Atmospheric Environment Vol. 26A, No. 1, pp. 113–123, 1992.
Printed in Great Britain.

0004–6981/92 $3.00 + 0.00
© 1991 Pergamon Press plc

OPTIMAL ESTIMATORS FOR AMBIENT AIR QUALITY LEVELS

STUART A. BATTERMAN
The School of Public Health, University of Michigan, Ann Arbor, MI 48109-2029, U.S.A.

(*First received* 1 *September* 1990 *and in final form* 14 *March* 1991)

Abstract—Procedures to estimate missing data, determine extrema, and derive uncertainties for data collected in ambient air monitoring networks are presented. The optimal linear estimators used obtain unbiased, minimum variance results based on the temporal and spatial correlation of the data and estimates of sample uncertainty. The first estimator interpolates missing data. The second estimator derives extrema, e.g. minimum and maximum concentrations, from the completed data set. Together the estimators can be used to check the validity of monitored observations, identify outliers, and estimate regional and local components of pollutant levels. The estimators are evaluated using data collected in urban air quality monitoring networks in Houston, Philadelphia and St Louis.

Key word index: Distribution, data imputation, optimal estimation, statistical models, uncertainty.

1. INTRODUCTION

Ambient air quality monitoring networks operating throughout the world over the last few decades have collected a vast amount of data. These data potentially are useful for many types of studies. However, several issues should be addressed before using historical data. These include the following. (1) How representative are the measurements? (2) How can sampling errors be estimated? (3) Can missing or invalid data be identified and estimated? (4) What are the extrema in pollutant levels? (5) Can 'local' and 'distant' (or 'background') components be separated? These issues may be critical in interpreting air quality data. Despite their importance, few methods which address them exist, and none are in common use.

This paper develops procedures to derive more meaningful information from ambient network data. The procedures use optimal estimation techniques which employ the spatial and temporal correlation of ambient measurements and related covariates, and estimates of sampling uncertainty. The procedures, which are quite general, can provide a practical way to enhance the usefulness of historical data.

The paper is organized as follows. Section 2 reviews aspects of ambient air quality sampling and statistical procedures used to analyse the collected data. Section 3 presents a conceptual framework for components of ambient pollutant levels and then gives the mathematical development of the estimation procedures. Section 4 applies and evaluates the procedures using three urban scale case studies. Section 5 discusses results and concludes the paper by suggesting further applications of the procedures.

2. BACKGROUND

Ambient air quality monitoring networks are established for purposes which include (1) the assessment of concentration levels and the compliance status with air quality standards; (2) the determination of health and environmental impacts; and (3) the selection and monitoring of emission abatement strategies. Many air quality monitoring networks consist of 5–20 sites obtaining hourly measurements of criteria pollutants CO, O_3, NO_x and SO_2, and 24-h measures of total suspended particulates (TSP). Networks which have been operating for two decades may have collected 10^7 observations. Recent concerns and new air quality standards have increased the number of pollutants monitored. Particulate matter less than 10 μm dia. (PM-10), lead, hydrocarbons and other contaminants may also be routinely measured.

Analysis of air quality data

Reports generated from collected data include monthly and annual summaries listing concentrations at various percentiles and averaging times. Many more sophisticated statistical analyses have been performed, although few procedures are used routinely. Applications of advanced analyses generally have been limited to special studies, e.g. trend analysis, exposure studies, receptor modeling, and dispersion model validation.

Table 1 classifies statistically-oriented analyses in the literature by two factors: the number of monitoring sites, and the number of variables. A wide range of analyses have been employed, including both standard and innovative methods. The following summary gives a cross-section of the literature. Single variable (i.e. single pollutant)—single site studies have included classical time series "Box and Jenkins-type" models for short term forecasts (e.g. McCollister and Wilson, 1975), spectral analyses indicating periodicities of pollutant data (e.g. Hayas *et al.*, 1982), regression models estimating pollutant distributions (e.g. Larsen, 1976), Poisson random process models

Table 1. Statistical methods applicable to network data

	Single monitoring site	Multiple monitoring sites
Single pollutant	Trend analysis Analysis of distributions Probability of exceedance Extreme value statistics Time series (ARIMA) Spectral analysis Markov-type models	Upwind/downwind analysis Kriging Spatial interpolation (1) Optimal estimation
Multiple variables	Correlation analysis Factor analysis Generalized linear models (2) Receptor models Cluster analysis Time series (ARIMAX)	Kalman filter models Co-kriging (3) Optimal estimation (4)

Notes: (1) Includes contouring, e.g. linear (planer) and non-linear interpolation.
(2) Includes linear and non-linear regression.
(3) No studies identified using co-kriging.
(4) Could use procedure discussed with exogenous variables.

(e.g. Baker *et al.*, 1984), probability of exceedances and return period models (e.g. Drufuca and Giugliano, 1977), Markov-type models based on up- and down-crossings of a threshold concentration (e.g. North *et al.*, 1984), and extreme value statistics (e.g. Roberts, 1979; Shively, 1990). Recently, the single monitoring site–multiple variable category has received the greatest attention due to the application of chemical mass balance regression-type receptor models (e.g. Henry *et al.*, 1984). Receptor model studies also have used principal component and factor analysis methods (e.g. Lowenthal and Rahn, 1987). Single site–multiple variable studies have employed time series models with pollutant and meteorological variables (e.g. Finzi *et al.*, 1980) and other procedures such as cluster analysis (Gether and Seip, 1979). Most multiple site studies have been performed for two purposes. Upwind/downwind analyses have been used to estimate contributions from distant and local emission sources (e.g. Batterman *et al.*, 1987). Various contouring routines have been used to derive pollutant isopleths over a region to estimate exposures and other impacts, including the use of kriging (Lefohn *et al.*, 1987; Venkatram, 1988; Haas, 1990). A few potential multiple site–multiple variable techniques are identified in the table; most applications use a Kalman filtering approach to reconcile models and data (e.g. Mulholland, 1989).

