

# Building models for daily pollen concentrations

## The example of 16 pollen taxa in 14 Swiss monitoring stations

Denis Hilaire · Mathias W. Rotach · Bernard Clot

Received: 18 February 2011 / Accepted: 8 February 2012  
© Springer Science+Business Media B.V. 2012

**Abstract** We describe a method for constructing prediction models for daily pollen concentrations of several pollen taxa in different measurement sites in Switzerland. The method relies on daily pollen concentration time series that were measured with Hirst samplers. Each prediction is based on the weather conditions observed near the pollen measurement site. For each prediction model, we do model assessment with a test data set spanning several years.

**Keywords** Aerobiology · Aeroallergen · Poisson regression · Data preprocessing · Boosting · Predictive modeling

### 1 Introduction

The ambient air contains a large number of biological particles called bioaerosols. Those particles can be

entire organisms, parts of organisms or substances produced by living organism. Some of them, such as spores and pollen, serve for reproduction or dissemination purposes. They are subjected to gravity but due to their size and density, air currents play a large role in their passive dispersion. Their concentration in the air can reach several thousands of particles per cubic meter.

The prediction of the concentration of bioaerosols and of their effects on other living organisms or ecosystems is an important task in aerobiology. Pollen forecasting abilities in particular are needed in health care for treating allergies and in other areas like agriculture, where the dispersion of genetically modified plants is an important topic. Since the 1950s biologists and physicians have therefore begun to monitor the levels of pollen taxa in the ambient air with volumetric systems. At the beginning of the 1990s, coordinated measurement networks developed in Europe. That was also the case for Switzerland where, on January 1, 1993, the pollen measurement network initiated by the Swiss Working Group in Aerobiology was integrated into MeteoSwiss, the Federal Office for Meteorology and Climatology.

The first monitoring site in Switzerland came into service in the late 1960s. The network has steadily grown in the 1970s and 1980s to include 14 measurement stations covering the different bioclimatological regions of the Swiss territory. At each of these stations, the daily average concentrations for several pollen taxa are recorded and later transferred to a data-warehouse that also contains the other parameters

---

D. Hilaire · M. W. Rotach · B. Clot (✉)  
Federal Office of Meteorology and Climatology  
MeteoSwiss, Chemin de l'Aérologie, 1530 Payerne,  
Switzerland  
e-mail: bernard.clot@meteoswiss.ch

D. Hilaire  
e-mail: denis.hilaire@meteoswiss.ch

M. W. Rotach  
e-mail: mathias.rotach@uibk.ac.at

M. W. Rotach  
Institute of Meteorology and Geophysics, University  
of Innsbruck, Innrain 52, 6020 Innsbruck, Austria

collected by the MeteoSwiss weather measurement and observation networks. Because of the measurement design plan (in many cases no measurement in part of autumn/winter) and interruptions in the measurement due to either technical failures or human errors, the data collected contain its share of missing values.

The initial goal of the present study had been to make model predictions for the missing values in the pollen concentration time series. Such completed time series can then be used instead of the original ones when a particular statistical technique fails because of the missing values. This is, for example, the case if summary statistics of the yearly time series are computed. To get rid of all the missing values in a pollen time series, also those that occur during a seasonal peak, it is necessary to build regression models for the daily pollen concentration based on other parameters. Such a model is needed for every combination of taxon and measurement station. As weather strongly influences pollen emission and dispersion (e.g., Isard and Gage 2001), weather parameters are well suited for that purpose.

These same regression models have also the potential to be used for prediction when used with weather forecasts. Clearly, they will only be capable of providing forecasts for the very site they had been trained for.

Several previous attempts to build statistical models for individual pollen taxa have been documented, like Bringfelt et al. (1982), Arizmendi et al. (1993), Norris-Hill (1995), Stark et al. (1997), Galán et al. (2001), Ranzi et al. (2003), Cotos-Yáñez et al. (2004), Makra et al. (2004), Castellano-Méndez et al. (2005), Smith and Emberlin (2006), Stach et al. (2008) and Voukantsis et al. (2010). The estimation and assessment on test samples of several dozens of pollen regression models taken together has, however, not been addressed before. This text describes how a large amount of such pollen regression models, using weather parameters as input, can be obtained and evaluated.

## 2 Data

### 2.1 Aerobiological data

In each of the 14 measurement stations, a volumetric spore trap (Hirst 1952), and the method described by Mandrioli et al. (1998), is used to collect the pollen

from the ambient air. Its pump aspirates 10 l of air per minute through an opening measuring  $14 \times 2$  mm. Behind this entry slot, a rotating drum with a plastic strip coated with silicone as adhesive is located. Once every week, the drum makes an entire rotation.

Inside the device, the pollen and other organic and inorganic particles that come through the orifice will stick on the part of the strip that is exposed at that time. The position on the band on which a given particle was deposited reflects the time of the day and of the week in which the particle went through the apparatus. In the laboratory, the strip is cut into seven separate pieces, one for each day of the week. The strips are then used to determine the daily pollen counts for the days of the week. For each day, pollen grains are identified and counted under the microscope.

The pollen data collected consisted of several pollen concentration time series. There is one series per pollen taxon for every 14 pollen measurement stations of Switzerland that are in service today. The names and the labels of all 14 measurement stations are given in Table 1; detailed description of the stations is given in Peeters et al. (1998).

The values in the time series are estimated daily pollen concentrations of 56 pollen taxa that are based on the occurrence of each species counted on a given number of lines on a microscope slide that comes from a spore trap. Each time the pollen count is multiplied by an appropriate factor to give the concentration

**Table 1** The 14 pollen measurement stations and their labels together with the corresponding reference weather stations

Pollen measurement stations		Reference weather stations
Basel	PBS	Basel-Binningen
Bern	PBE	Bern-Zollikofen
Buchs	PBU	Vaduz
La Chaux-de-Fonds	PCF	La Chaux-de-Fonds
Davos	PDS	Davos
Geneva	PGE	Geneva
Lausanne	PLS	Pully
Locarno	PLO	Locarno-Monti
Lugano	PLU	Lugano
Luzern	PLZ	Luzern
Münsterlingen	PMU	Güttingen
Neuchâtel	PNE	Neuchâtel
Visp	PVI	Visp
Zürich	PZH	Zürich

[pollen grains/m<sup>3</sup>]. This factor depends on the microscope settings used, on the air volume that went through the apparatus, on the number of lines read and on the size of the collection surface. A single slide represents the pollen count for a 24-h period starting at 08:00 in the morning.

The concentration data obtained with the procedure described above stem from a counting process. Thus, many statistical methods that are designed for continuous dependent variables need to be adapted by using some sort of data transformation. The sampling error for the estimated pollen concentration on the slide varies with the number of lines used for the counting. Since most of the time, due to time and budget restrictions, only two lines were read, the estimated count error can represent a substantial amount of the observed value. Values of 30% (of the real pollen total per slide) for the estimated count error are therefore quite common in pollen count data obtained with the used slide reading protocol. Comtois et al. (1999) made a more detailed treatise on that issue. In particular, these authors found that the relative error decreases with increasing pollen concentration and also decreases with the area of slide examined.

Among the available taxa, 10 tree taxa and 6 herb taxa were further considered for this analysis. They were chosen for their abundance on the Swiss territory and for their allergenic properties. The 16 species are listed in Table 2.

