Spatially correlated functional data[1]

Jorge Mateu
Department of Mathematics, University Jaume I, Castellon, Spain.
mateu@mat.uji.es

Abstract: Observing complete functions as a result of random experiments is nowadays possible by the development of real-time measurement instruments and data storage resources. Functional data analysis deals with the statistical description and modeling of samples of random functions. Functional versions for a wide range of statistical tools have been recently developed. Here we are interested in the case of functional data presenting spatial dependence, and the problem is handled from the geostatistical and point process contexts. Functional kriging prediction and clustering are developed. Additionally, we propose functional global and local marked second-order characteristics.

Keywords: Basis functions, Functional clustering, Functional kriging, LISA functions, Trace-variogram

1 Introduction

In many fields of environmental sciences the observations consist of samples of random functions. Since the early nineties, Functional Data Analysis (FDA) has been used to model this type of data (Ramsay and Dalzell, 1991). From the FDA point of view, each curve corresponds to one observation, that is, the basic unit of information is the entire observed function rather than a string of numbers. Functional versions for many branches of statistics have been given (Ramsay and Silverman, 2005).

The standard statistical techniques for modeling functional data are focused on independent functions. However, in several disciplines of applied sciences there exists an increasing interest in modeling correlated functional data: this is the case when samples of functions are observed over a discrete set of time points (*temporally correlated functional data*) or when these functions are observed in different sites of a region (*spatially correlated functional data*). In these cases some statistical methods for modeling correlated variables have been adapted to the functional context.

We can define a spatial functional process as $\left\{\boldsymbol{\chi}_s, s \in D \subset \mathbb{R}^d\right\}$ where s is a generic data location in the d-dimensional Euclidean space, the set $D \subset \mathbb{R}^d$ can be fixed or random, and $\boldsymbol{\chi}_s$ are functional random variables, defined as random elements taking values in an infinite dimensional space (or functional space). Typically $\boldsymbol{\chi}_s$

[1]Research partially supported by the Spanish Ministry of Education and Science through grant MTM2010-14961, and Bancaja grant P1-1B2008-27.

is a real function from $[a,b] \subset \mathbb{R}$ to $\mathbb{R}$. The nature of the set D allows to classify spatial functional data. Geostatistical functional data appear when D is a fixed subset of $\mathbb{R}^d$ with positive volume and n points $s_1, \ldots, s_n$ in D are chosen to observe the random functions $\boldsymbol{\chi}_{s_i}$, $i = 1, \ldots, n$. We say that we have a functional marked point pattern, when a complete function is observed at each point generated by a standard point process.

We focus here in the methodological issues opened around the geostatistical problems of spatial prediction and classification of functional data following Delicado *et al.* (2010) and Giraldo *et al.* (2010, 2011). In addition, and following Comas *et al.* (2011) and Mateu *et al.* (2008), we also present some issues concerning second-order characteristics in functional marked point patterns.

2 Materials and Methods

2.1 Geostatistical functional context

Let $\left\{\boldsymbol{\chi}_s(t), t \in T,\ s \in D \subset \mathbb{R}^d\right\}$ be a random function defined on some compact set T of $\mathbb{R}$. Assume that we observe a sample of curves $\boldsymbol{\chi}_{s_i}(t)$, for $t \in T$ and $s_i \in D, i = 1, \cdots, n$. It is usually assumed that these curves belong to a separable Hilbert space $\boldsymbol{H}$ of square integrable functions defined on T. We assume for each $t \in T$ that we have a second-order stationary and isotropic random process, that is, the mean and variance functions are constant and the covariance depends only on the distance among sampling sites. Formally, we assume that:

- $E(\boldsymbol{\chi}_s(t)) = m(t)$, for all $t \in T, s \in D$.
- $\mathrm{Cov}(\boldsymbol{\chi}_{s_i}(t), \boldsymbol{\chi}_{s_j}(u)) = C(h; t, u),\ s_i, s_j \in D, t, u \in T,\ h = \|s_i - s_j\|$, the Euclidean distance. If $t = u$, $\mathrm{Cov}(\boldsymbol{\chi}_{s_i}(t), \boldsymbol{\chi}_{s_j}(t)) = C(h; t)$.
- $\frac{1}{2}\mathrm{V}(\boldsymbol{\chi}_{s_i}(t) - \boldsymbol{\chi}_{s_j}(u)) = \gamma(h; t, u),\ s_i, s_j \in D, t, u \in T,\ h = \|s_i - s_j\|$. If $t = u$, $\frac{1}{2}\mathrm{V}(\boldsymbol{\chi}_{s_i}(t) - \boldsymbol{\chi}_{s_j}(t)) = \gamma(h; t)$.

The function $\gamma(h; t)$, as a function of h, is called the variogram of $\boldsymbol{\chi}(t)$.

