P 1.13 USE OF NEURAL NETWORKS TO SHORT-TERM FORECAST OF AIRBORNE POLLEN DATA

Bachisio Arca*[a], Grazia Pellizzaro [a], Annalisa Canu [a], Arnoldo Vargiu [b]
[a] CNR - IBIMET, Institute of Biometeorology, Laboratory for Monitoring of Agroecosystems, Sassari, Italy
[b] Aerobiological Observatory SS1, Sassari, Italy

1. INTRODUCTION

In the last two decades the allergic diseases inducted by allergenic pollen have dramatically increased, as well as the severity of allergic symptoms; consequently, the social cost of pollen related to diseases is very high. One of the main characteristics of the pollen allergies is its seasonal nature, due to the pollination period that characterizes each plant. This period is about the same every year, and this information may be used to support preventive allergic therapy. Several papers report that very low pollen concentrations are sufficient to produce allergic symptomatology: for example, severe symptoms of allergy were inducted in 90% of patients by pollen concentration of *Betula* greater than 30 grains/m^3 (Corsico, 1993). Therefore, a better application of preventive allergic therapy is highly dependent on improvements of the methods for early forecasting daily airborne pollen concentration. Several forecasting techniques and methods can be used for this purpose. Analytical models (Wolf et al., 1998; Moseholm et al., 1987), that are based on differential equations, describe emission and dispersion of pollen in the atmosphere combining many parameters related to plant observations and weather conditions. Often lack of parameter values makes the application of models difficult. Statistical models are based on the analysis of historical data of airborne pollen concentrations and meteorological data (Laaidi et al., 2003; Rodríguez-Rajo et al., 2003; Galán et al., 2001a; Galán et al., 2001b; Laaidi, 2001; Peeters, 1998; Nieddu et al., 1997). Time series analysis, the most common statistical prediction tool, do not assume knowledge of the structural relationships between variables involved in the process, but is simply based on the analysis of past observations of a variable to develop a model for future trends. The Box-Jenkins methodology (Box and Jenkins, 1976), that has dominated the area of time series forecasting since 1970, especially with the autoregressive integrated moving average (ARIMA) models, forecasts a variable by a linear combination of the previous state of the variable and the previous forecast errors. The Box-Jenkins methodology has been used also in pollen time series analysis and prediction (Belmonte, 2002; Katial, 1997; Stephen et al., 1990). The major limitation of ARIMA methodologies is the pre assumed linear form of the model (Zhang, 2003), that are not able to capture non linear patterns that affect many environmental phenomena. Artificial neural networks (ANNs) have been applied in time series analysis as tool for modeling the complex and non-linear phenomena; ANNs are able to learn from examples and to capture the functional relationships among time series values, even if the underlying relationship are unknown or hard to describe by the analytical approach (Zhang, 1998; Patterson, 1996). ANNs have been used to forecast pollen concentrations (Ranzi et al., 2003; Arizmendi et al., 1993; Bianchi et al., 1992) with different training algorithm (standard backpropagation, time delay backpropagation) and time step (daily and hourly time step). The aims of this study are (I) to develop a neural network model to short-term forecast airborne pollen concentration and (II) to analyze and compare the effect of the different model parameters on the forecasted values and (III) to improve the accuracy of airborne pollen forecasting for Graminaceae, on of the main allergenic plants of the Mediterranean area.

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

2. MATERIALS AND METHODS

In this study aerobiological and meteorological data collected from 1987 to 2001 in the urban area of Sassari (40° 44' lat. N, 8° 32' lon E, 150 m a.s.l.), Italy were used. 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 the spore trap, were: air temperature (T), air relative humidity (RH), wind speed (U), and rain intensity (P). The data set was divided into two sections: the first section (twelve years) was used for training (1987-1995) and for validating (1996-1998) the ANN models, the second one (1999-2001), was used for testing the ANN models.

The analysis was performed on Graminacee, one of the most important allergenic family in the studied area (Atzei et al., 1993; Atzei and Vargiu, 1990). The main pollen season was determined as the time interval between the dates when the sum of daily concentration reaches 2 % and 98 % of the total annual sum. Daily pollen count was then normalized to the annual sum of pollen count and a 5 day moving average was calculated. Short gaps in the pollen data were filled by linear interpolation to provide complete data sets for time series analysis.

ANN models were realized using a three-layer feed-forward topology and the backpropagation learning optimization algorithm; the processing element (PEs) of each layer are full-connected to the

Table 1 - Artificial neural networks parameters.

