GAMs and functional kriging for air quality data

Modelli Additivi Generalizzati e Kriging Funzionale per dati sulla qualitá dell'aria

Francesca Di Salvo, Antonella Plaia and Mariantonietta Ruggieri

Abstract Data having spatio-temporal structure are often observed in environmental sciences. They may be considered as discrete observations from curves along time and/or space and treated as functional. Generalized Additive Models (GAMs) represent a useful tool for modelling, for example, as pollutant concentrations describing their spatial and/or temporal trends.Usually, the prediction of a curve at an unmonitored site is necessary and, with this aim, we extend kriging for functional data to a multivariate context. Moreover, even if we are interested only in predicting a single pollutant, such as PM10, the estimation can be improved exploiting its correlation with the other pollutants. Cross validation is used to test the performance of the proposed procedure.

Abstract *Nell'analisi di dati ambientali si osservano spesso dati con struttura spazio-temporale. Questi possono essere considerati realizzazioni discrete da curve nel tempo e/o nello spazio e trattati come funzionali. I modelli Additivi Generalizzati (GAM) rappresentano uno strumento utile per modellare la concentrazione di inquinanti e descrivere il loro trend spaziale e/o temporale. Spesso, é necessaria la stima di un'intera curva in un sito non osservato. A questo scopo estendiamo il kriging per dati funzionali al contesto multivariato. Infatti, anche se possiamo essere interessati alla previsione di un singolo inquinante, per esempio il PM10, sfruttando la sua correlazione con gli altri inquinanti possiamo migliorarne la stima. Per verificare la validitá della metodologia proposta, utilizziamo una procedura di cross validation.*

Key words: FDA, GAM, OKFD

Francesca Di Salvo, Antonella Plaia, Mariantonietta Ruggieri
Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università di Palermo, e-mail: francesca.disalvo@unipa.it, antonella.plaia@unipa.it, mariantonietta.ruggieri@unipa.it

1 Introduction

Multivariate spatio-temporal data consist of a three way array with two dimension domains (both structured), time and space. In our case, a set of different pollutant levels, recorded for a month/year at different sites, is considered. In this kind of data we may recognize time series along one of the dimensions, spatial series along another and multivariate data along the third dimension.

Over the last few years, there has been an increasing interest within the statistical community for Functional Data Analysis (FDA) [10]. Such a methodology provides a suitable framework, especially in environmental studies, in which large amount of data are recorded over space and/or time. Recently, attention has focused on Spatial Functional Statistics considering spatially dependent functional data [2]. In this context, one of the main issues is the spatial prediction. The Functional kriging [4, 8] extends the ordinary kriging to the functional context, which allows predicting a curve at an unmonitored site by exploiting the curves related to other monitored sites. Suitable methodologies have also been developed in the more realistic cases of absence of stationarity [1], that is, for processes with non-constant mean function (non-stationary functional data). In order to take into account exogenous variables, such as meteorological information, Kriging with External Drift (KED), or regression kriging, is extended to the functional data, involving functional modelling for the trend (drift) and spatial interpolation of functional residuals [7].

In this paper, we want to consider a recurrent case, that is, when more than a single variable (pollutants) is recorded, two situations can be considered: a pollutant has to be predicted in a monitoring site where a) no other pollutants are recorded; b) other pollutants are recorded. Actually, even if we are interested in predicting a single pollutant, such as PM10, in an unmonitored site, exploiting its correlation with the other pollutants can improve the estimation. In this paper, only case a) will be analyzed. We integrate the Multivariate Spatial FDA with Ordinary Kriging for Functional Data (OKFD) in order to obtain the prediction. GAMs [6] are considered to convert data into functional as detailed in [3].

The paper is organized as follows: in Section 2 the methodology is described, Section 3 presents the data, and the performance of the spatial prediction is assessed, Section 4 reports the conclusions and further developments.

2 Methods

In order to predict a curve at an unmonitored site, we consider different steps.

