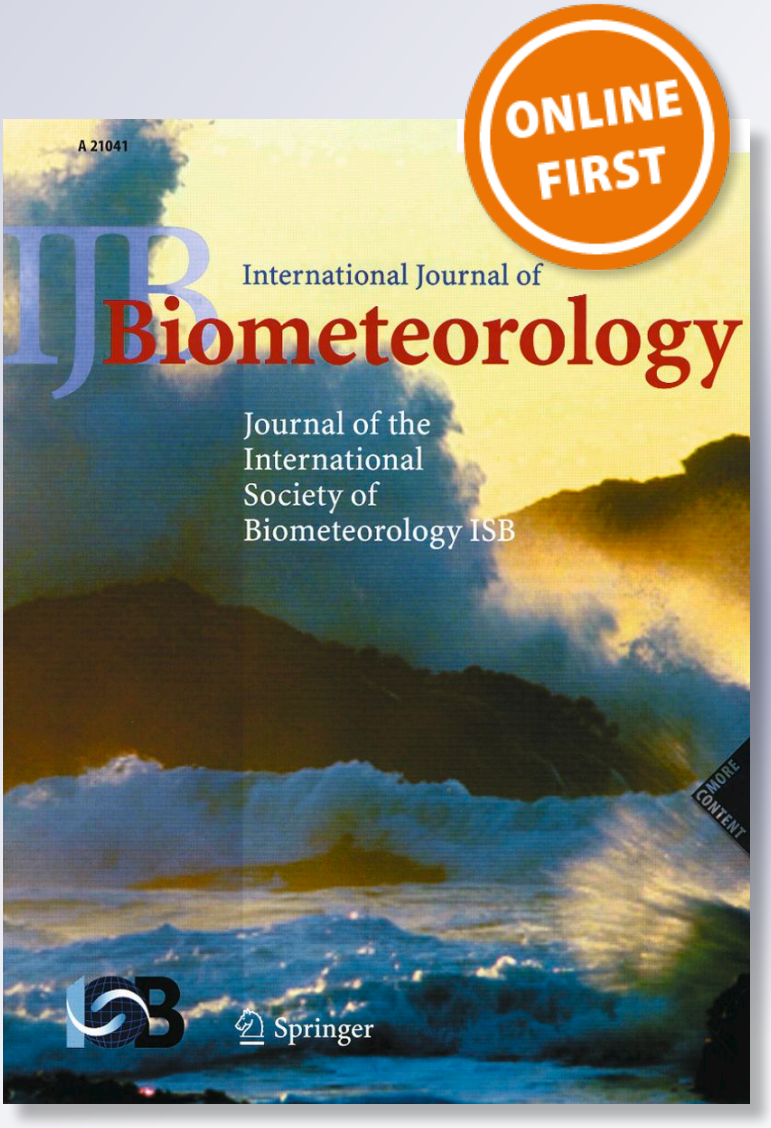
Year clustering analysis for modelling olive flowering phenology

J. Oteros, H. García-Mozo, C. Hervás-Martínez & C. Galán

International Journal of Biometeorology

ISSN 0020-7128

Int J Biometeorol
DOI 10.1007/s00484-012-0581-3



Springer

Int J Biometeorol
DOI 10.1007/s00484-012-0581-3

ORIGINAL PAPER

Year clustering analysis for modelling olive flowering phenology

J. Oteros · H. García-Mozo · C. Hervás-Martínez · C. Galán

Received: 12 April 2012 / Revised: 3 July 2012 / Accepted: 14 July 2012

Abstract It is now widely accepted that weather conditions occurring several months prior to the onset of flowering have a major influence on various aspects of olive reproductive phenology, including flowering intensity. Given the variable characteristics of the Mediterranean climate, we analyse its influence on the registered variations in olive flowering intensity in southern Spain, and relate them to previous climatic parameters using a year-clustering approach, as a first step towards an olive flowering phenology model adapted to different year categories. Phenological data from Cordoba province (Southern Spain) for a 30-year period (1982–2011) were analysed. Meteorological and phenological data were first subjected to both hierarchical and "K-means" clustering analysis, which yielded four year-categories. For this classification purpose, three different models were tested: (1) discriminant analysis; (2) decision-tree analysis; and (3) neural network analysis. Comparison of the results showed that the neural-networks model was the most effective, classifying four different year categories with clearly distinct weather features. Flowering-intensity models were constructed for each year category using the partial least squares regression method. These category-specific models proved to be more effective than general models. They are better suited to the variability of the Mediterranean climate, due to the different response of plants to the same environmental stimuli depending on the previous weather conditions in any given year. The present detailed analysis of the influence of weather patterns of different years on olive phenology will help us to understand the short-term effects of climate change on olive crop in the Mediterranean area that is highly affected by it.

Keywords Olive · Phenology · Aerobiology · Forecasting model · Clustering · Climate change

J. Oteros (✉) · H. García-Mozo · C. Galán
Department of Botany, Ecology and Plant Physiology, University of Córdoba, Agrifood Campus of International Excellence (CeiA3), 14071 Cordoba, Spain
e-mail: b42otmoj@uco.es

C. Hervás-Martínez
Computing and Numerical Analysis Department, University of Córdoba, Agrifood Campus of International Excellence (CeiA3), 14071 Cordoba, Spain

Introduction

This study was focussed on the southern Spanish region of Andalusia, which has by far the world's largest area given over to olive plantations (1,511,687 ha) and an annual olive output generally exceeding 5,000,000 t. Within this region, Cordoba province has the second-largest olive-growing area, with 343,812 ha producing an average of 1,000,000 t olive crop (Andalusia Statistical Yearbook 2010).

The olive is a temperate, spring-flowering tree. The bioclimatic requirements for flowering vary as a function of the tree's phenological status (Galán et al. 2001b; Aguilera and Ruiz 2009). Reproductive structures grow from undifferentiated buds within a few months of dormancy, this phase is termed differentiation. But differentiation must occur after induction: certain temperature requirements need to be met in order for bud break to take place; these may vary depending on the olive variety and on the degree of climate adaptation (Galán et al. 2005; García-Mozo et al. 2009). Although it is generally accepted that an initial induction occurs during summer months, a stress period during winter (induced by low temperatures) is also required to break bud dormancy (Andreini et al. 2008; Fernandez-Escobar et al. 1992; Orlandi et al. 2004; Rallo and Martin 1991). Flowering starts once a certain amount of heat has been accumulated but cold spells during endodormancy have been found to favour increased inflorescence formation, while the lack of cold leads to further development of branches

Published online: 11 August 2012

(Bonofiglio et al. 2009; Orlandi et al. 2005; IOOC 1996). In general, the Mediterranean climate is characterised by hot, dry summers and mild, rainy winters. However, it is also characterised by a marked year-on-year variations in weather patterns, due mostly to the alternation of wet and dry year periods.

