Int J Biometeorol (2013) 57:483–486
DOI 10.1007/s00484-012-0527-9

SHORT COMMUNICATION

# A principal component regression model to forecast airborne concentration of Cupressaceae pollen in the city of Granada (SE Spain), during 1995–2006

**Francisco M. Ocaña-Peinado · Mariano J. Valderrama · Paula R. Bouzas**

Received: 13 December 2011 / Revised: 16 January 2012 / Accepted: 21 January 2012 / Published online: 22 February 2012

**Abstract** The problem of developing a 2-week-on ahead forecast of atmospheric cypress pollen levels is tackled in this paper by developing a principal component multiple regression model involving several climatic variables. The efficacy of the proposed model is validated by means of an application to real data of Cupressaceae pollen concentration in the city of Granada (southeast of Spain). The model was applied to data from 11 consecutive years (1995–2005), with 2006 being used to validate the forecasts. Based on the work of different authors, factors as temperature, humidity, hours of sun and wind speed were incorporated in the model. This methodology explains approximately 75–80% of the variability in the airborne Cupressaceae pollen concentration.

**Keywords** Cupressaceae pollen concentration · Principal component · Karhunen-Loève expansion · Regression model

## Introduction

Forecasting airborne pollen levels is an interesting problem not only from an environmental point of view but also in health-care planning—mainly in vaccination strategies related to allergies among children and the elderly. Cupressaceae pollen has one of the highest pollen incidences in the Mediterranean area and is present in the atmosphere practically all year round, although it is predominant in the winter period, when no other plants are flowering, making this particle a powerful allergen. In Europe, allergy to Cupressaceae pollen was considered a rarity until 1975, but is now a recognised clinical entity (Belmonte et al. 1999). In order to develop a stochastic model to explain this phenomenon, several meteorological covariates must be taken into account, as was studied in Spain by several authors (e.g., Aira et al. 2001; Belmonte et al. 1999; Díaz de la Guardia et al. 2006; Galán et al. 1998; Sabariego et al. 2011; Tortajada and Mateu 2008).

The regression approach that we follow in this paper has been considered previously by several authors such as Stark et al. (1997), who applied a Poisson regression model for ragweed pollen; Brumback et al. (2000), who propose dan extension of the generalized linear models to the nonlinear framework; Smith and Emberlin (2005), who adjusted several regression models after considering pre-peak, peak and post-peak periods; Moseholm et al. (1987), Ocaña Peinado et al. (2008) and Rodríguez-Rajo et al. (2006), who use ARIMA processes; Valderrama et al. (2010), who proposed a two-step functional regression model; Makra and Matyasovszky (2011) and Makra et al. (2011), who consider nonparametric regression methods; and by Díaz de la Guardia et al. (2006) and Sabariego et al. (2011) using polynomial and multiple linear regression.

The aim of this paper was to select a set of variables suitable for modelling the stochastic process of Cupressaceae airborne pollen concentration during the pollination season. To do so, a dimensionality reduction on the basis of principal component analysis (PCA) was developed for both this process and for the above-mentioned climatic processes. The time predictive approach applied in this model means that the sample paths of the main

F. M. Ocaña-Peinado (✉) · M. J. Valderrama · P. R. Bouzas
Department of Statistics and Operations Research,
Faculty of Pharmacy, University of Granada,
18071 Granada, Spain
e-mail: fmocan@ugr.es

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

P. R. Bouzas
e-mail: paula@ugr.es

**Table 1** Mean, standard deviation and partial correlation coefficient between *P* (*t*) and the four climatic variables: *T* (*t*), *S* (*t*), *H* (*t*) and *W* (*t*) in the period 1995–2005

| | P (t) | T (t) | S(t) | H(t) | W (t) |
|---|---|---|---|---|---|
| Mean | 119.50 | 10.42 | 8.15 | 64.74 | 1.86 |
| Standard deviation | 244.75 | 3.34 | 2.69 | 8.21 | 3.25 |
| Partial correlation | ··· | 0.304 | 0.198 | −0.155 | 0.116 |

