Aerobiologia
https://doi.org/10.1007/s10453-018-9507-9

ORIGINAL PAPER

How to prepare a pollen calendar for forecasting daily pollen concentrations of *Ambrosia*, *Betula* and Poaceae?

B. Šikoparija · O. Marko · M. Panić · D. Jakovetić · P. Radišić

Received: 6 June 2017 / Accepted: 16 January 2018
© Springer Science+Business Media B.V., part of Springer Nature 2018

Abstract Forecasting daily airborne pollen concentrations is of great importance for management of seasonal allergies. This paper explores the performance of the pollen calendar as the most basic observation-oriented model for predicting daily concentrations of airborne *Ambrosia*, *Betula* and Poaceae pollen. Pollen calendars were calculated as the mean or median value of pollen concentrations on the same date in previous years of the available historic dataset, as well as the mean or median value of pollen concentrations of the smoothed dataset, pre-processed using moving mean and moving median. The performance of the models was evaluated by comparing forecasted to measured pollen concentrations at both daily and 10-day-average resolutions. This research demonstrates that the interpolation of missing data and pre-processing of the calibration dataset yields lower prediction errors. The increase in the number of calibration years corresponds to an improvement in the performance of the calendars in predicting daily pollen concentrations. However, the most significant improvement was obtained using four calibration years. The calendar models correspond well to the shape of the pollen curve. It was also found that daily resolution instead of 10-day averages adds to their value by emphasising variability in pollen exposure, which is important for personal assessment of dose-response for pollen-sensitive individuals.

Keywords Airborne pollen · Ragweed · Birch · Grass · Data-driven predictive modelling

1 Introduction

A recent review of allergenic pollen highlighted twelve pollen types: *Ambrosia*, *Alnus*, *Artemisia*, *Betula*, Amaranthaceae (includes former goosefoot family Chenopodiaceae), *Corylus*, Cupressaceae/Taxaceae, *Olea*, *Platanus*, Poaceae, *Quercus* and *Parietaria*/*Urtica* for their allergenic potency and abundance in the atmosphere (Skjøth et al. 2013). The majority of pollen sensitisations in Europe are caused by *Betula* and Poaceae (Bousquet et al. 2007). *Ambrosia* is the second most important cause of seasonal asthma and rhinitis in many areas of its native

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10453-018-9507-9) contains supplementary material, which is available to authorized users.

B. Šikoparija · O. Marko (✉) · M. Panić · P. Radišić
BioSense Institute - Research Institute for Information Technologies in Biosystems, University of Novi Sad, Novi Sad, Serbia
e-mail: oskar.marko@biosense.rs

D. Jakovetić
Department of Mathematics and Informatics, Faculty of Sciences Novi Sad, University of Novi Sad, Novi Sad, Serbia

Published online: 23 January 2018

distribution range (i.e. North America), and in the past decade, its clinical relevance has increased notably throughout Europe (Smith et al. 2013 and references therein). It is estimated that the number of allergic people in Europe will more than double by 2060 (Lake et al. 2017). Hence, the ability to predict the variability of daily pollen concentrations for the most important allergenic pollen would be beneficial for a great number of pollen-sensitive individuals.

Avoidance of airborne pollen is an important part of allergic symptoms management (Peden and Reed 2010). In order to be effective, avoidance requires temporally resolved forecasts on the quantity of the airborne pollen present in the atmosphere over the year. Forecasting in aerobiology is commonly based on either observation-oriented models or source-oriented models (Scheifinger et al. 2013; references therein). The former relate pollen records to one or more measurable variables (Norris-Hill 1995), while the latter use mathematical formulae of atmospheric transport and diffusion to calculate concentrations at various distances from a known source (Skjøth et al. 2010). Despite the advantages of both modelling approaches, their implementation requires a significant amount of data for calibration and running (i.e. pollen concentrations and meteorology); substantial information about the location, abundance and emission characteristics of the pollen source (Skjøth et al. 2010; Pauling et al. 2012); and significant computational resources. As a result, none of these models are available in many regions.

The emission of allergenic pollen from its source is directly linked to seasonality of the flowering phenophase (Dahl et al. 2013), which makes pollen calendars the simplest observation-oriented approach for predicting the time of occurrence of airborne pollen in a given area (Scheifinger et al. 2013). Although pollen calendars are commonly used all over the world to depict seasonal distribution of pollen in the atmosphere (e.g. El-Ghazaly and Fawzy 1988; Cadman 1990; O'Rourke 1990; Kok Ong et al. 1995; Kaya and Aras 2014; Jae-Won et al. 2012; Piotrowska-Weryszko and Weryszko-Chmielewska 2014; Martínez-Bracero et al. 2015; Calderón-Ezquerro et al. 2016), most of the calendars are produced using Spieksma's model (Spieksma and Wahl 1991). However, the 10-day temporal resolution of this model limits its application in predicting day-to-day variability of pollen concentrations.

The aims of this study are to build calendars for airborne *Ambrosia*, *Betula* and Poaceae pollen and to explore their performance in forecasting daily concentrations of these important allergens over the course of the season. The study also tests the suitability of different interpolation methods for completing short missing datasets (up to 7 days) that commonly occur in routine pollen monitoring using Hirst-type samplers (e.g. electric power loss, clock device stoppage).

2 Materials and methods

2.1 Airborne pollen data

Daily average concentrations of airborne *Betula*, Poaceae and *Ambrosia* pollen were collected in three Serbian cities: Novi Sad (45.25°N, 19.85°E), Sombor (45.77°N, 19.11°E) and Niš (43.32°N, 21.90°E) using the 7-day volumetric spore trap of the Hirst design (Hirst 1952). It was positioned above the local pollen sources to ensure regional representativeness of obtained data. The collected samples were prepared and analysed to retrieve daily average pollen concentrations expressed as particles per cubic metre of air (P m^{-3}) following European Aerobiology Society's minimum recommendations (Galán et al. 2014). The main part of the research was conducted using the data from Novi Sad that was collected between 2000 and 2015. This dataset was the largest and the most representative, since it included both seasons with high and low pollen records (Table 1). The datasets from Sombor and Niš were collected from 2009 to 2015 and from 2007 to 2011, respectively. There were at least 5 years in both of them, which exceeded the minimum number required for calibrating a calendar model. These two datasets were used for exploring the performance of the models in other locations.

Novi Sad is situated on the Pannonian Plain, one of the centres of European distribution of *Ambrosia artemisiifolia* L. (Smith et al. 2013), and is therefore surrounded by abundant sources of ragweed pollen. Conversely, local sources of *Betula* airborne pollen are limited to trees planted as ornamentals in the streets and parks, thus making the aerobiological situation notably dependent from atmospheric transport of birch pollen from distant sources. There are no large grasslands in the study region, and most of the

Table 1 Characteristics of the principal pollination period of *Betula*, Poaceae and *Ambrosia* in Novi Sad

Year	*Betula*			Poaceae			*Ambrosia*		
	Missing data	Concentrations sum	Average concentration	Missing data	Concentrations sum	Average concentration	Missing data	Concentrations sum	Average concentration
2000	0	791	9	1	306	3	1	2761	23
2001	0	1157	13	0	2349	19	16	12696	119
2002	14	4748	62	14	778	7	8	5466	48
2003	34	4442	79	11	676	6	5	7828	66
2004	12	2273	29	25	756	7	18	5862	56
2005	46	1610	37	46	380	4	25	3325	34
2006	24	8612	130	11	613	5	18	7576	72
2007	8	3422	42	8	528	4	0	6609	54
2008	0	9923	110	5	466	4	5	6114	52
2009	0	2011	22	5	556	5	5	5831	49
2010	0	16612	185	0	914	7	0	7100	58
2011	0	5092	57	32	521	5	20	5603	54
2012	0	10,455	116	5	337	3	5	5018	43
2013	0	1462	16	10	569	5	2	8113	67
2014	0	14,945	166	0	690	6	0	7638	62
2015	0	5357	60	0	495	4	0	8018	65
Average	9	5807.00	71	11	683.38	6	8	6597.38	58
Std. Dev.	14.29	4941.16	55.34	13.10	473.55	3.81	8.46	2263.74	20.69

Values for seasons with above-average amounts of recorded pollen are written in shaded cells

Poaceae pollen sources are limited to cereal crop fields and public and private green areas, which are heavily managed. Conditions in Sombor are very similar, as it is also located on the Pannonian Plain, less than 100 km from Novi Sad, while Niš is situated in south-eastern Serbia in the Nišava river valley. It is surrounded by hills, and thus, vegetation, pollen sources and pollen transmission are different to those observed on the Pannonian Plain.

