Rosaria Ignaccolo, Maria Franco Villoria
Kriging for functional data: uncertainty assessment
Editor: CUEC Cooperativa Universitaria Editrice Cagliaritana
2014
ISBN: 9788884678744

in

47th Scientific Meeting of the Italian Statistical Society P R O C E E D I N G S
1 - 6
47th SIS Scientific Meeting of the Italian Statistica Society
Cagliari
June 2014

Kriging for functional data: uncertainty assessment

Kriging per dati funzionali: valutazione dell'incertezza

Maria Franco-Villoria and Rosaria Ignaccolo

Abstract We predict a curve at an unmonitored site taking into account exogenous variables using a functional kriging model with external drift and, alternatively, an additive model with a spatio-temporal smooth term. To evaluate uncertainty of the predicted curves, a semi-parametric bootstrap approach is used for the first, while standard inference is used for the second. The performance of both approaches is illustrated on pollutant functional data.

Abstract *Allo scopo di predire una curva in un sito non monitorato, tenendo conto di variabili esogene, viene usato un modello di kriging funzionale con deriva esterna ed in alternativa un modello additivo con una componente spazio-temporale. L'incertezza viene valutata usando un approccio bootstrap semi-parametrico nel primo caso e inferenza classica nel secondo. La performance dei due approcci é illustrata su dati funzionali di inquinamento.*

Key words: kriging, P-spline, confidence bands, performance index

1 Introduction

Spatial prediction of particulate matter (PM) concentration is useful in order to assess air quality and health risk where no monitoring stations are available. The observed air pollutants and meteorological variables time series, collected at various locations of a monitoring network, can be treated as spatially dependent functional data. Several papers consider ordinary kriging models for functional data under the assumption of a constant mean (e.g. [4, 5, 9]). The more recent ones consider the

Maria Franco-Villoria, Rosaria Ignaccolo
Dipartimento di Economia e Statistica, Università degli Studi di Torino, Italy, e-mail: maria.francovilloria@unito.it, rosaria.ignaccolo@unito.it
Work partially supported by FIRB 2012 grant (project no.RBFR12URQJ) provided by the Italian Ministry of Education, Universities and Research.

mean as a function of longitude and latitude [1, 8] and (both scalar and functional) exogenous variables [6]. However, uncertainty evaluation about prediction remains an open issue, as kriging variance is constant in time. Given the difficulty to derive sampling distributions for functional data, confidence band calculation can be approached using resampling methods. We adapt the semi-parametric bootstrap approach for spatially correlated data proposed in [7] to the case of functional data. Confidence bands are obtained by ordering the bootstrapped predicted curves in two different ways, based on functional depth and on distance between curves. Alternatively, spatially dependent functional data can be modelled in a longitudinal perspective by means of an additive model that includes a smooth function of longitude and latitude, exploiting the close connection between penalized spline smoothing and kriging [12]. Uncertainty for a predicted curve can be assessed using classical inference. The two approaches are illustrated on pollutant functional data.

2 Case study: PM_{10} concentration in Piemonte

Our case study consists of daily PM_{10} concentrations (in $\mu g/m^3$) measured from October 2005 to March 2006 by the monitoring network of Piemonte region (Italy) in 24 sites (red triangles in Fig. 1(a)), and in 10 more sites used as validation stations (blue dots in Fig. 1(a)). Data were log transformed to stabilize the within-station variances and normalize the marginal distribution of PM_{10} data. The covariates considered are: 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). Since the ranges of the covariates are quite different, a standardization procedure is applied (for further details see [2]).

3 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* [4], that can be non-stationary and that is supposed to be decomposed as $\Upsilon_s = \mu_s + \varepsilon_s$. 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 site s_i, $i = 1, \ldots, n$, and point t, the model can be rewritten as a functional concurrent linear model $Y_{s_i}(t) = \mu_{s_i}(t) + \varepsilon_{s_i}(t)$ 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 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 $\varepsilon_{s_i}(t)$ represents the residual spatial functional process $\{\varepsilon_s(t), t \in T, s \in D\}$ at site s_i. Model (1) is fitted by means of a GAM representation (see [6]), and 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 a new site s_0 via ordinary kriging for functional data [5], according to which $\hat{e}_{s_0}(t) = \sum_{i=1}^{n} \xi_i e_{s_i}(t)$, with kriging coefficients $\xi_i \in \mathbb{R}$. More complex alternatives considering non-constant kriging coefficients are available [6]. Prediction at the unmonitored site s_0 is obtained by adding up the two terms, i.e. $\hat{Y}_{s_0}(t) = \hat{\mu}_{s_0}(t) + \hat{e}_{s_0}(t)$, where $\hat{\mu}_{s_0}(t) = \hat{\alpha}(t) + \sum_p \hat{\gamma}_p(t) C_{p,0} + \sum_q \hat{\beta}_q(t) X_{q,0}(t)$.

