$$

where N_{itj} is the number of disease cases; the probability p_{itj} and the mean λ_{itj} are linked to the linear predictor by $p(\eta^{\text{bin}}_{itj}) = \frac{\exp(\eta^{\text{bin}}_{itj})}{1+\exp(\eta^{\text{bin}}_{itj})}$, and $\lambda_{itj} = E_{it}\exp(\eta^{\text{poi}}_{itj})$ respectively, where E_{it} is equal to the monthly expected disease cases; hence, $\exp(\eta^{\text{poi}}_{itj})$ refers to relative risk. Since the linear predictors, η^{bin}_{itj} and η^{poi}_{itj}, share the same formulation, we henceforth drop the superscript. Thus the link predictors, η_{itj}, are defined as

$$\eta_{itj} = \alpha_j + f_1(X_{itl}) + \Sigma_q \beta_q X_{itq} + \theta_{k_i j} + \gamma_{t j} + \delta_{k_i t j}, \tag{1}$$

where α_j is the intercept for the j-th disease; $f_1(X_{itl})$ is a nonlinear smoothed function with first-order random walk prior (RW1) [2] for mean air temperature at 2m above the Earth's surface, $\Sigma_q \beta_q X_{itq}$ are the q-th covariates' linear (fixed) effects, namely the Normalized Difference Vegetation Index (NDVI), total precipitations, dew point temperature, and socio-economic variables, and $\theta_{k_i j}$, γ_{tj} and $\delta_{k_i tj}$ are random effects capturing spatial pattern, temporal trends and spatio-temporal interaction respectively for the k_i-th microregion, which groups several neighbouring municipalities. Relaxing to a lower spatial resolution is to compromise the computational cost.

To account for the correlation between spatial patterns and temporal trends of both diseases, we adopt the recently proposed multivariate M model [6]. We first denote $\Theta = \{\theta_{k_i j} : i = 1,...,I; j = 1,2\}$, $\Gamma = \{\gamma_{tj} : t =$

$1,...,T; j = 1,2\}$, and $\Delta_j = \{\delta_{k_itj} : i = 1,...,I; t = 1,...,T; j = 1,2\}$. Hence, we have $\Theta = \Phi_\theta M_\theta$ for spatial random effects, $\Gamma = \Phi_\gamma M_\gamma$ for temporal random effects, and $vec(\Delta_j) \sim N(0, \sigma^2_{\delta_j} Q^-_\delta)$ for spatio-temporal interactions, where M_θ and M_γ matrices' columns are the random effects accounting for spatial and temporal dependencies respectively. Each Δ_j captures the spatio-temporal interaction within the j-th disease, and Q^-_δ is defined depending on the type of space-time interaction [4].

	Bernoulli					Poisson				
Covariate	Mean	SD	2.5%	50%	97.5%	Mean	SD	2.5%	50%	97.5%
tp	-0.020	0.050	-0.117	-0.020	0.078	-0.400	0.011	-0.422	-0.400	-0.378
dew	-0.108	0.083	-0.272	-0.108	0.054	0.035	0.020	-0.005	0.035	0.074
NDVI	-0.114	0.025	-0.163	-0.114	-0.064	-0.603	0.007	-0.616	-0.603	-0.590
T_AGUA	0.059	0.025	0.011	0.059	0.108	-0.170	0.007	-0.184	-0.170	-0.156
IDHM	-1.498	0.042	-1.580	-1.498	-1.416	0.683	0.008	0.668	0.683	0.697

Table 1: Posterior estimates and the 95% Credible Intervals of the standardised covariates related to climate and socio-environmental variables in the state of Bahia. tp, dew, NDVI, T_AGUA and IDHM refer to total precipitation, dew point temperature, the normalized difference vegetation index, the population living in households with running water and Human Development Index respectively.

We have considered intrinsic conditional autoregressive prior (iCAR), Leroux prior (LCAR) and proper conditional autoregressive prior (pCAR). Here, we evaluate spatial models with the LCAR built on the scale of the Brazilian micro-regions. The covariance matrix of the separable spatial or temporal structure can be estimated via $M'_\theta M_\theta$ with a Wishart prior, i.e. $M'_\theta M_\theta \sim \text{Wishart}(J, \sigma^2_\theta I_J)$ and $M'_\gamma M_\gamma \sim \text{Wishart}(J, \sigma^2_\gamma I_J)$. The fixed effects M model (FE) assumes $\text{N}(0, \sigma^2)$ prior with a large σ for the elements in M; while a random effects M model (RE) makes inference on σ [6].

3. Results

As a pilot study, we have looked into the epidemiological monthly data in the state of Bahia during Jan 2015 - Jun 2019 with the FE models without interaction terms. The spatial structure of Dengue and Zika virus cases are consistent across the occurrence probabilities and relative risks with slight deviance in certain microregions, see Figure 1. This ascertains the close biological relationships between Dengue and Zika virus. The fixed effect parameters in both likelihoods shown in Table 1 are as expected. For example, the higher NDVI, the lower the diseases occurrence probability and relative risks since mosquito-borne diseases is more commonly found in urban areas, typically with lower NDVI. The extension of the model to the rest of Brazil is currently ongoing,

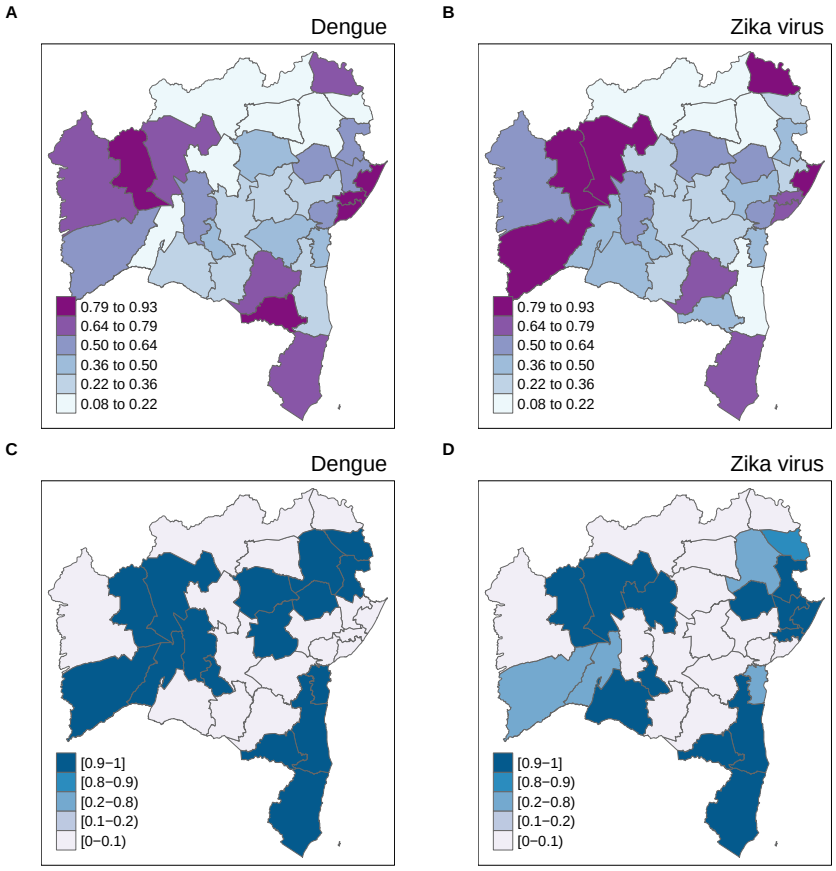


Figure 1: Posterior mean of the microregion-specific spatial pattern in the state of Bahia, the occurrence probability of disease, $p(\eta^{\text{bin}})$ (A and B), and the exceedance probabilities, $\mathbb{P}(\exp(\theta^{\text{poi}}_{k_i j}) > 1|\mathbf{Y})$ (C and D), for Dengue and Zika virus FE model with LCAR spatial and RW1 temporal priors.

as well as the evaluation of the predictive capability of the model.

4. Conclusions

The pilot study shows the spatio-temporal link between Dengue and Zika virus in the state of Bahia regarding the climatological factors via a Poisson hurdle model. The future extension to the entire Brazil will shed light on the ongoing debate about the interaction between global changes and arboviral disease epidemics, and offer methodological tools which can be useful in supporting early warning systems.

Acknowledgments

I wish to thank Prof Finn Lindgren at University of Edinburgh, Dr Monica Pirani and Prof Marta Blangiardo, at Imperial College London, and collaborator, Prof Francisco Chiaravalloti-Neto from the University of São Paulo. The project is supported by the Wellcome Trust Seed Award on Science "Spatio-temporal dynamics of arboviral diseases in Brazil in a changing climate.

References

[1] Bachl, F. E., Lindgren, F., Borchers, D. L., and Illian, J. B. (2019). Inlabru: an r package for bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, **10**(6):760766.

[2] Blangiardo, M. and Cameletti, M. (2015). Spatial and spatio-temporal Bayesian models with R-INLA. John Wiley & Sons.

[3] Botella-Rocamora, P., Miguel A., Martinez-Beneito, M. A. and Banerjee, S. (2015). A unifying modeling frame- work for highly multivariate disease mapping. *Statistics in medicine*, **34**(9):15481559.

[4] Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in medicine*, **19**(17-18):25552567.

[5] Rue, H., Martino, S. Chopin, N. (2009). Approximate bayesian inference for latent gaussian models using integrated nested laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, **71**:319 392.

[6] Vicente, G., Goicoa, T. and Ugarte, M.D. (2020). Bayesian inference in multivariate spatio-temporal areal models using inla: analysis of gender-based violence in small areas. *Stochastic Environmental Research and Risk Assessment*, **34**:14211440.

Spatial multi-resolution models for small forestry data sets

I. Marques[1] , P.F.V. Wiemann[2,3,*] and T. Kneib[2]

[1]*Chair of Spatial Data Science, Gttingen University, Germany; imarques@uni-goettingen.de*
[2]*Chair of Statistics, Gttingen University, Germany; pwiemann@uni-goettingen.de, tkneib@uni-goettingen.de*
[3]*Chair of Computational Statistics, Department of Statistics, TU Dortmund University, Germany*
[*]*Corresponding author*

Abstract. *In forestry science, researchers often have to deal with spatial datasets. In particular, experiments are often carried out in small plots that are sparsely distributed within a larger area. Within each plot the experiment is exactly replicated. Naturally, spatial dependencies within each plot, but also between plots, have to be considered. However, the distances within a plot and between plots are in different ballparks rendering a classical spatial model unsuitable. In this paper, we introduce multiple approaches to model this kind of data following the three guiding principles of comprehensibility, usefulness for small samples, and reduced computational complexity. We discuss an effective MCMC-based estimation procedure and present an application to biomass data from Northern Germany.*

Keywords. *Gaussian random field; MCMC; multi-resolution; replicates; spatial regression*

1. Introduction

Geophysical processes often yield datasets that are spatially irregular, showing a multi-resolution character over space where data are collected at different intensities in different parts of the domain. In such cases, spatial models that assume the same dependence structure over the whole space are not able to capture the true complexity of the dependence structure. The issue of how to model irregular spaced data has typically been handled by adding multiple spatial effects to a regression model, potentially subsequently, and with increasingly finer resolution [4, 3] - the so called multi-resolution models. Whenever ones talks about spatial models, computational feasibility is a crucial point. Multi-resolution models have been successful in spatial statistics due to their ability to flexibly capture dependence at multiple spatial scales while being computationally feasible [2].

