Forecasting binary longitudinal data by a functional PC-ARIMA model

- Ana M. Aguilera, Manuel Escabias, Mariano J. Valderrama
- Forecasting binary longitudinal data by a functional PC-ARIMA model
- *Computational Statistics Data Analysis, Volume 52, Issue 6, 2008, Pages 3187-3197*
- DOI: https://doi.org/10.1016/j.csda.2007.09.015

Volume 169
May 2022
ISSN 0167-9473
ELSEVIER
COMPUTATIONAL STATISTICS & DATA ANALYSIS
An Official Publication of
iasc
The International Association for Statistical Computing
A Section of The International Statistical Institute
CMStatistics
The network Computational and Methodological Statistics
Available online at www.sciencedirect.com
ScienceDirect

Available online at www.sciencedirect.com
ScienceDirect

Computational Statistics & Data Analysis 52 (2008) 3187–3197

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Forecasting binary longitudinal data by a functional PC-ARIMA model

Ana M. Aguilera*, Manuel Escabias, Mariano J. Valderrama

Department of Statistics and O.R., University of Granada, Spain

Available online 26 September 2007

Abstract

In order to forecast time evolution of a binary response variable from a related continuous time series a functional logit model is proposed. The estimation of this model from discrete time observations of the predictor is solved by using functional principal component analysis and ARIMA modelling of the associated discrete time series of principal components. The proposed model is applied to forecast the risk of drought from *El Niño* phenomenon.

Keywords: Logistic regression; Functional principal component analysis; ARIMA modelling

1. Introduction

Binary time series appear in many real applications modelling the occurrence of an event of interest in a certain time interval. Examples of these types of data are time evolution of flares in a lupus patient and the occurrence of transactions on a heavy trade stock market. In order to model binary time series data, the classic autoregressive Box–Jenkins model has been extended and different binary autoregressive models for the logistic link function have been proposed (Cox and Snell, 1989). Several generalizations of these models which allow nonparametric additive covariates have been recently introduced (Hyndman, 1999).

In practice we are interested in forecasting a binary time series from a related continuous time series observed only at a finite set of discrete time points. In the previous examples the risk of flare of a lupus patient during a period of time could be predicted in terms of the patient's daily stress level along such period; the occurrence of transactions in a time interval from the number of shares traded each minute during this interval; and the annual risk of drought in terms of monthly temperatures throughout the year. The final aim is to predict the annual risk of drought in a specific zone from monthly evolution of *El Niño* climatic phenomenon. Observe that in each of these examples the predictor variable represents the evolution of a certain magnitude in time, so that its realizations are functions (functional data) instead of vectors. Therefore this problem will be solved by developing a functional data analysis (FDA) approach.

Functional models have emerged in Statistics to describe situations in which the involved data are curves (realizations of a continuous-time stochastic process) observed at a finite set of time points. Examples of FDA are functional principal component analysis (FPCA) to reduce the dimension in a stochastic process, functional linear regression or functional

* Corresponding author. Departamento de Estadística e I.O., Facultad de Ciencias, Universidad de Granada, Campus Fuentenueva s/n. 18071-Granada, Spain. Tel.: +34 958246306; fax: +34 958243267.

E-mail addresses: aaguiler@ugr.es (A.M. Aguilera), escabias@ugr.es (M. Escabias), valderra@ugr.es (M.J. Valderrama).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.09.015

ANOVA to model a scalar response variable from a functional covariate, and functional canonical correlation analysis to investigate different modes of variability in two sets of curves. An excellent review of functional methods and their applications in a diverse range of subject areas has been developed (Ramsay and Silverman, 2005, 2002). The actual development on statistical methods for analyzing functional data and new trends have been recently revised (González-Manteiga and View, 2007).

In the general context of FDA, logistic regression has been recently extended to model a time-independent binary response in terms of a functional predictor (continuous time stochastic process). This model has been used to predict the probability of a high risk birth outcome from periodically stimulated foetal heart rate tracings (Ratcliffe et al., 2002). Usually, principal component (PC) regression and principal covariate regression are used to forecast a response variable from highly correlated predictors (see Heij et al., 2007 for a comparison of the forecast accuracy of these two methods). Different FPCA approaches for estimating the functional logit model have been proposed (Escabias et al., 2004). In the general context of functional generalized linear models, the EM algorithm has been used for estimating the model (James, 2002). An odds ratio interpretation of the relationship between a binary response and a functional predictor in terms of the estimated parameter function of the functional logit model has been established and an application with climatological data has been developed (Escabias et al., 2005). A PLS approach for estimating the functional logit model has been recently introduced (Escabias et al., 2007). With the same objective, a linear discriminant analysis for classification of functional data has been proposed (Preda et al., 2007). The estimation of this model is based on functional PLS regression (Preda and Saporta, 2005). Alternative nonparametric curves discrimination methods have been studied (Ferraty and Vieu, 2003).

