Spatial clustering of curves with an application of satellite data

Gaetan Carlo, Girardi Paolo[1], Pastres Roberto
Dipartimento di Scienze Ambientali, Informatica e Statistica
Università Ca' Foscari di Venezia- Venezia (Italy)

---

[1]Address for correspondence: Paolo Girardi, Dipartimento di Scienze Ambientali, Informatica e Statistica, Università Ca' Foscari - Venezia, Campus Scientifico, Via Torino, 155, I-30170 Mestre-Venezia, Italy.
E-mail: paolo.girardi@unive.it

Spatial clustering of curves with an application of satellite data

Gaetan Carlo, Girardi Paolo[1], Pastres Roberto
Dipartimento di Scienze Ambientali, Informatica e Statistica
Università Ca' Foscari di Venezia- Venezia (Italy)

Abstract

Water quality indicators are important to identify risks to the environment, society and human health. The European Community Water Framework Directive establishes guidelines for the classification of all water bodies across Europe and chemical and biological indicators were used to this scope. In particular, the Chlorophyll type A index (Chl-a) is a shared indicator of trophic status and monitoring activities may be useful to explain its spatial distribution and to discover local dangerous behaviors (for example the anoxic events). Differently by the classical approach based on an "average" values over a period, we propose a functional clustering model that takes into account temporal and spatial dependence of Chl-a concentrations in the Adriatic Sea for defining appropriate clusters of sites. We use satellite monthly data, during the period 2002-2012, and we model the spatial dependence among the sites by means of a Markov random field model. Compared to similar attempts in literature (Jiang and Serban, 2012) our formulation includes spatial covariates. This inclusion allows for more flexibility to obtain more homogeneous and representative clusters of sites in the Adriatic Sea. The estimation of the model and the identification of the number of clusters are carried out using a pseudolikelihood function. A small simulation study complements the real data analysis.

Keywords: Functional clustering, Markov random fields, Regression splines, Restoration-Maximization algorithm.

[1]Address for correspondence: Paolo Girardi, Dipartimento di Scienze Ambientali, Informatica e Statistica, Università Ca' Foscari - Venezia, Campus Scientifico, Via Torino, 155, I-30170 Mestre-Venezia, Italy.
E-mail: paolo.girardi@unive.it

1. INTRODUCTION

According to the European Water Framework Directive (WFD; 2000/60/EC), water bodies have to be monitored in order to achieve a "good ecological status" by 2015. Since the status of most European water bodies is affected by human activities, a regular monitoring activity includes the evaluation of the ecological status of each water body (such as bio-geochemical and hydromorphological parameters) with the scope to protect the environment, society and human health. A deterioration in water quality (i.e. eutrophication and cyanobacterial blooms) presents substantial risk and can have detrimental effects on the local economy.

In the sea water, the Chlorophyll type A indicator (Chl-a) is used as biomass indicator of primary producers (e.g. photosynthetic algae, from those unicellular to multicellular ones) in the water. The Chl-a level also increases as in eutrophic conditions, i.e. in presence of high concentration of nutrients and light availability. In such circumstance, marked algal blooms may be followed by nutrient depletion and a rapid decrease in algal biomass. The subsequent degradation may then lead to hypoxic or even anoxic events. The study related to the Chl-a concentrations and its trend may be useful to discover areas with different trophic status. This is particularly true for the Adriatic Sea (Marini et al., 2010; Giani et al., 2012). Time series of data can be easily gathered by remote satellite sensing techniques. However, few studies have classified areas on the basis of the entire temporal pattern of the time series. A common practice is to consider classification methods that compare only average values in a prefixed period (e.g. a month or a year). This limitation clearly leads to lose information about the temporal pattern of the observed parameter.

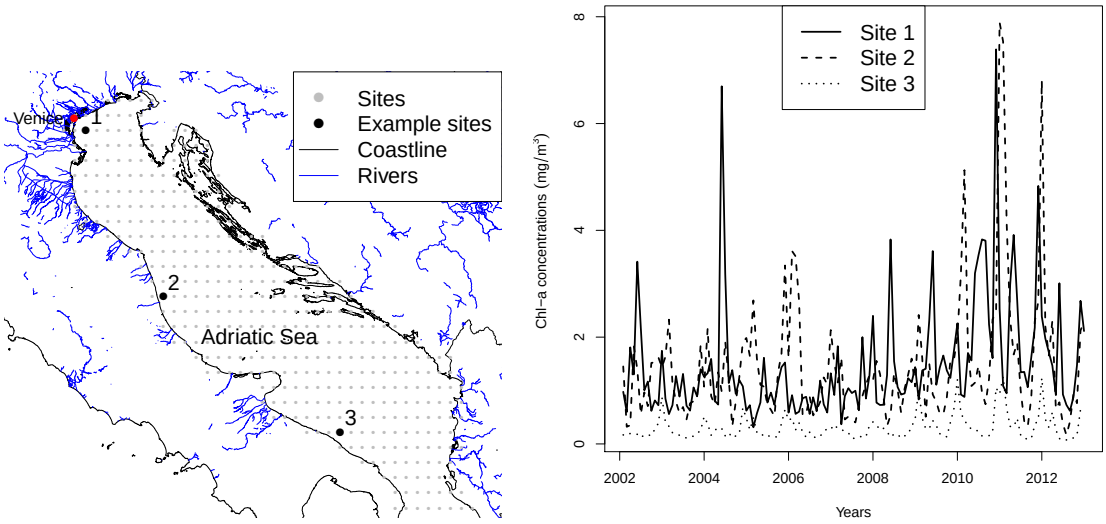
Recent advances in environmental statistics provide new clustering methods based on functional data analysis (FDA) (Ramsay and Silverman, 2005). In FDA each time series is viewed as observations of a continuous function collected at a finite series of time points. In this setting, observations are functions and the fundamental unit of interest is the entire function or curve constructed from the observations collected over time. This approach has been applied in several environmental contexts: to classify time series concerning air quality

indicators on sites as part of a monitoring network (Ignaccolo et al., 2008) or to regroup time series on water quality indicators (Pastres et al., 2011; Haggarty et al., 2012). Although with different approaches or formulations, these applications share the feature of grouping different sites together only when the observed time series have some common features, preserving sample information about the temporal pattern. However, they did not consider the spatial dependence in the clustering process and were limited to few sites.

In this respect a notable exception are two recent papers (Haggarty et al., 2015; Gaetan et al., 2016) where spatial dependence is accounted for by adjusting the L_2 dissimilarity measures calculated between two curves by means of a functional variogram (Giraldo et al., 2012). Another possibility still unexplored for environmental data is given by the bagging Voronoi classifiers (Secchi et al., 2013).

In this paper we follow a model based functional clustering approach (James and Sugar, 2003; Pastres et al., 2011; Haggarty et al., 2012, 2015) in which time series are supposed to be generated by a mixture of latent distributions. Because in our motivating example time series are collected on a regular grid or network we take into account the spatial dependence using a Markov random model as mixing distribution (Jiang and Serban, 2012). In addition we extend the modelling proposal of that by including different level of spatial dependence in the latent process, namely in the conditional mean and in the interacting potentials. This inclusion, motivated by our real problem, leads to more spatially homogeneous and consistent clustering of the water bodies in the Adriatic Sea based on Chl-a concentrations according to the current knowledge on the Chl-a dynamics in that zones.

The structure of the paper is the following. In Section 2 we highlight the main features of the data that we used. In Section 3 we introduce our model proposal and a small simulation study is presented in Section 4 for demonstrating the clustering performances. Section 5 reports our finding for the classification. A brief discussion about the results (Section 6) ends the paper.



(a) (b)

Figure 1: Observed grid-points (412 sites) in the Adriatic Sea and 3 examples of time series extracted from the dataset (Chl-a concentrations).

