Aerobiologia (2011) 27:247–259
DOI 10.1007/s10453-010-9194-7

ORIGINAL PAPER

Assessment of the daily ragweed pollen concentration with previous-day meteorological variables using regression and quantile regression analysis for Szeged, Hungary

László Makra · István Matyasovszky

Received: 5 August 2010 / Accepted: 9 December 2010 / Published online: 24 December 2010

Abstract Time-varying parametric linear and time-varying nonparametric regression models as well as a time-varying nonparametric median regression model are developed to predict the daily pollen concentration for Szeged in Hungary using previous-day meteorological parameters and the daily pollen concentration. The models are applied to rainy days and non-rainy days, respectively. The most important predictor is the previous-day pollen concentration level, and the only other predictor retained by a stepwise regression procedure is the daily mean global solar flux for rainy days and the daily mean temperature for non-rainy days. Although the variance percentage explained by these two predictors is higher for non-rainy (55.2%) days than for rainy (51.9%) days, the prediction rate is slightly better for rainy than for non-rainy days. Nonparametric regression yields substantially better estimates, especially for rainy days indicating a nonlinear relationship between the predictors and the pollen concentration. The explained variance percentage is 71.4 and 64.6% for rainy and non-rainy days, respectively. Concerning the mean absolute error, the nonparametric median regression provides the best estimate. The quantile regression shows that probability distribution of daily ragweed concentration is much more skewed for non-rainy days, while the more concentrated probability distribution for rainy days exhibits relatively stable ragweed pollen concentrations. The possible lowest limits of concentrations are also calculated. Under highly favorable conditions for peak concentrations, the pollen level reaches at least 350 grains m^{-3} and 450 grains m^{-3} for rainy and non-rainy days, respectively. These values again underline the excessive ragweed pollen load over the area of Szeged.

Keywords Time-varying linear regression · Nonparametric regression · Median regression · Quantile regression · Rainy days · Non-rainy days

L. Makra (✉)
Department of Climatology and Landscape Ecology, University of Szeged, P.O.B. 653, 6701 Szeged, Hungary
e-mail: makra@geo.u-szeged.hu

I. Matyasovszky
Department of Meteorology, Eötvös Loránd University, Pázmány Péter st. 1/A, 1117 Budapest, Hungary
e-mail: matya@ludens.elte.hu

1 Introduction

Ragweed originates in North America and has evolved in response to a dry climate and open environment. Among ragweed species, only seaside ragweed (*Ambrosia maritima*) is native in Europe (de Visiani 1842). The spread of ragweed in Europe began after the First World War (Comtois 1998). Seeds of different ragweed species were transferred to Europe from America by purple clover seed shipments and grain imports (Makra et al. 2005). Today ragweed is appearing in more and more

countries (Wopfner et al. 2005), its blooming lasts for a long time (in some regions for up to three months) (Wan et al. 2002), and it produces a lot of pollen (Fumanal et al. 2007). In Europe, highly polluted regions with ragweed pollen are the southern part of European Russia (Saar et al. 2000), Ukraine (Turos et al. 2009), and the Balkan Peninsula (Šikoparija et al. 2009). However, the three main regions in Europe infected by ragweed are the Carpathian Basin with peak values in Hungary (Makra et al. 2004, 2005), the Rhône-Alpes region (Laaidi et al. 2003) in France and the western part of the Po River Plain, i.e. southern Lombardy in Italy (Cecchi et al. 2006, 2007). The Carpathian Basin is unique in the sense that the peak values of ragweed pollen are the highest worldwide (Makra et al. 2005).

Clinical investigations have shown that the highly allergenic ragweed pollen is the main reason for the most widespread, most serious and most long-lasting pollinosis (Asero et al. 2006). After inhaling its pollen, the characteristic symptoms of pollinosis (coughing, sneezing, nasal discharge, inflammation of the mucous membranes of eyes, and nose) appear very quickly (Asero 2002). Hence, a prior knowledge of the levels of pollen in the air can be useful for making preparations for a future high pollen load, namely for the prevention and treatment of allergic symptoms (Jäger 2000; Bousquet et al. 2001).

The influence of meteorological parameters on the pollen concentration of different species has been studied by several authors (e.g. Giner et al. 1999; Galán et al. 2000, 2001; Jato et al. 2000; Makra et al. 2004). Temperature was found to be the most important meteorological parameter related to pollen counts (Galán et al. 2000; Peternel et al. 2006; Oh 2009). Humidity is another important parameter affecting pollen levels (Galán et al. 2000; Makra et al. 2004). However, the effect of humidity seems to be complex. Nightly relative humidity in excess of 60% adversely influences the pollen concentration during the day, but a relative humidity of over 80% in the morning is accompanied with increased pollen levels again (Giner et al. 1999). Galán et al. (2000) found the effect of humidity on the Urticaceae pollen is detrimental when heavy rainfall occurs in short bursts and beneficial when it comes from light rain showers over several days. Furthermore, it has been observed in Szeged that on the day following a rainy day (when the relative humidity is higher), the pollen concentration level can suddenly increase (Makra et al. 2004).

The strongest correlation between pollen concentration and meteorological parameters was obtained during non-rainy days (Fornaciari et al. 1992; Galán et al. 2000). Pollen levels on rainy days have scarcely been analyzed, as it is not very clear what role precipitation and relative humidity play here. Low Urticaceae pollen concentrations were recorded on days with heavy rainfall, while high pollen levels were observed on days with light rainfall (Galán et al. 2000). Also, ragweed pollen concentration levels (Peternel et al. 2006) as well as pollen grain levels from trees (pine, oak, alder, and birch), grasses and weeds (Japanese hop, sagebrush and ragweed) markedly decrease on rainy days (Oh 2009). To assess the daily pollen concentrations, previous-day meteorological parameters (e.g. Galán et al. 2000; Makra et al. 2004; Stennett and Beggs 2004) or those variables up to three days before (Stennett and Beggs 2004) are frequently used.

The previous-day pollen concentration has also been found to be a useful predictor. The influence of the pollen concentration for the 2 previous days (Galán et al. 2000) and the pollen concentration for the previous day (Galán et al. 2000; Angosto et al. 2005) has already been studied. According to Ruiz et al. (2008), the pollen concentration for the two previous days is the best predictive variable. The role of the mean seasonal variation of the daily pollen value and the average pollen value of the three preceding days were also found to be important (Makra et al. 2004).

Forecasting airborne pollen levels based on meteorological and pollen parameters is one of the most studied topics in aerobiology because of its crucial application in allergology. The commonly used tools for this problem are linear regression models (Fornaciari et al. 2002; Vázquez et al. 2003; Angosto et al. 2005; Smith and Emberlin 2005, 2006; Rodríguez-Rajo et al. 2005; Ruiz et al. 2008) and autoregressive models (ARIMA) (Rodríguez-Rajo et al. 2006; Ocana-Peinado et al. 2008). Other studies have used more advanced techniques such as neural network or neuro-fuzzy models (Ranzi et al. 2003; Sánchez Mesa et al. 2005; Aznarte et al. 2007). However, there is no evidence that these latter procedures perform better than the traditional techniques (Aznarte et al. 2007; Verma and Pathak 2009).

