12 STATISTICAL METHODS

Statistical methods are frequently used in air pollution studies. Several types of statistical models, methods and analyses will be discussed in this chapter.

The fundamental aspects of atmospheric diffusion are based on statistical theories (Taylor, 1921), as confirmed by the recent, promising development of Lagrangian Monte–Carlo techniques for pollutant dispersion simulations. These methods have been discussed in Chapter 8. Moreover, the newly developed theories of chaos (Berge et al., 1984; Grebogi et al., 1987), also based on stochastic methods, seem promising for the treatment of turbulence in fluid flows. These statistical aspects of chaos theory will not, however, be discussed further here.

In this chapter, we will address the following topics: 1) frequency distribution of air quality measurements and the characterization of extreme values, a subject that is important for regulatory purposes; 2) time series analysis, in the time and the frequency domain; 3) the joint application of deterministic and statistical techniques (e.g., by using Kalman filters); 4) receptor modeling techniques; 5) the statistical methodologies that can be used to evaluate the performance of air quality dispersion models; 6) interpolation and graphic techniques, such as Kriging, pattern recognition, cluster analysis, and fractal analysis; and 7) optimization methods.

A general distinction between statistical and deterministic approaches is that air pollution deterministic models initiate their calculations at the pollution sources and aim at the establishment of cause–effect relationships, while statistical models are characterized by their direct use of air quality measurements to infer semiempirical relationships. Statistical models are useful in situations such as real–time short–term forecasting, where the information available from measured concentration trends is generally more relevant than that obtained from deterministic analyses.

The reader can find some general information about basic statistical methods for air pollution data in Gilbert (1987). Statistical analysis with missing data — a problem often encountered in environmental applications — is discussed by Little and Rubin (1987). References and more specific applications are discussed in the sections below.

12.1 FREQUENCY DISTRIBUTION

As noted by Seinfeld (1986), "air pollution concentrations are inherent random variables because of their dependence on the fluctuations of meteorolo cal and emission variables." A random variable is characterized by two ma factors: its probability density function and its autocorrelation structure. Tl probability density function *pdf(c)* gives the probability *pdf(c)dc* that the conce tration c, of a certain species at a particular location during a certain time perio is between c and $c+dc$. The autocorrelation function quantifies the "serial" beha ior of the "time series" of concentrations, i.e., the relationship between the co centration value at t, and those at previous times, for a certain species at particular location. Nonstationary phenomena strongly complicate these statis cal characterizations. Often, if not always, concentration measurements posse highly nonstationary features. Under these conditions, both *pdf* and autocorrel tion vary with season and even the hour of the day (Zannetti et al., 1978).

The autocorrelation structure of concentration values plays a key role understanding the variation with time of concentration. For example, high pos tive autocorrelation means that peak concentration values tend to be followed l high values and that clean periods tend to be followed by low pollution periods a behavior that is typically measured in air pollution studies. The evaluation the *pdf*, however, has received more attention in air pollution statistical studi because its determination is useful in regulatory applications based on the co cept of air quality "standards," i.e., ambient concentration values that should n be exceeded. In other words, the knowledge, from direct measurements or oth techniques, of the *pdf* allows the calculation of the exact number of violations (expected violations) of a specified air quality standard, as illustrated in Fi ure 12–1), where the integral of *pdf(c)dc*, from the air quality standard value to infinity, gives the probability of exceedance of c_s.

Therefore, it is important to calculate or estimate the *pdf*. This operatic requires two steps: 1) the evaluation of *the form* of the *pdf* (e.g., a log–norm *pdf*) and 2) the evaluation of the parameters of this chosen form. Several fr quency distribution functions have been proposed and used to fit air quali measurements. Georgopoulos and Seinfeld (1982), Tsukatami and Shigemits (1980), and Marani et al. (1986) discuss and summarize several of them, inclu ing the following distributions:

- Log–normal
- Weibull
- Gamma

- Three-parameter log-normal
- Three-parameter Weibull
- Three-parameter beta
- Four-parameter beta
- Pearson

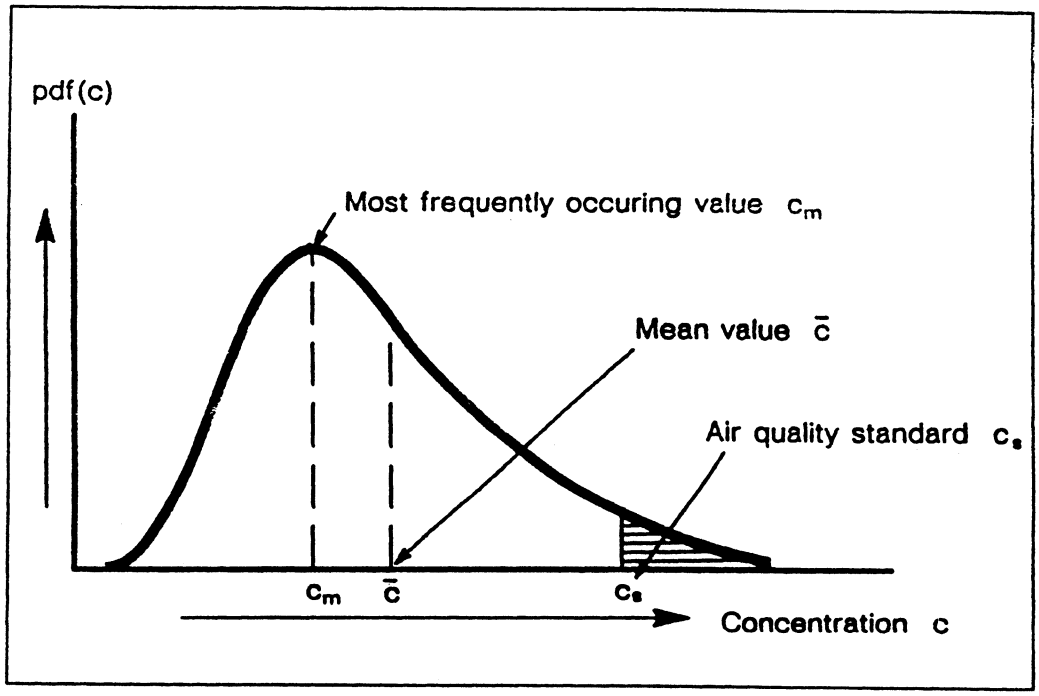

Figure 12–1. Example of application of the pdf to calculate the probability of exceedance of air quality standard c_s.

The most common distribution is the log-normal distribution, which is represented by

$$pdf(c) = \frac{1}{(2\pi)^{1/2}\, c\, \ln \sigma_g} \exp\left[-\frac{(\ln c - \ln \bar{c}_g)}{2 \ln^2 \sigma_g)}\right] \qquad (12\text{–}1)$$

where $\bar{c}_g$ is the geometric mean and σ_g is the geometric standard deviation of c. According to this distribution, the logarithms of the concentrations have a normal (i.e., Gaussian) distribution $N(\ln \bar{c}, \ln \sigma_g)$. Even though heuristic justifications have been provided for explaining the occurrence of log-normal distributions (Cats and Holtslag, 1980; Kahn, 1973), no *a priori* reason seems to exist for preferring one distribution above the others (Seinfeld, 1986).

