

Spatio-Functional data Analysis: clustering methods for models discovering

Elvira Romano ¹

Dipartimento di Studi Europei e Mediterranei, Facoltà di Studi Politici, Seconda Università degli Studi di Napoli, e-mail: elvira.romano@unina2.it

Abstract: In this paper we propose two clustering strategies for spatially referenced functional data. Both are model based approaches since a functional linear model is supposed to generate the elements of each cluster. A first proposal considers the problem of clustering curves that are realization of a random stochastic process. In this case functional sub-processes are discovered and their generative models are estimated in optimal unsampled locations of the space. A second one achieves simultaneously an optimal partition of the data and functional linear models which incorporate the spatial interaction among different functional variable in each cluster.

Keywords: Functional data Analysis , Geostatistical functional data, Clustering, Spatio-Functional Models

1. Introduction

There is a large number of applicative fields like meteorology, geology, agronomy, ecology, where data are curves observed in spatial varying way. This is leading, in these years, to the development of a new branch in statistics: Spatio-Functional Data Analysis (Ramsay, 2008). In this paper we focus on clustering methods for geostatistical functional data which are a kind of spatio-functional data (Dedicado et al., 2009). Clustering approaches in this framework can be categorized in functional methods and spatial methods. Following Functional Data Analysis aims, the first ones take into account only the functional nature of the data (Romano E., 2007); on the contrary the second ones are based on the characteristics of the space (Blekas et al., 2008). We propose two distinct clustering strategies according to a spatio-functional point of view.

A first approach is a special case of Dynamic Clustering Algorithm (Diday, 1971) based on an optimization criterion that minimizes the spatial variability among the curves in each cluster (Romano et al., 2009). The centroids of the clusters, are estimated curves in unsampled area of the space that summarizes spatio-functional behaviors. It is a modeling based approach since we suppose that the curves of each cluster are generated according to a spatio-functional linear model detected through a kriging-based spatial prediction.

The second one (Romano and Balzanella, 2009) is a clusterwise linear regression approach that attempts to discover spatio-functional linear regression models with two functional predictors an interaction term, and with spatially correlated residuals. This approach is such to establish a spatial organization in relation to the interaction among different functional data. The algorithm is a k-means clustering with a criterion based on the minimization of the squared residuals instead of the classical within cluster dispersion.

¹Address of correspondence: Via del Setificio 15, 81100 Caserta

In the following sections we introduce briefly spatio-functional data structure and the two procedures.

2. The methods

We adopt this general framework for the first proposal.

Let $\{\chi_s : s \in D \subset R^d\}$ be a random field where observations are functions defined on some compact set T of R . We assume to observe a sample of curves $\chi_{s_1}, \dots, \chi_{s_n}$ defined on T where s is a generic data location in the d -dimensional Euclidean space and the set $D \subset R^d$ is a fixed subset of R^d with positive volume. We assume for each $t \in T$ 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 between sampling points. Formally, we assume that:

- $E(\chi_s(t)) = m(t)$, for all $t \in T$, $s \in D$.
- $Cov(\chi_s(t), \chi_s(u)) = \sigma(t, u)$, $t, u \in D$ if $t = u$, $V(\chi_s(t)) = \sigma^2(t)$.
- $Cov(\chi_{s_i}(t), \chi_{s_j}(u)) = C(h_{ij}; t, u)$ where $h_{ij} = \|s_i - s_j\|$ and all $s_i, s_j \in D$ if $t = u$, $Cov(\chi_{s_i}(t), \chi_{s_j}(t)) = C(h, t)$.
- $\frac{1}{2}V(\chi_{s_i}(t), \chi_{s_j}(u)) = \gamma(h_{ij}; t, u)$ where $h_{ij} = \|s_i - s_j\|$ and all $s_i, s_j \in D$ if $t = u$, $\frac{1}{2}V(\chi_{s_i}(t), \chi_{s_j}(t)) = \gamma(h, t)$.

The function $\gamma(h, t)$ as function of h is called semivariogram of $\chi(t)$.