Background estimates

Often it is important to apportion pollutant contributions attributable to local and distant emission sources. Long-range transport by distant sources can provide a significant 'regional' or 'background' contribution which restricts the control options available to local authorities. Such situations can occur with PM-10, sulfate, ozone and other pollutants.

Approaches for separating local and regional components use either dispersion modeling or ambient monitoring. Both approaches require that monitoring and modeling errors are negligible or known. The key disadvantage of the dispersion modeling approach is the uncertainty of the predictions, which is about a factor of two for short-term averages (American Meteorological Society, 1981). Also, a suitable model, an accurate source inventory, and meteorological observations for a representative period are required. Thus, this approach is not recommended (EPA, 1984). The suggested approach uses ambient monitoring at upwind or isolated 'regional' sites. Upwind observations should exclude measurements affected by local sources. Regional sites should be located away from the area of interest and unaffected by local sources (EPA, 1984). More detailed guidelines for sites to monitor regional atmospheric deposition specify a minimum separation distance of 10 km from industrial and natural sources of emissions exceeding 10,000 t y^{-1} and population centers of 10,000 or more (ASTM, 1989). Separation distances should be increased 'dramatically' if the sampler lies in the prevailing downwind direction of emission sources.

Background estimates based on monitoring may have several deficiencies caused by insufficient temporal and spatial coverage in the network. Six examples are given. (1) Background contributions dominate some pollutants (e.g. PM-10 and SO_4), and there may not be enough monitoring sites to detect relatively small local impacts. (2) It may be difficult to designate particular sites as 'upwind' or 'regional' sites since some pollutants (e.g. PM-10, NO_x, HC) are emitted by many well-dispersed sources surrounding most monitors. (3) Wind shifts during sampling periods may invalidate upwind designations, especially for pollutants collected over long periods, e.g. 24-h particulate samples. (4) Pollutants sampled intermittently, e.g. TSP measured every sixth day, have temporal resolution too coarse to determine background. (5) The relative accuracy of measurements decreases with the low concentrations likely at regional sites (e.g. Evans and Ryan, 1983). (6) Missing data may bias results (Davison and Hemphill, 1987) especially when there are few monitoring sites. Any of these events may cause serious errors.

Accuracy and representativeness of data

A general goal of sampling is to obtain 'representative' measurements, defined by Geiger (1965) as having a wide range of validity. This goal is tempered by the need to obtain appropriate spatial and temporal resolution given time and cost constraints, and the need to accurately monitor concentrations at specified percentiles and averaging times. For example, air quality regulations focus on short-term peak concentrations such as the second highest concentration in a year. These peak concentrations or extrema can be difficult to measure accurately.

In most analyses, monitoring observations are assumed to be representative of ambient levels at the monitoring site for the averaging time of the measurement. With the exception of some receptor modeling techniques (e.g. Watson *et al*., 1984), observations also are assumed to be error-free. Errors, however, can arise from many sources including (1) analytic techniques; (2) sampler biases; (3) lack of sampling representativeness; and (4) miscellaneous sources, e.g. sample degradation, data entry mistakes, etc. In theory, errors can be partitioned into systematic and random elements, affecting accuracy and precision, respectively. Most concentrations are based on several components, e.g. sample volume and particle mass for particulate concentrations, each of which contributes systematic and random errors. As component errors may be additive or multiplicative, correlated or independent, or simply unknown, the total error is often uncertain. A good measure of the total random error is the sample variance of replicates (Draper and Smith, 1981), however, true repetitions in routine monitoring programs are rare. Assumptions of representativeness and accuracy may be particularly problematical for the extrema needed to determine regional and local contributions. Both the lowest and the highest concentrations may be prone to measurement anomalies.

Missing data

An additional concern is the completeness of the data. Missing (or invalid) data may result from instrument failure, calibration and maintenance problems. In the case studies described later, about one-quarter of the data was missing. A larger percentage was missing at specific monitoring sites, especially at rural sites which are difficult to service. Many networks achieve comparable records. Missing data increase the difficulty of establishing trends and determining compliance with ambient standards based on the number of exceedances (Davison and Hemphill, 1987). In multivariate applications such as receptor models, the omission of a single element may necessitate the rejection of the entire observation.

A number of methods to handle missing data have been developed in biostatistics where responses to surveys, for example, often contain large amounts of missing or incorrect data (e.g. Garfinkel, 1986; Little *et al*., 1989). Geographers have also confronted this problem (as reviewed by Bennett *et al*., 1984). Few applications of these or other methods have been used for air quality data. One approach for estimating or 'inputing' missing ozone data used ozone–temperature relationships (Davison and Hemphill, 1987). More general methods, as developed in the following section, would be helpful for other pollutants.

3. OPTIMAL ESTIMATORS

Statistical framework

A framework for ambient air concentrations is developed considering a single conservative (non-reactive) pollutant measured in an urban scale monitoring network. The concentration observed at site i and time t, $C_{i,t}$, consists of three components:

$$C_{i,t} = L_{i,t} + D_t + V_{i,t}, \tag{1}$$

where $L_{i,t}$ and D_t are local and regional components, respectively, and $V_{i,t}$ is measurement error. Local contributions result from emission sources situated within the urban area. These concentrations typically increase towards the source. The regional component, produced by long-range transport, has gradients that are negligible on the local scale. Thus, D_t is time varying, but constant in space at the urban scale.

The spatial and temporal correlation present in the data is used to improve the accuracy and robustness (insensitivity to outliers) of concentration estimates. A three-part procedure is used. First, an estimate of measurement uncertainty $V_{i,t}$ is derived. Next, an optimal estimation procedure estimates missing data. Lastly, the lowest and highest concentrations are estimated from the estimated data set.

Measurement uncertainty

Several approaches can be used to estimate error $V_{i,t}$. Random errors may be estimated using replicate observations, e.g. colocated samplers, while systematic errors can be estimated using reference or calibrated samplers. Alternatively, errors may be estimated by isolating uncertainties in the component measures and then propagating their effects, e.g. using Gaussian quadrature. Lastly, empirical means may be used. The following examples demonstrate these approaches.