The log-normal distribution has been studied by several authors, including Larsen (1971), whose pioneering work identified the following relations, sometimes referred to as Larsen's laws:

1. Pollutant concentrations are lognormally distributed for all averaging times

2. Median concentrations (50th percentile) are proportional to the averaging time raised to an exponent

3. Maximum concentrations are approximately inversely proportional to the averaging time raised to an exponent

An example of SO_2 concentration measurements mostly in agreement with the above laws is presented in Figure 12-2. Other data sets, however, may show less agreement (especially secondary pollutants, such as NO_2 and O_3).

Frequency distributions are mostly used to assess the probability of occurrence of *high* concentration values (e.g., exceedances). Therefore, it is most important that these distributions be accurate at their right tail than elsewhere. It is well known, however, that statistics of extreme values are the most affected by uncertainties, even though *ad hoc* methods have been proposed to handle extreme values (e.g., Roberts, 1979; Williams, 1984; Drufuca and Giuliano, 1978; Horowitz and Barakat, 1979; Chock and Sluchak, 1986; and Surman et al., 1987). It is unfortunate, from a statistical point of view, that U.S. air quality standards were chosen as values that can be exceeded only once a year. This choice has put the standards at the very end of the frequency distribution tail, where little confidence can be given to either measurements or theoretical estimates. Alternative choices of standards, i.e., the 95th percentile, would have been more "stable" and less questionable. Some countries are moving toward this alternative. For example, Italy introduced, in 1983, new SO_2 standards based on the 50th and 98th percentiles of the SO_2 daily average concentrations, instead of the previous 30-minute and daily not-to-be-exceeded standards.

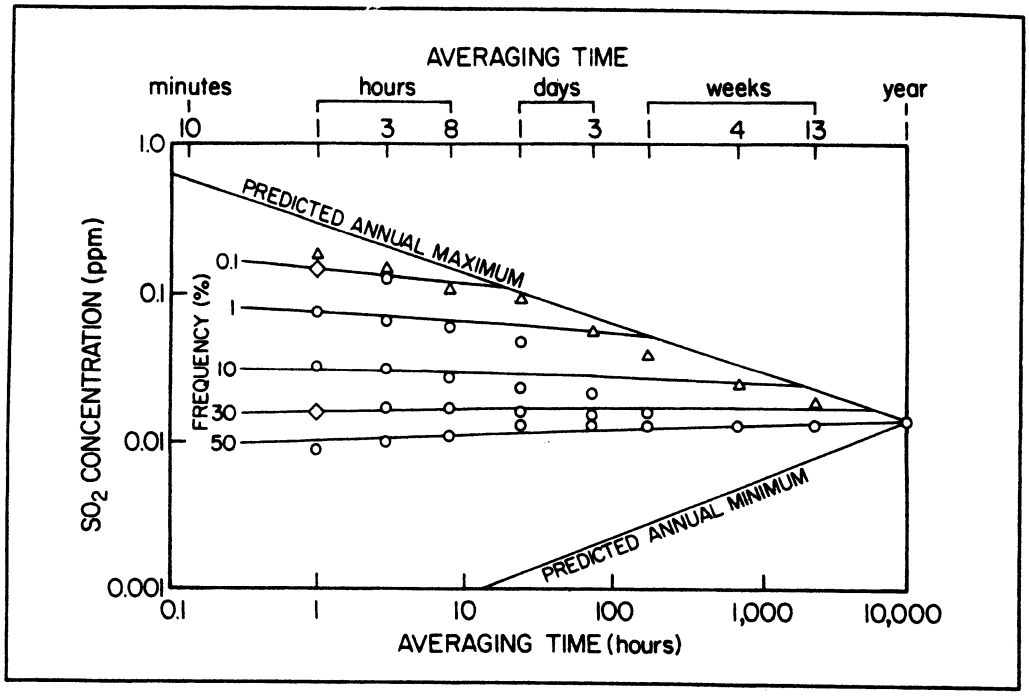

Figure 12–2. Sulfur dioxide concentration versus averaging time and frequency for 1980 at the United States National Aerometric Data Bank (NADB) Site 264280007 HO1, 8227 S. Broadway, St. Louis, Missouri. Source: Chart courtesy of Dr. Ralph Larsen, United States Environmental Protection Agency, Research Triangle Park, North Carolina (from Stern et. al., 1984). [Reprinted with permission from Academic Press.]

12.2 TIME SERIES ANALYSIS

Time series analysis methods aim at the analysis of data arranged in a time sequence, either in the time domain (e.g., Box–Jenkins methods) or in the frequency domain (e.g., spectral analysis). They include the following methods:

- Box–Jenkins approach
- Spectral analysis
- Regression analysis
- Trend analysis
- Principal components analysis

These statistical modeling approaches can be used in a "black box" mode when, for example, time series of concentrations are analyzed without any other

information, just to evaluate their intrinsic variations and without attempting any physical explanation. Or, they can be used in a "gray box" mode, in which other parameters, for example meteorological and emission terms, are included, in an effort to incorporate, within a statistical frame, deterministic relations.

These methods can be used either in a "batch" or a "real-time" mode. Batch simulations perform statistical analyses of past measurements, in an effort to establish empirical relationships. Real-time applications (e.g., Bacci et al., 1981; Finzi and Tebaldi, 1982) require the availability of on-line data and provide forecasts that can be used by decision-makers for real-time intervention strategies, in a effort to mitigate possible incoming concentration episodes.

The Box-Jenkins methodology (Box and Jenkins, 1970; new edition, 1976) is considered the most cost-effective approach for time-series analysis and has been frequently applied to evaluate meteorological and air quality measurement patterns. This theory has been described and summarized in several books and articles and will not be discussed here. The general form of the Box-Jenkins equation to describe a time series is

$$\phi_p(B)\ \Phi_P(B^s)\ \nabla^d\ \nabla_s^D\ z_t\ =\ \theta_q(B)\ \Theta_Q(B^s)\ a_t \qquad (12\text{-}2)$$

where ϕ_p is the autoregressive operation of order p, Φ_p is the seasonal autoregressive operation of order P with seasonality s, B is the backward operator, ∇ is the difference operator, ∇_s is the seasonal difference operator, z_t is the time series, θ_q is the moving average operator of order q, Θ_Q is the seasonal moving average operator of order Q and a_t is a Gaussian white noise. In general, however, simple forms of Equation 12-2 (with only three to four terms) are sufficient to well characterize z_t. Figure 12-3 shows an example of Box-Jenkins forecasting.

Important simulations and results, using the Box-Jenkins theory, have been provided by Chock et al. (1975); by Simpson and Layton (1983) for the forecasting of ozone peaks; by Tiao el al., (1975), who modified the Box-Jenkins approach for treating the effects of intervention strategies during the period in which the time series has been collected; by Roy and Pellerin (1982) for the evaluation of long-term trends and intervention analysis; by Zinsmeister and Redman (1980) for aerosol data; and by Murray and Farber (1982) for evaluating and historical visibility data base.

