P 3.7 AIRBORNE POLLEN FORECASTING: EVALUATION OF ARIMA AND NEURAL NETWORK MODELS

Bachisio Arca*[a], Grazia Pellizzaro [a], Annalisa Canu [a], Giuseppe Vargiu [b]
[a] CNR - IBIMET, Institute of Biometeorology – Section of Monitoring Agroecosystems, Sassari, Italy
[b] Osservatorio Aerobiologico SS1, Sassari, Italy

1. INTRODUCTION

Early forecasting of pollen concentration in the atmosphere is very important for medical applications due to the increasing occurrence of allergic diseases induced by allergenic pollen. Moreover, flowering and pollen dispersion are of great interest for agronomic studies of plant productions. Several statistical techniques have been used to forecast pollen concentration in the air and concern the prediction of the start of the pollen season, the maximum airborne pollen concentration, and the date when this occurs. Some forecasting techniques are based on the analysis of airborne pollen time series, due to the the autocorrelation of daily pollen count and, in addition, of meteorological variables involved in the phenomena (Moseholm et al, 1987; Katial, 1997). A time series is a set of measurements of a variable taken over time at equally spaced time intervals. The most frequently used time series models include the autoregressive integrated moving average (ARIMA) models (Box and Jenkins, 1976); recently, artificial neural networks (ANN) have been applied in time series modeling and forecasting (Arizmendi, 1993; Patterson, 1996; Luk et al., 2000), due to their good performances with complex and non-linear phenomena (Smith, 1992). The aims of this study are (I) to develop both the ARIMA and neural network models for forecasting the daily values of airborne pollen and the day of the maximum pollen concentration, (II) to analyze and compare the performance of these models and (III) to improve the accuracy of airborne pollen forecasting for the principal allergenic plants of the Mediterranean area.

2. MATERIALS AND METHODS

The study was carried out on aerobiological and meteorological data collected from 1986 to 2001 in the urban area of Sassari (40° 44' lat. N, 8° 32' lon E, 150 m a.s.l.), Italy; the pollen sampling device was a Burkard seven-day recording volumetric spore trap. The meteorological data collected by a weather station of the Sardinian Agrometeorological Service (S.A.R) located near th spore trap were air temperature, air relative humidity, wind speed and direction, and rain intensity. The data set was divided into two sections: the first section, composed of thirteen years of data was used to calibrate (1986-1995) and to test (1996-1998) the models, the second, composed of three years (1999-2001), was used to validate the models.

The analysis was performed on Graminacee and Oleaceae, the most important allergenic families in Sardinia (Atzei and Vargiu, 1990; Atzei et al., 1993). The main pollen season was determined by calculation of the time interval between the dates when the sum of daily concentration reaches 5 % and 95 % of the total annual sum; daily pollen count was then normalized to the annual sum of pollen count and a 5 days moving average was calculated. In addition, in order to stabilize the variance, the Oleaceae daily pollen count was transformed into natural logarithmic values.

For each species, Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) were used to calculate the significant autocorrelation existing in the airborne pollen data and to identify the components of the ARIMA models. ANN models were realized using a three-layer feed-forward topology and the backpropagation learning optimization algorithm. The time series used for learning, testing and validation the ANN were transformed into a series of seven day vectors by the embedding method. The best ARIMA and ANN models were identified calculating the t statistic (t) and the correlation coefficient (r) between observed and predicted values.

3. RESULTS AND DISCUSSION

The analysis of the ACF and PACF indicated that the time series of pollen follows a seasonal pattern with time lag of 365 days; a strong correlation at a lag of one day was also discovered. Therefore, the seasonal $ARIMA(1,1,1)_{365}$ and $ARIMA(2,1,1)_{365}$ models were tested; in this identification phase the model $ARIMA(1,1,1)_{365}$ furnished the best results respect to the lower number of parameters used. The time series of pollen data was used to forecast the daily values of airborne pollen concentration for three years (1999-2001).

