ACTA AGROBOTANICA
Vol. 65 (4), 2012: 69–74
DOI: 10.5586/aa.2012.023

THE ADVANCED STATISTICAL METHODS IN AEROBIOLOGICAL STUDIES

Agnieszka Grinn-Gofroń, Beata Bosiacka

Department of Plant Taxonomy and Phytogeography, University of Szczecin, Wąska 13, 71-415 Szczecin, Poland
e-mail: agofr@univ.szczecin.pl

Received: 30.05.2012

Abstract

Pollen and spore forecasting has become an important aim in aerobiology. The main goal is to provide accurate information on biological particles in the air to sensitive users in order to help them optimize their treatment process.

Many statistical methods of data analysis are based on the assumptions of linearity and normality that often cannot be fulfilled. The advanced statistical methods can be applied to the problems that cannot be solved in any other effective way, and are suited to predicting the concentration of airborne pollen or spores in relation to weather conditions. The purpose of the study was to review some advanced statistical methods that can be used in aerobiological studies.

Key words: aeroplankton, pollen and fungal spores, forecasting, artificial neural network (ANN), multivariate regression tree (MRT), canonical correspondence analysis (CCA).

INTRODUCTION

Modern epidemiological studies from various countries indicate that currently 15–20% of the average population suffers from allergic diseases. Pollen grains and fungal spores are one of the main sources of inhalation allergens. The small sizes of these allergens allow for deep penetration of the bronchial tree, which often leads to allergic reactions of the lower respiratory tract. Most people hypersensitive to this group of allergens present year-round symptoms, with periods of seasonal exacerbations. It is estimated that sensitization to pollen and fungal allergens relates to the growing number of people. This percentage is higher in the child population compared with the adult population (Burge, 2002; Beggs, 2004).

The important direction of aerobiological studies is to seek correlations between the characteristics of the pollen or spore season and weather variables. Modeling of the concentration of airborne particles is a relatively difficult issue. Due to the complexity of the study object (a large number of analyzed variables, very irregular changes in the concentration of airborne pollen or fungal spores of a large variety of species, nonlinear correlations between parameters), multi-dimensional techniques and other advanced statistical methods of exploring data are preferred.

Currently developed forecasting models

Until now, some forecasting models for selected pollen grains that cause allergies have been developed (Ranzi et al. 2003; Castellano-Méndez et al. 2005; Sánchez-Mesa et al. 2004; Rodríguez-Rajo et al. 2010) and only a few for fungal spores. Most of them are characterized by a relatively low percentage of variability explained by the model (approximately 30%) and are based mostly on simple descriptive statistics, among others: Pearson's or Spearman's correlation coefficients, or Duncan's multiple range test, and on multiple regression model (Katial et al. 1997; Angulo-Romero et al. 1999; Mitakakis et al. 2001; Troutt and Levetin, 2001; Stennett and Beggs, 2004).

In addition, they do not provide the values of the weather factors which are responsible for causing the threshold concentrations of allergenic pollen or fungal spores and do not specify which of them is the most important. Currently, the development of aerobiological forecasting is moving towards the inclusion of air pollution in the models, apart from weather and biogeographic parameters (preliminary studies have been conducted in Szczecin). It is very important to analyze all these factors so that the resulting models can accurately describe the complex dependencies that occur in nature.

Multivariate regression trees

From a statistical point of view, relatively high spore and pollen concentrations during seasons, compared to long periods of absence, as well as the rapid appearance of spores and pollen at the beginning of seasons suggest that there are some threshold meteorological conditions above which spores or pollen occur. In order to reveal the cut-off values of the environmental predictors, the multivariate regression trees (MRT) will be used (Breiman et al. 1984; De'ath and Fabricus, 2000). This method makes no assumptions about the form of the relationships (e.g. unimodal or linear) between species and their environmental predictors. Moreover, this method can be applied to complex ecological data with imbalance, nonlinear relationships between variables and high-order interactions (De'ath and Fabricus, 2000). MRT models species–environment relationships and forms clusters by repeated splitting of the data, with each split chosen to minimize the dissimilarity (sum of squared Euclidean distances, SSD) within clusters (Breiman et al. 1984; De'ath and Fabricus, 2000). Clusters and their dependence on environmental parameters are presented graphically as a tree. Each cluster (branch) represents a characteristic species composition, further defined by a range of environmental parameters. The overall fit of a tree is specified as relative error (RE; SSD in clusters divided by SSD of undivided data), while the predictive accuracy is assessed by cross-validated relative error (CVRE) (Breiman et al. 1984; De'ath and Fabricus 2000).

