Kriging uncertainty for functional data: a comparison study

M. Franco-Villoria[1] and R. Ignaccolo[1,*]

[1] *Department of Economics and Statistics, University of Torino (IT); maria.francovilloria@unito.it, rosaria.ignaccolo@unito.it*
[*] *Corresponding author*

Abstract. *Uncertainty evaluation for spatial prediction of curves remains an open issue in the functional data literature. We consider three different approaches that rely on semi-parametric bootstrapping, principal component analysis and classical inference for additive models respectively.*

Keywords. *Functional modelling; Functional random field; Prediction bands; Additive models; Bootstrap*

1 Introduction

Environmental data collected over time at various locations of a monitoring network can be considered as spatially dependent functional data (see e.g. the last two chapters of the book by Horvath-Kokozsca [8] or [5]). These kind of data has lead to the development of geostatistical techniques in a functional framework, i.e. ordinary [5, 7, 12] and universal [1, 11] kriging models to predict a curve at an unmonitored site, and more recently kriging with external drift [9]. The latter allows to introduce exogenous variables (both scalar and functional) in the mean function of the spatial functional process. However, uncertainty evaluation remains an open issue and we face it by considering three different approaches and illustrating their performance for spatial prediction of particulate matter (PM10) in the Piemonte region (Italy).

2 Functional Kriging with External Drift (FKED)

Assume that we observe a sample of curves Υ_{s_i}, $i = 1, \ldots, n$ taken as realizations of a functional random field $\left\{\Upsilon_s, s \in D \subseteq \mathbb{R}^d\right\}$ taking values in a separable Hilbert space of square integrable functions [5] and

consider the model

$$\Upsilon_s = \mu_s + \varepsilon_s. \tag{1}$$

The term μ_s is interpreted as a drift describing a spatial trend while ε_s represents a residual random field that is zero-mean, second-order stationary and isotropic. The drift can be expressed as

$$\mu_{s_i}(t) = \alpha(t) + \sum_p \gamma_p(t) C_{p,i} + \sum_q \beta_q(t) X_{q,i}(t) \tag{2}$$

where $\alpha(t)$ is a functional intercept, $C_{p,i}$ and $X_{q,i}$ are scalar and functional covariates at site s_i, $\gamma_p(t)$ and $\beta_q(t)$ are the covariate coefficients and $\varepsilon_{s_i}(t)$ represents the residual spatial functional process $\{\varepsilon_s(t), t \in T, s \in D\}$ at the site s_i. Once the functional regression model (2) has been fitted by means of a GAM representation (for details see [9]), the functional residuals $e_{s_i}(t) = Y_{s_i}(t) - \hat{\mu}_{s_i}(t)$ can be used to predict the residual curve at an unmonitored site s_0 via ordinary kriging for functional data [7], according to which $\hat{e}_{s_0}(t) = \sum_{i=1}^n \lambda_i e_{s_i}(t)$, with kriging coefficients $\lambda_i \in R$. More complex alternatives, where the kriging coefficients are not constant are available [9]. The prediction at the unmonitored site s_0 is obtained by adding up, as in the classical regression kriging, the two terms, i.e. $\hat{Y}_{s_0}(t) = \hat{\mu}_{s_0}(t) + \hat{e}_{s_0}(t)$.

3 Uncertainty evaluation

To evaluate the uncertainty of a predicted curve $\hat{Y}_{s_0}(t)$ at an unmonitored site s_0, we compare three different approaches. For curves predicted via the FKED model, we consider a semiparametric bootstrap for spatially dependent functional data and a principal components analysis (PCA) based bootstrap method. As an alternative, we also consider a generalized additive model to predict the curve $\hat{Y}_{s_0}(t)$, in which case standard inference results apply.

3.1 Semiparametric bootstrap for spatially dependent functional data

This bootstrapping method has been extended to the functional context following [10]. To obtain a bootstrap sample, we estimate and remove the drift μ_s following Model (2), then estimate the residuals covariance matrix through the trace-semivariogram [7] and use its Cholesky decomposition to uncorrelate the functional residuals. The B bootstrap samples are generated from the uncorrelated residuals using the smoothed bootstrap as suggested in [4], replacing the empirical distribution function by a smooth version of it to avoid appearance of repeated measures. The bootstrap samples are then fed into the FKED method to obtain B prediction curves at the unknown location. These are ordered based on their band depth [14], where the sample band depth (BD) of a curve $y(t)$ can be calculated as

$$BD_{n,2}(y) = \binom{n}{2}^{-1} \sum_{1 \le i_1 < i_2 \le n} I\{G(y) \subseteq B(y_{i_1}, y_{i_2})\}$$

i.e. the proportion of bands delimited by 2 curves containing the whole graph $G(y)$ of $y(t)$. Band depth can be modified to take into account whether a portion of the curve is in the band (see [14] for details). The lower/upper limits of a 95% prediction band are obtained by taking the pointwise (w.r.t. t) minimum/maximum of the 95% deepest curves (i.e. those closest to the center of the distribution).

