

## FUNCTIONAL MODELLING IN ENVIRONMETRICS

**Francisco A. Ocaña-Lara, Mariano J. Valderrama,  
Francisco M. Ocaña-Peinado and Manuel Escabias**

*Department of Statistics and Operation Research, Faculty of Pharmacy,  
University of Granada, 18071-Granada, Spain*

**Abstract.** An usual approach for functional regression of a stochastic process  $Y(t)$  versus a vector variable  $(X_1(t), X_2(t), \dots, X_n(t))$  consists of developing a multi-step marginal procedure for explaining  $Y(t)$  in terms of a certain  $X_i(t)$  and then the procedure is step-wise applied to the residuals. The relation among the coefficients of linear determination is derived and the equivalence between a functional linear regression and a principal component prediction model is also proved. Moreover, we estimate a functional model for pollen concentration in a certain interval of time from its past history as well as the environment temperature.

**Keywords:** Functional principal component, functional regression, pollen concentration

### 1. Introduction

Forecasting methods for time series may be self-explicative or based on dynamic regression. The former only consider the past history of the series itself, whereas the latter methods include information from an input process, with the residuals also being represented as an ARIMA model. Nevertheless, in practice, neither of the above techniques is suitable for forecasting peak pollen concentrations due to the very sparse nature of the data and especially because there are discontinuities in the atmospheric presence of pollen, which is limited to the first quarter of the year, reaching a peak by the second half of February.

The main aim of the present study is to derive, by means of a stepwise procedure, a regression model enabling us to forecast cypress pollen concentration in the interval spanning 15-28 February each year, on the basis of prior knowledge and of the air temperature during the previous month of January and the first half of February. Because of the continuous-time nature of both processes, a functional approach to the problem is adopted. Then the sample paths will be interpolated by cubic splines in order to work with regular functions concerning a given time interval.

Due to the seasonal behaviour of cypress pollen a long time series cannot be used; instead, it is truncated each year in an interval so that a set of sample curves are available. Then, we propose a functional regression model (FRM) based on functional principal component (FPC) analysis. For FPC selection, we will apply a criterion based on the proportion of error reduction of each pair (past and future) in the model, once they have been decreasingly ordered by their explained variance.

Our paper considers data provided by the Aerobiology Centre at the University of Granada (Spain). Cypress pollen grains were recorded throughout the day in a cylindrical device containing a suction pump and were then transformed into a volume scale ( $\text{grains/m}^3$ ). These data were averaged over 24 hours, such that the variable pollen concentration was represented by a daily data series. Both series of data, pollen and temperature, are very scattered, and so we employed a smoothing series obtained by logarithmic transformation.

Previous research by the present authors in this field were first published by Aguilera *et al* (1997) who introduced the Principal Component Prediction (PCP) method of functional forecasting of a stochastic process from its past sample paths and by Ocaña-Lara *et al* (1999) who studied some problems that arise when an inner product is proposed in the sample space in order to incorporate functional properties of the sample curves. In a more practical way, Valderrama *et al* (2002) fitted an ARIMA process to the FPC model and then applied it to climate data for the climate phenomena known as *El Niño*. Subsequently, Aguilera *et al* (2008) presented a binary response functional model including an environmental application to rainfall data.

## 2. Relation between FRM and PCP

Let us consider two second order and quadratic-mean continuous stochastic processes denoted as  $\mathbf{X} \equiv \{X(s), s \in S\}$  and  $\mathbf{Y} \equiv \{Y(t), t \in S \cup T\}$ , where  $S$  and  $T$  are compact intervals. We will divide  $\mathbf{Y}$  in two stochastic processes:  $\mathbf{Y}_S \equiv \{Y(s), s \in S\}$  and  $\mathbf{Y}_T \equiv \{Y(t), t \in T\}$  that we will call past and future respectively. Our aim is to estimate a forecasting model for  $\mathbf{Y}_T$  on the basis of the knowledge of  $\mathbf{X}$  and  $\mathbf{Y}_S$  by means of:

$$Y(t) = \alpha(t) + \int_S (\beta(s, t), \gamma(s, t)) \begin{pmatrix} X(s) \\ Y(s) \end{pmatrix} ds + \varepsilon(t), \quad t \in T, \quad (1)$$

where  $\varepsilon(t)$  is a white noise and the parameter functions  $\beta(s, t)$  and  $\gamma(s, t)$  are square-integrable in  $S \times T$ .

