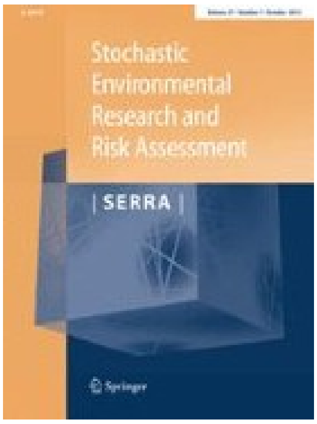
A dynamic regression model for air pollen concentration

- Francisco M. Ocaña-Peinado, Mariano J. Valderrama, Ana M. Aguilera
- A dynamic regression model for air pollen concentration
- *Stoch Environ Res Risk Assess 22, 59–63 (2008)*
- DOI: https://doi.org/10.1007/s00477-007-0153-y



Stoch Environ Res Risk Assess (2008) 22 (Suppl 1):S59–S63
DOI 10.1007/s00477-007-0153-y

ORIGINAL PAPER

A dynamic regression model for air pollen concentration

Francisco Ocana-Peinado · Mariano J. Valderrama · Ana M. Aguilera

Published online: 7 June 2007

Abstract A transfer function model with multiplicative intervention variable is proposed in this paper in order to forecast air pollen concentration using the temperature as input series. The inertia process is at the same time modelled by means of a principal component analysis (PCA) after a suitable time rescaling. The final model is tested with cypress pollen data recorded in Granada (Spain) along 11 years.

Keywords Transfer function model · ARIMA · Principal component analysis · Air pollen concentration

1 Introduction

Since Box and Jenkins introduced transfer function models (TFM), these have been broadly used in several fields, mainly in Economics, Engineering and also in the Environmental field, where this paper is focused.

Some recent contributions of the TFM in Environmetrics are the research of El-Din and Smith (2002), Vuorinen et al. (2004), Al-Awadhi (2005), Chih-Chiang et al. (2006), and Ping-Chun and Yim (2006).

It is known the existence of correlation between pollen concentration in the air and climatic variables, such as, temperature, hours of sun and humidity as example. The present research is focused on the derivation of a dynamic regression model using as climatic input the temperature. But due to the special characteristics of the pollen time series, whose occurrence season takes some months in a year, we propose to introduce an intervention variable I_t as follows:

$$y_t^* = I_t[\upsilon(B)x_t + N_t] \tag{1}$$

where $\upsilon(B)$ is the transfer function and the intervention variable takes the value zero out of the pollination interval in the year. The inertia process is estimated from the estimated transfer function as:

$$\widehat{N}_t = I_t(y_t - \widehat{\upsilon}(B)x_t)$$

Moreover, due to the fact that the estimated time series has several periods with zero values, we propose to model the inertia process by means of principal components instead of an ARIMA model as is usual by following the classic methodology, because the dependence structure is in fact broken. Therefore, Sect. 2 derives the multiplicative intervention model with a principal components inertia process, and Sect. 3 applies it to model the cypress pollen concentration in the air of Granada (Spain) with observations measured along 11 years.

2 Derivation of the theoretical model

The pollination season of a certain plant in a geographic area covers approximately the same interval $[t_0, t_1]$ all the years. So the intervention process associated to model Eq. (1) that we propose is:

$$I_t = \begin{cases} 1, & \text{if t} \in [t_0 + 365k, t_1 + 365k], \quad k = 0, 1, 2 \ldots \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

F. Ocana-Peinado (✉) · M. J. Valderrama ·
A. M. Aguilera
Department of Statistics and Operations Research,
University of Granada. Facultad de Farmacia,
Campus de Cartuja s/n, 18071 Granada, Spain
e-mail: fmocan@ugr.es

Because of all the values out of the interval $[t_0, t_1]$ in a year are zero, we can do a time rescaling so that the overall time series for the inertia process defined on the interval obtained by joining all the pollination intervals in only one, has not a real dependence structure so that the use of an ARIMA model for it does not look suitable but much more a model based on independent sample-paths such as a principal components model. We will recall the working interval as *short interval*.

In order to perform a PCA for the inertia process we consider the time series defined on the *short interval* structured as r realisations of itself in h different times that will be considered as equally spaces knots (Table 1).

Thus, we represent the time series in each one of its r sample-paths, as linear combination of the eigenvectors associated to its components:

$$\widehat{N}_t(\omega) = \sum_{i=1}^{h} u_i(t)\xi_i(\omega) \quad t = 1, 2, \ldots, h \tag{3}$$

where ξ_i are the principal components, and $(u_i(1),...,u_i(h))'$ denote the eigenvectors associated to them.