2. CHL-A CONCENTRATIONS OVER THE ADRIATIC SEA

Among the European waters, the Adriatic Sea reports some singular characteristics: it's almost land-locked basin separated from the central Mediterranean by the Strait of Otranto; although a fairly small size, it encompasses a significant diversity of properties, from eutrophic or oligotrophic in the northern to bloom conditions in the southern; the Italian coasts report the presence of fresh water from the rivers while in the Croatian coasts the rivers have a relatively short length and flow. Due these conditions, the Adriatic Sea is formed by an heterogeneous combination of ecosystems and its northern part is well recognized as degraded and under severe pressure (Lotze et al., 2006; Diaz and Rosenberg, 2008).

We have considered the monthly mean values of the Chl-a concentration in the Adriatic Sea from January 2002 to December 2012. Chl-a is averaged over a month because we are interested to long-term variations. Data were obtained by calibrating Ocean Colour data provided by different satellite missions, such as MERIS, SeaWiFS and MODIS, and the data set was made available by ACRI (http://hermes.acri.fr) in the framework of the GlobColour Project (Maritorena et al., 2010). For each month, we have extracted data with a resolution

of 192 x 240 grid points (longitude: 12.02^o E to 21.98^o E; latitude: 38.02^o N to 45.98^o N; 4Km scale). However only 8 217 points cover the area over the Adriatic Sea. Starting from this set of points (8 217) we obtain a valid measurement of the Chl-a concentration. However, satellite data are affected by a measurement error due to many physical factors and a quality control of data has been performed. In particular, we report that the low bathymetry and the tide level affect the quality of the measurement and lead to lack of several data. In our dataset 72.4% of extracted time series turns out to be without missing values. On the other hand 5% of time series contains eleven missing values or more. Therefore we decided to analyze a coarser resolution with an aggregation factor equal to 5, i.e. 5^2 is the number of grid points in the aggregating window. The resulting grid contains 412 sites over the Adriatic Sea (Figure 1-a). For each grid point the Chl-a concentration is calculated by averaging the values of the 25 starting curves. This aggregation increases the number of time series with complete data up to 88.1% and only 5% of time series contains four missing values or more.

This dataset presents several features which makes the clustering task interesting. Chl-a concentration is subjected to seasonal and annual changes. It is usually influenced by rivers and their input of nutrients, climatic conditions (sunlight, rain, temperature, etc..), water depth and hydrodynamics (see D'Ortenzio and Ribera d'Alcalà, 2009, for instance.)
In Figure 1(b) we show the temporal evolution of the Chl-a concentration in three sites located at the same distance from the coast. In general we expect to observe higher spring values and lower summer ones. However such seasonal pattern is masked by different grades of heteroscedasticity or completely modified as in site 1. Note also that the sites 2 and 3 have distinct mean level over the considered period due to the different sources of nutrients.

We think that a classification that takes into account the spatial dependence should permit a more homogeneous classification with, for example, a substantial noise reduction. In fact, meteorological conditions can induce severe measurement errors in the satellite images. Moreover the spatial dependence could vary according the positions of the sites. In our case we expect that spatial dependence is stronger offshore and weaker near to the sea coast. The reason is that the increasing load of nutrients from human activities (sewage effluents,

aquaculture farms and industrial facilities) or natural causes (river fluxes, lagoons) locally increments the Chl-a concentrations jointly with their spread. Thus different degrees of spatial dependence may be accounted for using external spatial information as the distance of a site from the coast, for instance.

3. METHODOLOGY

3.1 A functional spatial clustering model

Let $\mathbf{Y}_i = (Y_i(t_{i,1}), \ldots, Y_i(t_{i,m_i}))'$ a time series of m_i values collected on the site s_i, $i = 1, \ldots, n$. The value m_i corresponds to the number of observations included in the $i-th$ time series. We suppose to observe n time series and we want to cluster them in C clusters. In a mixture-model based approach to clustering the cluster membership of the i-th time series is represented by a latent random vector $\mathbf{Z}_i = (Z_{i,1}, \ldots, Z_{i,C})'$, where $Z_{i,c} = 1$ if the time series belongs to the cluster c, $Z_{i,c} = 0$ otherwise. Then a model for the complete data $(\mathbf{Y}_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$ is specified in a hierarchical way (McLachlan and Peel, 2000).

Given the cluster membership $\mathbf{Z}_i$, we suppose that $\mathbf{Y}_i$ is a discrete and noised measurement of a smooth time-varying curve $f_c(t)$, namely

$$Y_i(t_{i,j}) = f_c(t_{i,j}) + \varepsilon_i(t_{i,j}), \qquad j = 1, \ldots, m_i.$$

The smooth function $f_c(t)$ is described by a linear combination of K B-spline basis functions $B_k(t)$, $k = 1, \ldots, K$, evaluated at $K - 1$ equally spaced in knots, namely

$$f_c(t) = \sum_{k=1}^{K} \psi_{c,k} B_k(t)$$

where $\boldsymbol{\psi}_c = (\psi_{c,1}, \ldots, \psi_{c,K})'$ is a vector of unknown parameters.

In view of our application, it does not seem too restrictive to assume that $\varepsilon_i(t_{i,j})$ are temporally and spatially independent Gaussian random errors with zero mean and cluster specific variance σ_c^2. It turns out that time series $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are conditionally independent on $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ and

$$Y_i(t_{i,j}) | \mathbf{Z}_i \sim \mathcal{N}\left(\sum_{k=1}^{K} \psi_{c,k} B_k(t_{i,j}), \sigma_c^2 \right), \quad j = 1 \ldots, m_i. \tag{1}$$

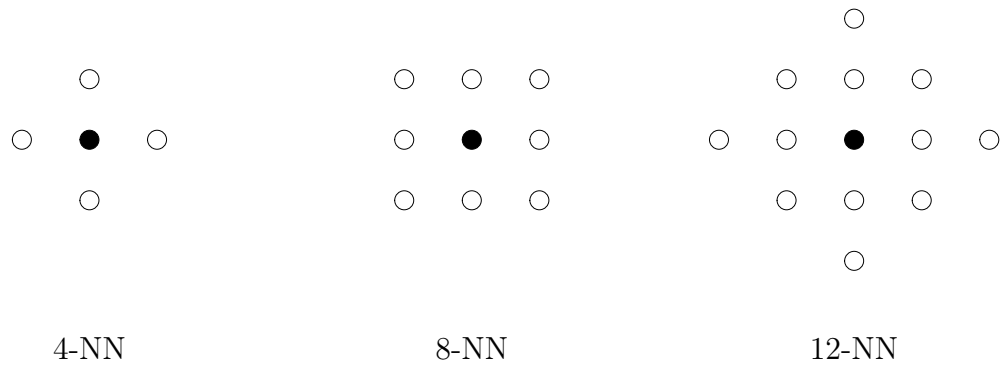


Figure 2: Three examples of nearest neighbourhoods on a 2D regular grid. Sites ($\circ$) are neighbours of the site ($\bullet$).

In the first attempts to clustering time series of water quality data (Pastres et al., 2011; Haggarty et al., 2012), the memberships $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ have been supposed independent and identically distributed multinomial variables with distribution

$$\Pr(\mathbf{Z}_i = \mathbf{z}_i) = \frac{\exp\left\{\sum_{c=1}^{C} \alpha_c z_{i,c}\right\}}{\sum_{c'=1}^{C} \exp\{\alpha_{c'}\}}, \tag{2}$$

In (2) we assume that the cluster C is the reference cluster and we set $\alpha_C = 0$ for the identifiability of the parameters. We refer to the equations (1) and (2) as Functional Clustering Model (FCM).

Jiang and Serban (2012) models possible spatial dependence among the memberships by means of a Markov Random Field (MRF). The distribution of MRF is specified in terms of the conditional distribution of a random variable in a site s_i given the values observed in the sites belonging ito its neighbourhood Examples of $k-$nearest neighbourhood (k-NN) of a site on a 2D regular grid are showed in Figure 2.