For the second one we deal with more specific framework since we assume to observe in data location $s_1, \dots, s_n$, the realization $\chi_{s_1}, \dots, \chi_{s_n}$ and $\theta_{s_1}, \dots, \theta_{s_n}$ of two random process $\chi_s(t)$ and $\theta_s(t)$ defined on T .

2.1 Dynamic Clustering for spatio-functional data

Our first proposal is to partition, through a Dynamic clustering algorithm, the random field $\{\chi_s : s \in D \subset R^d\}$ into a set of C clusters such that the obtained clusters contain spatially related curves. Dynamic clustering algorithm optimizes a criterion of best fitting between a partition P_c of a set of objects in C clusters and the way to represent the classes. Since we assumed that data are generated from a functional linear concurrent model (Ramsay and Silverman, 2005) we advise to optimize the following criterion:

$$\Delta(P, G) = \sum_{c=1}^C \sum_{i \in P_c} \int_T V \left(\chi_{s_i}(t) - \sum_{i=1}^{n_c} \lambda_i \chi_{s_i}(t) \right) dt \quad u.c. \sum_{i=1}^{n_c} \lambda_i = 1 \quad (1)$$

where n_c is the number of the elements in each cluster and the prototype $\chi_{s_c} = \sum_{i=1}^{n_c} \lambda_i \chi_{s_i}(t)$ is an ordinary kriging predictor for curves in the clusters. According to this criterion the kriging coefficients represent the contribute of each curve to the prototype estimate in an optimal location s_c . Thus, the parameters to estimate are: the kriging coefficients, the spatial location of the prototypes, the residuals spatial variance for each cluster. For a fixed value of s_c , this is a constrained minimization problem, due to the unbiasedness constraint. So it is necessary to solve a linear system by means of Langrange multiplier method. In this paper we refer to the method proposed by (Dedicado et al., 2007). We suppose to observe the spatio-functional data on a regular spatial grid where there are

unsampled locations. These last ones are candidates to be the location of the prototype. Since the location of the prototypes is a parameter to estimate, we evaluate, for each cluster, local kriging on unsampled locations of a regular spatial grid in order to obtain the best representative predictor. The prototypes are chosen as the best predictors in terms of the best spatio-functional fitting among the set of estimates on the unsampled locations of the grid. Once we have estimated the prototypes we allocate each new curve to the cluster according to a minimum-spatial variability rule. It is an euclidean weighted distance, where the weight accounts for the spatial component. Especially these weights are computed by averaging a set of kriging coefficients of the curves at a fixed distance h . The value of h is the spatial distance among the candidate curve and the prototype in the previous cluster. This leads to assign each curve to the cluster such that there is best prediction. Applying iteratively the assignment function followed by the allocation function under some conditions the algorithm converges to a stationary value.

2.2 Clusterwise linear regression for spatio-functional data

Clusterwise linear regression (Spaeth, 1979) supposes that points of each cluster are generated according to some linear regression relation. Given a spatio-functional dataset $(\chi_s(t), \theta_s(t), Y(s))_{s=1\dots S}$ we consider the following functional regression model with scalar response (Bar-Hen et al., 2008):

$$Y(s) = \mu + \langle A; \chi_s \rangle + \langle B; \theta_s \rangle + \langle C\theta_s; \chi_s \rangle + \epsilon_s \quad (2)$$

where

$$\langle A; \chi_s \rangle = \int_{[0,T]} A(t) \chi_s(t) dt \quad \langle B; \theta_s \rangle = \int_{[0,T]} B(t) \theta_s(t) dt \quad (3)$$

are the linear terms in χ_s, θ_s and

$$\langle C\theta_s; \chi_s \rangle = \int \int_{[0,T]^2} C(t, u) \chi_s(t) \theta_s(u) dt du \quad (4)$$

is the bilinear term and ϵ_s is a spatial stationary random field with spatial correlation function ρ_s .