ANN	1st	2nd	3rd
Input variables	C	C, RH	C
Input lags	15	15	15
Forecasting technique	one-lag	one-lag	multi-lag
n° of hidden units	5	5	5
n° of epochs	41000	15000	45000
Learning rate	0.1	0.1	0.1

C, pollen concentration; RH, relative humidity

PEs of the next layer. Several ANNs were developed in order to test the effect of different network geometry (number of time lags for each input variables, number of processing elements on hidden layer, etc.) and different combination of input parameters (pollen concentration, meteorological variables) on their performance. The ANN models were designed to forecast the pollen concentration by two different techniques: one-lag and multi-lag. In the one-lag technique only past values of input variables were used to perform forecast. In the multi-lag technique the ANN forecasted output is fed back as input for the next prediction step (time series projection) to forecast the pollen concentration several steps into the future. In Table 1 the structural and functional characteristic of the ANNs analyzed in this work are reported.

The following statistical parameters were used to evaluate the accuracy of ANN forecast: correlation coefficient for regression through the origin (r), regression coefficient for regression through the origin (b), root mean square error (RMSE) and RMSE normalized by the mean (NRMSE).

3. RESULTS AND DISCUSSION

Regarding the geometry of input layer this study showed low values of NRMSE up to 25 PEs in the input layer; the lowest value of NRMSE was obtained using 15 PEs (Figure 1). Since in time series

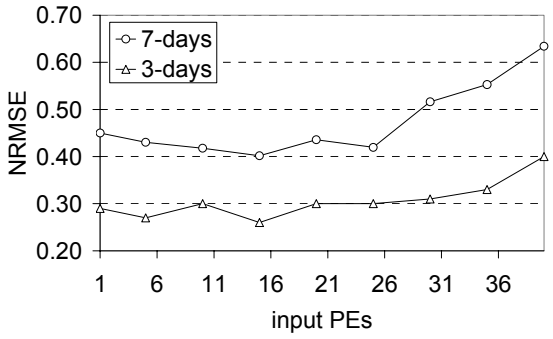

Figure 1 - NRMSE provided by the 1st ANN model with an increasing number of processing elements in the input layer (input PEs); Graminacee, years 1999-2001.

Table 2 - Statistics between daily pollen count observed and predicted by 1st ANN (Graminacee, 1999-2001).

Forecast ahead (days)	r	b	NRMSE	MAE
1	0.99 ***	0.99	0.09	0.74
2	0.97 ***	0.95	0.17	1.35
3	0.94 ***	0.94	0.24	1.93
4	0.89 ***	0.89	0.33	2.66
5	0.85 ***	0.85	0.40	3.33
6	0.82 ***	0.88	0.41	3.44
7	0.80 ***	0.86	0.44	3.72

r, correlation coefficient for regression through the origin; b, regression coefficient for regression through the origin; NRMSE, normalized root mean square error; MAE, mean absolute error; *, **, *** indicate significance at level $P \leq 0.05$, $P \leq 0.01$, $P \leq 0.001$ respectively; number of observations = 283

Table 3 - Statistics between daily pollen count observed and predicted by 2nd ANN model (Graminacee, 1999-2001).

Forecast ahead (days)	r	b	NRMSE	MAE
1	0.99 ***	0.99	0.09	0.68
2	0.97 ***	0.98	0.16	1.23
3	0.94 ***	0.94	0.25	2.06
4	0.91 ***	0.96	0.31	2.44
5	0.84 ***	0.98	0.36	2.98
6	0.84 ***	0.93	0.37	3.13
7	0.81 ***	0.90	0.41	3.42

number of observations = 283.

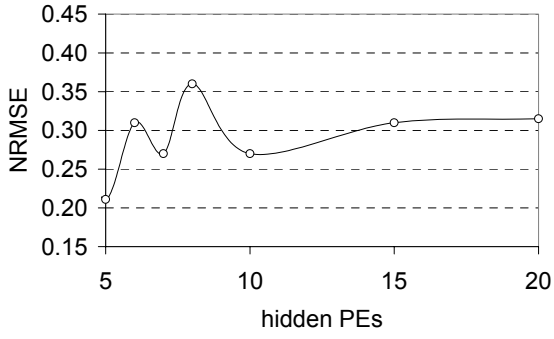

Figure 2 - NRMSE provided by the 1st ANN model with an increasing number of processing elements in the hidden layer (hidden PEs); Graminacee, years 1999-2001.

Table 4 - Statistics between daily pollen count observed and predicted by 3rd ANN model (Graminacee, 1999-2001).