We denote by $i \sim l$ that s_i and s_l are neighbours and the set $\partial i = \{l : i \sim l\}$ indicates the neighbourhood of s_i. Given a neighbourhood system, i.e. the set of the neighbourhood $\mathcal{G} = \{\partial i, \quad i = 1, \ldots, n\}$, the MRF specification assumes that

$$\Pr(\mathbf{Z}_i = \mathbf{z}_i | \mathbf{Z}_l = \mathbf{z}_l, l \neq i) = \Pr(\mathbf{Z}_i = \mathbf{z}_i | \mathbf{Z}_{\partial i} = \mathbf{z}_{\partial i})$$

with $\mathbf{Z}_{\partial i} = (\mathbf{Z}_l, l \in \partial i)$. Jiang and Serban (2012) have considered the Potts model (Potts, 1952), namely

$$\Pr(\mathbf{Z}_i = \mathbf{z}_i | \mathbf{Z}_{\partial i} = \mathbf{z}_{\partial i}) = \frac{\exp\left\{\sum_{c=1}^{C} (\beta v_{i,c}) z_{i,c}\right\}}{\sum_{c'=1}^{C} \exp\{\beta v_{i,c'}\}}, \tag{3}$$

where $v_{i,c} = \sum_{l \in \partial i} z_{l,c}$.

On a 2D regular grid the resulting model is isotropic and the parameter β plays the role of regularization parameter: the larger β is, the more the clustered regions are geometrically regular. The value $\beta = 0$ corresponds to no spatial interaction, i.e. a multinomial distribution (2) with equal class probabilities $\Pr(\mathbf{Z}_i = \mathbf{z}_i) = 1/C$. We refer to the equations (1) and (3) as the Spatial Functional Clustering Model (SFCM).

A limitation of the SFCM is that it does not allow for spatial covariates to be incorporated into the clustering approach. This limitation has been noted by James et al. (2012) and they suggested a simple extensions of (2) through

$$\Pr(\mathbf{Z}_i = \mathbf{z}_i | \mathbf{Z}_{\partial i} = \mathbf{z}_{\partial i}) = \frac{\exp\left\{\sum_{c=1}^{C} (\beta v_{i,c} + \delta_c x_i) z_{i,c}\right\}}{\sum_{c'=1}^{C} \exp\{\beta v_{i,c} + \delta_c x_i\}}, \tag{4}$$

where x_i is a spatial covariate. However there are situations where the spatial interaction is not identical for each cluster or is potentially modulated by a spatial covariate, x_i. For this reason we propose the following extension of the model (3)

$$\Pr(\mathbf{Z}_i = \mathbf{z}_i | \mathbf{Z}_{\partial i} = \mathbf{z}_{\partial i}) = \frac{\exp\left\{\sum_{c=1}^{C} z_{i,c} (\alpha_c + \boldsymbol{\beta}_c' \mathbf{v}_i + \boldsymbol{\gamma}_c(x_i)' \mathbf{v}_i + \delta_c x_i)\right\}}{\sum_{c'=1}^{C} \exp\{\alpha_{c'} + \boldsymbol{\beta}_{c'}' \mathbf{v}_i + \boldsymbol{\gamma}_{c'}(x_i)' \mathbf{v}_i + \delta_c x_i\}}, \tag{5}$$

where $\mathbf{v}_i = (v_{i,1}, \ldots, v_{i,C})'$. The parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_C)'$, $\boldsymbol{\beta}_c = (\beta_{c,1}, \ldots, \beta_{c,C})'$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_C)'$ and the functions $\boldsymbol{\gamma}_c(x) = (\gamma_{c,1}(x), \ldots, \gamma_{c,C}(x))'$, are supposed unknown. Setting the cluster C as reference cluster, the constraints $\alpha_C = \delta_C = 0$, $\boldsymbol{\beta}_C = \boldsymbol{\gamma}_C(x) = \mathbf{0}$ make the model identifiable.

The model is not stationary and it can be easily extended to consider several spatial covariates. Moreover it is a special case of the conditioned MRF model (Divino et al., 2000, equation 3). In that paper the authors considered $\gamma_c(x)$ as a smooth non parametric

function, but in our application we prefer to consider a simpler parametric setting, namely $\gamma_c(x) = \gamma_c h(x)$, where $h(x)$ is a known function. In the sequel we term the model defined by (1) and (5) Spatial with covariates Functional Clustering Model (ScFCM).

Finally we note that FCM and SFCM are nested models of ScFCM. In fact, if $\boldsymbol{\beta}_c = \boldsymbol{\gamma}_c(x) = \boldsymbol{\delta} = \mathbf{0}$, it reduces to the multinomial distribution (2). Under the constraints $\boldsymbol{\alpha} = \boldsymbol{\gamma}_c(x) = \boldsymbol{\delta} = \mathbf{0}$ and $\beta_{cc'} = \beta$, for $c = c'$, $\beta_{cc'} = 0$ for $c \neq c'$, it reduces to Potts model (3).

3.2 Model estimation and selection of number of clusters

The ScFCM is an instance of a mixture model for which parametric estimation is usually resolved by the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). However a standard application of EM is not suitable because exact computations of the expected values in the E-step as well as the objective function in the M-step are not practical (Qian and Titterington, 1991). Here we adapt a Restoration-Maximization (RM) algorithm proposed by James et al. (2012) to overcome this problem.

We denote by $\boldsymbol{\theta}_Y = (\boldsymbol{\psi}_1', \ldots, \boldsymbol{\psi}_C', \sigma_1^2, \ldots, \sigma_C^2)'$ the collection of observation distribution parameters and $\boldsymbol{\theta}_Z = (\alpha_1, \ldots, \alpha_{C-1}, \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_{C-1}', \boldsymbol{\gamma}_1', \ldots, \boldsymbol{\gamma}_{C-1}')'$ the collection of the MRF parameters. Moreover f indicates a probability mass function or a probability density function depending upon which random variable is considered.

Given the observations $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_n')'$ and the current values $\boldsymbol{\theta}_Y^{(k)}$, $\boldsymbol{\theta}_Z^{(k)}$, and $\mathbf{z}^{(k)}$,

M-step
- we obtain the estimate $\boldsymbol{\theta}_Y^{(k+1)}$ maximizing the conditional likelihood

$$L_Y(\boldsymbol{\theta}_Y | \mathbf{z}^{(k)}; \mathbf{y}) = \prod_{i=1}^{n} f(\mathbf{y}_i | \mathbf{z}_i^{(k)}; \boldsymbol{\theta}_Y);$$

- we obtain the estimate $\boldsymbol{\theta}_Z^{(k+1)}$ by maximizing the pseudolikelihood (Besag, 1974)

$$PL_Z(\boldsymbol{\theta}_Z; \mathbf{z}^{(k)}) = \prod_{i=1}^{n} f(\mathbf{z}_i^{(k)} | \mathbf{z}_{\partial i}^{(k)}; \boldsymbol{\theta}_Z);$$

R-step
- we restore $\mathbf{z}$ with $\mathbf{z}^{(k+1)} = ({\mathbf{z}'}_1^{(k+1)}, \ldots, {\mathbf{z}'}_n^{(k+1)})'$, using the Iterative Conditional Mode (ICM) algorithm (Besag, 1986), based on $\boldsymbol{\theta}_Y^{(k+1)}$ and $\boldsymbol{\theta}_Z^{(k+1)}$. Precisely, we obtain $\mathbf{z}_i^{(k+1)}$ in turn by

$$\begin{aligned}\mathbf{z}_i^{(k+1)} &= \text{argmin}_{\mathbf{w}} \Pr(\mathbf{Z}_i = \mathbf{w}|\mathbf{y}, \mathbf{z}_1^{(k+1)}, \ldots, \mathbf{z}_{i-1}^{(k+1)}, \mathbf{z}_{i+1}^{(k)}, \ldots, \mathbf{z}_n^{(k)}) \\ &= \text{argmin}_{\mathbf{w}} f(\mathbf{y}_i|\mathbf{w}; \boldsymbol{\theta}_Y)\, f(\mathbf{w}|\mathbf{z}_1^{(k+1)}, \ldots, \mathbf{z}_{i-1}^{(k+1)}, \mathbf{z}_{i+1}^{(k)}, \ldots, \mathbf{z}_n^{(k)}; \boldsymbol{\theta}_Z).\end{aligned}$$

We repeat the previous steps until the convergence of the parameters. An initial guess $\mathbf{z}^{(0)}$ can be obtained by the unsupervised functional clustering procedure described in Abraham et al. (2003).