The ARIMA model showed a good accuracy on Graminacee, where the difference between observed and expected dates of maximum airborne pollen concentration (peak dates) range from 1 to 4 days (Table 1); the model performed less well on Oleaceae (5-7 days) in all years of the validation data set. The Table 2 shows the statistics for the values of daily pollen count predicted by the ARIMA model. The Graminacee pollen count was very highly correlated

* *Corresponding author address:* Bachisio Arca, CNR, Institute of Biometeorology, Section Monitoring Agroecosystems, Via Funtana di Lu Colbu 4A, 07100 Sassari, Italy; e-mail : arca@imaes.ss.cnr.it

with observed data while the Oleaceae pollen count showed a low accuracy in 2000 ($P \leq 0.05$) and was not significant ($r=0.17$) in 2001. This loss in accuracy could be explained by the alternate bearing of Olea europaea L. (Galan et. al., 2001a; Galan et. al., 2001b; Atzei and Vargiu, 1990), which probably affected the time series of Oleaceae pollen, determined the high inter-annual variation of mean and variance, and reduced the convergence of the ARIMA model, although we performed the logarithmic transformation of time series. Moreover, this type of ARIMA models do not use the values of predictors variables, such as the meteorological parameters (air temperature and rainfall), that affecting the pollen production, the release and the scattering in the atmosphere (Galan et. al., 2001b).

The ARIMA models carried out in this work performed well in long-term prediction (Graminacee peak dates prediction), but short-term prediction is more useful in allergy prevention.

ANN models were designed to provide a short-term forecasting (1 to 7 days) of airborne pollen, after the first appearance of the pollen flight. The best performance during the learning and testing phases was obtained by the three-layer feedforward ANN with seven neurons in the input layer, ten neurons in the hidden layer and one neuron in the output layer, which provided the values of daily pollen count for the subsequent seven days. The statistical analysis of the results provided by ANN models on Graminacee pollen data (1999-2001) is summarized in Table 3, Table 4, Table 5 and Figure 1. The best results were provided in the first four days; the r coefficient decrease from four to seven days and ANN models underestimated the observed values of about 8-14 %. The poor results provided by the ANN models for predicting the Oleaceae pollen count (data not showed) confirm the above mentioned characteristics of Oleaceae pollen and the limitations in modeling this time series.

Table 1 - Day of year (DOY) with maximum pollen concentration observed and predicted by ARIMA$(1,1,1)_{365}$ model (1999-2001).

Years	Observed (DOY)	Predicted (DOY)	Differenc e(days)
Graminacee			
1999	134	133	1
2000	133	134	1
2001	129	133	4
Oleaceae			
1999	139	134	5
2000	130	136	6
2001	126	133	7

Table 2 – Summarizing table of statistics between daily pollen count observed and predicted by ARIMA$(1,1,1)_{365}$ model (1999-2001); regressions were forced through the origin.

Years	r	b	n
Graminacee			
1999	0.58 ***	0.7	71
2000	0.88 ***	0.9	139
2001	0.84 ***	1.1	79
Oleaceae			
1999	0.81 ***	1.3	29
2000	0.40 *	1.9	23
2001	0.17	0.6	27

r, correlation coefficient for regression through the origin; b, regression coefficient for regression through the origin; n, number of observations; *, **, *** indicate significance at level $P \leq 0.05$, $P \leq 0.01$, $P \leq 0.001$ respectively

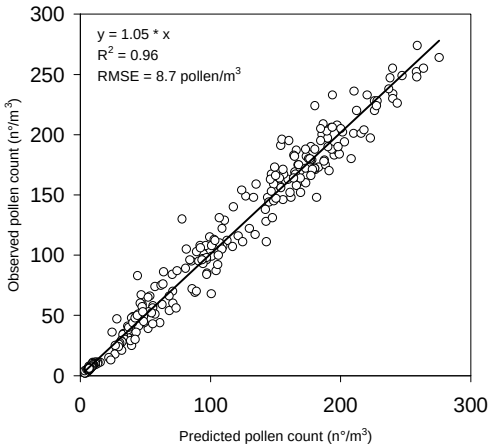

Figure 1 – Daily pollen count observed and predicted by ANN model three days ahead (Graminacee, year 2000).