Since 1981, federal regulations have required state and local agencies to assess the accuracy and precision of their ambient air quality measurement systems. Data collected in the Precision and Accuracy Reporting System (PARS) are based on blind audits using calibration gases for continuous instruments (gases), and colocated samples for manual instruments (TSP, Pb, and older gas measuring instruments). PARS results, expressed as a 95% confidence interval, typically show a relative accuracy of about 10% for most of the criteria pollutants. The precision of the measurements, obtained by repeated measures, is about 10% for O_3, 12% for TSP, 20 for SO_2, and 46% for NO_2 (Rhodes and Evans, 1988). These statistics represent many thousands of audits.

One theoretical study of errors in mass, flow rate and timing measurements suggests errors about half of that obtained in field evaluations (Evans and Ryan, 1983). Other examples of component errors estimate filter mass measurement errors (using beta gauge attenuation) of 3 $\mu g\, m^{-3}$ for 12-h samples (Jaklevic *et al.*, 1981), and biases between gravimetric and beta gauge measurements of <5% (Courtney *et al.*, 1982). With air volume errors of 5–10%, these figures yield a total error of 10–20% at typical particle concentrations.

An empirical estimate of sampling errors is the difference between the lowest two concentrations in the network, assuming that these concentrations result mainly from regional sources. While imperfect, this estimate may be useful in large monitoring networks where the two lowest concentrations can be considered replicates. In the case studies (described later), this procedure gave relative errors of 15–20%. A better, but rarely available measure, is the variance between measurements obtained from colocated samplers.

The three approaches yield relative errors in the range of 10–46%. In most cases, error statistics are not accurately known. Also, measurements obtained under unusual conditions may yield much larger errors. For example, erroneous particulate measurements can be caused by high loadings which clog filters, unusual size distributions, and high wind speeds which affect inlet performance.

Estimating missing data

This section develops an optimal linear estimator to estimate missing observations. Missing observations are considered unknown random variables. The statistics of these variables are based on available data, and are selected to preserve the observed spatial and temporal correlation.

Column vector $\mathbf{Z}_t$ is arranged to contain leading, lagging and simultaneous observations at all monitoring sites in the network:

$$\mathbf{Z}_t = [C_{1,t-m} \cdot\cdot C_{n,t-m} | \cdots | C_{1,t} \cdot\cdot C_{n,t} | \cdots | C_{1,t+m} \cdot\cdot C_{n,t+m}]', \qquad (2)$$

where $C_{i,t}$ is the concentration at site i and time t, n is the number of monitoring sites, m is the number of leading and lagging time periods, and the quote denotes transpose. The leading and lagging elements permit interpolations in time, while the simultaneous observations allow spatial averaging. As shown later, one lag and lead period is generally sufficient, so $m=1$ and $\mathbf{Z}_t$ includes $3n$ elements (lagging, simultaneous, and leading concentrations at n sites). As described earlier in Equation (1), observation $\mathbf{Z}_t$ includes the true pollutant level $\mathbf{X}_t$ plus error $\mathbf{V}_t$:

$$\mathbf{Z}_t = \mathbf{X}_t + \mathbf{V}_t. \qquad (3)$$

If some data are missing, the corresponding elements in vectors $\mathbf{Z}_t$ and $\mathbf{V}_t$ have missing values, but these can be estimated as the corresponding elements of vector $\hat{\mathbf{X}}_t$ and matrix $\mathbf{S}_t$, as described below.

Error covariance matrix $\mathbf{R}_t$ is defined as:

$$\mathbf{R}_t = E[\mathbf{V}_t \mathbf{V}_t']. \qquad (4)$$

Error $\mathbf{V}_t$ and covariance $\mathbf{R}_t$ must be estimated. If errors are uncorrelated, $\mathbf{R}_t$ is a diagonal matrix. Diagonal elements of $\mathbf{R}_t$ are set to the measurement variance. As discussed in section 2, measurement errors can be estimated in several ways. Here, errors are assumed to be time invariant, using a relative error of 30% and the mean concentration. The diagonal elements corresponding to missing observations are set to a much larger value, e.g. 1000 times the measurement variance, to represent the large (prior) variance of the missing data.

First and second moment statistics, namely, mean vector $\mathbf{M}$ and covariance matrix $\mathbf{P}$, are sample estimates from available data:

$$\mathbf{M} = T^{-1} \Sigma_t \mathbf{X}_t \qquad (5)$$

$$\mathbf{P} = T^{-1} \Sigma_t [(\mathbf{X}_t - \mathbf{M})(\mathbf{X}_t - \mathbf{M})'], \qquad (6)$$

where T is the number of observations used to estimate $\mathbf{M}$ and $\mathbf{P}$. Matrix $\mathbf{P}$ contains information regarding the spatial and temporal correlation of the data. Assuming unbiasedness ($E[\mathbf{V}_t]=0$) and uncorrelated errors ($E[\mathbf{X}_t \mathbf{V}_t']=0$), the best linear, unbiased and minimum variance estimate $\mathbf{X}$ of the missing observations is:

$$\hat{\mathbf{X}}_t = \mathbf{M} + \mathbf{P}(\mathbf{P} + \mathbf{R}_t)^{-1}(\mathbf{Z}_t - \mathbf{M}). \qquad (7)$$

This Bayesian estimator weights the information provided by the observations (the so-called influence vector $\mathbf{Z}_t - \mathbf{M}$) to yield the estimate $\mathbf{X}_t$. Results will be identical to mean $\mathbf{M}$ if there is zero correlation between observations, i.e. $\mathbf{P}=0$. The (posterior) error of estimation matrix S is:

$$\mathbf{S}_t = \mathbf{E}[(\mathbf{X}_t - \hat{\mathbf{X}}_t)(\mathbf{X}_t - \hat{\mathbf{X}}_t)'] = \mathbf{P} - \mathbf{P}(\mathbf{P} + \mathbf{R}_t)^{-1}\mathbf{P}. \quad (8)$$

The estimator in Equation (7) minimizes the diagonal terms $\mathbf{S}_t$ for the stated assumptions, as shown by Schweppe (1973). In Equation (7), missing data in $\mathbf{Z}_t$ are set to zero, however, solutions (for missing data) are insensitive to the value specified since the variance terms in $\mathbf{R}_t$ for corresponding elements are so large.