FPCA is a generalization of the classic principal component analysis (PCA) of a sample of data vectors for the reduction of dimension of a set of sample curves (Ramsay and Silverman, 2005). Different models based on FPCA to forecast a functional predictor in the future from its past evolution have been developed in recent years. Principal component prediction (PCP) models based on linear regression of each future PC on a reduced set of past PCs were introduced (Aguilera et al., 1997). PCP models have been adapted for predicting a continuous time series from unequally spaced discrete time observations (Aguilera et al., 1999a, b). A detailed study of PCP models can be found in Valderrama et al. (1997). Mixed FPC-ARIMA models are based on ARIMA modelling of the PCs of the sample functions obtained by cutting the observed time series in periods of the same amplitude (Valderrama et al., 2002). All these FPCA models allow not only to forecast a continuous time series in a whole future interval but also to reconstruct it between the discretization time points in the past. This mixed approach has been recently extended allowing application to a broader class of problems such as robust forecasting of mortality and fertility rates (Hyndman and Ullah, 2007). On the other hand, a time-localized frequency domain PCA method is proposed for signals that exhibit locally stationary behavior (Ombaoa and Hob, 2006).

A functional logit model based on mixed FPC-ARIMA modelling of the functional predictor, which allows us to forecast the time evolution of a binary response from discrete time observations of a continuous time series, is introduced in this paper. In Section 2, we formulate the problem in mathematical terms. A functional logistic regression (FLR) model for predicting binary longitudinal data, in terms of the functional predictor sample curves obtained by cutting the original time series in periods of the same amplitude, is proposed in Section 3. In order to solve the multicollinearity problem and to reduce dimension, FLR is estimated in Section 4 by using as covariates a reduced set of functional PCs of the functional predictor. ARIMA modelling of each PC series will be considered in Section 5 to forecast the predictor continuous time series in a future period of time followed by the forecast of the related binary response in this period. Finally, a climatological application with real data will be developed in Section 6. We aim to predict the risk of drought in a future period of time from monthly observations of *El Niño* phenomenon.

2. Problem formulation

Let us suppose that we have observations of a continuous time series $\{x(t)\}$ at discrete time points in the interval $(0, NT]$ and one observation Y_w of a related binary response Y at each period of amplitude T defined as $((w-1)T, wT]$, $w = 1, \ldots, N$. Let us denote by t_{wk} $(w = 1, \ldots, N; k = 1, \ldots, m_w)$ the observation time points of the original time series $\{x(t)\}$ at each period $((w-1)T, wT]$ and by $x_{wk} = x(t_{wk})$ the corresponding observed values. Thus, the purpose is to estimate a functional logit model to forecast the binary response in future periods $((w^* - 1)T, w^*T](w^* > N)$, from the forecasting of the series $x(t)$ provided by a mixed ARIMA-FPCA model in such periods.

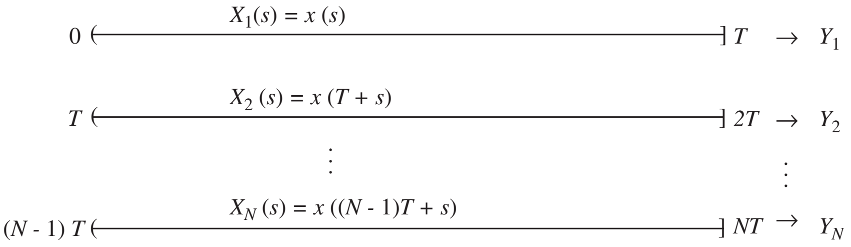

Fig. 1. Sample information obtained after cutting the original continuous time series.

In order to formulate and to estimate a functional logit model based on FPCA, we propose to cut the observed series $x(t)$ in N periods of amplitude T, so that we have N sample paths of the following functional predictor (continuous time process):

$$\{X_w(s) = x((w-1)T+s) : s \in (0,T];\ w = 1,\ldots,N\}, \tag{1}$$

and a sample of size N of the binary response given by $\{Y_w : w = 1,\ldots,N\}$ (see Fig. 1).

Let us observe that the choice of the amplitude T is simple enough in practice when there is a well-defined seasonal period as in the case of many real time series.

3. Functional logistic regression

The objective of the FLR model is to explain a binary response variable Y in terms of a functional variable $X(s)$ whose sample information is given by a set of curves measured without error. Let $\{X_w(s) : w = 1,\ldots,N\}$ be a sample of curves of a functional variable $\{X(s) : s \in (0,T]\}$ obtained by cutting the original predictor series $x(t)$ in periods of amplitude T, and let $\{Y_w : w = 1,\ldots,N\}$ be the random observations of the binary response variable Y associated with the sample curves. Therefore, the FLR model is given by $Y_w = \pi_w + \varepsilon_w$, where ε_w are zero mean independent random errors with variance $\pi_w(1-\pi_w)$, and π_w is the probability of response $Y=1$ for a specific curve $X_w(s)$ modelled as $\pi_w = \exp(l_w)/(1+\exp(l_w))$, with l_w being the logit transformation given by

$$l_w = \alpha + \int_0^T X_w(s)\beta(s)\,\mathrm{d}s, \quad w = 1,\ldots,N. \tag{2}$$

Here α is a real parameter and $\beta(s)$ is a parameter function that has to be estimated. In terms of the logit transformations, the model can be equivalently seen as a functional generalized linear model (James, 2002).

