A circular hidden Markov random field for the spatial segmentation of fire occurrence

J. Ameijeiras-Alonso, F. Lagona, M. Ranalli and R. M. Crujeiras

HOW TO CITE

Ameijeiras-Alonso, J., Lagona, F., Ranalli, M., Crujeiras, R. M. (2019). A circular hidden Markov random field for the spatial segmentation of fire occurrence. *Environmetrics* 30(2), e2501.

FUNDING

Research of Jose Ameijeiras-Alonso and Rosa M. Crujeiras has been funded by Project MTM2016-76969-P (Spanish State Research Agency, AEI), co-funded by the European Regional Development Fund (ERDF), IAP network from Belgian Science Policy. Part of the research was carried out by Jose Ameijeiras-Alonso during his visit to University of Roma Tre, supported by Grants BES-2014-071006 and EEBB-I-17-12716 from the Spanish Ministry of Economy, Industry and Competitiveness. Francesco Lagona is supported by the 2015 PRIN supported project “Environmental processes and human activities: capturing their interactions via statistical methods”, funded by the Italian Ministry of Education, University and Scientific Research.

A circular hidden Markov random field for the spatial segmentation of fire occurrence

Jose Ameijeiras–Alonso[1] Francesco Lagona[2] Monia Ranalli[2]
Rosa M. Crujeiras[1]

[1]Departamento de Estatística, Análise Matemática e Optimización
Universidade de Santiago de Compostela (Spain)

[2]Department of Political Sciences
University of Roma Tre (Italy)

Abstract

Motivated by studies of wildfire seasonality, we propose a non-homogeneous hidden Markov random field to model the spatial distribution of geo-referenced fire occurrences during the year, by representing occurrence times as circular data. The model is based on a mixture of Kato-Jones circular densities, whose parameters vary across space according to a latent non-homogeneous Potts model, modulated by geo-referenced covariates. It allows to segment fire occurrences according to a finite number of latent classes that represent the conditional distributions of the data under specific periods of the year, simultaneously accounting for unobserved heterogeneity and spatial autocorrelation. Further, it parsimoniously accommodates specific features of wildfire occurrence data such as multimodality, skewness and kurtosis. Due to the numerical intractability of the likelihood function, estimation of the parameters is based on composite likelihood (CL) methods. It reduces to a computationally efficient Expectation-Maximization algorithm that iteratively alternates the maximization of a weighted CL function with weights updating. The proposal is illustrated in a study of wildfire occurrences in the Iberian peninsula during a decade.

Keywords: spatial circular data, composite likelihood, fires, Kato-Jones density, land use, Markov random field.

1 Introduction

Wildfires represent a major environmental and economic issue in Southern Europe. In Spain and Portugal, for example, some areas have been recently devastated by wildfires. In mid June 2017, more than 45.000 hectares were devastated in the district of Leiria (central Portugal), with around 150 active fires in a week. Some years before, in 2006, the damaged region was Galicia (NW Spain), with almost two thousand active fires during the first fortnight of August which destroyed an area larger than the one affected by wildfires in the five previous years.

Studies on wildfires typically focus on specific aspects of wildfires, such as fire frequency, fire intensity and fire extension. This paper focuses on the spatial distribution of wildfire seasonality. As a part of fire regimes characterization, the spatial analysis of fire seasonality is crucial for understanding fires behavior at global and local spatial scales (Benali *et al.*, 2017). In addition, knowledge of seasonal patterns of fires across space is useful for designing appropriate precautionary and intervention measures against wildfires.

Basic information available for our study includes occurrence of geo-referenced wildfire events (day of the year) across a decade (2002-2012). On a yearly scale, event times can be placed on a unit circle corresponding to a period of 365 days which, without loss of generality, can be rescaled to the support $[0, 2\pi)$. The idea of studying seasonality by viewing event times as circular data is not new and it has been, for example, recently used by Shirota and Gelfand (2017) in a study of crime events. By taking this approach, geo-referenced event times can be represented as spatial circular data.

While methods for handling independent circular data are well established (Ley and Verdebout, 2017), the analysis of spatially dependent circular data is a new emerging area of research. Some proposals in this context rely on the extension of geostatistical models to the circular setting. Spatial wrapped-Gaussian processes (Jona Lasinio *et al.*, 2012) and spatial projected-Gaussian processes (Wang and Gelfand, 2014) have been proposed for modelling sea motion. Consistently with the geo-statistical paradigm, these models consider spatial circular processes that vary continuously over space. Alternative approaches are based on circular processes that vary discretely over space. Some proposals in this context make use of adaptations of spatial auto-regressive processes to the circular setting. For example, Modlin *et al.* (2012) introduce a circular autoregressive model, based on the wrapped normal distribution for studying hurricane wind directions. Lagona (2016) introduces a circular autoregressive model, based on the multivariate von Mises distribution, for characterizing sea wave directions. Other proposals consider hidden Markov random fields for spatial circular data (Lagona and Picone, 2016; Ranalli *et al.*, 2018). In this setting, spatial circular data are conditionally independent given a latent Markov random field (MRF), i.e. a multinomial process in discrete space which fulfils a spatial Markov property (Gaetan and Guyon, 2010) and which segments the study area according to a finite number of latent classes. This approach is useful when the interest is focused on segmenting the area under study according to a small number of classes, each one associated with a specific distribution of the data.

Hidden MRFs are popular models in spatial statistics, since the seminal paper by Besag (1975). They can be seen as a spatial extension of the hidden Markov models (HMMs) that are exploited in time series analysis. Circular extensions of HMMs have been widely used for the analysis of time series with circular components (Holzmann *et al.*, 2006; Lagona *et al.*, 2015; Mastrantonio *et al.*, 2015; Maruotti *et al.*, 2016). On the contrary, the widespread use of circular hidden MRF has been limited by the intractability of the likelihood function of these models. Following Alfó *et al.* (2008), Lagona and Picone (2016) adapt a mean-field approximation for Gaussian hidden MRFs to the circular setting and develop a computationally intensive Expectation-Maximization (EM) algorithm. Unfortunately, the method is numerically unstable and little is known about the distributional properties of the estimators. More recently, Ranalli *et al.* (2018) suggest composite-likelihood methods to estimate a (homogeneous) circular hidden MRF, showing that this method provides a good solution to balance statistical and computational efficiency, through an extensive simulation study.