Due to the data collection procedure, the newest data available are generally between 1 and 10 days old.

## 2.2 Missing values

Depending on the pollen station considered, the available pollen concentration time series had durations that ranged from 13 to 41 years. For all except one station, the measurements were only taken during the time frames that were expected to cover the natural pollen seasons. Thus, outside those time frames, the series contain long chains of missing values that in most cases can safely be assumed to be zero.

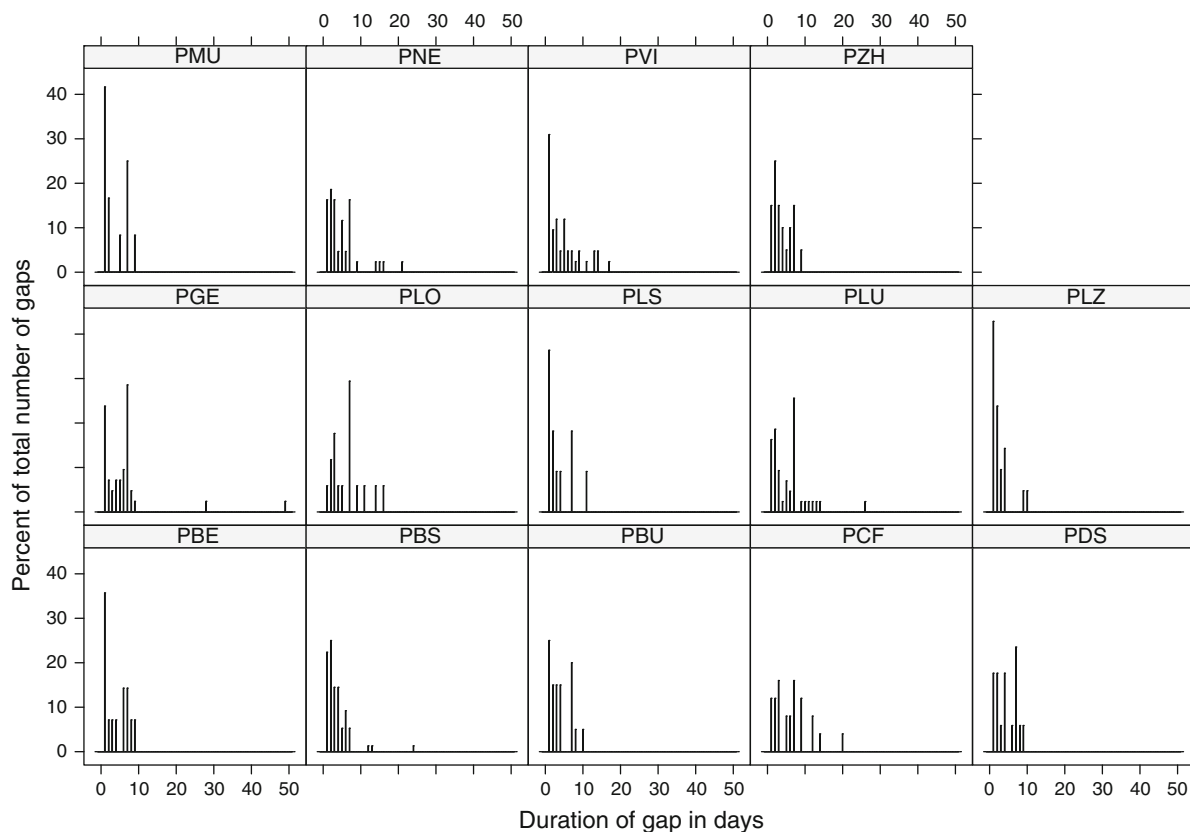
The Hirst apparatus used to collect the data and the human process involved in gathering that data are not 100% reliable. As a consequence, missing values for the pollen count can also appear during the pollen season. In that case, the data should be missing for all taxa at the station under consideration. The reason is that for every taxa taken together, the microscope slide is either available for counting or not.

**Table 2** The 10 tree pollen taxa and 6 herb pollen taxa used in the study

Taxa	
English	Latin
Trees	
Alder	<i>Alnus</i>
Birch	<i>Betula</i>
Hornbeam	<i>Carpinus</i>
Chestnut	<i>Castanea</i>
Hazel	<i>Corylus</i>
Beech	<i>Fagus</i>
Ash	<i>Fraxinus</i>
Plane	<i>Platanus</i>
Poplar	<i>Populus</i>
Oak	<i>Quercus</i>
Herbs	
Mugwort	<i>Artemisia</i>
Plantain	<i>Plantago</i>
Grasses	Poaceae
Sorrel, dock	<i>Rumex</i>
Nettle family	Urticaceae
Ragweed	<i>Ambrosia</i>

A histogram of the durations of the gaps occurring during the measurement time frames is given in Fig. 1. The plot shows that for durations <50 days—the gaps located between the seasons being not considered—values between 1 and 10 days are the most frequent. Thus, most intra-seasonal missing data points cannot be imputed with a conventional parametric prediction model using, for example, the pollen concentration of the previous day or of the day afterward without having to use the model several times on already altered data. The reason is that with parametric models a prediction cannot be made if there are missing values in the predictors.

Figure 2 reveals that some of the missing data also appear during the measurement time frame of the respective stations. The blank spaces between the solid chunks indicate the time frames in which no data are available. Most blank spaces visible at the zoom factor used in Fig. 2 are either due to the fact that they are outside the measurement time frame or that the measurement station did not yet exist. The longer inter-seasonal blank spaces in the data spanning the complete study time frame can be seen in Fig. 3. Among the 14 stations, Basel (PBS) has the longest time series available and Lausanne (PLS) the shortest.



**Fig. 1** Relative frequency histogram for all measurement sites of the duration of the gaps in the pollen time series in which no measures were available. Only missing periods shorter than

50 days were considered for the plot. The time frame considered goes from January 1, 1969, to December 31, 2009

In the 1990s, it was decided that in Geneva measures should be carried out throughout the year. The data collected there were then used to assess whether seasons may start earlier or later in a given year.

### 2.3 Weather data

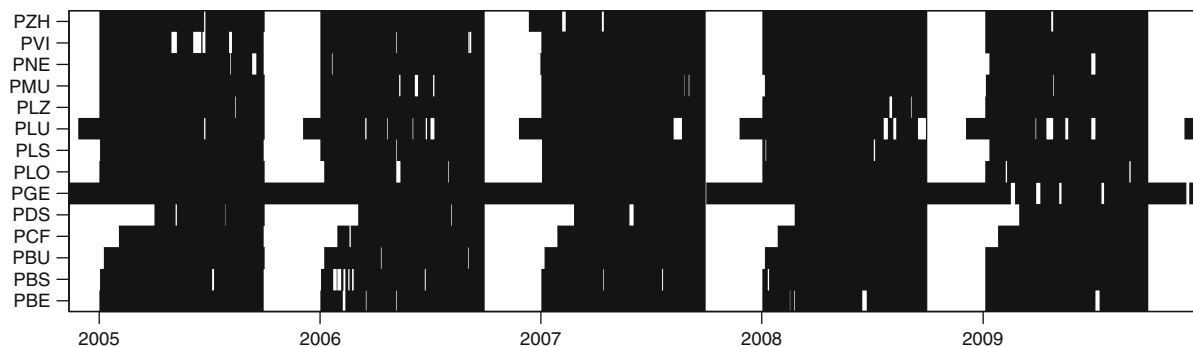
Besides seasonal aspects, the weather conditions play an important role in modulating airborne pollen concentrations and are therefore used when experts make pollen outburst forecasts. Typically, pollen particles are much less abundant in the air when the air humidity is too high or in cases of rainfall. Furthermore, other weather parameters, like temperature and wind direction and speed, are known to be important (Gregory 1961; Cox and Wathes 1995; Isard and Gage 2001). Weather parameters are usually not

measured at the pollen monitoring sites. Therefore, data of corresponding meteorological stations, located near the pollen monitoring sites, were used.