In this paper, we consider a design very common to forestry experiments. In these experiments, for a given spatial domain, several identically sized plots are considered. These plots are sparsely distributed over a larger domain. Within each plot, data collection is more intensive. Due to the irregular intensity of the data collection, a single-resolution model will most likely miss important features of the data. This paper aims at developing a spatial multi-resolution model that reflects this irregularity.

We are guided by a set principles that the final model must satisfy: (1) easily understandable, (2) ability to estimate within plot behavior for relatively "small" respective sample sizes, (3) reduced computational complexity. To guarantee Principle (2) is satisfied, we use a mother-children cell resolution with two resolutions

and assume all children (i.e., plots) to have the same spatial model. Thus, children cells are treated as replicates of each other. This concept can be easily extended to spatio-temporal models. Thus far, we limit the proposed models to identical sampling designs for all plots. However, limitation can easily be lifted by using a basis function approach. Finally, the principle of reduced computational complexity indicates a computational cost of factorizations smaller than $O(n^3)$ for a sample size of n.

2. The models

Consider a spatial domain $\mathcal{S} \subset \mathbb{R}^2$. Within $\mathcal{S}$, data is available at m equally sized areas $\mathcal{S}_i$, $i = 1, \ldots, m$. The areas $\mathcal{S}_i$ do not overlap and do not need to cover $\mathcal{S}$ entirely. We assume each area has the same number of observations y_{ij}, $j = 1, \ldots, n$ at the same locations within a plot. More precisely, let $\boldsymbol{s}_{ij}$ be location j in area i. We consider the model equation

$$y_{ij} = \boldsymbol{x}(\boldsymbol{s}_{ij})'\boldsymbol{\beta} + \gamma(\boldsymbol{s}_{ij}) + \varepsilon(\boldsymbol{s}_{ij}) \tag{1}$$

where $\boldsymbol{x}(\cdot)$ is a p-sized covariate vector and $\boldsymbol{\beta}$ is the associated coefficient vector. The latent variable $\gamma(\cdot)$ is a Gaussian random field (GRF). Moreover, $\varepsilon(\cdot) \sim N(0, \sigma^2)$ is an *i.i.d.* non-spatial error or nugget stochastically independent from $\gamma(\cdot)$.

The GRF is a spatial process $\{\gamma(\mathbf{s}) : \mathbf{s} \in \mathcal{S}\}$ for which all finite-dimensional distributions of the process are Gaussian. A natural candidate for $\boldsymbol{\gamma}$ is one that assumes a single spatial resolution, i.e. a global scale, such that $\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a global covariance matrix. For example, $\boldsymbol{\Sigma}$ can arise from the exponential covariance function

$$C(\boldsymbol{s}, \boldsymbol{s}') = \tau^2 \exp\left(-\kappa ||\boldsymbol{s} - \boldsymbol{s}'||\right), \tag{2}$$

where τ^2 represents the marginal variance of the spatial field and κ is related with the spatial range ρ. In this paper, we assume that all spatial covariance matrices can be linked to an exponential covariance function.

However, besides not being able to capture different levels of spatial variation in a global covariance function, this model would not respect Principle (3) of reduced computational complexity, since the computational complexity of factorizations is $O(k^3)$ where k is the number of rows in $\boldsymbol{\Sigma}$. In what follows, we present three models fit our data problem and satisfy the three guiding principles presented. The models differ in the way the GRF $\boldsymbol{\gamma}$ is specified. We elaborate how the models presented satisfy the third principle in Section 3. For the sake of simplicity, herein we write $\gamma(\boldsymbol{s}_{ij})$ as γ_{ij}.

M1: interacting resolutions model. In M1, the latent variable $\boldsymbol{\gamma}$ follows $\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^b \otimes \boldsymbol{\Sigma}^w)$ where $\boldsymbol{\Sigma}^b$ and $\boldsymbol{\Sigma}^w$ are covariance matrices representing the between plots and within plots correlations, and consequently are of size m and n, respectively. Both are linked to differently parametrized exponential correlations functions. Hence, M1 models two spatial resolutions that interact. The spatial dependence structure within each area $\mathcal{S}_i$ is not independent of all the other areas, but it depends on the spatial structure of the neighboring areas as well.

M2: independent resolutions model. In M2, we consider the superposition of two stochastically independent spatial processes $\boldsymbol{\gamma}^b$ and $\boldsymbol{\gamma}^w$. The prior distribution of $\boldsymbol{\gamma}^b = (\gamma^b_{11}, \ldots, \gamma^b_{1n}, \gamma^b_{21}, \ldots, \gamma^b_{mn})'$ is given by $\boldsymbol{\gamma}^b \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^b \otimes \boldsymbol{I}_n)$ and the prior distribution of $\boldsymbol{\gamma}^w = (\gamma^w_{11}, \ldots, \gamma^w_{1n}, \gamma^w_{21}, \ldots, \gamma^w_{mn})'$ is given by $\boldsymbol{\gamma}^w \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^w)$.

The spatial effect is composed of two latent variables, i.e., $\gamma_{ij} = \gamma^b_{ij} + \gamma^w_{ij}$.

M3: spatially correlated random intercept model. In M3, as in M2, the latent variable γ results from the superposition of two GRFs. However, M3 is simpler as each plot has a common spatially correlated random intercept, i.e., $\gamma_{ij} = \gamma^b_i + \gamma^w_{ij}$. Therefore, γ^b_i acts as a random intercept for area $\mathcal{S}_i$. Prior $\boldsymbol{\gamma}^b = (\gamma^b_1, \ldots, \gamma^b_m)'$ is normally distributed, e.g., $\boldsymbol{\gamma}^b \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_b)$ and the prior on $\boldsymbol{\gamma}^w = (\gamma^w_{11}, \ldots, \gamma^w_{1n}, \gamma^w_{21}, \ldots, \gamma^w_{mn})'$ is given by $\boldsymbol{\gamma}^w \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}_w)$. The GRFs $\boldsymbol{\gamma}^b$ and $\boldsymbol{\gamma}^w$ are stochastically independent.

The difference between the three models is most obvious when considering the covariance between two locations γ_{ij}, γ_{kl}. Precisely, for the three models, $\mathrm{Cov}(\gamma_{ij}, \gamma_{kl})$ is given by

$$\underbrace{(\Sigma^b)_{i,k}\,(\Sigma^w)_{j,l}}_{M1} \qquad \underbrace{\begin{cases} (\Sigma^b)_{i,k} + (\Sigma^w)_{j,l} & i = k, j = l, \\ (\Sigma^b)_{i,k} & i \neq k, j = l, \\ (\Sigma^w)_{j,l} & i = k, j \neq l, \\ 0 & i \neq k, j \neq l \end{cases}}_{M2} \qquad \underbrace{\begin{cases} (\Sigma^b)_{i,k} + (\Sigma^w)_{j,l} & i = k \\ (\Sigma^b)_{i,k} & i \neq k \end{cases}}_{M3}.$$

Multi-resolution spatial models typically assume that spatial models for different resolutions are independent. This is similar to M2 and M3, where we add the two spatial effects γ^b, for the m areas, and γ^w, for the n locations within each area. In M3, locations of different plots are solely correlated via the correlated random intercept whereas in M2 there is additional correlation if they are at the same location within a plot. In contrast, we see that the two spatial resolutions interact in M1.

3. Estimation

We base the estimation of the model parameters on Markov chain Monte Carlo (MCMC) using the software Liesel [5]. We consider a likelihood with marginalized γ and ε because of its typically better MCMC mixing properties [1]. However, this is computationally not feasible for M3 as we explain below. We use an Hamiltonian Monte Carlo (HMC) based update step for all parameters with variance and range parameters sampled in the log space. In our application, the number of areas m and the number of observations within one area n is relatively small.

Efficiency considerations for M1 and M2. In M1 and M2, we can exploit a method introduced by [6] to decrease the computational costs for evaluating the marginal likelihood from $O(n^3m^3)$ to $O(n^3 + m^3)$. Consider M1, where the likelihood with marginalized γ and ε is given by $\boldsymbol{y}|\boldsymbol{\beta}, \kappa, \sigma^2 \sim \mathcal{N}\left(\boldsymbol{X}'\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_{mn} + \Sigma^b \otimes \Sigma^w\right)$. The evaluation of its probability density function (pdf) requires the calculation of the determinant and inverse of $\boldsymbol{\Sigma}^b \otimes \boldsymbol{\Sigma}^w + \sigma^2\boldsymbol{I}_{mn}$ which is a $mn \times mn$ matrix, thus it has computational complexity of $O(n^3m^3)$. These tasks can be accomplished more efficiently by further exploiting the properties of the Kronecker product.

Instead of considering the marginal pdf from above, we consider the marginal pdf of the rotated data $\mathrm{vec}(\boldsymbol{U}_w^T\boldsymbol{Y}\boldsymbol{U}_b)$ where $\boldsymbol{U}_b$ originates from the eigenvalue decomposition $\boldsymbol{\Sigma}^b = \boldsymbol{U}_b\boldsymbol{S}_b\boldsymbol{U}_b^T$, similar $\boldsymbol{U}_w$, and $\boldsymbol{Y}$

is $n \times m$ matrix where each column corresponds to the data from one plot and each row corresponds to the same location in one plot. Consequently, the operator $\text{vec}(\boldsymbol{Y}) = \boldsymbol{y}$ defines the inverse operation, namely concatenating the columns. The rotated data can be interpreted having the covariance matrix $\boldsymbol{S}_b \otimes \boldsymbol{S}_w + \sigma^2 \boldsymbol{I}_{mn}$ for which the inverse is easily computable (for details refer to [6]). Similarly, the computational complexity can be reduced for M2.

