

Forecasting airborne pollen concentrations: Development of local models

Andrea Ranzi^{1,*}, Paolo Lauriola¹, Vittorio Marletto² & Franco Zinoni²

¹*Environmental Epidemiology Unit, Scientific Management – Regional Agency for Prevention and Environment of Emilia Romagna (ARPA-ER);* ²*Regional Meteorological Service – Regional Agency for Prevention and Environment of Emilia Romagna (ARPA-ER)*

(*Author for correspondence: E-mail: aranzi@mo.arpa.emr.it; Fax: +39-59-310193)

Received and accepted 17 June 2002

Key words: empirical models, forecast, grass pollen, neural networks, statistics

Abstract

People's sensitivity to allergies may represent one of the most important health factors of the next century to which attention must be paid in order to reduce the incidence of social costs and improve the quality of life. Taking into consideration the earnest requests of the medical-scientific community Emilia-Romagna ARPA (Regional Agency for the Prevention of the Environment) moved the attention from the monitoring to a short and medium term prediction of the concentration of allergenic pollens in the air in order to achieve a more effective therapeutic action. Our main objectives are to improve seasonal forecasts and to interpret anomalous years. A neural network model for grass pollen forecasting has been implemented. Input variables were meteorological situations, i.e., daily temperature (max., min. and average) and rainfall, in addition to combinations of individual variables and their thresholds. The output was daily pollen concentration. The model was able to understand and predict anomalous years. We demonstrate that the relationships between pollen concentrations and meteorological situations are independent from site. This means that such models can understand the differences in different areas.

1. Introduction

1.1 Dimension of the problem of allergenic illness

Many human diseases are caused by aerosols carried in the air we breath. Inhalation allergy, caused by airborne pollen, is very well known and produces many disorders of varying severity. Allergy due to pollen is growing so fast among European populations that it could be rated as one of the most widespread diseases of the next century. Also the severity of the allergic symptoms is rising, ranging from bearable, bothering, chronic forms, to increasingly acute forms.

Recently, a report presented to the European Parliament announced that allergy, reported only for 1% of the population in Europe at the beginning of this century, in one form or another now strikes about 20% of people and the trend is so strong that allergy will involve 35% of the population within next twenty

years with an assessed cost of about 30 billion Euro (UCB, 1999).

Rising sensitivity of the population to pollen and to other airborne allergenics seems correlated to the air quality, mainly influenced by traffic pollution. The forecast seems even worse when considering the ever increasing mobility for work and leisure.

The diffusion of pollen data and pollen forecasts should improve the prevention of allergic episodes, and the reduction of human and economic costs due to acute allergic symptoms.

Pollen allergy episodes risk forecasting should be more effective in attaining these aims.

Contacts between environmental officials, health officials and patients should raise the attention of the public towards the need for improving air quality, and applying pressure on local authorities for the full implementation of the relevant European policies.

Emilia-Romagna ARPA (Regional Agency for the Prevention of the Environment) has used its ten years experience and the collaboration with the Italian Society of Aerobiology to set up a monitoring network that nowadays is run by the provincial sections of ARPA.

The activities carried out to produce this monitoring service include: air-dispersed pollen monitoring, data management, development of a software for a weekly bulletin with last week observed data and qualitative forecasts. The data, relating to the pollen concentrations used in the analyses, come from the regional monitoring network of the provincial units of ARPA. The meteorological data used are collected from the automatic monitoring stations of the regional network of the Meteorological Service and the provincial agro-meteorological networks.

In this article we will present the development of a local forecasting model to predict grass pollen concentrations, to be used in the weekly pollen bulletin.

1.2 Pollen forecasting models

Short-term and long-term pollen forecasts are arguments of particular relief in applied aerobiological research. Now this field can count on a large amount of literature, in which authors present various different forecasting methods.

The first and more elementary level of forecast is based on the location of a medium trend of the flowering or pollen concentrations, calculated on a multi-year series of observations explained graphically by means of phenological or pollen calendars (calendar forecasts). However, in order to obtain a good level of accuracy in the forecasts, the possibility of modulating them in relation to the phenological and meteorological changes must be introduced through the application of forecasting models.