In studies of wildfire seasonality, circular hidden Markov fields provide a natural approach to segment an area of interest according to regions that are associated with specific seasonal patterns of fire events. In this paper, specifically, we assume that the distribution of the fires occurrence recorded as days on a yearly base (viewed as circular data) is well approximated by a mixture of Kato-Jones densities, whose parameters vary across space according to a non-homogeneous Potts model. The Kato-Jones density is a four-parameter unimodal density which flexibly accommodates skewness and kurtosis on the circle (Kato and Jones, 2015). The proposed non-homogeneous Potts model is a MRF, whose parameters depend on geo-referenced covariates. The model is estimated by extending the proposal by Ranalli *et al.* (2018) to the case of a non-homogeneous hidden MRF that is modulated by geo-referenced covariates.

The rest of the paper is organized as follows. Section 2 briefly describes the wildfire data that motivated this study. Section 3 presents the structure of the proposed non-homogeneous hidden MRF and Section 4 illustrates the CL methods that we suggest for estimation. Section 5 is devoted to the results that have been obtained by the proposed methods on the real data. Section 6 finally summarizes relevant points of discussion.

2 Wildfire occurrences in the Iberian peninsula

The identification of fire peaks in certain spatial areas should serve to warn the authorities in order to organize appropriate interventions or campaigns, especially when a critical period is highly probable (a critical day for fire risk is identified by the 30-30-30 conditions: more than 30^o, winds of more than 30 km/h and less than 30% of humidity).

In the Iberian peninsula, most fires occur during summer with a peak of activity in August-September. In terms of climatological and weather conditions, summer is the most suitable season for fires in latitudes over the Tropic of Cancer, because dry weather conditions are predominant. Authorities are certainly concerned with wildfire problems during summer season and rangers and firefighters groups are reinforced in this period. Nevertheless, wildfires do not only occur in summer. A recent example was October 2017, when northern Portugal counted 600 fires in five days and in Galicia more than one hundred fires occurred in a day. In this case, authorities were not prepared for fighting wildfires, since October is not considered as a *peak season*.

In addition to summer weather conditions, land management practices also influence fire seasonality, showing in this case preferential timings. Although some human fires are produced intentionally or unintentionally during dry months, there are other activities that cause fires in periods that are outside the principal peak of fires. These activities include, for example, agriculture burning for preparing fields (for harvest work) or for clearing the crop residues (after harvesting or in order to avoid future burnings during the climatological season of fires). The distribution of fire activity peaks during the year and across the study area helps to explain where and how human activity changed fire seasonality using fires as a land management tool (Ameijeiras-Alonso *et al.*, 2018).

For studying the distribution of the times of wildfires occurrences, we collected data of the fires in the Iberian Peninsula from 10 July 2002 to 9 July 2012, detected by the *MODerate resolution Imaging Spectroradiometer* (MODIS), launched into Earth orbit by NASA on board of the Terra (*EOS AM*) and the Aqua (*EOS PM*) satellites. MODIS identifies locations where fires are actively burning at the time of satellite overpass, using for this purpose an algorithm that summarizes a number of measures such as brightness, temperatures, cloud and water masks and the sun glint (Giglio *et al.*, 2003; Oom and Pereira, 2013).

During the study period, a total of 63127 fires were detected by MODIS in the inland territory of Spain and Portugal. By wrapping the day of occurrence of these fires around a

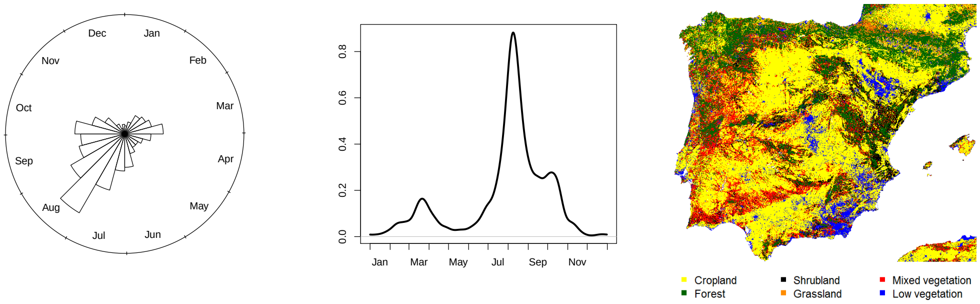


Figure 1: Distribution of fire occurrence times in the Iberian peninsula during the period 2002-2012. Left: rose diagram of the raw data. Middle: circular kernel density estimate. Right: land use of the Iberian peninsula. The ticks in the left and the middle pictures represent the beginning or the end of the month that is indicated between them.

circle of circumference 365 days, we obtain the rose diagram in Figure 1 (left side), where the area of each circular sector is proportional to the class frequency. The multimodality of these data is better visualized by the von Mises-kernel non-parametric density estimate of the data (Taylor, 2008), displayed in the middle of Figure 1. For simplicity, it is displayed as a function over the real line, although its support is the circle and so it should be wrapped around by joining both *ends* of the curve. As expected, most of the fires are concentrated around the first fortnight of August (climatological peak of fires). However, two additional peaks appear during the second fortnight of March and at the beginning of October. This alteration of fire seasonality is well documented in the forestry literature (see, e.g., Korontzi *et al.*, 2006; Le Page *et al.*, 2010; Magi *et al.*, 2012; Benali *et al.*, 2017) and the regularity of these multimodal patterns was recently studied by Ameijeiras-Alonso (2017) in a worldwide setting.

Ignoring the spatial dimension of the data, Figure 1 suggests that the distribution of fire occurrence over the year is a mixture of either two or three components that pertain to specific seasonal patterns of fire occurrences. We are interested in modeling the spatial distribution of these components through a hidden MRF. The model accommodate the spatial aspect of the data by a neighborhood structure between wildfire sites. Such a spatial neighborhood structure can be specified in different ways (Bivand *et al.*, 2008, chapter 9). A feasible strategy relies on defining the neighbors of each site i in terms of the Euclidean distance. Hence, the Iberian Peninsula was first partitioned by a square grid of resolution $300m^2$ and each fire event i was associated to a grid cell. Let $\mathbb{A} = \{a_1, a_2, \ldots\}$ be the set of centroids that were associated with at least one fire event. For each $a \in \mathbb{A}$, let $d_a^\star$ be the shortest Euclidean distance between a and any other centroid in $\mathbb{A}$. Then, for each site i that is associated with centroid a, we refer to $N(i) = \{j \in S : d(i,j) \leq d_a^\star \text{ and } j \neq i\}$ as the neighborhood of i.