5. CONCLUSIONS

Experimental results showed the capabilities of ARIMA methodology in the long-term forecasting, although the identification procedure involves many steps and often requires preliminary transformations of time series data.

The study also verifies the capabilities of ANN as a tool for short-term forecasting of some characteristics of the pollen season.

The objective of the further research is the introduction in both ARIMA and ANN models of the time series of meteorological information. As showed by other authors, ANN perform very well in complex and non linear interactions among variables and therefore ANN models can be used to support preventive allergenic therapy.

Table 3 – Summarizing table of statistics between daily pollen count observed and predicted by ANN model for the subsequent seven days (Graminacee, year 1999).

Forecast ahead (days)	r	b	n
1	0.97 ***	1.00	71
2	0.90 ***	1.01	"
3	0.82 ***	1.02	"
4	0.72 ***	1.03	"
5	0.64 ***	1.05	"
6	0.56 ***	1.07	"
7	0.51 ***	1.08	"

Table 4 – Summarizing table of statistics between daily pollen count observed and predicted by ANN model for the subsequent seven days (Graminacee, year 2000).

Forecast ahead (days)	r	b	n
1	0.99 ***	1.01	139
2	0.98 ***	1.02	"
3	0.96 ***	1.05	"
4	0.94 ***	1.07	"
5	0.92 ***	1.09	"
6	0.90 ***	1.11	"
7	0.88 ***	1.12	"

Table 5 – Summarizing table of statistics between daily pollen count observed and predicted by ANN model for the subsequent seven days (Graminacee, year 2001).

Forecast ahead (days)	r	b	n
1	0.98 ***	1.01	79
2	0.95 ***	1.03	"
3	0.89 ***	1.06	"
4	0.82 ***	1.08	"
5	0.74 ***	1.09	"
6	0.67 ***	1.12	"
7	0.62 ***	1.14	"

6. REFERENCES

Arizmendi, C. M., Sanchez J. R., Ramos N. E., Ramos G.J., 1993: Time series predictions with neural nets: application to airborne pollen forecasting, *Int. J. Biometeorol.*, **37**:139-144.

Atzei A.D., Chessa A.M., Del Giacco S., Locci F., Mulas M., Nieddu G., Tomasetti G., Vargiu A., Vargiu G., Zedda M.T., 1993: Distribuzione del genere olivo in Sardegna e suo impatto allergologico sulla popolazione. *G. Ital. Allergol. Immunol. Clin.*, **3**, 187-202.

Atzei, A. D., Vargiu G., 1990: Piante e allergie da polline, TAS, Sassari, Italy.

Box G., Jenkins G., 1976: Time series analysis forecasting and control. Holden-Day, San Francisco.

Galán C., Cariñanos P., García-Mozo H.,. Alcázar P., Domínguez-Vilches E., 2001b: Model for forecasting *Olea europea* L. airborne pollen in south-west Andalusia, Spain, *Int. J. Biometeorol.*, **45**, 59-63.

Galán C., García-Mozo H., Cariñanos P., Alcázar P., Domínguez-Vilches E., 2001a: The role of temperature in the onset of the *Olea europea* L. pollen season in southeastern Spain. *Int. J. Biometeorol.*, **45**, 8-12.

Katyal, R. K., Zhang Y., Jones R. H., Dyer P. D., 1997: Atmospheric mold spore counts in relation to meteorological parameters, *Int. J. Biometeorol.*, **41**:17-22.

Luk, K. C., Ball J. E., Sharma A., 2000: A study of optimal model lag and spatial inputs to artificial neural network for rainfall forecasting, *Journal of Hidrology*, **227**, 56-65.

Moseholm, L., Weeke E. R., Petersen B. N., 1987: forecast of pollen concentrations of poaceae (grasses) in the air by time series analysis, *Pollen et Spores*, Vol. XXIX, n° 2-3, 305-322.

Patterson, D.W., 1996: Artificial Neural Networks: theory and applications, Simon and Schuster, Singapore, pp. 477.