Because olive trees are anemophilous, flowering intensity was measured by means of airborne pollen detection. The amount of pollen detected by air samplers is considered to be direct proportional to the flowering intensity in anemophilous species that are within the effective area of the pollen sampler (Frenguelli 1998; Subba-Reddi and Reddi 1985). However, this assumption should be treated with some caution since, before being caught by the air sampler, the pollen has had to surmount the release process and transport to the pollen sampler, processes highly dependent on environmental conditions (García-Mozo 2011; Frenguelli 1998; Subba-Reddi and Reddi 1985). Weather conditions influence the flowering season, not only by regulating flower and pollen production but also, in the case of wind-pollinated plants, by regulating the timing of pollen season and therefore shaping pollen-curve dynamics. Olive is also characterised by exhibiting a high tendency toward variable flowering intensity that is highly related to the alternation in fruit production (Galán et al. 2004, 2008; Lavee 2006). Earlier research in the study area showed that olive flowering intensity was influenced mainly by temperature and rainfall over the preceding months (Galán et al. 2001a). Olive flowering may also vary, even within the study area, as a function of plant physiological status, which is itself governed by a number of factors including genetic background, internal reserves and local weather conditions (García-Mozo et al. 2009). The present study sought to test the hypothesis that the differentiation of year categories as a function of phenological and meteorological features, prior to the construction of phenological flowering-intensity models, would help to render these models more accurate. The different categories would serve as indicators of physiological status, as well as providing value information on reproductive phenology. Clustering analysis enabled different weather-related variables to be included in each category. It was assumed that year-clustering would provide a better fit for phenological models. For this purpose, the following three-step method was developed: (1) a clustering step using a K-means method, (2) a classification step using an artificial neural network (ANN) method, and (3) a forecasting step using the partial least squares regression (PLSR) method. Since olive is wind-pollinated, it was possible to precisely measure flowering intensity as a pollen index (PI) that records the sum of daily airborne pollen data throughout the entire pollination season (Galán et al. 2007).

In an earlier paper, we constructed some indices covering the main variables affecting olive flowering intensity in Cordoba: the thermal index (TI), the pre-flowering hydric index (PFHI), and the cyclicity index (CI) (Oteros et al. 2012). These were constructed using weather-related and phenological data, and various adjustment criteria were tested to ensure that models reflected the extreme weather events characteristic of the Mediterranean climate. These indices were used for clustering purposes in the present study. A number of authors have made use of clustering techniques in phenological and aerobiological research, e.g. for pollen back-trajectory analysis in the atmosphere (Hernández-Ceballos et al. 2011), and for daily airborne pollen forecasting (Makra et al. 2006; Sánchez-Mesa et al. 2002). The year-clustering approach has also been used in comparing classification methods and for general long-term forecasting, although not hitherto for specific, long-range forecasting purposes (Sánchez-Mesa et al. 2005). In the present study, each proposed year category would represent a distinct physiological situation, resulting from exposure of olive trees to different environmental conditions. With a view to selecting the model providing the best classification of future years in terms of weather-related features, several classification methods were tested. Fitted models to predict olive flowering intensity constructed for each year category were compared to a general model including all the years of the studied period.

This study should improve our knowledge of olive reproductive behaviour in a Mediterranean climate, i.e. in a region specially affected by climate change (IPCC 2007). Plant phenology is seen as one of the most important bio-indicators of climate change, since trends can provide considerable temporal and spatial information regarding ongoing changes (Galán et al. 2005). Analysis of the influence of weather patterns on olive phenology will help us to understand the short-term effects of climate change on olive crops. Also, advance information regarding olive-pollen intensity could be of particular value in a number of fields. Olive pollen is highly allergenic in the Mediterranean area (Dominguez et al. 1993; D'Amato et al. 2007; Barber et al. 2008), and patients with atopic or allergic asthma could be forewarned regarding likely pollen peaks. Pollen counts also provide a valuable bioindicator of flowering intensity and therefore of the volume of the forthcoming harvest, and are thus of value to farmers (Galán et al. 2004, 2008; García-Mozo et al. 2008; Ribeiro et al. 2008).

The first aim of this study was to develop a method for forecasting olive pollen season intensity by clustering years on the basis of phenological and meteorological data. A second objective was to compare the accuracy between different classification methods trying to improve the methodology.

Materials and methods

Study area and phenological data

The city of Cordoba is situated in south-west Iberian Peninsula (37°50′N, 4°45′W), 123 m a.s.l. The area has a Mediterranean climate with some continental features. The annual mean temperature is 17.8 °C and the annual average rainfall is 621 mm. Phenological data for the last 30 years (1982–2011) regarding flowering intensity were measured by analysing airborne pollen using a Hirst-type volumetric spore trap (Hirst 1952) placed on the roof of the Educational Sciences Faculty, at 15 m above ground level, which offers olive pollen data from a radius of 100 km around, well representing Cordoba province olive flowering phenology (Hernández-Ceballos et al. 2011). Pollen counts were obtained using a standard protocol published by the Spanish Aerobiology Network (REA) (Galán et al. 2007). Weather data for the last 30 years (1982–2011), provided by the Spanish State Meteorology Agency (AEMET), were taken at Cordoba Airport, located around 5 km south of the pollen-sampling site.

Year clustering

The first step was to assign the study years into clusters, on the basis of phenological and weather data. Phenological features of the intensity of flowering season were based on the variations of PI and average daily pollen counts during the pre-peak period of the pollen season (ADPP). Three biometeorological indices, proposed in an earlier study (Oteros et al. 2012), aimed at identifying the factors most influencing pollen season characteristics: TI (1), which takes into account the heat conditions in January and March by adjustment criteria; PFHI (2), which takes into account rainfall in February and March (both indices covering the heat and water conditions most affecting olive flowering intensity in Cordoba); and finally CI (3, 4), an autoregressive index that seeks to emulate cyclic changes in the pollen-index time-series. The development of the CI is summarized in two steps (3, 4), where CI_2 is the final CI.

$$TI = (T\min M)/(RanTM) \tag{1}$$

(TminM) is minimum March temperatures and (RanTM) is March temperature range.

$$PFHI = (RfF) + (RfM) \tag{2}$$

(RfF) is the sum of total rainfall in February and (RfM) is the total rainfall in March.

$$CI_1 = \begin{cases} AC + AV, \; if\,(PI_{n-1}) < \left(\frac{PI_{n-2}}{1.25}\right) + (PI_{n-2}) \\ \frac{AC+AV}{2}, \; if\,(PI_{n-1}) \geq \left(\frac{PI_{n-2}}{1.25}\right) + (PI_{n-2}) \end{cases} \tag{3}$$

(AC) is the number of years elapsing since the last appreciable crest and (AV) is the number of years elapsing since the last appreciable valley. The term "crest" was used to denote the change from a rising to a falling trend in the PI time series, while "valley" referred to the change from a falling to a rising trend (Oteros et al. 2012).

$$CI_2 = \begin{cases} \left(\frac{PI_{n-1}}{4.8}\right) + PI_{n-1}, \; if\,(CI_1) - (CI_{1_{n-1}}) > 0 \\ PI_{n-1}, \; if\,(CI_1) - (CI_{1_{n-1}}) = 0 \\ \left(\frac{PI_{n-1}}{1.8}\right) - PI_{n-1}, \; if\,(CI_1) - (CI_{1_{n-1}}) < 0 \end{cases} \tag{4}$$

Clustering analysis

First, the optimum number of natural groups of years (K) was determined by hierarchical clustering analysis; clusters were then generated by "K-means" conglomerate analysis. Both methods were developed using the software Unscrambler 9.7 (http://www.camo.com).