Since land use can alter the timing of fire occurrence, this information was included in the analysis. We obtained land use at a $300m^2$ resolution from the European Space Agency Climate Change Initiative project (*Land Cover version 2.0.7*; available at `http://www.esa-landcover-cci.org`), which describes the physical material at the surface of the earth, including various types of vegetation, bare rock and soil, water, snow and ice, and artificial surfaces. In our study, land cover before the fire event is included as a factor with six levels that are respectively associated with *cropland* (rainfed; irrigated or post-flooding), *forest* (tree cover; broadleaved, needleleaved or mixed leaf type; evergreen or deciduous; closed or open), *shrubland*, *grassland* (herbaceous cover, grassland), *mixed vegetation*, *low vegetation* (sparse vegetation; tree cover, flooded; urban areas; bare areas; water bodies; permanent snow and ice). In the Iberian peninsula a total of 16145 fires were produced in cropland , 25963 in forest, 9868 in shrubland, 1621 in grassland, 3096 in mixed vegetation and 6434 in low vegetation areas. The summary of 10-years land cover is shown, for the Iberian Peninsula, in Figure 1 (right).

3 A circular hidden Markov random field

The data that motivated this work are in the form of a spatial series of circular observations, say $\mathbf{y} = (y_i, i = 1, \dots, n)$, $y_i \in [0, 2\pi)$. Each observation is associated with the ith row vector $\mathbf{x}_i^\mathsf{T}$ of a design matrix $\mathbf{X}$. In our case study, y_i indicate the day of the year of the ith fire event, while $\mathbf{X}$ is a 6-column coding matrix that includes land cover before the fire event as a factor with 6 levels. In this section we describe a circular hidden MRF model that spatially segments these data according to K latent classes. The model can be seen as a mixture of K unimodal circular densities, whose parameters vary across space according to a latent multinomial process. We specify the mixture components as Kato-Jones densities (Kato and Jones, 2015). The latent multinomial process is instead specified by a non-homogeneous Potts model (Strauss, 1977), whose parameters are modulated by the design matrix $\mathbf{X}$. In Section 3.1 we briefly describe the Kato-Jones density and Section 3.2 is devoted to the Potts model. Finally, section 3.3 describes the proposed HMRF model, by integrating the Kato-Jones distribution with the latent Potts model.

3.1 The Kato-Jones density

The choice of the component density in a mixture model should be always driven by the purpose of the analysis. If the purpose is model-based classification, then the component should be unimodal and appropriate in light of the data (McNicholas, 2016, chapter 9, p. 157). A recent proposal that suits our needs is the four-parameter Kato-Jones density, namely

$$f(y; \mu, \gamma, \nu, \eta) = \frac{1}{2\pi}\left(1 + 2\gamma^2 \frac{\gamma\cos(\theta - \mu) - \nu}{\gamma^2 + \nu^2 + \eta^2 - 2\gamma(\nu\cos(\theta-\mu) + \eta\sin(\theta-\mu))}\right), \quad y \in [0, 2\pi), \tag{1}$$

where $(\nu, \eta) \neq (\gamma, 0)$ and

$$\begin{aligned} 0 &\le \mu < 2\pi \\ 0 &\le \gamma < 1 \\ (\nu - \gamma^2)^2 + \eta^2 &\le \gamma^2(1-\gamma)^2. \end{aligned} \tag{2}$$

The properties of this density are extensively described by Kato and Jones (2015). We summarize here the most practical advantages of this density. First, the meaning of the parameters is intuitively appealing. Precisely, μ indicates the circular mean, γ is the concentration parameter (if $\gamma = 0$ then the density reduces to a uniform density on the circle), ν measures kurtosis (the greater the value of ν the sharper the peakedness of the density) and, finally, η indicates skewness (the density is symmetric if and only if $\eta = 0$). Second, samples from the Kato-Jones distribution can be easily generated by drawing samples from the wrapped Cauchy distribution and then using the acceptance/rejection algorithm proposed by Kato and Jones (2015). Third, maximum likelihood is numerically tractable by using a maximization algorithm that allows for constraints such as that provided by the `nlminb` function in `R`. This function requires that the range of each parameter should not depend on other parameters, and a suitable re-parametrization is necessary, such as that suggested by Kato and Jones (2015).

3.2 A non-homogeneous Potts model

The Potts model is a multinomial process in discrete space (lattice) with K classes. Given a lattice that divides an area of interest according to n observation sites $i = 1, \dots, n$, a sample that is drawn from a spatial multinomial process is a segmentation of this area, obtained by associating each site with a segmentation label $k = 1, \dots, K$. Formally, each observation site i is associated with a multinomial random variable $\mathbf{U}_i = (U_{i1}, \dots, U_{iK})$ with one trial and K classes, where U_{ik} is a Bernoulli random variable that is equal to 1 if i is labelled by k and 0 otherwise. A specific segmentation of the area can be accordingly represented as a sample drawn from the multinomial process $\mathbf{U} = (\mathbf{U}_1, \dots \mathbf{U}_n)$. The Potts model is a spatial multinomial process which accounts for a neighborhood structure $N(i), i = 1, \dots, n$ among the observation sites, which associates each site with a set $N(i)$ of neighbors. Section 2 described the neighborhood structure that was exploited for our case study. By taking the class K as a reference, each segmentation $\mathbf{u}$ is associated with K sufficient statistics: the number of neighboring sites which share the same class $k \neq K$

$$n(\mathbf{u}) = \sum_{i=1}^{n} \sum_{j>i:j\in N(i)} \sum_{k=1}^{K-1} u_{ik} u_{jk},$$

and $K - 1$ sufficient statistics

$$n_k(\mathbf{u}) = \sum_{i=1}^{n} u_{ik} \quad k = 1, \dots, K-1$$

which indicate the number of neighboring sites that are associated with latent class k. Under the Potts model, the probability of a specific segmentation $\mathbf{u}$ is known up to K parameters $\alpha_1 \dots \alpha_{K-1}, \rho$ and it is given by

$$p(\mathbf{u}; \alpha, \rho) = \frac{\exp\left(\sum_{k=1}^{K-1} n_k(\mathbf{u})\alpha_k + n(\mathbf{u})\rho\right)}{W(\alpha, \rho)}, \tag{3}$$

where $W(\alpha, \rho)$ is the normalizing constant. The parameter ρ is an autocorrelation parameter: if it is positive (negative) then it penalizes segmentations with a few concordant (discordant) neighbors. In the image analysis literature, it often referred to a regularization parameter, given that large values of ρ are associated with segmentations where areas with the same label are geometrically regular. Each parameter α_k penalizes segmentations with a few sites that belong to class k. When $\rho = 0$, then (3) reduces to a multinomial distribution where the parameters α are class-specific log-odds:

$$\alpha_k = \log \frac{P(u_{ik} = 1)}{P(u_{iK} = 1)}.$$

Model (3) is an homogeneous Potts model, because the parameters α_k do not vary across the observation sites. We extend this setting by assuming that these parameters depend on site-specific covariates, through a multinomial logistic model:

$$\alpha_{ik} = \mathbf{x}_i^{\mathsf{T}} \beta_k \quad k = 1 \dots K-1, \tag{4}$$

where the regression coefficients β_k indicate the influence of the available covariates on the frequency of label k in the segmentation. As a result, we obtain the non-homogeneous Potts model

$$p(\mathbf{u}; \beta, \rho) = \frac{\exp\left(\sum_{i=1}^{n} \mathbf{x}_i^{\mathsf{T}} \beta + n(\mathbf{u})\rho\right)}{W(\beta, \rho)}, \tag{5}$$

where $W(\beta, \rho)$ is the normalizing constant. For each site i and each label k, let

$$n_k(\mathbf{u}_{\bar{N}(i)}) = u_{ik} \sum_{j \in N(i)} u_{jk}$$

be the number of sites in the neighborhood of i that are labelled by k, where $\bar{N}(i)$ indicates the neighborhood of i, completed by i. Under model (3), the conditional distribution of each site depends only on the labels taken by the neighboring sites, namely

$$p(u_{ik} = 1 \mid \mathbf{u}_1, \dots \mathbf{u}_{i-1}, \mathbf{u}_{i+1}, \dots \mathbf{u}_n) = \frac{\exp\left(\alpha_k + \rho n_k(\mathbf{u}_{\bar{N}(i)})\right)}{1 + \sum_{k=1}^{K-1} \exp\left(\alpha_k + \rho n_k(\mathbf{u}_{\bar{N}(i)})\right)}, \tag{6}$$

Accordingly, the Potts model is a Markov random field with respect to the neighborhood structure. Under the non-homogeneous random field (5), these conditional probabilities reduce to

$$p(u_{ik} = 1 \mid \mathbf{u}_1, \dots \mathbf{u}_{i-1}, \mathbf{u}_{i+1}, \dots \mathbf{u}_n) = \frac{\exp\left(\mathbf{x}^{\mathsf{T}} \beta_k + \rho n_k(\mathbf{u}_{\bar{N}(i)})\right)}{1 + \sum_{k=1}^{K-1} \exp\left(\mathbf{x}^{\mathsf{T}} \beta_k + \rho n_k(\mathbf{u}_{\bar{N}(i)})\right)}, \tag{7}$$

and the autocorrelation coefficient ρ can be viewed as an auto-regression coefficient that is associated with the the spatially-lagged outcome $n_k(\mathbf{u}_{\bar{N}(i)})$.

3.3 A circular hidden Markov random field

The proposed hidden MRF is specified by assuming that the observed data are conditionally independent, given a segmentation generated by the non-homogeneous Potts model (5). Precisely, we assume that the conditional distribution of the observed process, given the latent segmentation $\mathbf{u}$, takes the form of a product density, say

$$f(\mathbf{y} \mid \mathbf{u}; \theta) = \prod_{i=1}^{n} \prod_{k=1}^{K} f(y_i; \theta_k)^{u_{ik}}, \tag{8}$$

where the vector $\theta = (\theta_1, \dots, \theta_K)$ includes K label-specific parameters and $f(y; \theta_k)$, $k = 1, \dots, K$, are K Kato-Jones circular densities (1). The joint density of the observed data and the unobserved class memberships is therefore given by

$$f(\mathbf{y}, \mathbf{u}; \theta, \beta, \rho) = f(\mathbf{y} \mid \mathbf{u}; \theta) p(\mathbf{u}; \rho, \beta). \tag{9}$$

By integrating this distribution with respect to the segmentation $\mathbf{u}$, we obtain the likelihood function of the unknown parameters

$$L(\boldsymbol{\theta}, \rho, \beta) = \sum_{\mathbf{u}} f(\mathbf{y} \mid \mathbf{u}; \theta) p(\mathbf{u}; \rho, \beta). \tag{10}$$

When $\rho = 0$, then the model reduces to a latent class model with concomitant covariates. Otherwise, when $\rho \neq 0$, the model is a mixture of circular densities, whose parameters vary across space according to a Markovian process.

4 Parameter estimation through composite-likelihood methods

4.1 An EM algorithm

Direct maximization of the likelihood function (10) is unfeasible. As a result, we propose to estimate the parameters by maximizing a surrogate function, namely a composite likelihood (CL) function (Lindsay, 1988). Our proposal relies on covering the set $S = \{1 \dots n\}$ of the observation sites by all the pairs S_2 of neighboring sites. For each subset S_2, we define

$$L_{S_2}(\theta, \rho, \beta) = \sum_{\mathbf{u}_{S_2}} p(\mathbf{u}_{S_2}; \rho, \beta) \prod_{i \in S_2} \prod_{k=1}^{K} f(y_i; \theta_k)^{u_{ik}}$$

as the contribution of the data in S_2 to the CL function, where $\mathbf{u}_{S_2} = \{u_i : i \in S_2\}$ and

$$p(\mathbf{u}_{S_2}; \rho, \beta) = \frac{\exp\left(\sum_{i \in S_2} \mathbf{x}_i^{\mathsf{T}} \beta + n(\mathbf{u}_{\mathbf{S_2}}) \rho\right)}{W_2(\beta, \rho)},$$

with $W_2(\beta, \rho)$ being the normalizing constant, is a two-site non-homogeneous Potts model.