If we denote by  $\{\xi_i\}$  and  $\{\eta_j^T\}$  the sets of FPC of  $\mathbf{X}$  and  $\mathbf{Y}_T$  respectively, we can decompose both processes by means of the bi-orthogonal representations:

$$\begin{aligned} X(s) &= \mu_X(s) + \sum_{i=1}^{\infty} f_i(s) \xi_i, \quad s \in S \\ Y(t) &= \mu_Y(t) + \sum_{j=1}^{\infty} g_j(t) \eta_j^T, \quad t \in T \end{aligned} \quad (2)$$

being  $\mu_X(s)$  and  $\mu_Y(t)$  the mean functions of them and  $\{f_i\}$  and  $\{g_j\}$  the corresponding  $L_2$ -orthogonal eigenfunction system associated with the eigenvalues  $\{\lambda_i\}$  and  $\{\rho_j\}$  of the covariance operators of  $\mathbf{X}$  and  $\mathbf{Y}_T$ .

**Theorem 1.** Under certain regularity assumptions, the FRM

$$E[Y(t) / X] = \alpha(t) + \int_S \beta(s, t) X(s) ds$$

is equivalent to the PCP model

$$\eta_k^T = \sum_{j=1}^{\infty} \frac{E[\xi_j \eta_k^T]}{\lambda_j} \xi_j. \quad (3)$$

*Proof:*

We start from a general FRM for  $\mathbf{Y}$  given by:

$$\begin{aligned} Y(t) &= \alpha(t) + \int_S \beta(s, t) X(s) ds + \nu(t) = \\ &\mu_Y(t) + \int_S \beta(s, t) [X(s) - \mu_X(s)] ds + \nu(t) \end{aligned} \quad (4)$$

He *et al* (2003) proved that the function  $\beta(s, t)$  can be represented in terms of the eigenfunction system as:

$$\beta(s, t) = \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \frac{E[\xi_j \eta_k^T]}{\lambda_j} f_j(s) g_k(t). \quad (5)$$

Then, replacing expression (5) for  $\beta(s, t)$  in (4) we have that

$$E[Y(t) / X] = \mu_Y(t) + \sum_{i=1}^{\infty} \xi_i \int_S \beta(s, t) f_i(s) ds = \mu_Y(t) + \sum_{i=1}^{\infty} \sum_{k=1}^{\infty} \frac{E[\xi_j \eta_k^T]}{\lambda_j} \xi_j g_k(t).$$

and identifying now this expression with the one of  $Y(t)$  obtained from (2) we conclude identity (3).

## 3. Estimation of the FRM

To estimate the FRM we will adapt the orthogonal projection method for approximating eigenfunctions and hence the FPC by means of trigonometric functions as the basis of the linear approximating subspace. Then, a criterion for selecting the most suitable sets of FPC to be included in the model is developed on

the basis of the paper by Aguilera *et al* (1999b). Finally, because of we suppose that the residuals  $\upsilon \equiv \{\upsilon(t), t \in T\}$  are not white noise, but they depend on the process  $Y_s$ , we can also model  $\upsilon$  by means of the FRM:

$$\upsilon(t) = \int_s \gamma(s, t) Y(s) ds + \varepsilon(t) \tag{6}$$

where  $\varepsilon(t)$  is a white noise, and estimation is carried out following the aforementioned procedures. Relation among the linear coefficients of the partial and the complete FRM is given as follows:

**Theorem 2.** Let  $r_1$  and  $r_2$  be the linear correlation coefficients of the models (4) and (6) respectively and  $r$  the one of (1). Then:

$$r^2 = r_1^2 + r_2^2(1 - r_1^2).$$

*Proof:*

Let us consider three stochastic process  $X$ ,  $Y$  and  $Z$ . In order to obtain the relation among the linear determination coefficients in a two-step linear model such as:

$$Z = \hat{Z} + \varepsilon, \quad \varepsilon = \hat{\varepsilon} + \omega,$$

where  $\hat{Z}$  is estimated from  $Y$  and  $\hat{\varepsilon}$  from  $X$ , let us consider the coefficients of determination:

$$r_1^2 = \frac{\sigma_{\hat{Z}}^2}{\sigma_Z^2}, \quad r_2^2 = \frac{\sigma_{\hat{\varepsilon}}^2}{\sigma_\varepsilon^2}.$$