As in the functional linear model, it is impossible to obtain a direct estimation of the FLR model by using the usual likelihood or least-squares methods (Ramsay and Silverman, 2005). In addition, functional data are usually observed only at a finite set of time points so that its true functional form has to be reconstructed from its discrete time observations by using an approximating procedure. Therefore, the most used solution for solving this estimation problem is based on the assumption that the parameter function and the sample curves belong to a finite dimension space generated by a basis of functions $\{\phi_1(t),\ldots,\phi_p(t)\}$, so that they can be expressed in terms of the basis as

$$\beta(s) = \sum_{k=1}^{p} \beta_k \phi_k(s) \quad \text{and} \quad X_w(s) = \sum_{j=1}^{p} a_{wj}\phi_j(s). \tag{3}$$

Then, the functional model given by Eq. (2) is equivalent to the following multiple logit model

$$l_w = \alpha + \sum_{j,k=1}^{p} a_{wj}\psi_{jk}\beta_k,$$

that can be equivalently expressed in matrix form by $L = \mathbf{1}\alpha + A\Psi\beta$, with $L = (l_1,\ldots,l_N)'$, A being the matrix that has the basis coefficients of the sample curves as rows, $\Psi = (\psi_{jk})_{p\times p}$ that which has the L^2-usual inner products between the basic functions as entries, $(\psi_{jk} = \int_0^T \phi_j(s)\phi_k(s)\,\mathrm{d}t)$, and $\beta = (\beta_1, \ldots, \beta_p)'$ the vector of the parameter function basis coefficients (Escabias et al., 2004).

Before estimating by likelihood the vector β, we have to compute the matrix A of sample curves basis coefficients. Let $x_w = (x_{w1}, \ldots, x_{wm_w})'$ be the vector of observations of the wth sample curve $X_w(s)$ at m_w time points of the interval $((w-1)T, wT]$, $\forall w = 1, \ldots, N$. In order to approximate the basis coefficients of each sample path from its observations at a set of time points not necessarily the same for all curves, two different alternatives can be used depending on the observation error. When discrete-time observations are considered to be measured without error, $x_{wk} = X_w(t_{wk})$, an interpolation method can be used. For the case of nondecreasing stochastic processes, a monotone piecewise cubic interpolation of the sample paths has been proposed (Bouzas et al., 2006). On the other hand, if some error is considered in the observations, $x_{wk} = X_w(t_{wk}) + \varepsilon_{wk}$, least-squares approximation is usually used for estimating the basis coefficients for a specific curve as $a_w = (a_{w1}, \ldots, a_{wp})' = (\Phi'_w \Phi_w)^{-1} \Phi'_w x_w$, with $\Phi_{m_w \times p} = (\phi_j(t_{wk}))$. Taking into account the underlying nature of curves, different basis have been used in literature as for example, Fourier, wavelet or spline functions.

The problem is that likelihood estimation of the parameters of the logit model with design matrix $A\Psi$ is very inaccurate due to multicollinearity, meaning that the estimated parameter function cannot be used to establish the true relationship between the response and predictor variables (Escabias et al., 2004). In the case of the multiple logit regression model, this problem has been solved by using as predictors an optimum set of PCs of the original independent variables (Aguilera et al., 2006). On the other hand, a Gibbs-type sampling scheme for the Bayesian estimation of logistic models has been proposed (Fruhwirth-Schnattera and Fruhwirthb, 2006).

4. Functional PC estimation

In order to reduce dimension and to obtain better estimations of the parameter function, two different approaches have been proposed in Escabias et al. (2004) based on FPCA of sample paths, so that the FLR model is reduced to a multiple one with a reduced number of functional PCs as covariates. We perform an FPCA to the sample paths $X_w(s)$ with respect to the usual inner product in the space of square integrable functions on the interval $(0, T]$, denoted by $L^2((0, T])$.

By analogy with the finite case, functional PCs are uncorrelated generalized linear combinations of a functional variable with maximum variance. Therefore, functional PCs of $X_w(s)$ are defined as N-dimensional vectors ξ_j $(j = 1, \ldots, N-1)$ with components

$$\xi_{wj} = \int_0^T (X_w(s) - \bar{x}(s)) f_j(s)\,\mathrm{d}s, \quad w = 1, \ldots, N,$$

where $\bar{x}(s)$ is the sample mean of the sample curves, the weight functions $f_j(s)$ $(j = 1, \ldots, N-1)$ that define the functional PCs are the eigenfunctions of the sample covariance function $C(t, s)$ of the sample curves

$$\int_0^T C(t, s) f_j(s)\,\mathrm{d}s = \lambda_j f_j(t),$$

and their associated positive eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_{n-1} \geqslant 0$ are the variances of the corresponding PCs.

It can be shown that if the sample paths belong to a finite space of $L^2(0, T]$ generated by a basis, FPCA is equivalent to standard multivariate PCA of the matrix $A\Psi^{1/2}$ (Ocaña et al., 2007). If we denote by $\Gamma = (\xi_{ij})_{N \times p}$ the matrix whose columns are the PCs of the $A\Psi^{1/2}$ matrix, and G the one whose columns are their associated eigenvectors, then $\Gamma = (A\Psi^{1/2})G$ and the weight functions that define the functional PCs are given by

$$f_j(s) = \sum_{k=1}^{p} f_{jk} \phi_k(s), \quad j = 1, \ldots, p, \tag{4}$$

with $F = (f_{jk})_{p \times p} = \Psi^{-1/2} G$.

Functional PCA provides an orthogonal expansion of the functional predictor $\{X(s)\}$ in terms of a set of deterministic functions (the principal factors) and random variables (the PCs). This means that sample curves admit the following orthogonal representation in terms of the sample PCs:

$$X_w(s) = \bar{x}(s) + \sum_{j=1}^{p} \xi_{wj} f_j(s), \quad w = 1, \ldots, N, \tag{5}$$

so that by truncating it we obtain a reconstruction of the sample paths in terms of a reduced number of PCs that accumulates a certain percentage of the total variance defined as $TV = \sum_{j=1}^{N-1} \lambda_j$.

