Stepwise selection of functional covariates in forecasting peak levels of olive pollen

- Manuel Escabias; Mariano J. Valderrama; Ana M. Aguilera; M. Elena Santofimia; M. Carmen Aguilera-Morillo

- Stepwise selection of functional covariates in forecasting peak levels of olive pollen

- *Stoch Environ Res Risk Assess (2013) 27:367–376*
- DOI: 10.1007/s00477-012-0655-0

Stoch Environ Res Risk Assess (2013) 27:367–376
DOI 10.1007/s00477-012-0655-0

ORIGINAL PAPER

Stepwise selection of functional covariates in forecasting peak levels of olive pollen

Manuel Escabias · Mariano J. Valderrama · Ana M. Aguilera · M. Elena Santofimia · M. Carmen Aguilera-Morillo

Published online: 27 October 2012

Abstract High levels of airborne olive pollen represent a problem for a large proportion of the population because of the many allergies it causes. Many attempts have been made to forecast the concentration of airborne olive pollen, using methods such as time series, linear regression, neural networks, a combination of fuzzy systems and neural networks, and functional models. This paper presents a functional logistic regression model used to study the relationship between olive pollen concentration and different climatic factors, and on this basis to predict the probability of high (and possibly extreme) levels of airborne pollen, selecting the best subset of functional climatic variables by means of a stepwise method based on the conditional likelihood ratio test.

Keywords *Olea europaea* L. airborne pollen · Functional-logit-regression · Selection of functional predictors

M. Escabias (✉) · M. J. Valderrama · M. C. Aguilera-Morillo
Facultad de Farmacia, Universidad de Granada, Campus de Cartuja, 18071 Granada, Spain
e-mail: escabias@ugr.es

M. J. Valderrama
e-mail: valderra@ugr.es

M. C. Aguilera-Morillo
e-mail: caguilera@ugr.es

A. M. Aguilera
Facultad de Ciencias, Universidad de Granada, Campus Fuentenueva, 18071 Granada, Spain
e-mail: aaguiler@ugr.es

M. E. Santofimia
IES La Laguna, Consejería de Educación - Junta de Andalucía, Vicente Aleixandre S.N., Padul, 18640 Granada, Spain
e-mail: helenasantof@hotmail.com

1 Introduction

High concentrations of olive pollen occur every year in late spring and early summer in provinces of southern Spain (Jaén, Granada and Córdoba) where olives are the main crop; in Granada, 49.5 % of the total area dedicated to agriculture (127,208 Ha) is occupied by olive trees (see Alba et al. 2000). High levels of airborne olive pollen constitute a problem for many inhabitants of these areas because of the amount of allergies it causes; indeed, many people are forced to move away during this time of the year. According to earlier studies (D'Amato and Lobefalo 1989; Macchia et al. 1991), olive trees are the primary cause of pollen-related allergies in the Mediterranean region as a whole. In southern Spain, 71 % of the population is affected (Díaz de la Guardia et al. 2003).

In consequence, many authors have sought to forecast airborne olive pollen concentrations, using methods such as time series (Belmonte and Canela 2002) or have applied linear regression to determine the factors that influence on it (Vazquez et al. 2003; Diaz de la Guardia et al. 2003). Recently, more sophisticated methods have been used in this field, by Castellano-Mendez et al. (2005), who used neural networks, and by Aznarte et al. (2007), who combined fuzzy systems and neural networks to model Betula pendula in the air. A different approach was taken by Ocaña-Peinado et al. (2008), who used principal component analysis to model the inertia process of a transfer function model. Valderrama et al. (2010) developed a two-step functional model to forecast cypress pollen concentration. This latter paper offers a different standpoint for studying the relationship between olive pollen concentration and various other factors, from the perspective of functional data analysis.

Under functional data analysis methodology, the continuity and time-dependency of variables are used as tools

for modelling and forecasting variables. In the case of olive pollen, it would be very interesting to know the probability of high or extreme levels of airborne pollen concentration in order to minimise its allergenic effect. Functional logit regression fits this aim perfectly. The functional logistic regression model was defined to model and predict a binary response variable from a functional predictor. To this end, various functional models have been proposed (Cardot and Sarda 2005; Rossi et al. 2002). Ramsay and Silverman (1997) proposed diverse functional models based on basis expansion, and since then other authors have adopted these methods to predict a binary outcome from functional predictors (Ratcliffe et al. 2002; Escabias et al. 2004; Aguilera et al. 2008b).

As stated above, many studies about the factors that may influence the amount or time evolution of airborne olive pollen have been developed (Alba et al. 2000; Galán et al. 2005 and Valderrama et al. 2010). Its concentration on any given day is mainly influenced by the pollen concentration, temperature, hours of sunshine, humidity and rainfall during the immediately preceding days. The influence of the wind varies from one zone to another, with no well-defined pattern of association. The temperature may have influence in different ways; for example, different effects are exerted by the maximum, minimum and mean temperatures, and by different ranges of temperature; moreover, there may be a slighter influence below a certain temperature and a greater one above it. Furthermore, the influence is not the same everywhere; different areas are affected in different ways by different factors. Another fact to be taken into account is that there are differences between the effects of pollen concentration throughout the pollination season and its effects in the pre-peak period, i.e., the period preceding the maximum level of pollen concentration (Vázquez 2003). Sánchez-Mesa et al. (2002) defined a classification of successive years on the basis of the weather factors affecting grasses and their emission of pollen. Better knowledge of the climatic conditions that affect olive pollen emission would enable us to classify these years from a meteorological point of view and thus obtain a general model of olive pollen concentration.