Another technique for describing and forecasting pollen concentrations is to use atmospheric transport

models (e.g. Helbig et al. 2004; Schueler and Schlüntzen 2006; Sofiev et al. 2006; Vogel et al. 2008). Such models have a big advantage compared to the statistical models as they consider both local release and dispersion due to meteorological conditions and long-range transport, while statistical models are, by their nature, limited to the area where they are developed. However, regular use of atmospheric transport models is limited at present as they require additional knowledge compared to statistical models. This includes an inventory, a validated phenological model and a model that parameterizes daily pollen release (Skjøth et al. 2010). This information is actually not available for ragweed, and the feasible solution for providing pollen forecasts are therefore statistical models.

The purpose of this paper is to develop time-varying linear regression and time-varying nonparametric regression models as well as a time-varying nonparametric median regression to predict the daily pollen concentration for Szeged in Hungary using previous-day meteorological parameters and the daily pollen concentration. The models developed are applied to rainy and non-rainy days, respectively. As an extension of the median regression, a nonparametric quantile regression is also applied. The quantile regression estimates threshold levels have the property that the chance of exceeding these thresholds does not go above some preselected probability value. While the linear regression model specifies the mean conditioned on predictors, the quantile regression specifies the conditional quantile. Since any quantile can be used, it is possible to gain an insight into the whole range of the conditional distribution of the predictand.

2 Data and methodology

2.1 Location and data

Szeged (46.25N; 20.10E), the largest settlement in South-eastern Hungary, is located at the confluence of the rivers Tisza and Maros (Fig. 1). The area is characterized by an extensive flat landscape of the Great Hungarian Plain with an elevation of 79 m AMSL. The city is the centre of the Szeged region with 203,000 inhabitants. The climate of Szeged belongs to Köppen's **Ca** type (warm temperate climate) with relatively mild and short winters and hot summers (Köppen 1931). The pollen content of the air was measured using a 7-day recording "Hirst-type" volumetric trap (Hirst 1952). The air sampler is located on top of the building of the Faculty of Arts at the University of Szeged some 20 m above the ground (Makra et al. 2008).

In order to estimate actual daily ragweed pollen concentrations, previous-day values of eight meteorological variables were considered including mean temperature, mean wind speed, mean relative humidity, mean global solar flux, mean atmospheric sea-level pressure, minimum temperature, maximum temperature and precipitation amount. These data were collected by a monitoring station (operated under the auspices of the Environmental and Natural Protection and Water Conservancy Inspectorate of the Lower Tisza Region, Szeged, Hungary) located in the downtown, near to one of the busiest crossroads of Szeged. The type of instruments measuring meteorological parameters at the station are as follows. Air pressure: PS 71-I; temperature and humidity parameters: HMP-10; global solar flux: RS 81-I; wind speed: WS-12H+ and precipitation: Lambrecht. Temperature and humidity parameters are measured 3 m above the surface, while wind speed and global solar flux are recorded at a height of 6 m above the ground level. These variables and previous-day ragweed pollen concentrations as candidate predictors were applied for the period July 15–October 15 for years 1997–2006. This interval covers most of the ragweed pollination period in Szeged using the criterion of Galán et al. (2001). Namely, the start (end) of the pollen season is the earliest (latest) date on which at least 1 pollen grain m^{-3} is recorded and at least 5 consecutive (preceding) days also have 1 or more pollen grains m^{-3}. The mean of this yearly varying period is selected for the ten-year period examined. Main characteristics of the ragweed pollen season at Szeged can be found in Table 1. The datasets were divided into two parts—rainy and non-rainy days—and an analysis was performed on these two subsets separately. A day was taken rainy if daily precipitation total is at least 0.1 mm and non-rainy otherwise.

2.2 Statistical methods

Let the daily pollen concentrations from July 15 to October 15 be denoted by y_i, $i = 0, 1, \ldots, n$ at times $t_0, t_1, \ldots, t_n$. These latter values are scaled from July

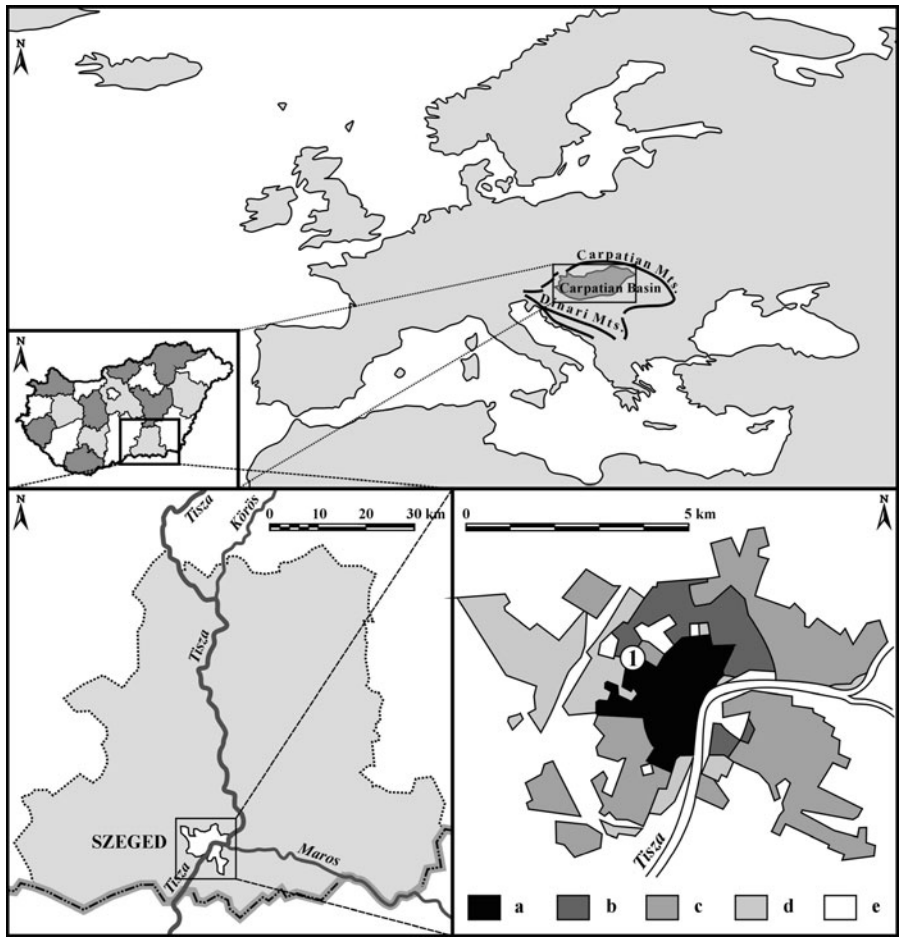

Fig. 1 Location of the Carpathian Basin (*upper large panel*), Csongrád county in Hungary (*upper panel*, *low left*), Szeged in Csongrád county (*low left panel*) and the urban web of Szeged (*low right panel*):[a: centre (2-to-4-story buildings); b: housing estates with prefabricated concrete slabs (5-to-10-story buildings); c: detached houses (1-to-2-story buildings); d: industrial areas; e: green areas, (1): monitoring station]