Under linear assumptions of a linear model, we can decompose:

$$\sigma_Z^2 = \sigma_{\hat{Z}}^2 + \sigma_\varepsilon^2, \quad \sigma_\varepsilon^2 = \sigma_{\hat{\varepsilon}}^2 + \sigma_\omega^2,$$

so that

$$\sigma_Z^2 = \sigma_{\hat{Z}}^2 + \sigma_{\hat{\varepsilon}}^2 + \sigma_\omega^2.$$

Then the coefficient of determination for the model  $Z = \hat{Z} + \hat{\varepsilon} + \omega$  is given by

$$r^2 = \frac{\sigma_{\hat{Z}}^2 + \sigma_{\hat{\varepsilon}}^2}{\sigma_Z^2} = \frac{\sigma_{\hat{Z}}^2}{\sigma_Z^2} + \frac{\sigma_{\hat{\varepsilon}}^2}{\sigma_\varepsilon^2} \cdot \frac{\sigma_\varepsilon^2}{\sigma_Z^2} = r_1^2 + r_2^2(1 - r_1^2)$$

that proves the theorem.

#### 4. A model for pollen concentration data

In this application we consider as the past interval  $S$  the period between 1 January and 14 February, and as the future interval  $T$  the one between 15 and 28 February in each year. Then, for estimating the model, we will consider nine sample paths corresponding to the years 1999 to 2007. The forecasting performance will be tested with the one corresponding to 2008. The orthogonal projection method is applied considering 25 trigonometric functions on the past interval  $S$  for temperature and 35 functions for pollen concentration, due to the higher scattering level of its trajectories, with 10 functions being considered on the future  $T$  for pollen concentration. Previously, the original series are transformed as follows:

$$\begin{aligned} X(s) &= \log\{\text{temperature in } s + 2\}, & s \in S, \\ Y(s) &= \log\{\text{pollen concentration in } s + 1\}, & s \in S, \\ Y(t) &= \log\{\text{pollen concentration in } t + 1\}, & t \in T. \end{aligned}$$

With the estimated model we predict the pollen concentration for the second half of February 2008, using the two-step method described in the previous section, first taking into account the temperature on  $S$  and then the pollen concentration also in  $S$  for this same year.

As we established in §2,  $\{\hat{\xi}_i\}$  will denote the estimated FPC of the process  $\mathbf{X}$  and  $\{\hat{\eta}_j^T\}$  the FPC of  $\mathbf{Y}_T$ . The significant coefficients of the regression equations in the PCP model are then given in Table 1. In the same way, by denoting  $\{\hat{\eta}_i^S\}$  the estimated FPC of  $\mathbf{Y}_S$  and by  $\{\hat{v}_j\}$  the ones of the residual process  $\mathbf{v} \equiv \{v(t), t \in T\}$ , the significant coefficients of the regression equations in the PCP model are given in Table 2. The resulting coefficients of linear determination were:

$$r^2 = 0.9591 \quad r_1^2 = 0.8115 \quad r_2^2 = 0.7831$$

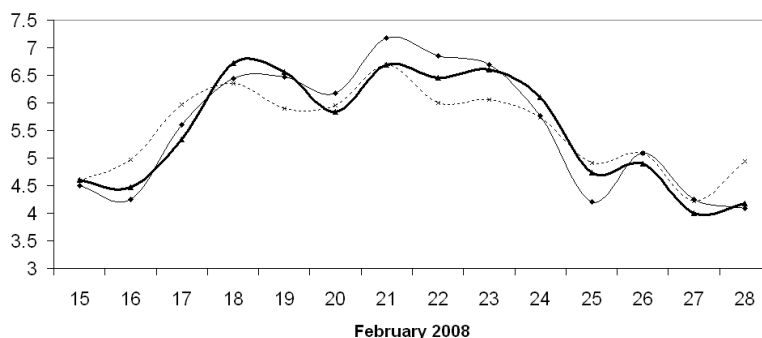
**Table 1.** Regression coefficients of the PCP model of  $\mathbf{Y}_T$  versus  $\mathbf{X}$ .

