Estimating sampling frequency in pollen exposure assessment over time†

Junxiang Luo,[a] Rakesh Shukla,*[a] Atin Adhikari,[b] Tiina Reponen,[b] Sergey A. Grinshpun,[b] Qi Zhang[a] and Grace K. LeMasters[a]

Received 18th April 2006, Accepted 13th June 2006
First published as an Advance Article on the web 27th June 2006
DOI: 10.1039/b605440d

A time series model was fitted to the pollen concentration data collected in the Greater Cincinnati area for the Cincinnati Childhood Allergy and Air Pollution Study (CCAAPS). A traditional time series analysis and temporal variogram approach were applied to the regularly spaced databases (collected in 2003) and irregularly spaced ones (collected in 2002), respectively. The aim was to evaluate the effect of the sampling frequency on the sampling precision in terms of inverse of standard error of the overall level of mean value across time. The presence of high autocorrelation in the data was confirmed and indicated some degree of temporal redundancy in the pollen concentration data. Therefore, it was suggested that sampling frequency could be reduced from once a day to once every several days without a major loss of sampling precision of the overall mean over time. Considering the trade-offs between sampling frequency and the possibility of sampling bias increasing with larger sampling interval, we recommend that the sampling interval should take values from 3 to 5 days for the pollen monitoring program, if the goal is to track the long-term average.

1. Introduction

The impact on human health of some naturally occurring ambient environmental exposures such as pollen is attracting considerable interest of researchers. The association between airborne pollen and allergic responses has been widely recognized.[1] It is necessary to measure the environmental exposure precisely in order to understand its health effects accurately. Usually, researchers monitor the exposure over time. Pollen monitoring networks usually collect daily pollen samples during the pollen season.[2–4] Due to the conflict between expanding environmental monitoring and constraints of limited budgets, a key question is: How can we optimize a sampling design for environmental exposure monitoring so that limited resources are not spent on unnecessary sampling and analysis?

Other studies have been conducted to establish cost-effective sampling programs for exposures over time. Peretz[5] fitted a nested unbalanced analysis of variance model to estimate the magnitude of the variability in workers' exposure to lead, benzene and dust over time by analyzing repeated measurements over time nested in worker, nested in factory, and nested in air contaminant. In our studies,[6–8] we also employed variance component analysis (VCA) to investigate the optimal temporal sampling allocation for ambient particles and aircraft maintenance workers' exposure to solvents, primarily 1,1,1-trichloroethane. This method requires a pilot study at first, and then variance components of season, month, week and day can be estimated through VCA. Based on those variance components, designs with different combinations of numbers of season, month, week, and day can be probed to achieve an optimal design with a specific precision in terms of standard error of estimated mean. This method assumes, however, that the sample collections in the pilot study should be adequately separated temporally so that the autocorrelation between adjacent samples could be ignored.

If the autocorrelation between data points that are temporally close to each other cannot be ignored, a time series approach that incorporates the autocorrelation may be preferable for optimizing the sampling design. In actual studies, time series data from the pilot study may be regularly spaced or irregularly spaced. In the former case, samples are evenly scattered over time, while in the latter case, these are not, due to irregular monitoring or missing data points. In our ongoing Cincinnati Childhood Allergy and Air Pollution Study (CCAAPS), pollen concentrations monitored over time in 2002 and 2003 in the Cincinnati metropolitan area show high autocorrelation. Furthermore, the pollen concentrations in 2002 were collected very irregularly, while the data in 2003 were regularly spaced. Usually, the statistical analysis of these two kinds of time series data presents challenges, especially for irregularly spaced time series. The variance of the sample mean $\hat{\text{var}}(\bar{Y})$ of independent measurements can be estimated by dividing the sample variance $\hat{\text{var}}(Y)$ by the number of measurements, N. In correlated time series data, however, the variance of the sample mean would be an underestimate if the correlation is present.

For regularly spaced time series, the "effectively" independent sample size, N_{eff}, was introduced for estimating the variance of sample mean in a paper by Somerville and Evans.[9]

[a] *Division of Epidemiology & Biostatistics, Department of Environmental Health, University of Cincinnati, Cincinnati, Ohio, USA. E-mail: Rakesh.Shukla@uc.edu; Fax: +1-513-558-6272; Tel: +1-513-558-0108*
[b] *Center for Health-Related Aerosol Studies, Department of Environmental Health, University of Cincinnati, Cincinnati, Ohio, USA*

† The HTML version of this article has been enhanced with colour images.

N_{eff} reflects the "effective" degrees of freedom for the data, which can never exceed N. The higher autocorrelation leads to the smaller N_{eff}. The estimated sample variance can be divided by N_{eff} to estimate the variance of sample mean over time. In Somerville and Evans's paper, the first order autoregressive structure, AR(1), was applied to estimate N_{eff} in a survey study of atmospheric fine mass; subsequently, the standard error of the mean was computed. However, the AR(1) correlation structure is not suitable for our pollen data collected in 2003, since the preliminary analysis shows that the ARMA(1,1) accounts for the data pattern better than AR(1). So, we extended the method and used ARMA(1,1), which seems more appropriate for assessing the autocorrelation pattern existing in the pollen concentrations over time.