The reference weather stations used are also listed in Table 1. Each weather station provides data for several weather parameters that are stored in the MeteoSwiss data warehouse with a time resolution of up to one value every 10 min. The data warehouse also contains aggregated values from several weather parameters on an hourly and a daily basis. For the regression models, weather data with a resolution of 1 h were used.

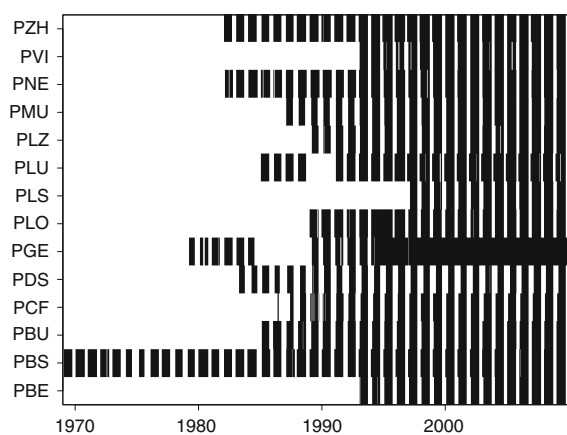
## 3 Materials and methods

Before building any model, it is important to gauge whether the missing data mechanism distorts the



**Fig. 2** Data availability from 2005 to 2009 for the pollen concentration time series. Missing values are given in *white*, while measured values are *dark*. We see that for the time series in

Geneva (PGE), measures were taken continuously. The function used to plot the series is based on a routine provided by Peng (2008)



**Fig. 3** Data availability over the entire time period for the pollen concentration time series taken from the 14 pollen measurement stations located on the Swiss territory. Missing values are given in *white*, while measured values are *dark*. The function used to plot the series is again based on code provided by Peng (2008)

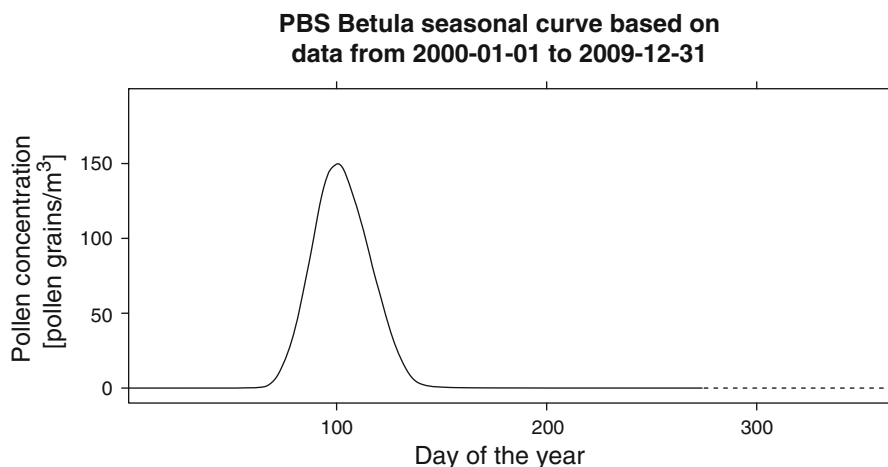
observed data. Because of the underlying data generation process, we could reasonably assume that the failure of a Hirst sampler did not depend on the amount of pollen it was exposed to or on the weather conditions around the device. The same is true for the human errors that resulted in missing values.

To get rid of the missing values that are known to be outside the natural season of a taxon, a first data preprocessing step was done. This was necessary to augment the data availability in the training data sets used to build the models. The idea here was to first get rid of the gaps that could be filled easily without relying on a more complex model.

### 3.1 Data preprocessing

There are several ways to characterize aerobiological time series like in Comtois and Sherknies (1991), Belmonte et al. (1999) and Kasprzyk and Walanus (2010). Belmonte and Canela (2002) proposed to use the Friedman Super Smoother (1984) as a nonparametric method to fit a smooth trend to pollen data time series. The idea here is to show with a seasonal curve how the pollen emissions behave during a given year when the variation due to the day-to-day weather conditions is taken out. The seasonal curve is an estimation of the expectation of the pollen concentration conditioned on the day of the year. Such a characterization is helpful for investigating the relationship between stations and taxa.

Besides using the Super Smoother with data from a single year, we can also use it to additionally generate a seasonal trend curve based on longer time frames of several years. Annual trend curves for some given taxa can vary substantially from year to year and are therefore not suitable to gain information that can be used for data imputation. A way of circumventing this problem was to superimpose the data points (day of the year/pollen concentration) of each year, before applying the Friedman Super Smoother on the obtained point cloud. The Friedman Super Smoother has several tuning parameters that determine the nature of the trend line fitted to the data. In our implementation, the default values for the tuning parameters of the `supsmu` function in the programming language R (R Development Core Team 2010) were used. Figure 4 shows an example of such a curve.



**Fig. 4** A seasonal Betula pollen concentration trend curve in Basel based on the years 2000–2009 using the Friedman Super Smoother. The dates are distinguished by the day of the year. A

*dashed line* shows the values for the interpolation that was used to fill in the inter-seasonal measurement gap in the curve

This kind of trend curves are based on several years and can be used to fill in missing values that are not in the inter-seasonal gap but still far away enough from the season peak. A way to implement this was to replace missing pollen concentration entries with zeros if the corresponding value of an associated trend curve, based on a given time frame of several years, was available and smaller than a predefined threshold. That threshold was set to 1 [count/m<sup>3</sup>] for all time series. The seasonal curves used for that purpose were based on data from the years 2000 to 2009.

Similarly, to generate a training data set, the inter-seasonal measurement gaps were also filled up. This was done in the following way. Whenever the values of the seasonal curve based on several years were <1 at the beginning and at the end of the available curve, a linear interpolation was done for the period where missing values occurred in the seasonal curve. That extended curve was then used as a tool to fill in some of the missing values in the raw data. An example for such an extension of a seasonal curve can be seen as a dashed line in Fig. 4.

Depending on the taxa, a substantial amount of missing values can thus be filled up without having to use more complex models that rely on weather parameters. An R function was written that takes a list of pollen tables from different stations and returns a list of tables where the missing values far outside the seasonal peak are replaced with zeros.

The pollen concentrations data were available in a daily resolution, which means that we had one value for each day (daily average concentration). The most straightforward attempt to build a regression model would have been to try to use that data in combination with daily averages of weather parameters, from the nearest weather station, to build a regression model. This approach is, however, problematic if we consider 2 days with roughly the same daily averages on the weather parameters but with different daily patterns in the hour-by-hour evolution of those parameters.

