The role of spatial dependence on the functional clustering based on the smoothing splines regression

Carlo Gaetan[1], Paolo Girardi[1], Roberto Pastres[1]

[1] *Department of Environmental Sciences, Informatics and Statistics*
Universitá Ca' Foscari - Venezia, Dorsoduro 2137
I-30121 Venezia, Italy
e-mail: carlo.gaetan@unive.it, e-mail: paolo.girardi@unive.it, e-mail: roberto.pastres@unive.it*
∗Corresponding author

Abstract. *The use of quality indicators of water is of crucial importance 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 is a shared indicator of trophic status and a monitoring activities may be useful to explain its spatial distribution and to discover local dangerous behaviours (for example the anoxic events). Despite of evaluation based on an "average" value over the hole period, we investigate and develop univariate functional clustering models to investigate the spatiotemporal structure of chlorophyll type A concentrations on the Adriatic Sea and to define appropriate groups. In addition, the importance of the spatial dependence was evaluated by comparing two different methodology. The data for the classification analysis is formed by glob-colours data during the period 2002-2012 (monthly values, 11 calendar years) provided by the ACRI server (http://hermes.acri.fr/) using satellite data source combining information of MERIS, Seaways and MODIS optical sensors.*

Keywords. *functional clustering,spatial dependence,clustering methods,smoothing splines*

1 Introduction

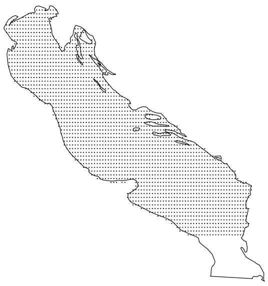
According to the European Water Framework Directive (WFD; 2000/60/EC), water bodies have to be monitored in order to achieving by 2015 a "good ecological status". Since the status of most European water bodies is affected by human activities, a regular monitoring activity includes the evaluation of the ecological status of each water body (such us bio-geochemical and hydromorphological parameter) 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. Long-term data records across multiple sites can be used to investigate water quality and risk factors statistically [Ferguson *et al.*(2008), Carvalho *et al.*(2011)]. In the recent years, there was a huge increase of raw data: the development of computational and measuring processes, as processing of satellite sensor data, implicates the available of a large amount of time series. In the sea water, the chlorophyll type A is used as biomass indicator of primary producers (e.g. photosynthetic algae, from those unicellular to multicellular ones) in the water. The level of chlorophyll type A also increases as in eutrophic conditions: 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. Particular behaviours of chlorophyll type A concentrations may be useful to discover areas related to different trophic status. It is a common practice to use the more traditional classification methods by considering the temporal averages (e.g. annual). This is clearly a limitation, since the whole information about the dynamics of the observed variables is lost. Recent literature provides some new classification methods based on Functional Data Analysis (FDA [James and Sugar(2003), Jacques and Preda(2013), Ramsay *et al.*(2011)]. In the environmental field the functional clustering model (FCM), developed by James and Sugar [James and Sugar(2003)], is a common used methods for classifying curves. With this approach, smooth functions are fit to the observed data for each determinant at each site, and then, a cluster analysis method is applied to these curves. In the context of the analysis of water bodies time series [Haggarty *et al.*(2012)] FCM was used to classify Scotland lakes using the temporal pattern of four water parameters. Similarly, Pastres et al. classified monitoring sites in the Venice Lagoon [Pastres *et al.*(2011)]. In addition, the same methodology way applied to site classification in the air quality network of Piemonte (Italy).[Ignaccolo *et al.*(2008)]. These latter approaches have the merit of grouping different sites together only when the observed trajectories share some common features, preserving sample information about the temporal dynamics of the variables of interest. However, they didn't considered the spatial location of observations in the clustering process. The incorporation of the spatial correlation can be useful to correctly classify the curves. Spatial correlation was included in several clustering approaches ranging from the environmental sciences (air quality studies [Guttorp *et al.*(1994)], water temperature [Akita *et al.*(2007)]) to social and economical sciences (spatial accessibility [Jiang and Serban(2012)]). For this purposes, we compare the results obtained from the FCM procedure with a technique proposed by Jiang and Serban (2012), called Functional Spatial Clustering Model (FSCM). This latter included terms of spatial dependence in the statistical model.

2 Materials, methods, results

2.1 Remote data sensor

The mean monthly values of the chlorophyll type A concentration in the Adriatic Sea were obtained for the period 2002-2012 (11 calendar years, 132 months) by the ACRI server (http://hermes.acri.fr/) using satellite data source combining information of MERIS, SeaWiFS and MODIS optical sensors. For each month, the ACRI server provided a dataset with resolution of 192 x 240 points (4km scale and 46080), but resolution was reduced to 96 x 120 (8 km scale; 11520 values). However, only 2168 points formed the grid related to the Adriatic Sea (Figure 1).

Figure 1: Coast of the Adriatic Sea and observed points.