The proportion of the total variability of the inertia process explained by the i-th component ξ_i, is the quotient between its associated eigenvalue, and the total variance of the process. Besides, the proportion of variance accumulated by the k principal components with highest variance is the sum of the proportions of the total variance explained by each one of them, due to the uncorrelated character of the ξ_i.

Then, we can approximate the inertia process in an optimal way, in function of the k first principal components:

$$\widehat{N}^*_t(\omega) = \sum_{i=1}^{k} u_i(t)\widehat{\xi}_i(\omega)$$

So the final estimated model can be written as follows:

$$\widehat{y}^*_t = I_t[\widehat{v}(B)x_t + \sum_{i=1}^{k} u_i(t)\widehat{\xi}_i] \tag{4}$$

We refer to it as the TF-PC model.

3 Application to pollination data

The above described model is now applied to data recorded from 1992 to 2002 in Granada, a city located at South-East of Spain. There, the pollination season is the interval between 15 January and 15 April. So, we consider the intervention process Eq. (2) with $t_0 = 15$ and $t_1 = 104$, being [0, 4015] the complete observation interval, that includes daily observations along 11 years, so that $k = 0, 1, 2, ..., 10$. In order to homogenise all the years we have subtracted one day for the leap years so that all of them consist of 365 days. Therefore we will work with the estimated inertia process defined on the interval [0, 990] obtained by joining the eleven pollination intervals in only one. The PCA will be performed from 66 sample-paths being each of them a vector with dimension 15.

The input process is the daily average temperature (in centigrade) and the output process the daily average cypress pollen concentration in the air (grains/m^3) obtained from measurements recorded by the Centre for Aerobiology of the Department of Botanic at University of Granada.

The transfer function has been estimated by using the ITSM software developed by Brockwell and Davis (2000). The following transfer function was obtained:

$$\widehat{v}(B)x_t = 3.6014\nabla x_t + 1.8051\nabla x_{t-1} \tag{5}$$

where one regular difference is applied to y_t (cypress pollen) and x_t (average temperature) in order to get stationarity. Equation (5) shows that cypress pollen has a positive dependence on the temperature in the same day and the temperature in the previous day.

Taking into account the length of the time series, inertia process modelling by PCA is realised by dividing the residual series in 66 sample-paths with length 15 ($r = 66$ and $h = 15$ in Table 1), this is to consider 6 sample-paths in a year. Applying PCA to inertia process, the following structure (Table 2) is obtained. In this case, the three first principal components that explain the 72.6% of the variability of the inertia process, are included in the model. So,

Table 1 Estimated residual series tabulated for applying PCA

	$t=1$	$t=2$	$\ldots$	$t=h$
$\omega=1$	$\widehat{N}_{11}$	$\widehat{N}_{12}$	$\ldots$	$\widehat{N}_{1h}$
$\omega=2$	$\widehat{N}_{21}$	$\widehat{N}_{22}$	$\ldots$	$\widehat{N}_{2h}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\omega=p$	$\widehat{N}_{p1}$	$\widehat{N}_{p2}$	$\ldots$	$\widehat{N}_{ph}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\omega=r$	$\widehat{N}_{r1}$	$\widehat{N}_{r2}$	$\ldots$	$\widehat{N}_{rh}$

Table 2 PCA of the inertia process for the cypress pollen

ξ_i	Eigenvalues	% of variance	Cum. %
1	7.2067	48.045	48.045
2	2.44071	16.271	64.316
3	1.22392	8.262	72.578
4	1.0334	6.890	79.468
5	0.7939	5.293	84.761
6	0.5604	3.736	88.497
7	0.3443	2.295	90.792
8	0.3375	2.250	93.042
9	0.3136	2.091	95.134
10	0.2498	1.665	96.799
11	0.2097	1.399	98.198
12	0.0975	0.650	98.848
13	0.0745	0.497	99.345
14	0.0556	0.371	99.716
15	0.0426	0.284	100.00

the TF-PC model given in Eq. (4) has the following expression:

$$\hat{y}_t^* = I_t[3.6014\nabla x_t + 1.8051\nabla x_{t-1} + \sum_{i=1}^{3} u_i(t)\hat{\xi}_i] \qquad (6)$$

where eigenvectors, $u_i(t)$ (i = 1,2,3), are vectors with dimension 15 × 1, and are showed in Table 3. Figure 1 shows the cypress pollen smoothing with the model Eq. (6) for the first half in 2002.