The algorithm has been implemented using standard fitting procedures in R language. Given the membership $\mathbf{z}_i^{(k)}$ the (conditional) likelihood of each time series $f(\mathbf{y}_i|\mathbf{z}_i^{(k)}; \boldsymbol{\theta}_Y)$ is a multivariate normal density with unknown parameters $\boldsymbol{\psi}_c$ and σ_c^2 that can be easily estimated using function `lm`. The pseudolikelihood $PL_Z(\boldsymbol{\theta}_Z; \mathbf{z})$ corresponds to the likelihood of a baseline multinomial logit model (Agresti, 2013) and $\boldsymbol{\theta}_Z$ can be estimated using function `multinom` in the recommended package nnet (Venables and Ripley, 2002). Finally the iterative conditional mode (ICM) algorithm has been coded in C and integrated into R by using RcppArmadillo package (Eddelbuettel and Sanderson, 2014).

The full implementation of the estimation procedure requires to select the number of basis functions representing the membership curves and the desided number of cluster C. For the FCM the use of likelihood based criteria like the Akaike Information Criteria (AIC) or the Bayesian Information Criteria (BIC) has been advocated in the literature (Fraley and Raftery, 1999; Pastres et al., 2011). However, for SFCM and ScFCM the evaluation of the likelihood

$$L(\boldsymbol{\theta}_Y, \boldsymbol{\theta}_Z) = \sum_{\mathbf{z}_1} \cdots \sum_{\mathbf{z}_n} f(\mathbf{y}_1|\mathbf{z}_1; \boldsymbol{\theta}_Y) \times \cdots \times f(\mathbf{y}_1|\mathbf{z}_n; \boldsymbol{\theta}_Y) f(\mathbf{z}_1, \ldots \mathbf{z}_n; \boldsymbol{\theta}_Z) \tag{6}$$

is not feasible because the sum involves all possible configurations of the hidden vectors . With n sites and C clusters, there are C^n possible configurations, which is huge, making this approach intractable. Stanford and Raftery (2002) have suggested to approximate the

required likelihood by the pseudolikelihood

$$PL(\boldsymbol{\theta}_Y, \boldsymbol{\theta}_Z) = \prod_{i=1}^{n} \sum_{\mathbf{z}_i} f(\mathbf{y}_i|\mathbf{z}_i; \boldsymbol{\theta}_Y) f(\mathbf{z}_i|\widehat{\mathbf{z}}_{\partial i}; \boldsymbol{\theta}_Z) \tag{7}$$

where $\widehat{\mathbf{z}}_i$, $i = 1, \ldots, n$, are the restored cluster memberships obtained at the convergence of the RM algorithm.

In view of our application we propose a practical and simpler solution. First of all we choose K, the number of basis functions, by minimizing the overall criterion

$$BIC^*(K) = \sum_{i=1}^{n} (m_i \log(s_i^2) + K \log(m_i)) \tag{8}$$

with $s_i^2 = m_i^{-1} \sum_{j=1}^{m_i} \{Y_i(t_{i,j}) - \sum_{k=1}^{K} \widehat{\psi}_{i,k} B_k(t_{i,j})\}^2$.

Instead of summing over all possible configurations of $\mathbf{z}_i$ in (7), we consider the ICM restoration of $\mathbf{z}$ and we get

$$PL^*(\boldsymbol{\theta}_Y, \boldsymbol{\theta}_Z) = \prod_{i=1}^{n} f(\mathbf{y}_i|\widehat{\mathbf{z}}_i; \boldsymbol{\theta}_Y) f(\widehat{\mathbf{z}}_i|\widehat{\mathbf{z}}_{\partial i}; \boldsymbol{\theta}_Z); \tag{9}$$

as further approximation of (6).

Then we select C, the number of clusters, by minimizing the pseudolikelihood criterion:

$$PLIC(C) = -2 \log PL^*(\widehat{\boldsymbol{\theta}}_Y, \widehat{\boldsymbol{\theta}_Z}) + (\dim(\widehat{\boldsymbol{\theta}}_Y) + \dim(\widehat{\boldsymbol{\theta}}_Z)) \log \left(\sum_{i=1}^{n} m_i \right). \tag{10}$$

We adopt a sequential approach to choosing C to maximize $PLIC(C)$. We begin by computing $PLIC(C)$ for $C = 1$, and then incrementally increase the value of C. We take the first local minimum of $PLIC(C)$ to be our choice for the number of clusters.

4. SIMULATION STUDY

In this small simulation study we want to assess the ability of the selection procedure introduced in the previous section to detect the 'true' number of clusters. In particular we consider a setup close to the real example we deal with and characterized by a non stationary spatial dependence.

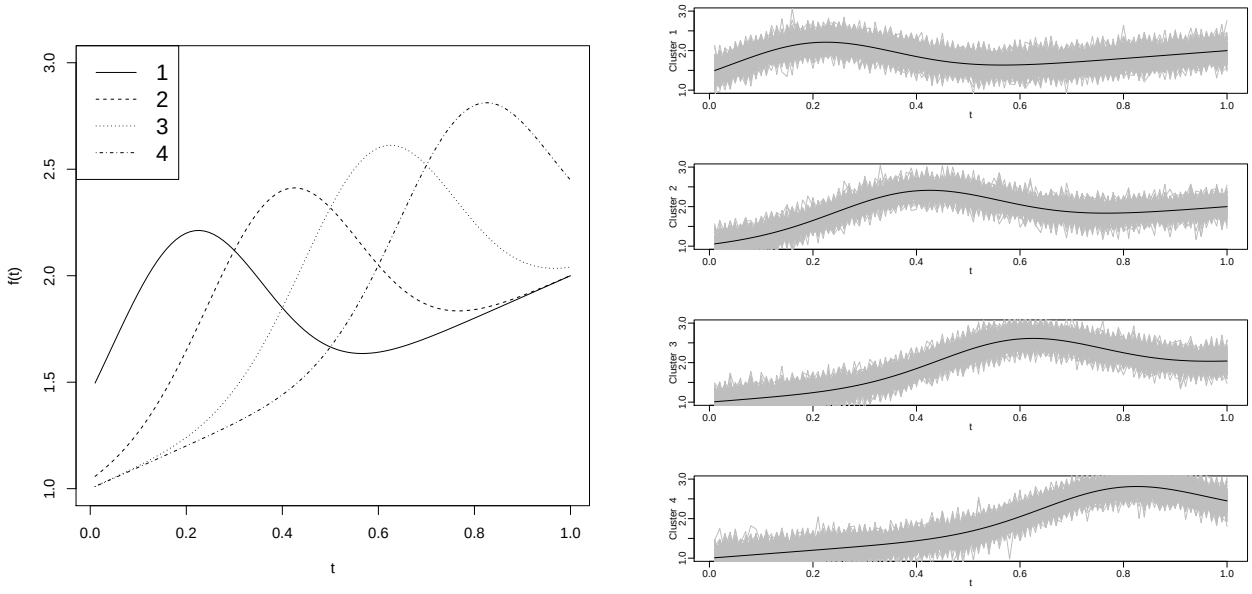


Figure 3: On the left, $f_c(t)$ curves for $C = 1, \ldots, 4$; on the right, the noised curves with error variance $\sigma^2 = 0.04$.

We simulate random curves, observed at one hundred time points $t = 0.01, 0.02, \ldots, 1$, from the model

$$Y(t) = f_c(t) + \varepsilon(t)$$

where $\varepsilon(t)$ is a Gaussian white noise with variance σ^2. The curves belong to C clusters with different patterns

$$f_c(t) = 1 + t + \exp\left\{20\left(t - \frac{c}{C+1}\right)^2\right\}, \qquad c = 1, \ldots, C, \tag{11}$$

according the values of the label c. An illustrative example of simulated curves is reported in Figure 3.