In the present paper, the phenomenon analysed—airborne pollen concentration—is a seasonal one. The main pollen season (MPS) is defined as the period in which the greatest airborne olive pollen concentration occurs. The pollen season as a whole, extends from April to mid/late June, with the greatest air concentrations occurring in May, when rainfall decreases and mean temperatures rise. Nevertheless, MPS has been defined in various ways. For example, Sanchez-Mesa et al. (2002) defined MPS as beginning when a mean value of at least 1 pollen grain/m^3/day was detected, and at least 1 grain/m^3 on the following days, with no more than one consecutive day of 0 grains/m^3; and the season ended when pollen concentration in the air decreased to 1 or 2 grains/m^3/day. Other authors define the beginning of the MPS with respect to the chilling period (Galán et al. 2005). It is important to stipulate the MPS very precisely so that the study is not affected by long tails at the beginning and end of the season, which would provoke serious errors in the statistical analysis.

In order to fit and predict peaks of airborne pollen concentration in the city of Granada (Spain), we analysed the different meteorological factors that affect this question, using a stepwise method based on a functional logistic regression model.

In summary, the main goals of this paper are to:

- Model the occurrence of airborne olive pollen peaks, using climatic functional variables.
- Predict as accurately as possible the probability of occurrence of airborne olive pollen peaks, by observing the time evolution of these climatic variables.
- Identify the climatic variables that best enable us to model the occurrence of olive pollen peaks.
- Analyze the forecasting performance of the model.

2 Theory of the functional logit regression model

A functional variable is one whose values depend on a continuous magnitude such as time. The statistical tool based on the analysis of functional data is the stochastic process. However in the approach based on basis expansion, as adopted in this paper, the functional (predictor) variables included in the model are not considered to be random. They are functional in the sense that they are evaluated at any time in the domain, instead of the discrete way, in which they were originally measured or observed (Ramsay and Silverman 2005). Thus, a functional data set is a set of curves $\{x_1(t),\ldots,x_n(t)\}$, with $t \in T$. Each curve can be observed at different time points of his argument t as $x_i = (x_i(t_0),\ldots, x_i(t_{m_i}))'$ for the set of times $t_0,\ldots,t_{m_i}$, $i = 1,\ldots,n$ and these are not necessarily the same for each curve.

Regardless of the stochastic nature of functional data and the form they are observed (continuously or discretely) the usual assumption in functional data analysis and considered in this work is that the curves belong to the squared integrable functions space $L^2(T)$ defined as

$$L^2(T) = \left\{ f : T \longrightarrow \mathbb{R} : \int_T f^2(t)dt < \infty \right\}.$$

With the usual scalar product

$$\langle f, g\rangle_u = \int_T f(t)g(t)dt, \quad \forall f, g \in L^2(T), \tag{1}$$

this is a separable Hilbert space.

Different approaches have been taken to the study of functional data, including the nonparametric methods proposed by Müller (2008) and Ferraty and Vieu (2006) and the basis expansion methods used by Ramsay and Silverman (2005). The latter method is adopted in the present study, in which we seek to reconstruct the functional form of curves in order to evaluate them at any time point t. This method assumes that the curves belong to a finite dimensional space generated a basis of functions $\{\phi_1(t), \ldots, \phi_p(t)\}$ and so they can be expressed as

$$x_i(t) = \sum_{j=1}^{p} a_{ij}\phi_j(t), \quad i = 1, \ldots, n. \tag{2}$$

The functional form of the curves is determined when the basis coefficients $a_i = (a_{i1}, \ldots, a_{ip})'$ are known. These can be obtained from the discrete observations either by least squares or by interpolation (see, for example, Escabias et al. 2005, 2007).

Depending on the characteristics of the curves and the observations, various classes of basis can be used (see, for example, Ramsay and Silverman 2005). In practice, those most commonly used are, on the one hand, the basis of trigonometric functions for regular, periodic, continuous and differentiable curves, and on the other, the basis of B-spline functions, which provides a better local behavior (see De Boor 2001).

In order to formulate the functional logit model let Y be a binary response random variable and let $\{X(t) : t \in T\}$ be a functional covariate related to Y. Given a curve $x(t)$ we can consider that the conditional distribution of the response to the given curve follows a Bernoulli distribution whose parameter depends on the curve. That parameter is the probability of success and the conditional expectation is expressed

$$\begin{aligned}\pi(x(t)) &= P\{Y = 1|X(t) = x(t)\} \\ &= E[Y|\{X(t) = x(t) : t \in T\}].\end{aligned}$$

Given $x_1(t), \ldots, x_n(t)$ a sample of curves of the functional predictor and $y_1, \ldots, y_n$ a sample of the response associated with the n curves, the model is expressed as

$$y_i = \pi_i + \varepsilon_i = \pi(x_i(t)) + \varepsilon_i, \quad i = 1, \ldots, n,$$

and in matrix form as

$$Y = \boldsymbol{\pi} + \boldsymbol{\varepsilon} \tag{3}$$

where $Y = (y_1, \ldots, y_n)'$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)'$, with

$$\pi_i = \frac{\exp\{\alpha + \int_T x_i(t)\beta(t)dt\}}{1 + \exp\{\alpha + \int_T x_i(t)\beta(t)dt\}}, \quad i = 1, \ldots, n, \tag{4}$$

$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ the vector of centered random errors, with unequal variances and a Bernoulli distribution, and $\beta(\cdot)$ the functional parameter to be estimated.

This model can also be expressed in terms of the *logit* transformations as

$$l_i = \ln\left[\frac{\pi_i}{1 - \pi_i}\right] = \alpha + \int_T x_i(t)\beta(t)dt, \quad i = 1, \ldots, n, \tag{5}$$

which presents the logistic model as a generalised functional linear model, as proposed by James (2002), with the logit transformation as the link function.

