Uncertainty evaluation for functional kriging: an application to the Canadian temperature data set

Rosaria Ignaccolo*
University of Torino, Torino, Italy - rosaria.ignaccolo@unito.it

Maria Franco-Villoria
University of Torino, Torino, Italy - maria.francovilloria@unito.it

Abstract

Uncertainty evaluation for functional spatial prediction remains an open issue. Prediction of a curve at an unmonitored location can be obtained using a functional kriging with external drift model that takes into account the effect of exogenous variables (either scalar or functional). To evaluate the uncertainty of a predicted curve, a semi-parametric bootstrap for spatially correlated data is adapted to functional data. Confidence bands are obtained by ordering the bootstrapped predicted curves in two different ways according to band depth and L^2 distance. The proposed approach is illustrated on a well known data set of Canadian temperature.

Keywords: bootstrap; P-spline; band depth; prediction bands.

1. Introduction

The use of geostatistical techniques for functional data was first addressed in the seminal paper by Goulard & Voltz (1993), but it has received increasing interest over the last few years, motivated by the availability of complex data sets, where data are collected over space and e.g. time or depth, giving rise to data that can be considered as spatially dependent curves (Horváth & Kokoszka (2012)). With the aim of predicting a whole curve at an unmonitored location, several authors have extended ordinary kriging to the case of functional data, where the mean function is assumed to be constant (see e.g. Delicado et al. (2010), Giraldo et al. (2011), Nerini et al. (2010)). A slightly more complex alternative that allows the mean function to depend on longitude and latitude was considered by Caballero et al. (2013) and Menafoglio et al. (2013). Further, Ignaccolo et al. (2014) proposed the so called functional kriging with external drift, that allows the mean function to depend on scalar and/or functional exogenous variables.
When predicting a curve, it is important to provide as well a measure of its uncertainty; however, literature concerning prediction uncertainty in the case of functional data is still very limited. An interesting approach to this matter, given the lack of distribution functions, is to use resampling methods. Ferraty et al. (2010) proposed the use of 'wild bootstrapping' for functional regression but with a scalar response, where the bootstrap samples are obtained multiplying the scalar errors by iid random variables constrained to satisfy certain properties. We propose a semi-parametric bootstrap approach for spatially correlated functional data that allows to obtain uncertainty bands for a predicted curve, in the case of kriging with external drift with functional response. We evaluate the performance of the proposed methodology on a well known data set.

2. Functional kriging with external drift (FKED)

Let $\Upsilon_s = \{Y_s(t); t \in T\}$ be a functional random variable observed at location $s \in D \subseteq \mathbb{R}^d$, whose realization is a function of $t \in T$, where T is a compact subset of $\mathbb{R}$. Assume that we observe a sample of curves Υ_{s_i}, for $s_i \in D$, $i = 1, \ldots, n$, that take values in a separable Hilbert space of square integrable functions. The set $\{\Upsilon_s, s \in D\}$ constitutes a functional random field or a *spatial functional process* (Delicado et al. (2010)), that can be non-stationary and whose elements are supposed to follow the model $\Upsilon_s = \mu_s + \epsilon_s$. 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. At the generic site s_i, $i = 1, \ldots, n$, and at point t, the model can be rewritten as a functional concurrent linear model $Y_{s_i}(t) = \mu_{s_i}(t) + \epsilon_{s_i}(t)$ with the drift

$$\mu_{s_i}(t) = \alpha(t) + \sum_p \gamma_p(t) C_{p,i} + \sum_q \beta_q(t) X_{q,i}(t) \tag{1}$$