First, we convert into functional the data, considering multivariate functional data as functions of space; for each of the P variables (pollutants in our case), with $p = 1, ..., P$, and for each time t, $t = 1, ..., T$, we model the spatial effects as:

$$\underbrace{y_{\mathbf{s}t}^{p}}_{data} = \underbrace{x_{t}^{p}(\mathbf{s})}_{signal} + \underbrace{\varepsilon_{t}^{p}(\mathbf{s})}_{noise} \tag{1}$$

where y is the response (pollutant concentration), $\mathbf{s}$ is the $(S \times 2)$ matrix of the geographic coordinates, $\mathbf{s} = (longitude, latitude)$, and ε is the random error that may follow any exponential family distribution [6]. We assume a smooth function to model spatial effects (for further details about models for smoothing data in space see [3]).

After modelling the spatial effect and obtaining an estimate $\tilde{x}_t^p(\mathbf{s})$ of $x_t^p(\mathbf{s})$, the residuals are obtained by difference: $e_{\mathbf{s}t}^p = y_{\mathbf{s}t}^p - \tilde{x}_t^p(\mathbf{s})$. They account for a residual spatial component, but also for the whole temporal variability not yet exploited.

$e_{\mathbf{s}t}^p$, like $y_{\mathbf{s}t}^p$, is the generic element of a three way array, and as such, can be considered as function of time $e_{\mathbf{s}}^p(t)$. In fact, if we want to predict y in an unobserved site, we have to account for the temporal variability in the residuals. The residual curve prediction at the unmonitored site s_0 can be obtained by linear combination of data residual curves:

$$\hat{e}_{s_0}^p(t) = \sum_{i=1}^{S} \lambda_i e_{s_i}^p(t), \tag{2}$$

where s_i $(i = 1, 2, ..., S)$ are the observed monitoring sites.

As a last step, the predicted residual $\hat{e}_{s_0}^p(t)$ and $\tilde{x}_t^p$ at s_0 are added to obtain the estimate $\hat{y}^p(t)$ at s_0.

To implement our proposal, all computations are coded in R (R Development Core 2015). The conversion to functional data is realized by using the fda package [11] and mgcv package [13], while the geoR package [5] is also used to implement the proposed kriging procedure.

3 Spatial prediction of pollutant curves

A spatio-temporal multivariate dataset related to air quality is here considered. In particular, our case study considers PM10 and the main daily gaseous pollutant concentrations (CO, SO2, NO2 and O3) aggregated by month and recorded during a year (2011) at 59 monitoring stations dislocated along the State of California (raw data are available at: http://www.epa.gov). The dimension of the array, initially $12 \times 59 \times 5$, is reduced to $12 \times 56 \times 5$, because three of the monitoring sites are excluded from the analysis and used for assessing the performance of the proposed modelling. A map of the monitored area, with the observed monitoring sites, is reported in Figure 1; the three sites chosen as the validation set are highlighted in red.

The concentrations of the pollutants are opportunely standardized and scaled in [0, 100] before performing any analysis. As detailed in [12], the linear interpolation introduced by [9] and used by US EPA (Environmental Protection Agency) is to be preferred. The standardization by segmented linear function, with respect to the standardization by threshold value, allows accounting for different effects of a pollutant on human health, as well as for short and long-term effects, as shown in [12].

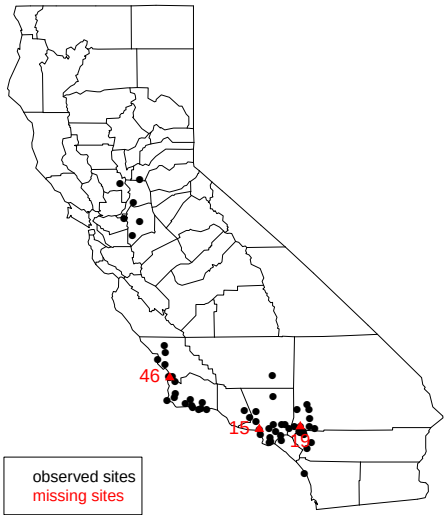

Fig. 1 Map of California with the considered sites.

3.1 Spatial prediction assessment

In order to assess the spatial prediction capability of the proposed procedure, we compare, graphically, observed and predicted data at the three validation sites.