Efficiency considerations for M3. The covariance matrix in the marginal likelihood of $\boldsymbol{y}$ in M3 does not allow to take advantage of the results found in [6]. Therefore, we use an alternating updating scheme considering $\boldsymbol{\gamma}^b$ as model parameters updated with a Gibbs step while the remaining parameters are still updated with an HMC step. Consequently, all matrix decompositions can be done in $O(n^3)$ or $O(m^3)$.

4. Empirical examples

In a proof of concept, we considered data generating processes following the exact model specifications. We can show that we reliably recover the model parameters. However, as discussed above variance parameter and range parameter of the GRF are not identifiable, so we consider their ratio, instead.

For our application, we study live biomass data collected in equally sized plots in Northern Germany. Within these plots the relative locations are identical. We compare the performance of our models and show their ability to capture spatial variation on different resolutions.

Acknowledgments

The authors acknowledge finical support from the German Research Foundation (DFG) through Grant 443179956. Further financial support from DFG via RTG 2300 is thankfully acknowledged by Isa Marques.

References

[1] Finley, A.O., Banerjee, S. and Gelfand, A.E. spbayes for large univariate and multivariate point-referenced spatio-temporal data models. *Journal of Statistical Software*, 63(13):128, 2015.

[2] Johannesson, G., Cressie, N. and Huang, H.C. Dynamic multi-resolution spatial models. *Environmental and Ecological Statistics*, 14(1):5–25, 2007.

[3] Katzfuss, M.. A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, 112(517):201–214, 2017.

[4] Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F. and Sain, S.. A multiresolution gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics*, 24(2):579–599, 2015.

[5] Riebl, H. and Wiemann, P.F.V. Liesel, 2022. https://github.com/liesel-devs.

[6] Stegle, O., Lippert, C., Mooij, J.M., Lawrence, N. and Borgwardt, K. Efficient inference in matrix-variate gaussian models with iid observation noise. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

An ensemble-based approach for the analysis of spatially misaligned data

R. Zhong[1,*] and P. Moraga[1]

[1]*Computer, Electrical and Mathematical Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia; ruiman.zhong@kaust.edu.sa, paula.moraga@kaust.edu.sa*
[*]*Corresponding author*

Abstract. *The spatial data misalignment problem occurs when data at different spatial scales need to be combined. In this work, we propose an ensemble-based approach for the analysis of spatially misaligned data that combines multiple statistical approaches including melding and downscaler models, and propagates uncertainty from individual models for better uncertainty quantification. A simulation study is conducted to assess the predictive performance of the models proposed. The ensemble-based approach is also used to predict fine particulate matter emissions (PM2.5) in the UK using data obtained from monitoring stations and satellite-derived environmental indicators. Results show that the proposed ensemble-based approach to combine multiple spatially misaligned data provides better predictions than individual models and can help decision-making in a wide range of disciplines.*

Keywords. *Spatial modeling; Spatial misalignment; Gaussian random process; Air Pollution.*

1. Introduction

In recent years, spatial data from satellite imagery, monitoring stations, and surveys have been collected in large quantities and at high spatial resolutions. The analysis of these data is crucial for decision-making in many disciplines such as the environment, climate and epidemiology. An ongoing challenge when analyzing these type of data is the spatial data misalignment problem which occurs when data at different spatial scales need to be combined. In this work, we propose a new ensemble-based approach for the analysis of spatially misaligned data that combines multiple statistical approaches including melding and downscaler models. This approach allows us to combine data at different resolutions to predict the variable of interest at points, areas and spatially continuous surfaces. Our approach improves prediction and propagates uncertainty from individual models for better uncertainty quantification. The predictive performance of the individual models and the ensemble-based approach are assessed by conducting a simulation study. Specifically, we generate spatial processes that can appear in real settings, fit the individual and ensemble approaches using measurements at several configurations of generated spatial data, and assess the performance of the approaches using spatial cross-validation designs. The new approach is also used to predict fine particulate matter emissions (PM2.5) in the UK using data obtained from monitoring stations and satellite-derived environmental indicators. Our results show that the new proposed ensemble-based approach result in a better final prediction. Moreover, the ensemble approach also shows robustness and generalization and avoids extremely deviant predictions. We believe our approach can

enhance the reliability of predictions of outcomes obtained by combining multiple spatially misaligned data and can help decision-making in a wide range of disciplines.

2. Methodology

The ensemble-based approach for the analysis of spatially misaligned data combines the outputs of two individual approaches, namely melding and downscaler approaches. The combination is done by the meta-learning method. The meta-learning uses algorithms to learn a second-level model (meta-learner) from the first-level models'(base-learner) outputs and generate the final predictions [4]. In our work, a stacked regression with spatial-varying coefficients is trained as meta-learners. The idea is to evaluate the performances of the base-learners on the spatial cross-validation designs to determine the coefficients in the combination. The model is as follows:

$$Y(x_i) = \tilde{\beta}_0(x_i) + \tilde{\beta}_1(x_i)Y_1(x_i) + \tilde{\beta}_2(x_i)Y_2(x_i) + \varepsilon(x_i),$$

where $Y_1(x)$ and $Y_2(x)$ represent the outputs of the melding and downscaler approaches, respectively, and $\varepsilon(x) \sim N(0,\sigma^2)$. The ensemble-based approach propagates uncertainty from individual models for better uncertainty quantification.

The base learner, melding and downscaler approaches are specified as follows. The melding approach assumes that underlying all point and areal level data, there is a spatially continuous variable that can be modeled using a Gaussian random field process $S = \{S(x) : x \in D \subset \mathbb{R}^2\}$ with $E[S(x)] = 0$ and stationary covariance function $Cov(S(x), S(x')) = \Sigma(x - x')$. Conditionally on S, $Y(x)|S(x) \sim N(\mu(x) + S(x), \tau^2)$, where $\mu(x)$ represents the large scale structure [1]. Then, point data observed at a finite set of sites $x_i \in D,\ i = 1,2,\ldots,I$ can be expressed as

$$E[Y(x_i)] = \mu(x_i) + S(x_i).$$

Areal data arise as block averages in blocks $B_j \in D,\ j = 1,2,\ldots,J$,

$$E[Y(B_j)] = |B_j|^{-1} \int_{B_j} (\mu(x) + S(x))dx,\ |B_j| > 0,$$

where $|B_j| = \int_{Bj} 1dx$ denotes the area of B_j.

The downscaler approach relates $Y(x_i)$, the point data at location x_i, to the areal data $Y(B_i)$, where B_i is the area that contains x_i [2]. Specifically,

$$Y(x_i) = \tilde{\beta}_0(x_i) + \tilde{\beta}_1(x_i)Y(B_i) + \varepsilon(x_i).$$

Here, $\varepsilon(x) \sim N(0,\sigma^2)$. $\tilde{\beta}_0(x)$ and $\tilde{\beta}_1(x)$ are spatially varying coefficients that can be expressed as a sum of an overall term and a spatial Gaussian random field as follows, $\tilde{\beta}_0 = \beta_0 + \beta_0(x)$, $\tilde{\beta}_1 = \beta_1 + \beta_1(x)$.

Inference is performed by using the integrated nested Laplace approximation (INLA) and the stochastic partial differential equation (SPDE) approaches with the `R-INLA` package [3].

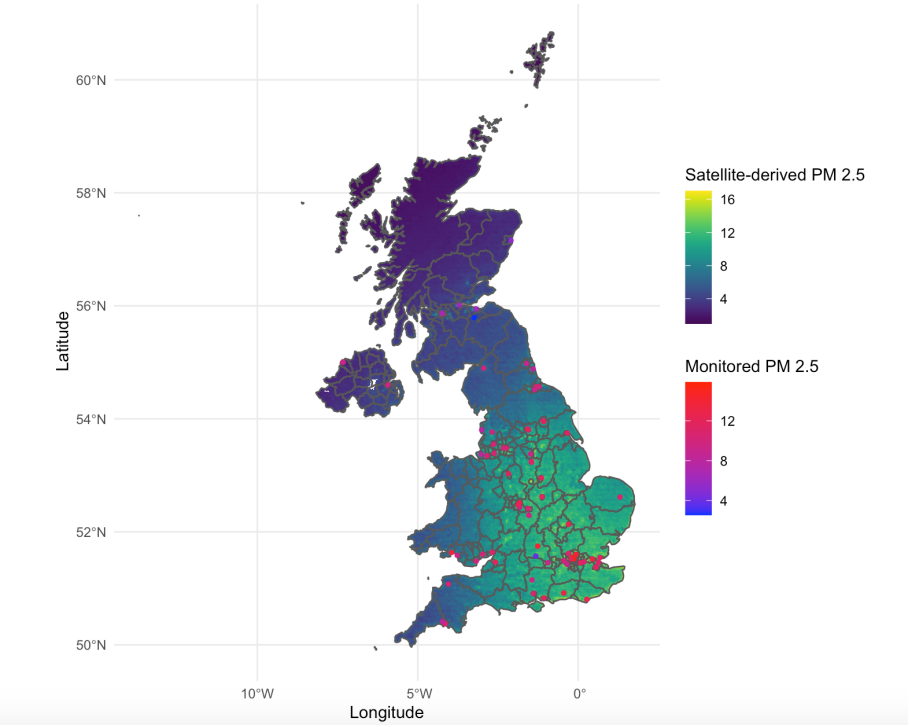


Figure 1: Fine particulate matter ($PM_{2.5}$), UK.

3. Simulation

We assess and compare the performance of the individual and ensembled-based approaches via simulation. First, we simulate a number of spatial processes that may reproduce some of the situations that can appear in real settings. Then, we take measurements of the simulated processes at different configurations of generated point and areal data. Then, we fit each of the models using the measurements taken at the generated data, and assess the performance of the models in each of the simulated scenarios using error measurements and coverage probabilities and using spatial cross-validation designs.

4. Application

The ensemble-based approach proposed allows us to combine spatially misaligned data in a wide range of applications. In this work, we use it to predict fine particulate matter ($PM_{2.5}$) in the UK using data obtained from monitoring stations and satellite-derived environmental indicators (Figure 1).

References

[1] Moraga, P., Cramb, S. M., Mengersen K. L. and Pagano, P. (2017). A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE. *Spatial Statistics* **21**, 27–41.

[2] Berrocal, V. J., Gelfand, A. E. and Holland, D. M. (2010). A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological, and Environmental Statistics* **15** 176–197.

[3] Lindgren, F. and Rue, H. (2015). Bayesian Spatial Modelling with R-INLA. *Journal of Statistical Software* **63** 1–25.

[4] Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and systems magazine* **6(3)** 21–45.