Specific environmental conditions are identified by the indicator species index (IndVal) calculated as the product of the relative density and relative frequency of occurrence in a particular branch of a tree. Species characteristic of the particular cluster for which the index is indicative of > 0.25 are defined as indicator taxa.

Artificial neural networks (ANN)

A considerable number of correlated variables, which often non-linearly influence pollen and/or concentration in air, suggest the application of one of the most advanced methods of data analysis: artificial neural networks (ANN). An artificial neural network is a mathematical model or computational model that is inspired by the structure and/or functional aspects of biological neural networks (Tadeusiewicz, 1993; Haykin, 1999). The neural network is a group of interconnected neurons, which usually consists at least three layers. The first is called the input, and the final - output. All layers located between them are called hidden layers. The data is placed on the input neurons of the first layer and then, by the existing connections, the output values of the previous layer are transmitted to the inputs of the next layer. The results of calculations are the results obtained at the output of the last layer.

A neural network consists of an interconnected group of artificial neurons and it processes information using a connectionist approach to computation (Tadeusiewicz, 1993).

The determination of valid parameter values for a specific network architecture (weights of connections between artificial neurons) is called neural network learning. This is a necessary stage in the construction of the neural model. The neural network learning is carried out using appropriate algorithms based on data collected by the user, describing the course of the studied phenomenon. During calculations, an available data set is divided into three parts: training set – Tr (given on the network in the learning process); validation set – Ve (allows to monitor the network learning); and testing set – Te (used to conduct a final evaluation of the obtained model) (Tadeusiewicz, 1993).

In most cases, an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. ANNs are non-linear statistical data modelling tools. They are usually used to model complex relationships between inputs and outputs or to find patterns in data.

Knowledge and experience gained during statistical data analyses and the development of models can be further used to elaborate the methodology of advanced, predictive model construction. It will include guidelines concerning:

- preparation of data basis (amount of data, length of research period, etc.);
- selection of adequate methods of experimental data analysis;
- practical aspects of using advanced, multivariate statistical techniques;
- criteria for the assessment of submodel performance and their selection;
- verification and validation of the forecasting model.

Canonical correspondence analysis (CCA)

To identify the relationships between fungal spore genera or pollen taxa and environmental factors, canonical correspondence analysis (CCA) can be used with the CANOCO package v.4.5 (ter Braak and Šmilauer, 2002).

Canonical ordination techniques are designed to detect the patterns of variation in the fungal spore genera or pollen taxa data that can be explained by the variables like: meteorological parameters or air pollution. The resulting ordination diagram also expresses the main relations between the pollen or spore concentrations and each of the environmental variables.

The fungal spore genera or pollen taxa are positioned as points in the CCA diagram. The environmental variables are represented by arrows pointing in the direction of maximum variation, with their length proportional to the rate of changes, and can be interpreted in conjunction with the fungal spore genera or pollen taxa point as follows. Each arrow determines an axis in the diagram and the fungal spore genera or pollen taxa points must be projected onto this axis. These projection points estimate the position of the optimum for the distribution of each fungal spore genus or pollen taxon along each environmental variable. The allergens-environment correlation coefficients with the ordination axes, as well as the correlation between the environmental variables and the ordination axes were used to interpret the CCA results.

The relative importance of each environmental variable for the prediction of fungal spore genera or pollen taxa composition along the CCA axes can be inferred from the signs and relative magnitude of the correlation coefficients (this magnitude is related to the rate of changes in allergen composition). The statistical significance of environmental factors to elucidate the allergenic pollen or spore concentration variability can be judged by the stepwise forward selection and Monte Carlo permutation test. The factor that best explains the variability of the entire set of variables is first selected and the sequence of the other factors is determined based on their decreasing significance for the total variability of the set in connection with the variables previously selected. The measure of fit is the sum of all canonical eigenvalues with each variable as the only additional variable. The program reports “extra fit” (Lambda-A), which is the change in the sum of all canonical eigenvalues (additional variance) when the successive variable is added.