The forecasting models available in literature are various, and they can be classified in different categories, based on the type of output provided (point, spatial and temporal forecasts) (Frenguelli and Bricchi, 1998; Norris-Hill, 1998). Recently, in spite of a greater complexity, models of pollen counts have had a great development (Moseholm et al., 1987; Stark et al., 1997; Puppi Branzi and Zanotti, 1992).

We can distinguish such forecasting models in: empirical models and simulation models. The main goal of empirical models is to provide reliable estimations; they are mainly based on statistical analysis. In order to realize and to calibrate these models, long

historical series of pollen and meteorological data are necessary.

Estimations are generally satisfactory, if the used historical series are consistent and the environmental conditions of the year don't differ greatly from those used for the calibration of the model.

The simulation models instead are so structured as to reconstruct the history of pollen grains from their production and emission into the atmosphere, to their dispersion and transport until the sampler: in this case the information in input are wider and regard both the phenology and distribution of the plants source and the atmospheric dynamics. The latter predictive models are surely more powerful than those of the empirical models, even though it should be emphasised that their high complexity and the difficulty in finding data could be an obstacle in their use.

In the present paper an empirical model based on neural network algorithms is presented.

1.3 Neural networks in time-series analysis

Autocorrelation analysis has been often employed in time series forecasting using statistical approaches such as ARMA models. This analysis is mainly used in detecting the autocorrelations between successive observations of time series, and used in the well-known ARIMA models with Box-Jenkins methods that are very efficient in forecasting linear time series.

Application of neural networks in time series forecasting (Chakraborty et al., 1992; Windsor et al., 1990) is based on the ability of neural networks to approximate nonlinear functions. The most popular treatment of input data is feeding the neural networks with either the data at each observation, or the data from several successive observations. Denote the data at instant k as $y(k)$, where y may be a vector, then the above treatment can be described as

$$y_{k+1} = NN(y_k), \text{ or } NN(y_k, y_{k-1}, \dots, y_{k-l})$$

respectively, where $NN()$ stands for the neural network forecaster and l is the number of successive observations. This treatment considers the time series as a nonlinear time series and tends to generate a nonlinear "auto-regression" model to fit the series.

Neural networks structure simulates the transmission of information in human brain (Figure 1). The cells or units receive a pattern of information in input, elaborate it through an activation function and then produce an output value. The connections link the

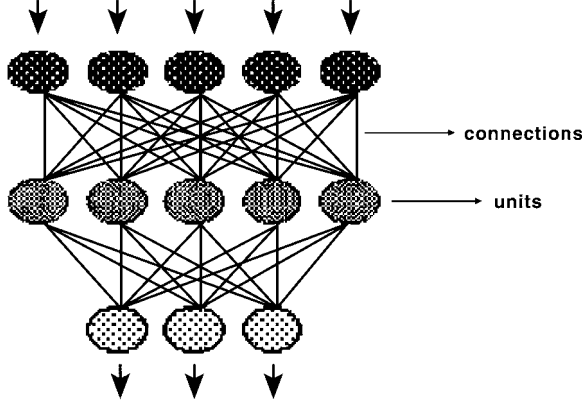


Figure 1. A schematic diagram of a feed-forward neural network with one hidden layer.

various cells and have a variable value, called synaptic weight.

The topology of these connections permit to have a high variety of networks. We used the most popular, the feed-forward MLP (multi-layer Perception), in which we have:

- an input layer, with a number of cells equal to the number of independent variables. They simply receive the input value and transmit it to the cells of the next layer (multiplied by the weight of each connection).
- a hidden layer (or more): at each hidden node (its number is variable and is a parameter of the network), a weighted linear combination of the inputs is summed and then a non-linear transformation is applied.
- an output layer, with a number of cells equal to the dependent (outcome) variable.

The training process of the neural network involves adjusting the weights till a desired input/output relationship is obtained. The majority of adaptation learning algorithms are based on the Widrow-Hoff back-propagation algorithm (Rumelhart et al., 1986).

The input of a particular element is calculated as the sum of the input values multiplied by the connection strength (synaptic weights).

$$H_j = \sum_{i=1}^m w_{ji} I_i$$

where H_j is a hidden layer cell, I_i is the i th input cell, m is the number of input cells, w_{ji} is the synaptic strength between i th input cells and j th hidden layer cell.