Forecast ahead (days)	r	b	NRMSE	MAE
1	0.99 ***	0.99	0.09	0.72
2	0.96 ***	0.96	0.18	1.51
3	0.92 ***	0.93	0.28	2.32
4	0.86 ***	0.89	0.37	3.10
5	0.81 ***	0.86	0.45	3.76
6	0.78 ***	0.84	0.49	4.13
7	0.77 ***	0.83	0.52	4.39

number of observations = 283.

forecasting the number of input nodes corresponds to the number of lagged observations used to discover the underlying pattern, a time series of 15 daily values of pollen concentration was used in the ANN models. These results were observed for both short (3 days) and medium (7 days) time steps (Figure 1).

Relative to PEs number in hidden layer, experimental results (Figure 2) did not show a clear trend relating the generalization ability of ANN to the number of hidden PEs. The poor generalization could be due to the overtraining as reported by several theoretical and empirical works (Maier and Dandy, 1998; Patterson, 1996). In this study the best results were obtained using 5 PEs; ANN performances were not significantly affected by using two hidden layer (data not shown).

In Table 2 results provided by ANNs using only values of daily pollen count (1st ANN) are reported; NRMSE values ranged from 0.09 to 0.38 and values of the regression coefficient showed an overestimation of pollen concentration ranging from 2 to 12%.

In Table 3 results provided by ANN using both daily pollen count and minimum air humidity (2nd ANN) are shown; NRMSE ranged from 0.09 to 0.36, with an overestimation ranging from 1 to 4%. Combination of pollen concentration and meteorological data did not

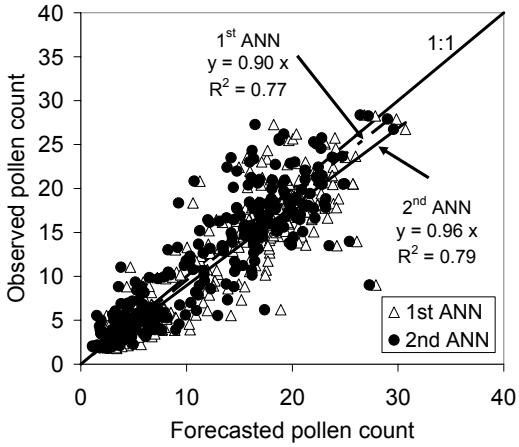

Figure 3 – Relationship between daily pollen count observed and forecasted 4-days ahead by the 1st and the 2nd ANN; Graminacee, years 1999-2001.

significantly affect ANN performances, except for minimum air humidity (RHmin) that was able to improve the performances of ANN models. The 2nd ANN furnished a better forecast accuracy for a 4-7 day forecast horizon compared to the 1st one (Table 2). Figure 3 show the lower scatter forecast and the lower overestimation of measured values provided by 2nd ANN in compared to the 1st one. A close agreement between daily pollen concentration observed and predicted by the 2nd ANN are also showed in Figure 4 and Figure 5. Our results were in agreement with Cotos-Yáñez et al. (2004) who suggested the combined use of meteorological parameters and pollen concentration, whereas the use of only meteorological parameters as predictor variables could be inadequate.

The multi-lag technique (3rd ANN) was used to test the effect of the recurrent use of predicted values, appended to the ANN inputs and used to predict future values. The NRMSE values ranged from 0.09 to 0.52 and values of the regression coefficient showed

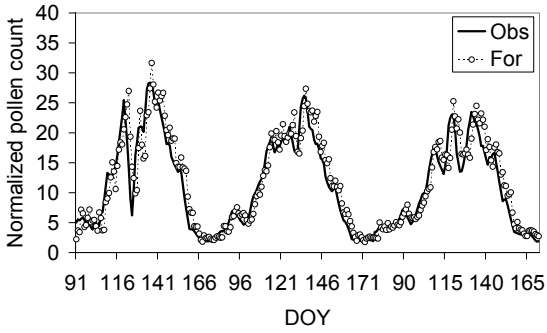

Figure 4 – Daily variation of pollen concentrations observed and predicted 3-days ahead by the 2nd ANN; Graminacee, years 1999-2001.