There is a distinct seasonality in occurrence of airborne *Betula*, Poaceae and *Ambrosia* pollen on the Pannonian Plain (Radišić 2013), which results from seasonality in flowering phenology of pollen sources. In order to avoid overestimation of correlation coefficients in long all-zero sequences of the dataset, the analysis focused on principal pollination periods of *Betula*, Poaceae and *Ambrosia* in the study region. The limits of these periods were determined as calendar months in which pollen is not accidentally recorded (i.e. at least for 1 day in 16-year average > 1 P m^{-3} is recorded): (1) for *Betula* March–May, (2) for Poaceae April–October, (3) for *Ambrosia* July–September. In comparison with commonly used percentage methods (Emberlin et al. 1997; Pathirane 1975), our approach includes long tails with low and intermittent records at the beginning and the end of the flowering period, which allows for testing models both during peak flowering period and outside of it.

2.2 Data analysis

Data analysis was performed in R (R core Team 2016), and charts were produced in Microsoft Excel and Mathworks MATLAB. The R Code along with instructions on how to use it for calibrating and testing calendar models is given in Online Resources 1–3.

2.2.1 Interpolation of missing data

Hirst-type samplers were used to continuously sample airborne particles until the drum has to be changed. Most devices in current operation (i.e. Lanzoni "VPPS2000" and "VPPS 2010," Burkard ltd., Burkard Scientific) allow for uninterrupted sampling for up to 7 days. This was convenient for long-term monitoring with unattended sampling, but could result in 7-day data gaps if the device stops working (i.e. due to electricity issues, issues with clock device that rotates the drum or due to clogged sampling orifice).

Missing values in the calibration dataset could affect the performance of the constructed prediction model. For example, when producing the calendar model, missing values will result in a decrease in the number of calibration years and therefore limit robustness of the model. In order to be able to predict pollen concentration for any day with the highest possible accuracy, it was decided to interpolate missing data in the calibration dataset. However, missing data in the validation dataset were not interpolated in order to avoid comparison of modelled to interpolated data.

Three common methods available in R Zoo package (Zeileis and Grothendieck 2005) were tested: (1) *last-observation-carried-forward* interpolation, (2) linear interpolation and (3) cubic spline interpolation. All of the possible 7-day sequences, for which the measured data were available, were consecutively declared as a gap week. After the missing values were interpolated, the performance of each method was evaluated by comparing the calculated and measured pollen concentrations by using the same indicators that were used for comparing predicted and measured concentrations. For all analysed pollen types, linear interpolation had the highest correlation coefficient, while NRMSE (normalised root-mean-square error) was the lowest (Table 2) and thus it was selected as the method of choice for the interpolation of the missing pollen data in the calibration dataset.

2.2.2 Calendar model

Standard calendar models predict the pollen concentration based on observed pollen concentrations on the same day in the calibration dataset. By calculating the mean or the median of available historic concentrations on the same day of the year, the model takes into account seasonal variability of the pollen concentration signal.

Standard calendar models typically follow Spieksma's model which presents concentrations as 10-day means (Spieksma and Wahl 1991). This approach overcomes season-specific short-term shifts in pollen occurrence, but hides day-to-day variability in pollen concentrations and so has limited potential for allergy management. In order to limit the influence of short-term shifts while keeping day-to-day variability in pollen signal, advanced calendar models were developed, in which the concentration for a given day was

Table 2 Performance of interpolation of 7-day data gaps using last-observation-carried-forward (LOCF), linear interpolation and cubic spline measured by average normalised root-mean-square error (NRMSE) and average Spearman's rank correlation coefficient (ρ)

Interpolation method	*Betula*		Poaceae		*Ambrosia*	
	ρ	NRMSE	ρ	NRMSE	ρ	NRMSE
LOCF	–	8.63	–	1.21	–	1.53
Linear	0.29	2.05	0.16	0.85	0.22	0.96
Cubic spline	0.20	69,741.83	0.08	47.23	0.11	36.12

calculated as the mean or the median using signals that were pre-processed by either moving average or moving median. This pre-processing resulted in four types of calendar models: mean of moving mean signals (MNMN), mean of moving median signals (MNMD), median of moving mean signals (MDMN) and median of moving median signals (MDMD). Different interval lengths (1, 2, …, 30 days) for calculating moving means/medians were tested in order to choose the most appropriate value resulting in the best performing calendar model. Odd window lengths are standard for moving mean and median filters, as they take the concentrations on the chosen date $\pm\, n$ days ($n \in N_0$). However, in R, they are defined for even numbers as well. In that case, the number of days in advance is for one greater than the number of previous days.

The relationship between the number of available years for the calibration dataset and the model performance for both standard calendar model and the best advanced calendar model were also analysed.

2.2.3 *Measures of model performance*

The performance of the calendar models developed in this study was assessed by comparing calculated and measured pollen concentrations. Correlation coefficients were used to analyse the strength of the relationship between two signals (i.e. simultaneous increase/decrease). The Shapiro–Wilk test (Razali and Wah 2011) indicated that daily pollen concentrations were not normally distributed and therefore Spearman's rank correlation coefficient (ρ) was calculated, which indicates how well the ranking order of predicted and measured concentrations correspond.

In order to provide comparable assessments of model performance with respect to predicting the magnitude of pollen concentrations, the root-mean-square error (RMSE) was calculated, which is a commonly used measure for comparing the predicted and measured values in aerobiology (Astray et al. 2010; Makra and Matyasovszky 2011; Kasprzyk 2009; Csépe et al. 2014), but also in other areas such as agriculture (Marko et al. 2016; Bornn and Zidek 2012), geosciences (Chai et al. 2009; McKeen et al. 2005) and telecommunications (Altiparmak et al. 2009). RMSE overestimates large deviations due to squaring. However, the context of the season is not accounted for. RMSE of 50 P m^{-3} may be rather insignificant for seasons with huge amounts of pollen, but it can also be very significant in case pollen concentrations were small during the season. This could be misleading when comparing the model performance for pollen types with notably different abundance (like Poaceae with *Ambrosia* or *Betula* in the study) and also for the same pollen type that shows notable seasonal variability (like *Betula* in the study). Therefore, for assessing the accuracy of the model to predict the magnitude of daily pollen concentrations we followed the suggestion by Makra et al. (2011) and normalised root-mean-square error by the mean concentration in the analysed year (NRMSE).

Leave-one-year-out cross-validation procedure was applied to test the performance of calendar models. Each year from the dataset was chosen at one point as the test dataset, while the remaining 15 years were taken as the calibration dataset. This procedure was repeated 16 times, so that each year would serve as the test year for which Spearman's ρ and NRMSE were calculated. In order to get a single measure of model performance, average of obtained Spearman's ρ and NRMSE was calculated. Independent-samples Kruskal–Wallis test was conducted to explore whether

there is a difference in distribution of Spearman's ρ and NRMSE between different calendar models.

3 Results

There was notable seasonal variability in the amounts of recorded pollen (Table 1). The lowest value was recorded in 2000 for all analysed pollen types. The year 2001 was the one with the highest concentrations of both Poaceae and *Ambrosia* pollen and was the second lowest for *Betula* pollen. In general, *Betula* and *Ambrosia* pollen were more abundant than Poaceae pollen in the atmosphere of Novi Sad. Above-average seasons of *Betula* and Poaceae pollen were better represented in the analysed 16-year dataset. The most distinctive pattern in season-to-season variability was observed for *Betula* pollen where both sums of concentrations and average concentration showed biannual variability starting from 2006, with high values in even and low values in odd years. Time series of measured pollen concentrations revealed notable variability in magnitude and temporal distribution of peaks, which is the most pronounced in *Betula* (Supplementary material Figure S1), followed by Poaceae (Supplementary material Figure S2) and *Ambrosia* (Supplementary material Figure S3). *Betula* and *Ambrosia* pollen signals resemble Gaussian shape, with most of the pollen recorded during about one-third of the analysed period. On the other hand, the Poaceae season was longer, with notable amounts of airborne Poaceae pollen recorded during the entire analysed period. The only exception was a distinctive peak at the very beginning of the season.