FLR model (2) can be equivalently expressed in terms of the PCs as (Escabias et al., 2004)

$$l_w = \alpha + \sum_{j=1}^{p} \xi_{wj}\gamma_j, \quad w = 1, \ldots, N. \tag{6}$$

In matrix form $L = \alpha\mathbf{1} + \Gamma\gamma$ and the parameter function basis coefficients are given $\beta = \Psi^{-1/2}G\gamma$.

Thus, the functional principal component logistic regression (FPCLR) model is obtained by truncating model (6) in terms of a subset of PCs. If we consider matrices G and Γ partitioned as follows:

$$\Gamma = (\Gamma_{(q)}|\Gamma_{(r)}), \quad G = (G_{(q)}|G_{(r)}), \ \ r + q = p$$

the FPCLR model is defined by taking as covariates the first q PCs

$$L_{(q)} = \alpha_{(q)}\mathbf{1} + \Gamma_{(q)}\gamma_{(q)},$$

where $\alpha_{(q)}$ is a real parameter and $L_{(q)} = (l_{1(q)}, \ldots, l_{N(q)})'$ with

$$l_{w(q)} = \ln\left[\frac{\pi_{w(q)}}{1-\pi_{w(q)}}\right] = \alpha_{(q)} + \sum_{j=1}^{q} \xi_{wj}\gamma_{j(q)}, \quad i = 1, \ldots, N. \tag{7}$$

Finally, the likelihood estimation of the parameter function given by

$$\widehat{\beta}_{(q)}(s) = \sum_{j=1}^{p} \widehat{\beta}_{j(q)}\phi_j(s), \tag{8}$$

with the coefficient vector $\widehat{\beta}_{(q)} = \Psi^{-1/2}G_{(q)}\widehat{\gamma}_{(q)}$ is more accurate than the one obtained with the original $A\Psi$ design matrix (Escabias et al., 2004).

Observe that we have considered as explicative variables of the FPCLR model the first q PCs. However, we know that the PCs with the largest variances are not necessarily the best predictors due to the fact that the PCs with the smallest variances could be highly correlated with the response. Consequently PCs must be introduced in the model according to their ability to explain the response. Two different methods to incorporate PCs in the model have been considered (Escabias et al., 2004): in order of their explained variability, and in the order given by a stepwise method based on conditional likelihood ratio tests. In order to choose the number q of PCs to be incorporated in the model, it has been proved in that a good stopping rule for these two criterion is to choose the model previous to a significant increment in the estimated variance of the reconstructed original parameters (Escabias et al., 2004).

5. Mixed FPC-ARIMA logit model

Once the FPCLR model has been estimated we plan to use it for predicting the probability of success in future periods of time. Notice that the covariates of this model are a reduced set of PCs, and in our case the sample values ξ_{wj} $(w = 1, \ldots, N)$ of each sample PC ξ_j can be seen as observations of a discrete time series at each period $((w-1)T, wT]$ of amplitude T where the original series $x(t)$ was observed. Therefore, in order to forecast the binary response in future periods $((w^*-1)T, w^*T]$ $(w^* > N)$, we propose the modelization of each PC by an ARIMA model (Box and Jenkins, 1970). The ARIMA-model-based methodology of programs TRAMO and SEATS has been recently applied for seasonal adjustment and trend-cycle estimation (Maravall, 2006).

The general expression of an ARIMA(p, d, q) model for the jth PC ξ_j is given by

$$\Phi(B)(1-B)^d \xi_{wj} = \theta(B)\varepsilon_{wj},$$

where B is the backward shift operator, $\Phi(B)$ is the autoregressive operator defined as $\Phi(B) = 1 - \Phi_1 B - \Phi_2 B^2 - \cdots - \Phi_p B^p$, $\theta(B)$ is the moving average operator given by $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_p B^p$, and ε_{wj} is a white noise process for each j $(j = 1, \ldots, q)$.

The prediction model proposed is based on ARIMA forecasting of each of the q PCs selected for estimating the functional logit model. After estimating in the usual way these q ARIMA models, we will be able to obtain forecasts for each PC in the future periods $((w^* - 1)T, w^*T]$, denoted by $\widetilde{\xi}_{w^*j}$.

Thus, the mixed FPC-ARIMA model provides a continuous forecasting of the original series $x(t)$ in the whole interval of time $((w^* - 1)T, w^*T]$ (Valderrama et al., 2002). This prediction is based on the PC reconstruction of the process $\{X(s)\}$ in terms of the predicted PC values

$$\widetilde{x}((w^* - 1)T + s) = \widetilde{X}^q_{w*}(s) = \bar{x}(s) + \sum_{j=1}^{q} \widetilde{\xi}_{w^*j} f_j(s), \quad s \in [0, T].$$

We propose to predict the probabilities of success in the future periods $((w^* - 1)T, w^*T]$ by using the logit transformations

$$\tilde{l}_{w*(q)} = \hat{\alpha}_{(q)} + \sum_{j=1}^{q} \tilde{\xi}_{w^*j} \hat{\gamma}_{j(q)},$$

in terms of the ARIMA forecasts of the PCs $\tilde{\xi}_{w^*j}$.

