APPLIED STOCHASTIC MODELS AND DATA ANALYSIS, VOL. 13, 61–72 (1997)

AN APPROXIMATED PRINCIPAL COMPONENT PREDICTION MODEL FOR CONTINUOUS-TIME STOCHASTIC PROCESSES

ANA M. AGUILERA, FRANCISCO A. OCAÑA AND MARIANO J. VALDERRAMA*

Department of Statistics and Operations Research, University of Granada, 18071-Granada, Spain

SUMMARY

In this paper, a linear model for forecasting a continuous-time stochastic process in a future interval in terms of its evolution in a past interval is developed. This model is based on linear regression of the principal components in the future against the principal components in the past. In order to approximate the principal factors from discrete observations of a set of regular sample paths, cubic spline interpolation is used. An application for forecasting tourism evolution in Granada is also included.

Appl. Stochastic Models Data Anal., Vol. 13, 61–72 (1997)
No. of Figures: 3 No. of Tables: 4 No. of References: 18

KEY WORDS: Principal components; Karhunen–Loève expansion; least squares linear prediction; cubic B-splines

1. INTRODUCTION

As a natural continuation of the paper by Aguilera *et al.*,[1] the purpose of the present paper is to develop a dynamic model for forecasting the degree of hotel occupation in Spanish cities in a future interval $[T_3, T_4]$ in terms of its evolution in a past interval $[T_1, T_2]$ $(T_2 < T_3)$.

In order to solve this problem, we have a data set corresponding to discrete-time observations of a set of regular curves which can be stochastically modelled as independent sample paths of a second order and quadratic mean continuous stochastic process, $\{X(t): t \in [T_1, T_4]\}$, whose sample paths have squares integrable on $[T_1, T_4]$.

The main problem is to solve the linear pure prediction problem, which consists in estimating the future, $\{X(s): s \in [T_3, T_4]\}$, of such a process given its recent past, $\{X(t): t \in [T_1, T_2]\}$. The forecasting model that we are going to provide is an extension of the topic of 'principal component regression' (see, for example, the work of Jackson[2]) to forecast the future as a function of some of the principal components associated with an infinite number of predictor variables.

*Correspondence to: M. J. Valderrama Department of Statistics and Operations Research, University of Granada, Campus de Cartuja 18071-Granada, Spain
Contract grant sponsor: DGICYT, Ministerio de Educación y Giencia, Spain; contract grant number: PS94-0136

CCC 8755–0024/97 020061–12$17.50

Received 11 June 1995
Revised 30 May 1996

By analogy with the finite case, Deville[3] defined the principal components associated with the process $\{X(t)\colon\ t\in[T_1, T_2]\}$ as uncorrelated generalized linear combinations of the process variables having maximum variance. That is, the ith principal component (p.c.) is defined as

$$\xi_i = \int_{T_1}^{T_2} (X(t) - \mu(t))\, f_i(t)\, \mathrm{d}t \tag{1}$$

where f_i, called the ith principal factor (p.f.), is the normalized eigenfunction corresponding to the ith largest eigenvalue λ_i of the covariance kernel $C(t, s)$ and denoting by $\mu(t)$ the mean function of the process.

Thus, the variance explained by the ith principal component is $V_i = \lambda_i/V$, and

$$V = E\left[\int_{T_1}^{T_2} (X(t) - \mu(t))^2\, \mathrm{d}t\right] = \sum_i \lambda_i$$

is the total variance of the process in $[T_1, T_2]$.

Then the process admits the following principal component decomposition, which is known, in the probabilistic context, as the Karhunen–Loève orthogonal expansion (see, for example, Todorovic[4]):

$$X(t) - \mu(t) = \sum_i \xi_i f_i(t), \quad t\in[T_1, T_2] \tag{2}$$

where the infinite series in (2) converges in quadratic mean to $X(t)$ uniformly in t.

Moreover, the series (2) truncated in the qth term is the best q-dimensional linear model for the process $X(t)$ in the least squares sense (see, for example, the work of Fukunaga[5]), $(\sum_{i=1}^{q} \lambda_i)/V$ being the variance explained by this model and $\sum_{i=q+1}^{\infty} \lambda_i$ the minimum mean square error.

In Section 2, a linear model for forecasting the process by means of linear regression of the p.c.'s in the future against the p.c.'s in the past is developed.