Statistically significant positive correlations ($p < 0.05$) between measured daily pollen concentrations and values calculated by standard calendar model were observed for all pollen types except for *Betula* in 2013 (Table 3). On average, the highest intensity of correlation was measured for *Ambrosia* airborne pollen ($\rho = 0.91$) followed by Poaceae ($\rho = 0.78$) and *Betula* ($\rho = 0.63$). In regard to predicting the magnitude of pollen concentrations, a major discrepancy for all pollen types was observed in 2000, when the overall pollen concentrations were rather low. Over the entire analysed period, the highest average NRMSE was calculated for *Betula* pollen (NRMSE = 3.96) followed by *Ambrosia* (NRMSE = 1.38) and Poaceae (NRMSE = 1.26). Using median instead of mean for producing standard calendar model did not notably influence average Spearman's ρ, but it did lower NRMSE of *Betula* pollen forecast (NRMSE = 2.77). Conversely, the difference was not so pronounced for Poaceae (NRMSE = 1.21) and *Ambrosia* (NRMSE = 1.48) pollen forecasts.

Pre-processing the calibration dataset with moving mean or moving median did not significantly influence the Spearman's ρ (Table 3). For the best advanced calendar model, the highest increase (approximately 10%) was observed when predicting *Betula* pollen ($\rho = 0.69$), while the values for Poaceae and *Ambrosia* remained unchanged at 0.78 and 0.91, respectively. The largest observed effect was in the reduction in NRMSE (Table 3). The 20-day MDMD model performed the best in predicting Betula pollen concentrations, while 18- and 11-day MDMN models performed best in predicting Poaceae and Ambrosia pollen concentrations, respectively (Fig. 1). In addition to the decrease in average NRMSE, advanced calendar models tended to under-represent measured concentrations in season's peak (Supplementary material Figures S1, S2, S3).

As the number of calibration years increases, the performance of the models improves, with respect to both Spearman's ρ and NRMSE. The critical number of years that produces the largest improvement (about 30% comparing to the single calibration year) is four. Including additional years above this value produces less pronounced improvements in model performance, in terms of both correlation coefficient and prediction of the magnitude (Fig. 2).

The results of independent-samples Kruskal–Wallis test showed that there was not a statistically significant difference in average Spearman's ρ (Fig. 3a–c) between calendar models developed for forecasting daily concentrations of airborne *Betula* ($p = 0.614$), Poaceae ($p = 0.944$) and *Ambrosia* ($p = 0.829$) pollen. Also, the improvement in NRMSE (Fig. 3d–f) when forecasting daily concentrations of airborne *Betula* ($p = 0.353$), Poaceae ($p = 0.669$) and *Ambrosia* ($p = 0.759$) pollen was not statistically significant.

Calendar models for Niš and Sombor yielded very similar values of NRMSE and Spearman correlation coefficient to those achieved in Novi Sad (Table 4). Similarly to the Novi Sad data, pre-processing with moving mean window proved to be optimal in terms of

Table 3 Performance of the standard and the best advanced calendar model when compared to measured pollen concentrations of *Betula*, Poaceae and *Ambrosia* airborne pollen in Novi Sad

	Betula						Poaceae						*Ambrosia*					
	SC[a]		10-day SC[b]		AC[c]		SC[a]		10-day SC[b]		AC[d]		SC[a]		10-day SC[b]		AC[e]	
	ρ	NRMSE	ρ	NRMSE	ρ	NRMSE	ρ	NRMSE	ρ	NRMSE	ρ	NRMSE	ρ	NRMSE	ρ	NRMSE	ρ	NRMSE
2000	0.69	11.89	0.69	10.73	0.80	1.93	0.70	2.16	0.72	2.02	0.72	1.90	0.93	3.53	0.93	3.24	0.93	2.68
2001	0.79	8.49	0.80	7.32	0.58	2.07	0.66	1.35	0.63	1.33	0.65	1.33	0.93	1.86	0.93	1.94	0.93	1.91
2002	0.54	3.17	0.59	3.03	**0.18**	2.85	0.83	0.91	0.83	0.85	0.85	0.84	0.84	1.20	0.83	1.19	0.85	1.10
2003	0.88	1.64	0.91	1.65	0.91	1.90	0.81	0.94	0.80	0.91	0.78	0.81	0.92	1.03	0.91	0.92	0.92	0.96
2004	0.67	2.83	0.72	2.51	0.68	1.90	0.68	1.09	0.68	1.02	0.64	1.00	0.92	1.02	0.92	1.14	0.92	1.03
2005	0.59	3.23	0.52	3.01	0.59	2.46	0.82	1.08	0.82	0.95	0.82	0.91	0.87	2.07	0.87	2.16	0.86	1.73
2006	0.79	1.79	0.77	1.81	0.78	2.14	0.77	1.45	0.77	1.44	0.75	1.44	0.93	1.37	0.93	1.20	0.93	1.34
2007	0.78	1.70	0.83	1.31	0.81	1.75	0.80	1.03	0.83	0.86	0.85	0.83	0.93	1.06	0.94	1.03	0.94	0.95
2008	0.80	1.94	0.82	1.92	0.51	2.26	0.86	1.15	0.86	1.18	0.87	1.21	0.86	1.52	0.85	1.41	0.86	1.26
2009	0.48	4.55	0.51	3.84	0.76	2.56	0.80	1.10	0.82	1.02	0.83	0.95	0.92	1.39	0.92	1.35	0.95	1.15
2010	0.56	2.32	0.63	2.37	0.90	2.55	0.78	1.18	0.80	1.12	0.78	1.12	0.93	1.10	0.93	1.30	0.93	1.23
2011	0.42	3.39	0.49	3.19	0.79	3.51	0.84	1.45	0.82	1.50	0.82	1.53	0.91	1.27	0.88	1.25	0.89	1.30
2012	0.69	2.58	0.77	2.62	0.88	2.87	0.80	1.69	0.80	1.66	0.79	1.37	0.92	1.25	0.93	1.40	0.93	1.20
2013	**0.09**	7.60	**0.10**	6.94	0.39	3.31	0.84	1.14	0.84	1.16	0.85	1.05	0.93	1.06	0.93	1.08	0.93	1.10
2014	0.82	2.76	0.87	2.75	0.58	2.90	0.76	0.87	0.76	0.94	0.76	0.94	0.92	1.22	0.93	1.30	0.92	1.30
2015	0.53	3.52	0.58	3.36	0.83	3.32	0.79	1.53	0.81	1.61	0.81	1.61	0.90	1.79	0.91	1.92	0.90	1.84
Average	0.63	3.96	0.66	3.65	0.69	2.52	0.78	1.26	0.79	1.22	0.79	1.18	0.91	1.48	0.91	1.49	0.91	1.38
Std. Dev.	0.20	2.89	0.20	2.53	0.20	0.56	0.06	0.34	0.06	0.34	0.07	0.32	0.03	0.63	0.03	0.58	0.03	0.45

Correlations that were not statistically significant ($p > 0.05$) are written in bold letters)

[a]Standard calendar—mean

[b]Standard calendar—mean at 10-day temporal resolution (Spieksma and Wahl 1991)

[c]Advanced calendar—median from 20-day moving median

[d]Advanced calendar—median from 18-day moving mean

[e]Advanced calendar—median from 11-day moving mean

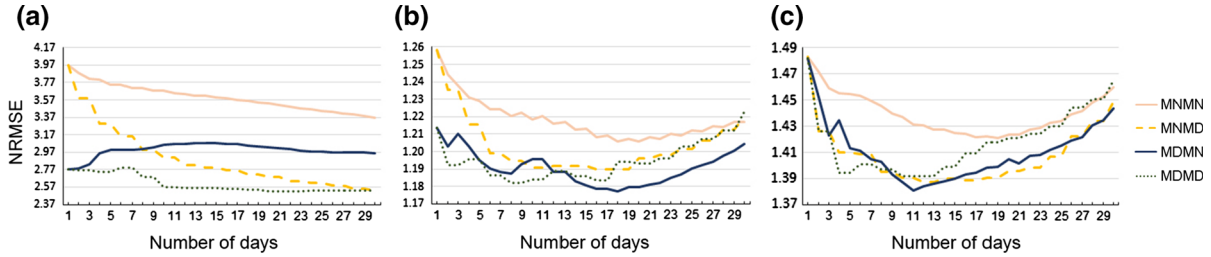

Fig. 1 Normalised root-mean-square error (NRMSE) of the advanced calendar models calibrated on 15-year dataset in relation to number of days used for calculating moving mean or moving median: **a** *Betula*, **b** Poaceae and **c** *Ambrosia* airborne pollen. Values for 1 day correspond to standard calendar model. Values for 1 day correspond to standard mean and standard median calendar model