The MSE is computed for 11 years as follows:

Table 3 Eigenvectors for the inertia process

$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$
0.2666	0.2138	0.1116
0.1740	–0.1202	–0.4354
0.2496	–0.2523	0.0505
0.1917	–0.3822	0.3096
0.2014	–0.3586	–0.0736
0.2248	–0.4181	0.0376
0.2389	–0.4106	0.195
0.3104	0.2642	0.0968
0.3060	0.1974	0.1384
0.3001	0.1684	0.2439
0.2804	0.1449	0.0278
0.3014	0.1890	–0.0444
0.3041	0.1911	0.0253
0.2841	0.1226	–0.3597
0.1695	–0.1122	–0.6587

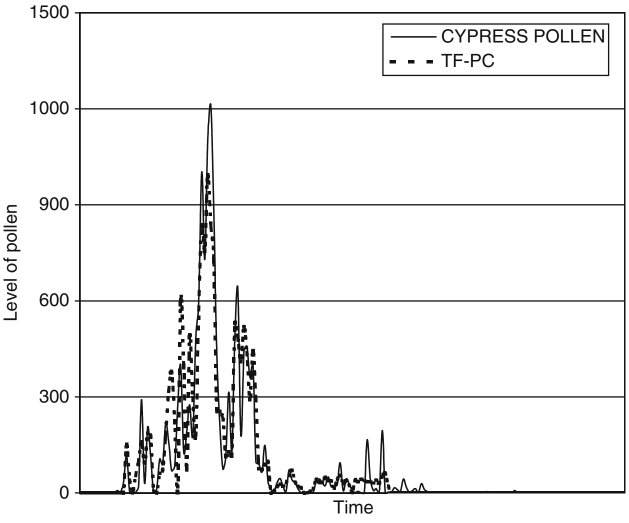


Fig. 1 TF-PC cypress pollen smoothing

$$\text{MSE} = \frac{1}{4015}\sum_{t=1}^{4015}(y_t - \hat{y}_t^*)^2 = 3870.5 \qquad (7)$$

In order to compare the performance of the model Eq. (6) with the classical Box–Jenkins methodology (1970), we will model the inertia process by a MA(3) structure, obtaining the Eq. (8), where a_t is a zero-mean white noise with variance σ_a^2 = 1248.7.

$$\hat{n}_t = a_t - 0.3759a_{t-1} - 0.2413a_{t-2} - 0.1008a_{t-3} \qquad (8)$$

The p-values associated to the parameters of this model are shown in Table 4. This MA(3) structure is added to the transfer function given in Eq. (5) obtaining the following model (TF-MA(3) model):

$$\hat{y}_t = 3.6014\nabla x_t + 1.8051\nabla x_{t-1} + \hat{n}_t \qquad (9)$$

where $\hat{n}_t$ is given from Eq. (8).

In order to forecast with the TF-PC model, we have to estimate the p.c.'s values for this period (that is, for ω = 67). These values are obtained by fitting ARIMA models to the three p.c.'s in the model from sample principal components, ω = 1,2,...,66. Table 5 shows the

Table 4 ARIMA parameters for the inertia process of the cypress pollen

Parameter	Estimate	Standard error	t	p-value
θ_1	0.3759	0.0316	11.8863	0.0000
θ_2	0.2413	0.0326	7.3869	0.0000
θ_3	0.1008	0.0316	3.1895	0.0014

Table 5 ARIMA fitted for the principal components of the inertia process

ξ_i	Model	$\widehat{\xi}_i(\omega = 67)$
1	ARIMA (0,0,1) × $(0,1,1)_6$	223.8561
2	ARIMA (0,0,1) × $(0,1,1)_6$	24.9598
3	ARIMA (0,0,1) × $(0,1,1)_6$	–44.1591

Table 6 TF-PC and TF-MA(3) forecasts for the cypress pollen

Time	Cypress Pollen	TF-PC Forec.	TF-MA(3) Forec.
15 Jan	0	3.3828	0.0000
16 Jan	1	5.0045	0.0000
17 Jan	0	0.6172	0.0000
18 Jan	0	6.3444	0.0000
19 Jan	0	7.3694	5.2644
20 Jan	0	5.7196	5.5187
21 Jan	14	18.4804	27.4718
22 Jan	5	10.5863	30.9754
23 Jan	74	67.3232	41.6654
24 Jan	13	106.3811	38.2344
25 Jan	8	17.9039	27.0491
26 Jan	13	18.7001	45.9859
27 Jan	77	124.1559	50.2603
28 Jan	81	81.9872	57.0223
29 Jan	268	196.9345	85.8732