In Section 3, the identification and estimation of such a prediction model from a set of independent sample paths are described. The natural estimators of the p.f.'s are the solutions of a second-kind integral equation whose kernel is the sample covariance function (Deville[3]). It usually happens in practice that the sample paths are only observed in a finite set of times. In this case the approximation of the PCA has been approached by many authors (for example, Deville,[3] Saporta,[6] Besse and Ramsay,[7] Castro *et al.*, [8] Bouhaddou *et al.*,[9] Besse[10] and Ramsay and Dalzell,[11] among others). The simplest numerical method for solving this problem consists in approximating the integral equation by a quadrature formula such as, for example, the trapezoid rule (Aguilera *et al.*[12]). A more complex method is to approximate the sample p.f.'s in a finite-dimension subspace of trigonometric functions (Aguilera *et al.*[13]). In the paper by Aguilera *et al.*,[1] the authors tested the accuracy of these approximation methods by simulating sample functions of random processes whose principal factors can be directly evaluated (for example, Brownian motion and Brownian bridge stochastic processes). Besse and Ramsay[7] solve this problem by making the assumption that the sample functions lie in a reproducing kernel Hilbert space. In this paper, the authors approximate the sample p.f.'s by means of cubic spline interpolation of the sample paths between the observed data (Aguilera and Valderrama[14]).

Finally, in Section 4 the proposed model is applied to forecasting tourism evolution in Granada. The forecasts are compared with those obtained by fitting a classic ARIMA model to the degree of hotel occupation series.

2. LINEAR PRINCIPAL COMPONENT PREDICTION

Let us consider that the process $X(t)$ is defined on a probabilistic space $(\Omega, \mathscr{A}, P)$. It will also be assumed without loss of generality that $\mu(t) = E[X(t)] = 0$ and it will be denoted by L_X^2 the closed linear manifold spanned by the r.v.'s $\{X(t)\colon T_1 \leqslant t \leqslant T_2\}$.

It is known that the mean square linear estimate of $X(s)$ $(s > T_2)$ given $\{X(t)\colon T_1 \leqslant t \leqslant T_2\}$ is the orthogonal projection of $X(s)$ onto the subspace L_X^2.

As the p.c.'s of the process $\{X(t)\colon t \in [T_1, T_2]\}$ make up a complete orthogonal family in L_X^2, the minimum mean square error linear estimator of $X(s)$ admits, for every $s > T_2$, the following expansion, which is convergent in quadratic mean (Deville[15]):

$$\tilde{X}(s) = \sum_{i=1}^{\infty} \frac{E[X(s)\xi_i]}{\lambda_i}\,\xi_i, \quad s > T_2 \tag{3}$$

On the other hand, principal component analysis gives the following orthogonal representation of the process in the future interval $[T_3, T_4]$:

$$X(s) = \sum_{j=1}^{\infty} \eta_j g_j(s) \tag{4}$$

where g_j and η_j denote the principal factors and components, respectively, in the interval $[T_3, T_4]$.

Then, using this representation of the process in the future, the mean square linear estimator $\tilde{X}(s)$ can be written as

$$\tilde{X}(s) = \sum_{j=1}^{\infty} \tilde{\eta}_j g_j(s), \quad s \in [T_3, T_4] \tag{5}$$

where $\tilde{\eta}_j$ is the least squares linear estimate of the p.c. η_j given the process variables $\{X(t)\colon T_1 \leqslant t \leqslant T_2]\}$, i.e.

$$\tilde{\eta}_j = \sum_{i=1}^{\infty} \beta_i^j \xi_i \tag{6}$$

where

$$\beta_i^j = \frac{E[\eta_j \xi_i]}{\lambda_i} \tag{7}$$

By truncating off each of the infinite series in equation (6), the following approximated linear prediction for each of the p.c.'s in the future is obtained:

$$\tilde{\eta}_j^{p_j} = \sum_{i=1}^{p_j} \beta_i^j \xi_i \tag{8}$$

Let us observe that last equation represents the least squares linear regression model of each p.c. η_j against the first p_j p.c.'s $\{\xi_i\}_{i=1}^{p_j}$ as predictors variables.

Finally, the following Principal Component Prediction (PCP) model for the process in the future is constructed:

$$\tilde{X}^q(s) = \sum_{j=1}^{q} \tilde{\eta}_j^{p_j} g_j(s), \quad s \in [T_3, T_4] \tag{9}$$

in terms of the first q p.c.'s in the future whose cumulative variance is as close as possible to one. This model will be denoted by PCP $(q; p_1, \ldots, p_q)$.

3. ESTIMATION

Now the proposed PCP model is going to be estimated from given data.

Let us suppose that we have discrete observations, denoted by

$$\{X_w(t_j): w = 1, \ldots, N; j = 0, \ldots, m\}$$

of N independent sample paths $\{X_w(t): w = 1, \ldots, N\}$ of the continuous-time stochastic process $\{X(t): t \in [T_1, T_4]\}$.