As ambient data are generally highly correlated, the estimate of the ith missing observation $\hat{X}_{i,t}$ often is very different from the mean, with a variance $S_{i,i,t}$ which is greatly reduced from the assumed prior variance. Conversely, if the measurement variance is zero for the ith (known) observation, then estimate $\hat{X}_{i,t}$ is unchanged from observation $Z_{i,t}$ and $S_{i,i,t}=0$.

Estimate $\hat{\mathbf{X}}_t$ is the mean of the conditional distribution given $\mathbf{Z}_t$ (Schweppe, 1973), that is, it is the expected value or average of the missing data based on many instances in which the same pollutant conditions prevailed. The actual pollutant levels for any single instance (if available) would differ from this mean and show a broader distribution than obtained from the estimator. Conditional simulation is used to obtain the original distribution by adding a random term to the mean, e.g. post-whitening:

$$\hat{X}'_i = \mathrm{E}[\hat{X}_{i,t} | \mathbf{Z}_t] + \alpha_{i,t}, \tag{9}$$

where $\alpha_{i,t}$ is a zero mean random variable with variance $S_{i,i,t}$ obtained in Equation (8). Because pollutant observations are roughly log-normally distributed with primarily multiplicative errors, the logarithms of observations are used in Equations (1)–(9). The final estimate employs conditional simulation and exponentiation:

$$\hat{X}''_{i,t} = \exp(\hat{X}'_{i,t} + S^{1/2}_{i,i,t} w), \tag{10}$$

where w is a normally distributed, unit variance random variable.

Estimating extrema

Extrema are found using an optimal estimation procedure similar to Equations (2)–(10). The procedure is developed for the highest three concentrations. Let $H_{1,t}$, $H_{2,t}$ and $H_{3,t}$, respectively, represent the highest, second and third highest concentrations in the network at time t. Redefining the symbols used earlier, $\mathbf{Z}_t$ contains the top three concentrations in the network for the current, and m leading and lagging periods:

$$\mathbf{Z}_t = [H_{1,t-m} H_{2,t-m} H_{3,t-m} | \ldots | H_{1,t} H_{2,t} H_{3,t} | \ldots | H_{1,t+m} H_{2,t+m} H_{3,t+m}]'. \tag{11}$$

The observation $\mathbf{Z}_t$ of these concentrations differs from the true concentration $\mathbf{X}_t$ by sampling error $\mathbf{V}_t$ (Equation (3)). Covariance matrix $\mathbf{P}$ in Equation (6) does not include sampling errors. Let $\mathbf{R}$ represent the error covariance matrix of sampling errors. If $\mathbf{R}$ is known and uncorrelated with the measurements, then $\mathbf{P}$ can be estimated as:

$$\mathbf{P} \simeq T^{-1} \Sigma_t (\mathbf{Z}_t - \mathbf{M})(\mathbf{Z}_t - \mathbf{M})' - \mathbf{R}, \tag{12}$$

where $\mathbf{M}$ is the sample mean. Observations (not estimated values) are used to compute $\mathbf{M}$ and $\mathbf{P}$. Matrix $\mathbf{R}$ is estimated as:

$$\mathbf{R} = k\mathbf{I}, \tag{13}$$

where $\mathbf{I}$ is an identity matrix and k is the sampling error which is assumed constant at all sites. Parameter k may be selected based on the expected relative error for extrema in the network, e.g. if the relative error was 30%, k is the square of the product of 0.3 and the mean concentration. Restrictions must be placed on $\mathbf{R}$ to ensure that $\mathbf{P}$ is positive semi-definite, and tests for positive definiteness may be made. Again, Equation (7) provides an unbiased minimum variance estimate of $\mathbf{X}$.

The lowest concentrations in the network are an estimate of the regional component D_t as described earlier in Equation (1). These concentrations are determined in the same manner as the peak concentrations except that the lowest three concentrations replace the three peak values. High and low extrema can be simultaneously estimated by including the highest and the lowest concentrations in the observation vector. Local impacts $L_{i,t}$ can then be estimated as the difference between estimated peaks X_t and the estimated background level D_t.

4. APPLICATION

Implementation and evaluation

The estimators described in section 3 were coded in FORTRAN and run on an 80386-based computer. A LU inversion algorithm (Press *et al.*, 1987) was used. Both single and double precision programs were written. A jack-knife procedure was used to evaluate the estimators' performance. A portion of the data in the case studies, selected randomly, was intentionally deleted. Deleted data were then predicted using the first estimator. Estimates were made if half or more of the elements in vector $\mathbf{Z}_t$ were available. This criterion provides a compromise between the reliability of the estimator and demands placed on the procedure. Predictions were compared to the actual observations using linear correlation coefficients, mean bias, and scatterplots. Because the actual errors were unknown, the estimators were tested with errors ranging from 0 to 60%.