Spatial detection and mapping of urban trees using remote sensing imagery and convolutional neural networks

L. Velasquez-Camacho[1,2,*], M. Etxegarai[2] and S. de-Miguel[1,3]

[1]*Department of Crop and Forest Sciences, University of Lleida; luisa.velasquez@eurecat.org, sergio.demiguel@udl.cat*
[2]*Eurecat, Centre Tecnologic de Catalunya, Unit of Applied Artificial Intelligence, Barcelona, Spain; maddi.etxegarai@eurecat.org*
[3]*Joint Research Unit CTFC − Agrotecnio − CERCA, Solsona, Spain*
[*]*Corresponding author*

Abstract. *Urban forests and trees are the main providers of ecosystem services for more than 50% of the world's population currently living in cities. However, there is currently a large information gap on urban trees worldwide, making it difficult to quantify these services accurately. Therefore, taking advantage of new computational technologies and remote sensing we have developed a tool for detection and mapping of urban trees. Our tool, based on convolutional neural networks and computer vision, allows to know how many and where are the trees in the cities. We use ground level and high resolution aerial images as input data, finding in them a massive source of information that allows us to simulate the traditional urban forest inventory. We have achieved an accuracy of 85% of the mapping of the trees visible in the images with an accuracy of 1 meter at the center of the canopy. This research is a breakthrough for urban forestry science, due to its applicability and ability to generate standardized information.*

Keywords. *Deep learning; Computer vision; Remote sensing; Urban forests*

1. Introduction

Urban trees have been in the focus of attention in recent years, as they provide ecosystem service to the urban society that represents more than 50% of the world's total population [16]. However, there is a gap in knowledge about urban forests and urban trees worldwide that prevents us from knowing precisely what the supply capacity of these services is [6]. This gap is mainly related to two factors: first, as in natural forests, the cost, and expertise to conduct a forest inventory are very high [6]; and second, urban trees, unlike natural areas, are scattered throughout the city [7, 12].

Through spatial analysis and taking advantage of remote sensing data, the scientific community has made progress in characterizing individual trees [14]. These data sources include high-resolution satellite imagery [13], aerial imagery [4, 11], aerial LiDAR (Light Detection and Ranging) data [1, 9] and most recently ground level images. Although advances in spatial analysis using Geographic Information Systems (GIS) have enabled the advancement of remote sensing in urban areas, ground level imagery requires new processing techniques. Computer vision and artificial intelligence offer a promising alternative to the processing of these images in a scalable, generalizable and fast way [7]. Our research is based on the customization of deep learning models,

enabling the geopositioning for the detection and classification of street trees through the analysis of ground level images, and mapping through triangulation process.

2. Methods

This approach is based on transfer of learning from two deep learning models: You Only Look Once (Yolo) [10] version 5x [5] and DeepForest [15]. These models are complex convolutional networks created for object detection and classification.

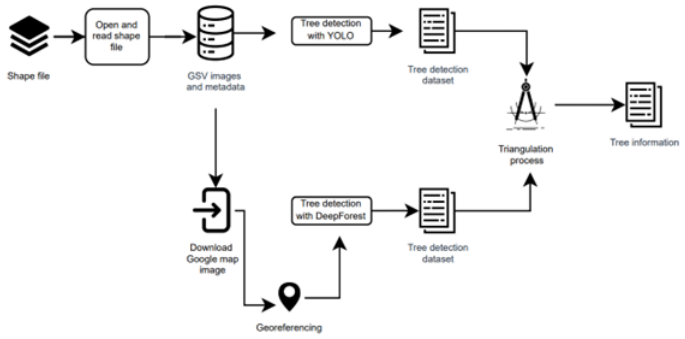


Figure 1: Flowchart of automatic urban tree mapping

The first algorithm used was YOLO, an open-source system which is fast from a computational processing point of view, because it was originally developed for video analysis [10]. Yolo-v5x was re-trained with 4500 ground level images that contain more than 25000 individual labeled trees. The model achieved a precision of 85% and a recall of 80% in the detection of urban trees in a ground level image.

The second model used for tree detection from above, in RGB satellite or aerial images, is DeepForest [15]. DeepForest is a model created and trained for the detection of canopies in natural environments, for this reason the training was shorter and simpler. We labeled 500 aerial images with 10000 individual trees for the specialization of this model in the detection of urban trees, reaching an accuracy of 80% and a recall of 75%.

To complete the mapping of urban trees, we have used a triangulation process based on the assignment of degrees within a circumference to each of the pixels of the ground level images (Figure 1), simulating a circumference at the point where the photograph was taken (known from the image metadata). This allows us to relate both images and obtain a total view of the tree (Figure 2). The model was tested in a 74 hectare plot in the city of Lleida (Catalonia), Spain, where the results obtained were contrasted with the official forest inventory.

3. Results

The tool test results are divided into two functionalities. First, the urban tree count. The model achieved a

Figure 2: Representation of degrees assignment to ground level and satellite images. a) Tree detection on satellite image and degrees assignment. b) Tree detect in four individual ground level images representing a circumference with degrees assignment.

count accuracy of 85% of the trees visible in the images with an accuracy of 1.2 meters in the assignment of coordinates to the center of the canopy.

The second result obtained is related to the ability to relate the trees detected in the ground-level with the satellite images, assigning a unique coordinate to each one. In this case the model reached an accuracy of 76% in the detection and improved average distance difference to the tree center, assigning the tree coordinate with an accuracy of 1 meter.

Figure 3: Example window of tree mapping results. Red dots represent the original forest inventory, yellow dots represent predicted the centroid of trees.

4. Discussion

Although there is research focused on the characterization of urban trees with traditional remote sensing data such as satellite imagery and LiDAR, the gap of forest inventory automation still remains, since it requires the incursion of ground level data. Some authors have started to use ground level imagery such as [7, 3] in urban vegetation characterization. However, mass data mapping has only been addressed by [3] and [8]. Both achieved accuracies above 75% in detecting and mapping urban trees. However, the accuracies achieved with respect to the distance to the center of the canopy are higher than 2.5 meters (average distance in traditional forest inventory). The spatial distribution of urban trees allows for coverage analysis [16], per capita trees [2], and others analysis. Therefore, this study represents a breakthrough in the automatic mapping of urban trees.

Acknowledgments

The author would like to thank the Lleida City Council for providing us with the urban forest inventory for the validation of this study. L. Velasquez-Camacho is a fellow Vicente Lopez PhD grant program from Eurecat

References

[1] Aval, J. Demuynck, J. Zenou, E. Fabre, S. Sheeren, D. Fauvel, M. Adeline, K. and Briottet, X. (2018). Detection of individual trees in urban alignment from airborne data and contextual information: A marked point process approach. *ISPRS Journal of Photogrammetry and Remote Sensing* **146**, 196–210.

[2] Baro, F. Caldern-Argelich, A. Langemeyer, J. and Connolly, JT. (2019).Under one canopy? assessing the distributional environmental justice implications of street tree benefits in barcelona. *Environmental science & policy* **102**, 54-64.

[3] Branson, S. Wegner, J. Hall, D. Lang, N. Schindler, K. and Perona, P. (2018).From googlemaps to a fine-grained catalog of street trees. *ISPRS Journal of Photogrammetry and Remote Sensing* **135**, 13-30.

[4] Hartling, S. Sagan, V. Sidike, P. Maimaitijiang, M. and P Carron, C. (2019) Urban tree species classification using a worldview-2/3 and lidar data fusion approach and deep learning. *Sensors* **1284**, 19.

[5] Jocher, G. Stoken, A. Chaurasia, A. Borovec, J.NanoCode012, TaoXie, Kwon, Y. Michael, K. Changyu, L. Fang, J. V, A. Laughing. tkianai. yxNONG. Skalski, P. Hogan, A. Nadar, J. imyhxy. Mammana, L. AlexWang1900. Fati, C. Montes, D. Hajek, J. Diaconu, L. Minh, M. albinxavi, M. fatig. oleg. wanghaoyang0106. (2021). ultralytics/yolov5: v6.0 - YOLOv5n Nano models, Roboflow integration, TensorFlow export, OpenCV DNN support.

[6] Ko, C. Lee, S. Jongsu, Y. Kim, D. Kang, J. (2021).Comparison of forest inventory methods at plot-level between a backpack personal laser scanning (bpls) and conventional equipment in Jeju island, South Korea. *Forests* **12(3)**, 308.

[7] Laumer, D. Lang, N. Van Doorn, N. Aodha, D. Perona, P. and Wegner, D. (2020). Geocoding of trees from street addresses and street-level images. *ISPRS Journal of Photogrammetry and Remote Sensing* **162**, 125–136.

[8] Lumnitz, S. Devisscher, T. Mayaud, J. Radic, V. Coops, N. and Griess, V. (2021). Geocoding of trees from street addresses and street-level images. *ISPRS Journal of Photogrammetry and Remote Sensing* **175**, 144–157.

[9] Matasci, G. Coops, N. Williams, D. and Page, N. (2018). Mapping tree canopies in urban environments using airborne laser scanning (als): a vancouver case study. *Forest Ecosystems* **5**, 1–9.

[10] Redmon, J. Divvala, S. Girshick, R. and Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition* **1**, 779–788.

[11] Schiefer, F. Kattenborn, T. Frick, A. Frey, J. Schall, P. Koch, B. and Schmidtlein, S. (2020).Mapping forest tree species in high resolution uav-based rgb-imagery by means of convolutional neural networks. . *ISPRS Journal of Photogrammetry and Remote Sensing* **170**, 205–2015.

[12] Stubbings, P. Peskett, J. Rowe, F. and Arribas-Bel, D. (2019). A hierarchical urban forest index using street-level imagery and deep learning. *Remote Sensing,* **11**,1395.

[13] Vahidi, H. Klinkenberg, B. Johnson, B. Moskal, L. and Yan, W. (2018). Mapping the individual trees in urban orchards by incorporating volunteered geographic information and very high resolution optical remotely sensed data: A template matching-based approach. *Remote Sensing* **10**, 1134.

[14] Velasquez-Camacho, L. Cardil, A. Mohan, M. Etxegarai, M. Anzaldi, G. and de Miguel, S. (2021). Remotely sensed tree characterization in urban areas: A review. *Remote Sensing* **13**, 4889.

[15] Weinstein, B. Marconi, S. Aubry-Kientz, M. Vincent, G. Senyondo, H. and White, P. (2020). Deepforest: A python package for rgb deep learning tree crown delineation. *Methods in Ecology and Evolution* **11**, 1743–1751.

[16] Xu, G. Jiao, L. Yuan, M. Dong, T. Zhang, B. and Du, C. (2020). How does urban population density decline over time? an exponential model for chinese cities with international comparisons. *Landscape and Urban Planning* **183**, 59–67.

