

Spatial and temporal modeling of daily pollen concentrations

Curt T. DellaValle · Elizabeth W. Triche ·
Michelle L. Bell

Received: 17 May 2010 / Revised: 24 January 2011 / Accepted: 24 January 2011 / Published online: 18 February 2011
© ISB 2011

Abstract Accurate assessments of pollen counts are valuable to allergy sufferers, the medical industry, and health researchers; however, monitoring stations do not exist in most areas. In addition, the degree of spatial reliability provided by the limited number of monitoring stations is poorly understood. We developed and compared spatial models to estimate pollen concentrations in locations without monitoring stations. Daily *Acer*, *Quercus*, and overall tree, grass, and weed pollen counts, in grains/m³, were obtained from 14 aeroallergen monitoring stations located in the northeastern and mid-Atlantic region of the United States from 2003 to 2006. Pollen counts were spatially interpolated using ordinary kriging. Mixed effects and generalized estimating equations incorporating daily and seasonal weather characteristics, pollen season characteristics and land-cover information were also developed to estimate daily pollen concentrations. We then compared observed values from a monitoring station to model estimates for that location. Observed counts and kriging estimates for tree pollen differed ($p=0.04$), but not when peak periods were removed ($p=0.29$). No differences between observed and kriging estimates of *Acer* ($p=0.46$), *Quercus* ($p=0.24$), grass ($p=0.31$) or weed pollen ($p=0.29$) were found. Estimates from longitudinal models also demonstrated good agreement with observed counts, except

for the extremes of pollen distributions. Our results demonstrate that spatial interpolation techniques as well as regression methods incorporating both weather and land-cover characteristics can provide reliable estimates of daily pollen concentrations in areas where monitors do not exist for all but periods of extremely high pollen.

Keywords Pollen · Allergy · Allergens · Spatial modeling

Introduction

Accurate reporting of aeroallergens is valuable to the public, especially sensitive individuals, because it allows them to limit exposure, reduce physical activity and/or prompt use of prophylactic medication (Levetin and Van de Water 2003). Research on aeroallergens and health, such as epidemiologic studies, depend on the quality and availability of exposure data. Unfortunately, very few aeroallergen monitoring stations exist, with 86 monitoring stations certified by the American Academy of Allergy, Asthma and Immunology (AAAAI) within 36 states of the continental U.S., providing daily allergen counts to the public (AAAAI 2010). In addition, the frequency of measurement can be highly variable across monitors. The scarcity of aeroallergen monitoring stations and infrequency of measurement limits the accuracy of reported allergen levels, especially in areas far from a monitoring station.

Previous work noted large variations in regional concentrations of pollen, potentially due to effects of vegetative land-cover patterns on the composition of regional pollen (Emberlin et al. 1999). Atmospheric flow also affects the dispersal of particles, which includes pollen (Nathan et al. 2005). Both the characteristics of particles (e.g., size, shape, density) and weather conditions, which can vary regionally,

C. T. DellaValle (✉) · M. L. Bell
School of Forestry and Environmental Studies, Yale University,
195 Prospect Street,
New Haven, CT 06511, USA
e-mail: curt.dellavalle@gmail.com

E. W. Triche
Department of Community Health/Epidemiology,
Brown University School of Medicine,
Box G-S121-2, Providence, RI 02912, USA

can affect atmospheric transport and dispersal of particles (Kuparinen et al. 2007). The spatial heterogeneity of pollen levels suggests that measurements from a sparse monitoring station network may not be adequate to estimate pollen concentrations over large spatial scales.

Several methods have been used to estimate aeroallergen concentrations for locations and time periods without monitors, such as spatial interpolation and pollen forecasting. However, these methods can be limited in their temporal and spatial scope. Alba et al. (2006) utilized the spatial interpolation method of kriging to estimate *Olea* pollen concentrations in the Andalusia region of Spain. No significant differences were found between observed and model estimates of weekly average pollen concentrations, but only when variability between monitoring stations was low.

Errors in pollen count estimates are potentially greater as distance between monitoring stations increases, due to spatial variation in concentrations. This may limit the ability of spatial interpolation models to accurately estimate pollen counts in areas with few monitoring stations. Due to the potential limitations of spatial interpolation in areas where monitors are far apart or demonstrate a high variability, other modeling methodologies that incorporate weather and land-use variables may be warranted.

Forecasting models have focused on predicting pollen concentrations over several days to weeks as well as characteristics of the pollen season such as overall severity or start date at specific locations (Levetin and Van de Water 2003; Weber 2003). Weather characteristics have been the primary input variables in forecasting models. Most notably, metrics of temperature and atmospheric water have demonstrated the ability to accurately predict pollen concentrations and characteristics of the pollen season (Smith and Emberlin 2005, 2006; Emberlin et al. 2007; Garcia-Mozo et al. 2006a; Stach et al. 2008a, b; Antepara et al. 1995). The purpose of forecasting models has been to describe temporal, and not spatial, variation in pollen levels. Variations in pollen counts have demonstrated relationships with changes in land-cover patterns (Emberlin et al. 1993, 1999). Understanding land-cover, which is an assessment of vegetative cover, would be valuable in describing regional variations in pollen.

This study presents a novel approach to estimate aeroallergen concentrations by incorporating land-cover and regional characteristics of the pollen season into forecasting models to improve assessment of spatial variability in estimation of daily pollen concentrations. The methods are designed to address regional variability in pollen levels that hinders accuracy of estimates using traditional approaches for areas without monitors. We compare model estimates to measurements from current aeroallergen monitors for daily pollen concentrations across

a large geographic region, New England and the Mid-Atlantic states. Our models, while not designed to describe pollen dispersion, are useful for estimating pollen concentrations in areas or time periods without aeroallergen monitors with the primary purpose of estimating exposures for health studies.

Materials and methods

Pollen, weather and land-cover data

Daily ambient pollen concentrations were obtained from the National Allergy Bureau (NAB), a section of the AAAAI Aeroallergen Network, from 14 aeroallergen monitoring stations in eastern U.S. cities listed in Table 1. Tree, grass, and weed pollen counts were collected from the period 2003–2006. We also performed separate analysis of the two most abundant individual pollen genera, *Acer* and *Quercus*. At some sites, fewer years of records were available (Table 1). Grass pollen was not recorded at Olean, NY, and genera specific counts were unavailable at the Olean, NY, and Brooklyn, NY, sites. Pollen was measured on unobstructed rooftops at least one story above ground using one of the following monitor types: Burkard volumetric spore trap or Rotorod sampler (Table 1), which have demonstrated comparable measurements of pollen particles (Frenz 1999).

Pollen groupings of tree, grass and weed have been used in epidemiologic studies investigating the effects of pollen on respiratory outcomes (Cakmak et al. 2005; Dales et al. 2004). Since a major goal of our research was to develop exposure estimates for health based studies, we considered general pollen groupings of tree, grass and weeds in addition to specific genera.

Daily minimum, maximum and mean temperature, dew point temperature, precipitation and wind speed data were collected from the National Climatic Data Center. Weather data were collected from the weather station in closest proximity to each aeroallergen monitoring station (mean 9.91 km; range 1.61–24.54 km).