The curves are located on a regular grid of 900 points $s_i = (i_1, i_2)$, $i_1 = 1, \ldots, 30, i_2 = 1, \ldots, 30$, according to the cluster membership $\mathbf{z}_i$ generated from the MRF with conditional distribution

$$\Pr(\mathbf{Z}_i = \mathbf{z}_i | \mathbf{Z}_{\partial i} = \mathbf{z}_{\partial i}, s_i) = \frac{\exp\left\{\sum_{c=1}^{C} [\gamma h(s_i) v_{i,c}] z_{i,c}\right\}}{\sum_{c'=1}^{C} \exp\{\gamma h(s_i) v_{i,c'}\}}, \tag{12}$$

where $h(s_i) = 1$ if $i_2 \leq 15$ and $h(s_i) = 0.5$ otherwise. The role of the function $h(\cdot)$ is to

account for different strength of spatial dependence (γ) according to the location s_i. In this specification the spatial dependence in the 15 bottom rows of the grid is stronger than in the remaining rows, encouraging the clustering in larger regions. The resulting spatial pattern with $C = 4$ clusters and $\gamma = 2$ is illustrated in Figure 4 where the simulated image is obtained running a Gibbs sampler.

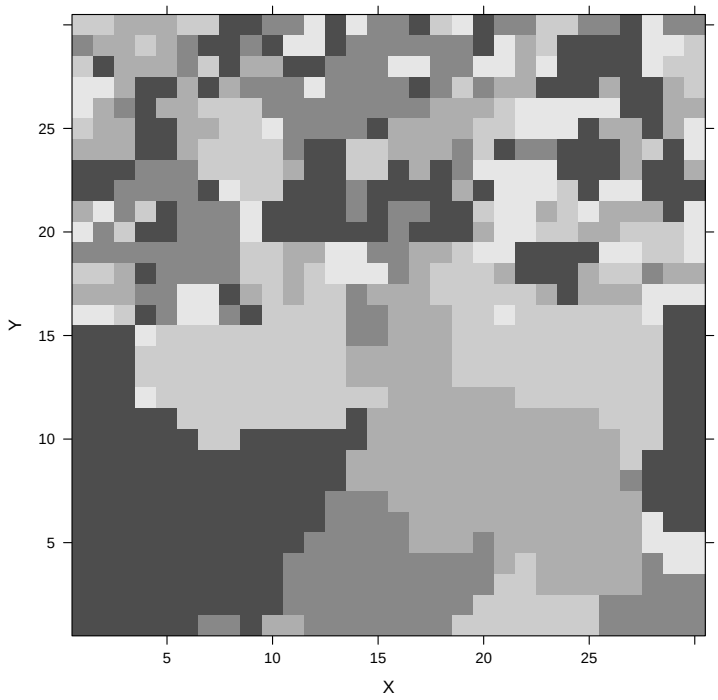


Figure 4: Simulation from the MRF (12) with $C = 4$ clusters and $\gamma = 2$.

In the simulation setting we have considered different combinations according three factors: the number of clusters C ($C = 2, 4$), the variance of the noise term σ^2 ($\sigma^2 = 0.01, 0.04$), and the strength of the spatial dependence γ ($\gamma = 2, 4$).

Table 1 reports the frequencies of the number of clusters identified in 100 Monte Carlo experiments by means of $PLIC(C)$. In the fitting procedure the number of basis function, K, is set to 9, according to overall BIC proposed in (8). The results from ScFCM have been compared to the tables obtained by using the SFCM with MRF (3) and the simple FCM.

It is encouraging that we obtain the best results under ScFCM. This is especially trustwor-

thy when the true number of cluster is $C = 2$. Moreover in such case the model seems robust to the reduction of the signal-to-noise ratio. Note that SFCM still has a discrete ability to classify the synthetic time series in all simulation instead the FCM yields to poorer results.

		$\gamma = 2$						$\gamma = 4$						
		Identified clusters						Identified clusters						
C	σ^2	1	2	3	4	5	6	1	2	3	4	5	6	
2	0.01	0	94	5	1	0	0	0	88	10	2	0	0	ScFCM
		0	91	9	0	0	0	0	79	20	1	0	0	SFCM
		0	0	0	0	0	100	0	0	0	0	6	94	FCM
2	0.04	0	95	5	0	0	0	0	93	7	0	0	0	ScFCM
		0	89	11	0	0	0	0	82	16	2	0	0	SFCM
		0	0	0	0	2	98	0	0	0	0	3	97	FCM
4	0.01	0	0	0	100	0	0	0	0	4	91	4	1	ScFCM
		0	0	0	100	0	0	0	0	3	84	11	2	SFCM
		0	0	0	0	3	97	0	0	0	0	2	98	FCM
4	0.04	0	0	0	100	0	0	0	0	3	88	7	2	ScFCM
		0	0	0	99	1	0	0	0	2	85	7	6	SFCM
		0	0	0	0	3	97	0	0	0	0	5	95	FCM

Table 1: Identified clusters in 100 Monte Carlo experiments.

5. CLUSTERING CHL-A CONCENTRATION DATA

Satellite Chl-a data show a log-normal distribution (Campbell, 1995) and we have considered in our analysis log-transformed values. Moreover the time series of each site has been standardized to have zero mean and unit variance. The standardization partially allows to remove climatic effects: sites belonging to the Adriatic Sea are characterized by several patterns of temperatures and precipitations producing a different variability and average

value of Chl-a concentrations.

We choose a cubic B-spline basis with equally spaced knots over the interval $[1, 132]$ to represent the mean value of the individual random curve. The number of basis function has been selected by means of (8). At $K = 30$ we have recorded the most pronounced minimum (see Figure 5) of $BIC^*(K)$. This choice corresponds approximately to one knot every four months and allows to take properly into account a long term trend jointly with a seasonal component.

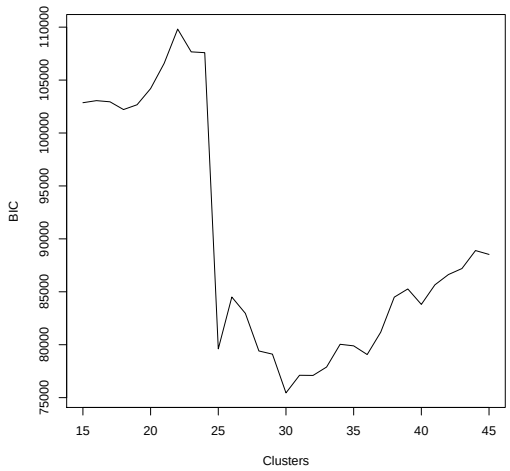


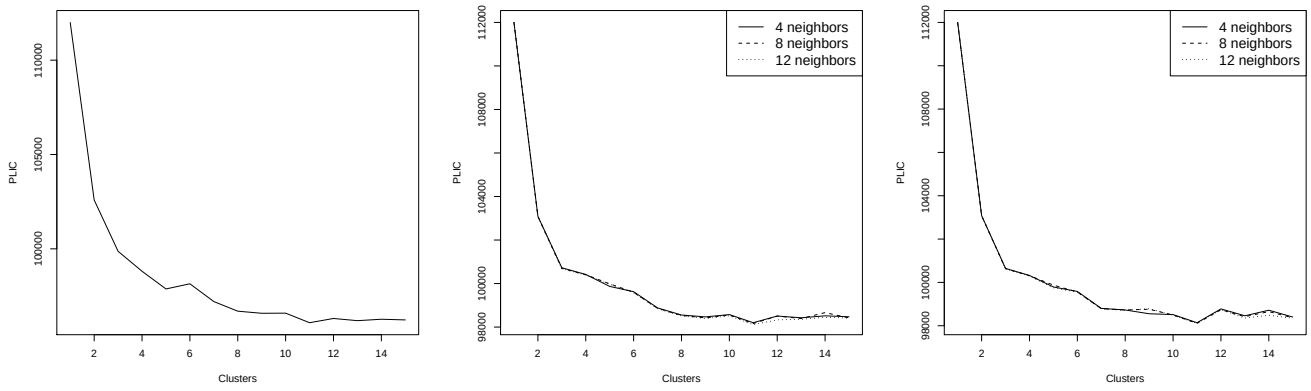
Figure 5: Values of BIC^* for different choices of the number of basis functions.