Hierarchical clustering analysis was performed using Ward's method, in which information is quantified as the sum of squared distances of each element with respect to the centroid of the cluster to which it belongs. To do this, we first calculated the mean vector of all variables, the multivariate centroid for each cluster. Next, we calculated the squared Euclidean distances between each element and the centroid (mean vector) of all clusters. Finally, distances for all elements were combined.

The "K-means" conglomerate method was used for cluster generation: "k" groups of years were generated on the basis of similar meteorological and phenological characteristics. The five variables used were standardised before performing cluster analysis, in order to remove any dependency on measurement units. Of the various types of cluster analysis available, this was deemed to the most appropriate, in that it provides a more flexible approach and does not assume any specific distribution of variables.

Characterisation of year categories

To characterise each category, the centroid of each cluster was analysed; weather and pollen-count patterns were examined for each of the years making up the cluster. The four year categories were designated C1–C4. A descriptive statistic analysis of several features of the years was performed using the software R 2.11.1. We analyzed the averages and standard deviations of meteorological parameters and aerobiological features of years.

Classification modelling

Models were constructed to classify other years on the basis of the weather conditions recorded prior to the pollen season, using the indices TI, PFHI and CI. Classification models were constructed using three different techniques: (1) linear discriminant analysis, (2) decision trees and (3) artificial neural networks.

The models were validated using the k-fold cross-validation test. K-fold cross-validation is similar to simple cross-validation but the original sample is partitioned randomly into k subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k−1 subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The optimal k used was 4 and that is the most appropriate k number considering that we work with four categories.

All classification models were built using the software Weka 3.7.4 (http://weka.softpedia.com/).

Linear discriminant analysis

The goal of linear discriminant analysis (LDA) is to establish a linear discriminant function, based on the original variables, which will discriminate between different classes. LDA seeks a linear combination or function, D, of the independent variables that maximizes the between-class variance relative to the within class variance. The latent variable thus obtained is called a canonical variate. For k classes, k−1 canonical functions can be calculated. Each time, LDA will select a direction leading to maximum discrimination between the given categories (Fisher 1936).

Decision trees

Decision tree (DT) building is a machine-learning method adapted for classification and prediction. DT-based methods express their results as a graphical presentation of decision rules. DT consists of a root, a number of internal nodes representing attributes and a sequence of branches representing attribute values. The tree ends in leaves reflect the appropriate target attribute and indicating a class. The descending order of the attribute is calculated on the basis of the gain ratio. The attribute with the highest information gain is selected for generating the root of the tree (Kirchner et al. 2004). The algorithm selected in our work as the most accurate was the termed LMT-I-1-M 15-W 0.0 using Weka 3.7.4.

Artificial neural network

In this subsection we consider standard sigmoidal ANN, or multilayer perceptron (MLP), as the base classification model. ANN can overcome the longer training time and the difficulty in determining hidden layer units of a backpropagation network to a large extent. An ANN is a three-layer feed-forward neural network. For determining the best ANN model, we apply an backpropagation algorithm to find the basis functions or nodes of the hidden layer: $B(x,W)=\{B_1(x,w_1), B_2(x,w_2), \ldots, B_m(x,w_m)\}$, corresponding to the nonlinear part of the discriminant functions, $f_l(x,\theta_l)$. We have to determine the number of basis functions "m" and the weight matrix "w". To apply evolutionary neural network techniques, we consider an ANN with softmax outputs defined in Eq. (5) and the standard structure: an input layer with a node for every input variable; a hidden layer with several sigmoidal nodes; and an output layer with J-1 nodes, where J is the number of classes.

If the output layer of the $ANN\left(\widehat{\theta}\right)$ classifier is interpreted from the point of view of probability, which considers the softmax activation function:

$$g_1(x,\theta_l)=\frac{\exp f_l(x,\theta_l)}{\sum_{l=1}^{Q}\exp f_l(x,\theta_l)} for\ l=1,\ldots,Q \tag{5}$$

where $g_l(x,\theta_l)$ is the probability a pattern x has of belonging to class l, $\theta_1=(\beta_1,w_1,\ldots,w_M)$, where $\beta_1=\left(\beta_1^1,\ldots,\beta_M^1\right)$ is the l-th vector of weights of a node in the hidden layer to the l output node, M is the number of hidden nodes, $w_j=\left(w_0^j,\ldots,w_k^j\right)$ for j=1.....M, is the vector of weights of the input layer to the hidden node j, and $f_l(x,\theta_1)$ is the output of the l-th output node for pattern x given by:

$$f_l(x,\theta_l)=\beta_0^l+\sum_{j=1}^{M}\beta_j^l\beta_j(x,w_j) for\ l=1,\ldots Q-1 \tag{6}$$

$$f_Q(x,\theta_Q)=0 \tag{7}$$

where $B_j(x,w_j)=\sigma_j\left(w_0^j+\sum_{i=1}^{k}w_i^jx_i\right)$ is the sigmoidal activation function of the nodes in the hidden layer.

The classification rule coincides with the optimal Bayes rule. In this way, by g classifiers, the classification rule assigns an individual to the class with the maximum probability, given vector measurement x: $C(x)=\widehat{l}$, where $\widehat{l}=\arg\max_l g_l\left(x,\widehat{\theta}_l\right)$, for l=1,…,Q. Detailed descriptions of various forms of neural networks are provided elsewhere (Bishop 1995; Haykin 1994; Gutiérrez et al. 2009; Martinez-Estudillo et al. 2006).

Results of classifications models (the confusion matrices and the percentages of correctly classified years) were compared in order to select the most effective method.

Modelling

To analyse the performance of clustering analysis, different models were tested for each year category. The performance of these models (significance level, standard error and RMSE) was compared with that of other forecasting models not involving conglomerate analysis. RMSE was obtained by the following expression (8), where Y = observed data and F = expected data:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{N}(Y_t - F_t)^2}{N}} \quad (8)$$

Forecasting models were constructed using the PLSR technique, taking PI, i.e. the annual sum of daily pollen counts, as dependent variable. Fitted predictive models were thus generated for each of the four categories (designated C1M to C4M), and a general model (GM) was constructed using all years. The models were built using the software Unscrambler 9.7.