6. Predicting the risk of drought

Drought is one of the main worries of governments all over the world who are interested in predicting the probability of its future occurrence. Time series analysis methods and state-space modelling have been the most used tools to predict future values of time variables from their observations obtained in the past. When the objective variable is binary, as in the occurrence of drought, classical methods of time series analysis do not work well, therefore we can find several attempts to model binary time series in literature (see Hyndman, 1999 among others). In our mixed FPC-ARIMA logit model we use an external time-dependent variable related to the objective to obtain future predictions.

The relationship between *El Niño* phenomenon and drought in the West Pacific as in Australia has been previously investigated (Lough, 1997; Nicholson and Kim, 1997). The objective of this application of the mixed FPC-ARIMA logit model, is to predict the probability of drought in the future from the prediction of the future evolution of *El Niño* phenomenon. The data consist of monthly observations of the sea surface Temperature in the Equatorial Pacific Ocean (TEPO) (monthly average) from 1950 to 2004 and the accumulated annual precipitation in Melbourne (Australia) over these years. There are several ways of measuring drought, one of which consists of considering that a period of time is dry if the rainfall in this period is lower than a certain percentile of the rainfall observed during a long time (Hyndman and Ullah, 2007). In this sense, in this application we consider that a year is dry in Melbourne (drought), when the observed accumulated rainfall for this year is lower than the 20th percentile of the annual accumulated rainfall observed in the last 54 years (from 1950 to 2004). We then define a binary variable that takes value 1 for one year if the precipitation observed in Melbourne that year is greater than the 20th percentile of the one observed during the period 1950–2004, and 0 if otherwise (Fig. 2 shows this series). Our objective is to predict the probability of absence of drought from the prediction obtained from the mean monthly TEPO by using a mixed FPC-ARIMA logit model.

The successful performance of mixed FPC-ARIMA models by predicting *El Niño* phenomenon has been previously analyzed (Valderrama et al., 2002). The prediction of the mean monthly TEPO from 2000 to 2001 by using mixed FPC-ARIMA models from the one observed from 1950 to 1999 has been studied. In order to predict the probability of absence of drought in years 2000–2004, we use the predictions of the temperature from 2000 to 2004 obtained from the one observed during 1950–1999 and the response (occurrence of rainfall greater than the 20th percentile) from 1950 to 1999.

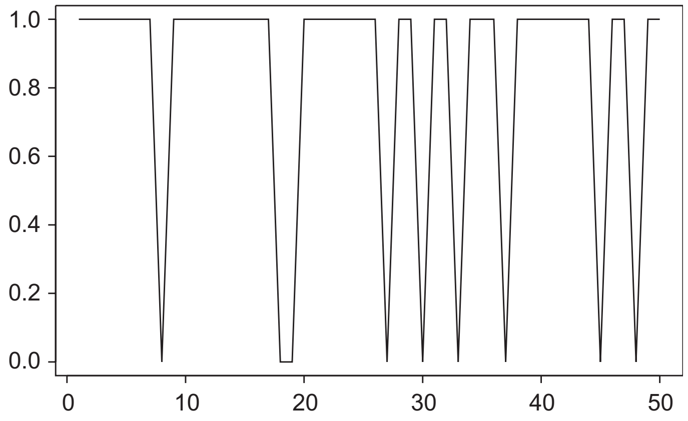

Fig. 2. Time series of the annual occurrence of greater than the 20th percentile of rainfall in Melbourne (Australia) from 1950 to 1999.

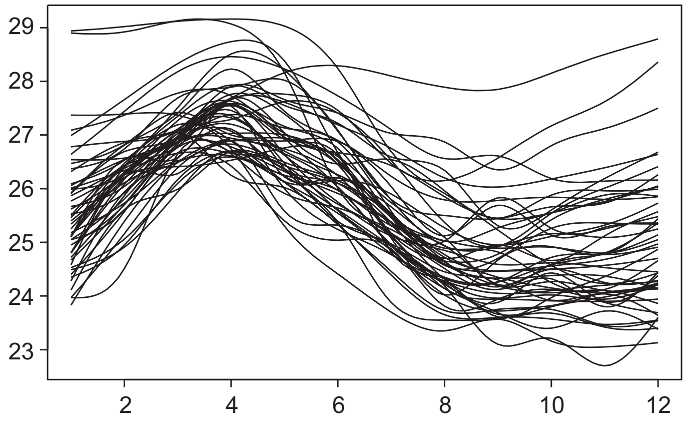

Fig. 3. Quasi-natural cubic spline interpolation of annual TEPO for all years from 1950 to 1999.

We used the discrete observations of the mean monthly TEPO and reconstructed the functional form of the annual evolution of such temperatures by interpolation of those observations for each year (1950–1999) (Valderrama et al., 2002). In this case we use quasi-natural cubic spline interpolation on the discrete observations (see Escabias et al., 2005 for a detailed discussion of this type of interpolation). Fig. 3 shows the interpolated temperatures for all the years.

Once the functional observations of annual evolution of TEPO and the response of absence of drought during the period 1950–1999 were obtained, we estimated FPCA of the functional observations. The variance explained by the PCs can be seen in Table 1 and the first four principal factors can be seen in Fig. 5. We can observe that the first four PCs accumulate more than 95% of the total variability (96.86%).