where $\alpha(t)$ is a functional intercept, $C_{p,i}$ is the p-th scalar (constant-in-time) covariate at site s_i, $X_{q,i}$ is the q-th functional covariate at site s_i, $\gamma_p(t)$ and $\beta_q(t)$ are the covariate coefficients and $\epsilon_{s_i}(t)$ represents the residual spatial functional process $\{\epsilon_s(t), t \in T, s \in D\}$ at the site s_i. Once the functional regression model (1) has been fitted by means of a GAM representation (for details see Ignaccolo et al. (2014)), 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 (Giraldo et al. (2011)), 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 (Ignaccolo et al. (2014)). The prediction at the unmonitored site s_0 is obtained by adding up, as in the classical regression kriging, the two terms, i.e. $\hat{Y}_{s_0}(t) = \hat{\mu}_{s_0}(t) + \hat{e}_{s_0}(t)$, where $\hat{\mu}_{s_0}(t) = \hat{\alpha}(t) + \sum_p \hat{\gamma}_p(t) C_{p,0} + \sum_q \hat{\beta}_q(t) X_{q,0}(t)$ depends on the covariate values $C_{p,0}$ and $X_{q,0}(\cdot)$ at the site s_0.

3. Uncertainty evaluation

The following bootstrapping algorithm is an extension of the work by Iranpanah et al. (2011) to the functional workframe. It allows to obtain uncertainty bands for a predicted curve $\hat{Y}_{s_0}(t)$ at an unmonitored site s_0.

Algorithm

1. Estimate the drift μ_s following Model (1) and subtract it from the observed data to obtain the functional residuals $\epsilon_{s_i}(t) = Y_{s_i}(t) - \hat{\mu}_{s_i}(t)$.
2. Estimate the residual covariance matrix Σ by means of the trace-semivariogram (Giraldo et al. (2011)), estimated as:
$$\hat{v}(h) = \frac{1}{2\,|N(h)|} \sum_{i,j \in N(h)} \int_T \left(\epsilon_{s_i}(t) - \epsilon_{s_j}(t)\right)^2 dt$$
where $N(h) = \{(s_i, s_j) : \|s_i - s_j\| = h\}$. Once estimated, the empirical trace-semivariogram provides a cloud of points $(h_g, \hat{v}(h_g)), g = 1, \ldots, G$ to which a parametric model (e.g. exponential, spherical, Matérn) can be fitted as in classical geostatistics. Decompose the estimated covariance matrix using its Cholesky decomposition as $\hat{\Sigma} = \hat{L}\hat{L}^T$ and transform the functional residuals to make them (spatially uncorrelated):
$$\zeta_{n \times M} = (\zeta(s_1), \ldots, \zeta(s_n))' = \hat{L}^{-1}_{n \times n} \left(Y_{n \times M} - \hat{\mu}_{n \times M}\right). \quad (2)$$
3. Generate B bootstrap samples $\zeta^*(s_1), \ldots, \zeta^*(s_n)$ from $\zeta(s_1), \ldots, \zeta(s_n)$ using the smoothed bootstrap as suggested in Cuevas et al. (2006); here F_n, the empirical distribution function of $\{\zeta(s_1), \ldots, \zeta(s_n)\}$, is replaced by a smooth version $\hat{F}_n$ to avoid appearance of repeated measures. In practice a booststrap sample $\zeta^*(s_1), \ldots, \zeta^*(s_n)$ from $\hat{F}_n$ can be obtained as
$$\zeta^*_{s_i}(t_j) = \zeta^0_{s_i}(t_j) + Z_i(t_j), \qquad j = 1, \ldots, M,$$
where $\zeta^0(s_i)$ is drawn from F_n, $(Z(t_1), \ldots, Z(t_M)) \sim MVN(0_M, \kappa\Sigma_\zeta)$, Σ_ζ is the covariance matrix of $\{\zeta(t)\}$ and κ is a smoothing parameter.
4. The final bootstrap sample $Y^*_{s_1}, \ldots, Y^*_{s_B}$ is obtained using an inverse transform:
$$Y^*_{s_i}(t) = \hat{\mu}_{s_i}(t) + \hat{L}\zeta^*_{s_i}(t).$$