Modelling was based on a linear transformation of the original descriptors to a small number of orthogonal factors (latent variables), attempting to maximize the covariance between the descriptors and the dependent variable; this procedure provides the optimal linear model in terms of forecasting. Here, each latent variable represented a key factor for olive flowering intensity.

Results

Clustering

To obtain the number of classes to be considered in the "K-means" algorithm, we applied hierarchical clustering, yielding the optimal number of groups that should be considered for "k-means" clustering equal to 4.

Once the optimal number of classes was defined, we applied a "4-means" clustering to group the study years into four clusters as a function of pollen-counts and weather-related characteristics. Categories were defined on the basis of meteorological characteristics and pollen parameters considered indicative of olive physiological and phenological status. PI is related directly to both types of parameters.

Centroid characteristics for each cluster are shown in Table 1; years belonging to a given category shared similar characteristics. The variables were bioclimatic indices, duly standardised in order to remove any dependency on measurement units. Figure 1 shows the relationship between years clustered into different categories and PI. In Fig. 1b we can see the structure of the dendrogram resulting from hierarchical cluster analysis.

Table 1 Centroids of each cluster. *CI* Cyclicity index, *PFHI* pre-flowering hydric index, *TI* thermal index, *PI* pollen index, *ADPP* average daily pollen counts during the pre-peak period of the pollen season

	Conglomerate			
	1	2	3	4
CI	−0.68	−0.36	0.57	1.52
PFHI	−0.53	0.47	−0.46	1.36
TI	−0.59	−0.07	0.27	1.29
PI	−0.83	−0.15	0.49	1.75
ADPP	−0.74	0.11	0.26	1.48

Category characterisation

Results for the descriptive analysis of years assigned to the various categories are shown in Table 2. Each category was defined in terms of the main features of the pollen season: the PI, pollen peak (PPk), pollen-season start-date (StD) and number of days with more than 100 grains (>100). Two major bioindices influencing the pollen season were included: PFHI and TI.

Weather-related variables for the months from January to May refer to the current pollen season, while variables for summer (June–September) and autumn (September–December) months refer to the previous pollen season, since these have a greater effect on flowering characteristics.

Category 1: Dry years

The years assigned to cluster 1 displayed lower PIs (55 % of the average), low daily pollen counts and low pollen peaks (Table 2). In terms of weather conditions, they were characterised by mild summers: maximum, mean and minimum temperatures from June to September were below the average for Cordoba. Autumns were dry and had a wider-than-average temperature range (12.5 ° C vs. average 11.9 °C). Rainfall between September and January in Category 1 (307 mm) was also lower than average for Cordoba (370 mm). The period from January to April was cold and dry, minimum temperatures being lower than average: January 0.8 °C (average 3.8 °C), March 0.7 °C (average 7.2 °C). May temperatures were 0.6 °C below average (27.1 °C). Rainfall from February to April was 25 % below average.

Weather data for 1982 lay at the shortest distance from the centroid of the cluster, signifying that they were the most representative for Category 1.

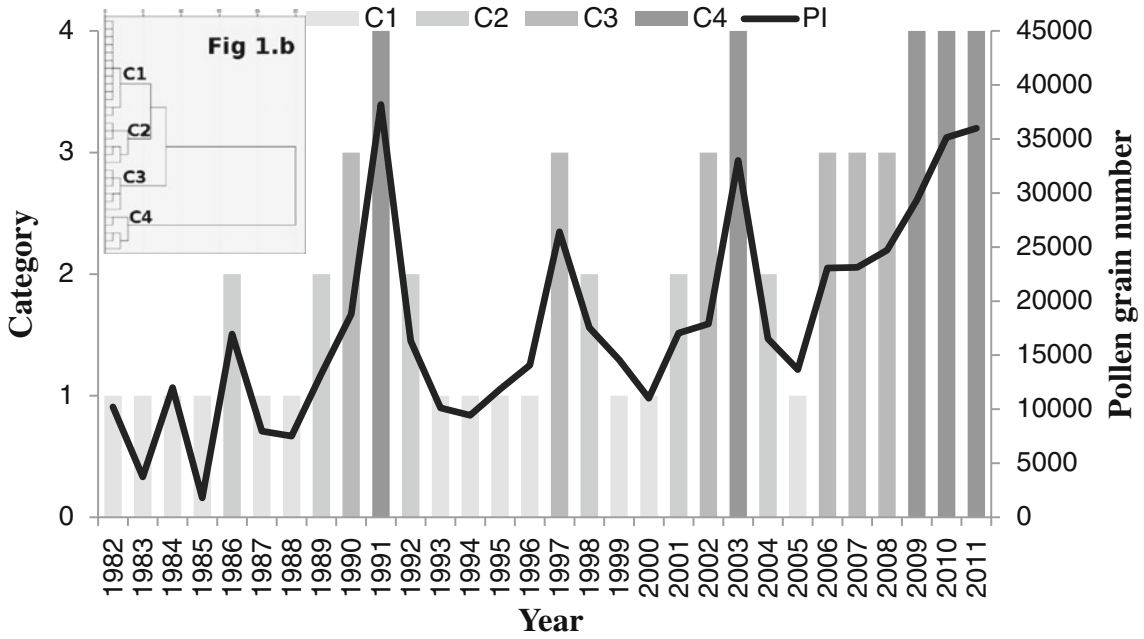


Fig. 1 Relationship between years clustered into different categories and pollen index (PI). Category 1 (C1), Category 2 (C2), Category 3 (C3), Category 4 (C4). *Inset* shows the dendrogram resulting from hierarchical clustering

Category 2: Cold years

Category 2 displayed average PIs and daily pollen counts. Like C1, however, the number of days with over 100 pollen grains was low. Summer temperatures were below the average for Cordoba. Maximum temperatures from October to December were 1 °C below average, while average temperatures were recorded in January and February; April temperatures were 1.1 °C below average. Rainfall during this period was within the average range for the area.

Category 3: Warm years

PIs in Category 3 years were 25 % above average, while daily pollen counts were average; however, the number of days with pollen counts exceeding 100 pollen grains/m^3 of air was higher than average. In Category 3 years, flowering started 5 days earlier than average. Summer and autumn temperatures were average, while rainfall from September to January was 80 mm above average. Temperatures from January to April were higher than the average for Cordoba: maximum temperatures from January to March were around 1 °C above average. Rainfall during this period was within the average range for the area.

Category 4: Wet years

Category 4 was characterised by very high PIs (193 % of local average), and high daily pollen counts. The pollen peak was higher than in the other categories. The pollen season was longer and the number of days with pollen counts exceeding 100 pollen grains/m^3 of air was greater.