Spectral analysis techniques (Jenkins and Watts, 1968) allow the identification of cycles in meteorological and air quality time-series measurements. In particular, two studies (Tilley and McBean, 1973; Trivikrama et al., 1976) first

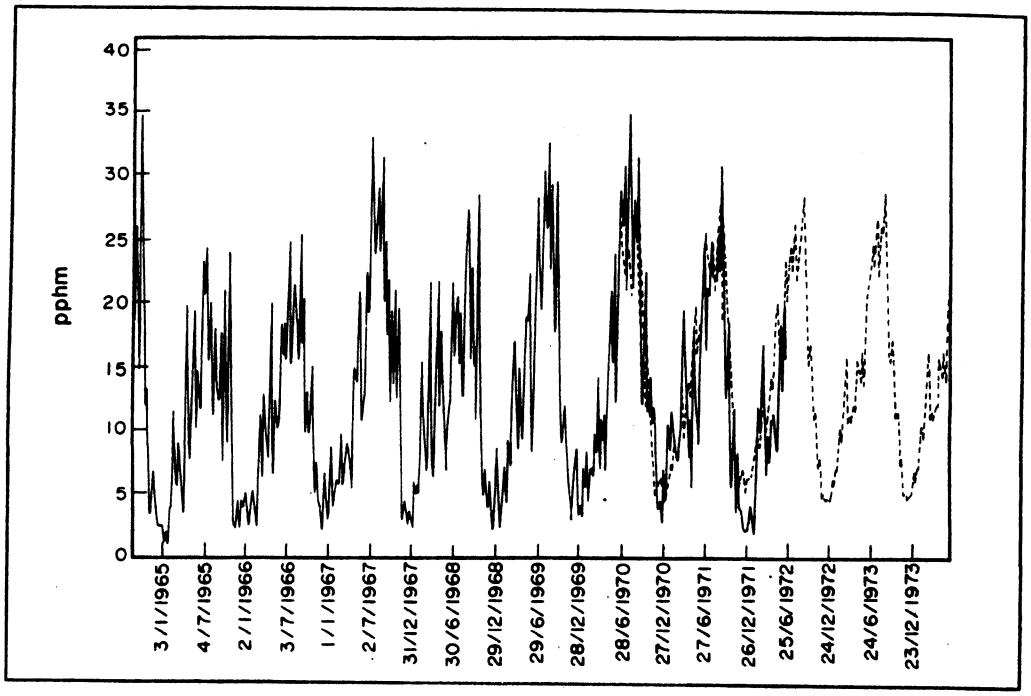

Figure 12–3. Univariate time series analysis of oxidant concentrations. The solid line indicates actual data and the dashed line forecasted data (from Chock et al., 1975). [Reprinted with permission from Pergamon Press.]

showed the existence of the following main oscillations of SO_2 and wind speed data: semidiurnal, diurnal, and three- to three-and-a-half-day period. Semidiurnal cycles have been ascribed to local phenomena (like the sea breeze), while the longest period seems to be caused by synoptic weather variations, which have a period close to three-and-a-half days in the study area (northeastern United States). Figure 12–4 shows an example of spectral analysis of SO_2, wind speed, temperature and pressure .

Regression analyses are a particular type of multiple time-series analysis, in which, for example, meteorological measurements are statistically related to air quality concentrations. Examples of multiple linear regression studies are presented in Figure 12–5, where visibility reduction is interpreted as a function of pollutant concentrations and meteorological conditions, and Figure 12–6, where oxidant concentrations are predicted as a multiple linear regression of the logarithm of the solar radiation intensity, wind speed and dry bulb temperature.

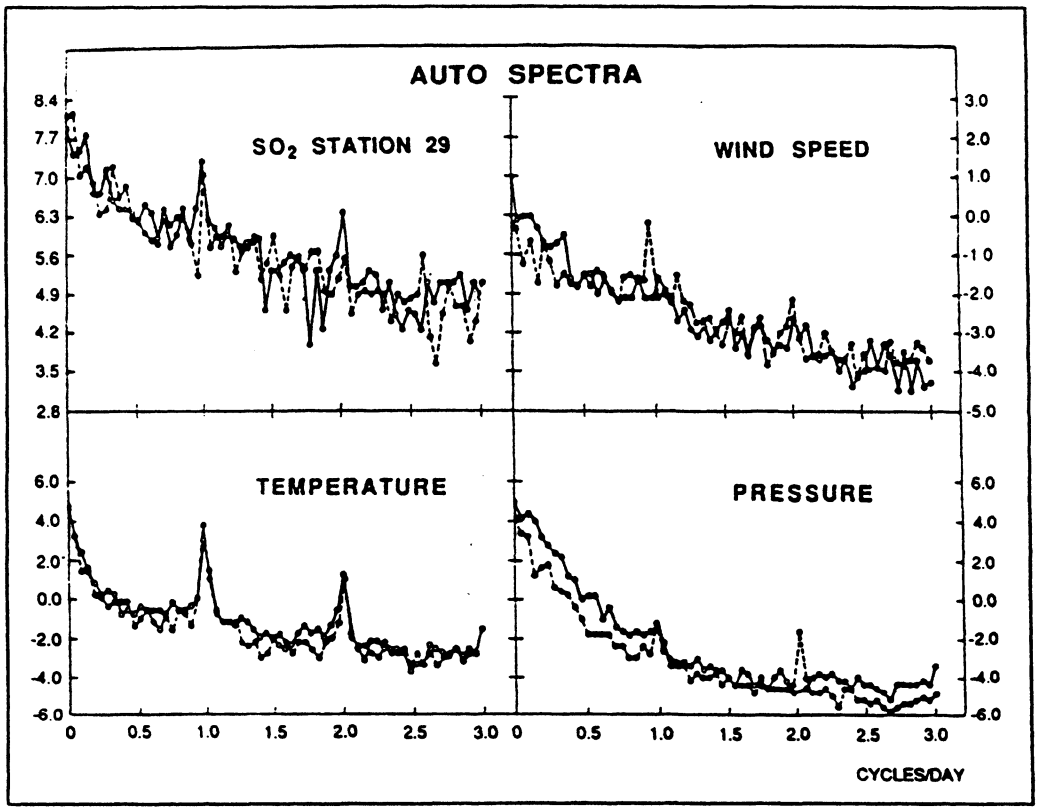

Figure 12–4. Logarithm of the power autospectra of SO_2 and meteorological hourly data (wind speed, temperature and pressure) during a six-month summer period (dots) and a six-month winter period (circles) (from Zannetti, et al., 1978a). [Reprinted with permission from Pergamon Press.]

Trend and seasonal variations are also assessed by multiple regression models, e.g., as performed by Buishand et al. (1988). Principal components have also been used to calculate pollutant distributions and predictions (Petersen, 1970; Henry and Hidy, 1979; Lin, 1982).