In order to forecast the process for $s \in [t_{k+1}, t_m]$ $(k = 1, \ldots, m-2)$, we will consider in the observed period $[t_0, t_m]$ two disjoint intervals, $[t_0, t_k]$ and $[t_{k+1}, t_m]$, as the past and the future, respectively, denoting $T_1 = t_0$, $T_2 = t_k$, $T_3 = t_{k+1}$ and $T_4 = t_m$.

Then the first step consists in estimating the principal factors in each interval.

3.1. Approximating the principal factors using cubic B-splines

The eigensystem (λ_i, f_i) of the covariance kernel $C(t, s)$ in $[T_1, T_2]$ is estimated by the corresponding one $(\hat{\lambda}_i, \hat{f}_i)$ of its usual unbiased estimator $\hat{C}(t, s)$, called the sample covariance kernel and defined as

$$\hat{C}(t, s) = \frac{1}{N-1} \sum_{w=1}^{N} \Big(X_w(t) - \bar{X}(t)\Big)\Big(X_w(s) - \bar{X}(s)\Big) \tag{10}$$

where $\bar{X}$ is the usual unbiased estimate of the mean μ. A detailed study of the properties of these estimators can be seen in the work of Deville.[16]

Therefore, the estimated eigenfunctions, denoted by $\hat{f}_i$ and called sample principal factors, are the solutions of the integral equation:

$$\int_{T_1}^{T_2} \hat{C}(t, s)\hat{f}_i(s)\,\mathrm{d}s = \hat{\lambda}_i \hat{f}_i(t), \quad t \in [T_1, T_2] \tag{11}$$

Now the problem consists in solving the integral equation (11) from discrete observations of the sample paths. For processes whose sample paths are regular, it has been confirmed by simulation that approximating the sample principal factors by those of the natural cubic splines interpolating the sample paths between the observed data gives accurate results (Aguilera,[17] Aguilera and Valderrama[14]). A brief summary of this approximation method is presented below.

Denoting by $\{B_p(t)\}_{p=-1}^{k+1}$ the cubic B-splines in $[T_1, T_2]$ with knots t_j $(j = 0, \ldots, k)$, the natural cubic spline interpolation of each sample function $X_w(t)$ at the points $(t_j, X_w(t_j))$ can be written as

$$IX_w(t) = \sum_{p=-1}^{k+1} a_{wp} B_p(t) \tag{12}$$

where the coefficients a_{wp} $(p = -1, 0, \ldots, k+1)$ are, for each $w = 1, \ldots, N$ the solutions of the linear system

$$\sum_{p=-1}^{k+1} a_{wp} B_p''(t_0) = (X)_w''(t_0)$$

$$\sum_{p=-1}^{k+1} a_{wp} B_p(t_j) = X_w(t_j) \quad j = 0, 1, \ldots, k$$

$$\sum_{p=-1}^{k+1} a_{wp} B_p''(t_k) = (X)_w''(t_k)$$

denoting by $X''(t)$ the second derivative in quadratic mean of the process.

Then, it can be proved (Aguilera[17]) that the principal factors of the interpolated sample paths are cubic splines

$$\hat{f}_i^I(t) = \sum_{p=-1}^{k+1} \alpha_{pi} B_p(t) \tag{13}$$

being the $(k+3)$-dimensional column vector $\underline{\alpha}_i$ the ith eigenvector of the $(k+3)\times(k+3)$-dimensional matrix $\mathbf{T}\hat{\mathbf{\Sigma}}\mathbf{T}'\mathbf{P}$, where $\mathbf{P}$ is the matrix whose elements are the natural inner products between B-splines, $\mathbf{T}$ is the matrix of interpolation over the B-splines, and $\hat{\mathbf{\Sigma}}$ is the sample covariance matrix of the random vector $(X''(t_0), X(t_0), \ldots, X(t_k), X''(t_k))'$.

Once we have computed the approximated principal factors, $\hat{f}_i^I$, their associated p.c.'s $\hat{\xi}_i^I$ are given by

$$\hat{\xi}_{wi}^I = \int_{T_1}^{T_2} (IX_w(t) - \overline{IX}(t)) \hat{f}_i^I(t) \mathrm{d}t = \sum_{l=-1}^{k+1} (a_{wl} - \bar{a}_l) \sum_{p=-1}^{k+1} P_{lp} \alpha_{pi} \tag{14}$$

where

$$\bar{a}_l = \frac{1}{N} \sum_{w=1}^{N} a_{wl}$$

In order to estimate the PCA of a continuous process we have developed our own software called PCAP program (Aguilera *et al.*[18]). It has been coded in Turbo Pascal using object oriented programming and provides different numerical methods for approximating the sample p.f.'s from discrete observations.