Log-Gaussian Cox Processes with Integro-Differential Equations: Modelling 112-Emergency Calls

D. Payares[1,*] , J. Platero[2] and J. Mateu[2]

[1]*ITC Faculty Geo-Information Science and Earth Observation, University of Twente, Enschede, The Netherlands; d.e.payaresgarcia@utwente.nl*
[2]*Department of Mathematics, University Jaume I, Castellon, Spain; platero@uji.es, mateu@mat.uji.es*
[*]*Corresponding author*

Abstract. *Spatial and spatio-temporal point processes for criminological applications are popular in literature due to crimes' clustering and context-dependent nature. Notably, most approaches use Log-Gaussian Cox processes (also called doubly stochastic) to model crime event data since the technique accounts for spatio-temporal dependence, covariates inclusion, and clustering phenomena. However, events with complex dispersion processes (e.g. crimes or derived emergency calls) cannot be trivially introduced into point process frameworks, mainly when data are recorded in continuous space but discrete-time. To account for events' complex temporal dynamics while consider discrete-time series in spatio-temporal data, we model emergency calls using a hybrid model between Log Gaussian Cox processes and Stochastic Integro-Differential Equations (LGCP-driven-SIDE). We show the advantages of the LGCP-driven-SIDE approach for inference and prediction using data from 112 emergency calls in Valencia from 2010 to 2020. The model accurately measures the influence of covariates, models the data's spatio-temporal dynamics, and predicts the system's future behaviour.*

Keywords. *Emergency calls; Log Gaussian Cox Processes; Spatio-temporal dynamics; Stochastic integro-differential equation.*

1. Introduction

Emergency calls arise when immediate action is required to deal with incidents such as accidents, wildfires, crimes or medical assistance. Typically, emergency calls provide information not only about the description of the incident but also about their location and time; the latter characteristics are essential for prompt response [1]. Authorities leverage the spatial and temporal behaviour of the emergency calls to allocate resources and infrastructure for effective response, high-risk event areas identification (e.g. crimes, traffic accidents), and contingency strategies development. As emergency calls often mirror crimes, spatio-temporal studies adapt popular point process methodologies for crime data to analyse distress signals. A popular approach are the Log-Gaussian Cox processes (LGCP). LGCP have several appealing properties that facilitate model estimation, interpretation and simulation [2]. Moreover, the intensity is an stochastic process that allows capturing stochastic spatial and space-time dependence [3]. Although specifying the spatio-temporal structure of data in LGPC is a straightforward task, the temporal dynamics of crime phenomena can not be trivially introduced; complex dispersion processes such as advection and diffusion [4, 5] need to be carefully incorporated into the modelling framework. To account for a system's complex temporal dynamics and to reinforce the discrete-time

series definition in LGCP, [6] introduced stochastic integro-difference equations (SIDE). Given that crimes (and thus, emergency calls) exhibit similar patterns in space and time as as conflict events, and they are recorded in discrete-time format, in this paper, we exploit [6] methodology to analyse the dynamics of 112-emergency calls in Valencia, Spain for ten years (2010 - 2020). This work aims to understand and forecast the spatio-temporal character of emergency calls in Valencia to guide mitigation strategies and policy responses to criminal activity. First, we model the emergency calls throughout the city neighbourhoods from 2010 to 2019, considering the data spatio-temporal behaviour and related geographical covariates. Then, we validate the model by predicting the emergency incidents for 2020.

2. Log-Gaussian cox processes and stochastic integro-differential equations

We chose the Log-Gaussian Cox process (LGCP), e.g., the logarithm of the event intensity is assumed to be a Gaussian Process, to model the emergency calls data. A discrete time division is considered as a discrete-time series of continuous-space LGCPs since the temporal range is discrete. Formally, let $k \in \mathcal{K}$, $\mathcal{K} = \{1, \ldots, K\}$ denote a discrete-time index set and $\{z_k(\mathbf{s})\}$, $z_k(\mathbf{s}) \sim \text{GP}\left(\mu_k(\mathbf{s}), \sigma_k^2 \Psi_k(\mathbf{s}, \mathbf{r})\right)$, a set of temporally correlated spatial Gaussian Processes (GPs), each with mean $\mu_k(\mathbf{s})$ and covariance function $\sigma_k^2 \Psi_k(\mathbf{s}, \mathbf{r})$. For each k, the point process intensity function is defined as $\lambda_k(\mathbf{s}) = \exp(z_k(\mathbf{s}))$. The mean function of $z_k(\mathbf{s})$ can be associated to explanatory variables to reduce prediction uncertainty. Let $\mathbf{d}(\mathbf{s})$ be a vector of spatially referenced covariates and $\mathbf{b}^T$ the corresponding regression coefficients; the intensity of the LGCP at time k then is given by $\lambda_k(\mathbf{s}) = \exp\left(\mathbf{b}^T \mathbf{d}(\mathbf{s}) + z_k(\mathbf{s})\right)$. The temporal dynamics of the intensity functions through $z_k(\mathbf{s})$ can be defined under the stochastic integro-difference equation (SIDE) framework. The SIDE is a flexible modelling tool that represents temporal dynamic effects such as diffusion and dispersal. Formally, the SIDE associates the spatio-temporal dependent variable $z_k(\mathbf{s})$ to $z_{k+1}(\mathbf{s})$ through the following integral equation

$$z_{k+1}(\mathbf{s}) = \int_D k_I(\mathbf{s}, \mathbf{r}) f_1(z_k(\mathbf{r})) \mathrm{d}\mathbf{r} + e_k(\mathbf{s}), \tag{1}$$

where $k_I(\mathbf{s}, \mathbf{r})$ is the mixing kernel in the integral and $e_k(\mathbf{s}) \sim \text{GP}(\mu_Q(\mathbf{s}), k_Q(\mathbf{s}, \mathbf{r}))$ is an added disturbance, modelled as a Gaussian field with mean $\mu_Q(\mathbf{s})$ and covariance function $k_Q(\mathbf{s}, \mathbf{r})$, and D is the spatial domain under investigation. The nonlinear mapping $f_1(\cdot)$ distorts the field in the sedentary stage; the identity $f_1(z_k(\mathbf{r})) = z_k(\mathbf{r})$ is adopted here. We can further decompose the kernel, the mean disturbance and the field through basis functions representation to reduce the computational burden and facilitate the inference between the LGCP and the SIDE. Then,

$$\begin{aligned} z_k(\mathbf{s}) &= \boldsymbol{\phi}(\mathbf{s})^T \mathbf{x}_k, \\ \mu_Q(\mathbf{s}) &= \boldsymbol{\phi}(\mathbf{s})^T \boldsymbol{\vartheta}, \\ k_I(\mathbf{s}, \mathbf{r}) &= \boldsymbol{\phi}(\mathbf{s})^T \boldsymbol{\Sigma}_I \boldsymbol{\phi}(\mathbf{r}), \\ k_Q(\mathbf{s}, \mathbf{r}) &= \boldsymbol{\phi}(\mathbf{s})^T \boldsymbol{\Sigma}_Q \boldsymbol{\phi}(\mathbf{r}), \end{aligned}$$

where $\boldsymbol{\phi}(\mathbf{s}) \in \mathbb{R}^n$ is the vector of basis functions, $\mathbf{x}_k \in \mathbb{R}^n$ and $\boldsymbol{\vartheta} \in \mathbb{R}^n$ are weights which reconstruct the spatio-temporal field and the disturbance mean respectively and where $\boldsymbol{\Sigma}_I \in \mathbb{R}^{n\times n}$ and $\boldsymbol{\Sigma}_Q \in \mathbb{R}^{n\times n}$ reconstruct the kernel covariance function and the disturbance covariance function, respectively. Under this decomposition, the SIDE (Eq. 1) can be rewritten as

$$\mathbf{x}_{k+1} = \mathbf{A}(\boldsymbol{\Sigma}_{\boldsymbol{I}})\mathbf{x}_k + \mathbf{w}_k(\boldsymbol{\vartheta}, \boldsymbol{\Sigma}_Q), \tag{2}$$

where $\mathbf{A}(\boldsymbol{\Sigma}_{\boldsymbol{I}}) \in \mathbb{R}^{n\times n}$ and $\mathrm{w}_k \in \mathbb{R}^n$ is a Gaussian colored noise term with mean $\mathrm{E}[\mathbf{w}_k] = \boldsymbol{\vartheta}$ and covariance $\mathrm{cov}[\mathbf{w}_k] = \boldsymbol{\Sigma}_{\boldsymbol{Q}}$. The states $\mathscr{X}_K = x_{0:k} = \{x_k\}_{k=0}^K$ and the unknown parameters $\boldsymbol{\theta} = \{\boldsymbol{\vartheta}, \boldsymbol{\Sigma}_I, \boldsymbol{\Sigma}_Q^{-1}\}$ need to be estimated from the data $\mathscr{Y}_k = \{y_k\}_{k=1}^K$ where each y_k is the set of coordinates of the logged events at the k-th time point.

3. Spatio-temporal dynamics of 112 emergency calls

Distances to banks, ATMs, bars, cafes and restaurants were included as the deterministic component of the intensity $\lambda_k(s)$ since they displayed significant association with the averaged spatial intensity of the emergency calls in Valencia. We found that emergency calls happen near economic facilities; offenders will find victims in these areas, which may represent a monetary benefit. While bars and cafes gather many potential victims, emergency calls are located away from these places. This is because bars and cafes have enhanced security systems as they are prone to crimes such as robbery and assault. Our findings indicate that emergency calls occur close to restaurants. Criminals particularly target restaurants for robbery, burglary and theft as these accumulate large sums of cash in daily operations.

Figure 1 display the weekly average fractional growth and decay of emergency calls in Valencia. As expected, most of the city has experienced an increase in emergency calls. Despite The city centre neighbourhoods having the highest number of events, these areas have not witnessed a significant increase in emergency calls. Contrarily, areas with sparse events report the neighbourhoods that have become 112 calls hot-spots over the study period. The intensity of emergency calls has decreased sparsely across Valencia.

Since we have accurately modelled the spatio-temporal dynamics of the emergency calls dynamics in Valencia, we can now estimate their behaviour in the future. We predicted the number of emergency calls in Valencia for the first 40 weeks of 2020. We do not predict the counts for the entire 2020 as data was not available after October. We transformed the counts into logarithms to stabilise the variances. Table 1 shows the Pearson correlation coefficients between the predicted and true counts and log counts. The coefficients show a strong correlation (0.87 for counts and 0.89 for log counts) between the estimated and reported values, demonstrating the model's predictive power.

Figure 2 shows the scatter plot between the log median prediction of the model and the log reported cases for 2020. Note how the circles concentrate around the ideal prediction indicating how closely our predicted data mirrors the observed data. However, although the median value closely resembles the actual value, the error bar plot shows that the median estimates for some neighbourhoods are highly uncertain. A high amount of uncertainty is linked to the existence of multiple zero observations in some neighborhoods.