The advanced statistical models developed for the city of Szczecin (MRT, ANN, CCA)

Numerical analyses and forecasting statistical models for some allergenic pollen and the most important spore type were created in Szczecin, Poland.

In the multilayer neural perceptron for daily data for the *Alternaria* genus, the most important meteorological factors were as follows: dew point temperature and three parameters of air temperature: maximum, minimum and average. The most important factors in the regression model for spore seasons are as follows: relative air humidity and the three parameters of air temperature. All these factors influence most significantly the amount of spores and their concentration in the air during the day. The occurrence of high concentration of *Alternaria* spores is a result of the mutual influence of many factors. These are: the level of air humidity on the day of the occurrence of high concentration, maximum air temperature and maximum wind speed two days prior the noted concentrations as well as average air temperature and average wind speed one day before. The level of *Alternaria* spores in the air increased up to 70–80% of humidity level and above that level it slightly decreased. Below 20°C, the concentration of spores of this type was low and above this temperature a rising trend was noted. The neural model developed for the *Alternaria* genus was found to be the most effective predicting tool and was useful for forecasting concentrations of this genus for the city of Szczecin (Grinn-Gofroń and Strzelczak, 2008a).

For the *Cladosporium* genus, the final model included classification (spore presence or absence) and regression for spore seasons with log(x+1) transformed *Cladosporium* spore concentration and the multilayer neural perceptron model assumed, as the most important weather variables affecting the concentration of spores of this genus, the following parameters: dew point temperature, maximum temperature, and average wind speed. The regression model for seasons confirmed the importance of the dew point and pointed out relative humidity as well as maximum and minimum temperature as particularly important. The average wind speed was in the eighth place in the ranking of importance. Similarly to the model for the *Alternaria* genus, high concentrations of spores of this genus were influenced by many mutually correlated factors: the dew point temperature recorded one day before the occurrence of high concentration, relative air humidity and the maximum temperature on a given day and one day later. The concentrations of spores of this genus increased up to 60–70% of the humidity level and above this point it began to decline. A monotonic increase in concentration was recorded for dew point temperature and minimum temperature one day before the observed phenomenon and a drop in maximum temperature one day later. The designed neural model, like the one for the *Alternaria* genus, is the most effective and it is used to estimate the concentrations for the city of Szczecin (Grinn-Gofroń and Strzelczak, 2008b).

The neural model developed for hourly concentrations of both above mentioned genera indicated the exact values of significant weather factors that determine the fluctuations in the concentration of spores in the circadian cycle. The neural multi-layer perceptron developed for the *Alternaria* genus calculated that air pressure with a value of 1.011 hPa affects the reduction in spore concentration in the air and at the humidity level lower than 36.5% the concentration of *Alternaria* spores was the highest. Average air temperature, wind speed and dew point were in the next places in the ranking of importance developed by the perceptron. This is perhaps the most popular network architecture in use today, due originally to Rumelhart and McClelland (1986) and discussed at length in most neural network textbooks (Bishop, 1995). This is the type of network in which the units each perform a biased weighted sum of their inputs and pass this activation level through a transfer function to produce their output, and the units are arranged in a layered feed forward topology. The network thus has a simple interpretation as a form of input-output model, with the weights and thresholds (biases) being the free parameters of the model. Such networks can model functions of almost arbitrary complexity, with the number of layers and the number of units in each layer determining the function complexity. The concentration of spores increases with air pressure and temperature and decreases with increasing humidity. The model did not show prediction for a specific time during the day when the concentration of *Alternaria* spores reached the highest values. A similar model set the limit value of air pressure to 1.008 hPa for the *Cladosporium* genus and it considered time as the second most important factor (the largest concentrations occur in the afternoon hours between 12.00 and 17.00). Spore concentrations increase with cloud cover, wind speed, and air pressure and decrease with increasing dew point temperature. Both models are characterized by high testability and quality and can be successfully used to predict hourly concentrations of both genera in the circadian cycle (Grinn-Gofroń and Strzelczak, 2009).