We propose to estimate the parameters by maximizing the following composite log-likelihood function

$$c\ell(\theta, \rho, \beta) = \sum_{S_2} \log L_{S_2}(\theta, \rho, \beta). \tag{11}$$

To this aim, we use an EM algorithm that iteratively generates a sequence of parameter values by alternating an E step and an M step, until convergence. During the E step, it computes the expected value of the complete-data composite log-likelihood with respect to the predictive distribution of the segmentation. This step reduces to the computation of the following expected value for each pair of sites

$$\mathbb{E} \log L_{S_2}(\theta, \rho, \beta) = \sum_{i \in S_2} \sum_{k=1}^{K} \hat{u}_{ik} \log f(y_i; \theta_k) + \sum_{\mathbf{u}_{S_2}} \hat{\mathbf{u}}_{S_2} \log p(\mathbf{u}_{S_2}; \rho, \beta),$$

where the $K \times K$ predictive probabilities $\hat{\mathbf{u}}_{S_2}$ are given by

$$\hat{\mathbf{u}}_{S_2} = p(\mathbf{u}_{S_2} \mid \mathbf{y}_{S_2}, \hat{\rho}, \hat{\theta}, \hat{\beta}) = \frac{p(\mathbf{u}_{S_2}; \hat{\rho}, \hat{\beta}) f(\mathbf{y}_{S_2}; \hat{\theta})}{\sum_{\mathbf{u}_{S_2}} p(\mathbf{u}_{S_2}; \hat{\rho}, \hat{\beta}) f(\mathbf{y}_{S_2}; \hat{\theta})}, \tag{12}$$

whereas $\hat{\rho}$, $\hat{\theta}$ and $\hat{\beta}$ are the parameter values that were available from the previous step of the algorithm. The normalizing constant of these probabilities is numerically tractable, as it involves a summation over two sites. Suitable marginalization of (12) provides the univariate probabilities $\hat{\mathbf{u}}_i = p(\mathbf{u}_i \mid \mathbf{y}_{S_2}, \hat{\rho}, \hat{\boldsymbol{\theta}}, \hat{\beta})$, with $i = 1, \ldots, n$.

During the M step, the algorithm maximizes the expected complete-data composite log-likelihood with respect to the unknown parameters. Because this function is the sum of two components that depend on different sets of parameters, the M-step reduces to the separate maximization of two functions, namely

$$\begin{aligned} Q(\theta) &= \sum_{S_2} \sum_{i \in S_2} \sum_{k=1}^{K} \hat{u}_{ik} \log f(y_i; \theta_k) & (13) \\ Q(\rho, \beta) &= \sum_{S_2} \sum_{\mathbf{u}_{S_2}} \hat{\mathbf{u}}_{S_2} \log p(\mathbf{u}_{S_2}; \rho, \beta). & (14) \end{aligned}$$

Maximization of both $Q(\theta)$ and $Q(\rho, \beta)$ can be carried out by a standard optimization routine that allows for parameter constrains, such as that provided by the `nlminb` function in `R`.

4.2 Computational details

The proposed composite likelihood has been defined by covering the study area with pairs of neighboring sites. Covering the area by larger subset might have been an option. However, the numerical tractability of EM algorithm dramatically decreases with the cardinality of the largest subset of the cover. A cover that includes subsets with two elements is therefore a natural strategy. When the cover includes all the subsets of two elements, equation (11) reduces to the pairwise likelihood function (Varin *et al.*, 2011). In a spatial setting, a pairwise likelihood can be further simplified by discarding all the pairs $\{i, j\}$ that do not include neighboring sites. Ranalli *et al.* (2018) provide an extensive simulation study that shows that this choice provides a computationally efficient EM algorithm, without a relevant loss in statistical efficiency.

It is well known that the EM algorithm suffers from two drawbacks: it is sensitive to the choice of starting points and it may converge to local maxima. These two aspects are strictly linked to each other. To avoid local maxima we follow a short-runs strategy, by running the EM algorithm from 50 random initializations, and stopping the algorithm without waiting for full convergence, i.e. when the relative increase in two consecutive composite log-likelihoods is less than 10^{-2}. The best solution is taken as starting point to run the EM algorithm until full convergence, that is when the difference in two consecutive composite log-likelihoods is less than 10^{-5}.

Standard errors could in principle be obtained by numerically approximating the observed Godambe matrix (Godambe, 1960), which is however known to present numerical instability. This computation requires both the numerical approximation of variability and sensitivity matrices, and the inversion of the variability matrix (i.e. the covariance of the CL score), that it is usually a large-size matrix. A feasible alternative can be found in parametric bootstrap methods, to obtain quantiles of the distribution of the estimates. In this paper, we re-fitted the model to $R = 500$ bootstrap samples, which were simulated from the estimated model parameters. We then computed the 2.5% and the 97.5% quantiles of the empirical distribution

Table 1: Composite integrated classification likelihood values, estimated by fitting a hidden MRF model with components $K = 2, \ldots 6$.

	Number of components				
	2	3	4	5	6
C-ICL	2128333	**1974793**	2280843	2033279	2346882

of each bootstrap estimate. Simulation of the circular hidden MRF is straightforward, by taking advantage of standard simulation routines available for the Potts model and the Kato-Jones distribution. Specifically, we exploited the Gibbs sampler algorithm (Feng *et al.*, 2012) that is available in the `R` package `PottsUtils` to simulate a configuration of segmentation labels. Given a configuration of segmentation labels, a circular observation y_i is drawn at each lattice site i, according to the appropriate Kato-Jones distribution, evaluated at $\theta = \hat{\theta}_k$, where k is the segmentation label that has been generated by the Potts model at location i. Kato and Jones (2015) suggest a simple acceptance/rejection algorithm to simulate from a their distribution and we follow their proposal.

5 Spatial segmentation of fire occurrence times

A collection of hidden MRF models have been estimated in order to obtain the spatial segmentation of the wildfire data described in Section 2, by varying the number of states K from 2 to 6. The number K of latent classes was chosen by selecting the model minimizing the composite integrated classification likelihood (C-ICL; Ranalli and Rocci, 2016), which extends the integrated classification likelihood to the CL framework. According to Table 1, a model with 3 components attains the minimum C-ICL value.

Table 2 (top panel) includes both the estimates and the bootstrap confidence intervals of the parameters of the 3 class-specific Kato-Jones densities that have been obtained by estimating a hidden MRF model with $K = 3$ components. These values should be interpreted by recalling the specific support of each parameter, displayed also in Table 2. In particular, we remark that the supports of the kurtosis and the skewness parameters depend on the value taken by the concentration parameter. By rescaling the mean parameters to the scale of a circle of circumference 365 days, the values $\mu_1 = 1.335$, $\mu_2 = 3.798$ and $\mu_3 = 4.800$ are respectively associated with the dates March 18th, August 8th and October 6th. The three latent classes can be therefore interpreted as three different periods of the year that respectively occur in Spring, Summer and Autumn, in the corresponding latitudes. The data are rather concentrated within these classes, as shown by the large values attained by the concentration parameters ($\gamma_1 = 0.861$, $\gamma_2 = 0.858$ and $\gamma_3 = 0.816$, in a unit scale). The kurtosis parameter indicates that the summer latent class is more peaked than the other two classes. This is an expected result, as summer is the season associated to most wildfires ($\nu_1 = 0.699$, $\nu_2 = 0.736$ and $\nu_3 = 0.666$). More interestingly, while the Spring component is left skewed, the shape of the other two components is quite symmetric and well approximated by a wrapped Cauchy density (when $\eta = 0$ and $\nu = \gamma^2$, the Kato-Jones distribution reduces to the two-parameter wrapped Cauchy density).