For a pollen-emitting plant and for airborne dispersal, it makes a difference if there is rainfall in the morning or in the evening, even if the mean rainfall for the day may be the same. Rainfall or high humidity will wash out pollen grains from the air and stop a plant from emitting pollen. If, for example, this happens in the morning for plants usually emitting pollen at that time of the day, one might expect to see a bigger impact on the pollen concentration of that particular day than when it happens in the evening, when pollen has already been dispersed. To also deal with those situations, it was therefore decided to use the weather parameters in an hourly resolution in our models. To do this, a matrix of predictors had to be set up in a adapted format.

In our data set, the pollen value from a single day represents the average concentration of the pollen that was collected from 08:00 to 08:00 am the next day. As we expected that the weather conditions observed a few hours before the daily measurement time frames

**Table 3** Variables used for creating the matrix of predictors together with the variables that were added later during the model building phase

	Unit	Variable name
Variables with hourly resolution		
Hourly mean temperature 2 m above soil	°C	<i>temperature</i>
Hourly mean relative humidity 2 m above soil	%	<i>humidity</i>
Hourly sum of evaporation	mm	<i>evaporation</i>
Hourly sum of precipitation	mm	<i>precipitation</i>
Hourly mean of global radiation	W/m <sup>2</sup>	<i>radiation</i>
Soil temperature 5 cm below soil	°C	<i>soil.temp</i>
Hourly mean of wind velocity	m/s	<i>wind.velocity</i>
Hourly mean of wind direction	°	<i>wind.direction</i>
Additional variables in matrix of predictors		
Sum of daily pollen values for the previous year	–	<i>p.y.cum.pollen</i>
Sum of daily pollen values already cumulated this year	–	<i>t.y.cum.pollen</i>
Daily mean temperature sum already cumulated this year	°C	<i>t.y.cum.temp</i>
Variables added during model building phase		
Lagged differences of <i>t.y.cum.temp</i>	°C	<i>diff.temp</i>
Mean relative humidity from 00:00 to 00:00 hours + 32 h	%	<i>mean.humidity</i>
Day of the year	d	<i>day</i>

could also have an important impact on the daily pollen count, we created predictor matrices with potential predictors. Each daily entry (each row) contained the 33 hourly weather values from 00:00 to 00:00 hours + 32 h for each of the eight weather parameters used (see the first part of Table 3).

Additionally, we created three other variables that were included in the matrix of predictors. The first one was the cumulated sum of daily pollen values for the previous year, known as Seasonal Pollen Index (SPI) (Mandrioli et al. 1998), labeled as *p.y.cum.pollen*. The fact is that for certain tree taxa a particularly high amount of pollen emitted in the previous year can result in a lower amount of pollen in the present year because some plants need an entire year to recover from such an effort. Examples in which such a cyclic behavior for the annual quantities of pollen has been observed which were already documented by Spieksma et al. (1995, 2003).

The second variable, *t.y.cum.pollen*, was the sum of daily pollen values already cumulated this year. Here, it was expected that for certain taxa we could find a yearly limit above which the available pollen reservoir would be empty—all the flowers having already emitted the pollen contained in their anthers and/or being faded.

The third variable added to the predictor matrix was the cumulated sum of daily mean temperatures,

*t.y.cum.temp*. This method simulates the growth of the plants before flowering and the production of pollen grains; such heat sums are commonly used in phenology (Schwartz 2003) and for predicting pollen seasons (e.g., Boyer 1973; Frenguelli and Bricchi 1998; Clot 2001; Rodriguez-Rajo et al. 2003).

Finally, three other variables were added at the model building phase. They were the differences in daily mean temperatures lagged by 1 day, *diff.temp*, the mean relative humidity during the time frame going from 00:00 to 00:00 hours + 32 h, *mean.humidity*, and the day of the year, *day*. The intention here was to have a matrix of potential predictors that was as big as possible in order to have a maximum of flexibility during the model building phase. By applying these steps for 16 taxa and 14 measurement sites, we ended up with 224 predictor matrices where each had 267 columns for the predictors and a varying number of rows for the number of observations.

### 3.2 Choosing a model building procedure

There were several requirements for the model building procedure when we wanted to use weather data to estimate the daily pollen count for each of the 224 pollen time series. Given that the initial goal was to have complete series whenever this is possible, a

regression model building procedure that can still make predictions even if a part of the predictor vector is missing was preferable. The reason is that missing values can also appear in the weather predictors.

Depending on the taxa and the measurement site considered, the predictors may contain several highly correlated and possibly irrelevant predictors. This can be a problem for many regression techniques. Additionally, the response variable, pollen concentration, results from a counting process (data values are discrete), and therefore, the model building algorithm used should preferably be able to do Poisson regression as well. Poisson regression is generally appropriate for count or rate data, where the rate is a count of events occurring to a particular unit of observation, divided by some measure of that unit's exposure.

The regression model obtained should also not rely too heavily on the pollen count of the previous day. There are two reasons for this. First, most of the holes we intend to fill occur in a row of missing values. If the information about the previous day is missing and if that variable is influential for the model created, most holes filled with such a model will be severely affected, if a prediction can be done at all. The second reason is that when doing predictions for the coming day by using hourly weather forecasts as predictors, in many cases the information about the pollen concentration of the previous day is available only several days later. The hindering fact is that the data collected in the Hirst sampler are only delivered to the laboratory once a week and not on a daily basis. This is obviously too late to be used as an input.

The construction of a model to predict the daily pollen concentration for a particular taxon, without using the pollen concentration of previous days, has been documented in the paper by Stark et al. (1997). In their approach, they used a generalized linear model, via the `glm` function in the statistical programming language S-PLUS, to build a linear Poisson regression model. That model used a set of transformed and untransformed weather variables as predictors. For their model, the predictors were obtained by creating a binary variable based on the hourly data of observed rainfall and by transforming other daily weather variables like temperature, wind and number of days after the start of the season. Their attempt to fit their model at the end of a season and then use the coefficients from that model to predict the pollen levels for the following years were, however,

unsuccessful. Their model was meant to be used for prediction only from day-to-day within each ragweed season. Furthermore, the model they proposed could not be used during the first 7 days of the season.

The reason to use transformations is to handle nonlinearities in the model. One can, for example, reasonably assume that in a certain range an increased temperature may have a positive effect on the pollen production, but it is, however, clear that above a certain threshold a further increase in temperature does not yield more pollen. Finding appropriate transformations is, however, extremely time-consuming and problematic when one has to create hundreds of prediction models. In our case, we decided to use a prediction model building procedure that would construct a model for each taxon and each measurement station by using the same set of predictors. To deal with such a problem, we could not rely on generalized linear models. Thus, another solution had to be sought.

Among today's available model building procedures that have a certain ability to handle irrelevant inputs and that can deal with missing values in the inputs, there are Friedman's multivariate adaptive regression splines (1991), Breiman's classification and regression Trees (1984) and algorithms that are based on trees. All these methods can model nonlinear relationships. Two recent regression methods that rely on trees are Gradient Boosting and Stochastic Gradient Boosting (Friedman 2001, 2002). Both algorithms are insensitive to strictly monotonic transformations in the individual predictors. This means that the basic structure of those sets of rules does not change if the input variables are transformed in a not too fancy way. In practical terms, this means that usually the model prediction performance and basic model structure will stay the same if, for example, the measurement units of the inputs are changed.

