Int J Biometeorol (2005) 49:310–316
DOI 10.1007/s00484-004-0247-x

ORIGINAL ARTICLE

M. Castellano-Méndez · M. J. Aira · I. Iglesias · V. Jato · W. González-Manteiga

Artificial neural networks as a useful tool to predict the risk level of *Betula* pollen in the air

Received: 30 September 2004 / Revised: 8 November 2004 / Accepted: 15 November 2004 / Published online: 13 January 2005

Abstract An increasing percentage of the European population suffers from allergies to pollen. The study of the evolution of air pollen concentration supplies prior knowledge of the levels of pollen in the air, which can be useful for the prevention and treatment of allergic symptoms, and the management of medical resources. The symptoms of *Betula* pollinosis can be associated with certain levels of pollen in the air. The aim of this study was to predict the risk of the concentration of pollen exceeding a given level, using previous pollen and meteorological information, by applying neural network techniques. Neural networks are a widespread statistical tool useful for the study of problems associated with complex or poorly understood phenomena. The binary response variable associated with each level requires a careful selection of the neural network and the error function associated with the learning algorithm used during the training phase. The performance of the neural network with the validation set showed that the risk of the pollen level exceeding a certain threshold can be successfully forecasted using artificial neural networks. This prediction tool may be implemented to create an automatic system that forecasts the risk of suffering allergic symptoms.

M. Castellano-Méndez (✉) · W. González-Manteiga
Department of Statistics and Operation Research,
Universidade de Santiago de Compostela,
Campus Sur. 15782, Santiago de Compostela A Coruña, Spain
e-mail: mariacmendez@hotmail.com
Tel.: +34-981-563100
Fax: +34-981-597054

M. J. Aira
Department of Vegetal Biology,
Universidade de Santiago de Compostela,
Campus Sur. 15782, Santiago de Compostela A Coruña, Spain

I. Iglesias · V. Jato
Departament of Vegetal Biology and Soil Sciences,
University of Vigo,
Campus As Lagoas, 32004 Ourense, Spain

Keywords Aerobiology · Allergenic risk · Binary data · *Betula* pollen · Error function · Neural networks · Pollen level · Probability function

Introduction

Birch is an anemophilous tree with high pollen production (Moore and Webb 1978; Lewis et al. 1983), whose allergenic capacity has been cited by numerous authors (Spieksma 1990; Norris-Hill and Emberlin 1991; D'Amato and Spieksma 1992). Its pollen is considered to be the main cause of pollinosis in North and Central Europe (Wihl et al. 1998; Spieksma et al. 1995), not only during its pollen season but also during previous and subsequent periods, since birch pollen can easily be transported over long distances (Wallin et al. 1991; Hjelmroos 1991). In such cases, the antigenic activity seems to be linked to allergens, characteristic of birch pollen grains, deposited on dust particles inside houses which can trigger the onset of allergic processes, even up to 2 months after maximum pollen concentrations in the air were recorded (Ekebom et al. 1996; Rantio-Lehtimäki et al. 1996). The prevalence of birch pollen allergy reaches 13 to 60% in populations affected by pollinosis in some localities in Galicia, Spain (Arenas et al. 1996; Aira et al. 2001), and 19% in Santiago de Compostela (Dopazo 2001).

Several researchers have carried out aeropalynological studies on this taxon, in order to construct a model of the seasonal and daily behaviour of birch pollen and to ascertain the influence of different meteorological parameters on pollen concentration (Spieksma et al. 1989; Atkinson and Larsson 1990; Norris-Hill and Emberlin 1991; Spieksma et al. 1995; Aira et al. 1998; Jato et al. 2000; Latalowa et al. 2002). In this way, models can be established in order to predict both the beginning and the severity of the pollen season. Different factors were used as predictors of the start of pollen season in these different models. The sum of the temperatures from a given date was utilized by Clot (2001), Caramiello et al. (1994), Ruffaldi and Greffier (1991). In other works phenological factors as chilling units and growing degree days have been used as predictors (Andersen 1991). Larsson (1993) used the method of accumulative activity and Laaidi (2001) both the sum of temperatures and a multiple regression model. Different works have been conducted with the aim of producing models to predict the mean pollen concentration 1 day in advance by using linear regression (Rodríguez-Rajo 2000; Méndez 2000) or time series (Moseholm et al. 1987). However, there are no studies which aim to establish the risk of allergy being caused by a given quantity of pollen grains in the air.