One advantage of regression models compared to other prediction models is that we can interpret the relationship between the explanatory variables and the response quantitatively. In the logit case this quantitative relationship is expressed in terms of odds ratios. Thus, the exponential of the integral over the interval $(t_0, t_0 + h)$ of the functional parameter multiplied by a constant K is the odds ratio of response $Y = 1$ versus $Y = 0$ when the sample path is constantly increased by K units in that interval (Escabias et al. 2005). For a more general increase $\Delta X(t) = g(t)$ the odds ratio would be

$$\theta[\Delta X(t) = g(t)/\Delta t = h] = \exp\left\{\int_{t_0}^{t_0+h} g(t)\beta(t)dt\right\}$$

Following Ramsay and Silverman (2005) in their explanation for the linear case, the estimation of the functional parameter $\beta(t)$ is impossible with the usual least squares methods (weighted in this case), since $\beta(t)$ contains an uncountable set of values and we would have at most a finite number of conditions. In other words, there could be infinite solutions with a perfect fit of the observations to the response. If we consider the sample paths $x_1(t), \ldots, x_n(t)$ expressed in terms of the basis $\{\phi_1(t), \ldots, \phi_p(t)\}$, in the form of Eq. 2 and the functional parameter in terms of a different basis $\{\varphi_1(t), \ldots, \varphi_q(t)\}$ as

$$\beta(t) = \sum_{k=1}^{q} \beta_k \varphi_k(t). \tag{6}$$

The functional logit model in terms of the logit transformations is then expressed as

$$l_i = \alpha + \sum_{j=1}^{p}\sum_{k=1}^{q} a_{ij}\psi_{jk}\beta_k, \quad i = 1, \ldots, n \tag{7}$$

with ψ_{jk} being the scalar products between the basis functions

$$\psi_{jk} = \int_T \phi_j(t)\varphi_k(t)dt, \quad j = 1,\ldots,p, \quad k = 1,\ldots,q.$$

The functional logit model is now a classical logit model which in matrix form and in terms of logit transformations is expressed as

$$L = X\beta$$

where $L = (l_1,\ldots,l_n)'$ is the vector of logit transformations. $X = (\mathbf{1} \mid A\Psi)$ is the design matrix, and | indicating the separation between the two boxes of the matrix. $\mathbf{1} = (1,\ldots,1)'$ is a $n-$ length vector of ones. Ψ is the matrix whose entries (ψ_{jk}) are the scalar products (defined in (1)) between basic functions setted abobe. A is the matrix of sample curve basis coefficients as rows. $\beta = (\beta_0,\beta_1,\ldots,\beta_q)'$ with $\beta_0 = \alpha$ are the basis coefficients of the functional parameter. These coefficients would be the parameters of the multiple model to be estimated.

A special case occurs when the same basis is used for both the functional parameter and the explanatory curves. Then Ψ is a square matrix that can be diagonal with a Fourier basis and tridiagonal with a B-splines basis. In this paper we propose the use of a stepwise method to choose the best functional predictors in order to model and predict airborne olive pollen peaks; to do so we show the generalisation of the functional logit model to the case of more than one functional predictor.

Let Y be a binary response and $X_1(t), X_2(t),\ldots,X_r(t)$ a set of functional predictors related to Y. Let us consider a sample of curves of the predictors $(x_{i1}(t),\ldots,x_{ir}(t))'$, $i = 1,\ldots,n$ and $y_1,\ldots,y_n$, the associated sample of the responses. Then, in terms of the logit transformations, the model is formulated as

$$l_i = \alpha + \int_T (x_{i1}(t)\beta_1(t) + \cdots + x_{ir}(t)\beta_r(t))dt, \quad i = 1,\ldots,n, \tag{8}$$

being $\beta_1(t),\ldots,\beta_r(t)$ the r functional parameters associated to the functional covariates.

By considering the basis expansion of the sample curves in terms of the basis $\{\phi_{11}(t),\ldots,\phi_{1p_1}(t)\},\ldots,\{\phi_{r1}(t),\ldots,\phi_{rp_r}(t)\}$ respectively in the form

$$x_{ih}(t) = \sum_{j=1}^{p_h} a_{ijh}\phi_{hj}(t), \quad i = 1,\ldots,n, \quad h = 1,\ldots,r. \tag{9}$$

and also for the functional parameters in the form

$$\beta_h(t) = \sum_{k=1}^{q_h} \beta_{hk}\varphi_{hk}(t), \quad h = 1,\ldots,r. \tag{10}$$

in terms of the basis $\{\varphi_{11}(t),\ldots,\varphi_{1q_1}(t)\},\ldots,\{\varphi_{r1}(t),\ldots,\varphi_{rq_r}(t)\}$, the logit model would be in matrix form and in terms of logit transformations

$$L = X\beta$$

where $L = (l_1,\ldots,l_n)'$ is a vector of logit transformations. $X = (\mathbf{1}|A_1\Psi_1|\cdots|A_r\Psi_r)$ is the design matrix of the model. $\mathbf{1} = (1,\ldots,1)'$ is a n-length vector of ones. Ψ_h is the $p_h \times q_h$ matrix of scalar products between basis $\{\phi_{h1}(t),\ldots,\phi_{hp_h}(t)\}$ and basis $\{\varphi_{h1}(t),\ldots,\varphi_{hq_h}(t)\}$, $h = 1,\ldots,r$ defined by Eq. (1). $A_1,\ldots,A_r$ are the matrices that have as rows the basis coefficients of the curves.

$\beta = (\beta_0|\beta_1|\cdots|\beta_r)'$ with $\beta_h = (\beta_{h1},\ldots,\beta_{hq_h})'$, $\beta_0 = \alpha$ and $h = 1,\ldots,r$ are the vectors of the basis coefficients of the functional parameters.

As in the case of a single functional predictor, the same basis can be used in all cases (predictors and parameters). This is done in the application of the airborne olive pollen model.