Following that, we fitted the FPCLR model in terms of different numbers of PCs by including them in the model by the two methods described in Section 4. Table 1 shows some goodness-of-fit measures for the model with different numbers of PCs and with the two methods of inclusion. From such table, we can observe that the estimated variance of the estimated parameter function significantly increases when we include the ninth PC in variability order and in the order given by the stepwise method (the third that enters in the model). The best model would therefore be the one that has the first two PCs included by stepwise method. See Escabias et al. (2004) for a detailed study of selecting PCs in FPCLR.

The ARIMA modelling of the PCs concludes that the first four follow a certain structure while the remaining PCs are white noises. ARIMA modelling of the first four PCs can be seen in Table 2. After modelling the PCs we obtained five steps ahead forecasts (years 2000–2004) of such time series and reconstructed the TEPO for these years from Eq. (5) through the forecasts obtained from the first two PCs, and through those obtained from the first four PCs. Fig. 4 shows the predictions of the TEPO evolution during each of these years and its prediction by the FPC-ARIMA model in terms of the first two and the first four PCs. We can visually appreciate the accuracy of the estimations obtained in

Table 1
Goodness-of-fit measures of the FPCLR models

Variability order					
N. PCs	Ac. Var	CCR	Est. Var	G^2	p-Value
1	68.61	82	2.52E − 01	39.17	0.81
2	91.50	86	1.10E + 00	31.59	0.96
3	94.88	90	3.35E + 00	30.28	0.96
4	96.86	92	9.96E + 00	28.25	0.98
5	98.21	90	1.65E + 01	27.65	0.97
6	98.86	86	4.63E + 01	26.62	0.98
7	99.22	88	1.17E + 02	26.45	0.97
8	99.46	86	3.55E + 02	26.18	0.97
9	99.67	100	2.57E + 13	0.00	1.00
10	99.82	100	6.77E + 13	0.00	1.00
11	99.93	100	6.14E + 13	0.00	1.00
12	100.00	100	7.37E + 13	0.00	1.00
13	100.00	100	5.51E + 13	0.00	1.00
14	100.00	100	6.96E + 13	0.00	1.00

Stepwise order					
N. PCs	PC	CCR	Est. Var	G^2	p-Value
1	ξ_2	84	0.83	38.92	0.82
2	ξ_1	86	1.10	31.59	0.96
3	ξ_9	90	163.01	26.04	0.99
4	ξ_4	88	286.88	20.93	1.00

N. PCs = number of PCs, Ac. Var = accumulated variance, CCR = correct classification rates, Est. Var. = estimated variance of the estimated parameter functions, G^2 = deviance statistic, p-value = p-value corresponding to the deviance statistic and PC = principal component entered in the model in the corresponding step of the stepwise method.

Table 2
ARIMA modelling of the first four PCs and their estimated parameters

PC	ARIMA model	Estimated parameters
ξ_1	$SARIMA(0,0,2) \times (0,1,2)_5$	MA(1) = 0.57, MA(2) = 0.35 SMA(1) = 1.32, SMA(2) = −0.54
ξ_2	$SARIMA(0,0,1) \times (0,1,1)_5$	MA(1) = 0.43, SMA(1) = 0.86
ξ_3	$ARIMA(1,0,1)$	AR(1) = 0.51, MA(1) = 0.82
ξ_4	$SARIMA(0,1,1) \times (0,1,1)_{11}$	MA(1) = 0.90, SMA(1) = 0.75

all cases.

Finally, we obtained the estimated probability of absence of drought in Melbourne (rainfalls over the 20th percentile) through the estimated FPCLR model given by Eq. (2), and by using the TEPO forecast obtained from the functional PCs for the period 2000–2004 in both cases: with the first two and the first four PCs. As accuracy measures for these forecasts, we have considered the usual mean square error (MSE) between observed (absence of drought) and estimated (probability of absence of drought) values and the correct classification rate with cut-point 0.5 (CCR). All these measures can be seen in Table 3 in which we can observe how well the model forecast drought with 80% of correctly classified observations if we use the first four PCs, and 100% if we use the first two.