Since the pollen concentration database collected in 2002 was irregularly spaced, standard time series approach is not appropriate. Instead, variogram analysis that is extensively utilized in geostatistics to evaluate spatial correlation can be successfully used here. Variogram may also be applied to irregularly spaced time series, since the information of temporal distances can be evaluated. There are several examples of applying variogram to sampling issues. Cameron and Hunter[10] developed a spatial and temporal algorithm (geostatistical temporal–spatial or GTS) to optimize long-term ground water monitoring networks through analyzing spatial and temporal variograms. Temporal redundancy was supposed to be reduced by lengthening the time between sample collections. Lwin[11] applied variogram analysis to sampling issues from continuous streams in mineral processing, where sampling units are increments of a volume or mass of specific mineral particles. Lwin developed a weighted sample mean to cope with irregularly spaced sampling intervals and then assessed the magnitude of sampling variation.

In the present paper, the pollen concentrations from the CCAAPS project were treated as pilot data. We assessed alternative designs with different sampling frequencies (in terms of inverse of sampling interval) with respect to the *cost* of sampling (in terms of the numbers of samples) and *precision* (in terms of the inverse of the standard error of the mean level of exposure over time). In an attempt to reduce the sampling frequency over time, we applied the modified time series analysis and variogram approach to pollen concentrations measured in 2003 and 2002 in CCAAPS, respectively.

2. Methods

2.1 Pollen concentration data

The primary goal of CCAAPS is to investigate if exposures to diesel exhaust particles (DEP) play a role in allergy and asthma in infants and young children. Because DEP may act as a confounder with allergens, this study required monitoring the ambient pollen concentrations over time. Pollen grains were collected with the button inhalable aerosol sampler daily from March 5 to the end of September in 2002 (Fig. 1) at two sites (Grooms and Taft) and from March 4 to November 25 in 2003 (Fig. 2) at Taft alone. The site of Grooms is located 12 miles north of downtown Cincinnati, while Taft is about three miles north of the downtown Cincinnati. At Grooms, the button sampler was installed on a wooden pole at the height of 3 m, whereas at Taft, the sampler located on the rooftop of a two-storied office building with the height of 7 m. At both sites, 24 hour air samples were collected with the inlet of the button sampler oriented towards the south-west, which is the predominant wind direction in Greater Cincinnati.[12] There were no tall buildings in the proximity to allow free air movement. After collection, pollens were counted under a microscope, and the counts were transformed into concentrations (pollen grains m^{-3}) by specific formulas. Sampling and analysis methods were described in detail by Adhikari *et al.*[12,13]

2.2 Statistical modeling and analyses

Time series analysis is a technique developed to analyze changes in a variable over time as an attempt to find patterns and relationships in the data. Data collected over time are usually auto-correlated (serially correlated), indicating that data points closer to each other often have higher correlation.[14]

For the pollen concentrations in 2002 and 2003, the data showed in Fig. 1 and Fig. 2 have periodical pattern, and it is usually assumed that the pollen concentration has a log-normal distribution. Thus, we assume the logarithmic transformed pollen concentration at time t denoted by Y_t, $t = 1,\ldots,n.$ to have the following model:

$$Y_t = \mu + \beta_1 t + \beta_2 \sin(2\pi t\nu) + \beta_3 \cos(2\pi t\nu) + \varepsilon_t \qquad (1)$$

Where t is the temporal distance (in days) from the first sampling date, $\mu + \beta_1 t$ accounts for the linear trend, sin and cos terms represent the seasonal periodicity, ε_t is the error term, μ, β_2 and β_3 are regression parameters, and ν denotes the periodicity parameter.

2.2.1. Regularly spaced time series technique

2.2.1.1 Variance of overall mean over time in pilot design. For the pollen concentrations in 2003, we assume the error term ε_t has an ARMA(1,1) structure. It means $\varepsilon_t = \phi\varepsilon_{t-1} + e_t - \theta e_{t-1}$, $e_t \sim \mathrm{NID}(0, \sigma^2)$. ARMA(1,1) is the mixture of the first-order autoregressive, AR(1), and the first-order moving average process, MA(1). The values of ϕ and θ are the autoregressive and moving average parameters, respectively, which measure the association between Y_t and Y_{t-1} after adjusting the seasonal cycle and linear trend. The autocorrelation in terms of parameters of ϕ and θ is given by

$$\begin{cases} \rho_i = (\phi - \theta)(1 - \phi\theta)/(1 + \theta^2 - 2\phi\theta),\ i = 1 \\ \rho_i = \phi\rho_{i-1},\ i \geq 2, \end{cases} \qquad (2)$$

where ρ_i represents the lag-i autocorrelation coefficient.

Our interest is to estimate the variance of overall sample mean denoted by $\mathrm{var}(\bar{Y})$. Due to $\mathrm{var}(\bar{Y}) = \sigma^2\{[n + 2\sum_{i=1}^{n}(n-1)\rho_i]/n^2\}$,[15] we express $\mathrm{var}(\bar{Y})$ and effective independent sample size, n_{eff} , for the data with ARMA(1,1) correlation structure as

$$\begin{cases} \mathrm{var}(\overline{Y}) = \sigma^2\left\{n + \frac{2\rho_1}{(1-\phi)^2}[n(1-\phi) - (1-\phi^n)]\right\}/n^2 \\ n_{\mathrm{eff}} = n^2/\left\{n + \frac{2\rho_1}{(1-\phi)^2}[n(1-\phi) - (1-\phi^n)]\right\} \end{cases} \qquad (3)$$

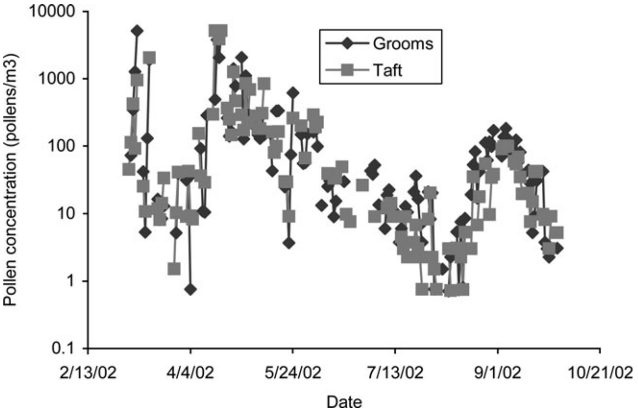

Fig. 1 Pollen concentration over time in 2002 by site.