Gradient Boosting is a machine-learning technique for regression problems. Going through several iterations, it produces a prediction model in the form of a linear combination of simple base learners that are typically regression or classification trees. The models are built in a stagewise forward manner and are generalized by minimizing an arbitrary differentiable loss function.

In other words, the resulting models can be seen as a constructed sum of sets of "If-then"-rules that apply to the input values and result in a number (in the case of regression) or in a class (in case of classification). In this context, the term 'stagewise forward' roughly means



that when constructing that sum, a simple model containing a handful of rules is grown to a more complex one containing a much higher number of rules.

Stochastic Gradient Boosting is a special variant of Gradient Boosting. With that technique, randomization is introduced into the algorithm by using only a random subsample, taken without replacement, of the training data set at each iteration. The goal is to avoid overfitting to the particular data set and thus obtain a better prediction capability. Typically, only half of the training data set is used to build the base learners in each iteration. A substantial improvement in prediction accuracy can usually be obtained with this modification. This was also true on our data.

In our case, we used an R implementation of Stochastic Gradient Boosting made available through the `gbm` package developed by Ridgeway (2007). The package allows for a wide variety of loss functions including Poisson deviation. It gives the control on a lot of metaparameters used to tune the model (see Sect. 3.4) and also offers several helpful forms of graphical outputs.

Another useful feature in `gbm` is that restrictions can be imposed on the kind of dependencies the target variable has with the predictors. It is, for example, possible to ensure that a particular predictor has a monotone increasing or monotone decreasing relationship with the outcome.

### 3.3 Model assessment

To compare Poisson models on the same aerobiological time series or to assess the quality of retained models between stations and taxa, it is necessary to have a criterion that can be used for model assessment. Because we had 224 models to build, we did not want to rely on a score function based on cross-validation since that would have needed even longer computations.

A simpler strategy that was applied for each individual time series is to use the first 75% of the available data as a training data set. The remaining 25% of the data set was set aside to be taken as a test data set. By doing so, it was possible to ensure that the test data set contained data for at least 3 pollen seasons. The default methodology when using `gbm` for Poisson regression is to track the sum of Poisson deviances on the training sample and on the test sample. Since it is difficult to interpret such a statistic

directly, we had to additionally implement another goodness-of-fit measure.  $R^2$  or pseudo- $R^2$ -based goodness-of-fit measures can also be computed on test data and are also easier to interpret than the sum of Poisson deviances. One such pseudo- $R^2$  measure, especially suited for dealing with count data in Poisson regression models, was proposed by Cameron and Windmeijer (1996). It is a pseudo- $R^2$  based on deviance residuals for the fitted Poisson model. For the observations  $y_i$ , the arithmetic mean of the observations  $\bar{y}$ , the predictions  $\hat{\mu}_i$  and the number of observations  $N$ , that statistic is given by

$$R_{\text{DEV},P}^2 = 1 - \frac{\sum_{i=1}^N [y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)]}{\sum_{i=1}^N y_i \log(y_i/\bar{y})}$$

and can equivalently be implemented as

$$R_{\text{DEV},P}^2 = \frac{\sum_{i=1}^N [y_i \log(\hat{\mu}_i/\bar{y}) - (\hat{\mu}_i - \bar{y})]}{\sum_{i=1}^N y_i \log(y_i/\bar{y})}$$

where  $y \log(y)$  is set to zero for  $y$  equal zero.

On the data used to train a model, that measure will generally lie between 0 and 1 but can still yield negative values on some test samples where the prediction model performs catastrophically. Higher values of that statistic will be associated with better fits, and for a perfect fit, the value of that measure becomes 1.

Contrary to the case in linear regression where the usual  $R^2$  statistic based on the training data will increase with each additional predictor added to the model, an  $R^2$  or pseudo- $R^2$  statistic computed from a test sample will not necessarily behave in that way.

When assessing prediction models, we also need to know how important the different predictors are. For tree-based regression or classification methods, Breiman (1984) introduced a measure called *relative influence* that can be used for exactly that purpose. That statistic is determined for each predictor and is measured in percent. It is a weighted sum of the number of times a particular variable is used in a split in the regression trees used in the model.

Given two models with similar prediction performances, the model having relative influences that are more evenly dispersed among the predictors was preferred. The reason is that the latter model is less likely to perform poorly in cases where only one of the most important predictors is missing.

Not every conceivable problem can be detected with an  $R^2$ -based measure alone, which is why the script that built the models also created residual plots based on the entire data set and a plot of the predicted and observed time series based on the test and training data for visual evaluation.

### 3.4 Building the models

Whenever we build models with Stochastic Gradient Boosting, several tuning parameters (metaparameters) have to be set. They define how the model building algorithm learns from the data. The metaparameters are the *bag fraction*, the *shrinkage factor*, the *minimum number of observations in a node*, the *number of trees built* and the *interaction depth*. The bag fraction is the fraction between the number of observations in the subsample used to train the base learner in each iteration and the total number of observations contained in the entire training data set. If that fraction is set to 1, Gradient Boosting is performed instead of Stochastic Gradient Boosting.

The shrinkage factor is a way of telling the algorithm how fast it should learn at each iteration. In general, increasing the shrinkage factor comes at the expense of prediction performance on unseen data points. The minimum number of observations in a node is used to control the minimal size of node for the regression trees built at each iteration. The number of trees built gives the number of regression trees used for the linear combination representing the model. Said in easier terms, it determines the complexity of the final model. Finally, the interaction depth gives the number of nodes the base learners constructed at each iteration must have. That metaparameter defines the complexity of the set of rules that are used to grow the final model. Except for the bag fraction where the usage of a value near 1/2 is considered proper, there is no general strategy to choose the other metaparameters and thus one has to use experience and trial and error to find a set of values that can be considered for building the model. An exact definition of the metaparameters and suggestions on how to set some of them is given by Friedman (2002).

The associated loss function to minimize in the algorithm was set to the Poisson deviance. Generally, there is no systematic way to choose the tuning parameters. Usually, this step has to be done by hand

guided only by experience and some rules of thumb. In our case, we set the maximum number of trees in the model to 1,200, the shrinkage factor to 0.008 and the interaction depth to 24. For the bag fractions, the values 1/5, 1/3 and 1/2 were considered. The bag fraction determines the size of the subsamples of the training data set that are actually used at each iteration during the model building. The possible values for the minimum number of observations in a node were set to 5, 10, 15 and 20. For each time series, all 12 possible combinations of these metaparameters were tried out to build models with 1,200 trees.

Because prediction models consisting of 1,200 trees may end up being too complex and possibly overfit the data, they were pruned using an out-of-bag (OOB) performance measure to determine the best number of trees. This means that the best number of trees was determined by evaluating in each iteration the reduction in Poisson deviance on those observations not used in selecting the next regression tree. Those observations left out at each iteration constitute a sequence of OOB samples. The ideal number of trees to use is then set to a value such that the reduction in Poisson deviance on the out-of-bag samples does not improve further for iterations that go above that number.