The bootstrap samples are then fed into the FKED method to obtain B prediction curves at the unmonitored location s_0. The 2.5^{th} and 97.5^{th} quantiles of the distribution of these B curves will determine the upper and lower limits of the uncertainty band for the predicted curve $\hat{Y}_{s_0}(t)$. To do so, the B curves need to be ordered. There is not a unique way of ordering functional data; here we consider two ordering techniques: band depth (Lopez-Pintado & Romo (2009)) and L^2 distance between curves.
The band in $\mathbb{R}^2$ delimited by $k = 2$ curves is defined as:
$$B(y_{i_1}, y_{i_2}) = \{(t, y(t)) : t \in T, min_{r=1,2} y_{i_r}(t) \leq y(t) \leq max_{r=1,2} y_{i_r}(t)\}.$$

The sample band depth (BD) of a curve $y(t)$ can be calculated as

$$BD_{n,2}(y) = \left(\begin{array}{c} n \\ 2 \end{array} \right)^{-1} \sum_{1 \leq i_1 < i_2 \leq n} I\left\{ G(y) \subseteq B(y_{i_1}, y_{i_2}) \right\}.$$

i.e. the proportion of bands delimited by 2 curves containing the whole graph $G(y) = \{(t, y(t)) : t \in I\}$ of $y(t)$. The bigger the band depth value, the more central the curve is. Band depth can be calculated for bands delimited by more than two curves, but the order induced when $k > 2$ remains stable (Lopez-Pintado & Romo (2009)), and using $k = 2$ is computationally more efficient. To avoid problems such as ties and crossing over of the curves delimiting the band, that may happen for $k = 2$, a modified version is available; the modified band depth (MBD) can be estimated as:

$$MBD_{n,2}(y) = \left(\begin{array}{c} n \\ 2 \end{array} \right)^{-1} \sum_{1 \leq i_1 < i_2 \leq n} \frac{\lambda\left(\{t \in T : min_{r=i_1,i_2} y_r(t) \leq y(t) \leq max_{r=i_1,i_2} y_r(t)\}\right)}{\lambda(T)}$$

where λ is the Lebesgue measure on T, with the difference that the MBD takes into account whether a portion of the curve is in the band (for further details see Lopez-Pintado & Romo (2009)).
Alternatively, the curves can be ordered based on their L^2 distance from the originally predicted curve $\hat{Y}_{s_0}(t)$ (Cuevas et al. (2006)), where distance between two curves x and y is calculated as:

$$||x - y|| = \left(\int_T (x(t) - y(t))^2 \, dt \right)^{1/2}. \tag{3}$$

The lower/upper limits of a 95% confidence band (based on band depth) 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). On the other hand, the 95% confidence ball (based on L^2 distance) is made of the 95% curves closest to the original FKED predicted curve.

4. Canadian temperature data set

We illustrate our proposal for uncertainty evaluation on a well known data set, the Canadian temperature data, that has been widely used in the functional data literature (see, for example, Ramsay & Silverman (2006), Giraldo et al. (2010), Menafoglio et al. (2013), Scheipl et al. (2013)). The data set consists of daily annual mean temperature collected at 35 meteorological stations in Canada's Maritimes Provinces over the period 1960-1994 (Figure 1). For further details on the dataset, the reader is referred to one of the references above. Data were converted to functional observations through smoothing by using penalized cubic B-splines with 120 basis functions and penalty parameter equal to zero. These values were chosen using functional cross-validation. We have chosen five locations at random and used them as validation stations. These can be seen in red in Figure 1(a). The FKED model, with longitude and latitude as covariates, was then fitted to the remaining 30 stations and predicted temperature curves were obtained for the 5 validation stations. For each of these sites, a bootstrap sample of predicted curves of size $B = 1000$ was obtained following the algorithm illustrated in Section 3. Band depth was calculated using the modified version *MBD*. The resulting prediction balls/bands are shown in Figure 2. The uncertainty bands are fairly narrow, as expected when observing the small variability between curves in Figure 1(b). Overall, the two uncertainty measures seem to agree well, although in some cases the confidence ball appears to be slightly narrower than the confidence band. The bands appear to be wider for Station 1 than for the remaining stations, specially in winter, suggesting greater uncertainty in that station. This may be due to the fact that this is an inland station, while the other four validation stations are closer to the coast.