We fit FCM, SFCM and ScFCM and we use the $PLIC(C)$ for identifying the number of clusters keeping the number of basis function fixed at $K = 30$. The number of clusters is chosen at the first local minimum (see Figure 6).

In Figure 8-a we present the spatial distribution of the clusters we have obtained by FCM. The sites are labelled using a greyscale according to the mean values of the clusters and the highest values of Chl-a concentrations are associated to the darkest colours.

The resulting classification seems simplistic. In particular the first cluster embraces a large portion of the Northern sites including the Venice and Grado lagoons in Italy and the Karavasta Lagoon in Albania.

However it is well documented (Giani et al., 2012; Marini et al., 2010) how the fresh water



(a) (b) (c)

Figure 6: $PLIC(C)$ values for different number of clusters: (a) FCM; (b) SFCM; (c) ScFCM. For the last two models the values according to different neighbourhood systems are depicted.

of dynamics related to the coast strongly influences the Chl-a dynamics. We note also that the main influencing factors (rivers, lagoons, discharges, etc...) are located in the North of the Adriatic sea. Our belief is that introducing this spatial information in the clustering could pool the sites in more spatially homogeneous clusters and give a more detailed gradient of the Chl-a concentration when we move away from the coastline.

Therefore we consider an instance of model (5). The Adriatic Sea is a semi-closed basin and for any site s_i the distance d_i from a coast can vary from 0 to 97 km. As a covariate we consider the indicator variable, x_i, such that $x_i = 1$ if distance d_i is less than 30 km and $x_i = 0$ otherwise (see Figure 7). The choice of the threshold 30 km is consistent to the results in the literature (Kourafalou, 2001). In this respect the variable x_i acts as modulating factor and we set $\gamma_c(x_i) = \gamma_c x_i$ into (5).

The resulting MRF has conditional distribution is given by

$$\Pr(\mathbf{Z}_i = \mathbf{z}_i | \mathbf{Z}_{\partial i} = \mathbf{z}_{\partial i}) = \frac{\exp\left\{\sum_{c=1}^{C}[(\beta + \gamma x_i)v_{i,c} + \delta_c x_i]z_{i,c}\right\}}{\sum_{c'=1}^{C}\exp\{(\beta + \gamma x_i)v_{i,c} + \delta_c x_i\}}. \tag{13}$$

We compare the specification (13) with (3) in which we assume that there is no modulating factor (i.e. $\gamma = 0$ and $\delta_c = 0$). Three neighbourhood systems have been considered: 4-NN, 8-NN and 12 NN where a large number of neighbours results in wider connected clusters.

According the $PLIC(C)$ values (Figures 6-b and 6-c) we select a number of clusters equal to $C = 9$ for SFCM and $C = 8$ for ScFCM with a slight preference for the 12-NN system. We find also that $PLIC(C)$ does not provide a sharp indication in favour of one of two models.

It is also interesting to note that SFCM and ScFCM yield to different partitions of the sites. We measure he agreement between two resulting partitions by means of the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985). An index value near to one (zero) indicates a good agreement (weak agreement). Looking at Table 2, the resulting partitions seem quite different.

Comparison	ARI
FCM - SFCM	0.44
FCM - ScFCM	0.54
SFCM - ScFCM	0.62

Table 2: ARI values obtained comparing FCM, SFCM and ScFCM.

Moreover the clustering procedures which include the spatial dependence (SFCM: Figure 8-b; ScFCM: Figure 8-c) seem to be able to better catch the influence of the freshwaters from the rivers defining a unique cluster (Cluster 1) closed to the Po river's mouth, the main river that discharges into Adriatic Sea. However we observe important differences in the partition of the Northern Adriatic Sea in Figure 8-c. The spatial extension of the Cluster 1 is coherent with the influence of the Alpine rivers (especially from Po river), while Cluster 2 and 3 stretch, in anti-clockwise direction, from the Dalmatian coast to the offshore sites of the Italian side, up the Intermediate Adriatic Sea; this finding is in agreement with the superficial sea stream circulation (Giani et al., 2012).

Finally Figure 9 shows the time-varying curves related to the obtained clusters by ScFCM model. In Cluster 1 the temporal pattern differs essentially from one of the other clusters because the seasonal component is nearly absent. This pattern is consistent with an eutrophic area due to the presence of fresh waters from the rivers and the rainfall regime. Cluster 2 and

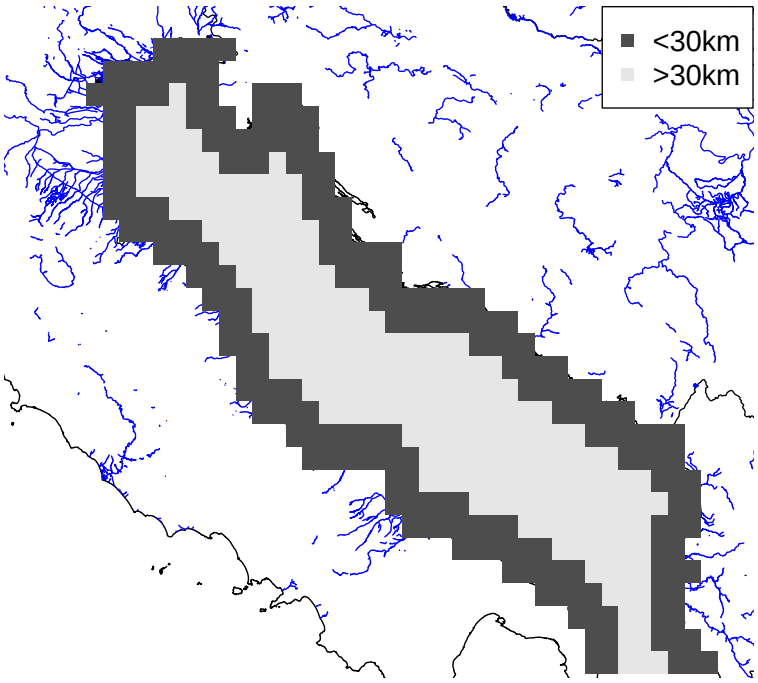


Figure 7: Spatial distribution of the sites (in dark colour) within 30 km from the coast.

3 reports a slight upward trend associated with an unsteady seasonal oscillation. Both clusters represent sites relatively influenced from the fresh waters of local rivers and the seasonal variability. They define a borderline condition where seasonal algal blooms and enriched fresh waters mix. The sites in the clusters from 4 to 8 exhibit a clear seasonal pattern with different cycles of positive peaks in the early spring and negative peaks during the summer period. Looking more closely Cluster 4 presents a constant seasonality in agreement with the annual stream of the rivers in the eastern side of the Apennines. The last three clusters (6, 7, 8), located far from the coasts, exhibit a flat trend and an evident seasonal pattern in the period 2002-2008. Afterwards all of them show an augmented inter-annual variability in the last part of the considered temporal period, associated to high peaks in the years 2010:2011.