Time series methods can be applied in two modes: a "fitting" mode and a "forecasting" mode. In the forecasting mode, the model parameters (e.g., the regression coefficients) are estimated from one set of measurements and, subsequently, the time series model is applied, with these estimated parameters, to another set of measurements to calculate the forecasting performance. In the fitting mode, the same set of measurements is used for both parameter estimation and model performance evaluation. Clearly, only the forecasting mode

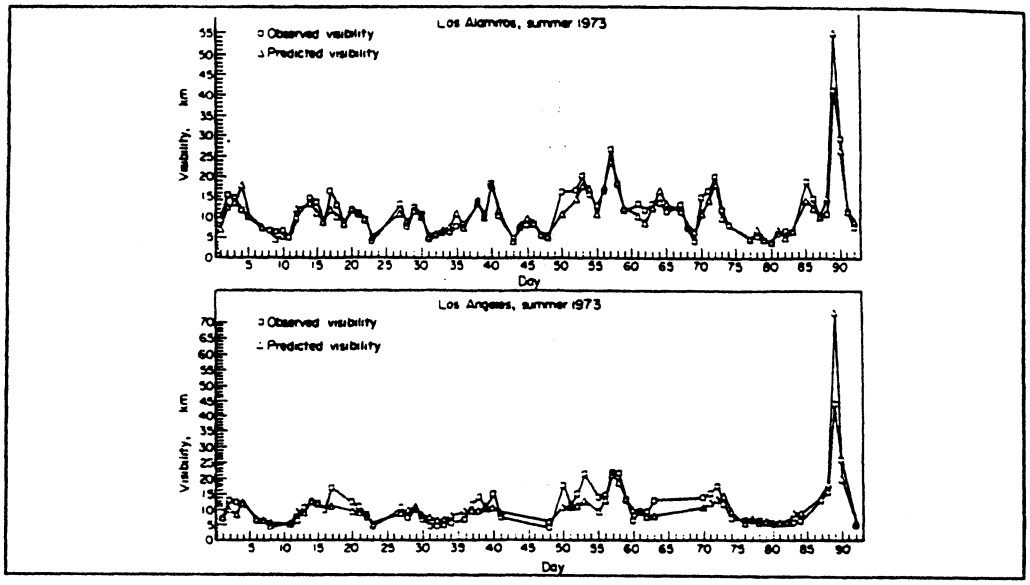

Figure 12–5. *Observed and predicted values for visibility (visual range in km) at Los Alamitos and Los Angeles, Summer 1973 (from Barone et al., 1978). [Reprinted with permission from Pergamon Press.]*

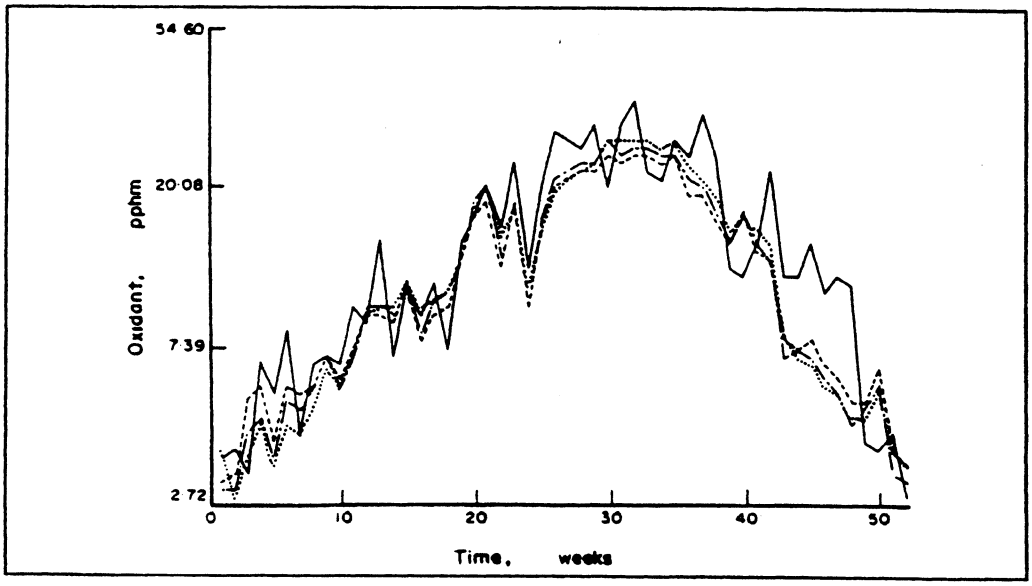

Figure 12–6. *Predicted and actual oxidant levels (in log scale) for 1970. —— actual values; —— Aitken's model; –•–•– OLS model; ••••• lagged model (long term prediction) (from Chock et al., 1975). [Reprinted with permission from Pergamon Press.]*

provides an unambiguous evaluation of model performance, while the fitting mode overestimates the model's forecasting ability.

A useful combination of fitting and forecasting can be obtained by applying time series models in an "adaptive" mode, in which, at each time t, the model parameters are re-estimated using the measurements of a "learning" period of duration T (i.e., from $t-T$ to t). In this way, if the duration T is appropriate, the model's performance can be maximized. Figure 12–7 shows an example of the adaptive technique using two simple models (AR(1) and AR(1)CS). Note that both models have an optimum T which gives a forecasting performance that exceeds even that of the fitting model.

12.3 MIXED DETERMINISTIC STATISTICAL MODELS (KALMAN FILTERS)

Semiempirical methods and real-time filters, especially the Kalman filters, have been frequently used for updating the forecasting capabilities of a predictor (generally a *deterministic* predictor) based upon the availability of real-time measurements of the system variables. (A common application of the Kalman filter is in navigation space guidance and orbit determination, where computations are dynamically changed according to real-time measurements of the flying object's position and velocity.) The Kalman filter technique and its application to air pollution problems are discussed below.

12.3.1 Introduction to Kalman Filters

The principle of least squares estimation originated at the beginning of the 19th century, but only a century and a half later, starting with the pioneering work of Wiener (1949), a substantial innovation allowed its "recursive" application, as explained below.