5. Conclusions and future work

Spatial functional analysis provides an interesting alternative to spatio-temporal modelling, allowing to predict a whole curve taking into account exogenous covariates and the underlying spatial structure. Here, we propose a bootstrap approach to estimate uncertainty bands for a predicted curve has been proposed, and its validity has been illustrated in a well known data set. We are currently running a simulation study to

evaluate the performance of the method proposed. Numerical results will be presented at the conference.

References

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.

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.

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

Ferraty, F., Van Keilegom, I. & Vieu, P. (2010) On the validity of the bootstrap in non-parametric functional regression. Scand. J. Statist., 37, 286-306.

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

Goulard, M. & Voltz, M. (1993). Geostatistical interpolation of curves: A case study in soil science. In A. Soares (Ed.), Geostatistics Troia 92, Volume 2, pp. 805–816, Kluwer Academic, Dordrecht.

Horváth, L. & Kokoszka, P. (2012). Inference for functional data with applications. Springer, New York.

Ignaccolo, R., Mateu, J. & Giraldo, R. (2014) Kriging with external drift for functional data for air quality monitoring. Stochastic Environmental Research and Risk Assessment, 28(5), 1171-1186.

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

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

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.

Nerini, D., Monestiez, P. & Manté, C. (2010). Cokriging for spatial functional data. J. Multivariate Anal., 101, 409-418.

Ramsay, J. & Silverman, B. W. (2006) Functional Data Analysis. Springer, New York.

Scheipl, F., Staicu, A. M. & Greven, S. (2013) Functional additive mixed models. arXiv:stat/1207.5947v5

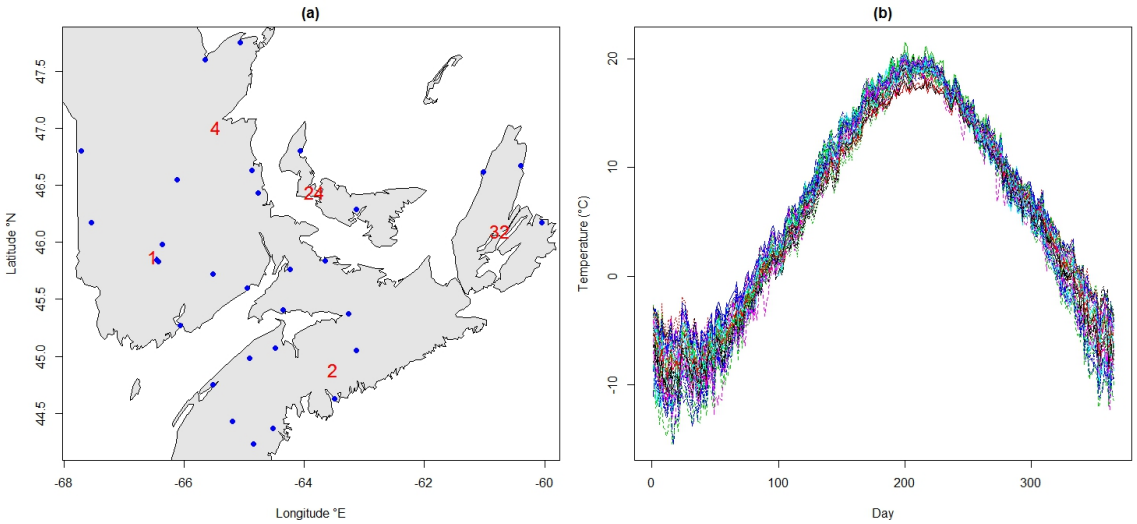

Figure 1: (a) Locations of the 35 meteorological stations in Canada's Maritimes Provinces area (validation stations numbered in red). (b) Temperature curves (raw data)