3 Selection of functional variables

The main contribution of this paper is in the selection of climatic functional variables predicting the occurrence of pollen peaks. This selection is based on a forward stepwise method, using conditional likelihood ratio tests. This section describes the particular tests involved in this selection method.

3.1 Goodness of fit test

The goodness of fit test is used to determine whether the proposed logit regression model fits the data. In the literature many such tests have been described, and some are compared in Hosmer et al. (1997) for the multiple case. In the present paper, we use the likelihood ratio test based on the Wilks statistic. This test is expressed as

$$H_0\colon \pi_i = \frac{\exp\{\alpha + \int_T x_i(t)\beta(t)dt\}}{1 + \exp\{\alpha + \int_T x_i(t)\beta(t)dt\}}, \quad i = 1,\ldots,n$$

$$H_1\colon \pi_i \neq \frac{\exp\{\alpha + \int_T x_i(t)\beta(t)dt\}}{1 + \exp\{\alpha + \int_T x_i(t)\beta(t)dt\}}$$

that is, we seek to determine whether the conditional expectation $E[Y|X = x_i(t)] = E[Y_i] = \pi_i$ can be expressed as a logit regression model.

In general the Wilks statistic follows a chi-square distribution and is defined as

$$-2\ln\Lambda = -2\ln\frac{\sup_{H_0} L(X_1,\ldots,X_n;\theta)}{\sup L(X_1,\ldots,X_n;\theta)} \underset{n\to\infty}{\overset{H_0}{\rightsquigarrow}} \chi^2_d$$

with d being the difference between the dimensions of the two parametric spaces. The Wilks statistic in logistic regression can be expressed as

$$G^2(M) = -2\ln\Lambda = 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{\widehat{\pi}_i}\right) + (1-y_i)\ln\left(\frac{1-y_i}{1-\widehat{\pi}_i}\right)\right] \underset{n\to\infty}{\overset{H_0}{\rightsquigarrow}} \chi^2_{n-q-1}.$$

with $d = n - q - 1$ and q the number of non-functional covariates of the model, that is, the number of bases considered for $\beta(t)$. The $G^2(M)$ statistic is termed *deviance* and plays the same role as the sum of squares of residuals in the linear regression model.

3.2 Conditional likelihood ratio test

This test is used to choose between two nested models and can be formulated as

H_0: Model M_h is verified

H_1: ModelM_h is not verified assuming model M,

where M is the logit model with possibly all the functional variables $X_1(t),\ldots,X_r(t)$ and which is assumed to be the true model, and where M_h is the model nested in M which is obtained after setting one functional parameter to zero $\beta_h(t) = 0$. This test could be formulated as

$$H_0: l_i = \alpha + \int_T \big(x_{i1}(t)\beta_1(t) + \cdots + x_{i(h-1)}(t)\beta_{h-1}(t) + x_{i(h+1)}(t)\beta_{h+1}(t) + \cdots + x_{ir}(t)\beta_r(t)\big)dt$$

$$H_1: l_i = \alpha + \int_T (x_{i1}(t)\beta_1(t) + \cdots + x_{ir}(t)\beta_r(t))dt$$

or equivalently

H_0: $\beta_h(t) = 0$

H_1: $\beta_h(t) \neq 0$

The test statistic can be calculated as the difference between the *deviance* (goodness of fit statistic) of each model $G^2(M_h/M) = G^2(M_h) - G^2(M)$ because

$$\begin{aligned} G^2(M_h/M) &= -2\ln\frac{L_{M_h}}{L_M} \\ &= 2(\mathcal{L}_M - \mathcal{L}_{M_h}) - 2\mathcal{L}_S + 2\mathcal{L}_S \\ &= [-2(\mathcal{L}_{M_h} - \mathcal{L}_S)] - [-2(\mathcal{L}_M - \mathcal{L}_S)] \\ &= G^2(M_h) - G^2(M), \end{aligned}$$

where $\mathcal{L}_M$ and $\mathcal{L}_{M_h}$ are the maxima of the log-likelihood functions of general model M and the particular one M_h respectively, and $\mathcal{L}_S$ is the maximum of the log-likelihood of the saturated model (Agresti 2002). This statistic can also be expressed as

$$G^2(M_h/M) = -2\ln\Lambda = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\widehat{\pi}_{i(M_h)}}{\widehat{\pi}_i}\right) + (1-y_i)\ln\left(\frac{1-\widehat{\pi}_{i(M_h)}}{1-\widehat{\pi}_i}\right)\right] \underset{n\to\infty}{\overset{H_0}{\rightsquigarrow}} \chi^2_d.$$

where $\widehat{\pi}_{i(M_h)}$ and $\widehat{\pi}_i$ are the predicted probabilities under the M_h and M models respectively. The degrees of freedom d of the chi-square distribution of the Wilks statistic are the difference between the number of parameters of each model. In the case of the functional logit model this difference is $d = q_h$.

The conditional likelihood ratio test allows the stepwise selection of functional variables. The forward method consists of examining each step of the variable that is not in the model and whose introduction would further reduce the *deviance*. Functional variables would be introduced into the model until there is no further significant reduction of the *deviance*.

Step 0. Fit the logit model with no variables, only the intercept term, and calculate its deviance. Set this model as the current model.

Step 1. Fit the models obtained by adding to the logit model each of the variables that are not in the current model, and select the variable that most reduces the deviance of the current model.

Step 2. If the difference between deviances is lower than a fixed chi-square quantile, stop. Otherwise, update the current model as the one with the selected variable and return to Step 1.

4 Forecasting peaks of airborne olive pollen from functional climatic variables

Our database contains daily observations of the concentration of airborne olive pollen, measured in grains per cubic meter of air, in the city of Granada from 1 January 1992 to 30 June 2003. The measurement of this concentration was performed following the standard methodology of the Spanish Aerobiology Society with a Hirst-type collector located at the Faculty of Sciences, University of Granada.