Case studies

Three case studies were used to evaluate the estimators. The first employed particulate data collected in St Louis, IL from May to September 1976 as part of the Regional Air Pollution Study (Strothmann and Schiermeier, 1979). In this study, dichotomous samplers at 10 sites collected 12-h samples in fine and coarse size fractions. Because of long gaps of missing data, fewer sites are used here (Table 2). Most sites were urban; coverage extended to about 45 km from the city center. The second study is the Philadelphia Area Field Study (Toothman, 1984) in which ambient data were collected from 14 July to 13 August 1982 at six sites. As in St Louis, dichotomous samplers collected 12-h particulate samples, also in two size fractions. This urban area was considerably larger than St Louis, yet the monitoring network was smaller. Most sites were urban and industrial. The study included a 'special studies' site with impacts from a nearby oil tank farm, local truck traffic and ongoing construction, and a rural site in New Jersey. The third case study used O_3 observations taken at 11 sites in Houston, TX from April to September 1987. In this study, monitoring sites ranged over a distance of

Table 2. Number of available and deleted data in case studies. Range is shown in parentheses

	Number monitoring sites	Number of observations per site Maximum possible	 Average available	 Average deletions
St Louis				
Fine particles	9	92	62 (58–79)	23 (16–32)
Coarse particles	7	92	62 (55–82)	22 (17–27)
Philadelphia				
Fine particles	6	62	54 (38–60)	7 (1–13)
Coarse particles	6	62	49 (30–95)	5 (4–6)
Houston				
Ozone average	11	152	120 (97–147)	7 (6–8)
Ozone peak	11	152	120 (97–147)	17 (9–22)
Average percentage of total		100%	78%	14%

60 km; most were located within urban Harris County. Daily averages and daily hourly peaks were calculated at each site from hourly observations. If fewer than 12-h were available at a site in a given day, the observation was considered to be missing.

The data capture rates of the three networks are summarized in Table 2. For the monitoring networks and time periods selected, network data capture averaged 78%. For single sites, 48–97% of the data were available.

Estimates of missing data

Table 2 also shows the number of observations deleted from each data set in order to evaluate the estimator. An average of 14% of the available observations was deleted. This posed a severe test of performance since an average of 22% of the data was already missing, thus an average of only 64% of the data was utilized to predict missing observations.

Linear correlation coefficients and biases between estimated and actual data are shown in Table 3. These calculations were obtained for each data set using 0, 1 and 2 lead/lags. A relative error of 30% was used in each case. Correlations were high, between 0.66 and 0.90, and biases were small. Despite the improvement expected, the number of leads/lags did not have dramatic effects. In fact, performance sometimes degraded as the number of lead/lag periods increased as illustrated by results for St Louis. This resulted as round-off errors increased with additional lags and negated the small (and diminishing) information provided with longer lead and lag periods. For example, with 2 leads/lags and $n = 10$ (10 monitoring sites), an ill-conditioned matrix of rank 50 must be inverted. While the double precision version of the program reduced these errors, more accurate inversion routines might be advantageous.

Means and standard deviations of the observations and estimates matched closely. Over 60–80% of the predictions were within 25% of the observation, and 80–90% were within 50%. Scatter plots show very good agreement with the Texas data set (Figs 1e and 1f). However, the variability of the particulate data is not fully reproduced, e.g. peak values are underestimated and low values are overestimated (especially Figs 1b and 1c). This analysis assumes that monitoring observations are error-free. Better agreement could be obtained by (1) changing the relative error; (2) changing the number of lead/lags; (3) altering the log-normal assumption; (4) decreasing the fraction of missing data; and (5) increasing post-whitening. For example, excellent agreement could be obtained ($r \geqslant 0.95$) for the Philadelphia fine fraction particulate data using 1 lead/lag, a relative error of 0.10, and assuming normal, rather than log-normal distributions. In practice, such parameters could be calibrated

Table 3. Correlation coefficient (r) and bias (b) of the estimator

	Philadelphia Fine	 Coarse	St Louis Fine	 Coarse	Houston Peak	 Average
Lags = 0						
r	0.84	0.84	0.73	0.66	0.90	0.90
b	0.88	2.11	4.10	4.74	0.36	0.05
Lags = 1						
r	0.84	0.82	0.68	0.67	0.86	0.89
b	0.07	1.00	3.25	5.65	0.39	0.04
Lags = 2						
r	0.81	0.82	0.33	0.76	0.88	0.82
b	0.61	1.23	3.50	3.68	0.06	0.24