The value of each cell is passed through a non-linear sigmoidal function $g(x)$: this function is used

because it permits the network to perform a high-dimensional, non-linear approximation to the data. Indeed, this allows the construction of quite arbitrary, non-predetermined functions by the network.

$$g(x) = \frac{1}{1 + e^{-x}}.$$

The output of a hidden layer was the input to the cells located in the final output layer of the network. Since the cell in the final output layer received inputs from all the previous layer cells, the output of whole network at a given p pattern O^p can be described as follows:

$$O_k^p = g \left(\sum_{j=1}^n w'_{kj} g \left(\sum_{i=1}^m w_{ji} I_i^p + \Theta_j \right) \right)$$

where m is the number of input cells, n is the number of hidden cells, w_{kj} are the weights between the hidden and the output layer, w_{ji} are the weights between the input and hidden layer cells; Θ_j is the threshold of the j th hidden layer cell; k is equal to 1 in case of a single output.

The error of the network for identifying a given pattern p is calculated as:

$$E = \sqrt{(O^p - d^p)^2}$$

where d^p is the desired output, O^p is the actual output of the network.

The weights were corrected according to the delta rule:

$$w_{ji}^{new} = w_{ji}^{old} + \gamma \frac{\partial E}{\partial w_{ji}}$$

where γ is a learning coefficient, $\partial E / \partial w_{ji}$ is the derivative of the error function with respect to the given weight w_{ji} .

2. Materials and methods

2.1 Data

Phenological and meteorological data, starting from 1988, have been used. Input variables were meteorological situations, i.e., daily temperature (min, max, mean) and rainfall, in addition to combinations of these variables and their thresholds. Thresholds were chosen with statistical techniques i.e., correlation with pollen concentrations, or based on knowledge. We used also auto-regressive inputs, i.e., past pollen

Table 1. Variables used in the neural network model

Inputs	Output
Meteorological data	Pollen concentration
Day of the year	
Temperature Sum (day degree) with thresholds	
Accumulated rainfall with thresholds	
N° of days with t° or rainfall > threshold	
N° of hours with t° > threshold	
Water deficit > threshold	
Pollen concentrations of past days	

concentrations. We calculate the moving average on 7 days for grass pollen data concentration and we selected phenological and meteorological data from 1/1 to 31/5 for each year. We used past grass pollen concentrations starting from three days before, with a temporal window of one week as auto-regressive component. Output of the model was the actual pollen concentration (Table 1).

The entire data set is divided into: a training, a validation and a test set. The training and the validation set are used to train the network: The validation set allows to avoid overfitting, a typical problem of such algorithm. The test set is used to test the network on unknown patterns.

2.2 Forecasting models

We used a feed-forward neural network model. Its topology consists of 25 neurons for the input layer, 8 for the hidden layer and 1 neuron for the output layer, i.e., the pollen concentration.

Determination of the number of neurons in the hidden layer was done automatically based on the rule that the number of weights should be less than a tenth of the number of training samples (Weigend et al., 1991).

Temporal window for auto-regressive component was determined by practical considerations, based on availability of past data at the moment of bulletin publication.

We apply the time-delay back-propagation training algorithm (TDNN), in which the time-delayed component were the past pollen concentration.

3. Results

The training algorithm network stopped after 850 epochs. Network error reached the value of 3×10^{-2} (Figure 2). The trained algorithm was tested on 1 year (test set) of the same site. Then we simulate forecasts for unknown years both in the same site and on in another site. Figures 3 and 4 show the satisfactory performances in both cases.

In order to test if the model was suitable for our main scope, i.e., the prediction to exceed thresholds, 4 thresholds exceeds have been tested: 10, 20, 50 and 100 pollen, and the error in delay or anticipation days have been calculated. Figure 5 shows that the mean errors never exceed 2 days, with an acceptable value of 4.

4. Discussion

Short-term and long-term pollen forecast is an argument of particular relief both in the field of aerobiological research and in the public health issues. In fact, forecasts of pollen concentrations and then of allergy episodes should help to improve the management of hospitals and other public health resources giving an expected reduction in budget.

Taking into consideration these facts, simple forecasting models based on the accumulation of heat (total day degrees) were developed. These models did not always give an improvement in the forecast when compared with information collected from previous years.