3.1 Uncertainty Evaluation: a bootstrap approach

To evaluate the uncertainty of a predicted curve $\hat{Y}_{s_0}(t)$, $t = 1, \ldots, M$, we extend the semi-parametric bootstrap approach for spatially correlated data proposed by [7] to the functional context. The bootstrapping algorithm can be summarized as follows:

1. Estimate and remove the drift μ_s following Model (1) to obtain the functional residuals $e_{s_i}(t) = Y_{s_i}(t) - \hat{\mu}_{s_i}(t)$.
2. Estimate the residuals covariance matrix Σ through the trace-semivariogram:

$$\hat{\upsilon}(h) = \frac{1}{2\,|N(h)|} \sum_{i,j \in N(h)} \int_T \left(e_{s_i}(t) - e_{s_j}(t)\right)^2 dt$$

   where $N(h) = \{(s_i, s_j) : \left\|s_i - s_j\right\| = h\}$. A parametric model (exponential or Matérn for example) can be fitted to the points $(h_g, \hat{\upsilon}(h_g)), g = 1, \ldots, G$, as it is done in classical geostatistics. Using the Cholesky decomposition $\hat{\Sigma} = \hat{U}\hat{U}^T$, the functional residuals can be transformed to become spatially uncorrelated:

$$\zeta_{n \times M} = \left(\zeta(s_1)', \ldots, \zeta(s_n)'\right)' = \hat{U}^{-1}_{n \times n} \left(Y_{n \times M} - \hat{\mu}_{n \times M}\right).$$

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 [3], replacing the empirical distribution function of $\{\zeta(s_1), \ldots, \zeta(s_n)\}$, denoted as F_n, by a smooth version $\hat{F}_n$.
4. The final bootstrap sample is determined using an inverse transform:

$$Y^*_{s_i}(t) = \hat{\mu}_{s_i}(t) + \hat{U}\,\zeta^*_{s_i}(t).$$

The bootstrap samples are then fed into the FKED method to obtain B prediction curves at the 10 validation sites. These B prediction curves need to be ordered to determine the upper and lower limits of the confidence band. Here we consider two ordering techniques based on band depth and L^2 distance between curves.

Band depth [11] can be defined for any set of k curves (here $k = 2$). The sample band depth (BD) of $y(t)$ can be calculated as the proportion of bands delimited by 2

curves containing the whole curve $y(t)$ [11]. The modified band depth (*MBD*), that takes into account whether a portion of the curve is in the band, is defined as

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

where λ is the Lebesgue measure on T (for details see [11]). 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. 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 FKED predicted curve $\hat{Y}_{s_0}(t)$ [3].

3.2 Uncertainty Evaluation: a GAM approach

We incorporate a smooth function of longitude, latitude and time, namely $f(lon, lat, t)$, in Model (1). By setting a penalized bivariate spline basis for longitude and latitude a spatial covariance structure is implicit in the model [10], allowing for spatial prediction. The model is linear in the coefficients and 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% confidence bands can be calculated as [12]:

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

Alternatively, uncertainty can be evaluated using the reformulation of the model as a mixed model so that bias is taken into account (see [12], p.138).

4 Results

Predictions at the 10 validation sites were obtained by means of FKED as summarized in Section 3 and a fully additive model as described in Section 3.2. Four performance indices were calculated to compare the spatial prediction of the two models considered and are summarized in Figure 1(b). The indices are: RMSE, Pearson correlation, Normalized Mean Bias Factor and weighted normalized MSE of the normalized ratios (WNNR), that takes into account peaks in observed data, where the normalized ratio is defined as $\exp(-|\ln(fitted/observed)|)$ (see [6] for details). All four indices suggest that FKED provides better predictions for the 10 validation sites. For each of these sites, a bootstrap sample of predicted curves of size 1000 was obtained following the algorithm illustrated in Section 3.1. Band depth was calculated using the modified version *MBD*. The resulting confidence balls/bands,

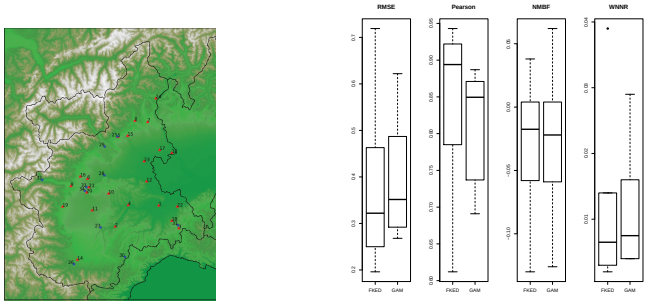
as well as 95% confidence bands according to Eq. (2), are shown in Figure 2 for two of the 10 sites, namely *25-Biella-Largo Lamarmora* and *30-Saliceto*. Overall, the two (band and ball) bootstrap based uncertainty measures seem to agree well. For *Biella-Largo Lamarmora*, nearly all (92.3%) the observed data lie within the bootstrap based confidence ball/band, while for *Saliceto*, there is a high percentage (32.4%) of observed values that lie outside the confidence ball/band. This reflects the performance of the corresponding kriging predictions, good in the first validation site (the predicted curve agrees well with the observed values), but not in the second. On the other hand, all data points lie outside the inference confidence bands, which are much narrower than the bootstrap ones. Part of this difference can be attributed to the fact that bootstrap based confidence bands are simultaneous, while inference based ones are pointwise.