The symptoms of *Betula* pollinosis can be provoked by different levels of *Betula* pollen in the air, depending on individual behaviour differences. Nevertheless, several threshold values have been mentioned as limiting for the provocation of symptoms. Ninety percent of clinically sensitive subjects showed symptoms when 80 pollen grains/m^3 was reached and the onset of severe symptoms was recorded with concentrations higher than 30 pollen grains/m^3 (Negrini et al. 1992). Corsico (1993) considered the same level as the threshold for the onset of allergic symptoms. Birch pollen is very abundant in the air of Santiago de Compostela during March and April and concentrations higher than 100 pollen grains/m^3 are frequent. The daily maximum levels are registered in the afternoon—between 12 h and 18 h—and they are coincident with the highest frequency of allergy symptoms (Dopazo 2001)

Birch is represented in Galicia by one species, *Betula alba* L. (Moreno 1990). The former is widely distributed in this area and as the dominant tree it forms altimontane oro-Cantabrian acidophilic forests, with a clearly Euro-Siberian distribution. They are found above an altitude of 1,150 m, being the last tree formations of the altitudinal sequence, with montane thermo-climates and hyper-humid ombro-climates (Izco 1994). Their limit, although fairly controversial, is situated in the Galician mountain ranges of Ancares and Caurel (Costa et al. 1990). In this same area, but on siliceous soils and with a greater Mediterranean influence, there are birch forests in the Galician-Portuguese altimontane layer and in the Ourense-Sanabrian supra-Mediterranean layer. In Galicia's mountains and foothills non-climatic birches may be found, in substitution for montane oak groves, which are located on acidic soils and with altitudinal limits between 600 and 1,100 m. In the Euro-Siberian region, birch may form part of riparian forests, along with *Alnus glutinosa*, *Salix atrocinera* and *Frangula alnus*. There are *Betula* as ornamental trees near the spore-trap, and this is one of the reasons why the concentrations of *Betula* pollen in Santiago are among the highest in Galicia.

The aim of this research is to predict the days of high allergenic risk during *Betula* pollination, using artificial neural networks, in order to alert allergists and the population with allergy problems to a potential risk situation. Artificial neural networks (ANNs) are a complete statistical tool for data analysis (Bishop 1995). The ANN's origin dates back to the middle of the last century when an interdisciplinary group of biologists, psychologists and engineers interested in understanding the functioning of the human brain was created (Rosenblatt 1958). ANNs try to artificially reproduce the human ability to take decisions by simulating the human brain's basic unit, the neuron, and the interconnections between neurons that allow them to work together and save information.

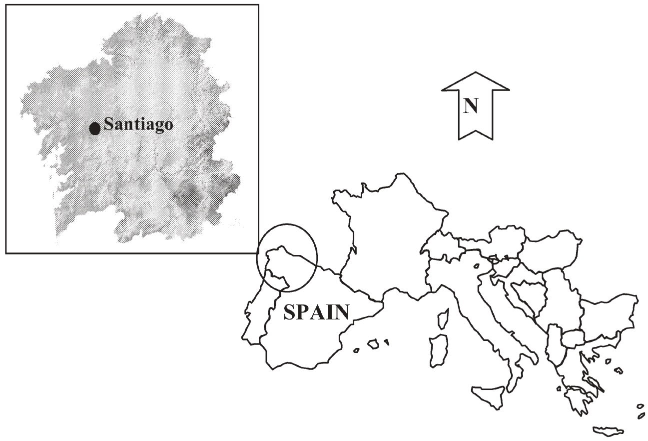

Fig. 1 Location of Santiago de Compostela in Europe

Recently, ANNs have been extended successfully to very different fields, from hydrology to finances. Neural networks was been also used in aerobiological studies, to achieve predictive models for the improvement of daily pollen concentration forecasts (Ranzi 2000; Hidalgo et al. 2002; Sánchez-Mesa et al. 2002)

Materials and methods

The study was conducted in the city of Santiago de Compostela, located in northwest Spain (Fig. 1). Pollen monitoring was carried out from 1993 to 2001 by means of a 7-day volumetric air sampler (Lanzoni VPPS 2000) situated approximately 25 m above the ground level. The methodology recommended by the Spanish Aerobiological Network (REA) was used to process and interpret the samples (Domínguez 1995).

Three data series were considered (Chakraborty 1992). The daily pollen, pollen$_t$, expressed as grains/m^3, and two exogenous meteorological series, the daily rainfall, DR$_t$, expressed as l/m^2, and daily mean temperature, DMT$_t$, expressed as °C.

The aim of this work was not forecast the pollen concentration but the level of allergenic risk. Given a level of pollen, lev, a new binary variable Y_t can be defined. This variable takes the value 1 if pollen$_t$ is over the quantity lev and 0 otherwise. The selected levels were 20, 30, 70 and 80 grains/m^3. For levels 20 and 30 the variable Y_t measures the risk of the onset of allergic symptoms, and for levels 70 and 80 Y_t measures the risk of severe symptoms for 90% of the most allergic population. The dependent or target variables are the Y_t values associated with each level.

The selected independent variables were the previous day's rainfall, DR$_{t-1}$, the previous day's mean temperature, DMT$_{t-1}$, and the previous day's pollen concentration, pollen $_{t-1}$.