Thus two statistical approaches have been used for our models including logistic regression and neural networks. Input variables were meteorological situations, i.e., daily temperature (max., min. and average) and rainfall, in addition to combinations of individual variables and their thresholds. The output of the first model was the day on which a pre-determined pollen threshold (50 pollen/mc) was reached. The output of the second model was daily pollen concentration.

The neural network model presents an auto-regressive component and an intrinsic nonlinearity that permits a better performance. However both models were able to understand and predict anomalous years and they demonstrate that the relationships between pollen concentrations and meteorological situations are independent from site. This means that these models can understand the differences in space. As regards differences in time, the neural network model

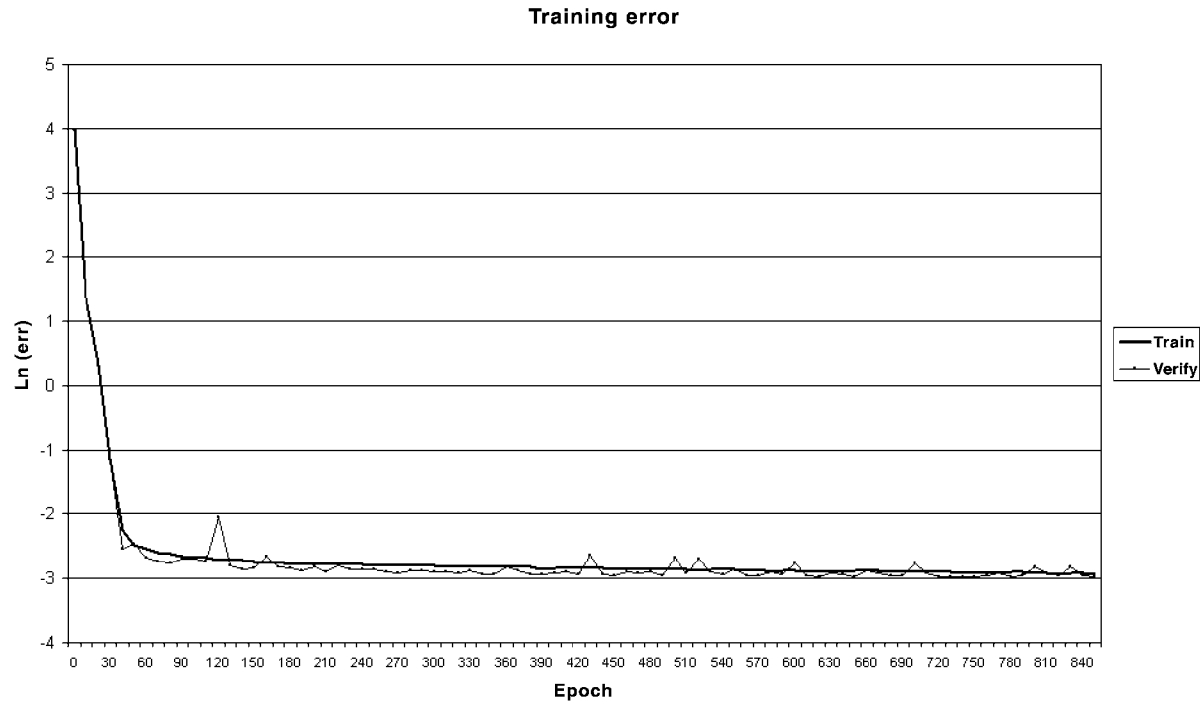


Figure 2. Performance of the network during training phase.

showed its intrinsic robustness in forecasting unknown years. Neural networks can be considered as non-linear semi-parametric regression techniques. They are semi-parametric: this means that the parameters to be set up refer to the topology of the network and to its learning procedure only; therefore they are independent from the characteristics of the system studied. The non-linearity is due to the activation function of each cell of the network, usually a sigmoidal function. The topology of such structure distributes the information to the various cells, amplifying or reducing each contribution by means of weights. This distribution of information allows these algorithms a better tolerance to error or lack of input data than traditional regression techniques.

These characteristics make neural networks particularly suitable in situations in which there are high nonlinear relations between variables and when the complete dynamics of the problem studied is not so clear. However this “blackbox-like” behavior could be a big problem in the interpretation of the results produced by the network. Another possible disadvantage is that the performance could not be reliable in situations very different from those contained in the learning data set.