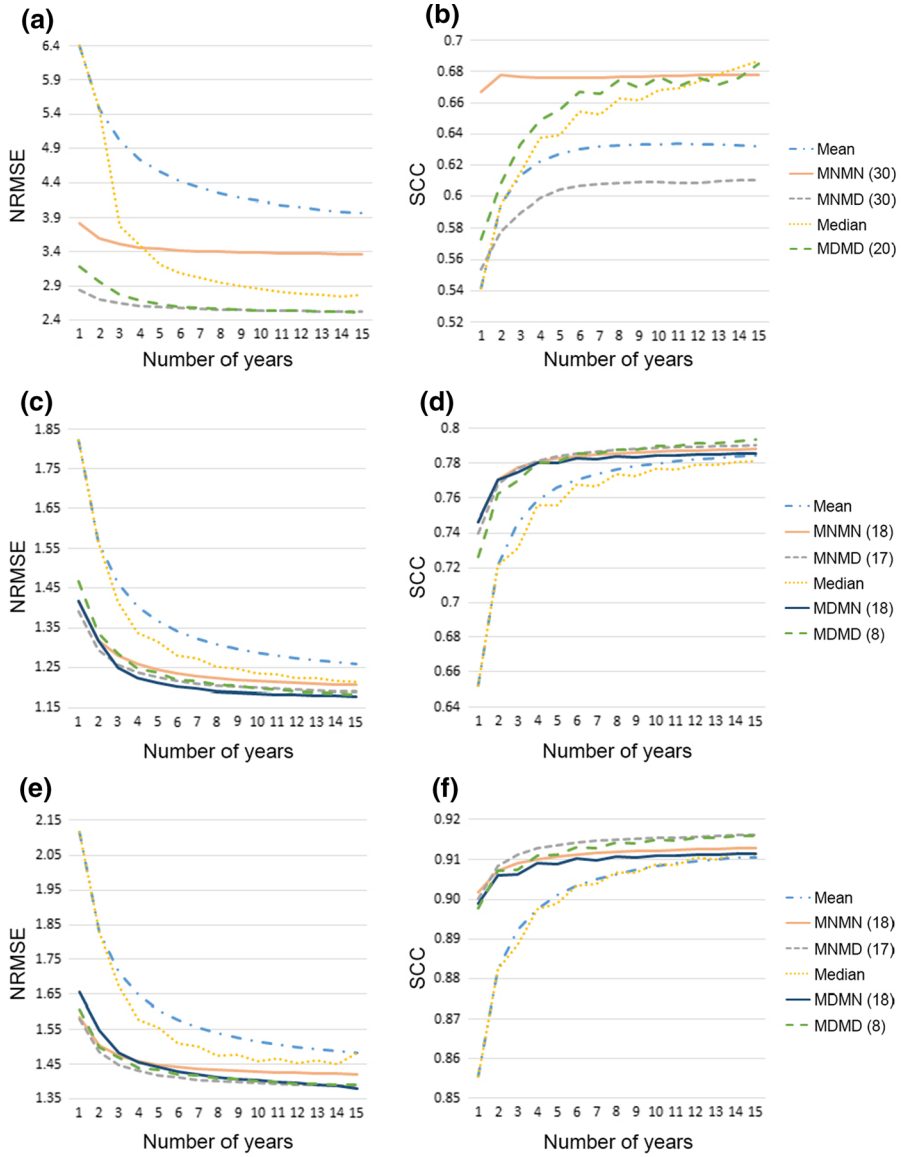

Fig. 2 Performance of calendar models depending on number of years available for training (measured by *NRMSE* normalised root-mean-square error and *SCC* Spearman's correlation coefficient) in predicting average daily concentrations of *Betula* (**a**, **b**), Poaceae (**c**, **d**) and *Ambrosia* (**e**, **f**) airborne pollen. Pre-filtering was performed by using previously determined optimal window length. Optimal window size for MDMN in *Betula* pollen prediction was 1. This makes optimal MDMN model equivalent to median, so it was omitted in graph **b**

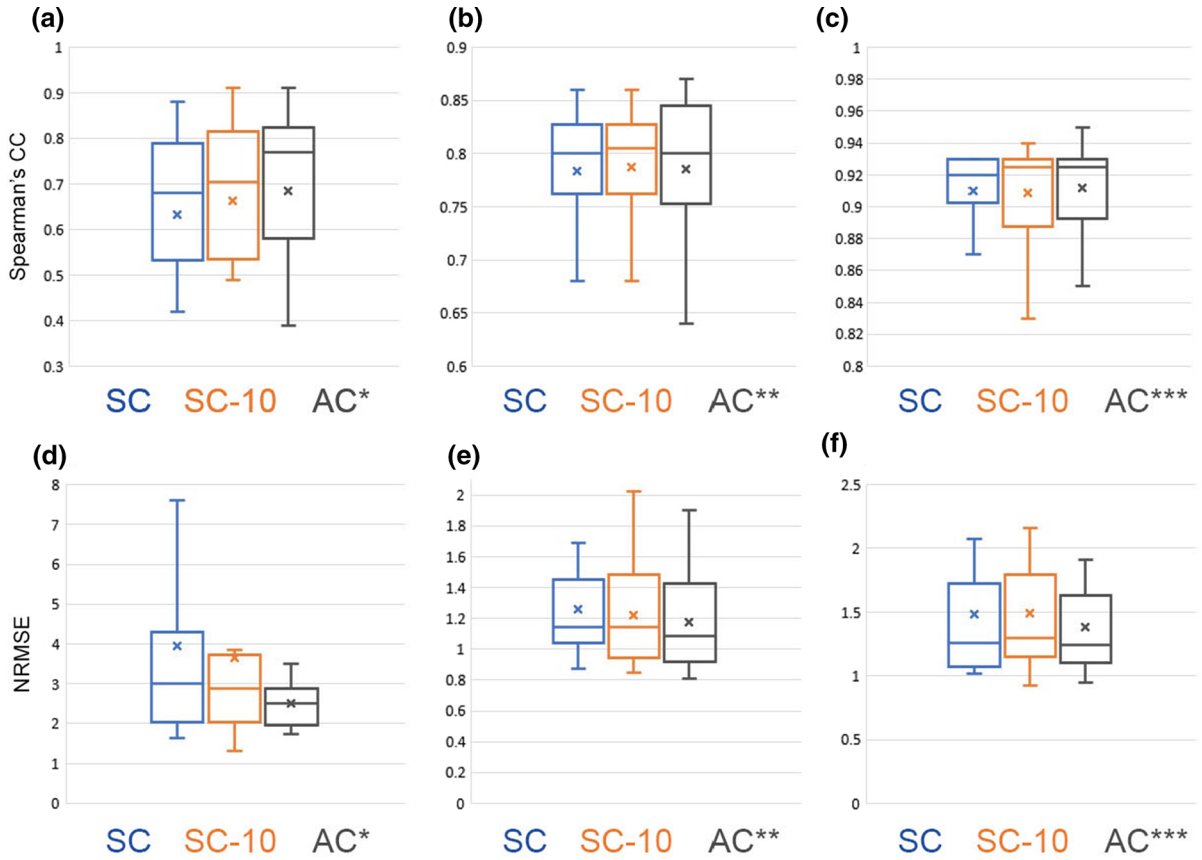

Fig. 3 Performance of the standard and the best advanced calendar model when compared to measured pollen concentrations of *Betula*, Poaceae and *Ambrosia* airborne pollen in Novi Sad. SC is the standard calendar (mean) model, SC-10 is the mean model at 10-day temporal resolution (Spieksma and Wahl 1991), and AC is the advanced calendar model. Upper and lower bounds of boxes are the upper and lower quartiles, the line in the middle is the median, the cross stands for the mean value, and the vertical line spans between minimum and maximum values. AC*—median from 20-day moving median. AC**—median from 18-day moving mean. AC***—median from 11-day moving mean