Let us consider, following the discussion by Young (1974), the linear regression problem

$$y = \mathbf{x}^T \mathbf{a} \tag{12–3}$$

in which a variable y is related to n other linearly independent variables $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$, through the unknown coefficients $\mathbf{a} = [a_1, a_2, \ldots, a_n]^T$. Then, if we have k observations of y and $\mathbf{x}$, we can obtain an estimate $\mathbf{a}'_k$ of $\mathbf{a}$ by using the least squares method, as follows

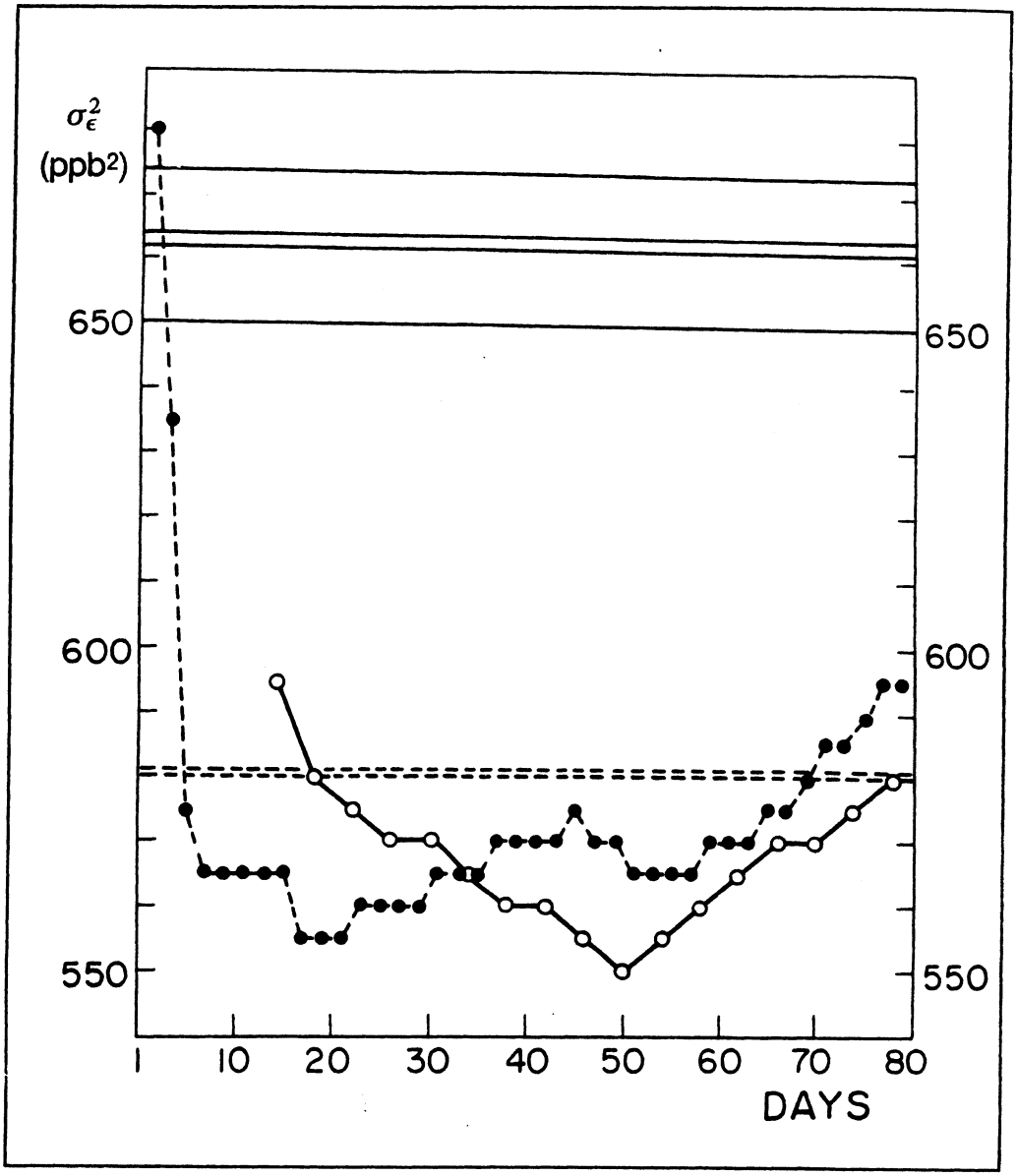

Figure 12–7. *Values of σ_ϵ^2, the forecasting error variance, for the AR(1) (dots) and AR(1)CS (circles) adaptive models applied to hourly SO_2 data (one–year analysis). The error variance is plotted against the length T of the learning period. The horizontal lines show comparable σ_ϵ^2 values of nonadaptive models for fitting (dashed lines) and forecasting (solid lines) cases. (Forecasting was obtained by using the parameters estimated during the same season one year before) (from Zannetti, 1978). [Reprinted with permission from the Air Pollution Control Association.]*

$$\mathbf{a}'_k = \left[\sum_{i=1}^{k} \mathbf{x}_i \mathbf{x}_i^T\right]^{-1} \sum_{i=1}^{k} \mathbf{x}_i y_i \qquad (12\text{–}4)$$

But if the number k is increasing, i.e., if new observations of y and $\mathbf{x}$ are progressively gathered, the updating of the estimate $\mathbf{a}'_k$ requires a repeated application of Equation 12–4. To avoid this expensive calculation, Plackett (1950) rewrote Equation 12–4 in a "recursive" form in which $\mathbf{a}'_k$ is a linear sum of the estimate obtained after k–1 observations (i.e., $\mathbf{a}'_{k-1}$), plus a correction term based on the newly–received information y_k and $\mathbf{x}_k$. This recursive form provides results mathematically identical to Equation 12–4.

The next step was provided by Kalman (1960), who expanded the work of Wiener (1949) and solved the general problem of estimating a set of parameters $\mathbf{a}_k$ in which:

- $\mathbf{a}_k$ represents, in a more general form, the "state" of a dynamic system
- parametric invariance (i.e., $\mathbf{a}_k = \mathbf{a}_{k-1}$, for all k) is not assumed any more
- the parameters $\mathbf{a}_k$ vary according to the general stochastic evolution scheme ("state equation" or "message model")

$$\mathbf{a}_k = \mathbf{F}(k, k-1)\mathbf{a}_{k-1} + \mathbf{G}(k, k-1)\mathbf{w}_k \qquad (12\text{–}5)$$

  where $\mathbf{F}(k, k\text{–}1)$ is an $n \times n$ transition matrix, $\mathbf{G}(k, k\text{–}1)$ is an $n \times m$ input matrix, and $\mathbf{w}_k$ is an $m \times 1$ vector of independent random variables with zero mean and covariance matrix $\mathbf{Q}$.

- "noisy" measurements $\mathbf{y}_k = [y_1, y_2, \ldots, y_p]_k^T$ are available that are linearly related to $\mathbf{a}_k$ by the "observation equation" or "observation model"

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{a}_k + \mathbf{v}_k \qquad (12\text{–}6)$$

  where $\mathbf{H}_k$ is a $p \times n$ coefficient matrix and $\mathbf{v}_k$ is the measurement error assumed to be a $p \times 1$ vector of independent random variables with zero mean and covariance matrix $\mathbf{R}$

Then, the equations of the Kalman filtering method allow a recursive computation of

- the estimates $\mathbf{a}'_{k+j}$ (j = 1, 2, ...) of $\mathbf{a}_{k+j}$ by considering only the effect of the most recent observation $\mathbf{y}_k$ instead of resolving at

each time the entire problem by the classical least squares regression technique

- the estimate of the covariance matrix of the forecasting error $\mathbf{a}_{k+j} - \mathbf{a}'_{k+j}$ and, therefore, an important indication of the accuracy of the estimates $\mathbf{a}'_{k+j}$ and information on their convergence

In the Kalman filter outlined above, the state vector $\mathbf{a}_k$ can be any numerical description of the state of a dynamic system, e.g., the location of a space ship, concentrations of pollutants in the atmosphere, or a velocity field representing groundwater dynamics. Then, the transition matrix $\mathbf{F}$ contains our (imperfect) deterministic representation of the phenomenon (e.g., a set of physical equations reduced into a linear matrix form(*)), and $\mathbf{y}_k$ are the limited measurements available. Then the Kalman filter, as outlined below, provides a method for forecasting the evolution of $\mathbf{a}_k$, which takes into account both the "deterministic" component $\mathbf{F}$ (predictor) and the continuous, on-line information (corrector) provided by the measurements $\mathbf{y}_k$.