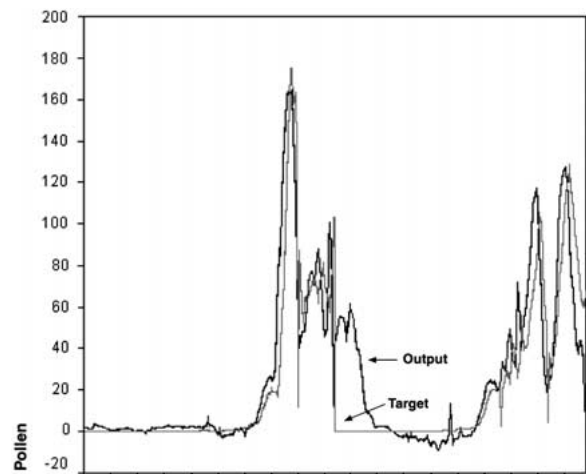


Figure 3. Simulation on Bologna site (two years).

While there has been many efforts in neural model development and learning algorithms, the design of the optimal network structure for a given task has remained a problem. Designing an optimal structure involves finding the structure with the smallest size network (parsimonious network), which produces optimal errors for trained as well as untrained cases. An often-cited rule of thumb (Weigend et al., 1991) is

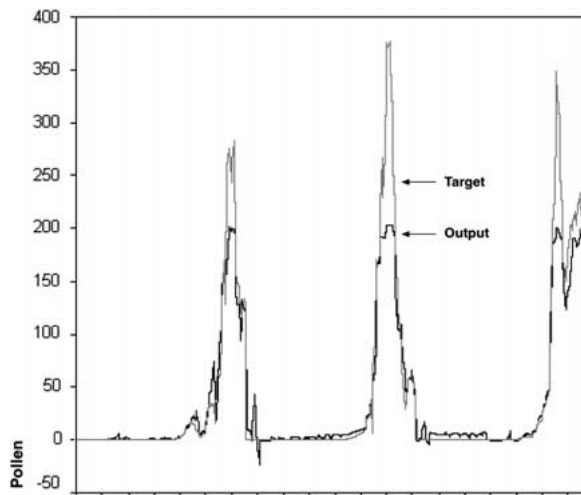


Figure 4. Simulation on Modena site (three years).

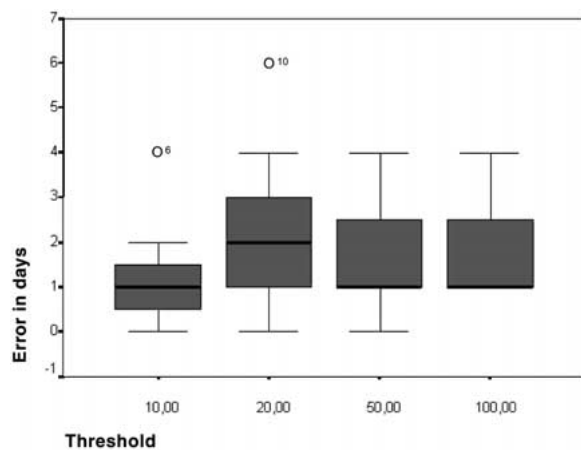


Figure 5. Test on pollen thresholds exceeding.

that the number of weights should be less than a tenth of the number of training samples. Another rule of thumb, known as the Baum-Haussler rule (Baum and Haussler, 1988), can be used to determine the number of hidden layer neurons.

In any nontrivial time series, new values depend not only on the immediately prior value but also on many preceding values. Using too few inputs can result in inadequate modelling, whereas too many inputs can excessively complicate the network, in terms of the training times and dimensionality of the network.

Autocorrelation analysis or Principal Component Analysis (PCA) could be effective in identifying input patterns and improving network performances

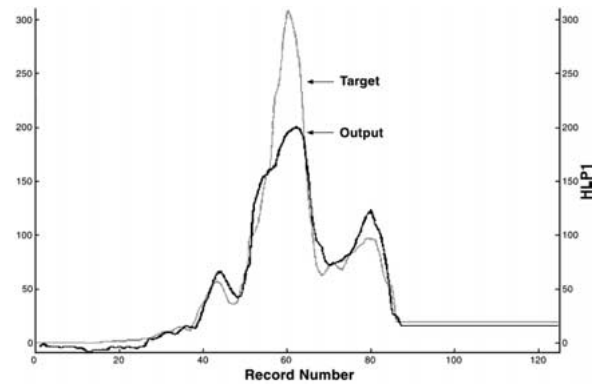


Figure 6. Performance of the neural network on the test set (one year).