6. DISCUSSION

In this paper we have proposed a model based approach to the functional clustering when time-varying curves are spatially dependent. In particular we have considered data on a regular

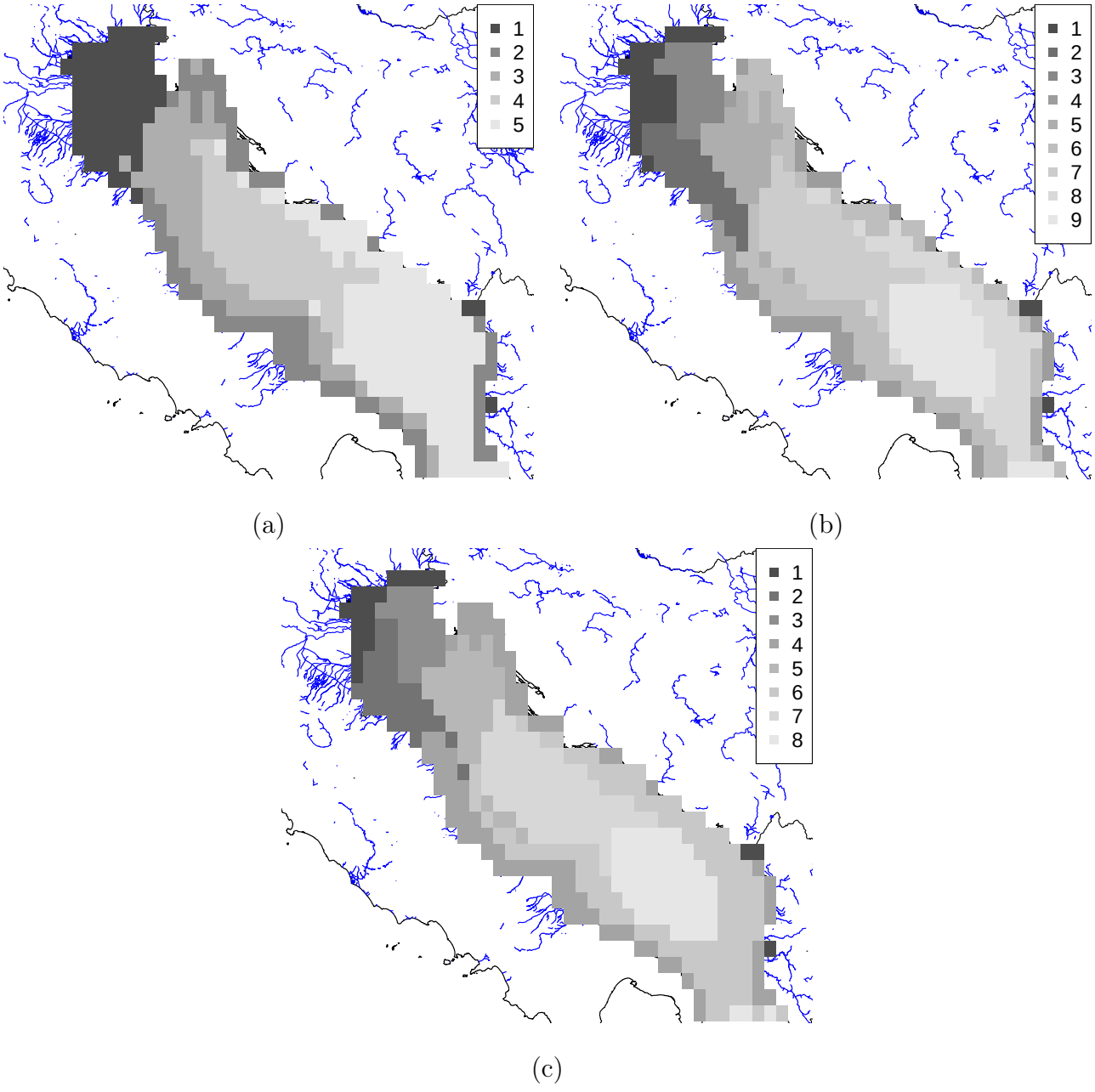


Figure 8: Spatial clustering resulting from (a)FCM model with 5 clusters, (b) FSCM with 9 clusters and (c) ScFCM with 8 clusters.

grid and spatial dependence has been modelled using non stationary MRF. With respect to existing methodological literature (James and Sugar, 2003) our model incorporates external information for modulating the spatial dependence. Such inclusion has been motivated

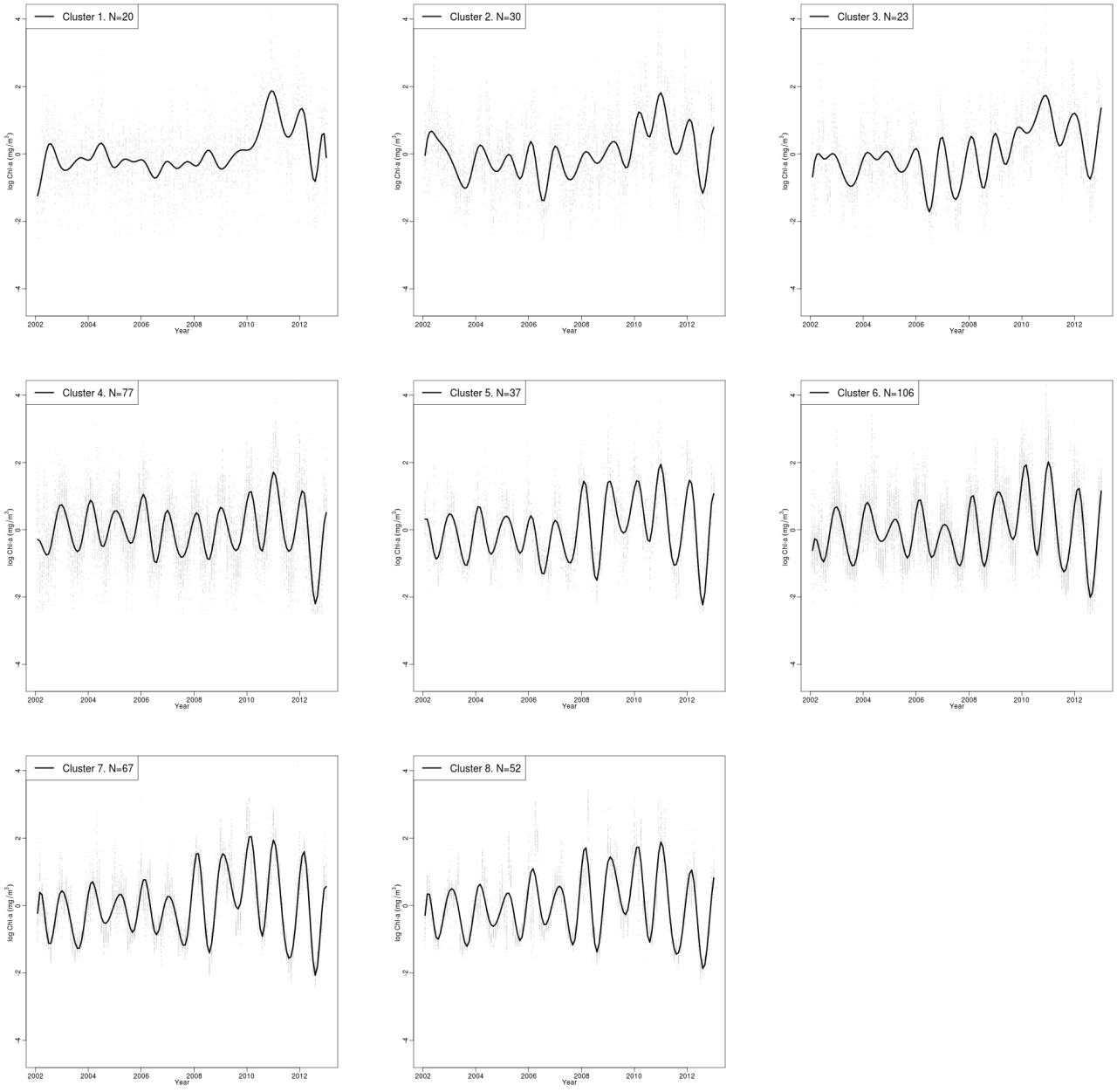


Figure 9: Log chl-a distribution of each cluster obtained by the ScFCM model with the number of sites belonging to each cluster.

by analysing a temporal dataset of satellite measures of Chl-a concentrations over the Adriatic Sea. The aim of the study was to provide a classification of the Adriatic Sea in homogeneous zones with respect to the temporal pattern. It is worth noting that the dataset has moderate size, namely 412 grid points monthly observed 132 times from January, 2002 to

December, 2012. All calculations have been carried out on a laptop equipped with 1.80 GHz 4 core-processor and 8 GB of RAM. One iteration of the proposed RM algorithm requires approximately 30 seconds. We have analysed a dataset with a reduced resolution in order to decrease the presence of missing data and outliers and to achieve an acceptable computational time. Actually the increasing resolution of satellite data from the new ESA Sentinel program produces high resolution datasets and the number of sites in a grid increases with the square of the resolution. Dealing with a dataset at the original resolution allows to avoid negative effects of data scaling (Raptis et al., 2003), but the computational time for estimating the model can be a serious problem. In fact parallel programming can only partially decrease the computational time, especially in presence of sequential iterative algorithms such as in our proposal.