Land-cover data for the region that contains Massachusetts, Connecticut, New York, New Jersey, Maryland, Delaware and Eastern Pennsylvania were obtained from the United States Geological Survey Land Cover Institute, National Land-cover Database 2001 (NLCD 2001; http://www.mrlc.gov/nlcd_multizone_map.php). NLCD 2001 is a 30-m resolution land cover database created from Landsat 5 and Landsat 7 satellite imagery. NLCD 2001 identifies 16 classes of land cover: open water, perennial ice/snow, developed (open space, low intensity, medium intensity, high intensity), barren land, deciduous forest, evergreen forest, mixed forest, shrub, grassland/herbaceous, hay/

Table 1 Allergen monitor locations, climate and average season dates

Monitor site	Years	Sampler type	Average temperature (°C)	Average precipitation (mm)	Average dates of pollen season				
					Grass pollen	Weed pollen	Tree pollen	<i>Acer</i> pollen	<i>Quercus</i> pollen
Rochester, NY	2003–2006	Burkard	9.6	71.5	5/10–9/19	5/23–10/23	3/11–7/25	3/12–5/3	4/22–6/1
Albany, NY	2003–2006	Burkard	9.4	81.5	5/20–9/18	8/7–9/19	4/2–6/17	4/2–5/11	5/5–6/3
Chelmsford, MA	2003–2005	Burkard	9.7	91.4	5/8–9/21	5/15–10/8	3/31–7/21	3/29–5/19	4/25–5/30
Olean, NY	2003–2006	Burkard	7.7	86.4	Not reported	6/8–10/10	4/7–7/5	Not reported	Not reported
Waterbury, CT	2003–2006	Burkard	9.4	109.0	5/29–9/21	8/4–9/26	4/3–7/3	4/11–5/29	4/16–6/12
Armonk, NY	2003–2006	Burkard	9.8	106.7	5/3–9/22	5/9–10/10	3/22–7/18	3/25–5/15	4/20–5/28
Newark, NJ	2005–2006	Burkard	12.8	98.7	5/26–10/16	4/26–10/28	3/27–6/3	4/17–5/4	4/22–5/21
Brooklyn, NY	2005–2006	Burkard	12.6	98.9	4/28–8/26	8/8–10/6	3/28–6/26	Not reported	Not reported
Philadelphia, PA	2004–2006	Burkard	13.3	88.9	4/11–9/26	5/5–10/26	3/12–6/22	3/12–4/29	4/14–5/29
York, PA	2004–2006	Rotorod	12.1	90.9	5/14–9/15	6/17–9/28	3/7–6/23	3/8–5/17	4/1–6/1
Cherry Hill, NJ	2003–2006	Burkard	13.3	97.3	4/18–10/7	5/15–10/23	3/30–7/14	3/13–4/28	4/14–5/27
New Castle, DE	2006	Burkard	14.7	90.6	5/5–9/25	4/26–10/11	3/6–9/13	3/3–5/10	4/12–5/30
Baltimore, MD	2003–2006	Rotorod	13.2	91.3	4/18–9/29	5/11–10/11	3/11–6/9	3/12–5/10	4/16–5/13
Washington, DC	2003–2006	Burkard	14.3	83.7	4/13–10/7	4/25–10/21	2/25–6/14	2/16–4/30	4/11–5/21

pasture, cultivated crops, woody wetland, and emergent herbaceous wetland. Land-cover was used to approximate densities of vegetative types, which are known to be important in pollen count estimation (Branzi and Zanotti 1992). This study assumes that 2001 land-cover is representative of the study period, 2003–2006.

Statistical analysis

We examined three methods for estimating ambient pollen concentrations in areas without monitors. These methods are (1) measurements from the nearest monitoring station, (2) spatial interpolation using ordinary kriging, and (3) longitudinal regression analysis.

Nearest monitor method

The nearest monitor method is the simplest approach and estimates pollen levels at a given point using measurements at the nearest monitor. We evaluated the appropriateness of this method by comparing how daily *Acer*, *Quercus*, tree, grass and weed pollen concentrations vary by distance. Pearson correlation coefficients for pollen counts on a given day were calculated for each monitor pair. Correlations were then analyzed as a function of distance between the sites.

Kriging

Daily *Acer* and *Quercus* pollen concentrations and tree, grass, and weed concentrations overall were estimated

through interpolation using an ordinary kriging technique. Autocorrelation due to the spatial structure of the data can be accounted for by an interpolation technique such as kriging (Robertson 1987; Liebhold et al. 1993; Legendre and Fortin 1989). Kriging allows for a weighting of values by distance based on a semivariogram and the spatial correlation of the data (Waller and Gotway 2004). Covariance structure of the model was determined by fitting a variogram using a weighted least squares approach. Several approaches to estimating the covariance structure for the kriging model were explored including spherical, Gaussian, exponential and Matern. Compared to other covariance structures, the Matern model produced a slightly better fit to the data as assessed by the variogram. Kriging surfaces were calculated using ArcGIS 9.3.1 (ESRI, Redwood, CA, USA).

Longitudinal models

Two model types, mixed effects and generalized estimating equations (GEE), were used to estimate daily ambient pollen concentrations. Separate mixed effects and GEE models were constructed for each dependent variable (daily *Acer*, *Quercus*, tree, grass and weed pollen). Each pollen type was regressed against daily and seasonal weather variables, characteristics of the pollen season (day of season and indicator variables for the peak and pre-peak pollen period), land-cover and variables describing the interaction between day of season and daily weather characteristics. Optimal covariance structure was determined by selecting

models that minimized the Akaike Information Criteria (AIC), when considering full models. Mixed effects models for *Quercus*, tree and weed pollen obtained a minimum AIC value when random effects of intercept and day of season were included. Random effects in mixed effects models for *Acer* and grass pollen that minimized the AIC value were random intercept and sine day of season. Final models were fit using backward, stepwise progression with inclusion criteria of $\alpha=0.05$. Longitudinal models were fit using SAS 9.1 (SAS Institute, Cary, NC, USA).

The following describes the independent variables considered in the mixed effects and GEE models.

Daily weather

Daily minimum, maximum and average temperature; dew point temperature; precipitation; and wind speed were considered in longitudinal models. Principal components analysis (PCA) was used to address high correlation between temperature and dew point temperature. PCA reduces the dimensionality of data by identifying directions of maximum variability. The resultant principal component (s) are uncorrelated linear combinations of original variables and represent the rotation that explains the maximum variation in 1-dimension. For all regression models, only the resultant first principal component was included, as it accounted for 92% of total variability between temperature variables.

Precipitation was considered in longitudinal models, but was not included in PCA analysis of weather variables for two reasons: (1) precipitation was not significantly correlated with temperature, Pearson correlation coefficient 0.07, and (2) previous research indicated that precipitation and temperature have different effects on pollen counts (Weber 2003; Laaidi 2001). Wind speed was not included in the PCA but was considered separately in longitudinal models.