(Chakraborty et al., 1992; Tiao and Tsay, 1989; Hann and Steurer, 1996).

In Figure 6 we can see the test set with the performance of the network. We can underline two facts:

1. Our main interest was to correctly simulate the starting season. Here, as in other simulations, we can see that this goal has been reached;
2. The network often underestimate the maximum value. This is a problem not yet solved. One possible reason is the absence of important predictor variables (as wind or relative humidity).

Nevertheless, for an efficacy preventive communication to allergenic people, it could be less important to exactly predict high values.

The correct estimation of thresholds overcoming is more important in disease prevention. We choose 4 thresholds next to the critical period for allergenic people, and we measured the error of delay or anticipation days. The mean error never exceeds 2 days: this represents a better result from a health prevention point of view.

To simulate a realtime case, we firstly tested the consistency of the network, introducing random errors on inputs: we can see that errors in meteorological forecasting are well tolerated, while errors in pollen forecasts are well tolerated for a few days only, and this is clear because such an error increases day by day because of its auto-correlation with the output variable.

5. Conclusion

A neural network model for grass pollen forecasting has been presented. The performance of this empirical

model was satisfactory, in particular regarding the prediction of the starting date of the pollen season.

We demonstrated that the relationships between pollen concentrations and meteorological situations are independent from site. This means that such models can understand the differences in area through phenological and meteorological information. In regards to differences in time, the neural network model showed its intrinsic robustness in forecasting unknown years.

Possible improvements in model performances could be the introduction of other predictor variables, such as wind and humidity.

The model is extendible to any area in which time series of pollen and meteorological data are available. The differences between areas could be interpreted by the model by means of changes in weight matrix and in meteorological thresholds values.

References

- Baum E.B. and Haussler D.: 1988, What size net gives valid generalization? *Neural Computation* **1**, 151–160.
- Chakraborty K., Mehrotra K., Mohan C.K. and Ranka S.: 1992, Forecasting the behavior of multivariate time series using neural networks. *Neural Networks* **5**, 961–970.
- Frenguelli G. and Bricchi E.: 1998, The use of the pheno-climatic model for forecasting the pollination of some arboreal taxa. *Aerobiologia* **14**, 39–44.
- Hann T.H. and Steurer E.: 1996, Much ado about nothing? Exchange rate forecasting: Neural networks vs. linear models using monthly and weekly data. *Neurocomputing* **10**, 323–339.
- Moseholm L., Weeke E. and Petersen B.N.: 1987, Forecast of pollen concentration of Poaceae (grasses) in the air by time series analysis. *Pollen and Spores* **29**, 305–322.
- Norris-Hill J.: 1998, A method to forecast the start of the Betula, Platanus and Quercus pollen season in North London. *Aerobiologia* **14**, 165–170.
- Puppi Branzi G. and Zanotti A.L.: 1992, Estimate and mapping of the activity of airborne pollen sources. *Aerobiologia* **8**, 69–74.
- Rumelhart D.E., McClelland J.L. and the PDP Research Group: 1986, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*. MIT Press, Cambridge.
- Stark P.C., Ryan L.M., McDonald J.L. and Burge H.A.: 1997, Using meteorological data to predict daily ragweed levels. *Aerobiologia* **13**, 177–184.
- Tiao G.C. and Tsay R.S.: 1989, Model specification in multivariate time series. *Journal of the Royal Statistical Society B* **51**, 157–213.
- UCB Institute of Allergy: 1999, *European Allergy White Paper Update*.
- Weigend A.S., Huberman B.A. and Rumelhart D.E.: 1991, Generalization by weight elimination with application to forecasting. *Advances in Neural Information Processing Systems* **3**, 875–882.
- Windsor C.G. and Harker A.H.: 1990, Multi-variate financial index prediction – a neural network study. *Proceedings of International Neural Network Conference*. Paris, France, pp. 357–360.