Table 4 Performance of calendar models in Niš and Sombor

Model		SC[a]		10-day SC[b]		Advanced calendar			
City	Pollen	ρ	NRMSE	ρ	NRMSE	Type	Window size	ρ	NRMSE
Niš	Betula	0.66	3.86	0.66	3.07	MNMN	24	0.68	3.42
	Poaceae	0.68	1.69	0.70	1.63	MNMN	11	0.70	1.63
	Ambrosia	0.83	2.13	0.82	1.94	MNMN	10	0.80	1.91
Sombor	Betula	0.64	4.55	0.63	4.21	MNMN	27	0.65	3.95
	Poaceae	0.87	1.57	0.87	1.52	MNMN	15	0.82	1.46
	Ambrosia	0.93	1.04	0.93	1.03	MNMN	13	0.80	0.98

MNMN mean of moving mean signals

[a]Standard calendar—mean

[b]Standard calendar—mean at 10-day temporal resolution (Spieksma and Wahl 1991)

NRMSE, but using the mean of previous years performed better than taking the median, due to the smaller number of years in the calibration dataset. Although the optimal window sizes are not the same in three different locations, it should be noted that their order was preserved. The optimal window size for *Betula* was the largest, followed by Poaceae and *Ambrosia*, mainly because of the variations in the beginning and the end of their pollen seasons. These variations were most pronounced in *Betula*, followed by Poaceae, and lastly *Ambrosia*, with the latter having the most stable season boundaries of the three analysed pollen types.

4 Discussion

4.1 Characteristics of *Betula*, Poaceae and *Ambrosia* pollen

The levels of *Betula* pollen measured in Novi Sad were well below the corresponding measurements in Northern Europe, but still notably higher than in the Mediterranean region (Skjøth et al. 2013). The biannual alternation of high and low seasons reflects the known biannual distribution of mast years in birch (Dahl et al. 2013 and references therein). As in many European regions, grass pollen season in Novi Sad results from flowering of different grass species and is therefore longer than *Betula* and *Ambrosia* seasons. Due to notable contribution of wheat crops in the region, the highest concentrations of Poaceae pollen were recorded in May and June. Novi Sad is situated on the Pannonian Plain, which is in the continental climatic zone (Lalic et al. 2011). Cold winters limit grass pollen occurrence to spring and summer, but at the same time, hot summers limit the amounts of pollen released, which is clearly observable in pollen indices measured in Mediterranean Europe (e.g. Spain and Greece) (Skjøth et al. 2013). The Pannonian Plain is one of the centres of European distribution of *Ambrosia artemisiifolia* L. (Smith et al. 2013). Therefore, *Ambrosia* pollen is regularly recorded in the air from July to October in the amount that exceeds values recorded in other parts of Europe (Skjøth et al. 2013).

4.2 Interpolation

There is a trade-off between the benefits of interpolating missing datasets and the issues that must be accommodated. Interpolation of intermittent missing values can be useful, as it improves the training dataset, thus making the model more robust. However, in regions where the infrequent transportation of pollen from distant sources causes sudden peaks, interpolation will not be able to adequately capture such changes. The tests confirmed that the *last-observation-carried-forward* method, which repeats the latest observation without any attempt to follow the signal's shape, is not appropriate for the interpolation of data gaps in the pollen concentration signal characterised by magnitude changes (e.g. sudden decreases during peak flowering due to rain). The second interpolation method taken into consideration was cubic spline method. While it is possible to achieve a smooth and accurate approximation of a given signal, the performance was poor when interpolating short gap intervals due to the exhibition of over-oscillatory behaviour (Surhone et al. 2010). The described tests indicated that linear interpolation was the most suitable among the examined approaches for completing short data gaps in the aerobiological dataset. Linear interpolation has been previously used for dealing with data gaps in aerobiology (Hilaire et al. 2012) and also to complete short gaps in meteorological datasets (Prank et al. 2013).

4.3 Performance of pollen calendars

Pollen calendars are commonly used in aerobiology as graphical representations of airborne pollen distribution in the study area and are used for comparing the shape of the seasons. Their ability to enable season comparison is limited by the methodology applied for their development. Pre-processing of the dataset affects the shape of the pollen signal, and thus, comparison of season intensity by using advanced calendar models is possible only if the same window length is applied. For calendars presented as 10-day means (Spieksma and Wahl 1991), comparison of season intensities is possible if the division of classes is the same. Common application of exponential division (Melgar et al. 2012) simplifies the comparisons, but using the same value for 10-day periods eliminates the variability, which is important for

personal assessment of dose-response for pollen-sensitive individuals. It is argued that both the standard and advanced calendars tested in this study can be used as models for forecasting daily concentrations of airborne pollen, if developed at daily temporal resolution.

It is difficult to compare the developed model to other regional models, since observation-oriented models are developed and evaluated for a limited territory. Unless these models adequately describe the emission, dispersion and deposition processes, it is not reasonable to apply them in other regions. It is common practice within the field to use categories (e.g. Stark et al. 1997; Sanchez-Mesa et al. 2002; Makra and Matyasovszky 2011; Castellano-Méndez et al. 2005) or statistical scores (e.g. Zink et al. 2012, 2017) rather than daily concentrations to test the performance of their prediction models. Nevertheless, there is no consensus on how to evaluate model performance in the prediction of exact daily pollen concentrations. Pearson's or Spearman's correlation coefficients are typically used in applications where the relationship between the two signals is of chief concern. However, the performance of magnitude prediction is evaluated using a range of indicators such as the coefficient of linear regression, mean-squared error, root-mean-square error, mean absolute error and the simple statistical comparison of the averages of two signals.

On average (over the range of validation years) calendar models yielded correlation coefficients similar to those seen in other studies that used advanced statistical and machine learning methods (Inatsu et al. 2014; Ritenberga et al. 2016). Due to regional-oriented assessment of its performance (Siljamo et al. 2012) calendar and numerical dispersion models cannot be directly compared when predicting daily *Betula* pollen concentrations. Multi-model ensemble simulations performed slightly better with respect to correlation coefficient and magnitude prediction (Sofiev et al. 2015). For Poaceae pollen, the correlations obtained by standard and advanced calendar models were within the range obtained by observation-oriented models that used meteorology data as an input (Stach et al. 2008; Voukantsis et al. 2010; Rojo et al. 2016; Rodríguez-Rajo et al. 2010; Hilaire et al. 2012). Regarding magnitude, the calendar models have worse performance in comparison with machine learning models that include meteorological parameters as predictors (Rodríguez-Rajo et al. 2010; Voukantsis et al. 2010). For *Ambrosia* pollen, the calendar models yielded a high correlation coefficient which was notably higher than the value obtained by multiple linear regression (Howard and Levetin 2014). Artificial neural network observation-oriented models that included meteorology (Csépe et al. 2014) were better, in respect of both correlation coefficient and predicting the magnitude of daily pollen concentrations. However, calendar models give predictions for more than 1 day ahead which makes them advantageous. Evaluation of SILAM operational numerical prediction model for *Ambrosia* focused on predicting total season amounts rather than daily concentration time series (Prank et al. 2013). Tests of COSMO-ART demonstrate good performance at individual observation sites, as well as regionally, which is evaluated based on statistical scores which do not include correlation and RMSE (Zink et al. 2017).

The observed year-to-year variability in recorded amounts of all analysed pollen types (as a result of either masting in birch or different management regimes in grass and ragweed) justifies the production of separate calendar models for below- and above-average seasons, which would improve forecasting of the magnitude of pollen concentrations. This requires an algorithm which determines in advance what kind of season is expected (below or above average). Without such algorithm, the application of multiple models is limited.

It is difficult to evaluate the performance of calendar models with respect to their application for management of allergic symptoms. For the prediction of the start and the end of airborne pollen occurrence season all models performed reasonably well. The biggest errors in prediction were observed for *Betula* for both performance measures, with the years 2013 and 2015 being the worst in terms of Spearman's correlation coefficient and NRMSE, respectively (Table 3). Flowering commences in spring which results in sensitivity to the extension of the preceding winter that may cause the extension of dormancy due to later achievement of forcing temperatures (Myking 1999). In order to address shifts in flowering times expected from the ongoing climate change, calendars have to be evaluated and updated on a regular basis. The biggest errors in predicting the magnitude of daily values for all models were observed at high concentrations during the most intensive part of the season.

Thresholds for the onset of allergic symptoms are expected to be at much lower concentrations for the majority of sensitive individuals (de Weger et al. 2013 and references therein), and thus, wrong predictions above the thresholds are not expected to significantly affect application of calendar models in management of the allergy.

It is believed that correlation coefficients overestimate the ability of pollen calendars to predict timing and intensity of sudden peaks. The large portion of achieved correlations tends to result from the overall shape of the pollen curve, as it emulates the pre-peak increase and the post-peak decrease, rather than day-to-day variability. It is argued that the time series of pollen concentration represents strongly non-stationary and non-ergodic processes due to the existence of the start and the end of the seasons (Sofiev et al. 2015; Ritenberga et al. 2016), which implies that NRMSE and correlation can be computed only within the main season. Despite limiting analysis to coincide with atmospheric pollen measurements, the effect of using concentration measurements from different fractions of each month may influence the correlation.

The results shown in this work can be interpreted in different manners. If correlation coefficients are the accuracy measure of choice, advanced calendar models did not manage to improve the prediction. However, if NRMSE is observed, it can be concluded that the advanced models brought a significant improvement to observation-based prediction, as they lower the error for as much as 40%, witnessed in birch pollen prediction. The opposing conclusions which can be drawn from this research raise a question of the applicability of various statistical measures in the domain of aerobiology. They can be used for comparison between the models, but none of them can impartially quantify the applicability of a model in aerobiological forecasting. Furthermore, its practical value for the end-users of forecasting systems is debatable. A possible solution for the problem would be to start from the end-users' point of view and observe the effects that different pollen concentrations have on them. In this way it would be possible to determine the exact amount of error that can be tolerated. Also, positive and negative errors could be treated differently, so that the effects of under- and overestimation could be separately analysed.

Observed variability in the magnitude and timing of concentration peaks is common in aerobiology. Previous studies have shown that atmospheric movement (i.e. speed and direction) and precipitation are responsible for day-to-day variability in airborne pollen concentrations (Barnes et al. 2001; Rodriguez-Rajo et al. 2003), and therefore, taking these environmental factors into consideration is required for further improvement of model performance.

Observation-oriented models are developed and evaluated for a limited territory and based on point receptors (i.e. pollen and meteorology). Therefore, the use of these models outside the region they were calibrated for is unpredictable (Sofiev et al. 2013) and depends on the model's ability to mechanistically describe natural processes behind records of airborne pollen. This is particularly emphasised for pollen calendars that do not take into consideration atmospheric conditions, which are known drivers of pollen emission and dispersion. For areas with strong and widespread local sources and where local transport dominates over regional and long-distance transport, pollen calendars based on diverse calibration dataset are expected to give representative forecasts even for larger geographical areas. It should be noted here that many authors highlighted the importance of previously recorded concentrations for short-term forecasting of *Betula* (Inatsu et al. 2014), Poaceae (Stach et al. 2008; Rodríguez-Rajo et al. 2010) and *Ambrosia* (Matyasovszky and Makra 2011; Csépe et al. 2014). Using auto-regression methods to push the model towards observations, standard calendar models are a good tool for setting the benchmark that will show to what extent adding meteorology among predictors increases model capability to predict processes and thus be more applicable in different regions.

5 Conclusions

In order to maximise the performance of short-term forecasts of airborne pollen concentrations, the models need to account for processes behind emission and dispersion. Although calendars do not consider these, the simple calendar model, produced as the mean of concentrations measured in the calibration dataset, describes well the shape of the pollen curve yielding correlation coefficients comparable to those obtained from more advanced observation-oriented models. The performance of forecasting the magnitude of pollen concentrations improves if median is used

instead of mean. Statistical analysis confirmed that with respect to difference in Spearman's ρ and NRMSE all calendars are equally suitable for forecasting daily concentrations of *Betula*, Poaceae and *Ambrosia* pollen over the course of the season. However, it is argued that either the standard calendar or advanced calendar is used instead of calendars produced following Spieksma's model (Spieksma and Wahl 1991), as these present day-to-day variations in pollen concentrations. Another conclusion is that preprocessing the calibration dataset or decreasing the temporal resolution causes the models to underrepresent peaks and day-to-day variability, thus limiting their use in analysis of dose-responses in allergy management. In any case, calendar models are the method of choice when meteorological data are not available and could also serve as the benchmark to assess whether introduction of meteorology in predictions improves performance and thus contributes to the description of processes behind emission and dispersion.

Acknowledgements This work was partly financed by: Ministry of Education, Science and Technological Development of Republic of Serbia (project numbers OI173002 and III44006), IPA Cross-border Cooperation programme Croatia – Serbia 2014-2020 (RealForAll project no. 2017HR-RS151), Provincial secretariat for finances, Autonomous Province Vojvodina (contract no. 102-401-337/2017-02-4-35-8) and the Swiss National Science Foundation (SCOPES JRP no. IZ73Z0_152348).

References

Altiparmak, F., Dengiz, B., & Smith, A. E. (2009). A general neural network model for estimating telecommunications network reliability. *IEEE Transactions on Reliability, 58,* 2–9.

Astray, G., Rodríguez-Rajo, F. J., Ferreiro-Lage, J. A., Fernández-González, M., Jato, V., & Mejuto, J. C. (2010). The use of artificial neural networks to forecast biological atmospheric allergens or pathogens only as Alternaria spores. *Journal of Environmental Monitoring, 12,* 2145–2152.

Barnes, C., Pacheco, F., Landuyt, J., Hu, F., & Portnoy, J. (2001). The effect of temperature, relative humidity and rainfall on airborne ragweed pollen concentrations. *Aerobiologia, 17,* 61–68.

Bornn, L., & Zidek, J. V. (2012). Efficient stabilization of crop yield prediction in the Canadian Prairies. *Agricultural and Forest Meteorology, 152,* 223–232.

Bousquet, P.-J., Chinn, S., Janson, C., Kogevinas, M., Burney, P., & Jarvis, D. (2007). Geographical variation in the prevalence of positive skin tests to environmental aeroallergens in the European Community Respiratory Health Survey I. *Allergy, 62,* 301–309.

Cadman, A. (1990). Airspora of Johannesburg and Pretoria, South Africa, 1987/88: I. Pollen calendars. *Grana, 29,* 311–317.

Calderón-Ezquerro, M. C., Guerrero-Guerra, C., Martínez-López, B., Martínez-Arroyo, A., & Trigo-Pérez, M. M. (2016). First airborne pollen calendar for Mexico City and its relationship with bioclimatic factors. *Aerobiologia, 32,* 225–244.

Castellano-Méndez, M., Aira, M. J., Iglesias, I., Jato, V., & González-Manteiga, W. (2005). Artificial neural networks as a useful tool to predict the risk level of Betula pollen in the air. *International Journal of Biometeorology, 49,* 310–316.

Chai, T., Carmichael, G. R., Tang, Y., Sandu, A., Heckel, A., Richter, A., et al. (2009). Regional NOx emission inversion through a four-dimensional variational approach using SCIAMACHY tropospheric NO_2 column observations. *Atmospheric Environment, 43,* 5046–5055.

Csépe, Z., Makra, L., Voukantsis, D., Matyasovszky, I., Tusnády, G., Karatzas, K., et al. (2014). Predicting daily ragweed pollen concentrations using computational intelligence techniques over two heavily polluted areas in Europe. *Science of the Total Environment, 476–477,* 542–552.

Dahl, A., Galán, C., Hajkova, L., Pauling, A., Sikoparija, B., Smith, M., et al. (2013). The onset, course and intensity of the pollen season. In M. Sofiev & C.-K. Bergman (Eds.), *Allergenic pollen: A review of the production, release, distribution and health impacts* (pp. 29–70). Berlin: Springer.

de Weger, L. A., Bergman, C.-K., Rantio-Lehtimaki, A., Dahl, A., Buters, J., Dechamp, C., et al. (2013). Impact of pollen. In M. Sofiev & C.-K. Bergman (Eds.), *Allergenic pollen: A review of the production, release, distribution and health impacts* (pp. 161–215). Berlin: Springer.