Table 2 (bottom panel) shows the estimated influence of land cover on the marginal distribution of the latent classes, by taking the class 3 as reference. As expected, latent class 2 is associated with positive estimates, indicating that the proportion of fires during summer

Table 2: Parameter estimates of a hidden MRF model with $K = 3$ states and relating bootstrap 95% confidence intervals (within brackets) and parameter supports (italics).

	Latent classes					
	1 (Spring)		2 (Summer)		3 (Autumn)	
Parameter	Est.	Conf. Int.	Est.	Conf. Int.	Est.	Conf. Int.
μ (mean)	1.335	(1.320,3.781)	3.798	(3.795,3.837)	4.800	(4.540,4.822)
		[0,2π)		*[0,2π)*		*[0,2π)*
γ (concentration)	0.861	(0.741,0.873)	0.858	(0.827,0.869)	0.816	(0.374,0.818)
		[0,1)		*[0,1)*		*[0,1)*
ν (kurtosis)	0.699	(0.628,0.762)	0.736	(0.718,0.755)	0.666	(0.297,0.672)
		[0.622,0.861]		*[0.614,0.858]*		*[0.516,0.816]*
η (skewness)	0.047	(-0.003,0.060)	0.000	(-0.010,0.016)	0.000	(-0.018,0.105)
		[-0.120,0.120]		*[-0.122,0.122]*		*[-0.150,0.150]*

	Latent classes (reference = latent class 3)			
Land cover	1 (Spring)		2 (Summer)	
Cropland	-0.636	(-0.777,0.080)	0.523	(-0.249,0.663)
Forest	-0.366	(-0.469,0.553)	0.956	(0.297,1.259)
Shrubland	0.066	(-0.075,0.740)	1.331	(0.433,1.604)
Grassland	0.390	(0.388,1.159)	1.474	(0.559,2.058)
Mixed vegetation	-0.377	(-0.453,0.336)	0.874	(-0.080,1.036)
Low vegetation	-0.713	(-0.789,0.883)	1.142	(0.498,1.620)
ρ (spatial)	0.952	(0.751,1.005)		

Table 3: Conditional latent class distribution of wildfire events given land use, estimated by a 3-state hidden Markov random field. Bold: estimated marginal latent class distribution and observed marginal land cover distribution.

		Latent classes		
Land cover		1	2	3
		0.113	**0.700**	**0.187**
Cropland	**0.256**	0.122	0.584	0.294
Forest	**0.411**	0.110	0.728	0.162
Shrubland	**0.156**	0.126	0.748	0.126
Grassland	**0.026**	0.151	0.687	0.162
Mixed vegetation	**0.049**	0.136	0.684	0.180
Low vegetation	**0.102**	0.060	0.813	0.127

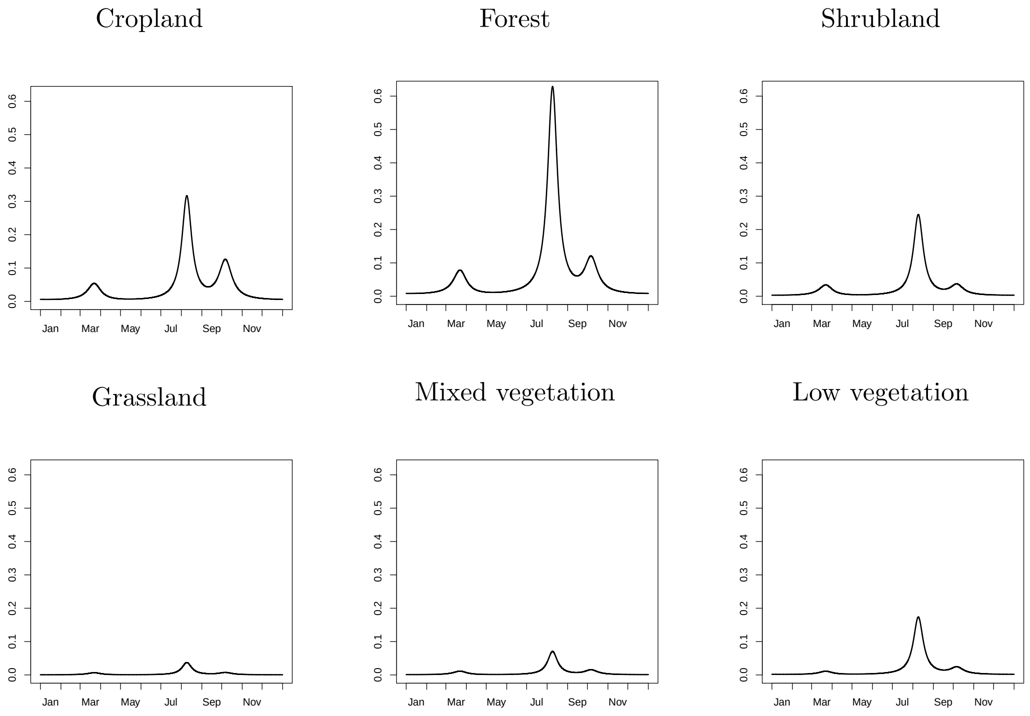


Figure 2: Estimated mixture of the three Kato-Jones densities, where each component is weighted according to the conditional latent class and land cover.

is larger than during autumn, regardless of land cover. More interestingly, the proportion of fires in spring is larger than in autumn only when fires occur in shrublands and grasslands. In all the other cases, the proportion of fires in spring is smaller than in autumn.