Artificial neural networks

The statistical method used for this study and to forecast of the risk level associated with *Betula* pollen is the artificial neural network technique (Ripley 1996). ANNs are "spoken data" methods, i.e., the

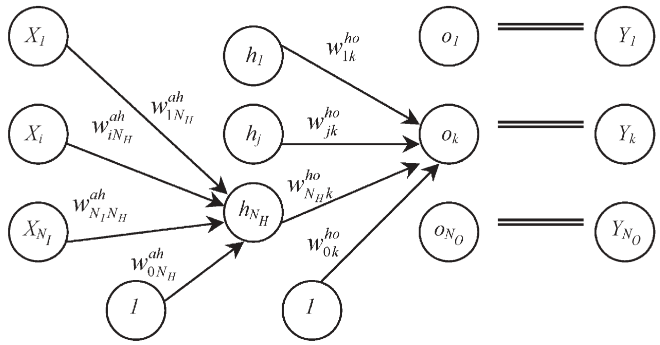

Fig. 2 Diagram of a multilayer perceptron with one hidden layer (N_i-N_h-N_o) with N_i input variables X_i, N_h nodes in the hidden layer with processed information h_j, N_o predictions or output o_k, N_o target variables y_k and the ANN's weights between connections w

structure of an ANN, and thus the relationship between the input and the output, depends on a historical observations set, named the *training set*, used for network 'learning'. The training set is a data collection related to past situations and associated with them, the neural network correct answer or a variable closely related to the unknown correct answer.

During the training phase the ANN will 'learn' the underlying relationship between the inputs and outputs by means of a learning algorithm that compares the networks outputs with the real outputs. The learning algorithm used was the*backpropagation* algorithm. (Rumelhart et al. 1986; Chauvin et al. 1995). After the training phase, when the ANN works with a new situation, it will behave in line with the learning set. So ANNs become interesting in unknown or complex structure data source phenomenon forecasting. Pollen dispersion is a very complex problem that involves a large amount of meteorological (wind direction, wind velocity, rainfall...), ecological (forest situation and selected species concentration in the vicinity of the prediction location...), and topographic (hills, valleys, rivers, towns, exact location) information, which is not always available. ANNs became a useful tool because it is not necessary to determine all of these characteristics. Instead, a general structure ANN and a diverse and extensive data set for training can be used. The performance of ANN general models are based on their universal approximation proprieties (Cybenko 1989; Hornik et al. 1989; Park and Sandberg 1991)

As the relationship between input variables and output or target variables, in this case meteorological or ecological features of the area, may change over time, it is useful to re-train the ANN periodically, increasing the training set with new data that reflect the changes in the variables with time.

One of the most popular ANN architectures is the *multilayer perceptron* (MLP) (Rosenblatt 1958). In an MLP, the nodes or basic unit of information processors, are distributed in layers. Only the nodes of consecutive layers can be connected. The first layer is named the *input layer*; the last is named the *output layer*, while the layers between them are known as *hidden layers*. An MLP with one hidden layer, represented in Fig. 2, has been considered.

ANN for binary response data

The target variable Y_t is a binary variable, i.e., it takes only two values, 1 or 0. We have focused on a special family of ANNs, with sigmoidal output nodes, appropriate for binary target variable data processing.

For prediction problems, the aim is to approximate the expected value of the target variable, conditioned by the independent variables. For binary target variables this expectation is the probability of Y_t taking the value 1, conditioned by the input variable values (Goldberger 1973; Agresti 1990). This conditional probability can be considered as a unknown function of the independent variables that takes values from 0 to 1 (McCullagh and Nelder 1989). Estimating this function is the aim of the ANN. Given a probability predictor a family of classifiers $F = \{C_p / p \in [0, 1]\}$can be constructed. Each one of these classifiers, determined by a given p, allows the building of a binary variable from the probability predictor by means of the following procedure. If the predicted probability is less than p the predicted value for Y_t will be 0, and otherwise will be 1. In order to obtain a Y_t estimation, $\hat{Y}_t$, we must select a classifier from this family, choosing a threshold p. The selected p was 0.5.

Using the ANN probability prediction, the forecast of Y_t may be obtained as follows: if the predicted probability is less than 0.5, the predicted value for Y_t will be 0, and otherwise will be 1.

Error function for the binary target variable

The density function estimation, as well as the probability estimation, are unsupervised learning problems. The real probabilities, as well as the real density, are not available; instead there is a binary variable, Y, with some kind of information about the probability value.

In the introduction it was explained that network learning is based on a training algorithm. This kind of algorithm compares the model output with the real target variable values and modifies the network parameters in order to minimize the differences between them. These differences are measured using an error function. Selecting the appropriate error function for the data is essential for training the network successfully. In binary dependent variable problems, during the training phase, the training algorithm will compare the binary target variable with the continous network output, and the estimation of the probability of the target variable takes the value 1.

When working with binary targets, the usual error function is the deviance (Hastie 1987), dev (Table 1), that somehow measures the credibility of the probability estimation considering the binary variable value.

As we have explained, a probability estimation provides us with a binary variable $\hat{Y}$, an estimation of the target binary variable Y. This is a two-class classification problem. So the total misclassification probability, *mcp*, can be considered as the error function; the *mcp* is the probability of having a target variable valued at 0 and the estimation $\hat{Y}$ valued at 1, MC_{I}, added to the probability of having a target variable valued at 1 and the estimation $\hat{Y}$ valued 0, MC_{II} (Table 1).

We can consider separately the type I error, $\mathrm{error}_{\mathrm{I}}$, and the type II error, $\mathrm{error}_{\mathrm{II}}$ (Table 1). The type I error is the probability of estimation $\hat{Y}$ being 0 conditioned to the target variable Y taking a

Table 1 Equations

Deviance	$\mathrm{Dev}(o, Y) = -2[Y Log(o) + (1 - Y)\mathrm{Log}(1 - o)]$, being o the probability estimation and Y the binary variable	1.1
Misclassification probability	$\mathrm{mcp} = P(\hat{Y} \neq Y) = MC_{\mathrm{I}} + MC_{\mathrm{II}} = P(\hat{Y} = 1,\ Y = 0) + P(\hat{Y} = 0,\ Y = 1)\mathrm{mcp} = P(Y = 0) \times P(\hat{Y} = 1 \mid Y = 0) + P(Y = 1) \times P(\hat{Y} = 0 \mid Y = 1)$	1.2
Error type I	$\mathrm{error}_{\mathrm{I}} = P(\hat{Y} = 0 \mid Y = 1)$	1.3
Error type II	$\mathrm{error}_{\mathrm{II}} = P(\hat{Y} = 1 \mid Y = 0)$	1.4
Relationship	$\mathrm{mcp} = P(Y = 1) \times \mathrm{error}_{\mathrm{I}} + P(Y = 0) \times \mathrm{error}_{\mathrm{II}}$	1.5
Empirical probability	$P(A) = \frac{\text{Number of cases where } A \text{ occurs}}{\text{Total number of cases}} = \frac{\mathrm{card}(\{A \text{ occurs}\})}{\mathrm{card}(\{\text{all examples}\})}$	1.6

Table 2 Artificial neural network equations. Matrix of parameters

ANN model	$\hat{Y}_t = \mathrm{threshold}(o_t - 0.5)$ $o_t = \mathrm{sigmoid}\left(-w_{01}^{\mathrm{ho}} + w_{11}^{\mathrm{ho}} \times h_1 + w_{21}^{\mathrm{ho}} \times h_2 - w_{31}^{\mathrm{ho}} \times h_3 + w_{41}^{\mathrm{ho}} \times h_4 - w_{51}^{\mathrm{ho}} \times h_5\right)$ $h_j = \mathrm{sigmoid}\left(w_{0j}^{\mathrm{ah}} - w_{1j}^{\mathrm{ah}} \times DR_{t-1} - w_{2j}^{\mathrm{ah}} \times \mathrm{DMT}_{t-1} + w_{3j}^{\mathrm{ah}} \times \mathrm{pollen}_{t-1}\right)$, with $1 \leqslant j \leqslant 5$ $\mathrm{sigmoid}\,(x) = \frac{\exp(x)}{1+\exp(x)}$ $\quad \mathrm{threshold}(x) = 1 \text{ if } x \geqslant 0 \text{ and } 0 \text{ if } x < 0$
Matrix of ANN parameters	$\begin{pmatrix} w_{01}^{\mathrm{ah}} & w_{02}^{\mathrm{ah}} & w_{03}^{\mathrm{ah}} & w_{04}^{\mathrm{ah}} & w_{05}^{\mathrm{ah}} & \\ w_{11}^{\mathrm{ah}} & w_{12}^{\mathrm{ah}} & w_{13}^{\mathrm{ah}} & w_{14}^{\mathrm{ah}} & w_{15}^{\mathrm{ah}} & \\ w_{21}^{\mathrm{ah}} & w_{22}^{\mathrm{ah}} & w_{23}^{\mathrm{ah}} & w_{24}^{\mathrm{ah}} & w_{25}^{\mathrm{ah}} & \\ w_{31}^{\mathrm{ah}} & w_{32}^{\mathrm{ah}} & w_{33}^{\mathrm{ah}} & w_{34}^{\mathrm{ah}} & w_{35}^{\mathrm{ah}} & \\ w_{01}^{\mathrm{ho}} & w_{11}^{\mathrm{ho}} & w_{21}^{\mathrm{ho}} & w_{31}^{\mathrm{ho}} & w_{41}^{\mathrm{ho}} & w_{51}^{\mathrm{ho}} \end{pmatrix}$

Table 3 ANN selected matrix of parameters

Level	Matrix of parameters	Level	Matrix of parameters
20	$\begin{pmatrix} -0.752 & 1.430 & -1.375 & -1.846 & -1.041 \\ -1.643 & 0.657 & -1.282 & 1.685 & 0.158 \\ -0.391 & -1.022 & 0.335 & -0.250 & 1.530 \\ 1.938 & 0.530 & -0.545 & 0.858 & 0.174 \end{pmatrix}$ $(-0.109 \quad 0.748 \quad 0.082 \quad -0.145 \quad 0.045 \quad -0.346)$	70	$\begin{pmatrix} -1.323 & 0.033 & 1.817 & -0.675 & 1.359 \\ 0.006 & 0.789 & -0.082 & -0.543 & 0.024 \\ -0.804 & -2.078 & 0.804 & -1.706 & -0.861 \\ -1.931 & 0.730 & -1.171 & 1.900 & 0.900 \end{pmatrix}$ $(-1.547 \quad 0.521 \quad 0.794 \quad -1.155 \quad 0.650 \quad 0.364)$
30	$\begin{pmatrix} 1.275 & 1.129 & 1.769 & 1.149 & 0.034 \\ 1.883 & 1.398 & -2.139 & 1.853 & 0.007 \\ 1.051 & -0.734 & 1.298 & 0.328 & -1.081 \\ -1.482 & 1.436 & 0.096 & 0.533 & 1.812 \end{pmatrix}$ $(-0.355 \quad -0.306 \quad 0.858 \quad -0.103 \quad -0.652 \quad 0.880)$	80	$\begin{pmatrix} 0.411 & -0.379 & 0.017 & 1.221 & 1.040 \\ 0.352 & 0.086 & 1.013 & -0.863 & -0.150 \\ -1.677 & -1.158 & -1.745 & 2.001 & -0.906 \\ 0.663 & 1.527 & 1.081 & -0.576 & 1.509 \end{pmatrix}$ $(-1.179 \quad 0.246 \quad 0.370 \quad 0.403 \quad 0.866 \quad 0.560)$

value of 1; in our problem this is the probability of predicting a below-threshold level day on a over-threshold level day (false security). The type II error, the probability of estimation of $\hat{Y}$ to be 1, conditioned to value 0 of the target variable Y , that is, the probability of predicting a over-threshold level day on a below-threshold level day (false alarm). Balancing both errors is necessary. Usually one of the errors is fixed, and the classifier that minimizes the other error is selected. All these error measures are estimated using the empirical probability formula (Table 1) over a collection of observations, F. Several equations will involve a function, namely the cardinal, card. The cardinal of a set A is the number of elements that belong to the set A.

In many real problems, e.g., ecological, epidemiological or medical problems, both error types, $\mathrm{error}_{\mathrm{I}}$ and $\mathrm{error}_{\mathrm{II}}$, do not have the same importance and one of them must be penalized. One reason may be the different proportion of cases with a target variable valued at 0. In many problems, the economic or health consequences of a false negative are very different from the consequences of a false positive, so it is necessary to penalize separately both errors.

The deviance does not distinguish between both kind of errors, so it leads the ANN performance to reach an equal balance between both empirical errors. In this problem, the proportion of 1-valued days is very small compared to the proportion of 0-valued days, so the deviance is not the best choice of error function so another loss function must be considered. We have defined the follow error function:

$$\mathrm{error}_{K_1 \cdot K_2}(o, Y) = K_1 \cdot Y \cdot (1 - o) + K_2 \cdot (1 - Y) \cdot o, \text{ with } K_1, K_2 \geqslant 0,$$

with o representing the ANN output, i.e., the estimated probability, and Y the binary target variable

This function penalizes both errors separately. The constants K_1 and K_2 will decode which error is more serious. During the training process this function can be minimized or equally:

$$\mathrm{error}_K(o, Y) = K \cdot Y \cdot (1 - o) + (1 - Y) \cdot o, \text{ with } K \geqslant 0$$

The constant K value determines the penalization. If K>1 the class 1 observation misclassification will be penalized; on the other hand if K<1 the class 1 observation misclassification will be penalized; finally if K=1 both misclassifications will be considered as equally serious.

Results and discussion

Four levels of pollen concentration were considered, lev = 20, 30, 70, 80 g/m^3. We have generated four artificial neural networks, one for each alert level. Explicit expressions of the four neural networks are showed in Tables 2 and 3.

The period between 2 January 1993 and 11 March 2000, is the data set used to train the neural networks. In order to evaluate the performance of the ANNs before a

Fig. 3 Predicted conditional probabilities against the target variable, for pollen levels 30 and 80 g/m^3 over a training set section and the validation set.The *dotted horizontal line* separates the two prediction zones. If the predicted conditional probability is above the line, the binary prediction will take the value 1, and if it is below the dotted line, the binary prediction will take the value 0