Besides the concentration of olive pollen, we recorded the daily values of the following meteorological variables: maximum, minimum and mean temperature, hours of sunlight, relative humidity, rainfall and wind speed. These data were obtained from the Spanish Institute of Meteorology.

The MPS coincides with the flowering of the tree, which in the case of the olive takes place in April, May, June and part of July. Various definitions have been given for the beginning and end of the MPS. For example, Sánchez-Mesa et al. (2002) define its start as the day on which an average value of at least one grain of pollen per cubic meter of air is recorded, and when this value is maintained for several consecutive days. The end of the MPS is defined

as the day on which there is a decrease to 1 or 2 grains per m^3. Other authors define the MPS on the basis of the days that exceed a certain average temperature. For our purposes, the MPS is defined as the period from 15 April to 14 July of each year. Figure 1 describes the different climatic variables observed in the MPS.

The purpose of this study is to model and forecast high levels of olive pollen for the period spanning 1 week, from the evolution of climatic variables during the preceding week. The MPS considered covers 91 days (13 weeks) per year, except in the last year considered (2003), when values were obtained from 15 April to 30 June, and so in this case data were available for 11 weeks. Thus, for each functional variable we had a total of 154 weeks' data.

For our purposes, the binary response variable was defined as taking the value of one in a week if there was at least 1 day when the pollen level exceeded 200 grains per m^3 of air. This is the level of pollen considered by the Spanish Society of Allergology and Clinical Immunology to be dangerous for patients with allergy problems. Of the 154 weeks observed, 43.51 % exceeded this threshold.

As stated above, the functional variables used to predict the peak of airborne olive pollen were the weekly curves for maximum, minimum and average temperature, hours of sunshine, relative humidity, rainfall, wind speed and the current concentration of airborne olive pollen. Although we had daily records for these variables, the curves were reconstructed by quasi-natural cubic spline interpolation (Escabias et al. 2005). Figure 2 shows interpolation curves for some of the weeks and some of the variables. The reconstruction of these curves can be expressed in terms of the basis of the cubic B-spline functions defined by the nodes $\{1, 2, 3, 4, 5, 6, 7\}$ which was the basis used for all the functional variables considered.

To meet our objectives, the multi-functional logistic regression model was used.

The first and most important step was to determine the variables, among those available, that best model the response through a multi-functional logistic model. To choose the functional variables that best predict a peak of pollen, the procedure used was a forward-stepwise selection method based on conditional likelihood ratio tests.

In order to assess the accuracy of the predictions made by different models, we selected a training sample of size 100 and a test sample of size 54, fitted the model with the training sample and predicted the test sample. Finally, both predictions, on the training sample and on the test sample, were evaluated by the area under the ROC curve and by the rate of correct classifications explained later.

After several tests it was found that depending on the training sample chosen, the functional variables that best predicted the occurrence of pollen peaks could vary. To decide which variables were the best predictors of our response, we repeated the sample selection (training and test) 500 times and observed the frequency of appearance of each variable in the model and the order in which it did so. The results of these repetitions are summarised as follows.

In the 500 replications of the experiment, the first variable to enter the model and therefore the most important predictor of pollen peaks was the current level of pollen. This corroborates the findings of various authors regarding the self-explanatory capacity of the pollen level (see, for example, Valderrama et al. 2010). In 74 of the 500 replications of the experiment, just two variables entered the model, firstly the level of pollen, and secondly one with the following frequencies:

Variable	Frequency
Maximum temperature	13
Minimum temperature	11
Mean temperature	15
Isolation	10
Humidity	11
Rainfall	11
Wind speed	14

In 22 of the 500 replications of the experiment, four variables entered the model. Because of this small percentage, we chose to reject these replications.

In the remaining 404 replications (80.8 %) three optimal functional variables were required to predict pollen peaks. The most important functional variables in this case were as follows:

100 % of the time, the first variable was the level of pollen. Of the remaining variables, the frequencies of appearance in second position in the model are summarised in the following table.

Variable	Frequency
Maximum temperature	16
Minimum temperature	82
Mean temperature	80
Isolation	47
Humidity	96
Rainfall	20
Wind speed	63

Thus, of these variables, humidity entered the model significantly more often than the others. The minimum and mean temperature were the other two most important