3.2. Estimating the PCP model

The identification and the estimation of the PCP model from discrete observations of a set of regular curves is summarized in the following steps.

Step 1. Approximate the sample principal factors in each interval by the above method, and denoting by

$$\hat{f}_i^I \ (i = 1, \ldots, k+3) \quad \text{and} \quad \hat{g}_i^I \ (i = 1, \ldots, m-k+2)$$

the sample p.f.'s in the intervals $[T_1, T_2]$ and $[T_3, T_4]$, respectively, and estimate their associated principal components in each interval according to expression (14). Similarly,

$$\hat{\xi}_{wi}^I \ (i = 1, \ldots, k+3) \quad \text{and} \quad \hat{\eta}_{wi}^I \ (i = 1, \ldots, m-k+2)$$

will denote the sample p.c.'s in the intervals $[T_1, T_2]$ and $[T_3, T_4]$ respectively, for every sample individual $w = 1, \ldots, N$.

Step 2. Choose the optimal number q of p.c.'s $\hat{\eta}_j^I$ to be introduced in the PCP model as response variables. The simplest criterion is used, which consists in specifying a cut-off (somewhere between 80 and 99%) and retaining the first q p.c.'s whose percentage of cumulative variance is greater than or equal to this cut-off. Some of the commonly used rules for deciding how many p.c.'s should be retained can be found in the work of Jackson.[2]

Step 3. Select the optimum p_j p.c.'s $\hat{\xi}_i^I$ to be introduced in the PCP model as predictors of each of the p.c.'s, $\hat{\eta}_j^I (j = 1, \ldots, p)$. It has been proved that there is no reason for the p.c.'s with the largest variances to be the best predictors (see, for example, the work of Jackson[2]). Indeed, it would be possible for some of the p.c.'s with the smallest variances in the past interval $[T_1, T_2]$ to be highly correlated with the p.c.'s in the future interval $[T_3, T_4]$.
In order to find the best predictors, compute the square of the sample linear correlation between each of the p.c.'s $\hat{\eta}_j^I$ and each of the p.c.'s $\hat{\xi}_i^I$. Then, for each p.c. $\hat{\eta}_j^I$, select those p_j p.c.'s $\hat{\xi}_i^I$ having significantly high correlation as its best predictors.

Step 4. Once the response and predictors p.c.'s have been identified, estimate in the usual form, the linear regression model for each p.c. $\hat{\eta}_j^I$ $(j = 1, \ldots, q)$ against its predictors $\hat{\xi}_i^I$ $(i = 1, \ldots, p_j)$ chosen in Step 3:

$$\tilde{\eta}_j^{Ip_j} = \sum_{i=1}^{p_j} \hat{\beta}_i^j \hat{\xi}_i^I$$

Step 5. Finally, estimate the identified PCP model as

$$\tilde{X}^q(s) = \bar{X}(s) + \sum_{j=1}^{q} \tilde{\eta}_j^{Ip_j} \hat{g}_j^I(s), \quad s \in [T_3, T_4] \tag{15}$$

Step 6. For every new sample individual w observed only at knots $\{t_j\}_{j=0}^k$, obtain a forecast $\tilde{X}_w^q(s)$ for every $s \in [T_3, T_4]$ after approximating its principal components in the past by expression (14).

4. FORECASTING TOURISM EVOLUTION

In this section, the proposed PCP model will be applied to forecast the degree of hotel occupation in Granada in the last four-month period of 1994 from its monthly observations during the 1974–1994 period. Finally, these forecasts will be compared with those obtained by fitting an adequate ARIMA model to this series and with the real observations.

4.1. Estimated PCP model

The real data set, which appears in Table I, represents the degree of hotel occupation in Granada at the end of each month during 1974–1994 and has been provided by the *Instituto Nacional de Estadística*: 'Movement of travellers in tourist establishments'. To forecast this process in the last third of 1994, the authors have considered two different periods in the year; $[T_1, T_2]$ and $[T_3, T_4]$ where $T_1 =$ January, $T_2 =$ August, $T_3 =$ September and $T_4 =$ December. The estimates have been computed using twenty sample paths corresponding to the monthly observations for each year from 1974 to 1993.

Firstly, assuming regular sample paths, the sample principal factors have been approximated in each period as proposed in Section 3 by using the PCAP program. The eigenvalues and the percentages of variance explained by their associated p.c.'s figure in Table II. Let us observe that, in the second period, the first two p.c.'s explain more than 95% of the total variability and that only the first principal component explains 89.34%. This implies that the PCP model should be constructed with no more than the first two p.c.'s as response variables.