Table 2 Descriptive analysis of each category (C1–C4). Pollen season characteristics: *PI* Pollen index, *PPk* pollen peak, *StD* starting date, *>100* number of days with more than 100 grains. Some important biometeorological indices influencing pollen season were included: *PFHI* pre-flowering hydric index, *TI* thermal index

	C1		C2		C3		C4		Total	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
PI	9,841	3,832	16,277	1,526	22,332	3,338	34,343	3,356	17,710	9,500
PPk	1,512	1,146	2,292	1,099	2,702	854	3,901	2,040	2,304	1,485
StD	115	7	114	13	108	13	116	9	113	10
>100	25	11	26	7	32	5	31	7	27	9
PFHI	67	32	140	53	72	55	205	87	106	73
TI	0.37	0.11	0.45	0.14	0.54	0.16	0.72	0.10	0.48	0.18

Summers were characterised by extreme temperatures: maximum, mean temperatures and minimum temperatures were 0.6 °C above average.

Autumns and winters were mild and rainy: the temperature range from September to January was 0.9 °C below average. Rainfall was higher than average over this period. The temperature range from January to April was 0.8 °C below average. Rainfall from February to April was 80 mm higher than the average for Cordoba.

Classifications

Three different methods were used to classify years within a given class: (1) discriminant analysis; (2) decision trees; and (3) neural networks.

Three variables were used for to test each analysed method: TI, PHI and CI. The three models were validated by 4-fold cross-validation. A total of 73.3 % of cases were classified correctly by LDA, a total of 76.6 % of cases were correctly classified by the DT method and a total of 80 % of cases were correctly classified by the ANN method. The neural network model thus proved to be the most effective of the three models tested. Table 3 shows the confusions matrices resulting from the classification methods tested.

Forecasting modelling

Different PLSR models were constructed to predict PI. A model was generated for each year category (C1M, C2M, C3M and C4M), and one GM was constructed for all study years.

A summary of statistical parameters, including R^2 and RMSE for the training model and cross-validation for the PI model, is shown in Table 4. As the results indicate, differentiation into several categories could be a useful tool for forecasting modelling.

The relationship between observed PI and the PI predicted by each PLSR model is shown in Fig. 2, which shows the PI predicted by the GM and the PI predicted by each specific category model.

The major coefficients for each model are shown in Fig. 3. The influence exerted by the weather-related variables examined here varied among categories. In overall terms, coefficients show that the general model of conditions prompting high PI included hot summers, rainy autumns, cold winters and mild, rainy springs. Category 1 years, characterised by drought, cold and low flowering intensity, were dependent largely on March rainfall and did not require lower winter temperatures. Category C2 years—with lower-than-average temperatures and medium-to-low flowering intensity—also had no need for very low winter temperatures. In C3 years, characterised by warm temperatures and middle-high flowering intensity, the high minimum temperatures in March had a negative effect. Coefficients for C4 years, which had warm temperatures, heavier rainfall and high flowering intensity, were similar to those recorded for the GM.

Discussion

This study sought to identify the weather-related parameters most influencing olive flowering intensity, expressed as PI, in the Mediterranean climate. Although in anemophilous plants PI is a good indicator of flowering intensity, this statement should be taken with some caution since before the pollen can be caught by an air sampler, it has had to surmount the release process and be transported to the pollen sampler—processes highly dependent on environmental conditions (García-Mozo 2011; Frenguelli 1998; Subba-Reddi and Reddi 1985). Nevertheless PI offers the possibility of obtaining a quantitative value for a wide area of study that can be compared objectively year by year.

A knowledge of factors most affecting flowering variations is of particular value for agricultural and environmental studies, and also for allergy sufferers, in the Mediterranean area due to its variable climatic characteristics. A great deal of research has attempted to determine the factors affecting phenology and flowering intensity in plant species, placing special attention on the effect of climate change (e.g. Linkosalo et al. 2010;

Table 3 Confusion matrices. Results of several classification methods: *LDA* Linear discriminant analysis, *DT* decision trees, *ANN* artificial neural network

		LDA				DT				ANN				Total
		Predicted category				Predicted category				Predicted category				
		1	2	3	4	1	2	3	4	1	2	3	4	
Observed Category	1	10	0	2	1	12	1	0	0	10	2	0	0	13
	2	0	4	0	1	4	2	0	0	3	3	0	0	5
	3	1	0	3	2	1	0	5	0	0	0	6	0	6
	4	1	0	0	5	0	0	1	4	0	0	1	4	6

Table 4 Pollen index (PI) partial least squares regression (PLSR) model summary. *GM* General model, *C_1M* category 1 model, *C_2M* category 2 model, *C_3M* category 3 model, *C_4M* category 4 model

Model	Training		Full cross-validation	
	R^2	RMSE	R^2	RMSE
GM	0.63	5,652	0.53	6,597
C_1M	0.86	1,391	0.53	2,741
C_2M	0.94	328	0.19	1,499
C_3M	0.97	498	0.84	1,464
C_4M	0.99	289	0.64	2,251

Morton et al. 2011; García-Mozo et al. 2010). Also some studies have focussed specifically on the olive tree, seeking in most cases to predict flowering onset and pollination intensity using linear regression techniques (Galán et al. 2001a; Ribeiro et al. 2006a)

Only a few authors (e.g. Reynolds et al. 2002; Yu et al. 2010) have applied PLSR to phenological research, but this technique has never been used in aerobiological studies. The most novel aspect of this study was that it differentiated between four different year categories using a year clustering approach prior to constructing the regression model. Clustering has been used previously in phenological and aerobiological research, not with the aim of predicting specific features of flowering intensity, but rather for general short- or long-term forecasting purposes (Sánchez-Mesa et al. 2002, 2005). The ANN model provided the most accurate classification of years into groups. Other authors have compared the effectiveness of discriminant analysis and neuronal network analysis for classifying years with different pollen intensities, and to make other predictions regarding pollen season characteristics. (Aznarte et al. 2007; Kasprzyk et al. 2011; Voukantsis et al. 2010; Rodríguez-Rajo et al. 2010; Sánchez-Mesa et al. 2005; Puc 2012).

The physiological response of plants to the same environmental stimuli may differ depending on the weather conditions in any given year (Galán et al. 2001b). Year-clustering took into account both weather-related variables and phenological variables to determine the potential physiological status of the olives in the study area. Four year-categories were generated, each comprising years sharing similar weather conditions, giving rise to similar phenological characteristics. Specific models to predict flowering intensity were generated for each year category. Though minor differences were apparent in each category model (CM), analysis of inter-model coefficients of variation revealed more marked differences.