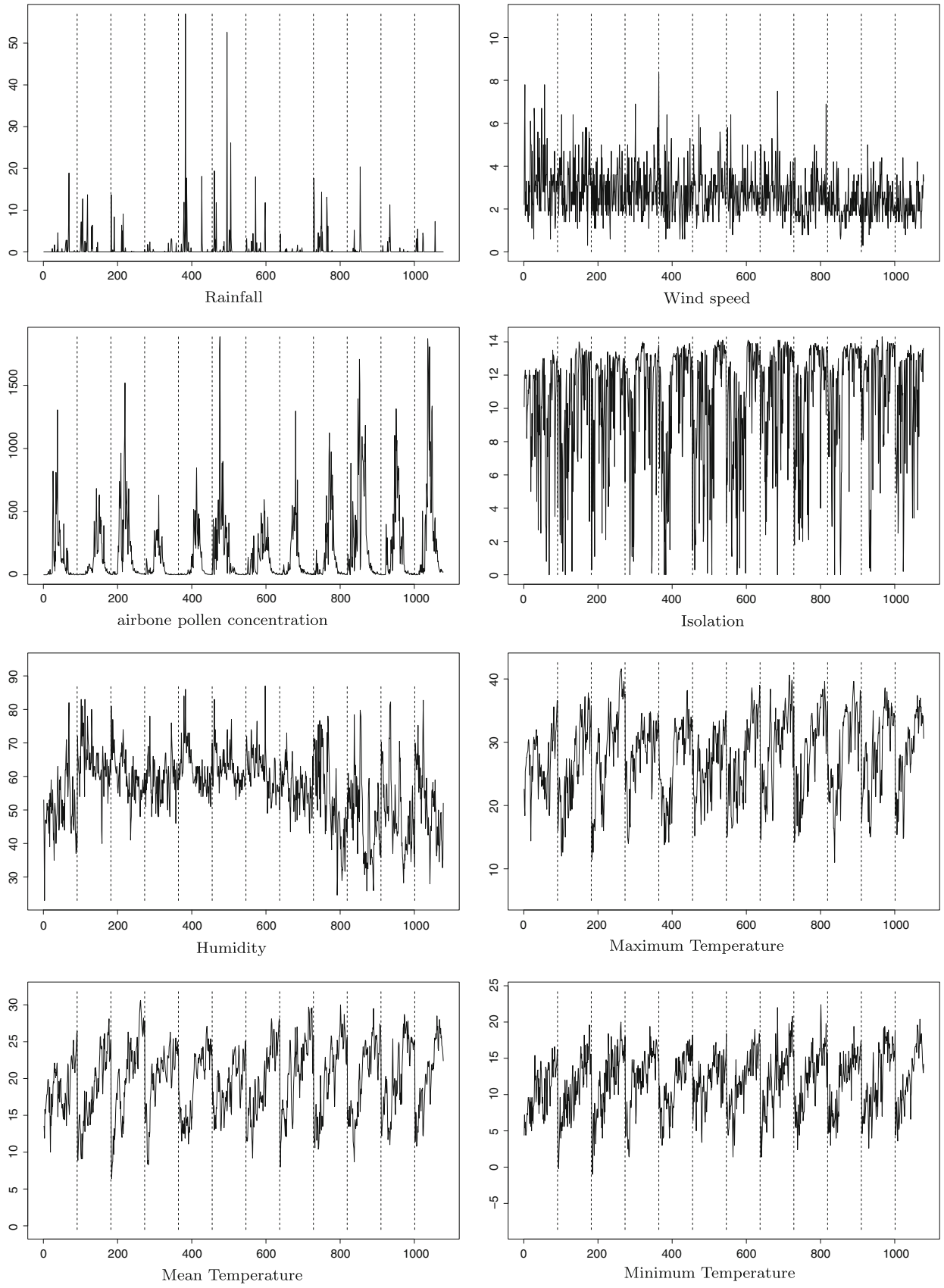

Fig. 1 Sample paths of functional covariates

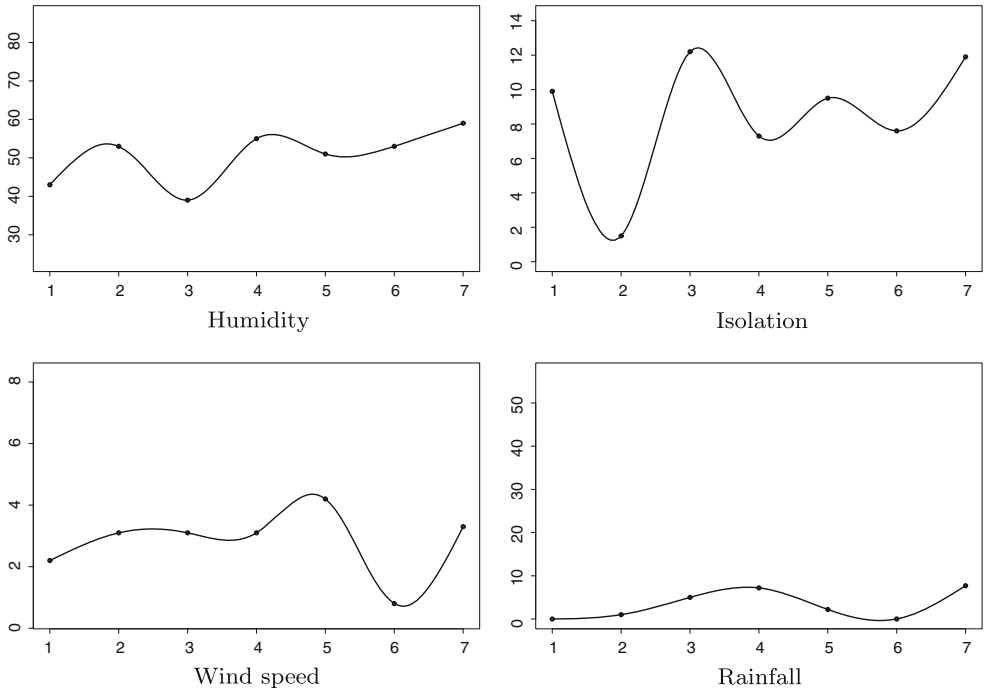

Fig. 2 Interpolated curves for some predictor variables

variables in frequency, with the minimum temperature being slightly the more significant. It should be noted that the average temperature used in this database is the average of the maximum and minimum temperatures. Of these two values, peak pollen levels are more strongly affected by the minimum temperature. In summary, after current pollen levels, the two most important variables for predicting peak pollen appear to be the minimum temperature and humidity.

The importance of these two variables is confirmed by the fact that the 96 occasions on which the second most important variable was humidity, the remaining variables entered in third place with the following frequencies:

Variable	Frequency
Maximum temperature	7
Minimum temperature	32
Mean temperature	8
Isolation	29
Rainfall	14
Wind speed	6

On the 82 occasions on which the minimum temperature entered in second place, the distribution of variables that entered in third place was:

Variable	Frequency
Maximum temperature	13
Mean temperature	9
Isolation	10
Humidity	32
Rainfall	10
Wind speed	8

To evaluate the predictive ability of the model, we used the rate of correct classifications and the area under the ROC curve.