The `gbm` package also offers the possibility to determine the optimal number of trees based on the test sample score. It was decided to rely on the OOB estimator for the optimal number of trees used in the model to prevent overoptimistic values for the  $R^2_{DEV,P}$  computed on the test samples.

The pruned models for all twelve possible combinations of metaparameters were then compared to each other with respect to their computed  $R^2_{DEV,P}$  value on the test data set. In the end, the best combination of metaparameters among the twelve possible ones was retained.

Initial prediction models for individual time series were constructed with `gbm` by using all the variables in the matrix of predictors. Later on, the three additional predictors *day*, *diff.temp* and *mean.humidity* (Table 3) were also included as predictors and other variables manually removed. For all hourly temperature variables, a monotone increasing relationship with the target variable was assumed, and for all hourly humidity variables, a monotone decreasing relationship was imposed on the model building algorithm.

## 4 Results

### 4.1 Models

Although the Stochastic Gradient Boosting algorithm is, to a certain extent, able to handle irrelevant or highly correlated inputs, the method is not completely immune to a bad choice in parameters. In general, irrelevant predictors or predictors that are a good proxy for other variables favored by the model building algorithm will end up having a small relative influence on the final model. Removing some predictors by hand, however, did slightly improve the models. There is no standard automated way for doing this. We started by removing some variables or groups of variables and verified whether it generally resulted in higher  $R_{DEV,P}^2$  values on the test samples. We first removed the variables *p.y.cum.pollen* and *t.y.cum.pollen*.

Respectively, all 33 evaporation, soil temperature, wind direction, precipitation, wind speed and global radiation variables were also discarded. With respect to precipitation that is usually considered to be a first order predictor (Norris-Hill 1995), it is noted that apparently the humidity, which also partly reflects precipitation, alone is sufficient. Furthermore, we reduced the resolution of the temperature and humidity parameters by removing the variables for the even hours 00:00, 2:00, 4:00, 6:00, . . . , 32:00. Finally, the last remaining hours 30:00 and 31:00 were also removed for the temperature and the humidity. From the 270 variables considered initially, only 34 were finally retained. Those predictors are shown in Table 4.

The 40 highest  $R_{DEV,P}^2$  scores for the models that were estimated with the final configuration of predictors are shown in Table 5. Among the 223 models estimated—the *gbm* model building algorithm broke down when trying to build a model for *Ambrosia* in Davos—39 had  $R_{DEV,P}^2$  test sample scores that were above 0.75. The impossibility to estimate a model for *Ambrosia* in Davos comes from the fact that there is almost no pollen of that taxon that is measured there. An example of the model behavior on the test sample for Poaceae in Geneva is given in Fig. 5.

### 4.2 Completed time series

To create completed aerobiological time series, it was decided to only retain the imputed missing values in

**Table 4** A list of the 34 variables retained in the final models used for each aerobiological time series

<i>temperature.H2</i>	<i>temperature.H4</i>	<i>temperature.H6</i>
<i>temperature.H8</i>	<i>temperature.H10</i>	<i>temperature.H12</i>
<i>temperature.H14</i>	<i>temperature.H16</i>	<i>temperature.H18</i>
<i>temperature.H20</i>	<i>temperature.H22</i>	<i>temperature.H24</i>
<i>temperature.H26</i>	<i>temperature.H28</i>	<i>temperature.H30</i>
<i>humidity.H2</i>	<i>humidity.H4</i>	<i>humidity.H6</i>
<i>humidity.H8</i>	<i>humidity.H10</i>	<i>humidity.H12</i>
<i>humidity.H14</i>	<i>humidity.H16</i>	<i>humidity.H18</i>
<i>humidity.H20</i>	<i>humidity.H22</i>	<i>humidity.H24</i>
<i>humidity.H26</i>	<i>humidity.H28</i>	<i>humidity.H30</i>
<i>t.y.cum.temp</i>	<i>day</i>	<i>diff.temp</i>
<i>mean.humidity</i>		

The variable *temperature.H2* corresponds to the temperature at 1:00 hours and the variable *temperature.H4* to the temperature at 3:00 hours and so on for the rest of the variables

those time series for which the prediction models had an  $R_{DEV,P}^2$  score of at least 0.75 on the test data set. A script was written that replaced each missing value with the corresponding prediction, based on the measured weather parameters in each station. Since the prediction model did not directly yield integer values for the predictions, it was therefore chosen to truncate them toward zero in order to get a reference data set that contained only nonnegative integer values. This way a reference data set with data ranging from the first measurement periods to the end of 2009 was created.

## 5 Discussion

### 5.1 Models

A station that is particularly badly represented in the top 40 is the station Davos (PDS) located at an altitude of 1,560 m. Pollen concentrations in that station are usually low and an important part of pollen found there is produced at lower altitude and brought there by wind transport. Visual inspection of plots, showing the predicted versus the observed values, indicated that a lot of the time series with low  $R_{DEV,P}^2$  scores for their models usually had low observed pollen levels. Pollen taxa that do not flower every year are especially difficult to model. The models for the *Fagus* taxon are

**Table 5** Prediction performance on the 25% test samples measured with  $R_{DEV,P}^2$  for each station and taxon

Rank	Taxa	Station	$R_{DEV,P}^2$ on test sample
1	<i>Betula</i>	PLO	0.873
2	<i>Castanea</i>	PLO	0.862
3	<i>Quercus</i>	PLU	0.858
4	Poaceae	PNE	0.853
5	<i>Castanea</i>	PLU	0.851
6	Poaceae	PGE	0.851
7	<i>Quercus</i>	PLS	0.838
8	<i>Betula</i>	PLU	0.835
9	<i>Populus</i>	PGE	0.831
10	Poaceae	PZH	0.829
11	<i>Quercus</i>	PLO	0.827
12	Urticaceae	PBE	0.823
13	Poaceae	PCF	0.812
14	Poaceae	PBE	0.812
15	<i>Quercus</i>	PGE	0.81
16	Urticaceae	PCF	0.806
17	<i>Betula</i>	PGE	0.805
18	<i>Betula</i>	PLS	0.804
19	Urticaceae	PLS	0.803
20	Urticaceae	PGE	0.803
21	Urticaceae	PNE	0.798
22	<i>Betula</i>	PMU	0.797
23	Poaceae	PLS	0.796
24	<i>Betula</i>	PCF	0.785
25	<i>Betula</i>	PBS	0.785
26	Urticaceae	PZH	0.782
27	Poaceae	PMU	0.78
28	<i>Populus</i>	PLS	0.78
29	<i>Fraxinus</i>	PLO	0.777
30	<i>Alnus</i>	PGE	0.775
31	<i>Corylus</i>	PGE	0.767
32	Poaceae	PDS	0.766
33	<i>Betula</i>	PLZ	0.764
34	<i>Quercus</i>	PBS	0.764
35	Poaceae	PBU	0.761
36	Poaceae	PBS	0.76
37	<i>Fraxinus</i>	PLU	0.759
38	<i>Populus</i>	PLO	0.757
39	<i>Betula</i>	PBU	0.753
40	Urticaceae	PBU	0.748