Starting from Equations 12-5 and 12-6, it is possible to develop (Jazwinski, 1970; Sage and Melsa, 1971) an unbiased linear minimum-error-variance algorithm (Kalman filter) to estimate the state of a linear time-varying dynamic system driven by white noise of zero mean and known variance. Under the further assumptions that $\mathbf{v}$, $\mathbf{w}$, and $\mathbf{a}$ are mutually uncorrelated, the relevant formulae, where $\mathbf{a}(t_2|t_1)$ is the estimate at time t_1 of $\mathbf{a}(t_2)$, are (**)

- predicted state

$$\mathbf{a}(t+1|t) = \mathbf{F}(t+1,t)\,\mathbf{a}(t|t) \tag{12-7}$$

- predicted error covariance matrix

$$\mathbf{V}_{\tilde{\mathbf{a}}}(t+1|t) = \mathbf{F}(t+1,t)\,\mathbf{V}_{\tilde{\mathbf{a}}}(t|t)\,\mathbf{F}^T(t+1,t) + \mathbf{G}(t)\,\mathbf{V}_{\mathbf{w}}(t+1)\,\mathbf{G}^T(t) \tag{12-8}$$

- filter gain matrix

$$\begin{aligned}\mathbf{K}(t+1) = {} & \mathbf{V}_{\tilde{\mathbf{a}}}(t+1|t)\,\mathbf{H}^T(t+1) \\ & [\mathbf{H}(t+1)\,\mathbf{V}_{\tilde{\mathbf{a}}}(t+1|t)\,\mathbf{H}^T(t+1) + \mathbf{V}_{\mathbf{v}}(t+1)]^{-1}\end{aligned} \tag{12-9}$$

(*) Nonlinear Kalman filters are also available, but will not be discussed here.

(**) In the formulae below, time is explicitly indicated by t, instead of using the subscript k as in the notation of the previous equations.

- filtered state after processing the observation $\mathbf{y}(t+1)$

$$\mathbf{a}(t+1|t+1) = \mathbf{a}(t+1|t) + \mathbf{K}(t+1)\,[\mathbf{y}(t+1) - \mathbf{H}(t+1)\,\mathbf{a}(t+1|t)] \quad (12\text{–}10)$$

- new error covariance matrix

$$\mathbf{V}_{\tilde{\mathbf{a}}}(t+1|t+1) = [\mathbf{I} - \mathbf{K}(t+1)\,\mathbf{H}(t+1)]\,\mathbf{V}_{\tilde{\mathbf{a}}}(t+1|t) \quad (12\text{–}11)$$

where $\mathbf{V}_{\tilde{\mathbf{a}}}(t_2|t_1)$ is the covariance of the error $\tilde{\mathbf{a}} = \mathbf{a}(t_2) - \mathbf{a}(t_2|t_1)$ and $\mathbf{I}$ is the identity matrix. See Zannetti and Switzer (1979a) for a rewriting of the above methodology in a computer-oriented recursive form referred to a forecast performed from 1 to p time steps ahead.

12.3.2 Applications of Kalman Filters to Air Quality Problems

Kalman filters have been used in air pollution problems to obtain more accurate predicted values in episode forecasting and control. This can be done by considering $\mathbf{a}_k = \mathbf{a}(t)$ as the vector of concentrations of a pollutant at the grid points of a grid dispersion model (Bankoff and Hanzevack, 1975; Melli et al., 1981) or at the pollutant monitoring points (Sawaragi et al., 1976) of the study area. The state vector $\mathbf{a}(t)$ might also be extended to include some additional adaptive parameters (Bankoff and Hanzevack, 1975), but we will not discuss this extension here. Then the transition matrix $\mathbf{F}$ becomes either the matrix of the spatial discretization (K-model) of the transport and diffusion equation (Bankoff and Hanzevack, 1975; Melli et al., 1981) (including time-dependent emission and meteorological inputs) or a multiple regression matrix (Sawaragi et al., 1976). Model inaccuracies and emissions and meteorology input errors are included in the system noise process $\mathbf{w}(t)$.

A hybrid (*) air quality application of the Kalman filter was developed by Zannetti and Switzer (1979a) who limited the dimension of the state vector $\mathbf{a}$ to the number of air quality monitoring stations, but incorporated the contribution of the meteorology by having the transition matrix $\mathbf{F}$ depend upon the time-varying meteorological conditions. They evaluated the coefficients of $\mathbf{F}$ in an "adaptive" manner, i.e., using a first-order Markov chain on a learning time period of given fixed length close to the forecasting time (see previous discussion at the end of Section 12.2 for a description of the "adaptive" forecasting technique).

(*) This approach is a hybrid one since the matrix $\mathbf{F}$ is not computed using a set of deterministic equations, but is calculated using statistical methods, in which a different matrix $\mathbf{F}$ is estimated for each meteorological class.

An important problem arises in the application of the Kalman filter to air pollution problems. In fact, it is necessary to avoid the high dimensionality of the resulting Kalman filter equations. For example, when **F** is the time-evolution transition matrix of the *K*-model, a simple spatial grid of 20 × 20 × 10 points produces Kalman filter matrices of dimension 4000 × 4000. Methods have been developed to simplify this problem. In particular, either the Green function can be used to reduce the equation of the *K*-model to a difference equation of relatively small dimension (Hino, 1974), or a discrete form of Chandrasekar-type equations (*) can be applied for the same goal (Desalu et al., 1974). Alternatively, the region can be partitioned into subregions (Bankoff and Hanzevack, 1975; Melli et al., 1981) and, if the subvectors of the subregions are not coupled (or are weakly coupled), the filter algorithm can be applied separately to each of the subvectors, thus reducing the size of the matrices that must be manipulated. Finally, a multiple linear regression model can be used (Sawaragi et al., 1976) for **F**, thus reducing the dimension of the filter to the number of monitoring stations in the area. However, this loses the "physical" information of the diffusion phenomenon.

An example of use of Kalman filters to forecast air pollution episodes is shown in Figure 12-8. Note that the forecasting performance of the method decreases when, instead of using the actual meteorological data, meteorological forecasts are used.

12.4 RECEPTOR MODELS

Receptor models are the dream of the air pollution experimentalist. A dream that, till now, has been only partially fulfilled. The basic concept of the receptor modeling approach is the apportionment of the contribution of each source, or group of sources, to the measured concentrations without reconstructing the dispersion pattern of the pollutants. In other words, while dispersion models compute the contribution of a source to a receptor as the product of the emission rate by a dispersion factor, receptor models start with *observed* ambient aerosol concentrations at a receptor and seek to apportion the observed concentrations among several source types (e.g., industrial, transportation, soil, etc.), based on the known chemical compositions (i.e., the chemical fractions) of source and receptor materials.