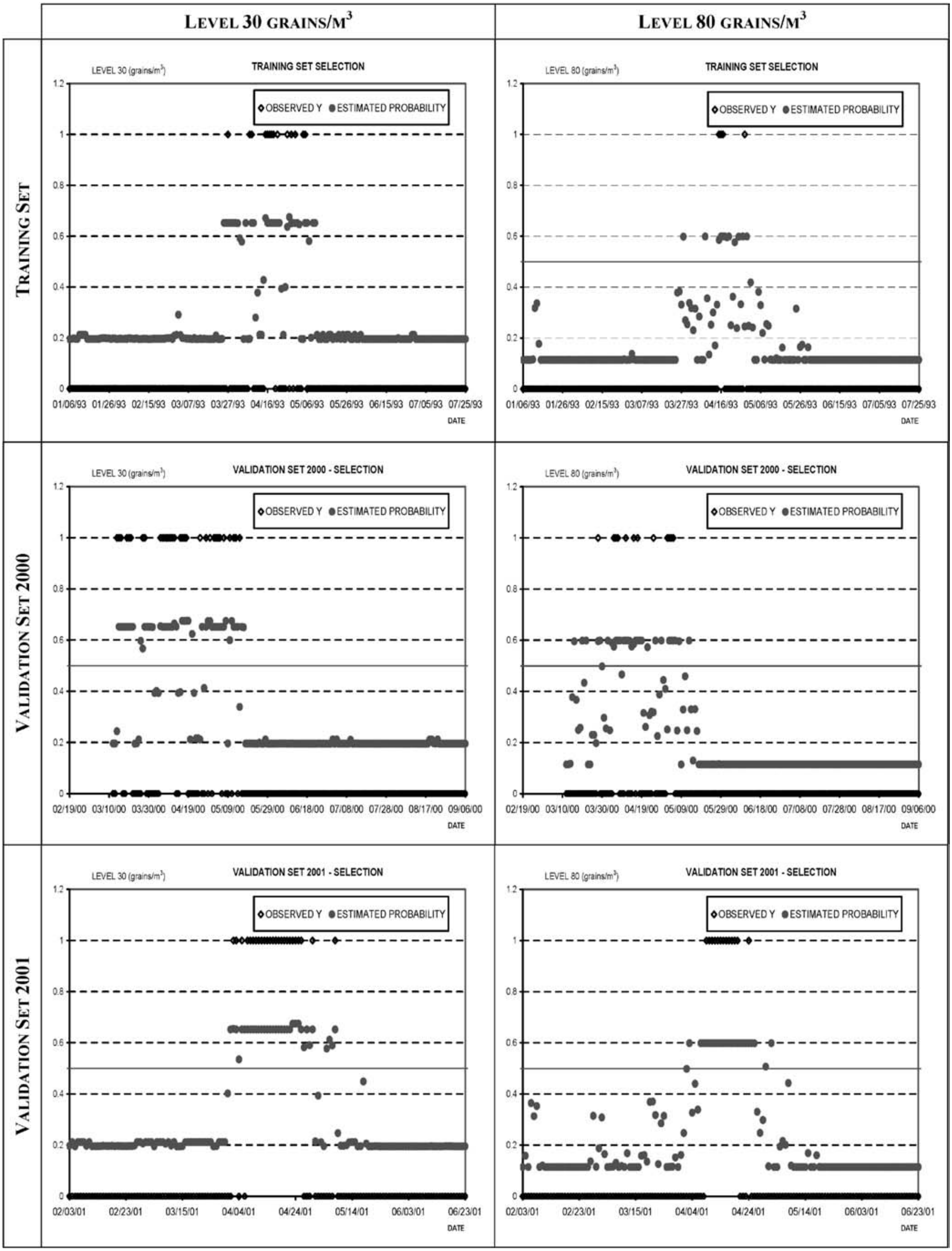
LEVEL 30 GRAINS/M3
LEVEL 80 GRAINS/M3
TRAINING SET
VALIDATION SET 2000
VALIDATION SET 2001
LEVEL 30 (grains/m3)
LEVEL 80 (grains/m3)
TRAINING SET SELECTION
VALIDATION SET 2000 - SELECTION
VALIDATION SET 2001 - SELECTION
OBSERVED Y
ESTIMATED PROBABILITY
DATE
01/06/93 01/26/93 02/15/93 03/07/93 03/27/93 04/16/93 05/06/93 05/26/93 06/15/93 07/05/93 07/25/93
02/19/00 03/10/00 03/30/00 04/19/00 05/09/00 05/29/00 06/18/00 07/08/00 07/28/00 08/17/00 09/06/00
02/03/01 02/23/01 03/15/01 04/04/01 04/24/01 05/14/01 06/03/01 06/23/01
0 0.2 0.4 0.6 0.8 1 1.2

Table 4 Equations part II

Binary value proportion	$Z_{\text{lev}} = \frac{\text{card}\{t/Y_t=0\}}{\text{card}\{t/Y_t=1\}}$	4.1
Good classification measures	$\text{GC}_{\text{I}} = 1 - \hat{P}\left(\hat{Y}=0\|Y=1\right) = \hat{P}\left(\hat{Y}=1\|Y=1\right) = \frac{\text{card}\left\{t/Y_t=\hat{Y}_t=1\right\}}{\text{card}\{t/Y_t=1\}}$	4.2
	$\text{GC}_{\text{II}} = 1 - \hat{P}\left(\hat{Y}=1\|Y=0\right) = \hat{P}\left(\hat{Y}=0\|Y=0\right) = \frac{\text{card}\left\{t/Y_t=\hat{Y}_t=0\right\}}{\text{card}\{t/Y_t=0\}}$	4.3

Table 5 Note that GC_{I} represents the empirical probability of good classification a over-level pollen day, and GC_{II} represents the empirical probability of good classification a below-level pollen day

		Lev = 20	Lev = 30	Lev = 70	Lev = 80
GC_{I}	Training set 1993–1999	0.95	0.95	0.86	0.86
	Validation set 2000	0.92	0.88	0.83	0.83
	Validation set 2001	1.00	1.00	1.00	1.00
GC_{II}	Training set 1993–1999	0.95	0.95	0.98	0.98
	Validation set 2000	0.93	0.92	0.92	0.93
	Validation set 2001	0.96	0.97	0.97	0.97

new situation, we have considered a validation set that started on 12 March 2000 and ended on 1 December 2001. The selected validation set contained two consecutive *Betula* pollination periods because of the biannual behaviour of the studied pollen.