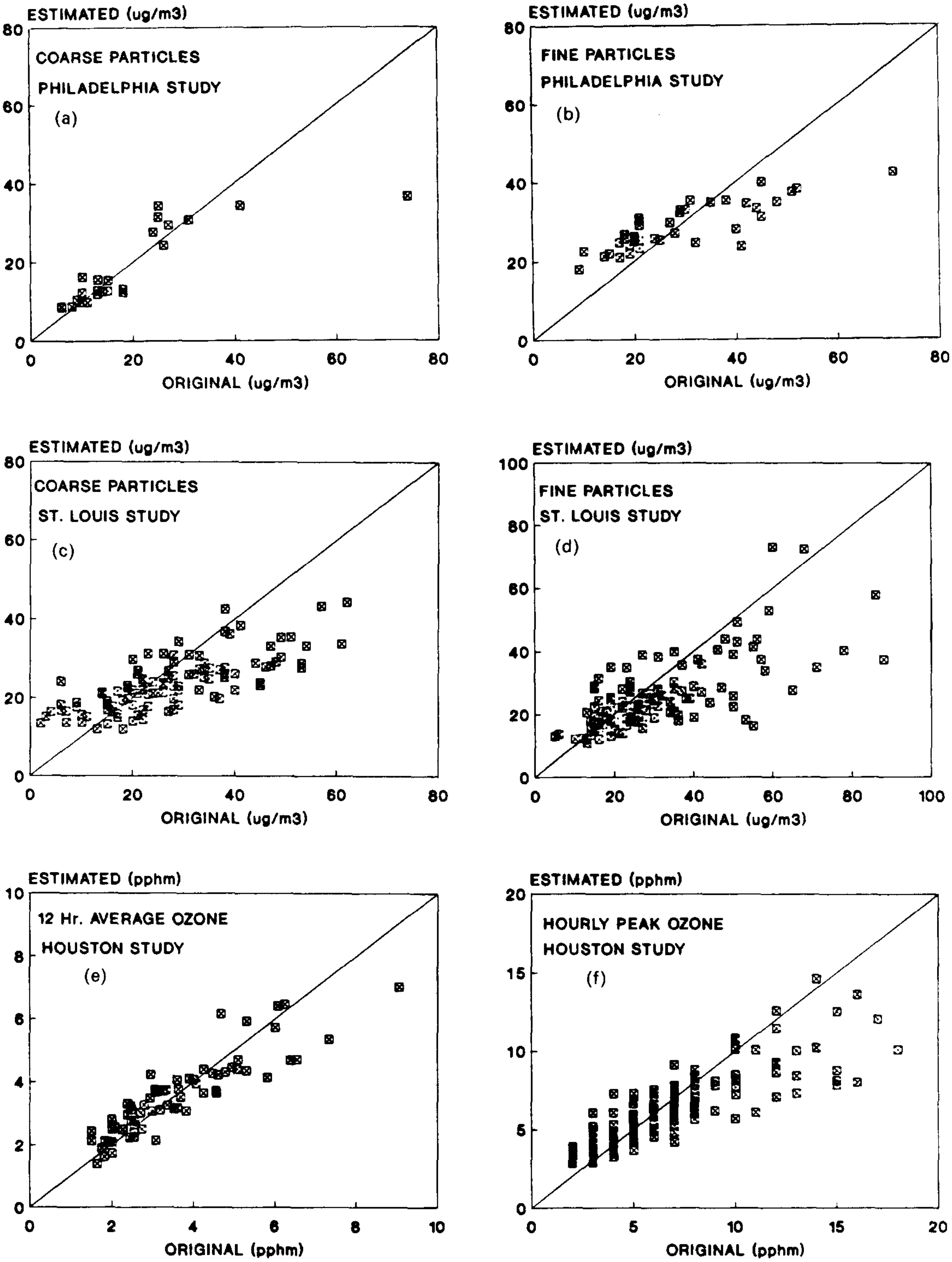

Fig. 1. Scatterplots of predicted vs estimated data for the three case studies.

using site-specific results. Even without such calibrations, the predictions preserve the actual spatial and temporal trends using solely the data's correlation structure.

The largest relative errors result from overprediction of very low observations, e.g. particulate concentrations averaging 6 and 9 $\mu g\,m^{-3}$ in coarse and fine fractions, respectively. These observations were several times smaller than those at other sites, and also smaller than preceding and following concentrations. These measurements are anomalies and possibly erroneous. The largest errors at high concentrations occur with the coarse fraction particulate data sets, probably a result of local influences which were uncorrelated with observations at other sites. Such cases cannot be predicted without highly detailed information. In general, the technique of deleting and then predicting each observation can provide a good check on data validity.

Extrema estimates

Extrema estimates were calculated for each data set, using various lead/lags and relative errors. Figs 2 and 3 show typical results for highly correlated (fine fraction) and poorly correlated (coarse fraction) data using the Philadelphia data for the period of 16–31 July 1982. All available data were used, and no deliberate deletions were made (as in the previous application). One lead/lag period is used. Pollutant observations and estimates are shown as points, while extrema estimates are drawn as lines. Three extrema estimates are shown. (The second and third highest and lowest extrema estimates have been omitted from the graph for clarity.) Extrema with 0% relative error provide an exact match between actual or expected extrema, e.g. the estimate is unchanged from the observation, and the two lines simply connect sequential maxima or minima. As the relative error increases to 20 and then 60%, the variation in pollutant levels is dampened, neither completely rising to the peak concentrations nor falling to the lowest values. In general, the upper and lower envelopes converge towards the

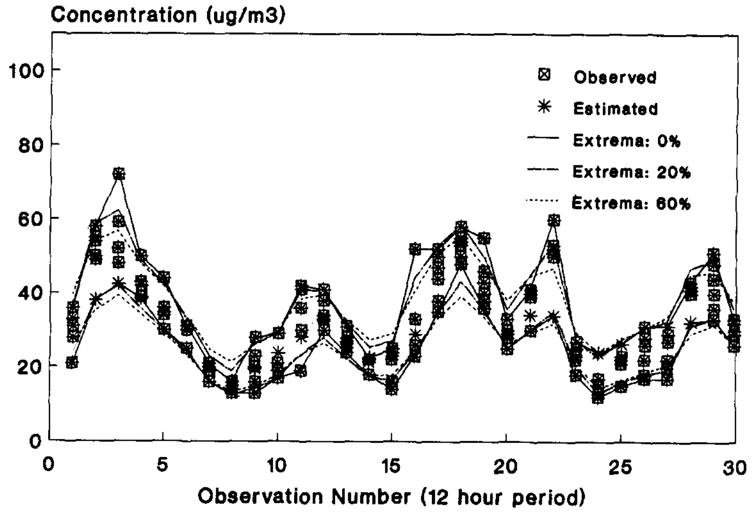

Fig. 2. Observations, estimates and extrema envelopes for fine fraction particulate data in Philadelphia.

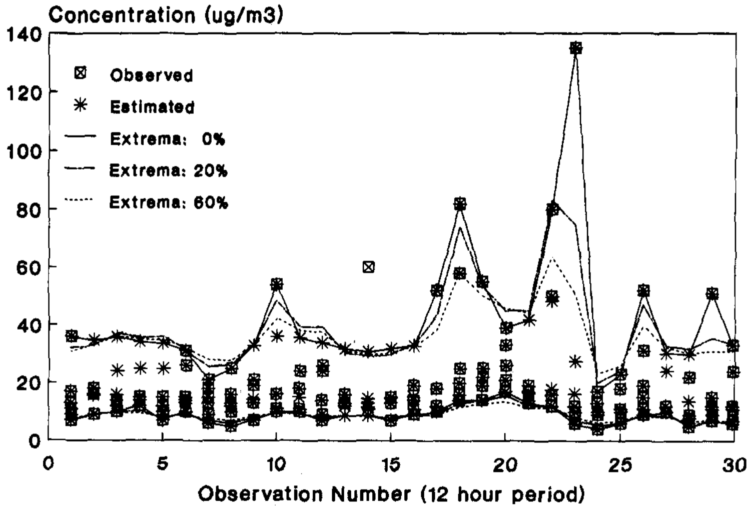

Fig. 3. Observations, estimates and extrema envelopes for coarse fraction particulate data in Philadelphia.

mean as the relative error increases, thus maximum concentrations are underpredicted and minimum concentrations are overpredicted. Local impacts $L_{i,t}$ are represented as the width of the envelope. Estimates of the lowest concentration appear smoothed or dampened in time, possibly reflecting the regional component which usually changes slowly with respect to the sampling frequency.