Principal component variables representing temperature and dew point temperature, precipitation and wind speed were considered for the current day (lag 0), previous day (lag 1), and, moving averages over multiple days (lag 1–2, 1–3, 1–5 or 1–7).

Seasonal variables

Weather patterns over time frames longer than days or weeks have demonstrated significant effects on pollen counts (Stach et al. 2008a, b; Emberlin et al. 2002, 2007; Laaidi 2001; Valencia-Barrera et al. 2002). Average temperature and precipitation 1 and 2 months prior to the average start of the pollen season of interest, during the previous growing season (March 15 to August 31), and winter (November 1 to February 28/29) prior to the pollen season of interest were considered.

Pollen seasons were defined (Emberlin et al. 1993; Goldberg et al. 1988) separately for each pollen type (*Acer*, *Quercus*, tree, grass, and weed) and year as the period during which 98% of cumulative pollen, of a certain type, for a given year was measured. Start dates were defined for each pollen type and year as the day on which 1% of the cumulative pollen, of a certain type, for a year had been measured. End dates for each pollen type and year were the days on which 99% of cumulative pollen for a year had been measured.

Other seasonal variables were day of the pollen season and sine day of the pollen season, as the time-series of pollen measurements in this region followed a sinusoidal pattern. Indicators of the peak and pre-peak periods of the pollen season, for each pollen type, were included. Previous studies observed that associations between weather and pollen counts during peak and non-peak periods can differ (Smith and Emberlin 2005). Pre-peak period was defined as the start date to the first inflection point on the cumulative distribution curve. Peak period ranged from the first inflection point on the distribution curve to the second inflection point or 85% of the cumulative pollen for a season, whichever comes earlier in the year. These definitions are based on earlier methods (Smith and Emberlin 2005).

Land-cover surrounding each monitoring station was calculated with buffer distances of 1, 5, 10, 20, 35, 50, 75 and 100 km. The proportion of each of the 16 classes of land-cover in the NLCD 2001 dataset was calculated for each radius size and monitor. Hierarchical cluster analysis was conducted to group sites with similar land-cover based on variations in the proportions of land-cover types around each monitoring station. Euclidean distances were calculated between all pairs of objects, i.e. land-cover classes at each monitor location. Monitor sites were then grouped into clusters according to Ward's method, a space conserving approach that minimizes internal sums of squares.

Model assessment

Cross-validation analysis was performed to evaluate model performance based on the Waterbury site for each model: kriging, mixed effects and GEE. This involved estimating pollen levels at the Waterbury location based on a dataset omitting Waterbury monitor values, and comparing to measurements. Cross-validation was completed only for days with measurements at Waterbury. This site was chosen because modeling techniques from this study will be applied to future research in the Connecticut region using existing health datasets. The Waterbury site is also situated in a representative area of this study region, with respect to the abundance of pollen. Wilcoxon tests were performed to compare differences between estimated and observed values.

Observed and estimated pollen counts were categorized by percentile as: *absent*=zero count; *very low*=any count >0 up to 50th percentile; *low*=50th–75th percentile; *moderate*=75th–90th percentile; *high*=90th–95th percentile and *very high*=95th percentile and higher. We calculated the percent of days with agreement (i.e., same category) for observed measurements and pollen counts estimated by kriging or longitudinal regression analysis.

Results

Nearest monitor approach

Correlation coefficients (Pearson r) of monitor pairs were calculated for each pollen type (*Acer*, *Quercus*, tree, grass, and weed pollen) and compared to distance between monitors to evaluate the appropriateness of exposure methods based the nearest monitor. Correlation coefficients between daily pollen counts of *Quercus*, tree, grass and weed at different monitors were inversely related to distance between monitoring sites (Fig. 1). Correlation of daily *Acer* pollen counts between monitor pairs was not a function of distance between monitors (Fig. 1). The average correlation of daily tree pollen counts between monitor pairs in this region (maximum distance 627 km) was 0.33

with range 0.002–0.79, with statistically significant ($\alpha=0.05$) correlations for 65% of monitor pairs. Even for stations within 50 km, correlations varied substantially (0.52–0.79), 100% of pairs with statistically significant correlations. Correlation of daily tree pollen counts between sites decreases more rapidly over the first 160 km (correlation decreased 1.7%/km) compared to stations >160 km apart (0.3%/km). Average declines in the correlation for weed and grass measurement between monitor pairs were 0.2 and 0.4%/km, respectively. The average correlation coefficient between monitoring stations in this region was 0.36 for weed measurements (73% of pairs significant) and 0.28 for grass (57% of pairs significant). The range of correlations between stations within 50 km for weed (0.24–0.75, 86% significant) and grass pollen (0.18–0.52, 86% significant) were relatively large.

Correlations of daily pollen counts between monitoring pairs were no stronger when examining individual genera of *Acer* and *Quercus*. Of pollen types considered in this study, measurements of *Acer* and *Quercus* pollen counts between monitoring pairs over the study area had the lowest average correlations, 0.18 (range <0.01–0.63, 21% significant) and 0.22 (range <0.01–0.70, 28% significant), respectively. These pollen types demonstrated little dependence on distance between monitors, with correlations declining an average 0.02%/km for *Acer* and 0.2%/km for *Quercus*.