(*) This alternative method completely bypasses direct calculation of the covariance matrices, while still retaining the properties of the Kalman filter.

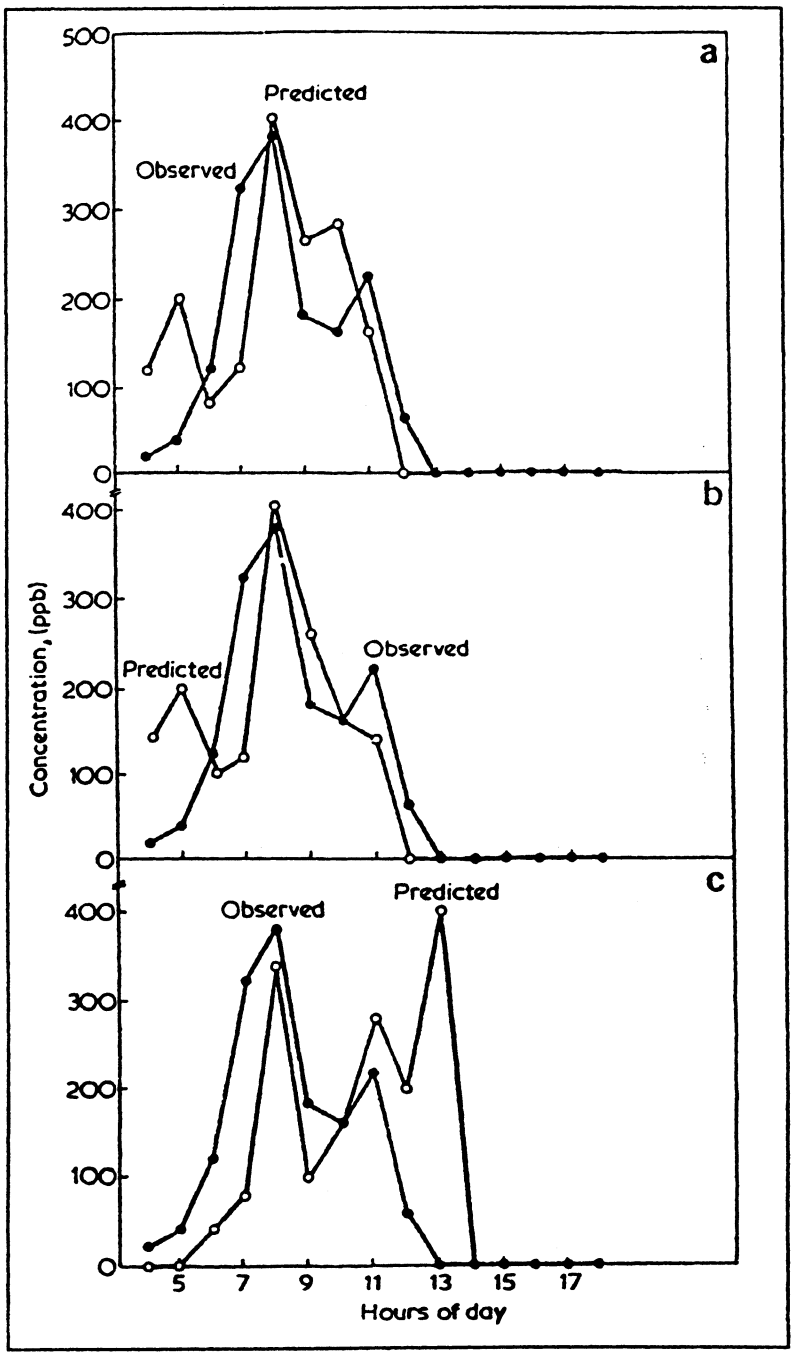

Figure 12–8. *Kalman predictions of a four hour–ahead episode with different meteorological inputs: (a) meteorology is forecast, (b) true meteorological values are used, (c) meteorology is persistent. From Fronza et al. (1979). [Reprinted with permission from Butterworth Scientific.]*

In mathematical notation, the concentration c_{ik} of the species i in the k–th sample at a certain monitoring station can be written as

$$c_{ik} = \sum_{j=1}^{p} a_{ij} D_{jk} E_{jk} \qquad (12\text{–}12)$$

where p sources (or groups of sources) are assumed to contribute to c_{ik}, a_{ij} is the fractional amount of the component i in the emission from the j–th source, D_{jk} is the atmospheric dispersion term and E_{jk} is the emission rate (i.e., $D_{jk} E_{jk} = S_{jk}$ is the total contribution of the source j to the k–th sample in the receptor). Dispersion models assume a_{ij}, D_{jk} and E_{jk} to be known (or obtainable from emission and meteorological data) and estimate the output c_{ik}. For receptor models, the concentrations c_{ik} and source "profiles" a_{ij} are measured instead, and the D_{jk} E_{jk} products are computed as a model result.

Receptor models can be classified into four categories (Henry et al., 1984):

- chemical mass balance (CMB)
- multivariate models
- microscopic models
- source–receptor hybrids

The first two categories are discussed below.

12.4.1 Chemical Mass Balance (CMB)

Chemical mass balance (CMB) models are based on a sample of n chemical properties of both source and receptor, thus giving n equations

$$c_i = \sum_{j=1}^{p} a_{ij} S_j, \; i = 1, 2, \ldots n \qquad (12\text{–}13)$$

Then, if $p \leq n$, the source contributions S_j can be computed by solving the overdetermined linear system (12–13).

Henry et al., (1984) identify five methods of calculating S_j:

1. the tracer property method, which simply assumes that each source j possesses a unique species i which is common to no other source

2. the linear programming method
3. the ordinary least-squares method, which estimates S_j by minimizing the sum of squares of the differences between the measured c_i values and those calculated by Equation 12-3 weighted by the analytical uncertainty of the c_i measurement
4. the effective variance least-square method, which includes the consideration of the errors in the a_{ij} terms and provides more reliable confidence intervals for the outputs S_j
5. the ridge regression, which is one approach useful in handling the "multicollinearity" problem; i.e., the case when two or more sources have similar chemical composition (in this case the least-squares solutions are mathematically unstable)

12.4.2 Multivariate Models

Multivariate models are used to solve Equation 12-12 in which multiple sampling data (k = 1, 2, ...) are considered. The objective of the multivariate models is to use the c_{ik} measurements for predicting

- the number p of sources affecting the monitoring station
- which a_{ij} is associated with which S_j
- when possible, both a_{ij} and S_{jk}

Multivariate methods include (Henry et al., 1984)

- factor analysis based on eigenvector analysis of the cross-product data matrix. (Caution should be used in applying this technique. As pointed out by Henry (1987), current factor analysis receptor models are "biased" in the statistical sense and, in inexperienced hands, can give large errors in source apportionment.)
- target transformation factor analysis, for extracting maximum information about the number and nature of sources with no or very limited *a priori* information other than the elemental composition data (see also Hopke, 1988)
- multiple linear regression, a linear least-squares fitting process that requires a tracer element to be determined for each source j (or each source category)
- extended Q-mode factor analysis, which is a CMB-type model (single sample) that uses multivariate methods to deconvolve the receptor composition into a sum of source compositions