The parameter K involved in the error function takes different values for the different levels. The K values have been selected around the value of the empirical proportion between the below-threshold level pollen days and the over-threshold level pollen days, Z_{lev}. [Table 4, (section 4.1)]. In fact, if Z_{lev} is select as K the ANN minimizes the misclassification probability. During *Betula* pollination the number of days with pollen in the air is less than the number of days when the pollen concentration is zero, so K is greater than one. Higher values of lev bear higher Z_{lev} values, and thus higher K values.

In order to show, compare and discuss the results we must consider two different complementary concordance measures. The proportion of good classification over the observations with the target variable value equals 1, GC_{I}, and the proportion of good classification over the observations with the target variable value equals 0, GC_{II} (Table 4). Table 5 shows the results obtained.

Figure 3 shows the predicted conditional probabilities against the target variable, for the levels 30 and 80 over a training set section and the validation set.

Pollen forecasting has become an important aim in aerobiology. The objective is to provide accurate information on pollen in the air to sensitive users in order to help them optimize their medication.

Usually, aerobiological information spread is made by using fixed categories in relation to threshold values. In this sense, our goal was to look for a model to allow us to know the probability of the *Betula* daily mean pollen concentration increasing over several threshold values, some of which have previously been cited as responsible for allergic symptomatology (Negrini et al. 1992; Corsico 1993).

Neural networks provided us a good result for forecasting the probability that a given value of *Betula* pollen concentration occurs. Between 83 and 92% of the occurrences in the year 2000, and 100% of the occurrences in the year 2000, of pollen concentration values reaching the thresholds considered (>20, >30, >70 and >80 g/m^3) were predicted in advance. Similarly, in the year 2000 between 92 and 93% of the occurrences of pollen concentrations below a threshold value were correctly predicted, while for 2001 this figure was between 96 to 97%. Therefore, neural networks are a good tool to predict the probability of pollen concentrations reaching or exceeding a threshold value, and thus help the dissemination of aerobiological information to the population suffering from allergic problems. This is a first step towards the automatization of the prediction system.

Acknowledgements This study was partially funded by the Xunta de Galicia's Environment Department (PGDIT00MAM38301PR). D.W. González-Manteiga's work was funded by BFM2002-03213 (European FEDER support included) from the Spanish Ministry of Science and Technology, and by PGIDIT03PXIC20702PN Dirección Xeral de I+D Xunta de Galicia.

References

Aira MJ, Jato V, Iglesias I (1998) *Alnus* and *Betula* pollen content in the atmosphere of Santiago de Compostela, north-western Spain (1993–1995). Aerobiologia 14:135–140

Aira MJ, Ferreiro M, Iglesias I, Jato V, Marcos C, Varela S, Vidal C (2001) Aeropalinología de cuatro ciudades de Galicia y su incidencia sobre la sintomatología alérgica estacional. Actas XIII Simposio de la A.P.L.E.Cartagena

Agresti A (1990) Categorical data analysis. Wiley, New York

Andersen TB (1991) A model to predict the beginning of the pollen season. Grana 30:269–275

Arenas L, González C, Tabarés JM, Iglésias I, Méndez J, Jato V (1996) Sensibilización cutánea a pólenes en pacientes afectos de rinoconjuntivitis-asma en la población de Ourense en el año 1994–95. 1st European symposium on aerobiology, Santiago de Compostela, 11–13 September 1996

Atkinson H, Larsson KA (1990) A 10-year record of the arboreal pollen in Stockholm, Sweden. Grana 29:229–237

Bishop C (1995) Neural networks for pattern recognition. Clarendon, Oxford

Caramiello R, Siniscalco C, Mercalli L, Potenza A (1994) The relationship between airborne pollen grains and unusual weather conditions in Turin (Italy) in 1989, 1990 and 1991. Grana 33:327–332

Chakraborty K, Mehrotra K (1992) Forecasting the behaviour of a multivariate time series using neural networks. Neural Netw 5:961–970

Chauvin Y, Rumelhart DE (1995) Backpropagation: theory, architectures, and applications. Lawrence Erlbaum Associates, Mawah, N.J.

Clot B (2001) Airborne birch pollen in Neuchâtel (Switzerland): onset, peak and daily patterns. Aerobiologia 17:25–29

Corsico R (1993) L'asthme allergique en Europe. In: Spieksma FTM, Nolard N, Frenguelli G, Van Moerbeke D (eds) Pollens de l'air en Europe. UCB, Braine-l'Alleud, pp 19–29

Costa M, Higueras J Morla C (1990) Abedulares de la Sierra de San Mamed (Orense, España). Acta B Malacitana 15:253–265

Cybenko G (1989) Approximations by superpositions of a sigmoidal function. Math Contr Signals Syst 1 2:303–314

D'Amato G Spieksma FTM (1992) European allergenic pollen types. Aerobiologia 8:447–450

Domínguez E (1995) La Red Española de Aerobiología. Monografía REA 1:1–8

Dopazo A (2001) Variación estacional y modelos predictivos de polen y esporas aeroalergénicos en Santiago de Compostela. Dissertation, University of Santiago de Compostela

Ekebom A, Verterberg O, Hjelmroos M (1996) Detection and quantification of airborne birch pollen allergens on PVDF membranes: immunoblotting and chemiluminescence. Grana 35:113–118