Table 1 Main characteristics of the ragweed pollen season

Year	Pollen season			Peak concentration		Total amount
	Start date	End date	Duration (day)	Pollen grains m^{-3}	Date	Pollen grains m^{-3}
1997	07.09.	10.19.	103	928	08.28.	7994
1998	07.26.	10.13.	80	332	09.06.	3857
1999	07.17.	10.21.	97	571	09.12.	8842
2000	07.02.	10.13.	104	608	09.04.	11592
2001	07.14.	10.10.	89	1125	08.26.	12277
2002	07.07.	10.13.	99	246	09.03.	4288
2003	07.26.	10.19.	85	404	08.27.	4760
2004	07.25.	10.12.	78	807	08.28.	6376
2005	07.15.	10.16.	92	603	09.04.	4957
2006	07.09.	10.18.	102	1385	09.06.	13854

15 for each particular year, i.e. $t_{u+vM} = t_u$, $u = 0, \ldots, M-1, v = 1, \ldots, N-1$, where $N = 10$ is the number of years and $M = 93$ is the length of the period examined for each year. Simultaneous values of $p+1$ number of predictors are written as $x_{ij}, i = 0, 1, \ldots, n, j = 1, \ldots, p+1$. Our

estimate $\hat{y}_i$. of y_i is defined by the time-varying linear regression

$$\hat{y}_i = \sum_{j=0}^{p} a_j(t_i) x_{ij}, \tag{1}$$

where $x_{i0} = 1$ and the rest of the predictors include the nine ($p = 9$) variables presented in Sect. 2.1. Based on the annual courses of both the pollen concentration and the predictors, the time-varying coefficients in Eq. (1) can be written as

$$a_j(t) = \alpha_{j0} + \sum_{k=1}^{2} \alpha_{jk} \cos(w_k t) + \beta_{jk} \sin(w_k t) \tag{2}$$

with $w_1 = 2\pi/365.25$ and $w_2 = 2w_1$. The role of the angular frequency w_1 is clear, while w_2 is introduced to describe the asymmetry of the annual course. Substituting Eq. (2) into Eq. (1), the ordinary least squares technique requires that we minimize the mean squared error

$$\text{MSE} = 1/n \sum_{i=1}^{n} \rho(y_i - \hat{y}_i) \text{ with } \rho(u) = u^2 \tag{3}$$

with respect to $\alpha_{jk}, j = 0, \ldots, p, k = 0, 1, 2, \beta_{jk}, j = 0, \ldots, p, k = 1, 2$. It might be expected that not every predictor should be included in the procedure to obtain a good estimator for the pollen concentration. The selection of optimal predictors is based on the well-known stepwise regression method (Draper and Smith 1981). However, our decision on the preserved variables is subjective as objective criteria used in stepwise regressions are applicable for independently and identically distributed samples, while our data viewed as values for random variables are neither independent nor identically distributed.

The least squares technique is a useful tool for data distributed nearly normally, but it is questionable with variables that have highly skewed distributions (e.g. pollen concentration data sets) and perhaps other estimation procedures should be applied. Here, Eq. (3) will be rewritten so as to minimize the mean absolute error

$$\text{MAE} = 1/n \sum_{i=1}^{n} \rho(y_i - \hat{y}_i) \text{ with } \rho(u) = |u|. \tag{4}$$

Note that solution of this problem approximates the conditional median instead of the conditional mean in the case of the least squares procedure. A further refinement of the methodology is introduced by rejecting the linear form in Eq. (1) and $\hat{y}_i = g\left(x_{i1}, \ldots x_{ip}, t_i\right)$, where g is an unknown function. Without any assumption on the analytical form of g, the conditional median can be estimated nonparametrically. This nonparametric technique consists of a non-linear local smoothing of data of the predictand. The locality, called bandwidth, is defined in the predictor space as a neighborhood of actual predictor variables, and those predictand values are involved in the smoothing that are accompanied with predictors lying in this neighborhood. Note that the nonparametric method can be applied to any data set satisfying some mild statistical dependence and smoothness conditions (Koenker 2005), but without obtaining an analytical functional form. A short description of the technique is in Appendix A. The bandwidth plays a crucial role in the accuracy of the procedure. Large bandwidths that allow large amounts of smoothing produce small variances with possibly large biases, while small bandwidths provide large variances with small biases. Thus, an optimal bandwidth that recognizes the trade-off between the bias and variance has to be estimated. Note that the concept of nonparametric median regression follows directly from the nonparametric regression model where the loss function $\rho(u) = u^2$ is used resulting in a linear local smoothing of data of the predictand (Cai 2007).

Koenker and Bassett (1978) recognized that the model of median regression can be extended to the quantile regression case. The τth quantile ($0 < \tau < 1$) of a random variable denotes that value below which the random variable takes values with probability τ. For instance, $\tau = 0.5$ corresponds to the median. While the linear regression model specifies the mean conditioned on predictors, the quantile regression specifies the conditional quantile. Since any quantile can be used, it is possible to gain an insight into the whole range of the conditional distribution of the predictand. The procedure is summarized in Appendix A.

Usually, a data set available is divided into a learning set and a validation set. The learning set is used to estimate parameters of the statistical model, and this model is then applied to the validation set. A general rule of thumb is to consider the learning set to be 80% of the total data and the validation set to be the remaining 20%. However, our data set is quite

short and it splits into two parts even worsens the case. Evidently, a "fortunate" choice of years for validation might provide too nice results, while an "unfortunate" choice might underestimate the potentials of the underlying statistical models. Note that the only parameter to be estimated in the nonparametric techniques is the bandwidth. Therefore, the validation should include just a proper bandwidth selection. Having N years of data, our validation makes it possible to use N-year validation set with $(N-1)$-year learning set. Taking the kth year from the entire data set, bandwidth is estimated with data omitting the kth year, and estimates for the kth year are then obtained using this bandwidth. The procedure is applied for $k = 1,\ldots N$, and thus these estimates for the entire data set are directly validated. However, a simplification working with the mean of annually varying bandwidths is made because the variability of the N number bandwidths is small. The time-varying parametric linear regression is not validated as it is used to have only preliminary results; the emphasis is on the nonparametric methods.

3 Results

3.1 Regression and median regression

The most important predictor is the previous-day pollen concentration, which accounts for 48.6 and 45.3% of the variance for rainy and non-rainy days, respectively. The only other predictor retained by a stepwise regression procedure is the daily mean global solar flux for rainy days and the daily mean temperature for non-rainy days (Table 2). The variance accounted for by these two predictors is 51.9 and 55.2%, respectively. Including the third most important predictor only gives an additional variance reduction of 0.3% (air pressure) and 0.7% (relative humidity) for rainy and non-rainy days, respectively. Ragweed likes dry and warm climates, with a tendency to the occasional drought, with abundant sunshine hours, small relative humidity and cloudiness during its pollination season, namely the summer and early autumn period (Béres et al. 2005). Hence, the importance of the temperature on non-rainy days is obvious. The relative prediction accuracy is higher for non-rainy days (55.2%) than for rainy (51.9%) days. However, as the residual variance is 6486 and 6058 (pollen grains m^{-3})^2 under the variance of 14477 and 12596 (pollen grains m^{-3})^2 for both cases, respectively, the prediction performance is slightly better for rainy days than for non-rainy days. These latter two variances are calculated as MSEs obtained from Eq. (1) with $p = 0$, i.e. as MSEs corresponding to the annual cycles of the pollen concentration on non-rainy and rainy days.

Applying nonparametric regression with the two predictors mentioned earlier produces substantially better estimates mainly for rainy days, clearly indicating that there is a nonlinear relationship between predictors and the pollen concentration. To be precise, the variance percentage accounted for by the predictors is 71.4 and 64.6% for rainy days and non-rainy days, respectively. A bootstrap test of Cai (2007) (see Appendix B) is used to check whether these improvements are significant when compared to parametric models. The test shows that nonparametric regression models have a gain at any reasonable significance levels (the test statistic obtained is substantially higher than the critical value at the 99% significance level). For each day examined, the explained variance percentage of 66.6% seems quite different from the 71.4% result for rainy days. However, the frequency of non-rainy days is around 2.3 times higher than the frequency for rainy days. The mean absolute error (MAE) provided by the nonparametric regression model is 26.8 pollen grains m^{-3} (25.6 pollen grains m^{-3} for rainy days and 27.2 pollen grains m^{-3} for non-rainy days). One might think that there is a contradiction here between the predictor selection based on linear regression and the nonlinear relationship between any predictor and the pollen concentration. Therefore, the predictor selection for nonparametric regression is repeated by minimizing the MSE with a technique similar to the stepwise regression. The procedure results in the

Table 2 Candidate predictors: previous-day pollen concentration (PC), previous-day mean temperature (T), previous-day mean wind speed (W), previous-day mean relative humidity (RH), previous-day global solar flux (GSF), previous-day mean sea-level pressure (SLP), previous-day minimum temperature (T_{min}), previous-day maximum temperature (T_{max}), previous-day precipitation amount (Pr)

Rainy day	**PC**	T	W	RH	**GSF**	SLP	T_{min}	T_{max}	Pr
Non-rainy day	**PC**	**T**	W	RH	GSF	SLP	T_{min}	T_{max}	Pr

Predictors involved in statistical models are in bold

same predictors obtained for time-varying linear regression.

When applying the median regression model, the MAE is expected to be lower (with a higher MSE) compared to the MAE obtained from the regression. Indeed, median regression gives an MAE value of 21.2 pollen grains m^{-3} (16.5 pollen grains m^{-3} for rainy days and 23.2 pollen grains m^{-3} for non-rainy days), which is 20.9% smaller than the MAE value obtained from nonparametric regression. Because prediction errors and not their squares are of practical interest, the nonparametric median regression model produces the best estimate among the prediction models discussed here. Figure 2 illustrates the prediction potential of applying the nonparametric median regression model for a ragweed infected area like Szeged.

3.2 Quantile regression

Generally, high pollen levels are of practical importance, so a quantile regression was performed with probability values $\tau = 0.5, 0.6, \ldots, 0.9$. Figures 3 and 4 show the annual cycles of these quantiles obtained after omitting the predictors (omitting the kernel K_p) (see Appendix A). Comparing the highest quantile values during the year, quantiles for rainy days are in general substantially smaller (principally for the 0.9th quantile) due to the less favorable conditions for pollen dispersion and the wash-out effect of the pollen grains from the air. But the 0.7th and 0.8th quantiles on rainy days are identical with or even

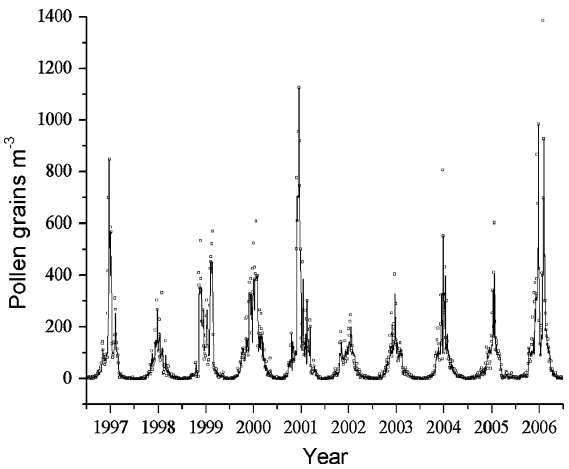

Fig. 2 The daily pollen concentration with its 1-day prediction (*solid*) obtained from nonparametric median regression

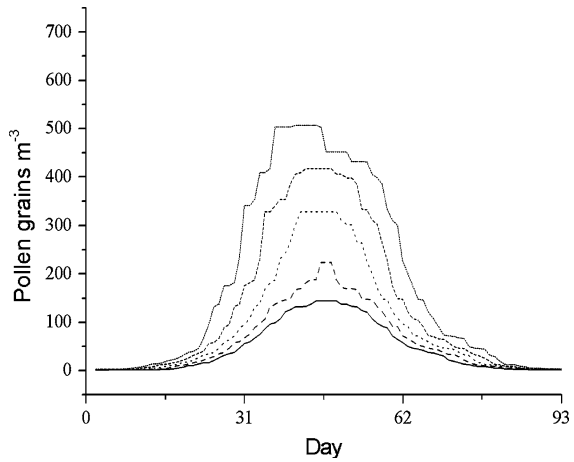

Fig. 3 The daily pollen concentration quantile trends for quantiles 0.5 (*solid*), 0.6 (*dashed*), 0.7 (*dots*), 0.8 (*short dashes*), 0.9 (*short dots*) for rainy days

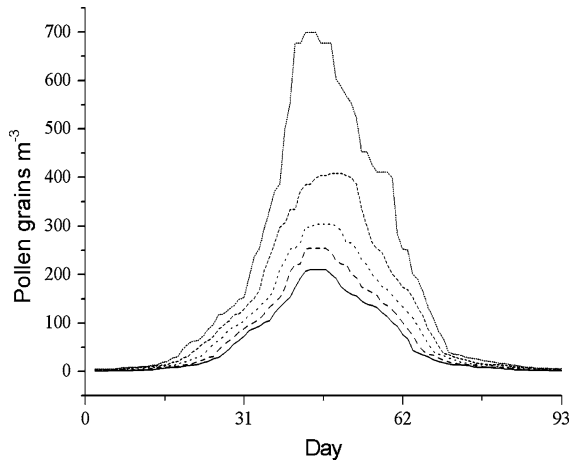

Fig. 4 The daily pollen concentration quantile trends for quantiles 0.5 (*solid*), 0.6 (*dashed*), 0.7 (*dots*), 0.8 (*short dashes*), 0.9 (*short dots*) for non-rainy days

larger than these quantiles on non-rainy days during most of the year. Therefore, the probability distribution of daily ragweed pollen concentration is much more skewed for non-rainy days, thus producing extremely high pollen levels, while the probability distribution for rainy days is more concentrated and gives relatively steady ragweed pollen concentrations. A weak tendency can also be observed that the highest values of the 0.9th quantile tend to occur earlier during the annual cycle for rainy days, while the highest values of the 0.8th quantile are liable to appear later for non-rainy days compared to the remaining quantiles. As ragweed generally likes dry

and warm climates with abundant sunshine (Béres et al. 2005) quantiles for non-rainy days are in general substantially larger due to the more favorable conditions. However, ragweed favors warm and rainy days at early stage of the pollen season as precipitation makes the plant development intensely (Béres et al. 2005). Therefore, rain accelerates the pollen emission reaching high pollen levels (high quantiles) earlier at the beginning of the pollen season. The curves are considerably smoother for non-rainy days, because the number of such days is around 2.3 times larger than the number of rainy days, and thus a considerably larger amount of data is available for the estimates.

The example given in Fig. 5 shows the $\tau = 0.75$th time-varying quantile conditioned on the predictors. The curve provides estimated pollen concentration thresholds exceeded with probability 0.25 under actual values of predictors. Evidently, quantile regression with other values of τ produces other pollen levels exceeded with probabilities $1 - \tau$, and thus different thresholds can be obtained for different probability levels. Unfortunately, the overall picture of pollen level quantiles conditioned on predictors and time cannot be shown here, as it would require four-dimensional curves.

A further question is what the lowest concentration limits that can be expected with different values of the chosen predictors are. The problem can be described using the concept of extremal quantile regression (Smith 1994), i.e. with quantile regression under $\tau = 0$. Main equations for the nonparametric extremal quantile regression are in Appendix A. The conditional extremal quantile is called the conditional boundary (Chernozhukov 2005) of the pollen concentration. Note that this conditional boundary can be calculated for any predictor values within a domain determined by the observed values. In order to illustrate the dependence of the extremal quantile on predictors, the date August 31 has been selected, as the highest pollen concentration levels are expected around late August—early September (see Table 1). A large number of data values (2,500) for the previous-day mean global solar flux (Fig. 6) and the previous-day mean temperature (Fig. 7) with the previous-day pollen concentration as predictors were generated within a domain defined by the smallest and largest values observed on that day. It is clear from Figs. 6 and 7 that rainy days are accompanied by a lower boundary. High (low) previous-day pollen levels imply high (low) actual boundaries, but the global solar flux on rainy days or temperature on non-rainy days also substantially influences this boundary because an increase in both quantities results in substantially raised boundaries. It is also apparent that the relationships found are not linear; in the case of linearity, the surfaces should be placed on planes. Conditions less favorable for pollination mean around 20 pollen grains m^{-3}, while high previous-day pollen

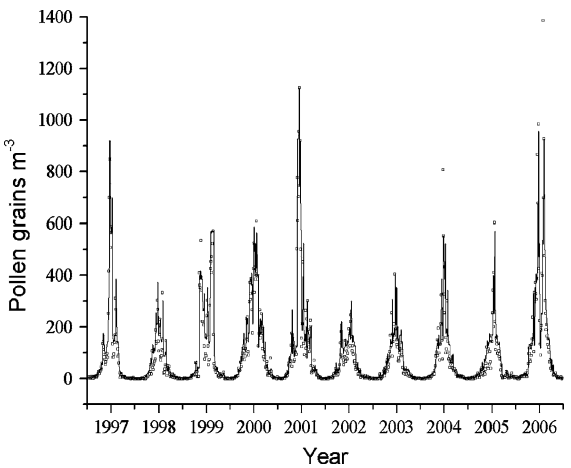

Fig. 5 The daily pollen concentration with its conditional quantile of 0.75

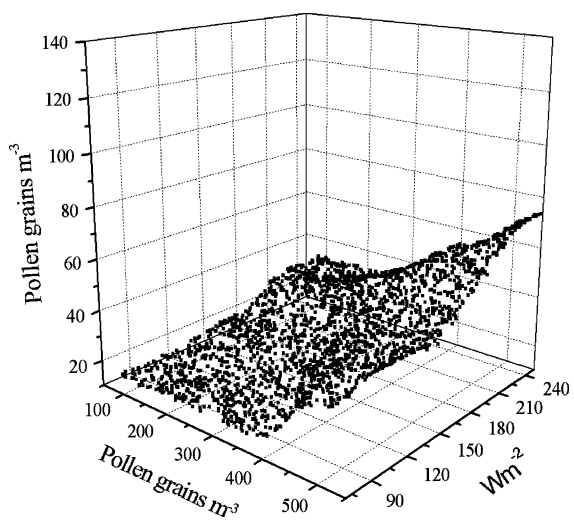

Fig. 6 Boundary of the daily pollen concentration level (pollen grains m^{-3}) for rainy days on August 31 conditioned on the previous-day pollen concentration (pollen grains m^{-3}) and the previous-day mean global solar flux (Wm^{-2})

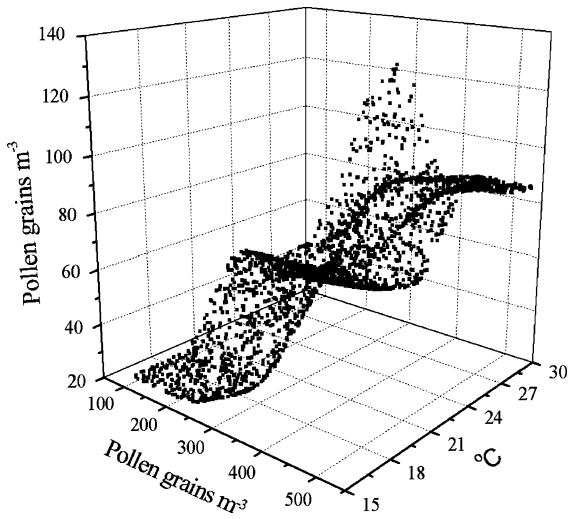

Fig. 7 Boundary of the daily pollen concentration level (pollen grains m^{-3}) for non-rainy days on August 31 conditioned on the previous-day pollen concentration (pollen grains m^{-3}) and the previous-day mean temperature (°C)

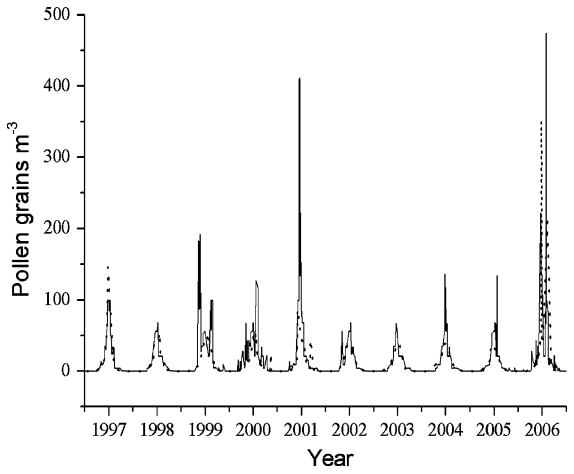

Fig. 8 The realized conditional boundary of the daily pollen concentration for non-rainy (*solid*) and rainy (*dotted*) days

concentrations with high global solar fluxes (rainy days) or with high temperatures (non-rainy days) result in just over 70 and 120 pollen grains m^{-3} for rainy days and non-rainy days, respectively. However, taking into account not just the above-mentioned day, a time sequence plot of the realized conditional boundary can also be obtained for the whole 10-year period (see Appendix A) with the observed values of predictors. Figure 8 tells us that the boundary can exceed even 350 and 450 pollen grains m^{-3} for rainy days and non-rainy days, respectively, under exceptionally favorable conditions for pollination. These pollen values again demonstrate how polluted the area of Szeged is by the ragweed plant.

4 Conclusions

Time-varying linear regression and time-varying nonparametric regression models as well as a time-varying nonparametric median regression were developed to predict the daily pollen concentration level for Szeged in Hungary using previous-day meteorological parameters and daily pollen concentration levels. The models were applied to rainy days and non-rainy days, respectively. Computations were carried out by computer programs developed by the authors.

The most important predictor found was the previous-day pollen concentration, which accounted for 48.6 and 45.3% of the variance for rainy days and non-rainy days, respectively. The only other predictor retained by the stepwise regression is the mean global solar flux for rainy days and the daily mean temperature for non-rainy days. The relative variance explained by these two predictors is higher for non-rainy days (55.2%) than for rainy days (51.9%). However, on examining the variances, the prediction rate is slightly better for rainy days than for non-rainy days. Including the third most important predictor only produces an additional variance reduction of 0.3% (air pressure) and 0.7% (relative humidity) for rainy days and non-rainy days, respectively. For non-rainy days, the role of the daily mean temperature is obvious: ragweed thrives best in a warm and dry climate. For rainy days, a small global solar flux means a low temperature that decreases the intensity of the pollen dispersion. The minimal role of the air pressure for rainy days is not surprising, but the small significance of the relative humidity for non-rainy days is quite apparent; the literature specifies this element as an important factor influencing daily pollen levels (Giner et al. 1999; Galán et al. 2000). It is to be mentioned that originally not only the previous-day pollen concentration and previous-day meteorological variables were taken into account as candidate predictors, but these variables were considered up to three previous days. However, the predictor selection procedure mentioned in Sect. 3.1 decided on predictors used in the paper.

Table 3 Root mean square error (RMSE) and mean absolute error (MAE) of the estimates and of the annual cycles (in parentheses) for the three statistical models (pollen grains m^{-3})

	Parametric linear regression		Nonparametric regression		Nonparametric median regression	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Rainy day	77.8 (112.2)	32.3 (58.4)	*61.7* (115.3)	25.6 (60.0)	63.2 (118.0)	**16.5** (36.8)
Non-rainy day	80.5 (120.3)	32.5 (54.8)	*67.5* (113.6)	27.2 (51.7)	68.8 (115.8)	**23.2** (50.3)
Every day	79.8 (117.9)	32.4 (55.9)	*66.1* (114.4)	26.7 (54.7)	67.1 (116.5)	**21.2** (46.2)

Bold and italic refer to the best estimate in view of MAE and RMSE, respectively

Applying nonparametric regression with predictors chosen above yields substantially better estimates mainly for rainy days, indicating a nonlinear relationship between the predictors and the pollen concentration. In particular, the variance percentage explained by the predictors is 71.4 and 64.6% for rainy days and non-rainy days, respectively.

The nonparametric median regression model provides a mean absolute error of 21.2 pollen grains m^{-3} (16.5 pollen grains m^{-3} for rainy days and 23.2 pollen grains m^{-3} for non-rainy days), which is substantially smaller than the mean absolute error obtained from regression. Because prediction errors and not their squares have a practical worth, the nonparametric median regression produces the best estimates among the prediction models mentioned earlier. These results are summarized in Table 3 showing the root of mean squared error (RMSE) and mean absolute error (MAE) corresponding to the three statistical models. Here, the term annual cycle refers to the case when the predictors are omitted from the models resulting in the estimation of the annual cycle of the mean, median, or quantiles (see Appendix A).

As regards the highest quantile values during the year, the quantile regression provides smaller quantiles (principally for the 0.9th quantile) for rainy days than for non-rainy days, which can be explained by less favorable conditions for the pollen dispersion and the wash-out effect of the pollen grains from the air on rainy days. The probability distribution of daily ragweed pollen concentration is much more skewed for non-rainy days that have the highest pollen levels, while the probability distribution for rainy days is more concentrated and results in relatively stable ragweed pollen concentration levels.

The lowest limits of possible concentrations under different values of the chosen predictors were also calculated. Rainy days, in contrast to the less favorable conditions for reaching peak concentration levels, produce over 350 pollen grains m^{-3}, while non-rainy days produce peak levels of over 450 grains m^{-3}. These values once again underline the very high ragweed pollen load over the area of Szeged.

Acknowledgments The authors would like to thank Miklós Juhász for providing pollen data of Szeged, and Zoltán Sümeghy for the digital mapping in Fig. 1. The European Union and the European Social Fund have provided financial support to the project under the grant agreement no. TÁMOP 4.2.1./B-09/1/KMR-2010-0003.

Appendix A

Nonparametric regressions

Having a data set $y_i, i = 1, \ldots, n$ for a predictand at times $t_1, \ldots, t_n$ and simultaneous values of p number of predictors written as $x_{ij}, i = 1, \ldots, n, j = 1, \ldots, p$, our estimate at given values of predictors $x_1, \ldots, x_p$ and time t is $\hat{g}(x_1, \ldots, x_p, t) = \hat{a}$ with $\hat{a}, \hat{b}_1, \ldots, \hat{b}_p$ that minimizes

$$1/n \sum_{i=1}^{n} \rho\left(y_i - (a + b_1(x_{i1} - x_1) + \cdots + b_p(x_{ip} - x_p))\right) \times K_p\left(\frac{x_{i1} - x_1}{h_1}, \ldots, \frac{x_{ip} - x_p}{h_p}\right) K\left(\frac{t_i - t}{h_t}\right)$$

with $\rho(u) = u^2$ for regression, $\rho(u) = |u|$ for median regression, furthermore with $\rho(u) = \rho_\tau(u) = (1 - \tau)|u|$ if $u < 0$, $\rho(u) = \rho_\tau(u) = \tau|u|$ if $u \geq 0$ for τ th quantile regression. Weights after the loss function ρ are generated by suitable kernel functions. K, the Epanechnikov kernel, is defined as $K(u) = 3/4(1 - u^2)$ in $[-1,1]$ and zero otherwise (Fan 1992),

while K_p is a p-variate kernel function. This latter kernel can be chosen as a multivariate normal density function using a suitable scaling of predictors such that $h_1 = h_2 = \cdots = h_p = h$ (Wand and Jones 1993). The bandwidths h and h_t playing a crucial role in the accuracy of the procedure is estimated by minimizing

$$1/n \sum_{i=1}^{n} \rho\left(y_i - \hat{y}_i^{(i)}\right)$$

with respect to h and h_t, where $\hat{y}_i^{(i)}$ is an estimate of y_i after omitting the ith data values from the estimation procedure (Yu and Jones 1998).

The nonparametric extremal quantile regression for $\tau = 0$ at given values of predictors $x_1, \ldots, x_p$ and time t is $\hat{g}(x_1, \ldots, x_p, t) = \hat{a}$ with $\hat{a}, \hat{b}_1, \ldots, \hat{b}_p$ that minimizes

$$1/n \sum_{i=1}^{n} \left(y_i - (a + b_1(x_{i1} - x_1) + \cdots + b_p(x_{ip} - x_p))\right) \times K_p\left(\frac{x_{i1} - x_1}{h}, \ldots \infty, \frac{x_{ip} - x_p}{h}\right) K\left(\frac{t_i - t}{b}\right)$$

such that $a + b_1(x_{i1} - x_1) + \cdots + b_p(x_{ip} - x_p) < y_i$ for every i satisfying $K_p\left((x_{i1} - x_1)/h, \ldots, (x_{ip} - x_p)/h\right) K((t_i - t)/b) > 0$.

Details of these methods are profoundly discussed e.g. in Fan and Yao (2005) and Koenker (2005). Note that omitting the predictors from these equations results in estimating the annual cycle of the mean, median, or quantiles.

Appendix B

Model comparison by a bootstrap technique

The null-hypothesis is that the time-varying parametric model is true. A comparison of this model with time-varying nonparametric model is based on the sums of squared residuals as

$$T = (\mathrm{SSE}_{np} - \mathrm{SSE}_p)/\mathrm{SSE}_p, \text{ with } \mathrm{SSE}_p = \sum_{t=1}^{n} r_{p,t}^2,$$

$\mathrm{SSE}_{np} = \sum_{t=1}^{n} r_{np,t}^2$, where $r_{np,t}$ and $r_{p,t}$ are the residuals of nonparametric and parametric models, respectively. The procedure can be summarized by the following five steps.

1. Form separate empirical distribution functions of $r_{np,t} - \overline{r_{np,t}}$ for every day of the year and generate a residual data set of length n using these distribution functions.
2. Generate a data set with the parametric model using the residual data obtained from step 1.
3. Apply the estimation procedure of nonparametric model to data set obtained from step 2, and calculate the actual T.
4. Repeat steps 1–3 L times resulting in $T_1, T_2, \ldots, T_L$, where L is a suitably large integer, say $L = 1000$.
5. Calculate $q_{1-\varepsilon}$, the $1 - \varepsilon$ quantile of the empirical distribution function of the sample $T_1, T_2, \ldots, T_L$. The critical value for accepting or rejecting the null-hypothesis will be $q_{1-\varepsilon}$, corresponding to a significance level of $(1 - \varepsilon)100\%$.

References

Angosto, J. M., Moreno-Grau, S., Bayo, J., & Elvira-Rendueles, B. (2005). Multiple regression models for predicting total daily pollen concentration in Cartagena. *Grana, 44*, 108–114.

Asero, R. (2002). Birch and ragweed pollinosis north of Milan: A model to investigate the effects of exposure to "new" airborne allergens. *Allergy, 57*, 1063–1066.

Asero, R., Wopfner, N., Gruber, P., Gadermaier, G., & Ferreira, F. (2006). Artemisia and Ambrosia hypersensitivity: Co-sensitization or co-recognition? *Clinical and Experimental Allergy, 36*, 658–665.

Aznarte, J. L., Sánchez, J. M. B., Lugilde, D. N., Fernández, C. D. L., de la Guardia, C. D., & Sánchez, F. A. (2007). Forecasting airborne pollen concentration time series with neural and neuro-fuzzy models. *Expert Systems with Applications, 32*, 1218–1225.

Béres, I., Novák, R., Hoffmanné Pathy, Zs., & Kazinczi G. (2005). Distribution, morphology, biology, importance and weed control of common ragweed (*Ambrosia artemisiifolia* L.). [Az ürömlevelű parlagfű (*Ambrosia artemisiifolia* L.) elterjedése, morfológiája, biológiája, jelentősége és a védekezés lehetőségei.] *Gyomnövények, Gyomirtás, 6*, 1–48. (in Hungarian).

Bousquet, J., Van Cauwenberge, P., Khaltaev, N., Ait-Khaled, N., Annesi-Maesano, I., Baena-Cagnani, C., et al. (2001). Allergic rhinitis and its impact on asthma. *Journal of Allergy and Clinical Immunology, 108*, S147–S334.

Cai, Z. (2007). Trending time-varying coefficient time series models with serially correlated errors. *Journal of Econometrics, 136*, 163–188.

Cecchi, L., Lorenzo, C., Morabito, M., Marco, M., Domeneghetti, M. P., Paola, D. M., et al. (2006). Long distance transport of ragweed pollen as a potential cause of allergy

in central Italy. *Annals of Allergy, Asthma & Immunology, 96*, 86–91.

Cecchi, L., Malaspina, T., Albertini, R., Zanca, M., Ridolo, E., Usberti, I., et al. (2007). The contribution of long-distance transport to the presence of Ambrosia pollen in central northern Italy. *Aerobiologia, 23*, 145–151.

Chernozhukov, V. (2005). Extremal quantile regression. *Annals of Statistics, 3*, 806–839.

Comtois, P. (1998). Ragweed (Ambrosia sp.): The Phoenix of allergophytes. In: F. Th. M. Spieksma (Ed.), Ragweed in Europe. Satellite symposium proceedings of 6th international congress on Aerobiology, Perugia, (pp. 3–5). Horsholm DK: Alk—Abello′A/S.

de Visiani, R. (1842), *Flora Dalmatica.* Vol. II.

Draper, N., & Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York: Wiley.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association, 87*, 998–1004.

Fan, J., & Yao, Q. (2005). *Nonlinear time series: Nonparametric and parametric methods*. New York: Springer.

Fornaciari, M., Bricchi, E., Greco, F., Fascini, D., Giannoni, C., Frenguelli, G., et al. (1992). Daily variations of Urticaceae pollen count and influence of meteoclimatic parameters in East Perugia during 1989. *Aerobiologia, 8*, 407–413.

Fornaciari, M., Pieroni, L., Orlandi, F., & Romano, B. (2002). A new approach to consider the pollen variable in forecasting yield models. *Economic Botany, 56*, 66–72.

Fumanal, B., Chauvel, B., & Bretagnolle, F. (2007). Estimation of pollen and seed production of common ragweed in France. *Annals of Agricultural and Environmental Medicine, 14*, 233–236.

Galán, C., Alcázar, P., Cariñanos, P., Garcia, H., & Domínguez-Vilches, E. (2000). Meteorological factors affecting daily urticaceae pollen counts in southwest Spain. *International Journal of Biometeorology, 43*, 191–195.

Galán, C., Cariñanos, P., García-Mozo, H., Alcázar, P., & Domínguez-Vilches, E. (2001). Model for forecasting Olea europaea L. airborne pollen in South-West Andalusia. *Spain. International Journal of Biometeorology, 45*, 59–63.