All the above receptor modeling techniques are still under theoretical and empirical development. Review papers are provided by Watson (1984), Henry et al., (1984), and Gordon (1988). Receptor models are becoming a major analysis tool and are much applied, especially for aerosol mass apportionment computations (e.g., see the article of Scheff el al., 1984, for the Chicago area, and the article of Chow et al., 1985, for Portage, Wisconsin). Receptor models seem extremely powerful and promising tools for analyses intended to complement but not to substitute for the information provided by dispersion modeling techniques. Actually, mixed dispersion–receptor modeling methodologies (e.g., Chow et al., 1985) seem to be the most promising. The need for this mixed approach is well indicated by the schematic illustration presented in Figure 12–9. Much investigation, however, is still required to assess the degree of reliability of receptor techniques and, especially, their sensitivity to input data errors.

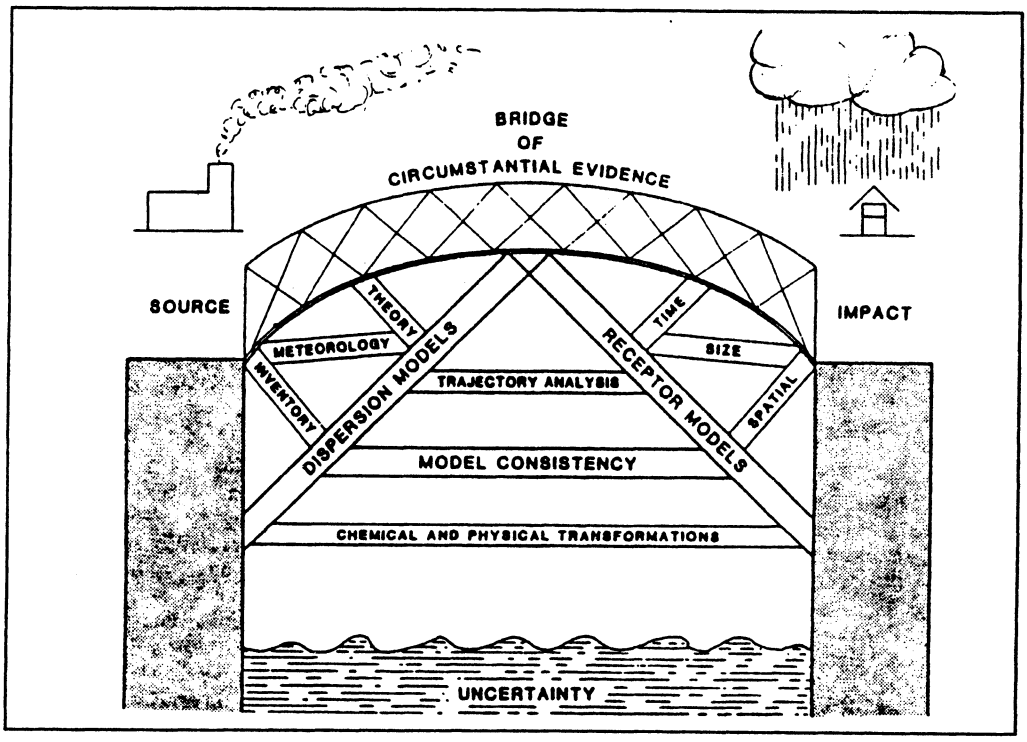

Figure 12–9. Schematic illustration of the need to use many different independent approaches to establish a strong bridge of circumstantial evidence quantitatively linking a source to its impact (from Cooper, 1983). [Reprinted with permission from Pollution Atmospherique.]

12.4.3 Receptor Models for Secondary Particulate Matter

The receptor modeling techniques presented above can simulate only primary particulate matter. Modifications have been recently proposed (Malm et al., 1989) to include adjustments that allow the simulation of secondary particulate matter (primarily sulfates) and deposition phenomena. In mathematical notation, we can rewrite Equation 12–12 as

$$c_{irt} = \Sigma_j a_{ijt} E_{jt} D_{jrt} a_{ijrt} \qquad (12\text{–}14)$$

where c_{irt} is the concentration of the species i in the sample at the receptor r during the time interval t, a_{ijt} is the fraction of emission of the species i from the source j, E_{jt} is the total emission from source j, D_{jrt} is the dispersion factor from the source j to the receptor r, and a_{ijrt} is the adjustment for gain or loss of the species i traveling from the source j to the receptor r.

With this approach, sulfur can be traced by a receptor model by including its emission as SO_2 and the transformation of some SO_2 to SO_4^{2-}. For example, for sulfur as SO_2, the term a can be defined as

$$a_{ijrt} = (1 - f_d)\,(1 - f_c) \qquad (12\text{–}15)$$

where f_d is the mass fraction of SO_2 that is deposited and f_c is the mass fraction of SO_2 that is chemically converted to SO_4^{2-}, both before reaching the receptor r. For sulfur as SO_4^{2-}, we have instead

$$a_{ijrt} = (1 - f_d')\, f_c \qquad (12\text{–}16)$$

where f_d' is the mass fraction of total sulfur that is deposited before reaching the receptor r and f_c is the same as above.

It is evident that a correct determination of f_d, f_d' and f_c requires correct assumptions on deposition and chemical transformation along the air parcel trajectory and, therefore, the application of some sort of deterministic method. The practical application of this technique, therefore, requires a hybrid approach, in which dispersion models still need to be applied to provide the input f_d, f_d' and f_c required by the receptor model.

12.5 PERFORMANCE EVALUATION OF DISPERSION MODELS

The performance of both Lagrangian and Eulerian dispersion models can be estimated by comparing their predictions against field measurements. Tracer

experiments are particularly helpful in evaluating the capability of these models to properly simulate transport and diffusion. Comparison between model outputs and measurements are performed using both qualitative data analysis techniques and quantitative statistical methods.

Initially (say, till a decade ago), this comparison was simple. The outputs of dispersion models were plotted against measurements and simple parameters such as the correlation coefficient were computed. High correlation values (a rare result) indicated that the model was good, low correlation (the most common case) that the model was poor. It is now clear that the problem is not so simple.

First of all, there are measurement errors, a fact that often seems forgotten in the common belief that monitoring data are "the real world." More importantly, even error-free measurements possess space and time limitations that prevent their use beyond their "representativeness" regions around the monitoring point. These representativeness regions are often very small and the comparison of measurements with grid-averaged model outputs is inappropriate. Second, certain statistical parameters, such as the correlation coefficient, can provide misleading results (e.g., Zannetti and Switzer, 1979b). Third, it has been shown that models possess intrinsic uncertainties (e.g., Venkatram, 1988a) that cannot be removed and that their outputs are "ensemble" averages, while measurements are just "realizations" (Lamb, from Longhetto, 1980). