Clustering of environmental functional data

Andrea Pastore and Stefano Tonellato
Università Ca' Foscari Venezia, Department of Economics, stone@unive.it

Roberto Pastres
Università Ca' Foscari Venezia, Department of Environmental Sciences, Informatics and Statistics

Abstract: Often environmental scientists face the problem of clustering different sites, areas or stations in a monitoring network in order to identify some common features among data collected at different locations. In a functional data analysis approach, each location can be seen as a specific individual, on which noisy observations from a continuous random function are collected at discrete times. The definition of suitable models for samples of such functional observations, can provide useful insights about the dynamics of the variables of interest. In such a context, a cluster can be defined as a group of individuals (i.e. locations, stations, areas etc.) where the observed trajectories share common salient features. We present some classification results in a water quality network and focus on some open issues.

Keywords: Cluster analysis, Functional data, Water quality.

1 Introduction

Often environmental scientists face the problem of clustering different sites, areas or stations in a monitoring network in order to identify some common features among data collected at different locations. It is a common practice to use standard classification methods such as k-means or hierarchical classifiers, by considering temporal (e.g. annual) averages of one or more variables measured at each site. This is clearly a limitation, since the whole information about the dynamics of the observed variables is lost. Moreover, such methods do not take into account the uncertainty that should characterise any partition based on sample information. The combination of functional data analysis (Ramsey and Silverman, 2005; Ferraty and Vieu, 2006) and probabilistic cluster analysis methods (Banfield and Raftery, 1993), which allow one to estimate the probability that a given object belongs to a given group, represents, in our opinion, an important step towards a better understanding of environmental data.

Here, we shall provide a classification of the sites of a water quality monitoring network located in Venice Lagoon, by using a trophic index (TRIX, Vollenweider et al. 1998). We apply a classification method based on functional data analysis,

introduced by James and Sugar (2003), which allows to take into account sample information about the temporal dynamics of the variable of interest, as well as quantify the uncertainty in the partition.

2 Classification of functional data

Grossly speaking, functional data analysis methods look at time series of data collected on each individual, in our case on each site, as measurements of a continuous function taken at a finite number of instants and corrupted by noise. Any observed trajectory can be seen as the noisy measurement of an unobservable curve, which is the object of interest. Following the classification method proposed by James and Sugar (2003), data are modelled as a mixture of Gaussian spline regressions, where each mixture represents a model for a specific cluster. Spline coefficients are the sum of a deterministic term, which represents the cluster effect on the mean of the variable, and a stochastic component, which represents an individual (site-specific) random effect. Parameters can be estimated via maximum likelihood. Mixture weights can be seen as prior membership probabilities of any site. The application of Bayes theorem, after plugging maximum likelihood estimates into the model, leads to *posterior* membership probabilities for each site in the network (Banfield and Raftery, 1993). A generic monitoring station is then allocated to the group which encompasses it with highest posterior probability. The number of groups, i.e. the number of mixture components, is selected by using BIC criterion.

3 Site classification in terms of water quality

The data. Venice Lagoon, with an extension of about 500 km^2 is one of the largest wet areas in Europe. It is a shallow water system with average depth of one meter crossed by a network of canals which determine a rather complex hydrodynamic circulation. As other European estuaries and lagoons, it is classified as a transition water body. Overall, the tributary discharge is about 30 m^3/sec. Rivers bring in freshwater, nutrients and pollutants, whereas tides bring in marine water. Internal hydrodynamics disperse the pollutants and, eventually, dissolved compounds are exported to the sea.

Data were collected at 30 monitoring sites which are shown on the map in figure 1. The first chacter of site labels identifies a particular category: letter B means that the site is located in a shallow area, letter C indicates that the site is located on a canal and letter M identifies sites located in the coastal area, next to the Lagoon. The same figure shows (in blue) the network of canals which are very influential in the Lagoon hydrodynamics and must be taken into account when interpreting classification results. Measurements were repeated in time at 38 subsequent instants (in the period ranging from January 16th, 2001 to December 17th, 2003) corresponding to neap tides. We considered a subset of the variables which have been

monitored, namely: chlorophyll-a (CHL-a), dissolved oxygen (DOX), total nitrate (NIT) and reactive phosphorus (PPO4). CHL-a and DOX can be taken as proxies for actual primary production. Even though in shallow lagoons and coastal areas, including the lagoon of Venice, macroalgae and seagrasses usually account for the major fraction of the production, phytoplanctonic production is extremely important, since the planktonic compartment represents a source of food for fish juvenils and shellfish. The concentrations of dissolved oxygen, total nitrate and reactive phosphorus provide information about the trophic potential of a water body. In fact, an excess of these chemicals could enhance the primary production of phytoplankton and macroalgae and cause the symptoms of eutrophication, as happened in Venice Lagoon in the 1970ies an 1980ies.