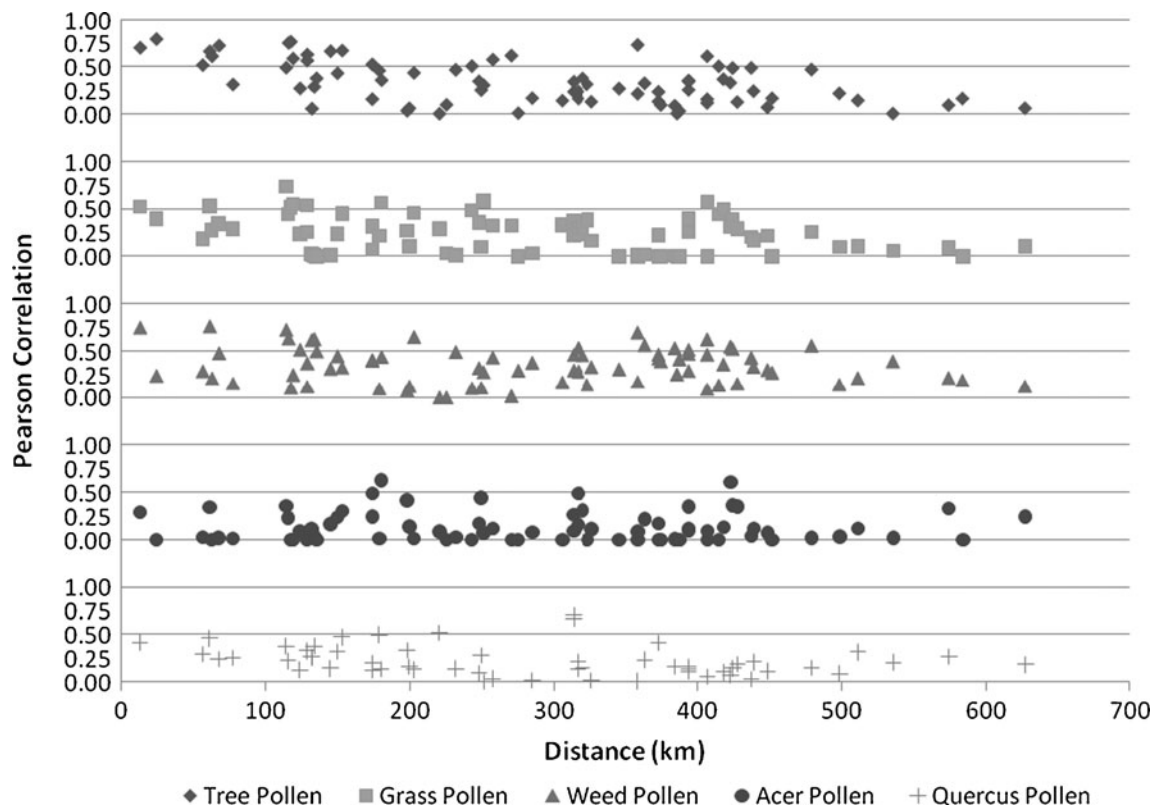


Fig. 1 Correlation of daily pollen counts by distance. Each point represents a monitor pair

Spatial interpolation (Kriging)

Figure 2 shows observed tree pollen counts with estimated values from kriging for Waterbury. For tree pollen, Wilcoxon tests indicated significant differences between observed and estimated counts from kriging ($p=0.04$). When tree pollen values classified as *very high* (≥ 95 th percentile) were removed, no significant difference remained between observed and estimated counts ($p=0.29$). Correlations between observed tree pollen counts at Waterbury and kriging estimates were 0.57 ($p<0.01$), which was lower than correlation between observed counts at Waterbury and measurements from the nearest monitor, Armonk, NY (67 km from Waterbury) ($r=0.72$, $p<0.01$).

Wilcoxon tests revealed no significant differences between observed pollen measurements and kriging estimates for *Acer* ($p=0.46$) or *Quercus* ($p=0.24$). Although no significant difference was detected in the underlying distributions, the low correlation between daily observed measurements and kriging estimates for *Acer* ($r=0.16$, $p=0.34$) reflects relatively poor agreement in daily measurements from neighboring monitoring stations. Correlation between observed *Acer* measurements from Armonk, the nearest monitor station, and Waterbury was 0.02 ($p=0.88$). Kriging estimates of *Quercus* pollen had a higher correlation ($r=0.35$, $p=0.07$) with observed measurements at Waterbury than the nearest monitor station, Armonk ($r=0.24$, $p=0.02$).

Peak concentrations for grass and weed pollen are not as extreme as tree pollen, which may add to the predictive ability of spatial interpolation models (Fig. 2). Test of differences between observed and estimated counts by kriging for grass ($p=0.31$) and weed pollen ($p=0.47$) confirm this observation.

Weed pollen counts at Waterbury were more highly correlated with observations at the nearest monitor station ($r=0.48$, $p<0.01$), Armonk, than with kriging estimates ($r=0.19$, $p=0.21$) for the Waterbury location. Grass pollen counts at Waterbury had a similar degree of correlation with counts at Armonk ($r=0.34$, $p=0.05$) and kriging estimates for the Waterbury location ($r=0.32$, $p<0.01$).

Low levels of agreement between pollen measurements at Waterbury and neighboring monitor stations may account for kriging estimates having a lower correlation with measurements at Waterbury than measurements from the nearest monitoring station (Armonk). Tree and grass pollen measurements at Albany (distance: 132 km) had correlations of 0.05 and 0.03 ($p=0.56$ and 0.51) with measurements at Waterbury, respectively. Weed pollen counts at Brooklyn (128 km) also demonstrated a relatively low correlation ($r=0.12$, $p=0.34$) with counts at Waterbury. The monitor at Waterbury recorded pollen measurements on 69.2% of days in the pollen season and Armonk 59.4%. Therefore there are days on which pollen estimates, using the nearest monitor method, would either be missing or obtained from one of these other neighboring stations.

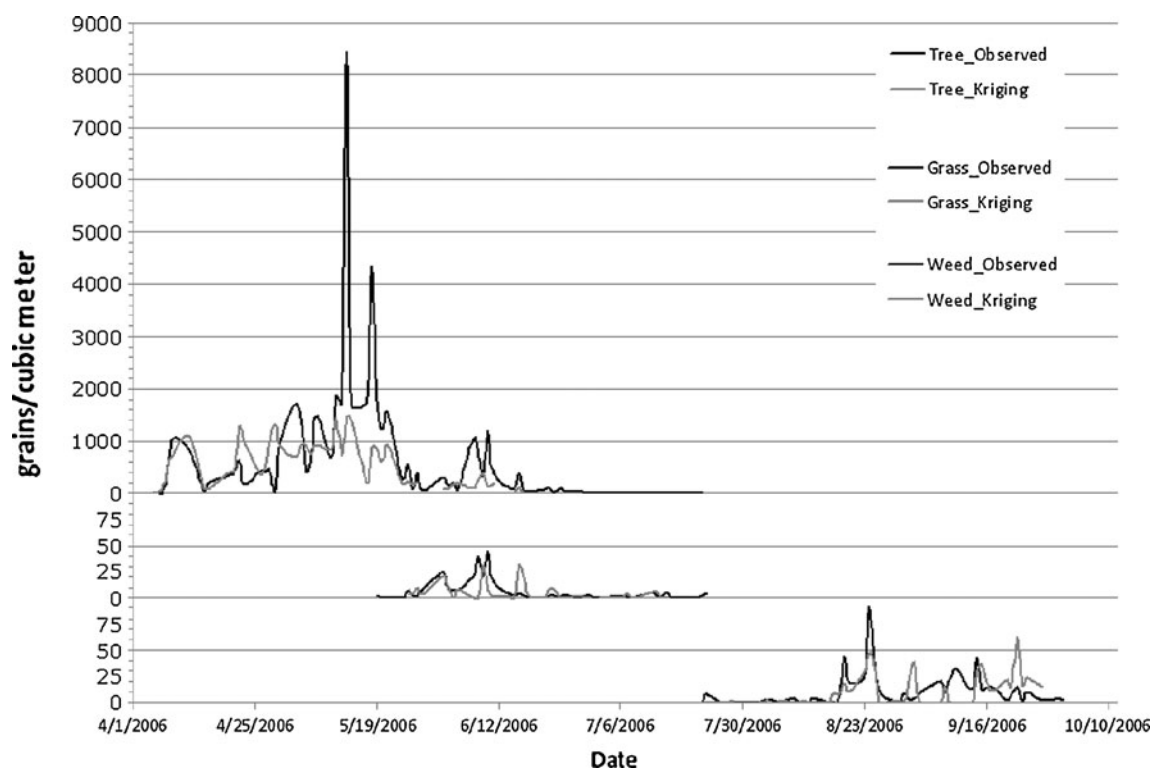


Fig. 2 Observed pollen counts compared to kriging estimates at Waterbury, CT

Agreement for categorized groupings (*absent*, *very low*, *low*, *moderate*, *high*, *very high*) of *Acer*, *Quercus*, tree, grass and weed pollen levels between observed measurements at Waterbury and kriging estimates are shown in Table 2, which provides the percent of estimates within the same concentration category (e.g., *low*) between measurements and an estimation method (nearest monitor, kriging, mixed effects model or GEE model). Results are provided for measurements at the Waterbury station and cumulatively for all 14 monitoring sites over the entire region.

Of days with measured tree pollen concentrations classified as *very high* at Waterbury, 81.5% also had *very high* concentrations estimated by the nearest monitor method (i.e., observations at Armonk) (Table 2). In contrast, only

20% of days with *very high* measured tree pollen concentrations at Waterbury were also estimated as *very high* by kriging. Agreement between observed tree pollen levels at Waterbury and either observed levels at Armonk (i.e. nearest monitor method) or estimated levels from kriging for other tree pollen level classifications (*absent*, *very low*, *low*, *moderate* and *high*) were comparable.

On 36.8% of days with no measureable (*absent*) *Acer* pollen at Waterbury there was also no measureable count at the nearest monitor station (Armonk). A similar proportion of days of *absent* count were found for *Quercus*, 38.9%. There were no days of *absent* *Acer* or *Quercus* pollen at the Waterbury station that were also estimated to be *absent* by kriging. Agreement between *Acer* measurements at Water-

Table 2 Agreement for closest monitor, kriging and longitudinal model methods

	Pollen type	Pollen category					
		Absent	Very low	Low	Moderate	High	Very high
Waterbury/Armonk ^a	Grass	51.5%	35.1%	17.7%	33.3%	0.0%	0.0% ^b
	Weed	16.7%	30.0%	28.1%	37.5%	25.0%	20.0%
	Tree	0.0%	72.2%	62.5%	32.5%	40.0%	81.5%
	<i>Acer</i>	36.8%	0.0%	22.7%	34.8%	0.0%	11.1%
	<i>Quercus</i>	38.9%	37.5%	51.1%	45.0%	0.0%	77.8%
Waterbury: kriging	Grass	0.0%	60.0%	22.2%	0.0%	0.0%	0.0% ^b
	Weed	0.0%	14.3%	0.0%	40.0%	33.3%	0.0%
	Tree	0.0%	84.2%	36.4%	38.9%	33.3%	20.0%
	<i>Acer</i>	0.0%	0.0%	27.3%	20.0%	0.0%	0.0%
	<i>Quercus</i>	0.0%	0.0%	40.0%	42.9%	0.0%	16.7%
Waterbury: mixed effects	Grass	74.4%	42.6%	12.5%	0.0%	0.0%	0.0% ^b
	Weed	8.3%	69.0%	23.7%	51.5%	0.0%	0.0%
	Tree	0.0%	74.2%	30.6%	72.2%	11.8%	12.0%
	<i>Acer</i>	0.0%	0.0%	76.2%	45.5%	40.0%	0.0%
	<i>Quercus</i>	4.0%	53.8%	86.7%	21.4%	0.0%	0.0%
Waterbury: GEE	Grass	58.6%	53.0%	7.7%	0.0%	0.0%	0.0% ^b
	Weed	8.3%	86.2%	62.5%	0.0%	0.0%	0.0%
	Tree	0.0%	74.2%	49.0%	25.0%	11.8%	8.0%
	<i>Acer</i>	0.0%	0.0%	75.0%	28.6%	16.7%	42.9%
	<i>Quercus</i>	0.0%	21.1%	71.9%	5.0%	0.0%	0.0%
Entire region: mixed effects	Grass	30.5%	57.7%	53.0%	40.0%	20.5%	8.1%
	Weed	43.9%	26.8%	53.6%	47.7%	16.5%	0.0%
	Tree	18.8%	77.9%	48.9%	34.1%	8.4%	6.9%
	<i>Acer</i>	6.2%	31.2%	68.9%	41.8%	15.1%	7.1%
	<i>Quercus</i>	1.0%	33.0%	79.7%	53.1%	18.9%	16.1%
Entire region: GEE	Grass	19.9%	67.9%	38.9%	31.9%	23.4%	14.0%
	Weed	3.8%	69.3%	36.3%	37.7%	0.0%	0.0%
	Tree	0.0%	82.6%	49.7%	24.0%	12.6%	4.8%
	<i>Acer</i>	0.0%	34.4%	73.6%	31.0%	10.0%	20.0%
	<i>Quercus</i>	0.0%	8.5%	82.6%	41.5%	13.3%	23.9%

^a Nearest monitor station: 67 km

^b No grass pollen reported in very high category at Waterbury

bury and kriging estimates was only observed for counts within the *low* (27.3%) and *moderate* (20.0%) classifications. Kriging estimates of *Quercus* within the *very high* classification demonstrated some agreement with days measured as *very high* at Waterbury (16.7%).

Overall, agreement between categorized grass pollen levels observed at Waterbury and either measured levels from the nearest monitor (Armonk) or kriging estimates were quite low (Table 2). Only for days on which grass levels were observed at the *very low* or *low* levels was there any agreement with kriging estimates, at 60 and 22%, respectively. However, 100% of days with *absent* counts at Waterbury were estimated by kriging as *very low*. Grass pollen counts in the *very high* category were not observed at the Waterbury location.

Agreement between observed weed pollen levels at Waterbury and either levels estimated by kriging or observed at the closest monitors was very poor (Table 2). No category of weed pollen had agreement higher than 43.8% with observed values.

Longitudinal models

We estimated daily *Acer*, *Quercus*, tree, grass and weed pollen using two forms of regression models, mixed effects models and GEE. Several weather variables, day of season, peak season indicators and land-cover were significant predictors ($\alpha=0.05$) of daily *Acer*, *Quercus*, tree, grass and weed pollen counts in mixed effects and GEE models. Principal component variables representing temperature characteristics on the present and/or previous day demonstrated positive associations with tree and grass pollen counts. However, the principal component variable representing same day temperature variables was negatively associated with weed pollen counts. Same and previous day precipitation were negatively associated with counts of all pollen types. Interaction of weather variables and time (either day of season or sine day of season) demonstrated the opposite direction of association as the main effect of weather for all pollen types. Land cover was also a significant predictor of pollen counts in all models, except for GEE models predicting weed pollen counts.

Estimated tree pollen counts from a mixed effects model had a correlation of 0.57 ($p<0.01$) with observed measurements over the entire study region and a correlation of 0.66 ($p<0.01$) with observed measurements at Waterbury. A correlation of 0.57 would be expected between monitoring stations that are only 2.8 km apart, based on extrapolation from correlations observed between all pairs of monitor stations in the study area. The correlation between Waterbury measurements and GEE model estimates was even higher ($r=0.71$, $p<0.01$). GEE model performance over the entire study region, $r=0.72$ ($p<0.01$), was comparable to

the correlation between GEE estimates and observations at Waterbury alone.

Correlations between observed measurements of *Acer* and *Quercus* and estimates from either mixed effects or GEE models were strong compared to the nearest monitor approach. Estimates of *Acer* from a mixed effects model demonstrated strong correlation with observed *Acer* counts over the entire study region ($r=0.56$, $p<0.01$) and at Waterbury ($r=0.64$, $p<0.01$). GEE model estimates had similar correlations with observed *Acer* counts over the entire study region ($r=0.60$, $p<0.01$) and at Waterbury ($r=0.57$, $p<0.01$). Correlations for mixed effect model estimates of *Quercus* pollen and observations were 0.66 ($p<0.01$) for measurements at Waterbury and 0.77 ($p<0.01$) for the entire study region. Results from GEE models were similar.

Table 2 shows accuracy of categorized tree pollen estimates for mixed effects and GEE models. Wilcoxon tests indicated significant differences between observed measurements at Waterbury and estimates by mixed effects ($p=0.02$) or GEE models ($p<0.01$). However, as with kriging, differences between observed measurements and estimates from mixed effects and GEE models improved when counts in the *very high* category were removed ($p=0.28$ and $p=0.03$, respectively). Neither GEE nor mixed effects models predicted any days of *absent* tree pollen for Waterbury. Over the entire study region, mixed effects models accurately predicted 18.8% of days with no observed tree pollen (50 of 266 observed days). Days with observed *low* levels of tree pollen at Waterbury were also estimated to have *low* levels by GEE on 49.0% of days, compared to 30.6% of days for the mixed effects model. Mixed effects model estimates were in higher agreement with days of observed *moderate* and *very high* tree pollen levels at Waterbury compared to GEE estimations. Accuracy of tree pollen estimates at the *high* and *very high* categories were much lower for mixed effects or GEE estimates compared to the nearest monitor method. The same is true, although less so, for comparisons to kriging estimates (Table 2).

As with overall tree pollen counts, accuracy of *Acer* and *Quercus* estimates from mixed effects and GEE models was greater for days with pollen counts classified as *very low* to *moderate* compared to days within the *high*, *very high* or *absent* categories. No significant differences between observed *Acer* measurements at Waterbury and estimates from either mixed effects ($p=0.77$) or GEE models ($p=0.32$). Daily *Quercus* estimates at Waterbury from mixed effects ($p=0.84$) and GEE models ($p=0.87$) did not differ from observed measurements. Over the entire study region, fit of *Quercus* pollen estimates from mixed effects ($p=0.16$) and GEE models ($p=0.02$) were not as strong. Differences between measurements of *Quercus* over the

entire study region and estimates from mixed effects and GEE models improved when counts in the *very high* category were removed ($p=0.64$ and $p=0.15$, respectively).

For grass pollen, model estimates had correlations of 0.56 and 0.44 ($p<0.01$ and <0.01) with observed counts at Waterbury for mixed effects and GEE models, respectively. Models for weed pollen had correlations of 0.46 and 0.70 ($p<0.01$ and <0.01) for mixed effects and GEE models, respectively. Wilcoxon tests revealed no difference between observed grass pollen measurements and estimates from mixed effects ($p=0.26$) or GEE models ($p=0.59$). No differences were found between observed weed pollen and mixed effects ($p=0.32$) or GEE estimates ($p=0.05$).

Over the entire study region, correlations between predicted and observed concentrations of grass pollen were 0.48 and 0.66 ($p<0.01$ and <0.01) for mixed effects and GEE models, respectively. For weed pollen correlations were 0.58 and 0.53 ($p<0.01$ and <0.01) for mixed effects and GEE models, respectively. For comparison, the correlation between observed grass pollen concentrations at Philadelphia and Cherry Hill, stations 13 km apart, was 0.52 ($p<0.01$). The correlation between observed weed pollen concentrations at the same two stations was 0.75 ($p<0.01$).

Agreement between observations and longitudinal model estimates for grass pollen at Waterbury decline with increasing concentration categories for mixed effects and GEE models (Table 2). Neither model type accurately estimated grass pollen levels on days with observed *moderate* or higher levels at Waterbury. Over the entire study region, however, a portion of days with observed grass pollen counts in the *moderate* or higher categories was estimated to have corresponding levels of grass pollen by both mixed effects and GEE models.

Days with observed weed pollen within the *moderate* category at Waterbury were estimated to have a corresponding level of weed pollen by mixed effect models on 51.5% of days. There were no days with observed *moderate* levels of weed pollen with the same level estimated by GEE models at Waterbury. Neither mixed effects nor GEE models accurately predicted days with observed *high* or *very high* levels at Waterbury. Over the entire study region, days with observed *high* weed pollen levels were accurately estimated by mixed effects models on 16.5% of days.

Discussion

The continental U.S. has only 86 aeroallergen monitors within the NAB network providing public reports of daily ambient allergen levels. Accurate estimation of ambient pollen concentrations in the vast areas where aeroallergen

monitors do not exist would be beneficial to those who suffer from allergies, the medical industry, and health researchers. However, the optimal manner of daily pollen estimation across a broad spatial scale is poorly understood. This study sought to investigate whether information on the spatial pattern of allergen measurements as well as local weather, pollen season characteristics and land-cover could be used to develop statistical techniques that could improve both temporal and spatial estimates of ambient pollen concentrations compared to the current monitoring network. Such models do not describe the dispersal patterns of pollen grains, but could be applied to improve public reporting of pollen levels in areas without monitors and improve ambient exposure assessments in epidemiologic studies.

There are of course pros and cons to each of the three methods presented for estimating pollen concentrations in areas where monitors do not exist. Previous studies utilized measurements from a single monitor to assess ambient pollen exposures, which may not fully represent spatial heterogeneity across a study region (Lewis et al. 2000). In this study, correlations of daily pollen counts between monitoring stations decreased with distance, with the most rapid rate of decline over the first 160 km of separation. Moreover, even for stations within 50 km of one another, the range in correlations of daily pollen counts was substantial (range for *Acer* pollen counts: 0.34–0.01; *Quercus* pollen: 0.46–0.25; tree pollen: 0.79–0.52; grass pollen: 0.52–0.18; and weed pollen: 0.75–0.24). The average distance between monitoring stations in this study area, which represents one of the most populous regions of the country, was 288 km. The relatively poor spatial coverage of monitor stations combined with the range in frequency of pollen measurements in this region (99.1% of days for Albany, NY, to 32.7% for Olean, NY) hinder the reliability of using measurements from monitoring stations to estimate pollen concentrations at locations far from the monitors, given the spatial heterogeneity in concentrations.

Estimates of pollen in areas without monitors by spatial interpolation techniques have been demonstrated to be "...acceptably reliable when data provided by all stations falls within a fairly narrow range..." (Alba et al. 2006). This is in agreement with the findings of this study, which found no difference in the distribution of pollen estimates from kriging models and observed measurements when extreme peak concentrations were excluded. Peak counts of tree pollen, which were substantially higher than peaks of other pollen types in this region, were consequently estimated poorly. However, agreement between days observed to be within a certain pollen category level (*absent*, *very low*, *low*, *moderate*, *high*, *very high*) that were also estimated to be within the same pollen category by kriging methods were poor. Previous studies have found limitations in spatial interpolation of phonological processes

based on a small number of sampling locations, as is found in this study (Garcia-Mozo et al. 2006b). For spatial interpolation techniques, the degree of potential error increases as the distance between stations increases and the number of stations decreases (Alba et al. 2006). Further, the frequency of monitoring at each station varies. Days and areas with fewer measurements potentially have greater error in interpolated estimates.