El-Ghazaly, G., & Fawzy, M. (1988). Short communications pollen calendar of alexandria (Egypt), 1981–1982. *Grana, 27,* 85–87.

Emberlin, J., Mullins, J., Corden, J., Millington, W., Broke, M., Savage, M., et al. (1997). The trend to earlier Birch pollen season in the UK: A biotic response to changes in weather conditions? *Grana, 36,* 29–33.

Galán, C., Smith, M., Thibaudon, M., Frenguelli, G., Oteros, J., Gehrig, R., et al. (2014). Pollen monitoring: Minimum requirements and reproducibility of analysis. *Aerobiologia, 30,* 385–395.

Hilaire, D., Rotach, M. W., & Clot, B. (2012). Building models for daily pollen concentrations. The example of 16 pollen taxa in 14 Swiss monitoring stations. *Aerobiologia, 28,* 499–513.

Hirst, J. M. (1952). An automatic volumetric spore trap. *Annals of Applied Biology, 39,* 257–265.

Howard, L. E., & Levetin, E. (2014). Ambrosia pollen in Tulsa, Oklahoma: Aerobiology, trends, and forecasting model development. *Annals of Allergy, Asthma and Immunology, 113,* 641–646.

Inatsu, M., Kobayashi, S., Takeuchi, S., & Ohmori, A. (2014). Statistical analysis on daily variations of birch pollen

amount with climatic variables in Sapporo. *Sola, 10,* 172–175.

Jae-Won, O., Ha-Baik, L., Im-Joo, K., Seong-Won, K., Kang-Seo, P., Myung-Hee, K., et al. (2012). The revised edition of korean calendar for allergenic pollens. *Allergy Asthma Immunology Research, 4,* 5–11.

Kasprzyk, I. (2009). Forecasting the start of Quercus pollen season using several methods–the evaluation of their efficiency. *International Journal of Biometeorology, 53,* 345–353.

Kaya, Z., & Aras, A. (2014). Airborne pollen calendar of Bartin, Turkey. *Aerobiologia, 20,* 63–67.

Kok Ong, E., Bir Singh, M., & Bruce Knox, R. (1995). Seasonal distribution of pollen in the atmosphere of Melbourne: An airborne pollen calendar. *Aerobiologia, 11,* 51–55.

Lake, I. R., Jones, N. R., Agnew, M., Goodess, C. M., Giorgi, F., Hamaoui-Laguel, L., et al. (2017). Climate change and future pollen allergy in Europe. *Environmental Health Perspectives, 125*(3), 385.

Lalic, B., Mihailovic, D. T., & Podrascanin, Z. (2011). Future state of climate in Vojvodina and expected effects on crop production (in Serbian with abstract in English). *Field and Vegetable Crop Research, 48,* 403–418.

Makra, L., & Matyasovszky, I. (2011). Assessment of the daily ragweed pollen concentration with previous-day meteorological variables using regression and quantile regression analysis for Szeged, Hungary. *Aerobiologia, 27,* 247–259.

Makra, L., Matyasovszky, I., Thibaudon, M., & Bonini, M. (2011). Forecasting ragweed pollen characteristics with nonparametric regression methods over the most polluted areas in Europe. *International Journal of Biometeorolgy, 55,* 361–371.

Marko, O., Brdar, S., Panic, M., Lugonja, P., & Crnojevic, V. (2016). Soybean varieties portfolio optimisation based on yield prediction. *Computers and Electronics in Agriculture, 127,* 467–474.

Martínez-Bracero, M., Alcázar, P., Díaz de la Guardia, C., Gonzalez-Minero, J., Ruiz, L., Trigo Pérez, M. M., et al. (2015). Pollen calendars: A guide to common airborne pollen in Andalusia. *Aerobiologia, 31,* 549–557.

Matyasovszky, I., & Makra, L. (2011). Autoregressive modelling of daily ragweed pollen concentrations for Szeged in Hungary. *Theoretical and Applied Climatology, 104,* 277–283.

McKeen, S., Wilczak, J., Grell, G., Djalalova, I., Peckham, S., Hsie, E.-Y., et al. (2005). Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004. *Journal of Geophysical Research, 110,* 1–16.

Melgar, M., Trigo, M. M., Recio, M., Docampo, S., Garcıa-Sanchez, G., & Cabezudo, B. (2012). Atmospheric pollen dynamics in Munster, north-western Germany: A three-year study (2004–2006). *Aerobiologia, 28,* 423–434.

Myking, T. (1999). Winter dormancy release and budburst in *Betula pendula* Roth. and B. *pubescens* Ehrh. ecotypes. *Phyton, 39,* 139–146.

Norris-Hill, J. (1995). The modelling of daily Poaceae pollen concentration. *Grana, 34,* 182–188.

O'Rourke, M. K. (1990). Comparative pollen calendars from Tucson, Arizona: Durham vs. Burkard samplers. *Aerobiologia, 6,* 136–140.

Pathirane, L. (1975). Graphical determination of the main pollen season. *Pollen and Spores, 17,* 609–610.

Pauling, A., Rotach, M., Gehrig, R., & Clot, B. (2012). A method to derive vegetation distribution maps for pollen dispersion models using birch as an example. *International Journal of Biometeorology, 56,* 949–958.

Peden, D., & Reed, C. E. (2010). Environmental and occupational allergies. *Journal of Allergy and Clinical Immunology, 125,* S150–S160.

Piotrowska-Weryszko, K., & Weryszko-Chmielewska, E. (2014). The airborne pollen calendar for Lublin, central eastern Poland. *Annals of Agricultural and Environmental Medicine, 21,* 541–545.

Prank, M., Chapman, D. S., Bullock, J. M., Soler, J. B., Berger, U., Dahl, A., et al. (2013). An operational model for forecasting ragweed pollen release and dispersion in Europe. *Agricultural and Forest Meteorology, 182,* 43–53.

R Core Team. (2016). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Radišić, P. (2013) Polen kao pokazatelj kvaliteta životne sredine. PhD Thesis. University of Novi Sad, Faculty of Sciences.

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics, 2,* 21–33.

Ritenberga, O., Sofiev, M., Kirillova, V., Kalnina, L., & Genikhovich, E. (2016). Statistical modelling of non-stationary processes of atmospheric pollution from natural sources: Example of birch pollen. *Agricultural and Forest Meteorology, 226–227,* 96–107.

Rodríguez-Rajo, F. J., Astray, G., Ferreiro-Lage, J. A., Aira, M. J., Jato-Rodriguez, M. V., & Mejuto, J. C. (2010). Evaluation of atmospheric Poaceae pollen concentration using a neural network applied to a coastal Atlantic climate region. *Neural Networks, 23,* 419–425.

Rodriguez-Rajo, F. J., Jato, V., & Aira, M. J. (2003). Pollen content in the atmosphere of Lugo (NW Spain) with reference to meteorological factors (1999–2001). *Aerobiologia, 19,* 213–225.

Rojo, J., Rivero, R., Romero-Morte, J., Fernández-González, F., & Pérez-Badia, R. (2016). Modeling pollen time series using seasonal-trend decomposition procedure based on LOESS smoothing. *International Journal of Biometeorology*. https://doi.org/10.1007/s00484-016-1215-y.

Sanchez-Mesa, J. A., Galan, C., Martinez-Heras, J. A., & Hervas-Martinez, C. (2002). The use of a neural network to forecast daily grass pollen concentration in a Mediterranean region: The southern part of the Iberian Penisula. *Clinical and Experimental Allergy, 32,* 1606–1612.

Scheifinger, H., Belmonte, J., Celenk, S., Damialis, A., Dechamp, C., Garcia-Mozo, H., et al. (2013). Monitoring, modelling and forecasting of the pollen season. In M. Sofiev & C.-K. Bergman (Eds.), *Allergenic pollen: A Review of the production, release, distribution and health impacts* (pp. 71–126). Berlin: Springer.

Siljamo, P., Sofiev, M., Filatova, E., Grewling, L., Jäger, S., Khoreva, E., et al. (2012). A numerical model of birch pollen emission and dispersion in the atmosphere. Model

evaluation and sensitivity analysis. *International Journal of Biometeorology, 57,* 125–136.