TRIX. TRIX is a widely used trophic index for marine coastal waters proposed by Vollenweider et al. (1998). It considers both factors that are direct expressions of productivity (chlorophill-a and dissolved oxigen) and nutritional factors (nitrogen and phosphorous). Some alternative formulations have been proposed. Here we consider the following one:

$$TRIX = \frac{\log_{10}(\text{CHL-a} \times \text{DOX} \times \text{NIT} \times \text{PPO4}) + 1.5}{1.2},\ TRIX \in [1, 10]$$

where DOX is the absolute deviation of oxygen from saturation and the other symbols indicate the concentrations, in mg/m^3, of the compounds mentioned above. The values of TRIX range from 1 to 10: low values indicate oligotrophy (scarcity of nutrients); high values indicate hypertrophy (exeedence of nutrients). A water body in a good trophic state should not exceed the value 5.

4 Results

In our application we identified two groups: the first one characterised by good values of TRIX and the second one exhibiting high TRIX values for the most part of the sample period. Figure 1 shows the raw data, the group specific mean trajectories and individual mean trajectories. The same figure shows a map where two spatial clusters are clearly identified. It is worth to note, however, that the posterior membership probabilities of sites $B11$, $C06$, $C01$, and $C05$ range between 0.54 and 0.87, indicating a rather strong uncertainty in their allocation to one of the two groups (for the remaining sites, the allocation probability was always higher than or equal to 0.99).

An explicit treatment of spatial dependence has not yet been developed for the class of models we have considered here. Important advances in this direction have been made in the Bayesian nonparametrics literature and research in this field is under way.

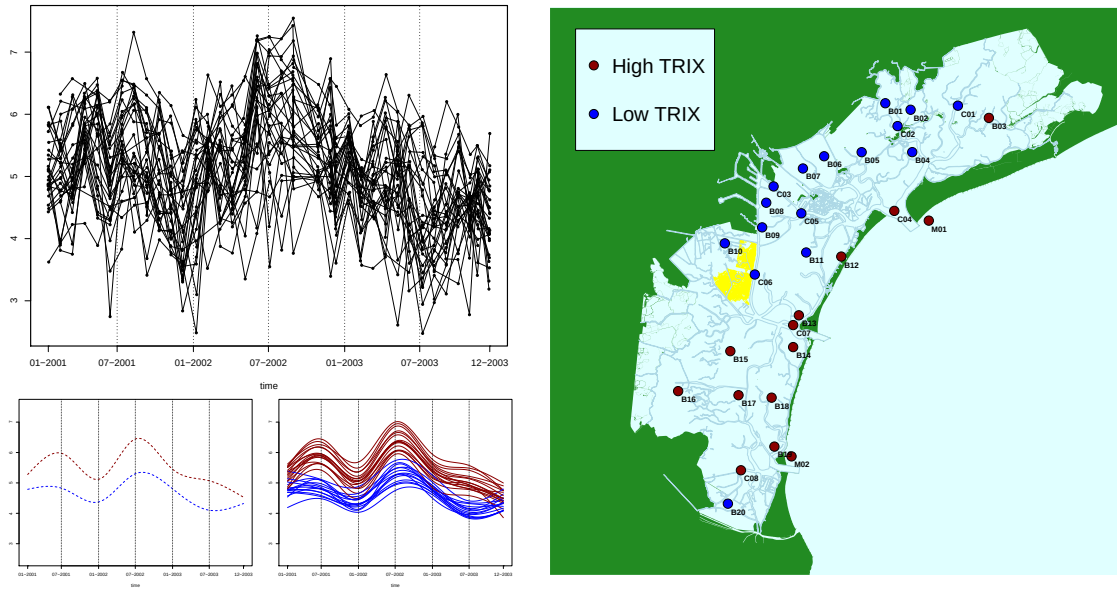

Figure 1: Plots of raw data, group specific mean trajectories, individual mean trajectories and map of monitoring sites (red="high TRIX"; blue="low TRIX").

References

Banfield J. D. and Raftery A. E. (1993) Model-Based Gaussian and Non-Gaussian Clustering, *Biometrics*, 49, 803-821.

Ferraty F. and Vieu P. (2006), *Nonparametric Functional Data Analysis: Theory and Practice*, Springer, New York.

James G. M. and Sugar C. A. (2003), Clustering for Sparsely Sampled Functional Data, *Journal of the American Statistical Association*, 98, 397-408.

Ramsey J. O. and Silverman B. W. (2005) *Functional Data Analysis*, Springer, New York.

Vollenweider R. A. , Giovanardi F., Montanari G. and Rinaldi A. (1998) Characterization of the trophic conditions of marine coastal waters with special reference to the NW Adriatic Sea: Proposal for a trophic scale, turbidity and generalized water quality index. *Environmetrics*, 9, 329-357.