**Table 3** Real pollen *P* (*t*), and forecast pollen $\widehat{P}(t)$, for the two sample paths in 2006 and their mean square error (MSE). *LL* and *UL* are lower and upper limits in the 95% confidence intervals for $\widehat{P}(t)$, respectively

| | 15 January–29 January | | | | 30 January–13 February | | | |
|---|---|---|---|---|---|---|---|---|
| t | $P(t)$ | $\widehat{P}(t)$ | LL | UL | $P(t)$ | $\widehat{P}(t)$ | LL | UL |
| 1 | 10 | 8 | 0 | 52.89 | 293 | 143 | 102.97 | 183.03 |
| 2 | 2 | 0 | 0 | 44.68 | 4 | 22 | 0 | 62.07 |
| 3 | 1 | 0 | 0 | 44.68 | 5 | 8 | 0 | 48.91 |
| 4 | 7 | 7 | 0 | 51.89 | 51 | 86 | 45.54 | 126.69 |
| 5 | 29 | 16 | 0 | 60.50 | 58 | 75 | 34.07 | 115.83 |
| 6 | 93 | 129 | 86.11 | 173.89 | 70 | 85 | 44.33 | 125.74 |
| 7 | 271 | 165 | 122.88 | 209.95 | 24 | 13 | 0 | 53.09 |
| 8 | 59 | 16 | 0 | 60.85 | 35 | 51 | 10.81 | 91.34 |
| 9 | 37 | 82 | 39.11 | 126.91 | 8 | 14 | 0 | 54.17 |
| 10 | 21 | 28 | 0 | 72.94 | 89 | 68 | 27.66 | 108.18 |
| 11 | 23 | 18 | 0 | 62.69 | 75 | 42 | 1.04 | 82.91 |
| 12 | 91 | 56 | 13.58 | 100.82 | 5 | 9 | 0 | 49.18 |
| 13 | 117 | 73 | 30.62 | 117.37 | 35 | 57 | 16.33 | 97.28 |
| 14 | 117 | 78 | 35.13 | 122.07 | 30 | 23 | 0 | 63.05 |
| 15 | 108 | 72 | 29.16 | 116.14 | 178 | 119 | 78.68 | 159.15 |
| | MSE=1,499.2 | | | | MSE=2,004.3 | | | |

processes were recorded 1 week in advance of the others. Multiple linear regression among the principal components (PCs) was then performed to obtain the predictive model.

The behavior of our methodology was tested by its application to data recorded by the Aerobiology Center at the University of Granada (southern Spain) over a period of 12 years (1995–2006).

## Materials and methods

This study was carried out in the city of Granada (SE Spain), in the Mesomediterranean bioclimatic level. All the data used in this paper were collected using methodology analogous to that of Díaz de la Guardia et al. (2006). Data were recorded for 11 years (1995–2005), from 15 January to 15 April (90 data days per year), but taking 2 weeks as the time interval applicable. Thus, six intervals were obtained for each year, i.e., 66 in total. During these 90 days the pollen intensity is very high because species of the genus Cupressus—very prevalent in urban vegetation—produce pollen on a massive scale (Díaz de la Guardia et al. 2006). Due to the predictive aim of this model, the pollen concentration process was considered 1 week in advance of the climatic processes, i.e., $I_1=[T_{i-1}, T_i]$ and $I_2=[T_i, T_{i+1}]$, for $i=1, 2, \ldots, 65$. The stochastic processes taken into consideration were as follows:

- Cupressaceae pollen concentration: {P (t), t ∈ $I_2$} expressed as number of pollen grains per cubic meter of air (grains/m$^3$).
- Daily average temperature: {T (t), t ∈ $I_1$} expressed in degrees centigrade (°C)
- Daily average relative humidity: {H(t), t ∈ $I_1$} expressed in percent (%)
- Daily hours of sun: {S(t), t ∈ $I_1$}
- Daily maximum wind speed: {W (t), t ∈ $I_1$} expressed in kilometers per hour (km/h)