3.2 PCA bootstrap

Once the drift has been estimated and removed following Model (2), a PCA analysis of the functional residuals $e_{s_i}(t)$ is carried out to rewrite $e_{s_i} = c_1^i V_1 + \ldots + c_k^i V_k$. Then B bootstrap samples are generated for each coordinate c_l, $l = 1, \ldots, k$ with empirical distribution function $F_n^{c_l}$ and the bootstrap residuals are constructed as $e_{s_i}^* = c_1^* V_1 + \ldots + c_k^* V_k$ [6]. These bootstrap replications are then used to obtain B predictions at the unknown location using the KFED. Prediction bands are obtained as in Section 3.1.

3.3 Inference based confidence intervals for a GAM model

In order to compare the bootstrapping approaches with classical inference for uncertainty evaluation, we consider a modelling strategy different from the FKED model introduced in Section 2. In this case, a smooth function of longitude, latitude and time is included in Model (2). By setting a penalized bivariate spline basis for longitude and latitude a spatial covariance structure is implicit in the model [13], allowing for spatial prediction. The model can be written in matrix form as $\hat{Y} = SY$ where $S = X(X'X + \eta P)^{-1}X'$ is the smoothing matrix, X is the design matrix, P is the penalty matrix and η is the smoothing parameter. At a new location s_0, the predicted value is given by $\hat{Y}_{s_0}(t) = S_{s_0} y$, where $S_{s_0} = X_{s_0}(X'X + \eta P)^{-1}X'$. Approximate 95% predictions bands can be calculated as [15]:

$$\hat{Y}_{s_0}(t) \pm 1.96\hat{\sigma}_\varepsilon \sqrt{1 + ||S_{s_0}||^2}. \tag{3}$$

4 Case Study

We consider the same case study as in [2, 3] and [9]. The data set consists of daily PM10 concentrations (in $\mu g/m^3$) measured from October 2005 to March 2006 by the monitoring network of Piemonte region (Italy) in 34 sites, 24 of which will be used to fit the model and the remaining 10 as validation sites. Covariates available include coordinates and altitude (scalar), daily maximum mixing height, daily total precipitation, daily mean wind speed, daily mean temperature and daily emission rates of primary aerosols (functional). The FKED model (Section 2) and the fully additive model (Section 3.3) will be fitted to these data and the three methods for uncertainty evaluation introduced here compared.

Acknowledgments. This work is part of the project *StEPhI* (http://stephiproject.it/), supported by FIRB 2012 grant (project no. RBFR12URQJ) provided by the Italian Ministry of Education, Universities and Research.

References

[1] Caballero W, Giraldo R, Mateu J (2013) A universal kriging approach for spatial functional data. *Stochastic Environmental Research and Risk Assessment* **27(7)**, 1553–1563.

[2] Cameletti M, Ignaccolo R, Bande S (2011) Comparing spatiotemporal models for particulate matter in Piemonte. *Environmetrics* **22**, 985–996.

[3] Cameletti M, Lindgren F, Simpson D, Rue H (2012) Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Adv Stat Anal* **97(2)**, 109–131.

[4] Cuevas A, Febrero M, Fraiman R (2006) On the use of the bootstrap for estimating functions with functional data. *Comput. Statist. Data Anal.* **51**, 1063–1074.

[5] Delicado P, Giraldo R, Comas C, Mateu J (2010) Statistics for spatial functional data: some recent contributions. *Environmentrics* **21**, 224–239.

[6] Fernández de Castro B, Guillas S, González Manteiga W (2005) Functional samples and bootstrap for predicting sulfur dioxide levels. *Technometrics* **47(2)**, 212–222.

[7] Giraldo R, Delicado P, Mateu J (2011) Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics* **18(3)**, 411–426.

[8] Horváth L, Kokoszka P (2012) *Inference for functional data with applications*. Springer, New York.

[9] Ignaccolo R, Mateu J, Giraldo R (2013) Kriging with external drift for functional data for air quality monitoring. *Stoch Environ Res Risk Assess* **28(5)**, 1171-1186.

[10] Iranpanah N, Mohammadzadeh M, Taylor CC (2011) A comparison of block and semi-parametric bootstrap methods for variance estimation in spatial statistics. *Comput. Statist. Data Anal.* **55**, 578–587.

[11] Menafoglio A, Secchi P, Dalla Rosa M (2013) A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. *Electronic Journal of Statistics* **7**, 2209–2240.

[12] Nerini D, Monestiez P, Manté C (2010) Cokriging for spatial functional data. *J. Multivariate Anal.* **101**, 409–418.

[13] Nychka D (2000) Spatial process estimates as smoothers. In: Schimek MG (ed.), Smoothing and Regression. *Approaches, Computation and Application* 393-424. Wiley, New York

[14] Lopez-Pintado S, Romo J (2009) On the concept of depth for Functional Data, *J. Amer. Statist. Assoc.* **104(486)**, 718–734.

[15] Ruppert D, Wand MP, Carroll RJ (2003) *Semiparametric Regression*. Cambridge University Press, New York.