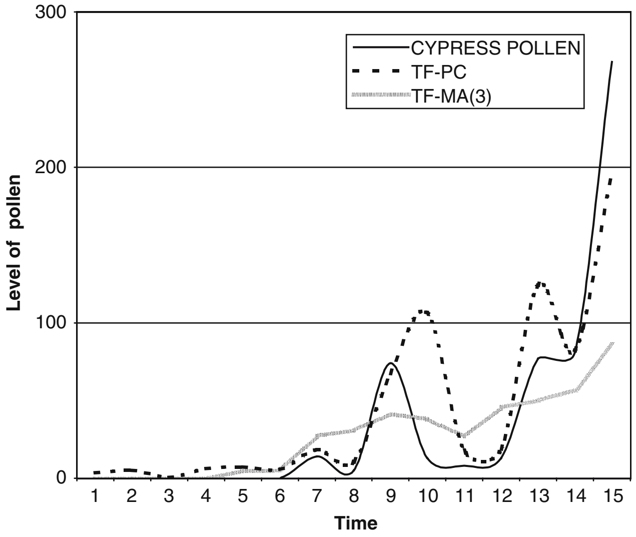


Fig. 2 TF-PC forecasts for cypress pollen

ARIMA models fitted for each principal component of the TF-PC model, and their forecasts values for $\omega = 67$.

Forecasts with both models Eqs. (6) and (9) and cypress pollen values in Granada the 15 days in the cypress pollination interval in 2003 are shown in Table 6 and represented in Fig. 2. The TF-PC model (MSE = 1091.7723) provides more accurate forecasts than the TF-MA(3) model (MSE = 2567.2484).

4 Concluding remarks

Along this paper we have designed a specific model to forecast pollen concentration in the air along the time. Because of the characteristics of this process, whose occurrence has a seasonal cycle, our model consists of an intervention on the transfer function model proposed by Box and Jenkins and afterward a PCA modelling of the inertia process resulting by time rescaling of the original time series that raises to a set of almost independent sample-paths.

The proposed methodology can be extended in two ways. The first one is the spatial-time modelling derived of taking into account pollen measurements recorded in several geographic stations. A second possible expansion would be the functional data approach by considering the continuous-time character of the process and its sample-paths. Some previous research on this way has been done by Valderrama et al. (2002), and Escabias et al. (2005).

Acknowledgments This paper was supported by Project MTM2004-05992 of the Ministerio de Ciencia y Tecnología, Spain, and the Research Group FQM-307 of the Consejería de Educación y Ciencia de la Junta de Andalucía. We acknowledge to the Aerobiology Centre of the Faculty of Sciences of the University Granada to offer us the climatic and pollen observations. We are also grateful with the anonymous referee, whose suggestions have contributed to improve the presentation of this paper.

References

Al-Awadhi SA (2005) Change in regime and transfer function models of global solar radiation in Kuwait. Environ Model Softw 20(9):1167–1174

Box GEP, Jenkins GM (1970) Time series analysis: forecasting and control. Holden Day, San Francisco

Brockwell PJ, Davis RA (2000) ITSM for windows. Springer, New York

Chih-Chiang L, Chu-Hui C, Tian-Chyi JY, Cheng-Mau W, I-Fang Y (2006) Integration of transfer function model and back propagation neural network for forecasting storm sewer flow in Taipei metropolis. Stochastic Environ Res Risk Assess 20(1–2):6–22

El-Din AG, Smith DW (2002) A combined transfer-function noise model to predict the dynamic behavior of a full-scale primary sedimentation tank. Water Res 36(15):3747–3764

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

Ping-Chun H, Yim JZ (2006) Wave height forecasting by the transfer function model. Ocean Eng 33(8–9):1230–1248

Valderrama MJ, Ocaña FA, Aguilera AM (2002) Forecasting PC-ARIMA models for functional data. In: Härdle W, Rönz B (eds) Proceedings in computational statistics, 2002. Physica-Verlag, Berlin, pp 25-36

Vuorinen I, Hanninen J, Kornilovs G (2004) Transfer-function modelling between environmental variation and mesozooplankton in the Baltic Sea. Progress Oceanogr 61(1):101–102