|            | $\xi_1$ | $\xi_2$ | $\xi_3$ | $\xi_4$ | $\xi_5$ | $\xi_6$ | $\xi_7$ | $\xi_8$ |
|------------|---------|---------|---------|---------|---------|---------|---------|---------|
| $\eta_1^T$ |         | 2.859   | 1.750   | -2.528  |         |         |         |         |
| $\eta_2^T$ | -1.219  |         |         |         | -2.347  |         |         |         |
| $\eta_3^T$ | 0.492   |         | -1.617  |         |         | -2.392  |         |         |
| $\eta_4^T$ |         | -0.636  |         |         | -1.066  | -1.513  | -1.307  |         |
| $\eta_5^T$ |         |         |         | 0.797   |         |         |         | -2.274  |

**Table 2.** Regression coefficients of the PCP model of  $\mathbf{v}$  versus  $\mathbf{Y}_T$ .

|       | $\eta_1^S$ | $\eta_2^S$ | $\eta_3^S$ | $\eta_4^S$ | $\eta_5^S$ | $\eta_6^S$ | $\eta_7^S$ | $\eta_8^S$ |
|-------|------------|------------|------------|------------|------------|------------|------------|------------|
| $v_1$ |            |            | -0.123     |            |            |            | 0.478      |            |
| $v_2$ |            | 0.105      | -0.172     |            |            | -0.176     |            | 0.248      |
| $v_3$ |            | 0.072      |            |            | -0.240     | 0.141      |            |            |
| $v_4$ |            |            |            |            |            | -0.147     |            | -0.275     |
| $v_5$ |            |            | 0.068      | -0.070     |            | -0.084     |            |            |
| $v_6$ |            |            |            | 0.072      |            |            |            |            |

Figure 1 shows the curve of pollen concentration in the future together with the forecasts obtained by the FRM (i) after the first-step modelling and (ii) incorporating in a second step the residual modelling to the FRM. We can observe that predictions with the complete model after the second step are the closest ones to the real values of pollen concentrations. Furthermore, one of the major advantages of the proposed model is that it allows forecast on the whole interval  $T$  taking into account only observed values on the past  $S$ .



**Fig. 1.** Pollen concentration in  $T$  (thick curve) and predictions obtained after the first-step modelling (dotted curve) and the second-step modelling (thin curve)

## Acknowledgements

This research was supported by Projects MTM2007-63793 from Dirección General de Investigación del MEC, Spain, and P06-FQM-01470 and FQM-307 from Consejería de Innovación, Ciencia y Empresa de la Junta de Andalucía, Spain.

## References

- Aguilera A.M., Ocaña F.A., Valderrama M.J., 1997. An approximated Principal Component Prediction Model for continuous time stochastic processes. *Appl. Stoch. Models Data Anal.*, **13** (1), 61–72.
- Aguilera A.M., Ocaña F.A., Valderrama M.J., 1999. Forecasting time series by functional PCA. Discussion of several weighted approaches. *Comput. Statist.*, **14**, 443–467.
- Aguilera A.M., Escabias M., Valderrama M.J., 2008. Forecasting binary longitudinal data by a functional PC-ARIMA model. *Comput. Statist. Data Anal.*, **52**, 3187–3197.
- He G.Z., Müller H.G., Wang J.L., 2003. Functional canonical analysis for square integrable stochastic processes. *J. Multivar. Anal.*, **85**, 54–77.
- Ocaña F.A., Aguilera A.M. and Valderrama M.J., 1999. Functional Principal Component Analysis by choice of norm. *J. Multivar. Anal.*, **71**, 262–276,
- Valderrama M.J., Ocaña F.A., Aguilera A.M., 2002. Forecasting PC-ARIMA models for functional data. *Proceed. Comput. Statist. 2002* (Härdle W., Rönz B., eds.), Physica-Verlag, 25–36.