2.2 Functional clustering procedures

The functional data analysis approach considers the time series of data collected for each site as a continuous smooth function collected at a finite series of time points. A brief description of the FCM procedure is provided in recent studies [Haggarty *et al.*(2012), Pastres *et al.*(2011)]. Our aim is to cluster N time series. Chlorophyll type a concentrations at site i at time t can be written as

$$Y_i(t) = g_i(t) + \varepsilon_i(t), \tag{1}$$

where $i = 1, \ldots, N$, $g_i(t)$ is the true value of the i-th curve at time t and $\varepsilon_i(t)$ is the corresponding measurement error. Omitting the index from the notation, the formula can be written in vector form $Y_i = g_i + \varepsilon_i$ and ε_i are independent and normally distributed $\sim N(0, \sigma^2 I)$. The curve g_i can be can be expressed as the sum of a group effect $(\lambda_0 + \Lambda\alpha_k)$ and a random independent site effect (γ_i). FCM can be rewritten:

$$Y_i(t) = S_i(\lambda_0 + \Lambda\alpha_k + \gamma_i) + \varepsilon_i(t), \tag{2}$$

$$\varepsilon_i \sim N(0, \sigma^2 I) \text{ and } \gamma_i \sim N(0, \Gamma) \tag{3}$$

where S_i is the spline basis matrix for the i-th curve evaluated at time points $t = 1, \ldots, t$. If we write the number of basis spline functions used to estimate the curves as p and the number of groups as G, λ_0 represents the p-dimensional vector of the overall mean for all sites, while α_k is an h-dimensional vector, which represents the group effect. The p x h matrix Γ where $h \leqslant min(p, G-1)$. FCM assumed that all random site effect, λ_i, have a common covariance structure represented by Γ. The distribution of the curve Y_i related to the group k has a normal distribution:

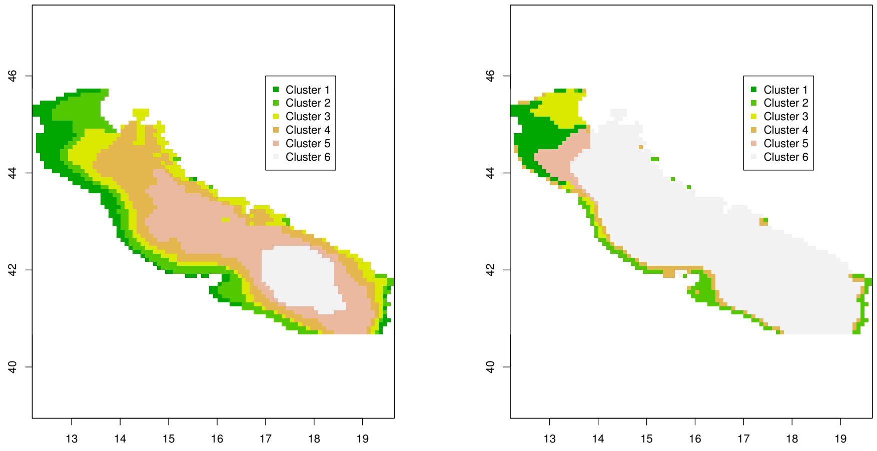
$$Y_i|(\text{curve i in cluster k}) \sim N(\lambda_0 + \Lambda\alpha_k, \Sigma_i), \tag{4}$$

where $\Sigma_i = \sigma^2 + S_i \Gamma S_i^T$. The probabilities that the curves belong to the k-th cluster was defined by a vector of probabilities π_k. The parameters that are required to be estimated to fit the FCM are $\alpha_1, \ldots, \alpha_g, \Gamma, \lambda_0, \Lambda, \sigma^2 and \pi_1, \ldots, \pi_G$. The parameter estimates can be obtained by maximum likelihood, which can be computed via the expectation-maximisation (EM) algorithm. The number of clusters is chosen by means of the Gap statistic [Tibshirani *et al.*(2001)]. In order to include a spatial dependence component in FCM, we can model the covariance matrix Γ with a function that takes in to account the spatial distribution of the sites. A procedure was proposed at this scope, called "Functional Spatial Clustering Model" (FSCM) was procedure proposed by Jiang and Serban. For the mathematical steps concerning the FSCM algorithm we suggests to refer to the paper of Jiang and Serban [Jiang and Serban(2012)]. The comparison between the two methods (FCM and FSCM) was assessed by the spatial representation of the estimated clusters. FSCM results were obtained by a free ad-hoc C++ program written by the authors. FCM analysis and graphs were performed by R [R Core Team(2013)].