Because of the erratic nature of the pollen data measured, a logarithmic transformation was applied in order to smooth them: X(t)=log[P (t)+1]. Then, to perform the PCA, all the variables considered were standardized. In the principal component analysis (PCA), the respective PCs for each stochastic process are denoted by:

$$\left\{\xi_i^{(X)}\right\}, \left\{\xi_i^{(T)}\right\}, \left\{\xi_i^{(H)}\right\}, \left\{\xi_i^{(S)}\right\}, \left\{\xi_i^{(W)}\right\}$$

so that the Karhunen-Loève expansion for {X(t), t ∈ $I_2$} is given by:

$$X(t) = \sum_{i=1}^{n} u_i' \xi_i, t \in I_2 \qquad (1)$$

**Table 2** Principal component analysis (PCA) for *X* (*t*) and the four climatic variables. *PC* Principal components, $\lambda_i$ eigenvalues of the PCA, $CV_i$ cumulated variance for each PC

| PC | $\lambda_i^{(X)}$ | $CV_i^{(X)}$ | $\lambda_i^{(T)}$ | $CV_i^{(T)}$ | $\lambda_i^{(H)}$ | $CV_i^{(H)}$ | $\lambda_i^{(S)}$ | $CV_i^{(S)}$ | $\lambda_i^{(W)}$ | $CV_i^{(W)}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 8.62 | 57.44 | 10.38 | 69.23 | 6.95 | 46.36 | 7.41 | 49.42 | 10.66 | 71.08 |
| 2 | 1.66 | 68.53 | 1.70 | 80.56 | 1.65 | 57.33 | 1.64 | 60.34 | 1.07 | 78.20 |
| 3 | 1.11 | 75.92 | 0.80 | 85.92 | 1.37 | 66.46 | 1.04 | 67.28 | 0.93 | 84.42 |
| 4 | 0.72 | 80.74 | 0.53 | 89.47 | 1.20 | 74.48 | 1.01 | 73.98 | 0.61 | 88.46 |
| 5 | 0.57 | 84.51 | 0.33 | 91.69 | 0.77 | 79.6 | 0.74 | 78.91 | 0.47 | 91.59 |

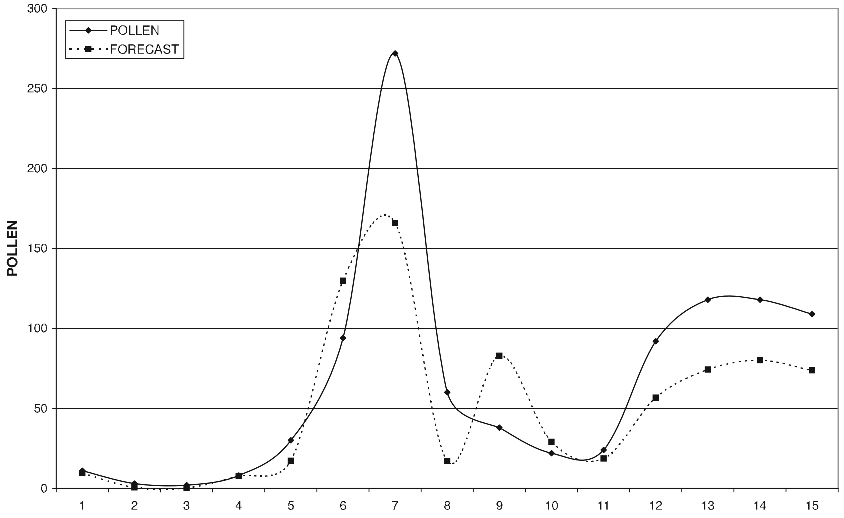

**Fig. 1** Pollen observed and forecast pollen values in the period 15 January–29 January

and the PCs are estimated by means of multiple linear regression as follows:

$$\widehat{\xi}_i^{(X)} = \gamma o + \sum_{j=1}^{n_i} a_j \xi_j^{(T)} + \sum_{j=1}^{n_2} b_j \xi_j^{(H)} + \sum_{j=1}^{n_3} c_j \xi_j^{(S)} + \sum_{j=1}^{n_4} d_j \xi_j^{(W)} \quad i = 1, 2, \ldots, n. \tag{2}$$

The criterion for selecting the number of PCs for each process to be included in the model is that they must have an explained variance greater than 1 (Kaiser 1958). The explicative variables to be introduced into the multiple regression model are then determined by the stepwise method. The goodness-of-fit for the model given by the expansion in Eq. 1 will be evaluated by the $R^2$ coefficient. The forecasts in 2006 will be tested by evaluating the mean square error (MSE):

$$MSE = \frac{1}{15} \sum_t \left[ X(t) - \widehat{X}(t) \right]^2 \tag{3}$$

A parametric 95% confidence interval for each forecast was performed in order to calibrate the degree of fit with the observed real values.