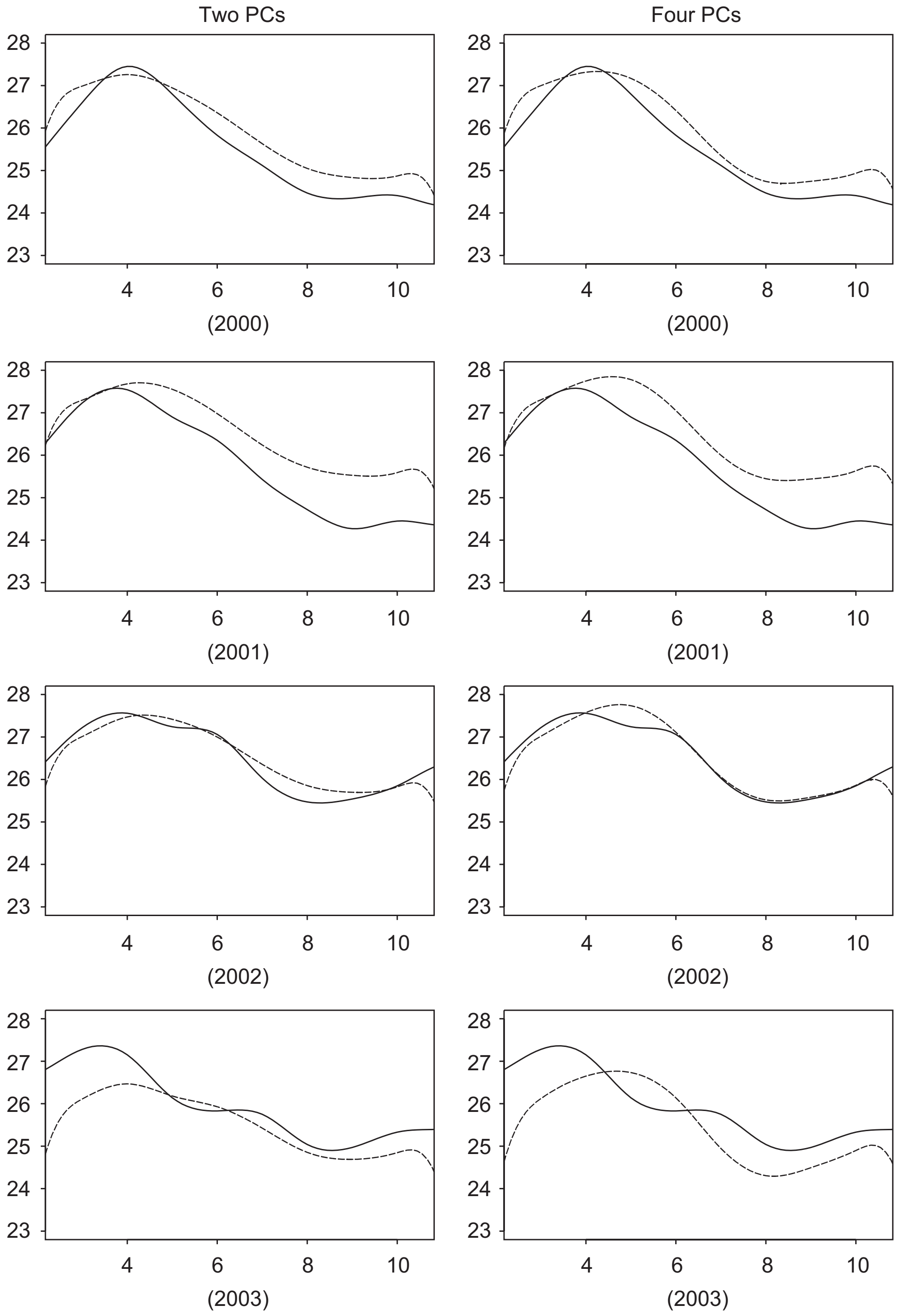

Fig. 4. Quasi-natural cubic spline interpolation of TEPO (dashed line) and its prediction by PC-ARIMA modelling with the first two and four PCs (solid line).

7. Conclusions

We have introduced a new methodology for predicting a binary response in the future from discrete time observations of a related continuous time series in the past. The proposed forecasting model is based on FLR of the binary longitudinal data on the sample curves obtained by cutting the continuous time predictor in periods of the same amplitude. In order to estimate the model and to solve multicollinearity, an FPCA approach that takes as explicative variables a reduced

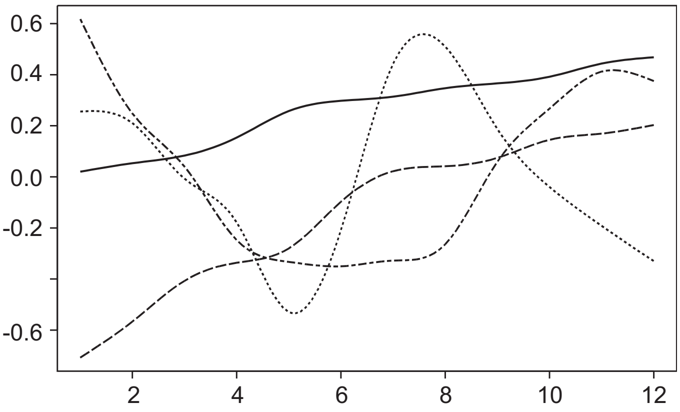

Fig. 5. First four principal factors: first principal factor in solid line, second in dashed line, third in dashed-dotted line and fourth in dotted line.

Table 3
Estimated probabilities and observed absence of drought for period 2000–2004

Year	Observed absence of drought	Estimated with four PCs	Estimated with two PCs
2000	1	0.93	0.66
2001	1	0.91	0.54
2002	0	0.66	0.29
2003	0	0.26	0.21
2004	1	0.84	0.58
MSE		0.11	0.13
CCR		80%	100%

Accuracy measures: mean squared error (MSE) and correct classification rates (CCR).

set of PCs of the functional predictor has been considered. ARIMA modelling of the predictor PCs is used to forecast the continuous time predictor in the future and then the associated binary response.

The forecasting performance of the proposed logit functional PC-ARIMA model has been tested by developing an application with real data where the annual risk of drought has been predicted from monthly observations of *El Niño* phenomenon. From this application we can conclude that the most accurate model (highest CCR $= 100\%$) is obtained by introducing PCs in the model in the order given by stepwise selection until a significant increment in the estimated variance of the estimated parameter function.