Table I. Degree of hotel occupation on Granada during 1974–1994

	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
1974	28·90	27·40	35·30	42·00	38·30	33·50	46·00	58·10	44·60	33·20	25·80	25·10
1975	26·95	30·37	32·90	37·33	39·09	33·58	44·92	53·66	40·73	30·66	23·38	23·92
1976	22·11	23·79	26·34	36·50	30·97	27·01	36·51	46·11	34·89	26·10	21·48	20·35
1977	23·33	31·69	36·37	48·07	35·58	34·14	50·18	59·72	46·28	37·70	32·40	26·42
1978	32·36	35·66	48·33	50·26	50·94	43·11	57·76	69·94	51·89	44·50	35·28	28·73
1979	33·02	34·92	41·46	54·82	45·73	39·93	47·66	59·50	43·05	43·34	33·73	28·89
1980	29·83	32·82	36·14	54·10	43·37	39·39	39·30	58·40	42·56	40·22	29·99	34·42
1981	28·64	28·23	34·02	45·53	46·49	34·19	38·97	57·46	43·99	39·79	29·94	26·96
1982	30·23	31·37	35·72	51·99	43·88	31·77	33·08	55·21	44·19	37·92	29·73	28·38
1983	24·94	31·96	41·73	44·01	39·66	35·42	41·61	56·22	45·91	38·77	29·80	28·57
1984	31·80	31·98	34·31	48·82	43·41	41·97	43·34	62·11	49·03	44·20	33·75	28·98
1985	26·90	32·97	41·15	49·84	45·96	40·49	44·29	56·83	50·06	43·41	30·85	24·71
1986	31·48	33·29	43·66	46·78	50·62	36·76	39·58	54·48	50·58	40·53	31·63	29·53
1987	28·41	32·90	39·42	55·23	49·00	40·40	38·64	52·11	47·07	45·39	33·19	33·82
1988	29·88	37·25	45·69	46·09	45·45	38·29	38·12	52·86	48·38	44·02	35·45	30·32
1989	34·68	38·11	51·57	51·05	51·40	39·57	37·96	54·39	46·39	46·44	37·37	36·00
1990	35·04	40·37	45·71	52·80	47·37	41·12	34·63	46·66	44·68	47·77	41·14	36·12
1991	31·71	35·78	49·39	47·24	45·57	38·63	36·12	47·56	47·02	43·62	35·70	33·74
1992	37·05	42·07	50·14	53·01	50·16	41·17	32·58	46·03	54·32	45·49	32·92	28·49
1993	25·61	37·23	38·13	42·86	43·02	33·54	30·93	41·69	40·16	42·55	29·66	35·79
1994	31·89	42·38	50·16	48·54	50·31	40·22	39·51	46·30	51·79	47·90	31·95	33·05

Table II. Principal values and percentages of variance explained by their associated principal components

	January–August		September–December	
No. of p.c.	Principal values	Explained variance (%)	Principal values	Explained variance (%)
1	110·6021	54·19	61·5292	89·34
2	64·6947	31·70	5·2366	7·60
3	15·2420	7·47	1·3688	1·99
4	5·9107	2·90	0·7401	1·07
5	3·9858	1·95	0·0000	0·00
6	1·8701	0·92	0·0000	0·00
7	1·1191	0·55		
8	0·6882	0·34		
	Total variance = 204·1126		Total variance = 68·8747	

Table III. Squares of the linear correlations between the two sets of p.c.'s

	$\hat{\xi}_1^I$	$\hat{\xi}_2^I$	$\hat{\xi}_3^I$	$\hat{\xi}_4^I$	$\hat{\xi}_5^I$	$\hat{\xi}_6^I$	$\hat{\xi}_7^I$	$\hat{\xi}_8^I$
$\hat{\eta}_1^I$	0·86229	0·00028	0·00327	0·00358	0·00548	0·00638	0·00002	0·02628
$\hat{\eta}_2^I$	0·01603	0·12264	0·00003	0·08266	0·00031	0·01371	0·00078	0·00441
$\hat{\eta}_3^I$	0·00349	0·15272	0·07279	0·06584	0·01923	0·02179	0·05881	0·06875
$\hat{\eta}_4^I$	0·00755	0·00028	0·08544	0·00201	0·09120	0·00774	0·05939	0·00197

Secondly, linear correlations between the principal components in the two different intervals have been computed. The squares of these correlation coefficients appear in Table III. Let us observe that, in this case, the p.c. with the largest variance in the first period is highly correlated with the first one in the second period. Moreover, the remaining correlations are too small to be considered.

Therefore, after estimating the first two p.c.'s $\hat{\eta}_1^I$ and $\hat{\eta}_2^I$ as

$$\tilde{\hat{\eta}}_1^{I_1} = 0{\cdot}6926\hat{\xi}_1^I; \quad \tilde{\hat{\eta}}_2^{I_1} = -0{\cdot}0996\hat{\xi}_2^I$$

we have considered the following PCP models:

$$PCP(1;\,1) \quad \tilde{X}^1(s) = \bar{X}(s) + \tilde{\hat{\eta}}_1^{I_1}\hat{g}_1^I(s)$$

$$PCP(2;\,1,\,1) \quad \tilde{X}^2(s) = \bar{X}(s) + \tilde{\hat{\eta}}_1^{I_1}\hat{g}_1^I(s) + \tilde{\hat{\eta}}_2^{I_1}\hat{g}_2^I(s), \quad s \in [T_3,\, T_4]$$

Although the linear correlation between $\hat{\eta}_2^I$ and $\hat{\xi}_2^I$ is not significant, we have fitted the model PCP (2; 1, 1) in order to make a comparison with the model PCP (1; 1).

In order to choose the best of these models, the mean-square prediction error

$$MSE(s) = \frac{1}{N-1}\sum_{w=1}^{N}(IX_w(s) - \tilde{X}^q(s))^2, \quad s \in [T_3,\, T_4]$$

has been computed and displayed in Figure 1 for the two PCP models.

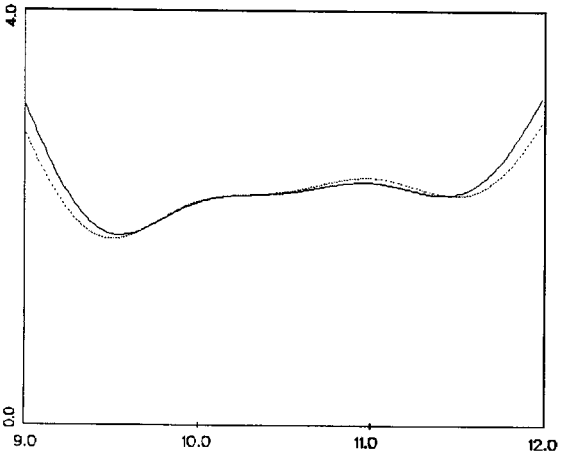

Figure 1. Mean-square prediction error of PCP(1; 1) (solid curve) and of PCP (2; 1, 1) (dotted curve)

In addition, in Figure 2 the natural cubic spline interpolation of the degree of occupation in the future is displayed for four different years superposed with its smoothing by the adjusted PCP models.

4.2. Estimated ARIMA model

Now, a suitable ARIMA model for the degree of occupation series from January 1974 to August 1994 has been built. It is clear that tourist data have a 12 month seasonality. This implies that lag 12 differencing should be used. Moreover, by differencing at lag 1 this series is converted to a stationary one. Following the Box–Jenkins methodology, it has been identified as the most adequate ARIMA model:

$$(1 - B^{12})(1 - B)X(t) = (1 - \Theta_1 B)(1 - \Theta_{12} B^{12})\varepsilon(t)$$

where B is the backshift operator. The left-hand side of this model corresponds to differencing at lags 1 and 12, and the right-hand side is a multiplicative MA at lags 1 and 12. We will denote this model by $\text{SARIMA}(0, 1, 1) \times (0, 1, 1)_{12}$. The parameters, which have been estimated by a conditional least squares method using program 2T of BMDP, are given by: $\hat{\Theta}_1 = 0{\cdot}5866$ and $\hat{\Theta}_{12} = 0{\cdot}6015$.

4.3. Discussion and conclusions

In order to make a comparison between the proposed forecasting models, we have supposed that the degree of hotel occupation in Granada is unknown in the last four months of 1994. The monthly forecasts provided by the adjusted models and the real values appear in Table IV. In Figure 2 the scatter-plot is displayed of the real degree of occupation superposed with its prediction by the three adjusted forecasting models.

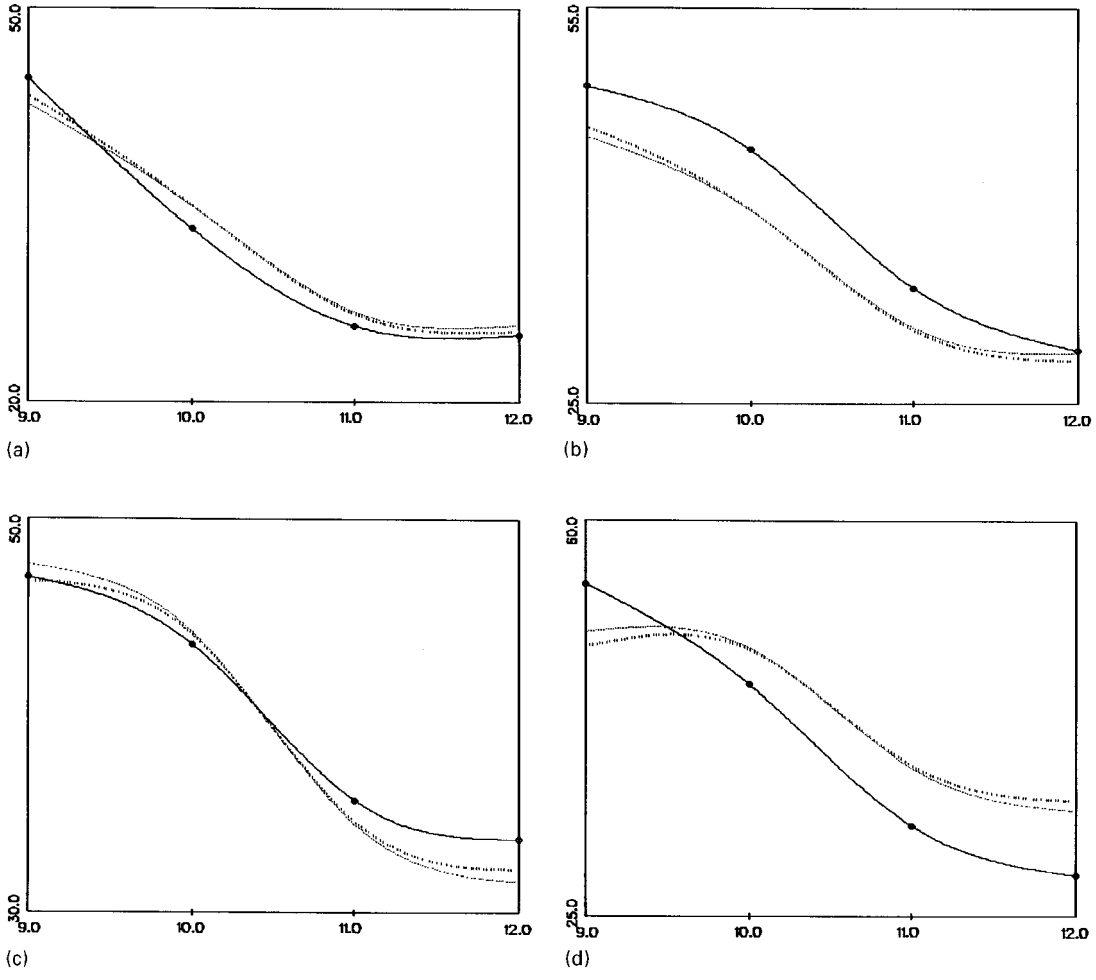

Figure 2. Scatter-plot, natural cubic spline interpolation (solid curves) and smoothing by PCP(1; 1) (thin dotted curves) and by PCP(2; 1, 1) (bold dotted curves) for: (a) 1974; (b) 1984; (c) 1991; (d) 1992

From Table IV and Figure 3, it is observed that the predictions given by the PCP models are better than those given by the SARIMA model, except for November where the SARIMA model gives a slightly better forecast. As expected, in Figure 1 it can be observed that the prediction error of the PCP models is bigger at the end of the interval. On the other hand, the differences between the forecasts and errors provided by the two PCP models are not significant. For this reason the best model to choose is that having the fewest parameters, that is PCP(1; 1).

To conclude, the authors have developed an approximated linear prediction which is valid even for non-stationary continuous processes. One of the main advantages of the PCP model is that it allows the interpolation of the process between the observations. In addition, the use of PCP avoids the multicollineality problem, which appears when using the ordinary least squares prediction model with the process variables $\{X(t_j): j = 0, \ldots, k\}$, which are highly correlated, as

Table IV. Real and forecasted degree of hotel occupation in Granada in 1994

	Sep.	Oct.	Nov.	Dec.
Real	51·79	47·90	31·95	33·05
PCP (1; 1)	49·34	47·32	36·95	33·42
PCP (2; 1, 1)	48·78	47·26	37·08	33·82
SARIMA $(0, 1, 1)\times(0, 1, 1)_{12}$	48·59	46·77	35·71	35·93