The general model (GM), constructed using data for all 30 study years, showed that rainfall in February and March exerted a major influence on flowering intensity, expressed as PI; similar findings have been reported by other authors (Recio et al. 1996; Galán et al. 2001a). The GM also indicated that low rainfall and higher minimum temperatures in summer led to increased flowering intensity. The amount of pollen available for spring flowering is influenced by weather conditions over the previous summer, by which time the mother cells designated to become pollen grains are already present. When the summer is characterised by abnormally high temperatures and low rainfall, high pollen counts tend to be recorded during the following spring (Mandrioli 1987). In the olive, the rate of flowering-bud differentiation is affected strongly by the prevailing weather conditions over the previous summer (Rallo and Cuevas 2004), which could prompt a higher rate of fruit abortion, which would increase floral induction (Dag et al. 2010).

The major influence of March–April temperatures has also been reported in other research carried out in the Mediterranean area. A number of authors have found that mild weather, i.e. narrow temperature ranges and high minimum temperatures are associated with greater flowering intensity

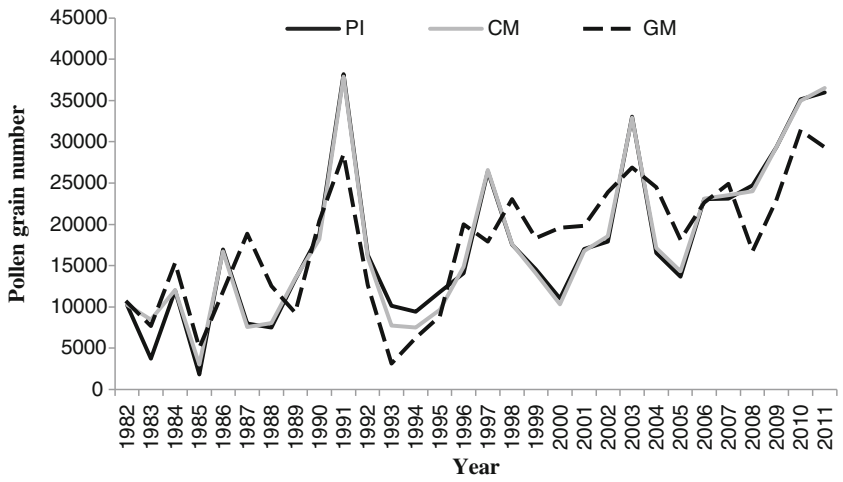


Fig. 2 Observed pollen index (*PI*), and pollen indices predicted by category model (*CM*) and general model (*GM*)

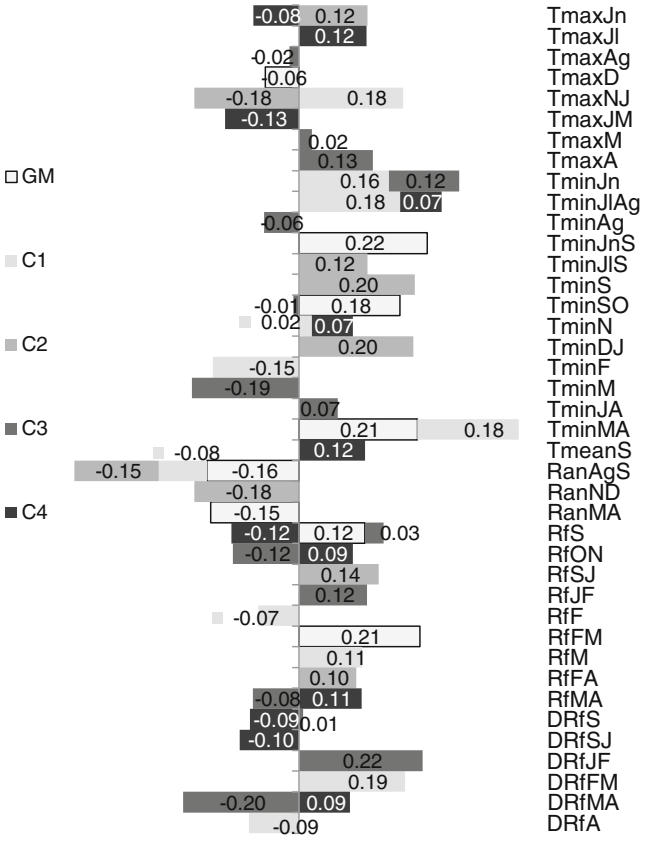


Fig. 3 Coefficients of important variables. The variables were defined prior to building the models—the coefficients show the relative importance of each variable on each model. Model variables: *Tmax* Average value of daily max temperatures; *Tmin* average value of daily minimum temperatures; *Tmean* average values of daily minimum and max temperatures; *Ran* average values of temperature range; *Rf* sum of total rainfall, *DRf* number of days with rainfall above 1 mm, both in individual months (single letters indicating month, e.g. *J* January, *F* February, *M* March…) and in different periods: *JnS* 1 June–30 September, *JlAg* 1 July–31 August, *JlS* 1 July–30 September, *SO* 1 September–31 October, *SJ* 1 September–31 January, *ON* 1 October–30 November, *ND* 1 November–31 December, *NJ* 1 November–31 January, *DJ* 1 December–31 January, *JF* 1 January–28 February, *JM* 1 January–31 March, *JA* 1 January–30 April, *FM* 1 February–31 March, *FA* 1 February–30 April, *MA* 1 March–30 April

(Galán et al. 2001a). The GM displayed a strongly negative coefficient for minimum temperatures in September and October, probably because this early cold impairs bud development. However, lower temperatures in winter, and particularly in December, are reported to favour flowering (Ribeiro et al. 2006b; Orlandi et al. 2010). Analysis of the optimal weather conditions for olive flowering suggests that these tend to occur in the Mediterranean climate, which would account for the high degree of adaptation of the olive in the study area.

Cluster analysis showed that C1 years were marked by dry springs and winters, with cooler temperatures throughout the year; flowering intensity was very poor. Temperature ranges in the summer have a negative effect on flowering intensity, possibly because in summers with below-average mean temperatures, a higher temperature range means that the minimum temperatures are even lower.

C2 years were also characterised by cold winters and springs, but did not display the low rainfall typical of C1 years. Flowering intensity was also poor. In the C2 years, winter temperatures had a more intense effect than in general: minimum temperatures in December to January had a more positive effect, while the wide temperature range in November and December exerted a negative influence; this behaviour can be understood in a context of cold years.

C3 years were marked by warm summers, winters and springs. Pollen counts were above the average for Cordoba, and the pollen season started earlier than in other year categories. Precipitation was lower than average in C3 years, and this could explain why rainfall in January and February exerted a positive effect on the flowering intensity, whereas rainfall from October to November and from March to April showed a negative effect. On the other hand, an influence of March temperatures was noted; maximum temperatures exerted little influence, while high minimum temperatures in March had a significant negative effect, perhaps because they were particularly high in C3 years.