Only the 40 best models are shown

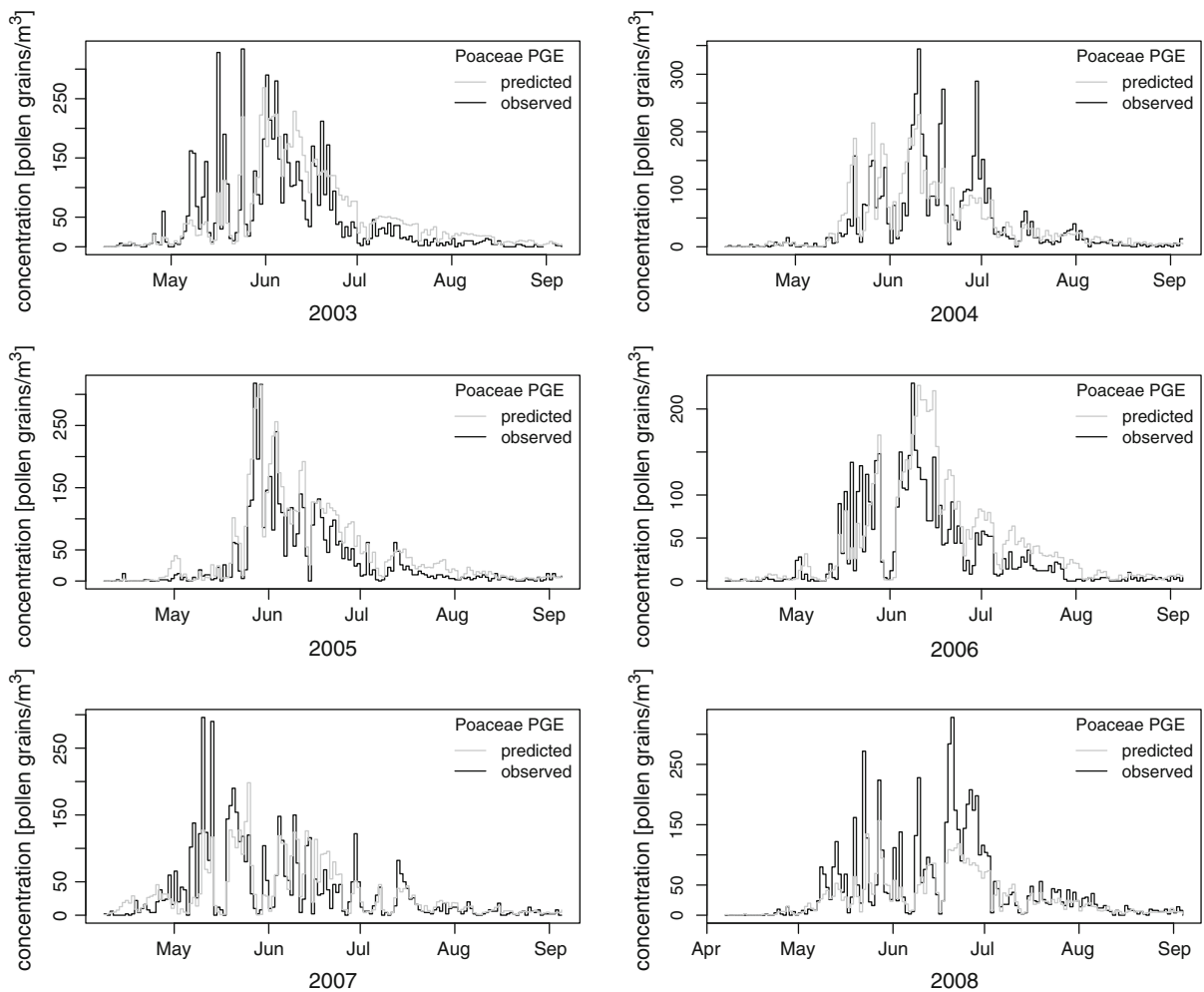
an example therefor. If accurate hourly forecasts for the weather predictors used in the models can be made available, prediction models for the daily pollen counts become conceivable for those combinations of taxa and stations with high  $R_{DEV,P}^2$  scores in Table 5.

The fact that  $R_{DEV,P}^2$  score of, for example, Urticaceae PBU is  $<0.75$  is not a reason to leave it out from further investigation. The score provides only one of several possible assessment criteria for a first evaluation of the models and should therefore be used as a guiding hand for further evaluation, not as an absolute criteria. In the end, the usefulness and weaknesses of a model can only be gauged in detail by taking a look at plots of observations vs. predicted values computed on the test data. Thus, further investigation might show that the model Urticaceae PBU can be useful for some applications. Because of space constraints, we restricted ourselves to show only the results for the first 40 models.

Some prediction models from Table 5, like the one depicted in Fig. 5, are ideal candidates to deliver daily pollen predictions for several years before needing to be retrained. In such a setting, the models can make predictions of the daily pollen counts for the coming 2–3 days using point information of weather parameters from numerical weather prediction as input. That latter time frame only depends on the time horizon that can be reasonably covered with the weather forecasts.

The  $R_{DEV,P}^2$  scores in Table 5 were computed for predictions made with weather measurements. However, an adequate model assessment in that context would require a further study and also the access to weather forecast data preferably spanning several years. The input parameters used in the models in Table 4 can all be made available up to 10 days in advance by numerical weather prediction programs commonly used by weather forecast offices. This would allow to use such models to make pollen forecasts based on numerical weather forecast.

The kind of models proposed here cannot be used to describe the transport of the pollen particles. This would require to implement the behavior of a pollen-emitting plant in a numerical weather prediction program. Such an approach is, however, difficult and requires a substantial amount of information about the location of the pollen sources (Pauling et al. 2011).



**Fig. 5** Predicted and observed Poaceae pollen concentrations in Geneva (PGE) from the years 2003 to 2008. All seasons shown here are taken from the test data. The training data used go from January 1979 to November 2002

## 5.2 Imputing missing values

Missing values in the original pollen time series can be replaced with model predictions, if a good enough model can be found for the time series in question. Choosing a minimal value for the  $R^2_{DEV,P}$  value computed on a test sample is the most simple, but somewhat arbitrary approach to solve that problem.

Another possible way of determining what constitutes a good enough model is to make that decision on a case-by-case basis, depending on the particular statistic one wants to compute on the completed time series. The sensitivity of a particular yearly statistic can, for example, be assessed by evaluating that statistic on predictions and original values from the

test data set on days where actual measures are available. If the discrepancy between the value computed on the model values and the value computed on the actual observed values is deemed sufficiently small for several years in a row contained in the test data set, the model can be used to complete the time series for that particular statistic. For each particular statistic to be computed on a yearly basis, one could then decide whether a particular model should be used to complete the raw time series or not.

## 5.3 Model building

The computations required for building the 224 models were simply split up among 23 cores on a

computer cluster. All prediction models could thus be computed in under a day. Since we used bag fractions smaller than one in our models, it is conceivable to run Monte Carlo simulations on the model building procedure to get an ensemble of models for each time series in order to try to gauge the uncertainty associated with each prediction. The drawback is, however, that one requires much more computing time to estimate the models and memory space to store them.

The 223 models that were obtained are not necessarily the best that could be achieved in terms of predictive performance on the test samples if a lot more time and energy would be invested in manually choosing the predictors individually for each model or if we would do a larger search in the set of metaparameters used in the prediction models. It is, however, doubtful that the gains that could be made would be worth the effort.