Giner, M. M., García, J. S. C., & Sellés, J. G. (1999). Aerobiology of Artemisia airborne pollen in Murcia (SE Spain) and its relationship with weather variables: annual and intradiurnal variations for three different species. Wind vectors as a tool in determining pollen origin. *International Journal of Biometeorology, 43*, 51–63.

Helbig, N., Vogel, B., Vogel, H., & Fiedler, F. (2004). Numerical modelling of pollen dispersion on the regional scale. *Aerobiologia, 20*, 3–19.

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

Jäger, S. (2000). Ragweed (Ambrosia) sensitisation rates correlate with the amount of inhaled airborne pollen. A 14-year study in Vienna, Austria. *Aerobiologia, 16*, 149–153.

Jato, M. V., Rodríguez, F. J., & Seijo, M. C. (2000). Pinus pollen in the atmosphere of Vigo and its relationship to meteorological factors. *International Journal of Biometeorology, 43*, 147–153.

Koenker, R. (2005). *Quantile regression*. Cambridge: Cambridge University Press.

Koenker, R., & Bassett, G. B. (1978). Regression quantiles. *Econometrica, 46*, 33–50.

Köppen, W. (1931). *Grundriss Der Klimakunde*. Berlin: Walter De Gruyter & Co.

Laaidi, M., Thibaudon, M., & Besancenot, J. P. (2003). Two statistical approaches to forecasting the start and duration of the pollen season of Ambrosia in the area of Lyon (France). *International Journal of Biometeorology, 48*, 65–73.

Makra, L., Juhász, M., Béczi, R., & Borsos, E. (2005). The history and impacts of airborne Ambrosia (Asteraceae) pollen in Hungary. *Grana, 44*, 57–64.

Makra, L., Juhász, M., Borsos, E., & Béczi, R. (2004). Meteorological variables connected with airborne ragweed pollen in Southern Hungary. *International Journal of Biometeorology, 49*, 37–47.

Makra, L., Tombácz, Sz., Bálint, B., Sümeghy, Z., Sánta, T., & Hirsch, T. (2008). Influences of meteorological parameters and biological and chemical air pollutants to the incidence of asthma and rhinitis. *Climate Research, 37*, 99–119.

Ocana-Peinado, F., Valderrama, M., & Aguilera, A. M. (2008). A dynamic regression model for air pollen concentration. *Stochastic Environmental Research and Risk Assessment, 22*, S59–S63. Supplement: Suppl. 1.

Oh, J. W. (2009). Development of pollen concentration prediction models. *Journal of Korean Medical Association, 52*, 579–591.

Peternel, R., Culig, J., Hrga, I., & Hercog, P. (2006). Airborne ragweed (Ambrosia artemisiifolia L.) pollen concentrations in Croatia, 2002-2004. *Aerobiologia, 22*, 161–168.

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

Rodríguez-Rajo, F. J., Jato, V., & Aira, M. J. (2005). Relationship between meteorology and Castanea airborne pollen. *Belgian Journal of Botany, 138*, 129–140.

Rodríguez-Rajo, F. J., Valencia-Barrera, R. M., Vega-Maray, A. M., Suarez, F. J., Fernandez-Gonzalez, D., & Jato, V. (2006). Prediction of airborne Alnus pollen concentration by using Arima models. *Annals of Agricultural and Environmental Medicine, 13*, 25–32.

Ruiz, S. S., Bustillo, A. M. G., Morales, P. C., & Cuesta, P. (2008). Forecasting airborne Platanus pollen in the Madrid region. *Grana, 47*, 234–240.

Saar, M., Gudzinskas, Z., Plompuu, T., Linno, E., Minkiene, Z., & Motiekaityte, V. (2000). Ragweed plants and airborne pollen in the Baltic states. *Aerobiologia, 16*, 101–106.

Sánchez Mesa, J. A., Galán, C., & Hervás, C. (2005). The use of discriminant analysis and neural networks to forecast the severity of the Poaceae pollen season in a region with a typical Mediterranean climate. *International Journal of Biometeorology, 49*, 355–362.

Schueler, S., & Schlüntzen, K. (2006). Modeling of oak pollen dispersal on the landscape level with a mesoscale atmospheric model. *Environmental Modeling & Assessment, 11*, 179–194.

Šikoparija, B., Smith, M., Skjøth, C. A., Radišič, P., Milkovska, S., Šimič, S., et al. (2009). The Pannonian plain as a source of Ambrosia pollen in the Balkans. *International Journal of Biometeorology, 53*, 263–272.

Skjøth, C. A., Smith, M., Šikoparija, 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, R. (1994). Nonregular regression. *Biometrika, 81*, 173–183.

Smith, M., & Emberlin, J. (2005). Constructing a 7-day ahead forecast model for grass pollen at north London, United Kingdom. *Clinical and Experimental Allergy, 35*, 1400–1406.

Smith, M., & Emberlin, J. (2006). A 30-Day-Ahead Forecast Model for Grass Pollen in North London, United Kingdom. *International Journal of Biometeorology, 50*, 233–242.

Sofiev, M., Siljamo, P., Ranta, H., & Rantio-Lehtimaki, A. (2006). Towards numerical forecasting of long-range air transport of birch pollen: theoretical considerations and a feasibility study. *International Journal of Biometeorology, 50*, 392–402.

Stennett, P. J., & Beggs, P. J. (2004). Pollen in the atmosphere of Sydney, Australia, and relationships with meteorological parameters. *Grana, 43*, 209–216.

Turos, O. I., Kovtunenko, I. N., Markevych, Y. P., Drannik, G. N., & DuBuske, L. M. (2009). Aeroallergen monitoring in Ukraine reveals the presence of a significant ragweed pollen season. *Journal of Allergy and Clinical Immunology, 123*(2), S95–S95.

Vázquez, L. M., Galán, C., & Domínguez-Vilches, E. (2003). Influence of meteorological parameters on olea pollen concentrations in Cordoba (South-western Spain). *International Journal of Biometeorology, 48*, 83–90.

Verma, K. S., & Pathak, A. K. (2009). A comparative analysis of forecasting methods for aerobiological studies. *Asian Journal of Experimental Sciences, 23*, 193–198.

Vogel, H., Pauling, A., & Vogel, B. (2008). Numerical simulation of birch pollen dispersion with an operational weather forecast system. *International Journal of Biometeorology, 52*(8), 805–814.

Wan, S. Q., Yuan, T., Bowdish, S., Wallace, L., Russell, S. D., & Luo, Y. Q. (2002). Response of an allergenic *Ambrosia psilostachya* (Asteraceae) to experimental warming and clipping: Implications for public health. *American Journal of Botany, 89*, 1843–1846.

Wand, M. P., & Jones, M. C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association, 88*, 520–528.

Wopfner, N., Gadermaier, G., Egger, M., Asero, R., Ebner, C., Jahn-Schmid, B., et al. (2005). The spectrum of allergens in ragweed and mugwort pollen. *International Archives of Allergy and Immunology, 138*, 337–346.

Yu, K., & Jones, M. C. (1998). Local linear quantile regression. *Journal of the American Statistical Association, 93*, 228–237.