Our modelling strategy helps to classify in a better way the areas affected by different trends. It introduces a certain grade of flexibility in the spatial dependence specification which may vary in accordance with the taken values of the spatial covariate.

Particular attention has to be paid to find the right possible covariate which can be inserted in the spatial dependence. In this study we have included a physical measure related to the spatial domain (distance of a site from a river mouth), but the proposed method may admit many choices and parameterizations.

The inclusion of non stationary terms modifies the clustering and yields to a zonification consistent with previous studies on the Adriatic Sea (Lazzari et al., 2012). In fact the clustering is driven by enriched waters of the rivers and highlights the differences between the eastern coast of the Adriatic Sea with respect to the Croatian coast. Only the Albanian coastal area reports the presence of a behaviour similar to the Italian coastal waters in accordance with the hydrological basin in that zone (Marini et al., 2010).

Comparing with simplest previous studies (D'Ortenzio and Ribera d'Alcalà, 2009) that tried to classify the sea superficial Chl-a concentrations in the Mediterranean Sea using a K-means procedure without exploiting the temporal patterns we have obtained a more consistent classification.

These promising results motivate future extensions of the model to cope with the information coming from several physical attributes as in Li et al. (2016), Ramos et al. (2012) and Haggarty et al. (2012).

REFERENCES

Abraham, C., Cornillon, P. A., Matzner-Løeber, E., Molinari, N., 2003. Unsupervised curve clustering using B-splines. Scandinavian Journal of Statistics 30, 581–595.

Agresti, A., 2013. Categorical Data Analysis, 3rd Edition. Wiley & Sons, New York.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society, Series B 36, 192–236.

Besag, J., 1986. On the statistical analysis of dirty pictures. Journal of the Royal Statistical Society, Series B 48, 259–302.

Campbell, J. W., 1995. The lognormal distribution as a model for bio-optical variability in the sea. Journal of Geophysical Research: Oceans 100, 13237–13254.

Dempster, A., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39, 1–38.

Diaz, R. J., Rosenberg, R., 2008. Spreading dead zones and consequences for marine ecosystems. Science 321, 926–929.

Divino, F., Frigessi, A., Green, P. J., 2000. Penalized pseudolikelihood inference in spatial interaction models with covariates. Scandinavian Journal of Statistics 27, 445–458.

D'Ortenzio, F., Ribera d'Alcalà, M., 2009. On the trophic regimes of the Mediterranean Sea: a satellite analysis. Biogeosciences 6, 139–148.

Eddelbuettel, D., Sanderson, C., 2014. RcppArmadillo: accelerating R with high-performance C++ linear algebra. Computational Statistics & Data Analysis 71, 1054–1063.

Fraley, C., Raftery, A. E., 1999. MCLUST: software for model-based cluster analysis. Journal of Classification 16, 297–306.

Gaetan, C., Girardi, P., Pastres, R., Mangin, A., 2016. Clustering chlorophyll-a satellite data using quantiles. Annals of Applied Statistics 10, 964–988.

Giani, M., Djakovac, T., Degobbis, D., Cozzi, S., Solidoro, C., Umani, S. F., 2012. Recent changes in the marine ecosystems of the northern Adriatic Sea. Estuarine, Coastal and Shelf Science 115, 1–13.

Giraldo, R., Delicado, P., Mateu, J., 2012. Hierarchical clustering of spatially correlated functional data. Statistica Neerlandica 66, 403–421.

Haggarty, R., Miller, C., Scott, E., Wyllie, F., Smith, M., 2012. Functional clustering of water quality data in Scotland. Environmetrics 23, 685–695.

Haggarty, R. A., Miller, C. A., Scott, E. M., 2015. Spatially weighted functional clustering of river network data. Journal of the Royal Statistical Society: Series C 64, 491–506.

Hubert, L., Arabie, P., 1985. Comparing partitions. Journal of Classification 2, 193–218.

Ignaccolo, R., Ghigo, S., Giovenali, E., 2008. Analysis of air quality monitoring networks by functional clustering. Environmetrics 19, 672–686.

James, G. M., Sugar, C. A., 2003. Clustering for sparsely sampled functional data. Journal of the American Statistical Association 98, 397–408.

James, G. M., Sun, W., Qiao, X., 2012. Comment on "Clustering random curves under spatial dependence" by H. Jiang and N. Serban. Technometrics 54, 123–126.

Jiang, H., Serban, N., 2012. Clustering random curves under spatial interdependence with application to service accessibility. Technometrics 54, 108–119.

Kourafalou, V. H., 2001. River plume development in semi-enclosed Mediterranean regions: north Adriatic Sea and northwestern Aegean Sea. Journal of Marine Systems 30, 181–205.

Lazzari, P., Solidoro, C., Ibello, V., Salon, S., Teruzzi, A., Béranger, K., Colella, S., Crise, A., 2012. Seasonal and inter-annual variability of plankton chlorophyll and primary production in the Mediterranean Sea: a modelling approach. Biogeosciences 9, 217–233.

Li, H., Deng, X., Dolloff, C., Smith, E., 2016. Bivariate functional data clustering: grouping streams based on a varying coefficient model of the stream water and air temperature relationship. Environmetrics 27, 15–26.

Lotze, H. K., Lenihan, H. S., Bourque, B. J., Bradbury, R. H., Cooke, R. G., Kay, M. C., Kidwell, S. M., Kirby, M. X., Peterson, C. H., Jackson, J. B., 2006. Depletion, degradation, and recovery potential of estuaries and coastal seas. Science 312, 1806–1809.

Marini, M., Grilli, F., Guarnieri, A., Jones, B. H., Klajic, Z., Pinardi, N., Sanxhaku, M., 2010. Is the southeastern Adriatic Sea coastal strip an eutrophic area ? Estuarine, Coastal and Shelf Science 88, 395–406.

Maritorena, S., d'Andon, O. H. F., Mangin, A., Siegel, D. A., 2010. Merged satellite ocean color data products using a bio-optical model: Characteristics, benefits and issues. Remote Sensing of Environment 114, 1791–1804.

McLachlan, G. J., Peel, D., 2000. Finite Mixture Models. Wiley & Sons, New York.

Pastres, R., Pastore, A., Tonellato, S. F., 2011. Looking for similar patterns among monitoring stations. Venice Lagoon application. Environmetrics 22, 712–724.

Potts, R., 1952. Some generalized order-disorder transitions. Mathematical Proceedings of the Cambridge Philosophical Society 48, 106–109.

Qian, W., Titterington, D. M., 1991. Estimation of parameters in hidden Markov models. Philosophical Transactions: Physical Sciences and Engineering 337, 407–428.

Ramos, E., Juanes, J. A., Galván, C., Neto, J. M., Melo, R., Pedersen, A., Scanlan, C., Wilkes, R., van den Bergh, E., Blomqvist, M., Karup, H. P., Heiber, W., Reitsma, J. M.,

Ximenes, M. C., Silió, A., Méndez, F., González, B., 2012. Coastal waters classification based on physical attributes along the NE Atlantic region. an approach for rocky macroalgae potential distribution. Estuarine, Coastal and Shelf Science 112, 105–114.

Ramsay, J., Silverman, B., 2005. Functional Data Analysis, 2nd Edition. Springer, New York.

Raptis, V., Vaughan, R., Wright, G., 2003. The effect of scaling on land cover classification from satellite data. Computers & Geosciences 29, 705–714.

Secchi, P., Vantini, S., Vitelli, V., 2013. Bagging Voronoi classifiers for clustering spatial functional data. International Journal of Applied Earth Observation and Geoinformation 22, 53–64.

Stanford, D. C., Raftery, A. E., 2002. Approximate Bayes factors for image segmentation: the pseudolikelihood information criterion (PLIC). IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 1517–1520.

Venables, W. N., Ripley, B. D., 2002. Modern Applied Statistics with S, 4th Edition. Springer, New York.