Another problem appears when a too extensive automated search is done on the metaparameters. First, it is costly to try out an extensive list of combinations of metaparameters. Secondly even if such a search can be done in a reasonable amount of time, one runs the risk of overtuning the prediction models. This would result in model assessments that are slightly too optimistic.

**Acknowledgments** The authors would like to thank Katrin Zink for helping to improve the readability of the text.

## References

- Arizmendi, C., Sanchez, J., Ramos, N., & Ramos, G. (1993). Time series predictions with neural nets: Application to airborne pollen forecasting. *International Journal of Biometeorology*, 37(3):139–144.
- Belmonte, J., & Canela, M. (2002). Modelling aerobiological time series. Application to Urticaceae. *Aerobiologia*, 18(3):287–295.
- Belmonte, J., Canela, M., Guardia, R., Guardia, R., Sbai, L., Vendrell, M., et al. (1999). Aerobiological dynamics of the Urticaceae pollen in Spain, 1992–98. *Polen*, 10, 79–91.
- Boyer, W. (1973). Air temperature, heat sums, and pollen shedding phenology of longleaf pine. *Ecology*, 54(2), 420–426.
- Breiman, L. (1984). *Classification and regression trees*. London/Boca Raton, FL: Chapman & Hall/CRC.
- Bringfelt, B., Engström, I., & Nilsson, S. (1982). An evaluation of some models to predict airborne pollen concentration from meteorological conditions in Stockholm, Sweden. *Grana*, 21(1), 59–64.
- Cameron, A., & Windmeijer, F. (1996). R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business & Economic Statistics*, 14(2), 209–220.
- Castellano-Méndez, M., Aira, M., 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. *International Journal of Biometeorology*, 49(5), 310–316.
- Clot, B. (2001). Airborne birch pollen in Neuchâtel (Switzerland): Onset, peak and daily patterns. *Aerobiologia*, 17(1), 25–29.
- Comtois, P., Alcazar, P., & Neron, D. (1999). Pollen counts statistics and its relevance to precision. *Aerobiologia*, 15(1), 19–28.
- Comtois, P., & Sherknies, D. (1991). Pollen curves typology. *Grana*, 30(1), 184–189.
- Cotos-Yáñez, T., Rodríguez-Rajo, F., & Jato, M. (2004). Short-term prediction of betula airborne pollen concentration in Vigo (NW Spain) using logistic additive models and partially linear models. *International Journal of Biometeorology*, 48(4), 179–185.
- Cox, C., & Wathes, C. (1995). *Bioaerosols handbook*. USA: Lewis publishers.
- Frenguelli, G., & Bricchi, E. (1998). The use of the phenoclimatic model for forecasting the pollination of some arboreal taxa. *Aerobiologia*, 14(1), 39–44.
- Friedman, J. (1984). *A variable span smoother*. Department of Statistics. Technical report. Stanford, CA: Stanford University.
- Friedman, J. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1), 1–67.
- Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.
- Friedman, J. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4), 367–378.
- Galán, C., Cariñanos, P., García-Mozo, H., Alcazar, P., & Domínguez-Vilches E. (2001) Model for forecasting Olea europaea L. airborne pollen in South–West Andalusia, Spain. *International journal of biometeorology*, 45(2), 59–63.
- Gregory, P. (1961). *The microbiology of the atmosphere*. London: Leonard Hill.
- Hirst, J. (1952). An automatic volumetric spore trap. *Annals of Applied Biology*, 39(2), 257–265.
- Isard, S., & Gage, S. (2001). *Flow of life in the atmosphere: An airscape approach to understanding invasive organisms*. East Lansing: Michigan State University Press.
- Kasprzyk, I., & Walanus, A. (2010). Description of the main poaceae pollen season using bi-Gaussian curves, and forecasting methods for the start and peak dates for this type of season in rzeszów and ostrowiec św. (SE Poland). *Journal of Environmental Monitoring*, 12(4), 906–916.
- Makra, L., Juhász, M., Borsos, E., & Béczi, R. (2004). Meteorological variables connected with airborne ragweed pollen in southern Hungary. *International Journal of Biometeorology*, 49(1), 37–47.
- Mandrioli, P., Comtois, P., & Levizzani, V. (1998). *Methods in aerobiology*. Bologna: Pitagora Editrice.
- Norris-Hill, J. (1995). The modelling of daily Poaceae pollen concentrations. *Grana*, 34(3), 182–188.

- Pauling, A., Rotach, M., Gehrig, R., & Clot, B. (2011). A method to derive vegetation distribution maps for pollen dispersion models using birch as an example. *International Journal of Biometeorology*, 1–10.
- Peeters, A., Clot, B., Gehrig, R., et al. (1998). Luftpollengehalt in der schweiz. Zürich, Switzerland: SMA–Schweizerische Meteorologische Anstalt.
- Peng, R. (2008). A method for visualizing multivariate time series data. *Journal of Statistical Software* 25(1), 1–17.
- R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07-0.
- Ranzi, A., Lauriola, P., Marletto, V., & Zinoni, F. (2003). Forecasting airborne pollen concentrations: Development of local models. *Aerobiologia* 19(1), 39–45.
- Ridgeway, G. (2007). *gbm: Generalized boosted regression models*. R package version 1.6-3.
- Rodriguez-Rajo, F., Frenguelli, G., & Jato, M. (2003). Effect of air temperature on forecasting the start of the betula pollen season at two contrasting sites in the south of Europe (1995–2001). *International Journal of Biometeorology*, 47(3), 117–125.
- Schwartz, M. (2003). *Phenology: An integrative environmental science*. Dordrecht: Kluwer Academic Publishers.
- Smith, M., & Emberlin, J. (2006). A 30-day-ahead forecast model for grass pollen in north London, United Kingdom. *International Journal of Biometeorology*, 50(4), 233–242.
- Spieksma, F., Corden, J., Detandt, M., Millington, W., Nikkels, H., Noland, N., et al. (2003). Quantitative trends in annual totals of five common airborne pollen types (betula, quercus, poaceae, urtica, and artemisia), at five pollen-monitoring stations in western Europe. *Aerobiologia*, 19(3), 171–184.
- Spieksma, M., Emberlin, J., Hjelmroos, M., Jäger, S., & Leuschner, R. (1995). Atmospheric birch (Betula) pollen in Europe: Trends and fluctuations in annual quantities and the starting dates of the seasons. *Grana*, 34(1), 51–57.
- Stach, A., Smith, M., Prieto Baena, J., & Emberlin, J. (2008). Long-term and short-term forecast models for poaceae (grass) pollen in Poznan, Poland, constructed using regression analysis. *Environmental and Experimental Botany*, 62(3), 323–332.
- Stark, P., Ryan, L., McDonald, J., & Burge, H. (1997). Using meteorologic data to predict daily ragweed pollen levels. *Aerobiologia*, 13(3), 177–184.
- Voukantsis, D., Niska, H., Karatzas, K., Riga, M., Damialis, A., & Vokou, D. (2010). Forecasting daily pollen concentrations using data-driven modeling methods in Thessaloniki, Greece. *Atmospheric Environment*, 44(39), 5101–5111.