Acknowledgments

This research has been supported by projects MTM2007-63793 from *Dirección General de Investigación*, *Ministerio de Educación y Ciencia*, Spain and P06-FQM-01470 from Consejería de Innovación Ciencia y Empresa, Junta de Andalucía, Spain. Thanks to the Australian Commonwealth Bureau of Meteorology: Climate and Consultancy Section in the Victorian Regional Office of the Bureau of Meteorology for providing us with the precipitation data.

References

Aguilera, A.M., Ocaña, F.A., Valderrama, M.J., 1997. An approximated principal component prediction model for continuous-time stochastic processes. Appl. Stochastic Models Data Anal. 13 (1), 61–72.

Aguilera, A.M., Ocaña, F.A., Valderrama, M.J., 1999a. Forecasting with unequally spaced data by a functional principal component approach. Test 8 (1), 233–253.

Aguilera, A.M., Ocaña, F.A., Valderrama, M.J., 1999b. Forecasting time series by functional PCA. Discussion of several weighted approaches. Comput. Statist. 14, 442–467.

Aguilera, A.M., Escabias, M., Valderrama, M.J., 2006. Using principal components for estimating logistic regression with high-dimensional multicollinear data. Comput. Statist. Data Anal. 50 (8), 1905–1924.

Bouzas, P.R., Valderrama, M.J., Aguilera, A.M., Ruiz-Fuentes, N., 2006. Modelling the mean of a doubly stochastic Poisson process by functional data analysis. Comput. Statist. Data Anal. 50 (10), 2655–2667.

Box, G.E.P., Jenkins, G.M., 1970. Time Series Analysis. Forecasting and Control. Holden Day, San Francisco.

Cox, D.R., Snell, E.J., 1989. Analysis of Binary Data. second ed. Chapman & Hall, London.

Escabias, M., Aguilera, A.M., Valderrama, M.J., 2004. Principal component estimation of functional logistic regression: discussion of two different approaches. J. Nonparametric Statist. 16, 365–384.

Escabias, M., Aguilera, A.M., Valderrama, M.J., 2005. Modelling environmental data by functional principal component logistic regression. Environmetrics 16 (1), 95–107.

Escabias, M., Aguilera, A.M., Valderrama, M.J., 2007. Functional PLS logit regression model. Comput. Statist. Data Anal. 51 (10), 4891–4902.

Ferraty, F., Vieu, P., 2003. Curves discrimination: a nonparametric functional approach. Comput. Statist. Data Anal. 44, 161–173.

Fruhwirth-Schnattera, S., Fruhwirthb, R., 2006. Auxiliary mixture sampling with applications to logistic models. Comput. Statist. Data Anal. 51 (7), 3509–3528.

González-Manteiga, W., View, P., 2007. Statistics for functional data. Comput. Statist. Data Anal. 51 (10), 4788–4792.

Heij, C., Groenen, P.J.F., Dijk, D., 2007. Forecast comparison of principal component regression and principal covariate regression. Comput. Statist. Data Anal. 51 (7), 3612–3625.

Hyndman, R.J., 1999. Nonparametric additive regression models for binary time series. In: Proceedings, 1999 Australasian Meeting of the Econometric Society, 7–9 July 1999, University of Technology, Sydney.

Hyndman, R.J., Ullah, Md.S., 2007. Robust forecasting of mortality and fertility rates: a functional data approach. Comput. Statist. Data Anal. 51 (10), 4942–4956.

James, G.M., 2002. Generalized linear models with functional predictors. J. Roy. Statist. Soc. Ser. B 64 (3), 411–432.

Lough, J.M., 1997. Regional indexes of climate variation temperature and rainfall in Queensland, Australia. Internat. J. Climatology 17 (1), 55–66.

Maravall, A., 2006. An application of the TRAMO-SEATS automatic procedure; direct versus indirect adjustment. Comput. Statist. Data Anal. 50 (9), 2167–2190.

Nicholson, S.E., Kim, E., 1997. The relationship of the *El Niño* southern oscillation to African rainfall. Internat. J. Climatology 17 (2), 117–135.

Ocaña, F.A., Aguilera, A.M., Escabias, M., 2007. Computational considerations in functional principal component analysis. Comput. Statist. 22 (3), 449–465.

Ombaoa, H., Hob, M.R., 2006. Time-dependent frequency domain principal components analysis of multichannel non-stationary signals. Comput. Statist. Data Anal. 50 (9), 2339–2360.

Preda, P., Saporta, G., 2005. PLS regression on a stochastic process. Comput. Statist. Data Anal. 48 (1), 149–158.

Preda, P., Saporta, G., Lévéder, C., 2007. PLS classification of functional data. Comput. Statist. 22 (2), 223–235.

Ramsay, J.O., Silverman, B.W., 2002. Applied Functional Data Analysis. Springer, New York.

Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis. second ed. Springer, New York.

Ratcliffe, S.J., Leader, L.R., Heller, G.Z., 2002. Functional data analysis with application to periodically stimulated foetal heart rate data. II: functional logistic regression. Statist. Med. 21 (8), 1115–1127.

Valderrama, M.J., Aguilera, A.M., Ocaña, F.A., 1997. Predicción Dinámica Mediante Análisis de Datos Funcionales. La Muralla-Hespérides, Madrid.

Valderrama, M.J., Ocaña, F.A., Aguilera, A.M., 2002. Forecasting PC-ARIMA models for functional data. In: Härdle, W., Rönz, B. (Eds.), Proceedings in Computational Statistics. Physica-Verlag, New York, pp. 25–36.