Table 1 Summary of correct classification rates and ROC areas

	Training samples		Test samples	
	ROC	CCR	ROC	CCR
Mean	0.99	98.6	0.78	75.98
ST. Dev	0.029	3.344	0.069	6.085
Mín	0.84	83.0	0.59	57.41
Q_1	1.00	100	0.74	72.22
Q_2	1.00	100	0.78	75.93
Q_3	1.00	100	0.82	79.63
Máx	1.00	100	0.97	90.74

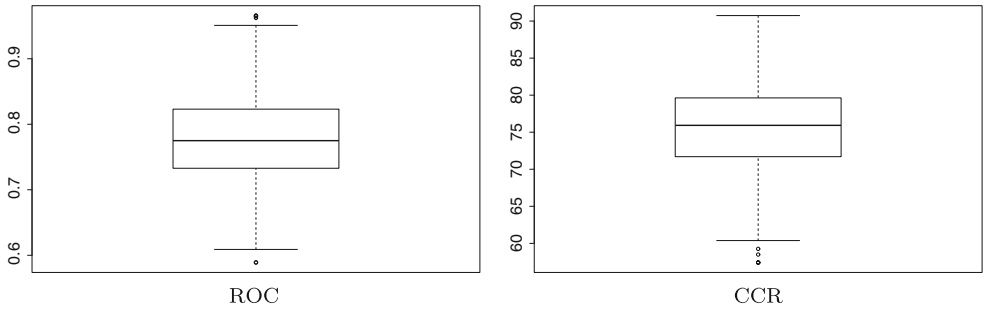

Fig. 3 Distribution of ROC areas and CCR in the test samples

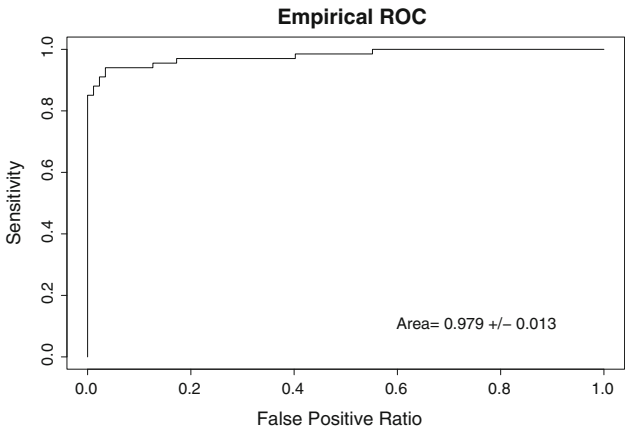

Fig. 4 ROC curve of functional logit model for the complete database (all curves)

The correct classification rate (CCR) is one of the most commonly used measures in logistic regression to assess the goodness of predictions. To calculate the CCR a cutoff point p_c (usually $p_c = 0.5$) is chosen and a prediction is considered to be correctly classified when the estimated probability $\widehat{\pi}_i \geq p_c$ and $y_i = 1$ or $\widehat{\pi}_i < p_c$ and $y_i = 0$, otherwise it is considered classified incorrectly. Thus, the CCR is defined as the ratio between the number of observations correctly classified and the total number of sample observations.

Although a cutoff value of 0.5 is usually used, it would be more appropriate to use the cutoff point that maximises the CCR (Hosmer and Lemeshow 1989), which is usually very close to the proportion of ones in the sample.

The ROC curve is a graph that evaluates the model's ability to discriminate. The fitted logistic regression model predicts the value of the response depending on whether the predicted probability is greater than or equal to the cut-point chosen to discriminate. The logistic regression model is considered a good predictor if it predicts as a success those individuals actually observed to be successes and predicts as a failure those individuals observed to be failures. The ROC curve plots the true positive rate ($y = 1,\ \widehat{y} = 1$) against the false positive rate ($y = 0,\ \widehat{y} = 1$) for different cutoff points. The nearest point to the unit is the best discrimination point and the area under the curve is a measure of the capacity to discriminate. The closer this measure is to one, the better it is, and an acceptable value would be 0.7 or higher.

As indicated above, the prediction models with these three functional variables were evaluated using the area under the ROC curve and the CCR for both the fitted values (training sample) and the predictions of the test sample. As summary measures of the replications, we calculated the mean and the quartiles of the distributions of the areas under the ROC curve and the CCR for the training sample and the sample test (Table 1, Fig. 3). Thus, we conclude that the functional logit model performs very well with respect to predicting the peak values of airborne olive pollen concentration, with a mean CCR of nearly 99 % in the training samples and 76 % in the test samples. The ROC area showed a good prediction ability for the model, with a mean area of 0.99 in the training samples and 0.78 in the test samples.

In order to evaluate the goodness of fit of the logit model, we obtained the deviance statistics $G^2(M)$ and the p-values of the chi-square, with $100 - 27 - 1$ degrees of freedom for each of the 404 models in which the three selected variables entered the stepwise selection method. In only 6 % of the replications did the goodness of fit test show that the logit model was not a good model. For the model with the 154 curves, the deviance statistic was $G^2(M) = 53.585$, which produced a p-value for the chi-square statistic distribution with $154 - 27 - 1 = 126$ degrees of freedom of 1, and so we accept that the model that uses these variables to predict peaks of airborne olive pollen performs adequately. Finally, the prediction measures for this model showed it to perform well, with an area under the ROC curve of 0.9792 (see Fig. 4) and a CCR of 94.8052 %.

5 Conclusions

The aims of this paper were to model and forecast the occurrence of airborne olive pollen peaks, on the basis of climatic functional variables, and to analyze the accuracy

of these predictions. The methodology developed consists of deriving a functional logit regression model whose covariates are selected by a stepwise procedure.