The multivariate regression tree proved to be the best model for the *Ganoderma* genus. The factors determining high concentrations of spores in the air are dew point temperature with a value higher than 8.7ºC and average air temperature above 15.4ºC. Moreover, for this genus a neural multilayer perceptron was developed which recognized relative air humidity, apart from dew point temperature, as an important factor determining high concentrations. The results, derived from the perceptron, relating to the dew point appeared to be interesting. In temperatures below 11ºC, the model predicted the lack of spores in the air and their presence over the above-mentioned temperature. The models built for spores of the *Ganoderma* genus are so far the first models developed in the world. In comparison with the *Alternaria* and *Cladosporium* genera, they are characterized by lower verifiability. Because this genus belongs to Basidiomycotina, the spore release processes are influenced by additional factors (biological, ecological) that are more important than meteorological ones (Grinn-Gofroń and Strzelczak, 2011).

The neural network model of the relationship between *Betula* pollen and meteorological factors created in Poland (Puc, 2012) can be used at any moment of time and it provides a forecast for the whole season using only meteorological variables. The neural multilayer perceptron developed for *Betula* pollen concentrations indicated maximum temperature and humidity as the most important variables. The remaining variables: mean and maximum wind speed, daily precipitation, mean and minimum air temperature as well as dew point temperature, were of lesser importance, but all of them were statistically significant.

The results of canonical correspondence analysis showed that in the pollen taxa-environmental relations four meteorological variables were statistically significant: dew point temperature explained 18.8–21.7% of total variation, average air temperature – 17%, relative humidity – from 0.3 to 7.2%, and average wind speed – from 0.1 to 3.3% of total variation. According to intra-set correlations of meteorological standardized variables with the first two axes of CCA, the first axis was defined by temperature, and the second axis – by the relative humidity. The dispersion of pollen taxa in the CCA ordination diagram indicated some clusters. According to the variables that had the largest contribution to the total variation of pollen taxa composition and following the gradient of increasing temperature, the first cluster contained the following pollen taxa: Chenopodiaceae, *Artemisia* and *Urtica*, which reached their optima of high or moderate temperature; the second cluster: *Plantago*, *Tilia*, Poaceae, *Rumex*, *Secale* and *Pinus*, related to moderate temperature; the third cluster: *Aesculus*, *Platanus*, *Quercus*, *Rosa* and *Morus*, connected with moderate temperature; the fifth cluster: *Alnus*, *Corylus*, *Ulmus*, *Salix*, Cupressaceae, *Populus*, *Carpinus* and *Fraxinus*, related to the lowest temperature (Puc and Bosiacka, 2011).

All these above described methods are a much more effective tool in predicting concentrations of pollen grains and fungal spores and should be used in aerobiological studies.

REFERENCES

Angulo-Romero J., Mediavilla-Molina A., Dominquez-Vilches E. 1999. Conidia of *Alternaria* in the atmosphere of the city of Cordoba, Spain in relation to meteorological parameters. Int. J.Biometeorol.43;45–49.http://dx.doi.org/10.1007/s004840050115

Beggs P.J. 2004. Impacts of climate change on aeroallergens: past and future. Clin. Exp. Allergy, 34; 1507–1513. http://dx.doi.org/10.1111/j.1365-2222.2004.02061.x

Bishop C. 1995. Neural Networks for Pattern Recognition. Oxford: University Press

Breiman L., Friedma J.H., Olshen R.A., Stone C.G. 1984. Classification and regression trees. Wadsworth, Belmont, CA.

Burge H.A. 2002. An update on pollen and fungal spore aerobiology. J. Allergy Clin. Immunol. 110; 544–552. http://dx.doi.org/10.1067/mai.2002.128674

Castellano-Méndez M., Aira M.J., Iglesias I., Jato V., González-Manteiga W. 2005. Artificial neural networks as a useful tool to predict the risk level of *Betula* pollen in the air. Int. J. Biometeorol. 49; 310–316. http://dx.doi.org/10.1007/s00484-004-0247-x

De'ath G., Fabricus K.E. 2000. Classification and regression trees: a powerful and simple technique for ecological data analysis. Ecology, 81; 3178–3192. http://dx.doi.org/10.2307/177409

Grinn-Gofroń A., Strzelczak A. 2008a. Artificial neural network models of relationships between *Alternaria* spores and meteorological factors in Szczecin (Poland). Int. J. Biometeorol. 52; 859–868. http://dx.doi.org/10.1007/s00484-008-0182-3

Grinn-Gofroń A., Strzelczak A. 2008b. Artificial neural network models of relationships between *Cladosporium* spores and meteorological factors in Szczecin (Poland). Grana. 47; 304–314.