We can use a family of point-wise linear predictors for $\boldsymbol{\chi}_{s_0}(t)$, $t \in T$, given by

$$\hat{\boldsymbol{\chi}}_{s_0}(t) = \sum_{i=1}^{n} \lambda_i(t) \boldsymbol{\chi}_{s_i}(t),\ \ \lambda_1(t), \ldots, \lambda_n(t) : T \to \mathbb{R}, \tag{1}$$

For each $t \in T$, the predictor (1) has the same expression as an ordinary kriging predictor. This predictor is called the point-wise linear predictor for functional data. This modeling approach is consistent with the functional linear concurrent model (FLCM) as mentioned in Ramsay and Silverman (2005) in which the influence of each covariate on the response is *simultaneous* or *point-wise*. In our context, the covariates are the observed curves at n sites of a region and the functional response is an unobserved function on an unsampled location. Consequently, the objective function is

$$E\|\hat{\boldsymbol{\chi}}_{s_0}(t) - \boldsymbol{\chi}_{s_0}(t)\|^2 = \int_T E\left(\hat{\boldsymbol{\chi}}_{s_0}(t) - \boldsymbol{\chi}_{s_0}(t)\right)^2 dt.$$

The predictor (1) is unbiased if $E(\hat{\boldsymbol{\chi}}_{s_0}(t)) = m(t)$, for all $t \in T$, that is, if $\sum_{i=1}^{n} \lambda_i(t) = 1$ for all $t \in T$. In this case $E\left(\hat{\boldsymbol{\chi}}_{s_0}(t) - \boldsymbol{\chi}_{s_0}(t)\right)^2 = V\left(\hat{\boldsymbol{\chi}}_{s_0}(t) - \boldsymbol{\chi}_{s_0}(t)\right)$.

We then present an approach for spatial prediction based on the functional linear point-wise model adapted to the case of spatially correlated curves. First, a smoothing process is applied to the curves by expanding the curves and the functional parameters in terms of a set of basis functions. The number of basis functions is chosen by cross-validation. Then, the spatial prediction of a curve is obtained as a point-wise linear combination of the smoothed data. The prediction problem is solved by estimating a linear model of coregionalization to set the spatial dependence among the fitted coefficients. We extend an optimization criterion used in multivariable geostatistics to the functional context. We also extend cokriging analysis and multivariable spatial prediction to the case where the observations at each sampling location consist of samples of random functions, that is, we extend two classical multivariable geostatistical methods to the functional context. Our cokriging method predicts one variable at a time as in a classical multivariable sense, but considering as auxiliary information curves instead of vectors. We also propose an extension of multivariable kriging to the functional context by defining a predictor of a whole curve based on samples of curves located at a neighborhood of the prediction site. In both cases a non-parametric approach based on basis function expansion is used to estimate the parameters, and we prove that both proposals coincide when using such an approach.

Finally, noting that classification problems of functional data arise naturally in many applications, we present methods to detect groups when the functional data are spatially correlated. Our methodology allows to find spatially homogeneous groups of sites when the observations at each sampling location consist of samples of random functions. In univariable and multivariable geostatistics various methods of incorporating spatial information into the clustering analysis have been considered. Here we extend these methods to the functional context in order to fulfill the task of clustering spatially correlated curves. In our approach we initially use basis functions to smooth the observed data, and then we weight the dissimilarity matrix among curves by either the trace-variogram or the multivariable variogram calculated with the coefficients of the basis functions.

2.2 Point pattern functional context

Despite of the relatively long history of point process theory few approaches have been performed to analyse spatial point patterns where the features of interest are functions (i.e. curves) instead of qualitative or quantitative variables. Examples of point patterns with associated functional data include forest patterns where for each tree we have a growth function, curves representing the incidence of an epidemic over a period of time, and the evolution of distinct economic parameters such as unemployment and price rates all for distinct spatial locations. The study of such configurations permits to analyse the effects of the spatial structure on individual functions. For instance, the analysis of point patterns where the associated curves depend on time may permit the study of space-time interdependencies of such dynamic processes. However, note that time has not necessarily to be the dependent argument. Here point patterns with associated curves will be called functional marked point patterns.

Following Comas *et al.* (2011), we formulate and illustrate a new second order characteristic to analyse functional marked point patterns, the functional mark correlation function. This new statistic is a counterpart version of the mark correlation function where instead of a test function relating a quantitative mark we consider a test function involving two whole functions. This permits to analyse the spatial dependence in the functional marks. An additional mark configuration is considered by defining local characteristics in terms of LISA functions (Mateu *et al.*, 2008), and we exploit these functions to obtain functional information of the point pattern.

References

Comas C., Delicado P., Mateu J. (2011) A second order approach to analyse spatial point patterns with functional marks, *Test*, DOI: 10.1007/s11749-010-0215-1.

Delicado P., Giraldo R., Comas C., Mateu J. (2010) Statistics for spatial functional data: some recent contributions, *Environmetrics*, 21, 224-239.

Giraldo R., Delicado P., Mateu J. (2010) Continuous time-varying kriging for spatial prediction of functional data: An environmental application, *Journal of Agricultural, Biological, and Environmental Statistics*, 15, 66-82.

Giraldo R., Delicado P., Mateu J. (2011) Ordinary kriging for function-valued spatial data, *Environmental and Ecological Statistics*, DOI: 10.1007/s10651-010-0143-y.

Mateu J., Lorenzo G., Porcu E. (2008) Detecting features in spatial point processes with clutter via local indicators of spatial association, *Journal of Computational and Graphical Statistics*, 16, 968-990.

Ramsay J., Dalzell C. (1991) Some tools for functional data analysis, *Journal of the Royal Statistical Society, Series B*, 53, 539-572.

Ramsay J., Silverman B. (2005) *Functional data analysis*, Second edition, New York: Springer.