The whole statistical analysis was carried out using SPSS 15.0 software (SPSS, Chicago, IL).

## Results

A complete set of data corresponding to 12 years (1995–2006) was provided by the Airbiology Unit of the University of Granada. The peak day for P (t) usually occurred in February with quantities higher than 1,000 grains/m$^3$. This study considers the pollination interval from 15 January to 15 April during 1995–2005, while data for 2006 were used to compare forecast and real values.

Table 1 shows the mean, standard deviation and partial correlation coefficient between Cupressaceae pollen concentration and the four climatic variables during 1995–2005. All correlations are significant at the 0.01 level. The

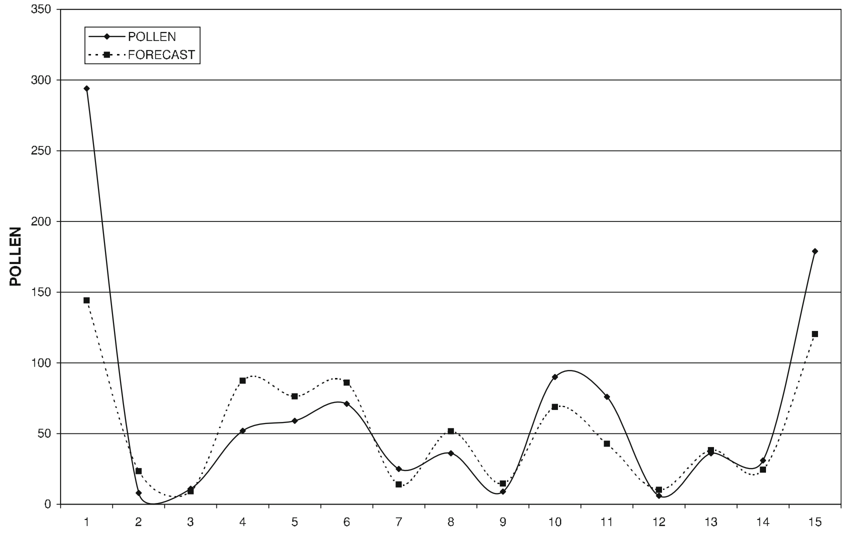

**Fig. 2** Pollen observed and forecast pollen values in the period 30 January–13 February

PCA of the transformed process $\{X(t), t \in I_2\}$ and the four climatic processes mentioned above are shown in Table 2. Note that the accumulated percentage of explained variance is approximately 75–80%. On the basis of the criterion to select the number of PCs, three are considered for X(t), and the linear regressions (Eq. 2) in terms of the remaining ones are the following:

$$\widehat{\xi}_1^{(X)} = 0.508\xi_1^{(T)} + 0.275\xi_1^{(S)} + 0.373\xi_1^{(W)} R^2 = 87.44\%$$
$$\widehat{\xi}_2^{(X)} = 0.384\xi_1^{(T)} + 0.271\xi_2^{(T)} + 0.802\xi_1^{(W)} R^2 = 83.77\%$$
$$\widehat{\xi}_3^{(X)} = 0.764\xi_2^{(T)} - 0.286\xi_1^{(H)} + 0.181\xi_1^{(S)} R^2 = 58.67\%$$

Predictions for the sample paths of $X(t)$ in 2006 were obtained by replacing the above-mentioned PCs in the expansion in Eq. 1. Associated to the general model proposed in this expansion, the $R^2$ coefficient obtained was 75.35%.