C4 comprised years with hot summers and mild autumns, and heavy rainfall in both seasons. Weather conditions were most conducive to high flowering intensity, as evident in high PIs, high pollen peaks and high daily pollen counts. The category-specific model indicated that summer temperatures exerted a positive effect on PI, perhaps because only extremely high summer temperatures can produce a significant fruit abortion that can affect the next spring flowering. Spring and autumn rainfall also had a positive influence, although in the context of very wet years; September rainfall did not have any positive influence.

The biological responses of plants to climatic variables are rather complex. As stated above, the same weather conditions can act differently on different plants depending on their physiological status but plants are also highly influenced by adaptation to geographical features and local climates (García-Mozo et al. 2009). The models obtained in this study are derived from an empirical approach in a specific Mediterranean area and for these reasons should be applied to different regions with extreme caution. Although the methodology developed in the present work could be considered applicable in other areas, the proposed models were built for a specific region, and therefore are not strictly transferable to different locations.

In overall terms, specific cluster-based CMs proved more effective than the general model. Category differentiation enhanced the effectiveness of phenological modelling for two main reasons:

1. Weather conditions act differently on each year category, since plant physiological status at any given time varies from one category to another.

2. The years comprising each category share similar flowering intensity characteristics, so forecasting models are subject to less error.

These category-specific models proved to be more effective than general models, and therefore better suited to the variability of the Mediterranean climate, probably because plants respond differently to the same environmental stimuli depending on the weather conditions in any given year. Moreover, analysis of the influence of weather patterns on olive phenology will help us to understand the short-term effects of climate change on olive crop in the Mediterranean area that is highly affected by it.

Conclusions

Predictive models obtained using clustering analysis are more effective than general models because the years comprising each category share similar aerobiological characteristics and because weather conditions can act differently on flowering intensity depending on previous meteorological context that could have been decisive for configuring the physiological status of plant.

PLSR proved valuable for generating phenological forecasting models.

The autoregressive and biometeorological indices used to take into account the effects of extreme weather events yielded optimal results in the construction of an effective classification model. ANNs proved a useful tool for effective classification.

We corroborated the finding that summer weather conditions play a major role in olive flowering intensity.

Acknowledgements The authors are grateful to the European Social Fund for co-financing with the Spanish Science Ministry the "Ramón y Cajal" contract of G.M. Authors are grateful to the Spanish Inter-Ministerial Commission of Science and Technology (MICYT), FEDER funds, for funding the TIN2011-22794 and CGL2011-24146 projects. The authors also thank the Andalusia Regional Government for funding projects P10-RNM5958 and P08-TIC-3745. Finally, the authors appreciate the availability of meteorological data from the Spanish Meteorological Agency (AEMET) and from the Andalusian Government Agroclimatic Information Network (RIA).

References

Aguilera F, Ruiz Valenzuela L (2009) Study of the floral phenology of *Olea europaea* L. in Jaén province (SE Spain) and its relation with pollen emission. Aerobiologia 25(4):217–225

Andalusia Statistical Yearbook (2010) http://www.juntadeandalucia.es/institutodeestadisticaycartografia/anuario/anuario10/index.htm

Andreini L, Bartolini S, Guivarc'h A, Chriqui D, Vitagliano C (2008) Histological and immunohistochemical studies on flower induction in the olive tree (*Olea europaea L.*). Plant Biol 10:588–595

Aznarte JL, Benítez JM, Nieto D, de Linares C, Díaz de la Guardia F (2007) Forecasting airborne pollen concentration time series with neural and neuro-fuzzy models. Expert Syst Appl 32:1218–1225

Barber D, de la Torre F, Feo F, Florido F, Guardia P, Moreno C, Quiralte J, Lombardero M, Villalba M, Salcedo G, Rodríguez R (2008) Understanding patient sensitization profiles in complex pollen areas a molecular epidemiological study. Allergy 63:1550–1558

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

Bonofiglio T, Orlandi F, Sgromo C, Romano B, Fornaciari M (2009) Evidences of olive pollination date variations in relation to spring temperature trends. Aerobiologia 25:227–237

IPCC (2007) Synthesis Report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Pachauri RK, Reisinger A (eds). IPCC, Geneva, Switzerland

D'Amato G, Cecchi L, Bonini S, Nunes C, Annesi-Maesano I, Behrendt H, Liccardi G, Popov T, van Cauwenberge P (2007) Allergenic pollen and pollen allergy in Europe. Allergy 62:976–990

Dag A, Bustan A, Avni A, Tzipri I, Lavee S, Riov J (2010) Timing of fruit removal affects concurrent vegetative growth and subsequent return bloom and yield in olive (*Olea europaea* L.). Sci Hortic 123:469–472

Dominguez-Vilches E, García-Pantaleón F, Galán C, Guerra F, Villamandos F (1993) Variations in the concentrations of airborne *Olea* pollen and associated polinosis in Cordoba (Spain): a study of the 10-year period 1982–199. J Invest Allergol Clin Immunol 3(3):121–129

Fernández-Escobar R, Benlloch M, Navarro C, Martín GC (1992) The time of floral induction in olive. J Am Soc Hortic Sci 117(2):304–307

Fisher RA (1936) The use of multiple measurements in taxonomic problems. Ann Eugenics 7:179–188

Frenguelli G (1998) The contribution of aerobiology to agriculture. Aerobiologia 14:95–100

Galán C, Cariñanos P, García-Mozo H, Alcázar P, Domínguez-Vilches E (2001a) Model for forecasting *Olea europaea* L. airborne pollen in South-West Andalusia, Spain. Int J Biometeorol 45:59–63

Galán C, García-Mozo H, Cariñanos P, Alcázar P, Domínguez-Vilches E (2001b) The role of temperature in the onset of the *Olea europaea L.* pollen season in southwestern Spain. Int J Biometeorol 45:8–12

Galán C, Vázquez L, García-Mozo H, Domínguez E (2004) Forecasting olive (*Olea europaea*) crop yield based on pollen emission. Field Crop Res 86:43–51

Galán C, García-Mozo H, Vázquez L, Ruíz L, Díaz de la Guardia C, Trigo M (2005) Heat requirement for the onset of the *Olea europaea L.* pollen season in several sites in Andalusia and the effect of the expected future climate change. Int J Biometeorol 49:184–188

Galán C, Cariñanos P, Alcázar P, Domínguez-Vilches E (2007) Spanish Aerobiology Network (REA): Management and Quality Manual. Servicio de publicaciones de la Universidad de Cordoba, Cordoba

Galán C, García-Mozo H, Vázquez L, Ruiz L, Díaz de la Guardia C, Dominguez E (2008) Modeling olive crop yield in Andalusia, Spain. Agron J 100:98–104

García-Mozo H (2011) The use of aerobiological data on agronomical studies. Ann Agric Environ Med 18:159–164

García-Mozo H, Chuine I, Perez-Badía R, Galán C (2008) Aerobiological and meteorological factors' influence on olive (*Olea europaea L.*) crop yield in Castilla-La Mancha (Central Spain). Aerobiologia 24:13–18