5 Discussion and Future Work

Spatial functional analysis provides an alternative to spatio-temporal modelling capable of predicting a whole curve taking into account exogenous covariates. However, uncertainty evaluation remains an open issue. Here, two different approaches are illustrated. The resulting confidence bands differ considerably, with the bootstrap based ones being too wide while inference based ones are too narrow. Despite being easy to implement, the bootstrap approach seems to be non ideal; it can be computationally expensive and in our case study there are time periods at which the confidence region becomes unjustifiably wider. In the case of the fully additive model, the width of the (inference based) interval remains fairly constant, being slightly wider at the beginning and end of the predicted curve. Improving uncertainty estimation through a mixed model approach is part of our ongoing research.

References

1. Caballero, W., Giraldo, R., Mateu, J.: A universal kriging approach for spatial functional data. Stochastic Environmental Research and Risk Assessment **27(7)**, 1553–1563 (2013)
2. Cameletti, M., Ignaccolo, R., Bande, S.: Comparing spatiotemporal models for particulate matter in Piemonte. Environmetrics **22**, 985-996 (2011)
3. Cuevas, A., Febrero, M., Fraiman, R.: On the use of the bootstrap for estimating functions with functional data. Comput. Statist. Data Anal. **51**, 1063–1074 (2006)
4. Delicado, P., Giraldo, R., Comas, C., Mateu, J.: Statistics for spatial functional data: some recent contributions. Environmentrics **21**, 224–239 (2010)
5. Giraldo, R., Delicado, P., Mateu, J.: Ordinary kriging for function-valued spatial data. Environmental and Ecological Statistics **18(3)**, 411–426 (2011)
6. Ignaccolo, R., Mateu, J., Giraldo, R.: Kriging with external drift for functional data for air quality monitoring. Stochastic Environmental Research and Risk Assessment (2013) doi: 10.1007/s00477-013-0806-y



(a) (b)

Fig. 1 (a) Locations of 24 PM_{10} monitoring sites (red triangles) and 10 validation stations (blue dots). (b) Boxplot of performance indices distribution over the validation stations.

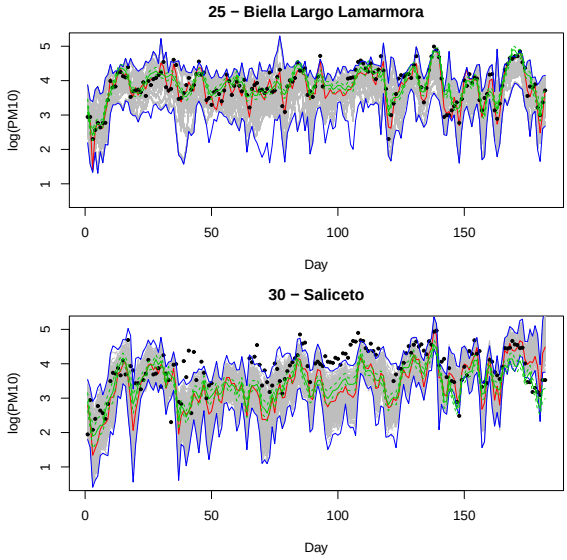


Fig. 2 Original data (black dots), GAM predicted curve (green) and corresponding 95% confidence band (dashed green line), FKED predicted curve (red), 95% confidence ball (grey) based on L^2 distance and 95% confidence band (blue) based on *MBD* for locations *25* and *30* in Figure 1(a).

7. Iranpanah, N., Mohammadzadeh, M., Taylor, C.C.: A comparison of block and semi-parametric bootstrap methods for variance estimation in spatial statistics. Comput. Statist. Data Anal. **55**, 578–587 (2011)
8. Menafoglio, A., Secchi, P., Dalla Rosa, M.: A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. Electronic Journal of Statistics **7**, 2209–2240 (2013)
9. Nerini, D., Monestiez, P., Mant, C.: Cokriging for spatial functional data. J. Multivariate Anal. **101**, 409-418 (2010)
10. Nychka, D.: Spatial process estimates as smoothers. In: Schimek, M.G. (ed.), Smoothing and Regression. Approaches, Computation and Application, pp. 393-424. Wiley, New York (2000)
11. Lopez-Pintado, S., Romo, J.: On the concept of depth for Functional Data. J. Amer. Statist. Assoc. **104(486)**, 718–734 (2009)
12. Ruppert, D., Wand, M.P., Carroll, R.J.: Semiparametric Regression. Cambridge University Press, New York (2003)