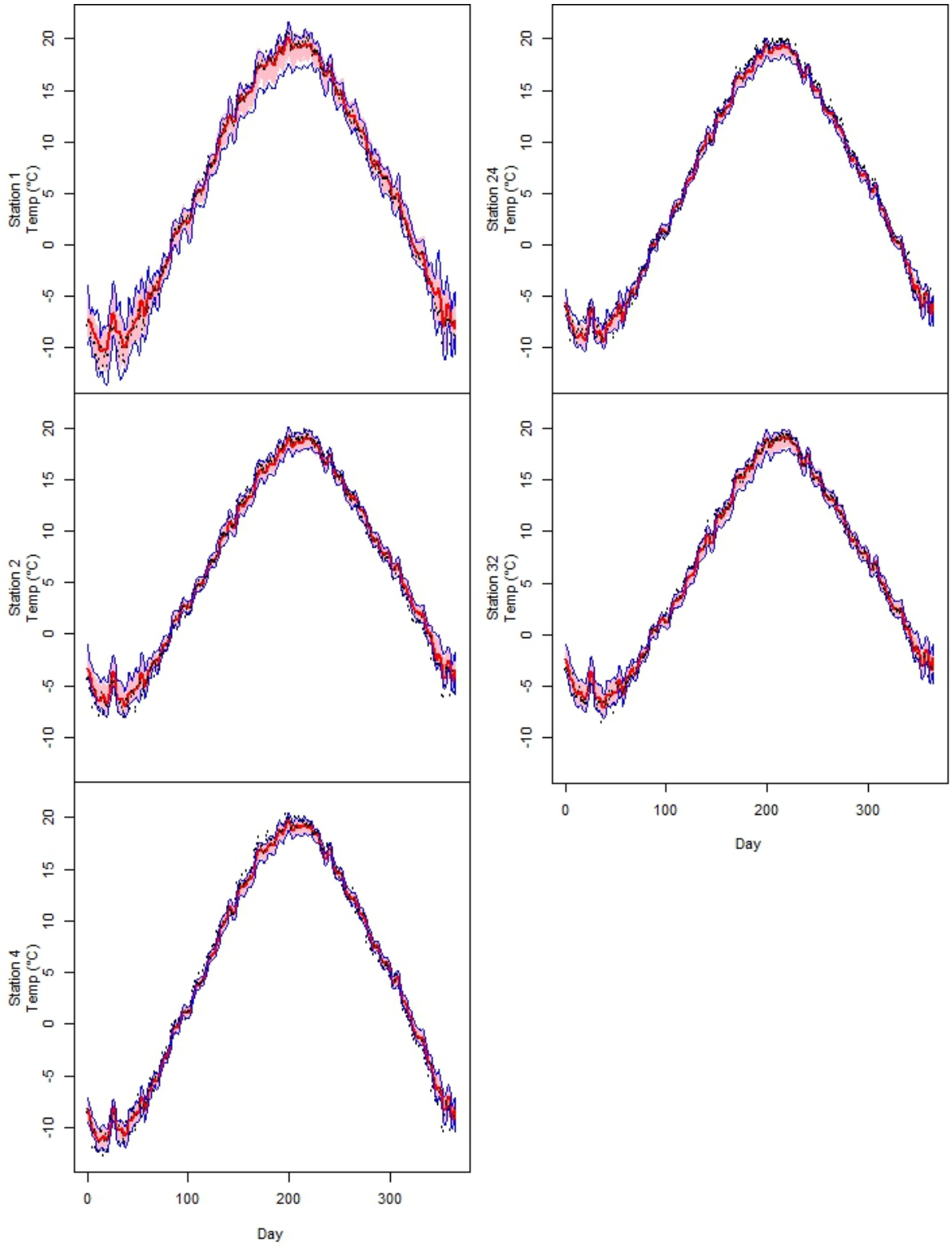

Figure 2: Original data (black dots), FKED predicted curve (red line), 95% prediction ball (pink) based on L^2 distance and 95% prediction band (blue line) based on *MBD* for validation stations