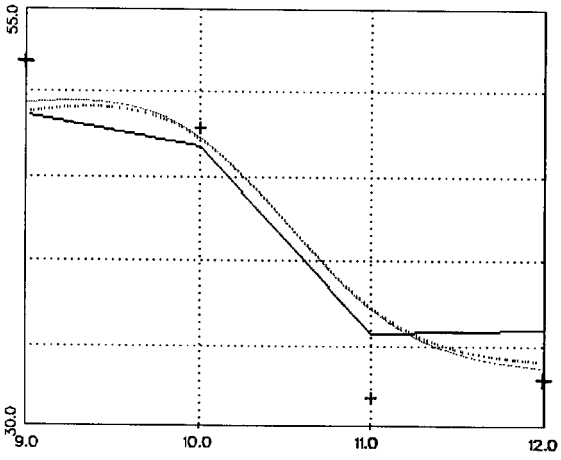

Figure 3. Scatter-plot (crosses) and forecast by SARIMA (solid curve), by PCP (1; 1) (thin dotted curve) and by PCP (2; 1, 1) (bold dotted curve) for the degree of hotel occupation in Granada, September–December 1994

predictors. Finally, if the principal components can be easily interpreted, the resultant regression equations may be more meaningful.

ACKNOWLEDGEMENTS

This research was supported in part by Project PS94-0136 of DGICYT, Ministerio de Educación y Ciencia, Spain.

REFERENCES

1. A. M. Aguilera, R. Gutiérrez, F. A. Ocaña and M. J. Valderrama, 'Computational approaches to estimation in the principal component analysis of a stochastic process', *Appl. Stochast. Models & Data Anal.*, **11**, 279–299 (1995).
2. J. E. Jackson, *A User's Guide to Principal Components*, Wiley, 1991.
3. J. C. Deville, 'Méthodes statistiques et numériques de l'analyse harmonique', *Annales de l'INSEE*, **15**, 3–101 (1974).
4. P. Todorovic, *An Introduction to Stochastic Processes and their Applications*, Springer-Verlag, 1992.
5. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1990.
6. G. Saporta, 'Data analysis for numerical and categorical individual time-series', *Appl. Stochast. Models & Data Anal.*, **1**, 109–119 (1985).

7. P. Besse and J. O. Ramsay, 'Principal component analysis of sample functions', *Psychometrika*, **51**(2), 285–311 (1986).
8. P. E. Castro, W. H. Lawton and E. A. Sylvestre, 'Principal modes of variation for processes with continuous sample curves', *Technometrics*, **28**(4), 329–337 (1986).
9. O. Bouhaddou, C. Obled and T. P. Dinh, 'Principal component analysis and interpolation of stochastic processes: methods and simulation', *J. Appl. Statistics*, **14**(3), 251–267 (1987).
10. P. Besse, 'Approximation spline de l'ánalyse en composantes principales d'une variable aléatoire hilbertienne', *Annales de la Faculté des Sciences de Toulouse*, **12**, 329–346 (1991).
11. J. Ramasay and C. Dalzell, 'Some tools for functional data analysis', *J. R. Statist. Soc. B*, **53**, 539–572 (1991).
12. A. M. Aguilera, M. J. Valderrama and M. J. Del Moral, 'Un método para la aproximación de estimadores en ACP, Aplicaciónal proceso de Ornstein-Uhlenbeck', *Revista de la Sociedad Chilena de Estadística*, **9**(2), 57–77 (1992).
13. A. M. Aguilera, M. J. Del Moral and M. A. Piñar, 'On the empirical behaviour of a stochastic process', *Proc. 6th Int. Symp. on Applied Stochastic Models and Data Analysis*, Crete. World Scientific, Singapore (J. Janssen and C. H. Skiadas, eds.) Vol. I, 1993, pp. 5–16.
14. A. M. Aguilera and M. J. Valderrama, 'Principal component analysis of a stochastic process for discrete data: interpolation by B-splines', *Proc. 49th Session of International Statistical Institute*, Vol. I, 1993, pp. 11–12.
15. J. C. Deville, 'Analyse et prevision des series chronologiques multiples non stationnaires', *Statistique et Analyse des Données*, **3**, 19–29 (1978).
16. J. C. Deville, 'Estimation of the eigenvalues and of the eigenvectors of a covariance operator', *Note Interne de l'INSEE*, 1973.
17. A. M. Aguilera, 'Métodos de Aproximación de Estimadores en el ACP de un Proceso Estocástico', Tesis doctoral, Universidad de Granda, 1993.
18. A. M. Aguilera, F. A. Ocaña and M. J. Valderrama, 'A computational algorithm for PCA of random processes', *Proc. COMPSTAT. Software Descriptions*, 1994, pp. 39–40.