SAS/ETS is a component of the SAS system (SAS, 9.1). It includes SAS procedures for econometric analysis and time series analysis, The MODEL procedure in SAS/ETS analyzes models in which the relationships among the variables comprise a system of one or more nonlinear equations and the error term can be a time series process. Therefore, PROC MODEL is appropriate to fit the model in eqn (1) and estimate the parameters (μ, β_1, β_2, β_3, ν, ϕ, θ and σ^2).

2.2.1.2 Cost and precision for potential designs. Based on the parameter estimates from the pilot study, we calculate the variance of the mean value for a potential design with specific sampling frequency (*i.e.* pollen concentrations monitored once every k days). The variance $\mathrm{var}(\bar{Y}_k)$ can be estimated by

$$\hat{\mathrm{var}}(\bar{Y}_k) = \frac{\hat{\sigma}^2}{n_k^2}\left[n_k^2 + 2\sum_{i=1}^{n_k}(n_k - i)\rho_i^*\right] \tag{4}$$

where $\rho_1^* = \frac{(\phi-\theta)(1-\phi\theta)}{(1+\theta^2-2\phi\theta)}\phi^{k-1}$, $\rho_i^* = \phi^{k-1}\rho_{i-1}^*$, $i \geq 2$; $n_k = n/k$ denotes the number of the samples monitored over time in this potential design.

Since, the precision in the estimate of the overall mean is defined as the inverse of the standard error of the overall mean, and since the sampling cost is directly proportional to the total number of samples monitored, we obtained the relative cost (RC) and relative precision (RP) of any specific potential design with sampling scenario of one measurement for every k days *vis à vis* the pilot design.

$$\begin{cases} \mathrm{RC} = \frac{\text{number of samples(design)}}{\text{number of samples(pilot)}} \times 100\% = \frac{n_k}{n} \times 100\% \\ \mathrm{RP} = \frac{\text{precision(design)}}{\text{precision(pilot)}} \times 100\% = \frac{\mathrm{SE(pilot)}}{\mathrm{SE(design)}} \times 100\% = \sqrt{\frac{\mathrm{var}(\bar{Y})}{\mathrm{var}(\bar{Y}_k)}} \\ \times 100\% \end{cases} \tag{5}$$

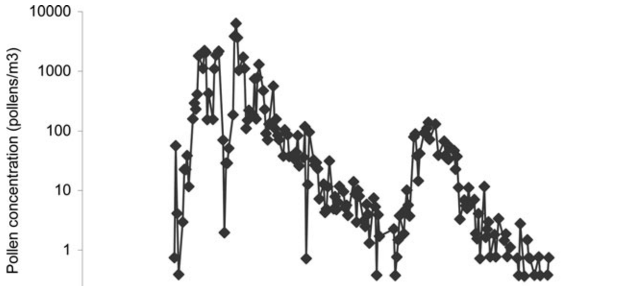

Fig. 2 Pollen concentration over time in 2003 at Taft.

2.2.2. Temporal variogram for irregularly-spaced time series

2.2.2.1 Temporal variogram. A variogram is a function used in geostatistics for describing the spatial or temporal correlation among observations. It is directly related to the covariance function. The variogram is important as it is used to fit a model of the spatial/temporal correlation in any observed phenomenon. In geostatistics, the variogram[16–20] is defined as a measure of the continuity of spatial phenomena expressed as half the variance of the difference between measured quantities at different locations x and $x + h$. If $Z(x)$ is denoted as a measured quantity at location x, the mathematical definition of the variogram is

$$\begin{cases} r(h) = \frac{1}{2}\mathrm{var}\{Z(x+h) - Z(x)\} \\ r(h) = \frac{1}{2}E[\{Z(x+h) - Z(x)\}^2] \text{ when} E\{Z(x+h) - Z(x)\} \\ \quad = 0 \end{cases} \tag{6}$$

In variogram analysis, $Z(x)$ is assumed to be second-order stationary, which means it satisfies two conditions simultaneously. Firstly, $E\{Z(x)\}$ exists and does not depend on location x. It means $E\{Z(x)\} = \mu$, for all x. Secondly, for each pair of quantities, $\{Z(x), Z(x + h)\}$, the covariance exists and depends only on the separation vector h. The stationarity of the covariance implies the stationarity of the variance. It means $\mathrm{var}\{Z(x)\} = \sigma^2$, for all x.

The variogram may also be applied to dealing with irregularly spaced time series in time domain. Let measurement at time t_i be denoted by Y_{t_i}, ($i = 1,2,\ldots,n$) having a distribution with $E(Y_{t_i}) = \mu$ and $\mathrm{var}(Y_{t_i}) = \sigma^2$. Define temporal variogram as $r_h = \frac{1}{2}E[\{Y_{t_i} - Y_{t_j}\}^2]$, where $h = t_i - t_j$ is the distance in time between two measurements. The variogram is estimated by the average of observed half-squared-differences between pairs of measurements corresponding to that particular distance class h.

Variogram analysis consists of the experimental variogram calculated from the data and the variogram model fitted to the data. The variogram model is chosen from a set of mathematical functions that describe spatial relationships. The appropriate model is chosen by matching the shape of the curve of the experimental variogram to the shape of the curve of the mathematical function.