Forecasting models for aeroallergens have focused on both short- and long-term predictions (Smith and Emberlin 2005, 2006). Other types of forecasting models addressed characteristics of pollen seasons such as overall season load, season start date and duration of the pollen season (Garcia-Mozo et al. 2006a; Antepara et al. 1995; Laaidi 2001; Emberlin et al. 1993, 2002; Fuertes-Rodriguez et al. 2007; Galan et al. 2001; Garcia-Mozo et al. 2000). Common among forecasting models to predict pollen levels is the use of weather variables, primarily temperature and measures of atmospheric moisture. While forecasting models have addressed temporal variation, they have largely failed to account for spatial variability.

Previous studies indicated that understanding land-use patterns is important to understanding regional aeroallergen exposures (Emberlin et al. 1993, 1999). Therefore, this study built on the temporal forecasting model framework by incorporating land cover and regional pollen season characteristics to address spatial variability in pollen estimation. Incorporation of pollen season characteristics, such as day of season and timing of peak periods, were also used to account for regional variability. Inclusion of such variables does add a layer of uncertainty with spatial estimation in that season start dates and peak periods must be estimated in areas where monitors do not exist, which are the areas of interest. Although season start dates have been reliably estimated (Emberlin et al. 2002, 2007; Laaidi 2001; Fuertes-Rodriguez et al. 2007; Garcia-Mozo et al. 2000), calculation of these parameters may be problematic in regions with limited data.

Correlations between observed and estimated pollen counts were generally higher for GEE model estimates compared to mixed effects models. Overall, estimates from either longitudinal regression model had correlations with observed measurements that were as high or higher than would be expected on average between sites even a few kilometers apart using the nearest monitor approach. However, regression models poorly estimated pollen levels on days with *absent* or *very high* levels, although the models performed well at less extreme values. The proportion of days with observed *very low*, *low*, *moderate* and *high* levels of tree and grass pollen that were also estimated to have a corresponding pollen level was generally similar for both mixed effects and GEE models. Only mixed effects models estimated any days of *high*

levels of weed pollen on which there were also observed *high* levels. Neither longitudinal model type estimated any days of *very high* weed pollen concentrations.

The main advantage of longitudinal models is that they estimate pollen concentrations on every day, regardless of whether monitoring data are available. Major disadvantages of these models are the amount of information required (weather data, characteristics of the pollen season, and land cover) and failure to capture very high concentrations.

Development of these models was limited by a number of factors. One limitation was in the reporting of the pollen counts themselves. Guidelines for placement of sampling equipment state they “be situated on an unobstructed rooftop at least one story above ground with no local pollen and/or mold spore sources” (AAAAI 2010). Height of equipment was not uniform across monitor stations. This is problematic because the types and quantity of pollen measured at a single location differ depending on the height of the monitoring equipment (Ranta et al. 2008). Time of day at which measurements are taken is not reported. Timing of measurements could affect measured concentrations, especially if it is inconsistent within or between stations.

Missing observations are another concern, as the frequency of measurement varies by site. No systematic differences in frequency of measurement were readily apparent at any given site. In other words, the proportion of days on which measurements were taken at a given monitor was consistent for all months in the pollen season.

Our focus was on the development of methods to estimate ambient pollen exposures and not on describing the dispersion of pollen. Given that general groupings of tree, grass and weed pollen have been used in epidemiologic studies (Cakmak et al. 2005; Dales et al. 2004), we investigated methods to estimate those pollen groupings as well as *Acer* and *Quercus* genera. It should be noted tree, grass, and weed pollen may represent different pollen compositions for various regions. Our analysis focused on a geographic region with similar meteorological conditions and genera of flora. General tree, grass or weed pollen groupings may not be appropriate for spatial-temporal models to estimate exposure over large or disparate spatial areas. Determining spatial-temporal scales and degrees of pollen heterogeneity for which exposure modeling is appropriate warrants further investigation.

Furthermore, the location of monitoring sites is not random. Monitoring sites are primarily located in populated areas and are not located in a random fashion as one would design experimentally if the goal were to describe the spatial distribution of pollen. There is also a paucity of monitoring sites in this region, as well as most of the U.S.

Multiple approaches are available to estimate pollen concentrations, including new methods such as those

presented here. Accuracy of pollen count estimates based on measurements from the nearest monitor station varied by distance to that monitoring station and type of pollen. The temporal nature of the exposure of interest is also a consideration. Frequency of measurement is a limiting factor in estimating daily pollen by closest monitor or spatial interpolation. However, if one's concern is exposures averaged over several days or weeks, then monitors or spatial interpolation methods may provide the best estimates. Spatial interpolation is also advantageous in describing spatial heterogeneity because it provides a continuous surface of estimates. The number of pollen measurements available and distance between stations are limiting factors. Longitudinal models are able to produce daily pollen estimates for any location, provided weather information is available, regardless of the availability of monitoring data. Additionally, these models may best describe pollen concentrations in areas a great distance from monitoring stations. Finally, the nature and degree of uncertainty a researcher is willing to accept in exposure estimates can affect which method is optimal. For example, our findings indicate that estimation of peak values, by any modeling technique, has greater uncertainty than estimates during non-peak periods.

While the models created in this study are only applicable to the study region, or regions with similar flora compositions, the methodology has value in application for other areas. These models could provide more accurate assessments of ambient exposures in studies investigating the health effects of pollen or in providing daily pollen estimates to populations in regions not near monitoring stations.

Acknowledgements We thank the National Allergy Bureau, and specifically Jerome Schultz, for providing daily ambient pollen measurements used in this study.

Ethical standards This research was conducted without human or animal subjects and did not violate any legal or ethical standards.