After evaluating the repetitions of the the stepwise method, we conclude that the optimum set of variables to explain the occurrence of peaks of airborne olive pollen in a given week are the current level of pollen concentration, together with the minimum temperature and the humidity in the previous week. Both the CCR and the area under the ROC curve reflect the good performance of the functional logit model to predict these peak values.

The functional logit model is a particular case of the functional generalized linear model proposed by James (2002) with the logit link that works well in the prediction of occurrence of olive pollen peaks in a week from the evolution of meteorological variables in the previous week. In a similar way authors plan to extend their study to model the number of olive pollen peaks in a week from the same functional covariates, by considering a Poisson link in the same general class of functional generalized linear models.

Acknowledgments This research was supported by Projects MTM2010-20502 from Dirección General de Investigación del MEC, Spain, and FQM-307 from Consejería de Innovación, Ciencia y Empresa de la Junta de Andalucía Spain. The authors are grateful to the Aerobiology Research Group at the University of Granada for providing data for our study. The authors also want to thank the referees their suggestions that have allowed to prepare this improved version of the paper.

References

Agresti A (2002) Categorical data analysis. Wiley, New York

Aguilera AM, Escabias M, Valderrama MJ (2008a) Discussion of different logistic models with functional data. Application to Systemic Lupus Erythematosus. Comput Stat Data Anal 53: 151–163

Aguilera AM, Escabias M, Valderrama MJ (2008b) Forecasting binary longitudinal data by a functional PC-ARIMA model. Comput Stat Data Anal 52:3187–3197

Alba F, Díaz de la Guardia C, Comtois P (2000) The effect of meteorological parameters on diurnal patterns of airborne olive pollen concentration. Grana 39:200–208

Aznarte JL, Benítez JM, Nieto D, Linares C, Díaz de la Guardia C, Alba F (2007) Forecasting airborne pollen concentration time series with neural and neuro-fuzzy models. Expert Syst Appl 32: 1218–1225

Belmonte J, Canela M (2002) Modelling aerobiological time series. Application to Urticaceae. Aerobiologia 18:287–295

Cardot H, Sarda P (2005) Estimation in generalized linear models for functional data via penalized likelihood. J Multivar Anal 92: 24–41

Castellano-Méndez M, Aira MJ, Iglesias I, Jato V, González-Manteiga W (2005) Artificial neural network as a useful tool to predict the risk level of Betula pendula in the air. Int J Biometeorol 49:310–316

D'Amato G, Lobefalo G (1989) Allergenic pollen in the southern Mediterranean area. J Allergy Clin Immunol 83:116–122

De Boor C (2001) A practical guide to Splines. Springer, New York

Díaz de la Guardia C, Alba F, Sánchez F, Trigo MM, Galán C, Ruíz L, Sabariego, S (2003) Aerobiological analysis of *Olea europaea* L. pollen in different localities of southern Spain. Grana 42: 234–243

Escabias M, Aguilera AM, Valderrama MJ (2004) Principal component estimation of functional logistic regression. J Nonparametric Stat 16(3–4):365–384

Escabias M, Aguilera AM, Valderrama MJ (2005) Modeling environmental data by functional principal component logistic regression. Environmetrics 16:95–107

Escabias M, Aguilera AM, Valderrama MJ (2007) Functional PLS logit regression model. Comput Stat Data Anal 51:4891–4902

Ferraty F, Vieu P (2006) Nonparametric functional data analysis. Springer, New York

Galán C, García-Mozo H, Vázquez L, Ruíz L, Díaz de la Guardia C, Trigo MM (2005) Heat requirement for the onset of the *Olea europaea* L. pollen season in several sites in Andalusia and the effect of the expected future climate change. Int J Biometeorol 49:184–188

Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S (1997) A comparison of goodness-of-fit tests for the logistic regression model. Stat Med 16:965–980

Hosmer DW, Lemeshow S (1989) Applied logistic regression. Wiley, New York

James JM (2002) Generalized linear models with functional predictors. J R Stat Soc Ser B 64(3):411–432

Macchia L, Caiffa MF, D'Amato G, Tursi A (1991) Allergenic significance of Oleaceae pollen. In: D'Amato G, Spieksma FThM, Bonini S (eds) Allergenic Olea pollen in southern Spain pollen and pollinosis in Europe. Blackwell Scientific Publication, Oxford. pp 87–93

Müller HG (2008) Functional modeling of longitudinal data. In: Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G (eds) Longitudinal data analysis (Handbooks of modern statistical methods). CRC, New York. pp 223–252

Ocaña-Peinado FM, Valderrama MJ, Aguilera AM (2008) A dynamic regression model for air pollen concentration. Stoch Environ Res Risk Assess 22(1):59–63

Ramsay JO, Silverman BW (1997) Functional data analysis. Springer-Verlag, New York

Ramsay JO, Silverman BW (2005) Functional data analysis, 2nd edn. Springer-Verlag, New York

Ratcliffe SJ, Heller GZ, Leader LR (2002) Functional data analysis with application to periodically stimulated foetal heart rate data. II: functional logistic regression. Stat Med 21:1115–1127

Rossi N, Wang X, Ramsay JO (2002) Nonparametric item response function estimates with the EM algorithm. J Behav Educ Sci 27: 291–317

Sánchez-Mesa JA, Galán C, Martínez-Heras JA, Hervas-Martínez C (2002) The use of a neural network to forecast daily grass pollen concentration in a Mediterranean region: the southern part of the Iberian Peninsula. Clin Exp Allergy 32:1606–1612

Valderrama MJ, Ocaña FA, Aguilera AM, Ocaña-Peinado FM (2010) Forecasting pollen concentration by a two-step functional model. Biometrics 66(2):578–585

Vázquez LM, Galán C, Domínguez-Vilches E (2003) Influence of meteorological parameters on olea pollen concentrations in Córdoba (South-western Spain). Int J Biometeorol 48(2):83–90