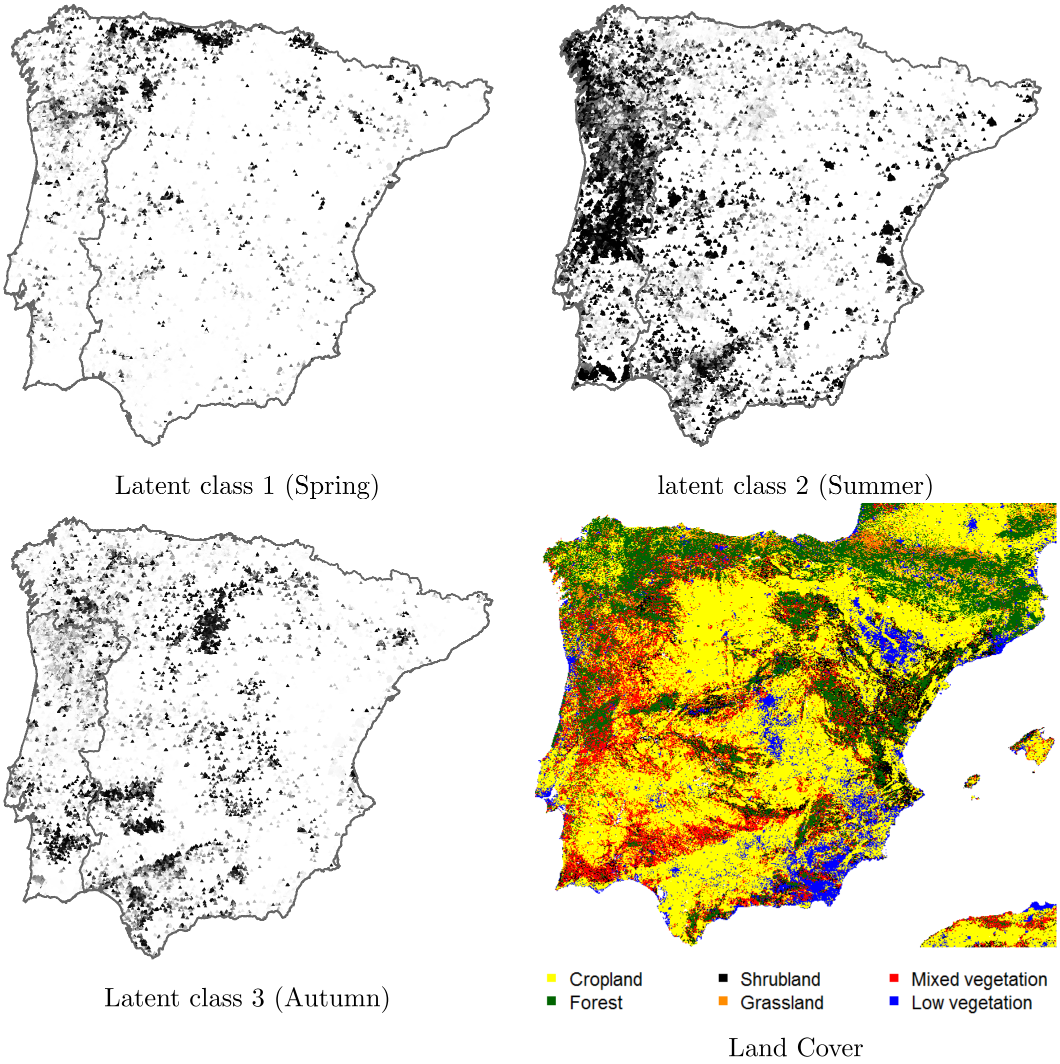
The estimated model segments the Iberian Peninsula according to three latent classes that are associated with three well distinct seasonal patterns of fire occurrences. Each fire event can be allocated to a given class by comparing the posterior probabilities $\hat{\mathbf{u}}_i$ and associating each event with the most likely class. By counting the events that the model allocated to each class across land cover, we obtain the estimated marginal distribution of the latent classes (Table 3, first bold row) and the conditional distribution of latent classes within each land type. These probabilities were integrated with the parameters of the top panel of table 2 to compute the distribution of wildfire occurrence days within each land cover, shown in Figure 2. This figure decomposes the marginal distribution of the data (figure 1) into land-specific components, showing that wildfire seasonality varies across lands of different types.

Figure 3 displays the final segmentation of the wildfire occurrence data across the study area, obtained by associating each geo-referenced event with the posterior probability of being in each class. In particular, dark colors indicate a higher posterior probability of latent class membership. It can be clearly seen from top and bottom-left panels that the wildfire distribution across space is quite different for the three latent classes. Although we do not aim at drawing thorough conclusions about fire dynamics, this figure depicts some relevant association between fire occurrence and land use. Wildfires occurrence in Portugal can be divided in three regions: the north-east district of Bragança, affected by Spring fires; the south-central region of Alentejo, which suffers from Autumn fires; and the rest of the country, with a high incidence of summer fires. Actually, these summer fires are spatially concentrated in the western part of the peninsula, both in Portugal and Galicia. With a closer look at the Spanish territory, the Autumn component seems to be related with regions where winter cereal crops are frequent. This is the case of the central part of Castile and León and the south-central and south-eastern part of the Iberian Peninsula (south-central region of Alentejo, south of Extremadura, around Sevile and Córdoba in Andalusia and some regions in Castile-La Mancha), which are also the predominant areas where barley, wheat and rye are harvested (the area harvested for each crop in the Iberian Peninsula can be obtained at `http://www.earthstat.org`). In addition, maize and sunflower crops play an important role in these areas, because also maize (and potato) is harvested in the western part of the Iberian Peninsula that is characterized by summer fires.

6 Concluding remarks

Scaling georeferenced event times to a unit circle allows to view fire occurrences as a spatial series of circular data. Circular spatial series however require special methods that address the circular nature of the data in a spatial setting. Our proposal is based on a hidden Markov random field for circular data. It segments the study area according to latent classes that represent specific seasons of fire occurrence. By taking this approach, we were able to indicate the most likely places where fires could occur in specific periods of the year and to capture the association between fire occurrences and land use within each season of the year.

From a technical viewpoint, this approach offers a number of advantages. First, it flexibly accommodates multimodality, skewness and kurtosis, simultaneously accounting for spatial auto-correlation. Second, it provides a parsimonious representation of the data distribution



Latent class 1 (Spring)

latent class 2 (Summer)

Latent class 3 (Autumn)

Land Cover

Figure 3: Top and bottom-left: locations of wildfires in the Iberian Peninsula in the period 2002-2012. Each point is associated to a grey level that is proportional to the posterior probability of latent class membership (darker levels indicate higher probabilities). Bottom-right: land cover in the study area.

by means of a small number of latent classes that offer an intuitively appealing interpretation of fire regimes. Third, despite of the huge sample size, model estimation is computationally feasible.

The model provides a platform for a number of possible extensions. For example, it could be exploited to investigate at what extent covariates of land morphology and weather conditions might influence the seasonality of fires. Furthermore, it could be extended to a cylindrical setting, by simultaneously modeling the time of fire occurrence and the fire intensity. Although further research is needed to explore these possibilities, in its present form the model is already capable to capture relevant aspects of fires seasonality and provide crucial information for policy planning.