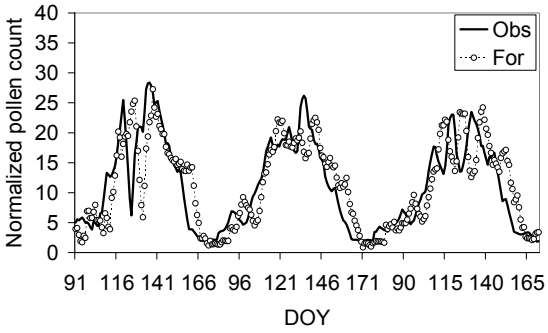

Figure 5 – Daily variation of pollen concentrations observed and predicted 7-days ahead by the 2nd ANN; Graminacee, years 1999-2001.

an overestimation of pollen concentration ranging from 1 to 17% (Table 4). This approach leads to worse prediction than the one-lag technique, according to other works (Zhang et al., 1998).

5. CONCLUSIONS

In this work the performance of ANN models for short-term forecasting of Graminacee airborne pollen concentration in Mediterranean area was evaluated. The results obtained showed a good agreement between daily pollen concentrations measured and estimated by ANNs. On the whole testing dataset (1999-2001), ANN models gave MAE (not normalized values) ranging from 1.93 to 8.53 pollen/m^3 for three days forecast and from 6.63 to 15.97 pollen/m^3 for seven days forecast. The one-lag forecasting technique furnished the best results, although required more time to perform the learning phase. Moreover, the introduction of relative humidity parameter, in combination with pollen concentration, allowed to improve forecast accuracy. The performance of ANNs can be attributed to their structural and functional characteristic, such as the nonlinear model capability and the universal function approximation. These characteristic suggest the use of ANNs as tool to forecast pollen concentration and to support the preventive allergic therapy, even if, as other statistical models, ANNs require local calibration to produce reliable results.

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

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. Allerg. Immun. Clin.*, **3**, 187-202.

Belmonte J., Canela M., 2002: Modelling aerobiological time series. Application to Urticaceae. *Aerobiologia*, **18**, 287-295.

Bianchi M.M., Arizmendi C.M., Sanchez J.R., 1992: Detection of chaos: New approach to atmospheric pollen series analysis. *Int. J. Biometeorol.*, **36**, 172-175.

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

Corsico R., 1993: L'asthme allergique en Europe In F.T.M. Spieksma, N. Nolard, G. Frenquelli and D. Van Moerbeke (eds): *Pollens de l'air en europe*. UCB, Braine-l'Alleud, 19-29.

Cotos-Yáñez T.R., Rodríguez-Rajo F.J., Jato M.V., 2004: Short-term prediction of Betula airborne pollen concentration in Vigo (NW Spain) using logistic additive models and partially linear models. *Int. J. Biometeorol.*, **48**, 179-185.

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.

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.

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.

Laaidi M., 2001: Forecasting the start of the pollen season of Poaceae: evaluation of some methods based on meteorological factors. *Int. J. Biometeorol.*, **45**, 1-7.

Laaidi M., Thibaudon M., Besancenot J.P., 2003: Two statistical approaches to forecasting the start and duration of the pollen season of Ambrosia in the area of Lyon (France). *Int. J. Biometeorol.*, **48**, 65-73.

Maier H.R., Dandy G.C., 1998: The effect of internal parameters and geometry on the performance of back-propagation neural networks: an empirical study. *Environmental Modelling*, **13**, 193-209.

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.

Nieddu G., Chessa I., Canu A., Pellizzaro G., Sirca C., Vargiu G., 1997: Pollen emission from olive trees and concentrations of airborne pollen in an urban area of North Sardinia. *Aerobiologia*, **13**, 235-242.

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

Peeters A.G., 1998: Cumulative temperatures for prediction of the beginning of ash (Fraxinus excelsior L.) pollen season. *Aerobiologia*, **14**, 375-381.

Ranzi A., Lauriola P., Marletto V. and Zinoni F., 2003: Forecasting airborne pollen concentrations: Development of local models. *Aerobiologia*, **19**, 39-45.

Rodríguez-Rajo F.J., Frenguelli G., Jato M.V., 2003: Effect of air temperature on forecasting the start of the Betula pollen season at two contrasting sites in the south of Europe (1995-2001). *Int J. Biometeorology*, **47**, 117-125.

Stephen E., Raftery A.E., Dowding P., 1990: Forecasting spore concentrations: A time series approach. *Int. J. Biometeorol.*, **34**, 87-89.

Wolf F., Puls K.E., Bergmann K.C., 1998: A mathematical model for mugwort (Artemisia vulgaris L.) pollen forecasts. *Aerobiologia*, **14**, 359-373.

Zhang G., Patuwo E., Hu M.Y., 1998: Forecasting with artificial neural networks: The state of the art. *Int. J. of Forecasting*, **14**, 35-62.

Zhang G.P., 2003: Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, **50**, 159-175.