2.3 Results

Chlorophyll type A concentrations appeared skewed and it was log-normal distributed; a logarithmic transformation was applied to the raw data[Arun Kumar *et al.*(2014)]. The time series of chlorophyll type-a concentrations in each site was modelled by a 6 basis of B-splines (1/2 per year) in order to take to account only the main trend over the hole period. The Gap-statistic detected the presence of six clusters and we estimated this latter by means FCM and FSCM procedures. The spatial representation of the clusters produced for both techniques (FCM and FCSM) were represented in Figure 2. The six clusters was labelled as 1,2,3,4,5 and 6 in descending order (from curve with high values to low ones). The classification performed by both the methods reported different results, but generally the clusters were regrouped according to the distance from the coast and any possible sources of perturbations (i.e. rivers, lake, lagoon,etc...). However the two classifications appeared substantial different, in particular in the northern part of the Adriatic Sea.

Figure 2: Spatial classification based on chlorophyll type A concentrations by FCM and FSCM procedures.



(a) FCM procedure

(b) FCSM procedure

3 Discussions

This work discussed about the use of functional clustering procedures and their implementation on spatial data, such us satellite data provided by remote sensor. Functional clustering techniques seem to be very attractive and flexible for researchers. In presence of spatial data, more consistent results may be obtained including the spatial pattern. The classification based on FCM and FSCM procedures showed the presence of clusters which were generally distributed according to the distance from the coast. But if we considered the FCSM procedure, the clusters were aggregated in a more appropriated way, defining, for example, a large cluster in front of the Po and Adige rivers, that reported the highest values of concentrations. These results was consistent with the presence of important rivers in the northern part that contributed to increase the concentration of nutrients and, consequently, the level of plankton concentrations [Lazzari *et al.*(2012)]. In addition, the clusters obtained by the FSCM procedure reflected the presence of a cyclonic marine current in the northern part of the Adriatic Sea[Giani *et al.*(2012)].

References

[Akita *et al.*(2007)] Akita, Y., Carter, G., and Serre, M. L. (2007). Spatiotemporal nonattainment assessment of surface water tetrachloroethylene in New Jersey. *Journal of Environmental Quality*, **36**(2), 508–20.

[Arun Kumar *et al.*(2014)] Arun Kumar, S., Babu, K., and Shukla, A. (2014). Comparative analysis of chlorophyll-a distribution from seawifs, modis-aqua, modis-terra and meris in the arabian sea. *Marine Geodesy*, (just-accepted), 00–00.

[Carvalho *et al.*(2011)] Carvalho, L., Scott, E. M., Codd, G. A., Davies, P. S., Tyler, A. N., *et al.* (2011). Cyanobacterial blooms: statistical models describing risk factors for national-scale lake assessment and lake management. *Science of the Total Environment*, **409**(24), 5353–5358.

[Ferguson *et al.*(2008)] Ferguson, C., Carvalho, L., Scott, E., Bowman, A., and Kirika, A. (2008). Assessing ecological responses to environmental change using statistical models. *Journal of Applied Ecology*, **45**(1), 193–203.

[Giani *et al.*(2012)] Giani, M., Djakovac, T., Degobbis, D., Cozzi, S., Solidoro, C., and Umani, S. F. (2012). Recent changes in the marine ecosystems of the northern adriatic sea. *Estuarine, Coastal and Shelf Science*, **115**, 1–13.

[Guttorp *et al.*(1994)] Guttorp, P., Meiring, W., and Sampson, P. D. (1994). A space-time analysis of ground-level ozone data. *Environmetrics*, **5**, 241–254.

[Haggarty *et al.*(2012)] Haggarty, R., Miller, C., Scott, E., Wyllie, F., and Smith, M. (2012). ciao. *Environmetrics*, **23**(8), 685–695.

[Ignaccolo *et al.*(2008)] Ignaccolo, R., Ghigo, S., and Giovenali, E. (2008). Analysis of air quality monitoring networks by functional clustering. *Environmetrics*, **19**(7), 672–686.

[Jacques and Preda(2013)] Jacques, J. and Preda, C. (2013). Funclust: A curves clustering method using functional random variables density approximation. *Neurocomputing*, pages 164–171.

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

[Jiang and Serban(2012)] Jiang, H. and Serban, N. (2012). Clustering random curves under spatial interdependence with application to service accessibility. *Technometrics*, **54**(2), 108–119.

[Lazzari *et al.*(2012)] Lazzari, P., Solidoro, C., Ibello, V., Salon, S., Teruzzi, A., Béranger, K., Colella, S., and Crise, A. (2012). Seasonal and inter-annual variability of plankton chlorophyll and primary production in the mediterranean sea: a modelling approach. *Biogeosciences*, **9**(1).

[Pastres *et al.*(2011)] Pastres, R., Pastore, A., and Tonellato, S. (2011). Looking for similar patterns among monitoring stations. venice lagoon application. *Environmetrics*, **22**(6), 712–724.

[R Core Team(2013)] R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Ramsay *et al.*(2011)] Ramsay, J., Ramsay, T., and Sangalli, L. (2011). Spatial functional data analysis. In F. Ferraty, editor, *Recent Advances in Functional Data Analysis and Related Topics*, Contributions to Statistics, pages 269–275. Physica-Verlag HD.

[Tibshirani *et al.*(2001)] Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society Series B*, **63**, 411–423.