2.2.2.2 Variance of weighted overall mean over time in pilot study. For the pollen concentrations in 2002, since the time series at Grooms and Taft are very irregularly spaced, the traditional regularly spaced time series approach is not applicable, where the commonly used variogram analysis is appropriate. In traditional variogram analysis, the spatial autocorrelation takes the distance between points in space into account. Similarly, variogram analysis can be employed on irregularly spaced time series data dealing with the temporal autocorrelation and considering the temporal distance between samples collected over time. We call the variogram analysis for the data in the time domain 'temporal variogram analysis'.

We fit the model presented above by eqn (1) and call the first part, $\mu + \beta_1 t + \beta_2 \sin(2\pi t\nu) + \beta_3 \cos(2\pi t\nu)$, a drift. Then, the variogram that measures the temporal autocorrelation is estimated on the error item ε_t, so it is called a residual variogram. In other words, the residual variogram is drift-adjusted variogram. It is unnecessary to assume that the error item ε_t has any correlation pattern such as ARMA(1,1), but it should be stationary, which means the expectation and variance of ε_t are constant over time. The residual variogram is defined as $r_h = \frac{1}{2}E[\{\varepsilon_{t_i} - \varepsilon_{t_j}\}^2]$; $h = |t_i - t_j|$ is the temporal distance between two samples. The sample version of r_h (called a sample residual variogram or experimental residual variogram) is estimated as follows:

$$\hat{r}_h = \frac{1}{2n_h} \sum_{i,j:|(t_j - t_i) - h| \le \delta} (\varepsilon_{t_j} - \varepsilon_{t_i})^2, \text{ for all } i,j : (t_j \ge t_i) \qquad (7)$$

The VARIOGRAM function in the R package is employed to obtain the sample residual variogram. The R package, similar to S-plus, is a language and computing environment for statistical analyses and graphics. The R package is available as free standing software under the terms of the Free Software Foundation's GNU General Public License in source code form.

Theoretically, r_h is directly related to the autocorrelation function ρ_h

$$r_h = \begin{cases} 0, & h = 0 \\ \sigma^2[1 - \rho_h], & h \ne 0 \end{cases} \qquad (8)$$

where ρ_h represents lag-h autocorrelation coefficient

The residual variogram at $h = 0$ is by definition 0. It often occurs, however, that the residual variogram of ε_t cannot be represented by a completely continuous function since its value at $h = 0$ may not be lower than a positive quantity σ_0^2. The phenomenon appears often in geological and environmental sampling and is called the nugget effect or nugget variance. Any apparent nugget variance usually arises from measurement errors or variation within the shortest sampling interval.[18] Thus, a realistic r_h is represented by

$$r_h = \begin{cases} \sigma_0^2, & h = 0 \\ \sigma_0^2 + \sigma^2[1 - \rho_h], & h \ne 0 \end{cases} \qquad (9)$$

With the sample residual variogram in hand, we estimate $\hat{\sigma}_0^2$ and $\hat{\sigma}^2$ in eqn (9) by modeling r_h. In general, three types of models are used widely: linear, spherical, and exponential. They are defined as follows, respectively:

- Linear: $r_h = \begin{cases} \sigma_0^2 + \sigma^2 h/b, & h \le b \\ \sigma_0^2 + \sigma^2, & h > b \end{cases}$

- Spherical: $r_h = \begin{cases} \sigma_0^2 + \sigma^2\left[1.5\frac{h}{b} - 0.5\left(\frac{h}{b}\right)^3\right], & h \le b \\ \sigma_0^2 + \sigma^2, & h > b \end{cases}$ (10)

- Exponential: $r_h = \sigma_0^2 + \sigma^2\,[1 - \exp(-h/b)]$

The estimator of nugget effect $\hat{\sigma}_0^2$ is the intercept at $h = 0$, and $\hat{\sigma}_0^2 + \hat{\sigma}^2$ is the sill of the fitted curve. The sill exists on the curve because the autocorrelation coefficient $\rho_h \rightarrow 0$ when temporal distance between pairs of samples is beyond b (*i.e.*, b serves as the range).

The next step is to estimate the variance of the overall mean for the irregularly spaced time series in the pilot study. Lwin[11] showed that a weighted mean $\bar{Y}_w$ gave more reasonable information than the algorithm mean $\bar{Y}$ about the average value in an irregularly spaced time series data. Thus, we need to estimate var($\bar{Y}_w$). We define

$$\bar{Y}_w = \sum w_i\, Y_{ti},\ i = 1, \ldots, n. \qquad (11)$$

where $w_1 = \Delta t_2/(2T)$; $w_i = (\Delta t_{i+1} + \Delta t_i)/(2T)$, $i = 2,\ldots,n-1$; $w_n = \Delta t_n/(2T)$; $\Delta t_i = t_i - t_{i-1}$ denotes the temporal distance of two consecutive samples; and $T = t_n - t_1$.

Suppose the nugget effect comes from the measurement errors, thus the variance of $\bar{Y}_w$ is given by

$$\begin{aligned} \mathrm{var}(\bar{Y}_w) &= \sum_{i=1}^{n} \mathrm{var}(w_i Y_{t_i}) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_i w_j \,\mathrm{cov}(Y_{t_i}, Y_{t_j}) \\ &= \left(\sum_{i=1}^{n} w_i^2\right)(\sigma_0^2 + \sigma^2) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_i w_j(\sigma_0^2 + \sigma^2 - r_{(t_j - t_i)}) \\ &= \sigma_0^2 + \sigma^2 - 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_i w_j r_{(t_j - t_i)} \end{aligned} \qquad (12)$$

2.2.2.3 Cost and precision for potential designs. For the design with sampling interval equal to k days, $\hat{\rho}_h$ can be computed by eqn (9), using $\hat{\sigma}_0^2$, $\hat{\sigma}^2$, and $\hat{r}_h$ from the fitted model in eqn (10). Thus, the variance of the overall mean var($\bar{Y}_k$) for the potential design can be estimated by applying $\hat{\sigma}_0^2$, $\hat{\sigma}^2$ and $\hat{\rho}_h$, to the following formula:

$$\hat{\mathrm{var}}(\bar{Y}_k) = \frac{\hat{\sigma}_0^2 + \hat{\sigma}^2}{n_k^2}\left[n_k + 2\sum_{i=1}^{n_k}(n_k - i)\rho_i^*\right] \qquad (13)$$

where $\rho_i{}^* = \hat{\rho}_{k*i}$ denotes the autocorrelation between measures with $k*i$ days lag, and n_k is the sample size of the specific design. The relative cost and relative precision can be calculated in the same way as described in Section 2.2.1.2, after obtaining var($\bar{Y}_\mathrm{k}$).

3. Results

3.1 Time series approach for regularly-spaced observations

Fig. 3 presents a plot of the predicted values from the fitted model *versus* the observed log-transformed pollen

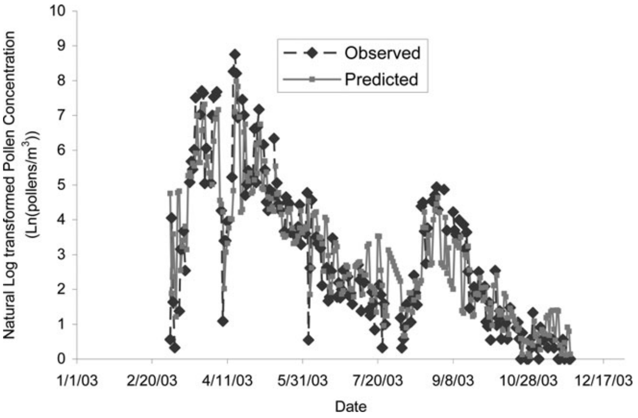

Fig. 3 Model predicted and observed pollen concentrations in 2003.

concentrations. It shows that the model accounts for most of the data pattern and its periodicity. The plots of residuals from the fitted model *versus* the predicted values and residuals *versus* the time did not indicate heterogeneity of variance. Table 1 shows the estimates of model parameters. The estimate of θ is 0.23 (with the P value = 0.06), which confirms that the ARMA(1,1) correlation structure is more reasonable for the data than AR(1). The adjusted R^2 (0.73) also suggests that the fitted model accounted for most of the trend and variation among the pollen concentrations in 2003. Negative estimates of linear coefficient β_1 (-0.02, $p < 0.001$) and coefficient β_3 (-0.59, $p = 0.03$) show that the pollen concentrations of 2003 have some linear decreasing trend with time and displayed periodicity.

Table 2 lists the variances of estimates of overall mean for the pilot design and other potential designs with sampling interval equal to k, $k = 1, \ldots, 10$ days. Those variances were calculated using eqn (3) and (4). The relative costs and relative precisions were obtained in terms of number of samples and the inverse of the standard error of the overall mean, respectively. Due to high autocorrelation ($\phi = 0.91$) in the data, relative precisions are still very high for those potential designs. For instance, if we monitor the pollen concentrations once every 3 days, the estimate of the overall mean would still have 99.4% precision compared to the daily monitoring program (pilot design). If monitoring is done every 10 days the precision is 93.7% while the cost is 90% less.

Due to the large value of the autocorrelation parameter (*e.g.*, $\phi = 0.91$ in pollen data for 2003), we can still obtain very high precision for the estimate of sample mean even though the sampling frequency reduces to one data point per 10 days. It is possible, however, that, the bias may become serious in reduced samples, since it usually exists in systematic sampling strategies. In order to check bias in our study, we calculated sample means with different sampling frequencies based on the log-transformed pollen concentrations obtained in 2003. We also computed the sample means with different starting points within a selected sampling interval. For example, if the sampling interval was 3 days, we may use the first, second or third date (*i.e.* 3/5/2003, 3/6/2003, or 3/7/2003) as the starting point. Table 3 shows the sample means with different starting points within different sampling intervals. Daily sampling (with sample mean equal to 3.04) represents the conditions implemented in our pilot study. Results show that sampling with every other day provides very close sample means to that of a daily sampling scenario. With the sampling interval increasing, the bias of the sample mean also rapidly increases. In order to control the bias, we recommend the sampling interval should be less than 5 days.

3.2 Temporal variogram approach for irregularly-spaced observations

The pollen concentrations were not regularly monitored at the Grooms and Taft locations from 5th March 2002 to 30th September 2002. The samples were collected for 125 out of 210 days at Grooms, 117 out of 210 days at Taft. Obviously, the two time series are irregularly spaced, so temporal variogram analysis was applied to these data. We also fitted the model given in eqn (1), and then estimated sample residual variogram from the fitted model.

The function of VARIOGRAM in the library of Gstat of the R package was used to estimate the sample residual variograms. Subsequently, we used functions of FIT.VARIOGRAM and VGM to model the sample residual variograms. Three models (spherical, exponential, and linear) were selected after checking the curve patterns of the sample residual variograms. The estimates of parameters for each of the three models are shown in Table 4 and the corresponding graphs are plotted in Fig. 4.

Using the estimates of $\hat{\sigma}_0^2$, $\hat{\sigma}^2$, and $\hat{b}$, $\hat{r}_h$, was calculated for different h values. Then $\hat{\text{var}}(\bar{Y}_w)$ was calculated for the pilot design by using eqn (12), and $\hat{\text{var}}(\bar{Y}_k)$ was determined for potential designs (sampling interval equal to k days) by implementing eqn (13), respectively. The relative costs and relative precisions of potential designs were obtained as compared to the pilot design (see Table 5). In the pilot study, we irregularly monitored pollen concentrations on 125 days from March to November of 2002 at the site of Grooms and 117 days at Taft. At Grooms, if daily concentrations were collected, the precision (in terms of the inverse of standard error) of the estimate of the overall mean increased by 2.3% assuming the residual variogram follows spherical model; the cost, however, increased by 67.2%, compared to the pilot design. If

Table 1 Estimates of the parameters in the model for pollen concentrations in 2003

Parameter[a]	μ	β_1	β_2	β_3	ν	ϕ	θ	σ^2
Estimate	5.4	−0.02	−0.21	−0.59	0.03	0.91	0.23	1.31
P value	<0.0001	<0.0001	0.637	0.027	<0.0001	<0.0001	0.06	—

[a] Parameters in the model given in eqn (1), where μ = intercept; β = regression parameters; ν = the periodical parameter; ϕ = autoregressive parameter; θ = moving average parameter; σ^2 = the scale of the noise of the time series.

Table 2 Relative precisions for designs with various sampling intervals (2003)

k^a/day	$N_k{}^b$	$N_{\text{eff}}{}^c$	var($\bar{\text{Y}}$)	Relative cost (%)	Relative precision (%)
1 (pilot)	267	14.0	0.0928	100.0	100.0
2	133	13.9	0.0937	50.0	99.5
3	89	13.8	0.094	33.3	99.4
4	66	13.6	0.0959	25.0	98.4
5	53	13.5	0.0966	20.0	98.0
6	44	13.3	0.0981	16.7	97.3
7	38	13.2	0.0987	14.3	97.0
8	33	12.9	0.1008	12.5	95.9
9	29	12.6	0.1035	11.1	94.7
10	26	12.3	0.1057	10.0	93.7

[a] Sampling interval. [b] Sample size for design with sampling interval k. [c] Effective independent sample size with sampling interval k.

Table 3 Sample means of log-transformed pollen concentrations (2003) determined for different sampling intervals

	Starting point				
Sampling interval (k)	1	2	3	4	5
1	3.04				
2	3.04	3.05			
3	3.09	3.03	3.01		
4	3.07	3.13	3.01	2.96	
5	2.75	2.96	3.25	3.05	3.22

one sample was collected every three days, the cost decreases by 45% with a loss of only 0.8% in precision (again, assuming the spherical model for residual variogram). Similar calculations were applied to other designs and variogram models as presented in Table 5. The results indicate that there is no large difference between models with respect to sampling precision. Similar to the results of the data analysis of the pollen concentrations measured in 2003, temporal variogram analysis to the data collected in 2002 revealed that the relative sampling precision was still very high even if the sampling interval increases to 5–10 days. Therefore, unnecessary samples largely exist in the pilot study due to the high autocorrelation.

4. Discussion and conclusion

In survey sampling theory, fixed-interval samples (*e.g.*, those taken every k days) are called systematic samples. A systematic sample includes one of the first k units (days in our study) chosen at random, and every unit at intervals of k after the first. Each of the k possible samples is a cluster of units covering the population. Systematic sampling is often used in monitoring programs because it is easy to apply. This study managed to obtain an optimal systematic sampling strategy through applying the time series analysis and the temporal variogram approach to autocorrelated data. The results reported here have important implications for planning sampling networks for ambient exposures. As frequency of sampling in systematic sampling increases, more redundancy exists because of the autocorrelation. Increasing the time interval between sampling reduces temporal redundancy. In order to obtain optimal sampling frequency while balancing cost and precision, a time series analysis from a pilot study is needed in order to apply the information obtained from a pilot study to potential designs. The present study utilized two different methods (traditional time series analysis and temporal variogram approach) for pollen concentration data collected during 2003 and 2002 monitoring. We have extended the method used by Somerville and Evans[9] and selected ARMA(1,1) to represent the autocorrelation structure for the 2003 pollen concentration database after taking seasonality into account. To the best of our knowledge, no previous study has employed correlation structures such as ARMA (p, q) ($p \geq 2$, $q \geq 2$) to investigate the issue of sampling frequency. This is partly because of the difficult algorithms of calculating variance of overall mean over time with complex correlation structures. In time series analysis, the data should be second order stationary (constant expectation and unchanged variance over time), which is a challenging assumption. For instance, since the log-transformed 2003 pollen concentration data of this study exhibited seasonal periodicity, a technique removing the trend was used by fitting a specific model to ensure that the residual from the model is a stationary time series. This technique, however, requires satisfied assumption of variance. Other techniques can cope with the time series with heterogeneity of variance. For example, Belmonte and Canela[21] developed two satisfactory methods, a Gaussian mixture model (parametric) and the Friedman super smoother (non-parametric), for fitting a smooth trend to annual series of mean daily concentration of Urticaceae pollen, in which the deviations are non-stationary, at six sampling sites. Those methods are useful for time series data for forecasting and extracting signals, but these methods may be difficult to implement in sampling frequency analysis.

Analysis of irregularly spaced data sets is more complicated than that of regularly spaced ones. Our investigation used a temporal variogram approach to explore directly the characteristic of autocorrelation for irregularly spaced time series data (*i.e.* pollen concentrations monitored in 2002). Other analytic approaches can be applied to these data. For instance, one might consider using regularization techniques to make a

Table 4 Estimates of the parameters of variogram models

		Grooms				Taft	
	Parameters	Spherical	Exponential	Linear	Spherical	Exponential	Linear
Nugget effect	$\hat{\sigma}_0^2$	0.88	0.79	0.92	0.72	0.42	0.58
Variance	$\hat{\sigma}^2$	2.17	2.54	2.19	2.42	2.96	2.65
Range	$\hat{b}$	22.52	13.77	20.15	24.13	14.22	22.30

Table 5 Relative precisions for designs with various sampling intervals (2002)

				Spherical		Exponential		Linear	
Site	K^b	$N_k{}^c$	Relative cost (%)	Variance of overall mean[a]	Relative precision (%)	Variance of overall mean[a]	Relative precision (%)	Variance of overall mean[a]	Relative precision (%)
Grooms	Pilot	125	100	0.183	100.0	0.325	100.0	0.217	100.0
	1	209	167.2	0.175	102.3	0.317	101.3	0.209	101.9
	2	104	83.2	0.18	100.8	0.322	100.5	0.214	100.7
	3	69	55.2	0.186	99.2	0.328	99.5	0.221	99.1
	4	52	41.6	0.19	98.1	0.332	98.9	0.224	98.4
	5	41	32.8	0.198	96.1	0.342	97.5	0.231	96.9
	6	34	27.2	0.205	94.5	0.349	96.5	0.241	94.9
	7	29	23.2	0.212	92.9	0.356	95.5	0.245	94.1
	8	26	20.8	0.213	92.7	0.354	95.8	0.249	93.4
	9	23	18.4	0.22	91.2	0.363	94.6	0.255	92.2
	10	20	16	0.234	88.4	0.381	92.4	0.262	91.0
Taft	Pilot	117	100.0	0.212[a]	100.0	0.382[a]	100.0	0.28[a]	100.0
	1	210	179.5	0.206	101.4	0.376	100.8	0.274	101.1
	2	105	89.7	0.21	100.5	0.378	100.5	0.277	100.5
	3	70	59.8	0.214	99.5	0.381	100.1	0.281	99.8
	4	52	44.4	0.22	98.2	0.388	99.2	0.287	98.8
	5	42	35.9	0.223	97.5	0.388	99.2	0.289	98.4
	6	35	29.9	0.227	96.6	0.392	98.7	0.292	97.9
	7	30	25.6	0.232	95.6	0.396	98.2	0.295	97.4
	8	26	22.2	0.24	94.0	0.405	97.1	0.303	96.1
	9	23	19.7	0.247	92.6	0.411	96.4	0.312	94.7
	10	21	17.9	0.249	92.3	0.411	96.4	0.31	95.0

[a] Var($\bar{Y}_k$) for design with sampling interval k; var($\bar{Y}_w$) for the pilot design. [b] Sampling interval. [c] Sample size with sampling interval k.

given irregularly sampled data series onto a regular grid, in order to use conventional tools for further analysis. The techniques require some form of interpolation or estimation (*e.g.*, linear interpolation and smooth interpolation), which effectively constructs an "underlying" continuous function representing the discrete data. The goal is to use an interpolation/estimation method, which preserves the relevant information as much as possible. After interpolation, traditional time series approaches can be used on the regularly spaced time series data. We, however, employed a temporal variogram approach directly with 2002 pollen concentrations data without interpolation, since almost half of the data points were missing. In temporal variogram analysis, it is important to select variogram models. Unfortunately, there are no effective model selection criteria. Instead, one selects models based on the curve of sample variogram. Models frequently used in literature include linear, spherical, exponential, circular, Gaussian, logarithmic, spline. We tried three types of models (linear, spherical and exponential) in the present paper and found no large differences between them with respect to sampling precision.

In this study, the precision is expressed as the inverse of standard error of the estimate of overall mean across time. This index is easy to apply and was also used in our previous studies.[6–8] Some other investigations (*e.g.*: in Cameron and Hunter's paper[10]) employed the kriging technique, which evaluates average kriging variance to optimize sampling program and has been frequently applied to spatial data.

We found from the pollen concentration data collected in 2002 and 2003 that relative precision exceeded 95% with a 5 day sampling interval in each scenario. It means large cost savings can be realized with a slight loss of precision. Furthermore, the relative precision remains rather high even if sampling frequency is decreased to once every 10 days. However, the bias for various sampling frequencies (Table 3) increased markedly when sampling interval was above 5 days. We recommend the future pollen sampling program may choose the sampling frequency between "once in 3 days" and "once in 5 days". This allows one to perform the field sampling in a cost-efficient way.

In our study, we assume that the relationships found in the pollen concentration data from the pilot study would reflect the characteristics of the data in future sampling. This assumption may however, need to be verified since there may be other factors changing over time that can affect the periodicity and autocorrelation in the data. If autocorrelation decreases to some lower values, reducing sampling frequency too much may lead to an imprecise estimate of the overall mean.

The application of this approach is most suitable to estimating the long-term mean for some ambient exposures of interest. They may reflect exposures in a particular season or may reflect exposures on an annual basis. The present approach can be used to make sampling plan to detect trends in overall mean levels of exposure to aeroallergens (or any particular allergen) over time. The approach described here is not restricted to exposures to airborne pollen only. For instance, Somerville and Evans[9] successfully used time series analysis in assessing the effect of sampling frequency on detecting trends in the mass concentration of atmospheric fine particles. In situations, however, when one is directly studying the health effects of susceptible individuals due to short-term sporadic high exposures to certain allergens, further research and

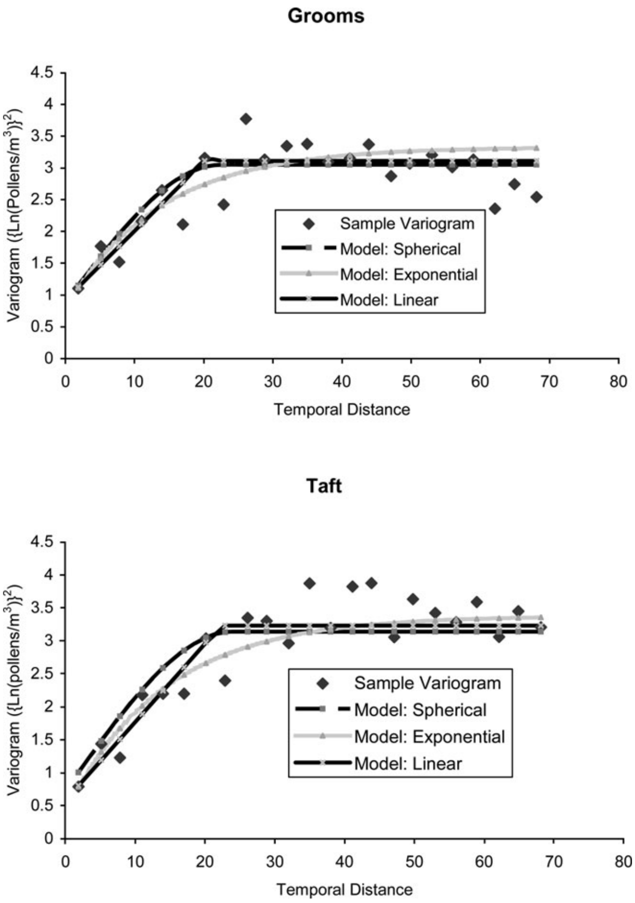

Fig. 4 Sample residual variograms associated with three models by site.

modifications would be needed to optimize sampling strategies for estimating short-term means.

The methodology presented here is primarily driven by the results of the pilot study. It is implicit in the approach that the results of the pilot study are indeed applicable to the subsequent full-scale study. For any future monitoring studies of a similar nature, where the goal is to monitor long term averages and not the daily fluctuations in exposures, we recommend that analyses presented in this paper can and should be done on a pilot study before the full-scale implementation of the chosen sampling plan.

Acknowledgements

This study was funded by the NIEHS Grant ES-0111170-01 as a part of the Cincinnati Childhood Allergy & Air Pollution Study [CCAAPS].

References

1 D. Myszkowska, D. Stepalska, K. Obtulowics and G. Porebski, *Aerobiologia*, 2002, **18**, 153–161.
2 E. Levetin, *Aerobiologia*, 1998, **14**, 21–28.
3 R. Pasken and J. A. Pietrowicz, *Atmos. Environ.*, 2005, **39**, 7689–7701.
4 P. C. Stark, L. M. Ryan, J. L. McDonald and H. A. Burge, *Aerobiologia*, 1997, **13**, 177–184.
5 C. Peretz, P. Goldberg, E. Kahan, S. Grady and A. Goren, *Ann. Occup. Hyg.*, 1997, **41**, 485–500.
6 G. K. Lemasters, R. Shukla, Y. Li and J. Lockey, *J. Occup. Environ. Med.*, 1996, **38**, 39–45.
7 D. Martuzevicius, J. Luo, T. Reponen, R. Shukla, A. L. Kelley, H. Clair and S. A. Grinshpun, *J. Environ. Monit.*, 2004, **7**, 67–77.
8 R. Shukla, J. Luo, G. K. LeMasters, S. A. Grinshpun and D. Martuzevicius, *J. Environ. Monit.*, 2005, **7**, 603–607.
9 M. C. Somerville and E. G. Evans, *Atmos. Environ.*, 1995, **18**, 2429–2438.
10 K. Cameron and P. Hunter, *Environmetrics*, 2002, **13**, 629–656.
11 T. Lwin, *Int. Mineral Process.*, 2003, **69**, 49–74.
12 A. Adhikari, D. Martuzevicius, T. Reponen, S. A. Grinshpun, S.-H. Cho, S. K. Sivasubramani, W. Zhong, L. Levin, A. L. Kelly, H. G. St. Clair and G. K. LeMasters, *Atmos. Environ.*, 2003, **37**, 4723–4733.
13 A. Adhikari, T. Reponen, S. A. Grinshpun, D. Martuzevicius and G. K. LeMasters, *Environ. Pollut.*, 2006, **140**, 16–28.
14 R. H. Shumway, D. S. Stoffer, *Time Series Analysis and Its Applications*, Springer-Verlag, New York, 2000, ch. 2, pp. 89–212.
15 Z. Sen, *Int. J. Climatol.*, 1998, **18**, 1725–1732.
16 A. Journel and C. Huijbregts, *Mining Geostatistics*, Academic Press Limited, London, 1978, ch. 2, pp. 26–147.
17 P. Diggle, K. Liang and S. Zegger, *Analysis of Longitudinal Data*, Oxford University Press, New York, 1994, ch. 3, 33–54.
18 R. Webster and M. A. Oliver, *Geostatistics for Environmental Scientists*, 2001, vol. 12, ch. 4–6, pp. 47–134.
19 E. J. Pebesma, *Comput. Geosci.*, 2004, **30**, 683–691.
20 D. J. Gorsich and M. G. Genton, *Math. Geol.*, 2000, **32**, 249–270.
21 J. Belmonte and M. Canela, *Aerobiologia*, 2002, **18**, 287–295.