Referring to data of two different sites (chosen from the validation set), and two different pollutants, Figure 2 shows the observed concentrations (black), their conversion into functional (blue) and our predicted curve $\hat{y}(t)$ (red). In both cases, the predicted curve is very close to the smoothed and observed data.

4 Conclusions and further developments

In this paper we propose an integration of Multivariate Spatial FDA with OKFD, in order to exploit correlations among variables to predict one of them. In order to assess the spatial prediction capability of the proposed procedure, we have considered a three way array ($time \times space \times variables$) containing the concentrations of 5 main pollutants recorded in 59 monitoring sites in California (USA) over a year. We focus on predicting a single pollutant in an unmonitored site. The performance of the proposed procedure has been evaluated, as yet, only graphically, comparing ob-

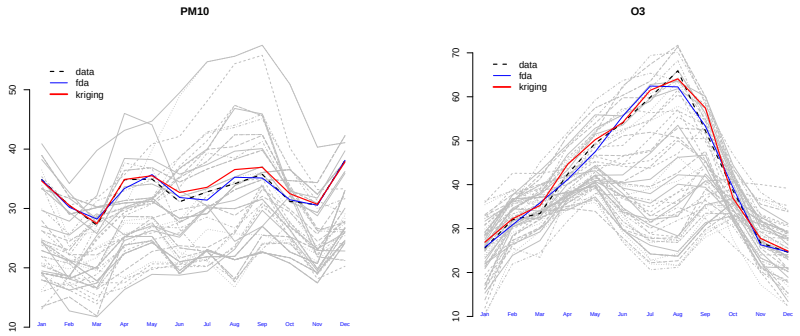

Fig. 2 Observed concentrations (black), their conversion into functional (blue) and predicted curve (red) for the site 15 (left) and the site 19 (right).

served and predicted data at the three validation sites. A more detailed performance evaluation will be carried out considering some performance indexes. In particular, recorded and estimated data in these monitoring sites will be compared by computing the correlation coefficient ρ (the higher the better) and the root mean square deviation RMSD (the lower the better). Moreover, as a further step, the situation b) described in Section 1 will be explored in a future work.

References

1. Caballero, W., Giraldo, R., Mateu, J., A universal kriging approach for spatial functional data. Stoch Environ Res Risk Assess (2013). doi:10.1007/s00477-013-0691-4.
2. Delicado, P., Giraldo, R., Comas, C., Mateu, J., Statistics for spatial functional data: some recent contributions. Environmentrics, 21, 224-239 (2010).
3. Di Salvo, F., Ruggieri, M., Plaia, A., Functional Principal Component analysis for multivariate multidimensional environmental data. Environmental ad Ecological Statistics, 22 (4), 739-757 (2015).
4. Giraldo, R., Delicado, P., Mateu, J., Ordinary kriging for function valued spatial data. Environ Ecol Stat, 18 (3), 411-426 (2011).
5. Ribeiro Jr, P.J., and Diggle, P.J., geoR: Analysis of Geostatistical Data. R package version 1.7-5.1. (2015).
6. Hastie, T., and Tibshirani, R., Generalized Additive Models. Chapman & Hall/CRC. Boca Raton (1990).
7. Ignaccolo, R., Mateu, J., Giraldo, R., Kriging with external drift for functional data for air quality monitoring. Stoch Environ Res Risk Assess, 28, 1171-1186 (2014).
8. Nerini, D., Monestiez, P., Mant, C., Cokriging for spatial functional data. J Multivar Anal, 101, 409-418 (2010).
9. Ott, W.R., and Hunt Jr W.F., A quantitative evaluation of the pollutant standards index. J Air Pollut Control Assoc, 26, 1051-1054 (1976).
10. Ramsay, J.O., and Silverman, B.W., Functional Data Analysis. Second Edition. Springer-Verlag (2005).

11. Ramsay, J.O., Wickham, H., Graves, S., Hooker, G., fda: Functional Data Analysis. R package version 2.4.4. (2014).
12. Ruggieri, M., and Plaia, A., An aggregate AQI: comparing different standardizations and introducing a variability index. Science of the Total Environment, 420, 263-272 (2012).
13. Wood, S.N., mgcv: mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation, R package version 1.8-11 (2016).