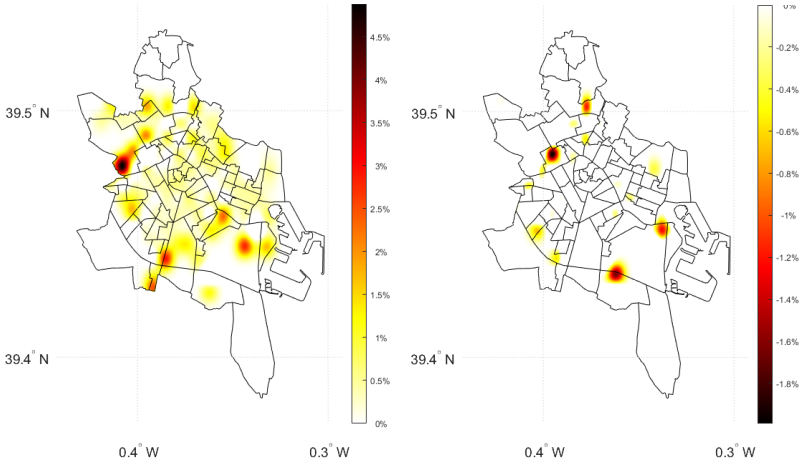


Figure 1: Posterior mean fractional growth/decay of emergency calls per week in Valencia (2010-2019).

Prediction	ρ	$p-value)$
counts vs predicted counts	0.8724	< 0.001
log counts vs predicted log counts	0.8994	< 0.001

Table 1: Pearson correlation coefficient between the SIDE model predictions and ground true values (counts and log counts) for the first 40 weeks of 2020.

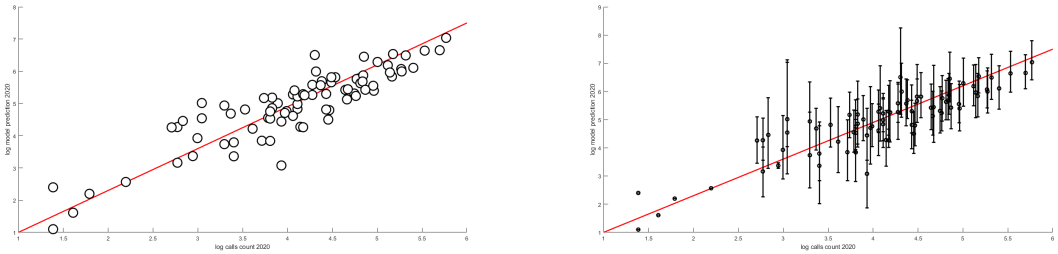


Figure 2: Scatter plot and error bar plot (99% confidence intervals) between the log median predictions and log actual values for 2020. Each circle represents a neighbourhood in Valencia. The red line refers to the ideal prediction.

References

[1] Dunnett, S., Leigh, J., and Jackson, L. (2019). *Optimising police dispatch for incident response in real time*. Journal Of The Operational Research Society, **70** (2): 269 – 279.

[2] Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. (2011). *Self-exciting point process modeling of crime*. Journal Of The American Statistical Association, **106** (493): 100 – 108.

[3] Shirota, S. and Banerjee, S. (2019). *Scalable inference for space time gaussian cox processes*. Journal Of Time Series Analysis, **40** (3, SI): 269 – 287.

[4] Tang, Y., Zhu, X., Guo, W., Ye, X., Hu, T., Fan, Y., and Zhang, F. (2017). *Non-homogeneous diffusion of residential crime in urban china*. Sustainability, **9** (6).

[5] Short, M. B., D'Orsogna, M. R., Pasour, V. B., Tita, G. E., Brantingham, P. J., Bertozzi, A. L., and Chayes, L. B. (2008). *A statistical model of criminal behavior*. Mathematical Models and Methods In Applied Sciences, **18** (1): 1249 – 1267.

[6] Zammit-Mangion, A., Dewar, M., Kadirkamanathan, V., and Sanguinetti, G. (2012). *Point process modelling of the afghan war diary*. Proceedings of the National Academy of Sciences, **109** (31): 12414 – 12419.

Assessing the Effect of Model-based Geostatistics Under Preferential Sampling for Spatial Data Analysis

A.V. Ribeiro-Amaral [1,*] and P. Moraga[1]

[1]*King Abdullah University of Science and Technology, CEMSE Division. Thuwal 23955-6900, Saudi Arabia; andre.ribeiroamaral@kaust.edu.sa and paula.moraga@kaust.edu.sa*
[*]*Corresponding author*

Abstract. *Geostatistics is concerned with the estimation and prediction of spatially continuous phenomena using data obtained at a discrete set of spatial locations. In geostatistics, preferential sampling occurs when the locations are not independent from the spatial process of interest, and common geostatistical approaches may yield wrong inferences if preferential sampling is not taken into account. In this work, we briefly review common geostatistical models and inference procedures for a preferential sampling setting. We conduct a simulation study to assess the performance of different models in different scenarios and demonstrate that although models that take preferential sampling into account may be needed in some situations, they may perform worse than the usual geostatistical approaches when they are used to solve problems that do not do preferential sampling. In summary, although preferential sampling is important, careful consideration of the obtained data is needed to determine the most appropriate modeling approach for each studied problem.*

Keywords. *Spatial Statistics; Geostatistics; Preferential Sampling; INLA; Air Pollution.*

1. Introduction

Many different problems in spatial statistics can be seen as problems that belong to the geostatistics domain, that is, problems that are characterized by the study of an underlying spatial process that has been observed at a discrete set of locations. For instance, one can be interested in studying how air pollution is distributed in a given region. For these problems, it is usually assumed that the sampling process is independent of the process of interest. However, this may not always be the case, and in situations where this assumption does not hold, we say we have *preferential sampling*, as in [1].

Trying to accommodate the dependence between the sampling process and the process of interest into the modeling approach is not trivial, and methods that take this dependence into account to obtain valid inferences have been developed [1, 2]. Before, models that account for preferential sampling were fitted by rewriting the likelihood function in a way that it could be seen as an expectation, which allowed people to approximate it by Monte Carlo methods [1]. However, more recently, a Bayesian approach relying on the Integrated Nested Laplace Approximation (INLA) and Stochastic Partial Differential Equation (SPDE) methods started to be employed.

2. Geostatistical Model

Geostatistics refers to the analysis of a data set sampled from a spatially continuous domain, say $\mathcal{A}$, at a discrete set of locations $\{x_i\}_{i\in I}$, such that $I = \{1,2,\cdots,n\}$, for $n \in \mathbb{N}$, of a process $S(\cdot)$, where $x_i \in \mathcal{A}$, $\forall i$. For most of the applications, $\mathcal{A}$ is usually a subset of $\mathbb{R}^d$, with $d = \{1,2\}$. Moreover, if we let $X = (X_1,\cdots,X_n)$ be a random vector representing the locations where the process $S(X) = (S(X_1),\cdots,S(X_n))$ is observed, and if the distribution of X is determined in such a way that X is not stochastically independent from $S(X)$, then we say we are dealing with *preferential sampling*, as in [1].

A geostatistical model to predict a spatially continuous process can be defined as follows. Suppose that Y_i denotes the observed value of a noisy version of a spatial process $S(x_i)$ at some given location x_i, for $i \in \{1,\cdots,n\}$, in the following manner:

$$Y_i = \mu + S(x_i) + \varepsilon_i, \tag{1}$$

where ε_i are independent Gaussian zero-mean random variables with $\text{Var}(\varepsilon_i) = \sigma^2_\varepsilon$. Also, $S(x_i)$ can be assumed to have zero mean; in this case, $\mathbb{E}(Y_i) = \mu$, for all $i \in \{1,\cdots,n\}$. Moreover, let $x = (x_1,\cdots,x_n)$ be a realization from $X = (X_1,\cdots,X_n)$ and $S(x) = (S(x_1),\cdots,S(x_n))$ a realization from $S(X) = (S(X_1),\cdots,S(X_n))$. In this case, although Model (1) is fairly common, it usually assumes that X is stochastically independent from $S(X)$, which is not reasonable in many situations when preferential sampling is presented and can yield invalid inferences.

To allow for the stochastic dependence between X and $S(\cdot)$, we will consider, based on [1], the following additional assumptions for Model (1): [(a)] S is a stationary and isotropic Gaussian random field with mean zero, variance σ^2, and covariance function $r(h;\theta) = \text{corr}(S(x_1),S(x_2))$, for $h \neq 0$, such that h is the Euclidean distance between x_1 and x_2, [(b)] $X|S \sim$ Poisson Process$(\lambda(x))$ with intensity $\lambda(x) = \exp\{\alpha + \beta \cdot S(x)\}$, for $\alpha,\beta \in \mathbb{R}$, and [(c)] conditional on S and X, $Y = (Y_1,\cdots,Y_n)$ is a vector of independent Gaussian random variables, such that $Y_i \sim \text{Normal}(\mu + S(x_i), \sigma^2_\varepsilon)$, $\forall i \in \{1,\cdots,n\}$.

2.1 Inference

Since the original paper that introduced the preferential sampling idea was published by [1], people have been working on this class of problems using different approaches. Here, we will present two, namely the original idea, and the a method that uses INLA and the SPDE techniques.

1. Start by recalling that, if we consider Model (1) and if we want to predict the value of the process in, say, x_0, we can use, for instance, the Best Unbiased Linear Predictor (BLUP), namely *Kriging*. However, in order to do this, we have to be able to estimate the parameters of the model. In particular, if $S(x)$ is a Gaussian random field with a covariance structure described by $\Sigma(\theta)$, this can be done through the Maximum Likelihood method.

   However, if X and $S(X)$ are not independent, then the likelihood function $\mathcal{L}(\theta)$, given the data, is

$$\mathcal{L}(\theta) = [X,Y] = \int [X,Y,S]dS = \int [Y|S,X][X|S][S]dS. \tag{2}$$

Therefore, to determine θ that maximizes $\mathcal{L}(\theta)$, one has to solve the integral in Equation (2). And for this problem, [1] has proposed a way to approximate $\int [Y|S,X][X|S][S]dS$ using a Monte Carlo method. Finally, from the approximated likelihood function, they could estimate the parameters by determining θ that maximizes $\mathcal{L}_{\text{Approx.}}(\theta)$.