Acknowledgments

We thank Prof. José M. C. Pereira and his research team from the University of Lisbon who kindly provided the wildfire data that are used in this paper. Jose Ameijeiras-Alonso and Rosa M. Crujeiras gratefully acknowledge the support of Projects MTM2016-76969-P (Spanish State Research Agency, AEI) and MTM2013-41383-P (Spanish Ministry of Economy, Industry and Competitiveness), both co-funded by the European Regional Development Fund (ERDF), IAP network from Belgian Science Policy. Part of the research was carried out by Jose Ameijeiras-Alonso during a visit to Roma Tre University supported by grants BES-2014-071006 and EEBB-I-17-12716 from the Spanish Ministry of Economy, Industry and Competitiveness. Francesco Lagona is supported by the 2015 PRIN supported project "Environmental processes and human activities: capturing their interactions via statistical methods", funded by the Italian Ministry of Education, University and Scientific Research.

References

Alfó, M., Nieddu, L., and Vicari, D. (2008). A finite mixture model for image segmentation. *Statistics and Computing*, **18**, 137–150.

Ameijeiras-Alonso, J. (2017). *Assessing Simplifying Hypotheses in Density Estimation*. Ph.D. thesis, Universidade de Santiago de Compostela.

Ameijeiras-Alonso, J., Crujeiras, R. M., and Rodríguez-Casal, A. (2018). *Applied Directional Statistics: Modern Methods and Case Studies*, chap. Directional statistics for wildfires. Chapman and Hall/CRC Press. Unpublished book.

Benali, A., Mota, B., Carvalhais, N., Oom, D., Miller, L. M., Campagnolo, M. L., and Pereira, J. (2017). Bimodal fire regimes unveil a global-scale anthropogenic fingerprint. *Global Ecology and Biogeography*.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, **24**, 179–195.

Bivand, R. S., Pebesma, E. J., and Gómez-Rubio, V. (2008). *Applied Spatial Data Analysis with R*. Springer, New York.

Feng, D., Tierney, L., and Magnotta, V. (2012). MRI tissue classification using high-resolution Bayesian hidden Markov normal mixture models. *Journal of the American Statistical Association*, **107**, 102–119.

Gaetan, C., and Guyon, X. (2010). *Spatial Statistics and Modeling*. Springer, New York.

Giglio, L., Descloitres, J., Justice, C. O., and Kaufman, Y. J. (2003). An enhanced contextual fire detection algorithm for MODIS. *Remote Sensing of Environment*, **87**, 273–282.

Godambe, V. P. (1960). An Optimum Property of Regular Maximum Likelihood Estimation. *The Annals of Mathematical Statistics*, **31**, pp. 1208–1211.

Holzmann, H., Munk, A., Suster, M., and Zucchini, W. (2006). Hidden Markov models for circular and linear-circular time series. *Environmental and Ecological Statistics*, **13**, 325–347.

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

Kato, S., and Jones, M. C. (2015). A tractable and interpretable four-parameter family of unimodal distributions on the circle. *Biometrika*, **102**, 181–190.

Korontzi, S., McCarty, J., Loboda, T., Kumar, S., and Justice, C. (2006). Global distribution of agricultural fires in croplands from 3 years of Moderate Resolution Imaging Spectroradiometer (MODIS) data. *Global Biogeochemical Cycles*, **20**.

Lagona, F. (2016). Regression analysis of correlated circular data based on the multivariate von Mises distribution. *Environmental and Ecological Statistics*, **23**, 89–113.

Lagona, F., and Picone, M. (2016). Model-based segmentation of spatial cylindrical data. *Journal of Statistical Computation and Simulation*, **86**, 2598–2610.

Lagona, F., Picone, M., and Maruotti, A. (2015). A hidden Markov model for the analysis of cylindrical time series. *Environmetrics*, p. in press.

Le Page, Y., Oom, D., Silva, J., Jönsson, P., and Pereira, J. (2010). Seasonality of vegetation fires as modified by human action: observing the deviation from eco-climatic fire regimes. *Global Ecology and Biogeography*, **19**, 575–588.

Ley, C., and Verdebout, T. (2017). *Modern Directional Statistics*. Chapman and Hall/CRC Press, Boca Raton, Florida.

Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics*, **80**, 221–239.

Magi, B., Rabin, S., Shevliakova, E., and Pacala, S. (2012). Separating agricultural and non-agricultural fire seasonality at regional scales. *Biogeosciences*, **9**, 3003–3012.

Maruotti, A., Punzo, A., Mastrantonio, G., and Lagona, F. (2016). A time-dependent extension of the projected normalregression model for longitudinal circular data basedon a hidden Markov heterogeneity structure. *Stochastic Environmental Research and Risk Assessment*, **30**, 1725–740.

Mastrantonio, G., Maruotti, A., and Jona-Lasinio, G. (2015). Bayesian hidden Markov modelling using circular-linear general projected normal distribution. *Environmetrics*, **26**, 145–158.

McNicholas, P. D. (2016). *Mixture Model-Based Classification.* Chapman and Hall/CRC Press, Boca Raton, Florida.

Modlin, D., Fuentes, M., and Reich, B. (2012). Circular conditional autoregressive modeling of vector fields. *Environmetrics*, **23**, 46–53.

Oom, D., and Pereira, J. M. C. (2013). Exploratory spatial data analysis of global MODIS active fire data. *International Journal of Applied Earth Observation and Geoinformation*, **21**, 326–340.

Ranalli, M., Lagona, F., Picone, M., and Zambianchi, E. (2018). Segmentation of sea current fields by cylindrical hidden Markov models: a composite likelihood approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, p. forthcoming.

Ranalli, M., and Rocci, R. (2016). Standard and novel model selection criteria in the pairwise likelihood estimation of a mixture model for ordinal data. *Analysis of Large and Complex Data. Studies in Classification, Data Analysis and Knowledge Organization. Editors: Adalbert F.X. Wilhelm Hans A. Kestler. (in print).*

Shirota, S., and Gelfand, A. E. (2017). Space and circular time log Gaussian Cox processes with application to crime event data. *Annals of Applied Statistics*, **11**, 481–503.

Strauss, D. J. (1977). Clustering on coloured lattices. *Journal of Applied Probability*, **14**, 135–143.

Taylor, C. C. (2008). Automatic bandwidth selection for circular density estimation. *Computational Statistics & Data Analysis*, **52**, 3493–3500.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, **21**, 1–41.

Wang, F., and Gelfand, A. E. (2014). Modeling Space and Space-Time Directional Data Using Projected Gaussian Processes. *Journal of the American Statistical Association*, **109**, 1565–1580.