García-Mozo H, Orlandi F, Galán C, Fornaciari M, Romano B, Ruiz L, Diaz de la Guardia C, Trigo MM, Chuine I (2009) Olive flowering phenology variation between different cultivars in Spain and Italy: modelling analysis. Theor Appl Climatol 95:385–395

García-Mozo H, Mestre A, Galán C (2010) Phenological trends in southern Spain: a response to climate change. Agric For Meteorol 150:575–580

Gutiérrez PA, Hervás C, Carbonero M, Fernández JC (2009) Combined projection and kernel basis functions for classification in evolutionary neural networks. Neurocomputing 72(13–15):2731–2742

Haykin S (1994) Neural networks. A comprehensive foundation. MacMillan College, New York

Hernández-Ceballos M, García-Mozo H, Adame JA, Domínguez-Vilches E, Bolívar JP, De la Morena BA, Pérez-Badía R, Galán C (2011) Determination of potencial sources of *Quercus* airborne pollen in Cordoba city (southern Spain) using back-trajectory analysis. Aerobiologia 27:261–276

Hirst JM (1952) An automatic volumetric spore trap. Ann Appl Biol 39:257–265

IOOC(1996) Biology and physiology of the olive. World Olive Encyclopedia, International Olive Oil Council (IOOC), Madrid, Spain

Kasprzyk I, Grinn-Gofrón A, Strzelczak A, Wolski T (2011) Hourly predictive artificial neural network and multivariate regression tress models of *Ganoderma* spore concentrations in Rzeszów and Szczecin (Poland). Sci Total Environ 409:949–956

Kirchner K, Tölle KH, Krieter J (2004) The analysis of simulated sow herd datasets using decision tree technique. Comput Electron Agric 42:111–127

Lavee S (2006) Biennial bearing in olive (*Olea europaea L.*). Olea FAO olive Netw 25:5–13

Linkosalo T, Ranta H, Oksanen A, Siljamo P, Luomajoki A, Kukkonen J, Sofiev M (2010) A double-threshold temperatures sum model for predicting the flowering duration and relative intensity of *Betula pendula* and *B. pubescens*. Agric For Meteorol 150:579–1584

Makra L, Juhász M, Mika J, Bartzokas A, Béczi R, Sümeghy Z (2006) An objective classification system of air mass types for Szeged, Hungary, with special attention to plant pollen levels. Int J Biometeorol 50:403–421

Mandrioli P (1987) Biometeorology and its relation to pollen count. Adv Aerobiol 51:37–41

Martínez-Estudillo AC, Hervás-Martínez C, Martinez-Estudillo FJ, García-Pedrajas N (2006) Hybridization of evolutionary algorithms and local search by means of a clustering method. IEEE T Syst Man Cy B 36(3):534–546

Morton J, Bye J, Pezza A, Newbigin E (2011) On the causes of variability in amounts of airborne grass pollen in Melbourne, Australia. Int J Biometeorol 55:613–622

Orlandi F, Garcia-Mozo H, Vazquez-Ezquerra L, Romano B, Dominguez E, Galan C, Fornaciari M (2004) Phenological olive chilling requirements in Umbria (Italy) and Andalusia (Spain). Plant Biosys 138:111–116

Orlandi F, Vazquez L, Ruga L, Bonofiglio T, Fornaciari M, Garcia-Mozo H, Domínguez E, Romano B, Galán C (2005) Bioclimatic requirements for olive flowering in two Mediterranean regions located at the same latitude (Andalucia, Spain, and Sicily, Italy). Ann Agric Environ Med 12:47–52

Orlandi F, García-Mozo H, Galán C, Romano B, Díaz de la Guardia C, Ruíz L, Trigo MM, Domínguez-Vilches E, Fornaciari M (2010) Olive flowering trends in a large Mediterranean area (Italy and Spain). Int J Biometeorol 54:151–16

Oteros J, García-Mozo H, Hervás C, Galán C (2012) Biometeorological and autoregressive indices for predicting olive pollen intensity. Int J Biometeorol. doi:10.1007/s00484-012-0555-5

Puc M (2012) Artificial neural network model of the relationship between *Betula* pollen and meteorological factors in Szczecin (Poland). Int J Biometeorol 56:395–401

Rallo L, Cuevas J (2004) Fructificación y producción, capítulo 5, in: Barranco D, Fernández-Escobar D, Rallo L, El Cultivo del Olivo. 5ª Ed. Madrid, España. Junta de Andalucía y Ediciones Mundi-Prensa, pp 159–183

Rallo L, Martin GC (1991) The role of chilling in releasing olive floral buds from dormancy. J Am Soc Hortic Sci 116(6):1058–1062

Recio M, Cabezudo B, Trigo M, Toro F (1996) *Olea europaea* pollen in the atmosphere of Málaga (S.Spain) and its relationship with meteorological parameters. Grana 35:308–313

Reynolds MP, Thethowan R, Crossa J, Vargas M, Sayre KD (2002) Physiological factors associated with genotype by environment interaction in wheat. Field Crop Res 75:139–160

Ribeiro H, Santos L, Abreu I, Cunha M (2006a) Influence of meteorological parameters on *Olea* flowering date and airborne pollen concentration in four regions of Portugal. Grana 45:115–121

Ribeiro H, Cunha M, Abreu I (2006b) Comparison of classical models for evaluating the heat requirements of olive (*Olea europaea L.*) in Portugal. J Integr Plant Biol 48(6):664–671

Ribeiro H, Cunha M, Abreu I (2008) Quantitative forecasting of olive yield in Northern Portugal using a bioclimatic model. Aerobiologia 24:141–150

Rodríguez-Rajo FJ, Astray G, Ferreiro-Lage JA, Aira MJ, Jato-Rodriguez MV, Mejuto JC (2010) Evoluation of atmospheric Poaceae pollen concentration using a neural network applied to a coastal Atlantic climate region. Neural Netw 23:419–425

Sánchez-Mesa JA, Galán C, Martínez-Heras JA, Hervás-Martínez C (2002) The use of 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

Sánchez-Mesa JA, Galán C, Hervás C (2005) The use of discriminant analysis and neural networks to forecast with a typical Mediterranean climate. Int J Biometeorol 49:355–362

Subba-Reddi C, Reddi NS (1985) Relation of pollen release to pollen concentrations in air. Grana 24:109–113

Voukantsis D, Karatzas K, Damialis A, Vokou D (2010) Forecasting airborne pollen concentration of Poaceae (grass) and Oleaceae (olive), using artificial neural networks and genetic algorithms, in Thessaloniki, Greece. The International Joint Conference on Neural Networks (IJCNN), pp 1–6

Yu H, Eike L, Xu J (2010) Winter and spring warming result in delayed spring phenology on the Tibetan Plateau. Proc Natl Acad Sci USA 107(51):22151–22156