2. An alternative approach to estimate the model parameters and make prediction for Model (1) is to use the INLA and SPDE approaches, which can be easily implemented with the `R-INLA` package [5]. In a nutshell, INLA is a method for approximating Bayesian inference in latent Gaussian models [4]. In particular, models are of the form

$$\begin{aligned} y_i|S(x_i),\theta &\sim \pi(y_i|S(x_i),\theta), \text{ for } i \in \{1,\cdots,n\} \\ S(x)|\theta &\sim \text{Normal}(\mu(\theta), Q(\theta)^{-1}) \\ \theta &\sim \pi(\theta), \end{aligned}$$

where $y = (y_1,\ldots,y_n)$ is the vector or observed values, $x = (x_1,\ldots,x_n)$ is a Gaussian random field, and $\theta = (\theta_1,\ldots,\theta_k)$, for some $k \in \mathbb{N}$, is a vector of hyperparameters. $\mu(\theta)$ and $Q(\theta)$ represent the mean vector and the precision matrix, respectively.

From the above formulation, notice that our Model (1) satisfies all described conditions to be classified as a latent Gaussian model, and therefore we can take advantage of the the INLA method. To fit Model (1) model using INLA, we will take an SPDE approach. As showed in [6], a Gaussian random field with Matérn covariance matrix can be expressed as a solution to the following SPDE:

$$(\kappa^2 - \Delta)^{\alpha/2}(\tau S(x)) = \mathcal{W}(x),$$

where Δ is the Laplacian, $\mathcal{W}(s)$ is a Gaussian white-noise random process, α controls the smoothness of the random field (in particular, $\alpha = \nu + d/2$, such that ν is the smoothness parameter from the Matérn model and d is the dimension), τ controls the variance, and κ is a scale parameter. Based on this result, [3] proposed a new approach to represent a Gaussian random field with Matérn covariance, as a Gaussian Markov Random Field (GMRF), by representing a solution to the SPDE using the finite element method. This representation implies a sparse precision matrix for the spatial effects, allowing the implementation of fast computational methods to do inference.

3. Simulation

In this section, we conduct a simulation study to compare the performance of different geostatistical models that account and do not account for preferential sampling under different scenarios. In our simulation study, we consider the unit square $[0,1]\times[0,1]$ as the study region, and simulate values from a Gaussian process $S(\cdot)$ with mean $\mu \in \mathbb{R}$ and Matérn covariance function.

We wish to assess the performance of the models in scenarios that have been simulated with and without preferential sampling, and to acomplish this, we generate n points x_i, such that $i = 1,\cdots,n$, where we obtain measurements of the simulated processes. In preferential sampling scenarios, points are a realization from $X|S \sim$ Poisson Process$(\lambda(x))$ with intensity $\lambda(x) = \exp\{\alpha + \beta \cdot S(x)\}$, such that $\beta > 0$. On the other hand, in non-

preferential sampling scenarios we consider $X \sim$ Poisson Process$(\lambda(x))$ with constant intensity $\lambda(x) = \exp\{\alpha\}$, that is, we set $\beta = 0$.

Different scenarios for the data generation procedure were considered, but here we will present just two of them: one corresponding to a non-preferential sampling setting ($\beta = 0$) and another one to a preferential sampling setting ($\beta > 0$). In all cases, we set $\mu = 0$ and $\sigma^2_\varepsilon = 1$. Also, after generating data, we fitted Model (1) under the assumption that X and $S(X)$ are **independent** (`A1`), and we fitted Model (1) under the assumption that X and $S(X)$ are **dependent** (`A2`). Both models were fitted with `R-INLA` [5].

For instance, Figure 1 shows simulated scenario under preferential sampling with the corresponding predictions made based on models `A1` and `A2`. Visual inspection suggests better results for model `A2`.

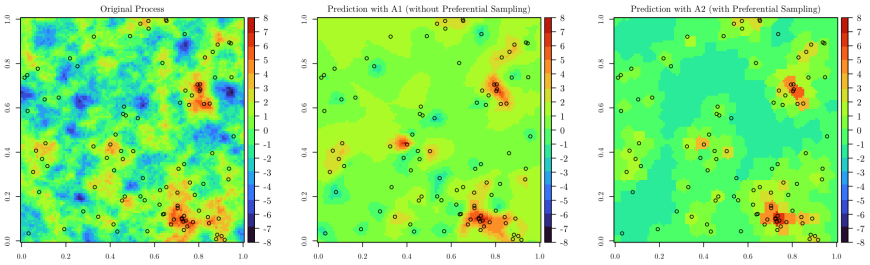


Figure 1: Simulated S and $X|S$ processes (left) and predictions using models `A1` (center) and A2 (right).

3.1 Results

For the 2 simulated scenarios, we predicted the processes values on the region of interest using a model that do not account, and a model that do account for preferential sampling, namely models `A1` and `A2`, respectively. Then, we assessed the performance of the models by using the Mean Squared Error (MSE), which can be computed as $\text{MSE} = n^{-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, such that y_i and $\hat{y}_i$ correspond to the observed and predicted values, respectively, at location x_i, $i = 1, \cdots, n$. We simulated $m = 50$ data sets from each of the 2 scenarios and computed the mean and quantiles of the MSE values to have stable results (Table 1).

Scenario	Mean (and SD) of MSE_{A1}	5^{th} and 95^{th} perc. of MSE_{A1}	Mean (and SD) of MSE_{A2}	5^{th} and 95^{th} perc. of MSE_{A2}
Non-pref. sampl.	3.16 (0.33)	2.70—3.75	3.59 (0.48)	3.03—4.50
Pref. sampl.	5.63 (1.07)	4.19—7.80	3.44 (0.59)	2.80—4.80

Table 1: Computed statistics for the MSEs for models `A1` and `A2` in the two scenarios.

From Table 1, we can see that, for the scenario in which preferential sampling was **not** considered for the data generation procedure, `A1` performed *slightly* better than `A2`; however, for data generated with preferential sampling, model `A2` performed *much* better than model `A1` (with respect to the MSE). In this regard, some

knowledge of the sampling process may be needed in order to choose an appropriate model for our problem. Thus, although preferential sampling may play an important role in the modeling procedure, careful consideration of the obtained data is required to determine the appropriate technique.

References

[1] Diggle, P. J., Menezes, R. and Su, T.-l. (2010). Geostatistical inference under preferential sampling. Journal of the Royal Statistical Society: Series C (Applied Statistics) *59*, 191–232.

[2] Gelfand, A. E., Sahu, S. K. and Holland, D. M. (2012). On the effect of preferential sampling in spatial prediction. Environmetrics *23*, 565–578.

[3] 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*, 423–498.

[4] Moraga, P. (2019). Geospatial health data: Modeling and visualization with R-INLA and Shiny. CRC Press.

[5] 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*, 319–392.

[6] Whittle, P. (1963). Stochastic-processes in several dimensions. Bulletin of the International Statistical Institute *40*, 974–994.

A simple test to detect preferential sampling in Geostatistics

I. Natário[1,2*] , A. Monteiro[2] , I. Figueiredo[3,5] , P. Simões[4,2] and M.L. Carvalho[5]

[1] *Department of Mathematics, NOVA School of Science and Technology, Portugal; icn@fct.unl.pt*
[2] *NOVA MATH - Center for Mathematics and Applications (CMA), NOVA University of Lisbon, Portugal; andreiaforte50@gmail.com*
[3] *Portuguese Institute for Sea and Atmosphere (IPMA), Portugal; ifigueiredo@ipma.pt*
[4] *Military Academy Research Center - Military University Institute (CINAMIL), Portugal; pc.simoes@campus.fct.unl.pt*
[5] *Center of Statistics and its Applications of the University of Lisbon (CEAUL), Portugal; mlucilia.carvalho@gmail.com*
[*] *Corresponding author*

Abstract. *Geostatistics infers about a spatially continuous phenomenon, sampled in a finite number of locations, where it usually measured with error. Preferential sampling exists whenever there is stochastic dependence between the spatial and sampling processes. Ignoring this problem drives to incorrect and biased estimates. Thus, identifying this problem is very important, but not always easy to implement and understand. In this work, a quite simple test, easy to implement, is presented for this purpose overcoming the previous concerns, based on the correlation between the number of sampled points and the values of the corresponding measures. Simulation studies were run, both on regular and irregular shaped regions, given different levels of preferentiability, including or not a relation with a covariate. These results were quite encouraging, although some issues still need to be better worked out, which became clearer when the test was applied to a real set of fish capture data.*

Keywords. *MLC test; Preferential sampling; Geostatistics.*

1. Introduction

A common geostatistical model is:

$$Y_i = \mu + S(x_i) + \varepsilon_i, \qquad x_i \in \mathcal{D},\ i = 1,\ldots,n, \tag{1}$$

where $\boldsymbol{Y} = (Y_1,\ldots,Y_n)$, $Y_i \in \mathbb{R}$, $i = 1,\ldots,n$, are imperfect observations of the true surface of the phenomenon of interest $\{S(x) : x \in \mathcal{D} \subseteq \mathbb{R}^2\}$, taken in locations $\boldsymbol{X} = (X_1,\ldots,X_n)$, according to some sampling scheme, $\boldsymbol{X} = \boldsymbol{x}$; S is modeled as a stationary Gaussian process with constant mean (zero) and variance (σ^2) and correlation function depending only on the distance between locations as, for example, the Matérn family of flexible correlations functions, [3]; $\mu \in \mathbb{R}$ is a constant mean parameter, ε_i are i.i.d. random errors with $E[\varepsilon_i] = 0$ and $Var(\varepsilon_i) = \tau^2$, the nugget variance.
Model (1) assumes that the sampling process $\boldsymbol{X}$ is independent of the spatial process S, but it happens frequently that the choice of the sampling locations is done in a preferential way, according to the gradient of what is being measured, [2]. This is the case of fishery data, for which sampling locations are defined by the fishermen's desire to maximize capture.

Ignoring the problem of preferential sampling drives to misleading inferences [2], incorrect estimates [5] and large biases in the prediction of the spatial process S [7], since common statistical methods condition response inference on fixed locations [6]. Therefore, prior to inference, one should test whether preferential sampling is a problem in data to use [4].
Taking preferential sampling effect into account in modelling was first considered in [2], by joint modeling the observations $\boldsymbol{Y}$ in (1) and the locations $\boldsymbol{X}$, conditionally on their mutual dependence on S. $X|S$ is taken as Log-Gaussian Cox model for point patterns with intensity function

$$\lambda(x) = \exp(\alpha + \beta S(x)). \tag{2}$$

Preferentiability, if any, is captured by magnitude of the parameter β, but this does not really set up a test to conclude about this problem, it is more an indication about it. Tests for assessing wether the sampling locations selected to monitor a spatial process depend stochastically on the process they are measuring have been proposed previously, as [6] and works herein referred, usually based on marked point processes and hence beyond reach to most researchers in practice.
It is presented in this work a very simple test for preferential sampling, based on the correlation between the number of sampled points and the values of the corresponding measures. It is straightforward to implement and the first results here presented, based on simulations and also real data, are quite interesting.