References

- AAAAI (American Academy of Allergy, Asthma & Immunology) (2010). National Allergy Bureau: Reading the Charts. http://www.aaaai.org/nab/index.cfm?p=reading_charts. Accessed 25 January 2010
- Alba F, Nieto-Lugilde D, Comtois P, de la Guardia CD, De Linares C, Ruiz L (2006) Airborne-pollen map for *Olea europaea* L. in eastern Andalusia (Spain) using GIS: estimation models. *Aerobiologia* 22(2):109–118. doi:10.1007/s10453-006-9024-0
- Antepara I, Fernandez JC, Gamboa P, Jauregui I, Miguel F (1995) Pollen allergy in the Bilbao area (European Atlantic seaboard climate) - pollination forecasting methods. *Clin Exp Allergy* 25(2):133–140
- Branzi GP, Zanotti AL (1992) Estimate and mapping of the activity of airborne pollen sources. *Aerobiologia* 8(1):69–74
- Cakmak S, Dales RE, Judek S, Coates F (2005) Does socio-demographic status influence the effect of pollens and molds on hospitalization for asthma? Results from a time-series study in 10 Canadian cities. *Ann Epidemiol* 15(3):214–218. doi:10.1016/j.annepidem.2004.06.001
- Dales RE, Cakmak S, Judek S, Dann T, Coates F, Brook JR, Burnett RT (2004) Influence of outdoor aeroallergens on hospitalization for asthma in Canada. *J Allergy Clin Immunol* 113(2):303–306. doi:10.1016/j.jaci.2003.11.016
- Emberlin J, Savage M, Jones S (1993) Annual variations in grass-pollen seasons in London 1961–1990 - trends and forecast models. *Clin Exp Allergy* 23(11):911–918
- Emberlin J, Mullins J, Corden J, Jones S, Millington W, Brooke M, Savage M (1999) Regional variations in grass pollen seasons in the UK, long-term trends and forecast models. *Clin Exp Allergy* 29(3):347–356
- Emberlin J, Detandt M, Gehrig R, Jaeger S, Noland N, Rantio-Lehtimäki A (2002) Responses in the start of *Betula* (birch) pollen seasons to recent changes in spring temperatures across Europe. *Int J Biometeorol* 46(4):159–170. doi:10.1007/s00484-002-0139-x
- Emberlin J, Smith M, Close R, Adams-Groom B (2007) Changes in the pollen seasons of the early flowering trees *Alnus* spp. and *Corylus* spp. in Worcester, United Kingdom, 1996–2005. *Int J Biometeorol* 51(3):181–191. doi:10.1007/s00484-006-0059-2
- Frenz DA (1999) Comparing pollen and spore counts collected with the Rotorod Sampler and Burkard spore trap. *Ann Allergy Asthma Immunol* 83(5):341–347
- Fuertes-Rodriguez CR, Gonzalez-Parrado Z, Vega-Maray AM, Valencia-Barrera RM, Fernandez-Gonzalez D (2007) Effect of air temperature on forecasting the start of Cupressaceae pollen type in Ponferrada (Leon, Spain). *Ann Agric Environ Med* 14(2):237–242
- Galan C, Carinanos P, Garcia-Mozo H, Alcazar P, Dominguez-Vilches E (2001) Model for forecasting *Olea europaea* L. airborne pollen in South-West Andalusia, Spain. *Int J Biometeorol* 45(2):59–63
- Garcia-Mozo H, Galan C, Gomez-Casero MT, Dominguez-Vilches E (2000) A comparative study of different temperature accumulation methods for predicting the start of the *Quercus* pollen season in Cordoba (South West Spain). *Grana* 39(4):194–199
- Garcia-Mozo H, Galan C, Jato V, Belmonte J, de la Guardia CD, Fernandez D, Gutierrez M, Aira MJ, Roure JM, Ruiz L, Trigo MM, Dominguez-Vilches E (2006a) *Quercus* pollen season dynamics in the Iberian Peninsula: Response to meteorological parameters and possible consequences of climate change. *Ann Agric Environ Med* 13(2):209–224
- Garcia-Mozo H, Galan C, Vazquez L (2006b) The reliability of geostatistic interpolation in olive field floral phenology. *Aerobiologia* 22(2):97–108. doi:10.1007/s10453-006-9026-y
- Goldberg C, Buch H, Moseholm L, Weeke ER (1988) Airborne pollen records in Denmark, 1977–1986. *Grana* 27(3):209–217
- Kuparinen A, Snäll T, Vanska S, O'Hara RB (2007) The role of model selection in describing stochastic ecological processes. *Oikos* 116(6):966–974
- Laaidi M (2001) Forecasting the start of the pollen season of Poaceae: evaluation of some methods based on meteorological factors. *Int J Biometeorol* 45(1):1–7
- Legendre P, Fortin MJ (1989) Spatial pattern and ecological analysis. *Vegetatio* 80(2):107–138
- Levetin E, Van de Water PK (2003) Pollen count forecasting. *Immunology and Allergy Clinics of North America* 23(3):423–+. doi:10.1016/s0889-8561(03)00019-5
- Lewis SA, Corden JM, Forster GE, Newlands M (2000) Combined effects of aerobiological pollutants, chemical pollutants and

- meteorological conditions on asthma admissions and A & E attendances in Derbyshire UK, 1993–96. *Clin Exp Allergy* 30 (12):1724–1732
- Liebhold AM, Rossi RE, Kemp WP (1993) Geostatistics and geographic information-systems in applied insect ecology. *Annu Rev Entomol* 38:303–327
- Nathan R, Sapir N, Trakhtenbrot A, Katul GG, Bohrer G, Otte M, Avissar R, Soons MB, Horn HS, Wikelski M, Levin SA (2005) Long-distance biological transport processes through the air: can nature's complexity be unfolded in silico? *Divers Distrib* 11 (2):131–137. doi:[10.1111/j.1366-9516.2005.00146.x](https://doi.org/10.1111/j.1366-9516.2005.00146.x)
- Ranta H, Sokol C, Hicks S, Heino S, Kubin E (2008) How do airborne and deposition pollen samplers reflect the atmospheric dispersal of different pollen types? An example from northern Finland. *Grana* 47(4):285–296. doi:[10.1080/00173130802457230](https://doi.org/10.1080/00173130802457230)
- Robertson GP (1987) Geostatistics in ecology - interpolating with known variance. *Ecology* 68(3):744–748
- Smith M, Emberlin J (2005) Constructing a 7-day ahead forecast model for grass pollen at north London, United Kingdom. *Clin Exp Allergy* 35(10):1400–1406
- Smith M, Emberlin J (2006) A 30-day-ahead forecast model for grass pollen in north London, United Kingdom. *Int J Biometeorol* 50 (4):233–242
- Stach A, Emberlin J, Smith M, Adams-Groom B, Myszkowska D (2008a) Factors that determine the severity of *Betula* spp. pollen seasons in Poland (Poznan and Krakow) and the United Kingdom (Worcester and London). *Int J Biometeorol* 52 (4):311–321. doi:[10.1007/s00484-007-0127-2](https://doi.org/10.1007/s00484-007-0127-2)
- Stach A, Smith M, Baena JCP, Emberlin J (2008b) Long-term and short-term forecast models for Poaceae (grass) pollen in Poznan, Poland, constructed using regression analysis. *Environ Exp Bot* 62(3):323–332. doi:[10.1016/j.envexpbot.2007.10.005](https://doi.org/10.1016/j.envexpbot.2007.10.005)
- Valencia-Barrera RM, Comtois P, Fernandez-Gonzalez D (2002) Bioclimatic indices as a tool in pollen forecasting. *Int J Biometeorol* 46(4):171–175. doi:[10.1007/s00484-002-0138-y](https://doi.org/10.1007/s00484-002-0138-y)
- Waller LA, Gotway CA (2004) Applied spatial statistics for public health data. John Wiley & Sons, Inc, Hoboken, NJ
- Weber RW (2003) Meteorologic variables in aerobiology. *Immunology and Allergy Clinics of North America* 23 (3):411–+. doi:[10.1016/s0889-8561\(03\)00062-6](https://doi.org/10.1016/s0889-8561(03)00062-6)