Skjøth, C. A., Šikoparija, B., Jäger, S., & EAN. (2013). Pollen Sources. In M. Sofiev & K.-C. Bergmann (Eds.), *Allergenic pollen* (pp. 9–27). Berlin: Springer.

Skjøth, C. A., Smith, M., Sikoparija, B., Stach, A., Myszkowska, D., Kasprzyk, I., et al. (2010). A method for producing airborne pollen source inventories: An example of Ambrosia (ragweed) on the Pannonian Plain. *Agricultural and Forest Meteorology, 150,* 1203–1210.

Smith, M., Cecchi, L., Skjoth, C. A., Karrer, G., & Sikoparija, B. (2013). Common ragweed: A threat to environmental health in Europe. *Environment International, 61,* 115–126.

Sofiev, M., Belmonte, J., Gehrig, R., Izquierdo, R., Smith, M., Dahl, A., et al. (2013). Airborne pollen transport. In M. Sofiev & C.-K. Bergman (Eds.), *Allergenic pollen: A review of the production, release, distribution and health impacts* (pp. 127–159). Berlin: Springer.

Sofiev, M., Berger, U., Prank, M., Vira, J., Arteta, J., Belmonte, J., et al. (2015). MACC regional multi-model ensemble simulations of birch pollen dispersion in Europe. *Atmospheric Chemistry and Physics, 15,* 8115–8130.

Spieksma, F Th M, & Wahl, P. G. (1991). Allergenic significance of Artemisia (Mugwort) pollen. In G. D'Amato, F. Th, M. Spieksma, & S. Bonini (Eds.), *Allergenic pollen and pollinosis in Europe*. Oxford: Blackwell Scientific Publications.

Stach, A., Smith, M., Baena, J. C. P., & 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,* 323–332.

Stark, P. C., Ryan, L. M., McDonald, J. L., & Burge, H. A. (1997). Using meteorologic data to predict daily ragweed pollen levels. *Aerobiologia, 13,* 177–184.

Surhone, L. M., Tennoe, M. T., & Henssonow, S. W. (2010). *Runge's phenomenon* (p. 80). Saarbrücken: Betascript publishing.

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,* 5101–5111.

Zeileis, A., & Grothendieck, G. (2005). Zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software, 014*(6), 1–26.

Zink, K., Kaufmann, P., Petitpierre, B., Broennimann, O., Guisan, A., Gentilini, E., et al. (2017). Numerical ragweed pollen forecasts using different source maps: A comparison for France. *International Journal of Biometeorology, 61*(1), 23–33.

Zink, K., Vogel, H., Vogel, B., Magyar, D., & Kottmeier, C. (2012). Modeling the dispersion of *Ambrosia artemisiifolia* L. pollen with the model system COSMO-ART. *International Journal of Biometeorology, 56,* 669–680.

How to prepare a pollen calendar for forecasting daily pollen concentrations of *Ambrosia, Betula* and Poaceae?

Šikoparija, B.[1], Marko, O.[1*], Panić, M.[1], Jakovetić, D.[2], Radišić, P.[1]

[1]*BioSense Institute - Research Institute for Information Technologies in Biosystems, University of Novi Sad, Novi Sad, Serbia*

[2]*Department of mathematics and informatics, Faculty of Sciences Novi Sad, University of Novi Sad, Novi Sad, Serbia*

*Corresponding author: BioSense Institute - Research Institute for Information Technologies in Biosystems, University of Novi Sad, Novi Sad, Serbia, e-mail: oskar.marko@biosense.rs

Online Resource 1: R Code for Prediction of Pollen Concentrations based on Pollen Calendars

Instruction Manual

Forecasting daily airborne pollen concentrations is of great importance for management of seasonal allergies. In the paper "How to prepare a pollen calendar for forecasting daily pollen concentrations of Ambrosia, Betula and Poaceae" (Šikoparija et al., 2017) we explored performance of pollen calendar as the most basic observation-oriented model for predicting daily concentrations of *Ambrosia*, *Betula*, and Poaceae airborne pollen.

Different forecasting models were programmed and tested in R and the code is located at https://github.com/oskarmarko/Forecasting-pollen-concentrations.

1. Reading the database

Here is an example of the input data format:

Day of Year	**RSNIS1 2011**	**RSNIS1 2010**	**RSNIS1 2009**	**RSNIS1 2008**	**RSNIS1 2007**
	POAC	**POAC**	**POAC**	**POAC**	**POAC**
1	0.00	-1.00	0.00	0.00	0.00
2	0.00	-1.00	0.00	0.00	0.00
3	0.00	-1.00	0.00	0.00	0.00
4	0.00	-1.00	0.00	0.00	0.00
5	0.00	-1.00	0.00	0.00	0.00
…	…	…	…	…	…

Pollen concentrations are expressed in pg/m^3, while the unknown values are denoted with -1.

In *main*, *sheet_id* should be the number of the sheet associated with the pollen type of interest, as usually the measurements of different pollen types are located in multiple sheets. Before reading the database, please adjust the *XLConnect* function accordingly:

```
sheet_id = 3 #1 betu 2 poac 3 ambr

poll0 = data.matrix(readWorksheetFromFile("RSNIS.xls", sheet = sheet_id,
                    header = FALSE, startCol = 2,
                    startRow = 3, endCol = 6,
                    endRow = 368))
```

2. Interpolation of missing values

In this step, missing values denoted with -1 are imputed using linear interpolation.

```
poll = interpolatePoll(poll0)
```

3. Defining the season boundaries

Season boundaries should be set in order to realistically assess the classifiers. Using the whole year as the season will yield a very high correlation coefficient, as long streams of 0s at the beginning and the end of the year are correlated. The

```
#Season boundaries (start and end day of the season)
if (sheet_id == 1) {
  poll = poll[60:151,]
}

if (sheet_id == 2) {
  poll = poll[91:304,]
}

if (sheet_id == 3) {
  poll = poll[182:273,]
}
```

4. Model outputs

The following function are used for forecasting values for the next season. They have two arguments. The first is the pollen matrix, while the second one is the sliding window size (for advanced calendar models). The name of the function consists of two parts. The first (mean/med) is the measure used for averaging the concentrations in the previous years, while the second part (mean/med after PreFilt) is the measure used for averaging pollen signals within a year, using sliding windows.

Advanced calendar models

meanPreFiltMean(poll, wdw)
meanPreFiltMed(poll, wdw)
medPreFiltMean(poll, wdw)
medPreFiltMed(poll, wdw)

Standard calendar models

Note that the standard calendar models are derived as the mean or median of the previous years' concentrations without prefiltering with sliding windows. Using the sliding window of size 1 is equal to no prefiltering, so the mean and median models can be derived using the following calls.

For mean:
meanPreFiltMean(poll, 1) or meanPreFiltMed(poll, 1)

For median:
medPreFiltMean(poll, 1) or medPreFiltMed(poll, 1)

Spieksma model

spieksma(poll)

5. Model statistics

Statistics are calculated using the following functions. For NRMSE, *measure* parameter should be 0 and for Spearman correlation coefficient it should be 1.

Advanced calendar models

meanPreFiltMedStats (poll, wdw, measure)
medPreFiltMedStats (poll, wdw, measure)
meanPreFiltMeanStats (poll, wdw, measure)
medPreFiltMeanStats (poll, wdw, measure)

Standard calendar models

For mean:
meanPreFiltMedStats (poll, 1, measure) or meanPreFiltMeanStats (poll, 1, measure)

For median:
medPreFiltMeanStats(poll, 1, measure) or medPreFiltMedStats(poll, 1, measure)

Spieksma model

spieksma(poll, measure)

6. Choosing the best advanced calendar model

If using advanced calendar models, the optimal sliding window size should be chosen. For this reason, the user must first test different window sizes using *bestWindow* function. There can the maximal window size be specified (*max_size*). This means that all window sizes from 1 to *max_size* will be tested. Again, *measure* should be 0 for NRMSE and 1 for Spearman correlation coefficient.

w_statistics = bestWindow(poll, max_size, measure)

The statistics are outputted as a list and the individual models (mean prefiltered with mean, mean prefiltered with median, median prefiltered with mean and median prefiltered with median) can be accessed through:

w_statistics $mnmn
w_statistics $mnmd
w_statistics $mdmn
w_statistics $mdmd