2. MLC Test

This section presents a test for the null hypothesis of stochastic independence between the sampling locations $\boldsymbol{X}$ selected to monitor a spatial process $\{S(x), x \in \mathcal{D}\}$ and the process itself.
The idea of this test is that if there is dependence between sampling locations and spatial process (and thus its measures), it is expected that, considering a partition of the area of interest, the number of sampled points in each partition division is correlated with the mean value of the corresponding measures.
Therefore, considering the domain of interest $\mathcal{D}$, the sampling locations $\boldsymbol{X} = (X_1, \ldots, X_n)$ and the observations $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, the test considers a regular partition of $\mathcal{D}$ formed by d equal sized squares and computes the number $N_{P,j}$ of sampling points within the jth square, $j = 1, \ldots, d$, and the mean value of the observations within the jth square, $\bar{Y}_j$, $j = 1, \ldots, d$. The test proceeds by performing a Spearman correlation test between the number of points in each cell $\{N_{P,j}\}$ and the average measures in each cell $\{\bar{Y}_j,\}$, corrected for ties. The use of the Spearman correlation test do not restrict measures $\boldsymbol{Y}$ to be Gaussian or even of a continuous type. This test was named *Means and Locations Correlation test*, MLC test.
The square side size is data dependent and it is proposed here, as a rule of thumb, for regularly shaped regions, to be $\frac{h}{12}$, where h is the maximum distance between all the n points. This is an open question, has been considered in other contexts requiring grid definitions in spatial statistics, and needs to be further studied, maybe under MAUP or the sampling theorem.

3. Simulation Study

The first simulation study presented considers an almost regularly shaped domain of interest where it was

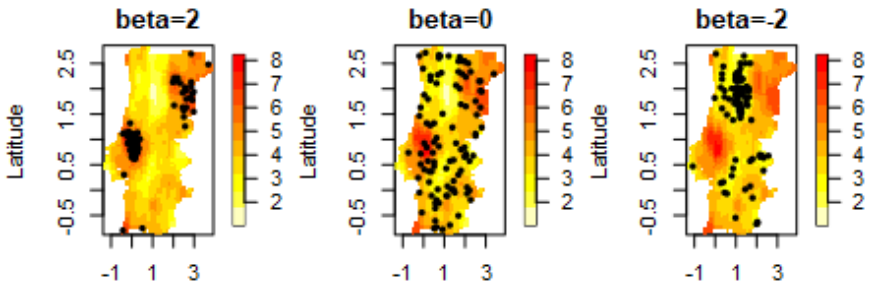


Figure 1: Realization of a spatial process with $\mu = 4$ and preferential and non preferential simulated data.

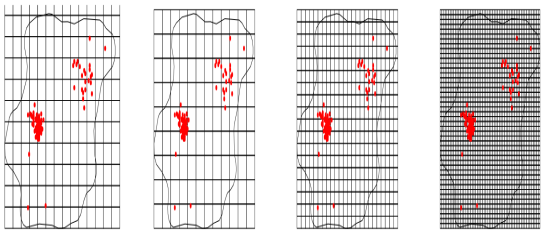


Figure 2: Considered grids and first simulated data set, grid side sizes: 0.44, 0.5, 0.25 and 0.1.

simulated a true realization $s^*(x)$ of a spatial process described by a Matérn field with $\kappa = 1$, $\sigma^2 = 2.5$ and scale parameter $\phi = 0.7$, [3]. Additionally, the true field mean was allowed to be $\mu = 4$ or given by the spatial covariate Euclidean distance to coast. Based on the same field realization, 100 points were sampled with and without preferability, $\beta = 2$ and $\beta = -2$ for preferability and $\beta = 0$ for non-preferability, from joint models (2) with intensity given by $(\beta s^*(x))$ for locations, and (1) with Gaussian mean given by $(s^*(x))$ and nugget variance $\tau^2 = 0.2$. In Figure 1 is a field realization and simulated data for no covariate case. Five different grids were considered for the test, varying the grid side size: $\frac{h}{12}$, where h is the maximum distance between all the n points, 0.5; 0.25; 0.1; and 0.05 - Figure 2 for grids and first data set. For all the grids considered, the test rejected at a 10% significance level the preferential sampling hypothesis only for the non preferential simulated data, both with and without covariate.

Next, a single grid side size was fixed to $\frac{h}{12}$, where h is the maximum distance between all the n points. Using this rule, 50 different spatial patterns $s_k^*(x)$, $k = 1,\ldots,50$, were generated from the same spatial process defined before. Also, the true field mean was allowed to be $\mu = 4$ or $dist(x)$. Then the degree of preferability was allowed to vary through a choice of β values in $\{-2,-1.5,-1,-0.5,0,0.5,1,1.5,2\}$, for simulating 100 points of data, as before, with a point intensity process here given by $\left(\beta s_k^*(x)\right)$, in a total of 50 replicas for each combination $(\beta, s_k^*(x))$. The preferential sampling hypothesis was not rejected, at a 10% significance level, for almost 100% replicas for $\beta \in \{-2,-1.5,-1.0,1,1.5 \text{ and } 2\}$, for both with and without covariate. For $\beta = -0.5$ and $\beta = 0.5$ these values dropped to 90%. For $\beta = 0$, the test rejected the preferential sampling hypothesis for all replicas, as expected.

The second simulation study presented is based on a real data set of captures of Black Scabbardfish off the Portuguese coast, a deep water species that by that reason has its captures confined into a very irregular shaped region, posing an extra challenge to perform tests that depend on the geometry of the regions. The real data set,

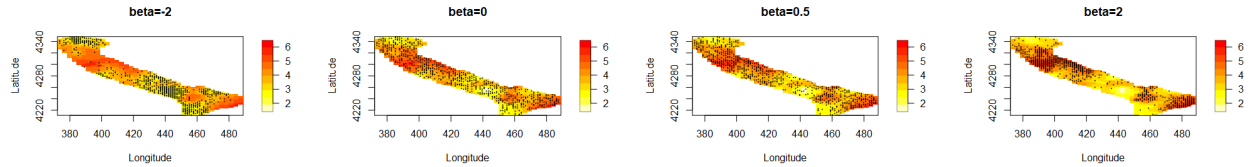


Figure 3: A realization of a spatial process with $\mu = 4$ and preferential and non preferential simulated data, left to right corresponding to $\beta = -2$, $\beta = 0$, $\beta = 0.5$ and $\beta = 2$.

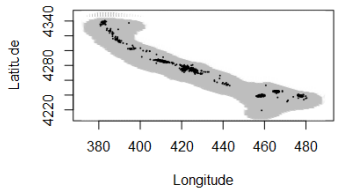


Figure 4: Black scabbardfish location of captures off the southern Portuguese coast from 2009 to 2013.

corresponding to captures in a period of 5 years, comprehended data on a total of 733 observed locations and estimated a preferential parameter β of 0.5 for model (2).
In the simulation study, 50 different spatial patterns $s_k^*(x)$, $k = 1, \ldots, 50$, were generated from a spatial process defined by a Matérn field with $\kappa = 1$, $\sigma^2 = 1.5$ and scale parameter $\phi = 15$ and true field mean of $\mu = 4$. Allowing the degree of preferability to vary, for β values in $\{-2, -0.3, 0, 0.3, 2\}$), 100 points were simulated accordingly as before, using a nugget variance of $\tau^2 = 0.01$. See Figure 3 for field realization and simulated data for one of the replicas and for the different values of β considered.
The preferential sampling hypothesis was not rejected, at a 10% significance level, for almost 100% replicas for $\beta \in \{-2,$ and $2\}$. For $\beta = -0.5$ and $\beta = 0.5$ these values dropped to 90%. For $\beta = 0$, the test rejected the preferential sampling hypothesis for all replicas.

4. Real Data and Discussion

Finally the MLC test was performed to the real data already mentioned, provided by the Instituto Português do Mar e da Atmosfera (IPMA) on Black Scabbardfish captures in a period of 5 years, as described in [1] and displayed in Figure 4. As stated before, a degree of preferentiality was estimated as $\beta = 0.5$, for model (2). Three different grid side sizes were chosen: to be $\frac{h}{12} = 12.0$ or $\frac{h}{24} = 6.0$, where h is the maximum distance between all the n points; and a size defined by a point density criterion of value 2.4. The MLC test resulted in the non rejection of the preferential sampling assumption for the case of $\frac{h}{24}$ and rejection on the other cases.

The test for preferential sampling presented in this work is very simple to implement and the simulation results obtained are quite promising, even for quite irregular shaped regions. More simulation studies are run-

ning and also the theoretic properties of the test are being studied. There are still some issues to address, as the question of grid size, how to handle the borders of the regions of interest, and evaluation of test performance in the presence of isotropy and non stationarity.

Acknowledgments

This work is funded by national funds through the FCT - Fundação para a Ciência e Tecnologia, I.P., under the scope of the projects PREFERENTIAL, PTDC/MAT-STA/28243/2017; UIDB/00297/2020 and UIDP/00297/2020 (Center for Mathematics and Applications); and UIDB/00006/2020 (CEAUL).

References

[1] André, L.M., Figueiredo, I., Carvalho, M.L., Simões, P. and Natário, I. (2020). Spatial modelling of black scabbardfish fishery off the portuguese coast. In *Int. Conf. Comp. Science and Appl.*, 332–344. Springer.

[2] Diggle P.J., Menezes R., Su T.L. (2010). Geostatistical Inference under Preferential Sampling (with discussion). *J. R. Stat. Soc. Ser. C*, **59**: 191–232 (RSS read paper).

[3] Diggle P.J., Ribeiro Jr. P. (2007) *Model-based Geostatistics*. Springer.

[4] Krainski E.T., Gómez-Rubio V., Bakka H., Lenzi A., Castro-Camilo D., Simpson D., Lindgren F., Rue H. (2018). Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA. CRC.

[5] Pennino, M.G., Paradinas, I., Illian, J.B., Muñoz, F., Bellido, J.M., López-Quílez, A., Conesa, D. (2019). Accounting for preferential sampling in species distribution models. Ecological Evolution, **9**: 653–663.

[6] Watson, J. (2021). A perceptron for detecting the preferential sampling of locations and times chosen to monitor a spatio-temporal process. Spatial Statistics, **43**:100500.

[7] Watson, J., Zidek, J.V., Shaddick, G. (2019). A general theory for preferential sampling in environmental networks. Ann. Appl. Stat., **13**:2662–2700.

Author Index