Grinn-Gofroń A., Strzelczak A. 2009. Hourly predictive ANN and MRT models of *Alternaria* and *Cladosporium* spore concentrations in Szczecin (Poland). Int. J. Biometeorol. 53; 555–562.

Grinn-Gofroń A., Strzelczak A. 2011. The effects of meteorological factors on the occurrence of *Ganoderma* sp. spores in the air. Int. J. Biometeorol., 55: 235–241. http://dx.doi.org/10.1007/s00484-010-0329-x

Haykin S. 1994. Neural networks: A comprehensive foundation. Macmillan, New York.

Katial R.K., Zhang Y.M., Jones R.H., Dyer P.D. 1997. Atmospheric mold spore counts in relation to meteorological parameters. Int. J. Biometeorol. 41; 17–22. http://dx.doi.org/10.1007/s004840050048

Mitakakis T.Z., Clift A., McGee P.A. 2001. The effect of local cropping activities and weather on the airborne concentration of allergenic of *Alternaria* spores in rural Australia. Grana, 40; 230–239. http://dx.doi.org/10.1080/001731301317223268

Puc M., Bosiacka B. 2011. Effects of meteorological factors and air pollution on urban pollen concentrations. Polish J. Environ. Stud. 20 (3); 611–618.

Puc M. 2012. Artificial neural network model of relationship between *Betula* pollen and meteorological factors in Szczecin (Poland). Int. J. Biometeorol. 56; 395–401. http://dx.doi.org/10.1007/s00484-011-0446-1

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

Rodríguez-Rajo F.J., Astray G., Ferreiro-Lage J.A., Aira M.J., Jato-Rodriguez M.V., Mejuto J.C. 2010. Evaluation of atmospheric Poaceae pollen concentration using a neural network applied to a coastal Atlantic climate region. Neural Networks, 23; 419–425. http://dx.doi.org/10.1016/j.neunet.2009.06.006

Rumelhart D.E. and McClelland J. (eds.). 1986. Parallel Distributed Processing, Vol 1. Cambridge, MA: MIT Press.

Sánchez-Mesa J.A., Galán C., César Hervás C. 2004. The use of discriminant analysis and neural networks to forecast the severity of the Poaceae pollen season in a region with a typical Mediterranean climate. Int. J. Biometeorol., 49; 366–352. http://dx.doi.org/10.1007/s00484-005-0260-8

Stennett P.J., Beggs P.J. 2004. *Alternaria* spores in the atmosphere of Sydney, Australia, and relationships with meteorological factors. Int. J. Biometeorol. 49; 98–105. http://dx.doi.org/10.1007/s00484-004-0217-3

Tadeusiewicz R. 1993. Neural networks. Akademicka Oficyna Wydawnicza, Warszawa, Poland.

ter Braak., Šmilauer. 2002. CANOCO Reference Manual and User's Guide to Canoco for Windows: Software for Canonical Community Ordination (version 4.5). Microcomputer Power; Ithaca, NY, USA.

Troutt C., Levetin E. 2001. Correlation of spring spore concentrations and meteorological conditions in Tulsa, Oklahoma. Int. J. Biometeorol. 45; 64–74. http://dx.doi.org/10.1007/s004840100087

Zaawansowane metody statystyczne stosowane w badaniach aerobiologicznych.

Streszczenie

Ziarna pyłku i zarodniki grzybów należące do aeroplanktonu stanowią ważny przedmiot badań aerobiologicznych. Głównym celem takich badań jest dostarczenie dokładnych informacji na temat zawartości biologicznych cząsteczek w powietrzu aby pomóc zoptymalizować proces leczenia alergii i astmy u osób wrażliwych. Wiele metod statystycznej analizy danych opartych jest na założeniach liniowości i normalności, często założenia te nie mogą być spełnione. Zaawansowane metody statystyczne mogą być stosowane wtedy, kiedy inne metody nie mogą być użyte. Dodatkowo niektóre z nich konstruują modele prognostyczne, które przewidują skuteczne stężenie w powietrzu pyłku lub zarodników w zależności od warunków atmosferycznych. Celem badania było dokonanie przeglądu zaawansowanych statystycznych metod, które mogą być stosowane w badaniach aerobiologicznych.