Peak coarse fraction concentrations (Fig. 3) occur at the Fireboat site, due to local impacts which are largely uncorrelated to concentrations at other sites. This is clearly shown by the extrema estimated with $\geqslant 20\%$ error which do not approach these peaks. Conversely, the fine fraction data is highly correlated and the envelope maintains a nearly constant width. In both cases, upper extrema are sensitive to the relative error used, while lower extrema are not. In part, this results from the use of a logarithmic transformation.

As discussed, the lowest and highest concentrations are prone to measurement anomalies, unusual meteorological factors and source conditions. Large differences between observed and estimated extrema may indicate such events; definitely, these cases need further investigation. This simple strategy can isolate atypical extrema which may affect estimates of local and regional impacts.

Computational demands

The computational demands of estimation depend on m, the number of leads/lags, and n, the number of sites. Covariance matrix $\mathbf{P}$ and mean vector $\mathbf{M}$ are determined once for each data set. To interpolate missing data for each sampling period requires inversion of a rank $n(2m+1)$ matrix and several matrix multiplications. About 2 s of computer time were required for each sampling period using a fast (33 MHz) 80386/7 microcomputer and a high accuracy LU decomposition inversion algorithm. Extrema estimates require a single inversion for the entire data set. Only simple matrix operations are needed for each observation, thus these estimates are quickly calculated.

5. DISCUSSION AND CONCLUSION

This paper has developed linear estimators to estimate data and extrema with the purposes of handling missing data and accounting for errors. Extrema are estimated from the full (estimated) data set. Estimators of the lowest concentration in the network may represent regional levels if the monitoring system includes sites which are largely unaffected by local sources. Peak estimates, provided by the same estimator, indicate the contribution of local sources. Both estimates should be more robust than observations from single stations since spatial and temporal information from all sites is utilized.

In application to three diverse data sets in St Louis, Philadelphia and Houston, the estimators provided reliable results. The high spatial and temporal correlation present in ambient pollutant levels at the urban scale makes such estimators practicable. The estimators may be less appropriate for observations with little correlation, e.g. some coarse fraction particulates, networks with intermittent sampling, or networks covering very large spatial scales. The Bayesian estimator in Equation (7), which also can be viewed as a linear contraction operator, tended to reduce the scatter in the original data. In general, this is an undesirable property. However, the original dispersion of the data can be restored by changing the degree of post-whitening, altering the relative error, or by using a different data transformation. Such network-specific calibrations may further increase the accuracy of the estimators.

The estimators view historical observations as imperfect (error-containing) random variables, a fundamentally different perspective than the usual assumptions that the observations are representative and error-free. Spatial and temporal information has been used in the opposite manner to select sites in the optimal design of air monitoring networks (e.g. Shindo *et al.*, 1990). Results obtained in the case studies imply that monitoring observations, to varying degrees, are redundant in providing site-specific information since observations at some sites can be used to predict concentrations at other sites.

The behavior of the estimator depends on the relative strengths of the temporal and spatial correlation. If temporal correlation is dominant, a missing observation both preceded and followed by valid observations at the same site is estimated using primarily a weighted sum of leading and lagging observations at that site. If leading and lagging observations are missing, the estimate is derived from simultaneous observations at other sites. If spatial correlation is dominant, results depend on simultaneous measurements taken at other sites and to a lesser extent on leading and lagging observations. If many simultaneous measurements are missing, leading and lagging observations and the constant (mean) are emphasized. In each case, weights given to leading and lagging observations can be significant, and the coefficients are site-specific and depend on the data available. In comparison with the estimators for missing values, extrema estimates primarily depend on simultaneous observations. Thus, these estimators might be simplified to use observations at only the current time. The estimator automatically determines the weightings so as to minimize the variance of the estimate. The estimation procedure is flexible and applicable to other types of data.

Several refinements to the estimation procedures are possible. Although not attempted here, additional variables could be used to augment the pollutant variables and improve performance. For example, ambient temperature could be used to help predict O_3 concentrations. More accurate estimators might disaggregate by season, wind direction, or other features

—if the data are sufficient to estimate covariance matrix **P**. In the case studies, however, estimates based on seasonal data, first, second and third order lags were similar. In the case studies, estimates were obtained when up to 50% of the data were missing. A more restrictive parameter would improve performance. A stepwise procedure might be used to determine how many sites and how many lead/lag periods are necessary for estimation. While additional data would be expected to improve results, it also introduces greater numerical errors in matrix inversion. An automated procedure could be used to test various or all possible subsets of sites and variables, and have the added benefit of showing the sensitivity of results to these factors.

The estimators have several applications. They can be used to estimate missing data, thus providing a more complete data set. Although the estimators performed well, the use of estimated data must be carefully considered in interpreting results, especially if few data are available. Second, they may be used to check the validity of observations. Suspiciously low or high observations may be easily identified. An automated bootstrapping procedure is suggested to accomplish this task. Third, the estimators may provide more robust estimates of extrema from which local and regional contributions may be determined. Such estimates can be used in trend analysis and to determine compliance with air quality standards.

Acknowledgements—The author sincerely thanks Dr Hap Hemphill for providing the Houston data set, and Mr Adarsh Kulshretha for his assistance in performing the computational work. The two anonymous reviewers provided a number of helpful suggestions.

REFERENCES

American Meteorological Society (1981) Air quality modeling and the clean air act: recommendations to EPA on dispersion modeling for regulatory applications. Boston, MA.