Real and forecast values for the first three sample paths in 2006 of $\{P(t), t \in I_2\}$ are shown in Table 3, together with their associated MSE. A 95% confidence interval for each forecast is also included in Table 3. Figures 1 and 2 show these forecasts for the two first sample paths.

## Discussion

Pollen is an important component in the development of allergic diseases. This research examined Cupressaceae pollen found in the atmosphere of Granada in the period 1995–2005. Due to the prevalence of this pollen in Granada during the winter, the relationship between of the most important meteorological parameters on daily pollen counts was researched in order to investigate the conditions that influence the prevalence of Cupressaceae pollination.

In agreement with the results of other studies in the Mediterranean area (Díaz de la Guardia et al. 2006; Galán et al. 1998; Sabariego et al. 2011; Tortajada and Mateu 2008), meteorological variables such as daily average temperature, daily humidity, daily hours of sun and daily maximum wind speed, were revealed as predictors to construct a PC multiple regression model.

From Tables 1–3 and Figs. 1 and 2 we can observe that the model proposed in this paper captures the trends in an optimal way, and allows the anticipation of the appearance of peaks in the Cupressaceae airborne pollen process. However, pollen levels are related not only to meteorological variables,; human activities such as pruning, watering, or introduction or elimination of plants can modify pollen values.

**Acknowledgments** This work was supported partially by projects MTM2010-20502 of Dirección General de Investigación y Gestión del Plan Nacional I+D+I and grants FQM-307 of Consejería de Innovación de la Junta de Andalucía, both in Spain. The authors are grateful to Dra. Consuelo Díaz de la Guardia (Botanic Department of the University of Granada) for providing the data to perform this research.

## References

Aira MJ, Dopazo A, Jato MV (2001) Aerobiological monitoring of Cupressaceae pollen in Santiago de Compostela (NW Iberian Peninsula) over six years. Aerobiologia 17:319–325

Belmonte J, Canela M, Guardia R et al (1999) Aerobiological dinamics of the Cupressaceae pollen in Spain, 1992–1998. Polen 10:27–38

Brumback BA, Ryan LM, Schwartz JD et al (2000) Transitional regression models, with application to environmental time series. J Am Stat Assoc 95:16–27

Díaz de la Guardia C, Alba F, De Linares C et al (2006) Aerobiological and allergenic analysis of Cupressaceae pollen in Granada. J Investig Allergol Immunol 16:24–33

Galán C, Fuillerat MJ, Comtois P et al (1998) Bioclimatic factors affecting daily Cupressaceae flowering in southwest Spain. Int J Biometeorol 41:95–100

Kaiser HF (1958) The Varimax criterion for analytic rotation in factor analysis. Psychometrika 23:187–200

Makra L, Matyasovszky I (2011) Assessment of the daily ragweed pollen concentration with previous-day meteorological variables using regression and quantile regression analysis for Szeged, Hungary. Aerobiologia 27:247–259

Makra L, Matyasovszky I, Thibaudon M et al (2011) Forecasting ragweed pollen characteristics with nonparametric regression methods over the most polluted areas in Europe. Int J Biometeorol 55:361–371

Moseholm L, Weeke ER, Petersen BN (1987) Forecast of pollen concentrations of Poaceae (Grasses) in the air by time series analysis. Pollen Spores XXIX:305322

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

Rodríguez-Rajo FJ, Fernández-Gonzlez D, Vega-Maray AM et al (2006) Biometeorological characterization of the winter in the north west Spain based on Alnus pollen flowering. Grana 45:288–296

Sabariego S, Cuesta P, Fernández-González F et al (2011) Models for forecasting airbone Cupressaceae pollen levels in Central Spain. Int J Biometeorol. doi:10.1007/s00484-011-0423-8

Smith M, Emberlin J (2005) Constructing a 7-day ahead forecast model for grass pollen at north London, United Kingdom. Clin Exp Allergy 35:1400–1406

Stark PC, Ryan LM, McDonald JL et al (1997) Using meteorological data to predict daily ragweed pollen levels. Aerobiologia 13:177–184

Tortajada B, Mateu I (2008) Cupressaceae pollen in the atmosphere of Valencia (East of Spain) and relationship with meteorological parameters. Polen 18:51–59

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