Our aim is to find simultaneously an optimal partition of $(\chi_s(t), \theta_s(t), Y(s))_{s=1\dots S}$ in C clusters and regression coefficients within each cluster which maximize the overall fit. We assume that exists a group-variable $\mathcal{G} : \chi_s \mapsto \{1, \dots, C\}$, such that

$$E(Y(s) | \chi_s = \chi_s, \theta_s = \theta_s, \mathcal{G} = c) = \mu^c + \langle A^c; \chi_s \rangle + \langle B^c; \theta_s \rangle + \langle C^c \theta_s; \chi_s \rangle \quad (5)$$

where $\{\mu^c, A^c, B^c, C^c\}_{c=1,\dots,C}$ are the regression estimated functions given by the generalized least squares criterion. If S data points have been recorded, the clusterwise linear regression algorithm finds simultaneously an optimal partition of the S points and the regression models $\{\mu^c, A^c, B^c, C^c\}_{c=1,\dots,C}$ associated to each cluster, which optimize the criterion :

$$\Delta^s(P, G) = \sum_{c=1}^C \sum_{i \in P_c} [Y_i(s) - (\mu^c + \langle A^c; \chi_s \rangle + \langle B^c; \theta_s \rangle + \langle C^c \theta_s; \chi_s \rangle)]^2 \quad (6)$$

that is minimizing the sum of the squares errors over the C clusters SSE_c . Starting with an initial partition $P^0 = \{P_1^0, \dots, P_C^0\}$ the algorithm constructs iteratively a sequence of partitions and regression coefficients as follows:

- for each $c = 1, \dots, C$, $\{\mu^c, A^c, B^c, C^c\}_{c=1, \dots, C}$ are computed by generalized least squares estimators of the functional linear regression;
- let $(P_l, \{\mu^c, A_l^c, B_l^c, C_l^c\}_{c=1}^C)$ be known. For each cluster $c = 1, \dots, C$

$$P_{l+1}^c = \{(Y_i(s), \chi_s(t), \theta_s(t)) | SSE_c < SSE'_c, \forall c \neq c'\} \quad (7)$$

The sequence $(P_l, \{\mu^c, A_l^c, B_l^c, C_l^c\}_{c=1}^C)_{l \geq 0}$ is such that $\forall l \geq 0$

$$\Delta^s(P_l, \{\mu^c, A_l^c, B_l^c, C_l^c\}_{c=1}^C) \geq \Delta^s(P_{l+1}, \{\mu^c, A_{l+1}^c, B_{l+1}^c, C_{l+1}^c\}_{c=1}^C) \quad (8)$$

This leads the algorithm to the convergence.

3. Conclusions

From the testing of the methods on real datasets we can conclude that they seems promising. We intend to perform more extensive simulation studies, namely to investigate the performance on larger data sets with contamination in order to preserve the results from possible harmful effects of the data.

References

- Bar Hen A., Bel L., Cheddadi R. and Petit R. (2008) Spatio-temporal functional regression on paleoecological data, *Functional and Operatorial Statistics*, 54–56.
- Blekas K., Nikou C. and Galatsanos N. and Tsekos N.V. (2007) Curve clustering with spatial constraints for analysis of spatiotemporal data, in: *ICTAI '07: Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence*, IEEE Computer Society, Washington, DC, USA, 529–535.
- Delicado P., Giraldo R., Comas C. and Mateu. J. (2009) Statistics for spatial functional data, in: *Environmetrics*.
- Delicado P., Giraldo R. and Mateu. J. (2007) Geostatistics for functional data: An ordinary kriging approach, in: *Technical Report*.
- Ramsay J. and Silverman B. (2005) *Functional Data Analysis*, Springer, New York.
- Romano E. (2007) *Dynamical curves clustering with free knots spline estimation*, PHD Thesis, University Federico II Naples.
- Romano E. and Balzanella A. (2009) A clusterwise regression strategy for spatio-functional data, *Seventh Scientific Meeting of the CLAssification and Data Analysis Group of the Italian Statistical Society CLADAG*, (submitted).
- Romano E., Balzanella A. and Verde R. (2009) Clustering spatio-temporal data with a functional approach, *IFCS 2009: Book of abstract of the International Federation of Classification Societies 2009 Conference*.
- Spaeth H. (1997) Clusterwise linear regression, *Computing*, 22, 367–373.