ASTM (1988) Proposed standard guide for choosing locations and sampling methods to monitor atmospheric deposition. American Society of Testing and Materials, Boston, MA.

Baker M. B., Eylander M. and Harrison H. (1984) The statistics of chemical trace concentrations in the steady state. *Atmospheric Environment* **18,** 969–975.

Bennett R. J., Haining R. P. and Griffith D. A. (1984) The problem of missing data on spatial surfaces. *Ann. Ass. Am. Geograph.* **74,** 138–156.

Batterman S., Fay J. and Golomb D. (1987) Significance of regional source contributions to urban PM-10 concentrations. *J. Air Pollut. Control Ass.* **37,** 1286–1292.

Courtney W. J., Shaw R. W. and Dzubay T. G. (1982) Precision and accuracy of a beta gauge for aerosol mass determinations. *Envir. Sci. Technol.* **16,** 236–239.

Davison A. C. and Hemphill M. W. (1987) On the statistical analysis of ambient ozone data when measurements are missing. *Atmospheric Environment* **21,** 629–639.

Draper N. R. and Smith H. (1981) *Applied Regression Analysis.* John Wiley, New York.

Drufuca G. and Giugliano M. (1977) The duration of high SO_2 concentrations in an urban atmosphere. *Atmospheric Environment* **11,** 729–735.

EPA (1984) Proposed rule for 40 CFR, Part 50, Federal Register, pp. 10,407–10,462, 20 March.

Evans J. S. and Ryan P. B. (1983) Statistical uncertainty in aerosol mass concentrations measured by virtual mass impactors. *Aerosol Sci. Technol.* **2,** 531–536.

Finzi G., Fronza G. and Spirito A. (1980) Multivariate stochastic models of sulphur dioxide pollution in an urban area. *J. Air Pollut. Control Ass.* **30,** 1212–1215.

Garfinkel R., Kunnathur S. and Liepins A. S. (1986) Optimal imputation of erroneous data: categorical data, general edits. *J. Oper. Res.* **34,** 744–751.

Geiger R. (1965) *The Climate near the Ground.* Harvard University Press, Cambridge, MA.

Gether J. and Seip H. M. (1979) Analysis of air pollution data by the combined use of interactive graphic presentation and a clustering technique. *Atmospheric Environment* **13,** 87–96.

Haas T. C. (1990) Kriging and automated variogram modeling within a moving widow. *Atmospheric Environment* **24A,** 1759–1769.

Hayas A., Gonzalez C. F., Pardo G. and Martinez M. C. (1982) Application of spectral analysis to atmospheric dust pollution. *Atmospheric Environment* **16,** 1919–1922.

Henry R. C., Lewis C. W. and Hopke P. K. (1984) Review of receptor model fundamentals. *Atmospheric Environment* **18,** 1507–1515.

Jaklevic J. M., Gatti R. C., Goulding F. S., Loo B. W. and Thompson A. C. (1981) Aerosol Analysis for the Regional Air Pollution Study. PB 81-157 141, Environmental Sciences Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC.

Larsen R. I. (1976) A mathematical model for relating air quality measurements to air quality standards. Report AP-89, U.S. Environmental Protection Agency, Research Triangle Park, NC.

Lefohn A. S. *et al.* (1987) An evaluation of the kriging method to predict 7-h seasonal mean ozone concentrations for estimating crop losses. *J. Air Pollut. Control Ass.* **37,** 595–602.

Little R. J., Rubin D. B. and Strawderman W. E. (1989) *Statistical Analysis with Missing Data.* John Wiley, New York.

Lowenthal D. H. and Rahn K. A. (1987) Application of the factor-analysis receptor model to simulated urban- and regional-scale data sets. *Atmospheric Environment* **21,** 2005–2013.

McCollister G. M. and Wilson K. R. (1975) Linear stochastic models for forecasting daily maxima and hourly concentrations of air pollutants. *Atmospheric Environment* **9,** 417–423.

Mulholland M. (1989) An autoregressive atmospheric dispersion model for fitting combined source and receptor data sets. *Atmospheric Environment* **23,** 1443–1458.

North M., Hernandez E. and Garcia R. (1984) Frequency analysis of high CO concentrations in Madrid by stochastic process modeling. *Atmospheric Environment* **18,** 2049–2054.

Press W. H. *et al.* (1987) *Numerical Recipes, The Art of Scientific Computing.* Cambridge University Press, New York.

Rhodes R. C. and Evans E. G. (1988) Precision and accuracy assessments for state and local air monitoring networks: 1986. EPA/600/S4-88-007. Environmental Monitoring Systems Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, NC.

Roberts E. M. (1979) Review of statistics of extreme values with applications to air quality data: Part I. Review. *J. Air Pollut. Control Ass.* **29,** 632–637.

Schweppe F. C. (1973) *Uncertain Dynamic Systems.* Prentice-Hall, Englewood Cliffs, NJ.

Shindo J., Ot K. and Matsumoto Y. (1990) Considerations on air pollution monitoring network design in the light of spatio-temporal variations of data. *Atmospheric Environment* **24B,** 335–342.

Shively T. S. (1990) An analysis of the long-term trend in ozone data from two Houston, Texas monitoring sites. *Atmospheric Environment* **24B,** 293–301.

Strothmann J. A. and Schiermeier F. A. (1979) Documentation of the regional air quality study and related investigations in the St. Louis air quality control region. EPA-600/4-79-076, Environmental Sciences Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC.

Toothman D. (1984) Development of an emission inventory for urban particle model validation in the Philadelphia AQCR. Environmental Sciences Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, NC.

Venkatram A. (1988) On the use of kriging in the spatial analysis of acid precipitation data. *Atmospheric Environment* **22,** 1963–1975.

Watson J. G., Cooper J. A. and Huntzicker J. J. (1984) The effective variance weighting for least squares calculations applied to the mass balance receptor model. *Atmospheric Environment* **18,** 1347–1355.