Goldberger AS (1973) Correlations between binary choices and probabilistic predictions. J Am Stat Assoc 68:84

Hastie T (1987) A closer look at the deviance. Am Stat 41:16–20

Hidalgo PJ, Mangin A, Galán C, Hembise O, Vázquez LM, Sánchez O (2002). An automated system for surveying and forecasting *Olea* pollen dispersion. Aerobiologia 18:23–31

Hjelmroos M (1991) Evidence of long-distance transport of *Betula* pollen. Grana 30:215–228

Hornik KM, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neural Netw 2:359–366

Izco J (1994) O bosque Atlántico. In: Vales C (ed) Os Bosques Atlánticos Europeos. Bahía, La Coruña, pp 13–49

Jato V, Aira MJ, Iglésias MI, Alcázar P, Cervigón P, Fernández D, Recio M, Ruíz L, Sbai L (2000) Aeropalynology of birch (*Betula* sp.) in Spain. Pollen 10:39–49

Laaidi M (2001) Regional variations in the pollen season of *Betula* in Burgundy: two models for predicting the start of the pollination. Aerobiologia 17:247–254

Larsson K (1993) Prediction of the pollen season with a cumulated activity method. Grana 32:111–114

Latalowa M, Mietus M, Uruska A (2002) Seasonal variations in the atmospheric *Betula* pollen count in Gdansk (southern Baltic coast) in relation to meteorological parameters. Aerobiologia 18:33–43

Lewis WH, Vinay P, Zenger VE (1983) Airborne and allergenic pollen of North America. Johns Hopkins University Press, Baltimore

McCullagh P, Nelder JA (1989) General linear models, 2nd edn. Chapman & Hall, London

Méndez J (2000) Modelos de comportamiento estacional e intradiurno de los pólenes y esporas de la ciudad de Ourense y su relación con los parámetros meteorológicos. Dissertation, University of Vigo

Moreno G (1990) Flora Ibérica. In: Castroviejo S. (ed) Real Jardín Botánico, vol. 2. C.S.I.C., Madrid

Moore PD, Webb JA (1978) An illustrated guide to pollen analysis. Hodder and Stoughton, London

Moseholm L, Weeke ER, Petersen BN (1987) Forecast of pollen concentration of Poaceae (grasses) in the air by time series analysis. Pollen Spores 2–3:305–322

Negrini AC, Voltolini S, Troise C, Arobba D (1992) Comparison between Urticaceae (*Parietaria*) pollen count and hay fever symptoms: assessment of a threshold value. Aerobiologia 8:325–329

Norris-Hill J, Emberlin J (1991) Diurnal variation of pollen concentration in the air of north-central London. Grana 30:229–234

Park J, Sandberg IW (1991) Universal approximation using radial basis function networks. Neural Comput 3:246–257

Rantio-Lehtimäki A, Pehkonen E, Yli Panula P (1996) Pollen allergic symptoms in the off season? Compostela Aerobiol 96:91–92

Ranzi A, Lauriola P, Marletto V Zinozi F (2000) Forecasting airborne pollen concentrations: development of local models. In: Abstracts of the European Symposium on Aerobiology, Vienna, 5–9 September 2000, p 43

Ripley BB (1996) Pattern recognition using neural networks. Cambridge University Press, Cambridge

Rodríguez-Rajo FJ (2000) El polen como fuente de contaminación ambiental. Dissertation, University of Vigo

Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev 65:386–408

Ruffaldi P Greffier F (1991) Birch (*Betula*) pollen incidence in France (1987–1990). Grana 30:248–254

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representations by error propagation. In: Rumelhart DE, McClelland JL (eds) Parallel distributed processing: explorations in the microstructure of cognition, vol 1. MIT Press, Cambridge, Mass., pp 318–362

Sánchez-Mesa JA, Galán C, Martínez-Heras JA, Hervás-Martínez C (2002) The use of a neural network to forecast daily grass pollen concentration in a Mediterranean region: the southern part of the Iberian Peninsula. Clin Exp Allergy 32:1606–1612

Spieksma FTM (1990) Pollinosis in Europe: new observations and developments. Rev Paleobot Palynol 64:35–40

Spieksma FTM, Frenguelli G, Nikkels AH, Mincigrucci G, Smithius LOMJ, Bricchi E, Dankaart W, Romano B (1989) Comparative study of airborne pollen concentrations in central Italy and The Netherlands (1982–1985). Grana 28:25–36

Spieksma FTM, Emberlin JC, Hjelmroos M, Jäger S, Leuschner RM (1995) Atmospheric birch (*Betula*) pollen in Europe: trends and fluctuations in annual quantities and the starting dates of the seasons. Grana 34:51–57

Wallin JE, Segerström V, Rosenhall L, Bergmann E, Hjelmroos M (1991) Allergic symptoms caused by long distance transported birch pollen. Grana 30:256–268

Wihl JA, Ipsen B, Nüchel PB, Munch EP, Janniche EP, Lövenstein H (1998) Immunotherapy with partially purified and standardized tree pollen axtracts. Allergy 43:363–369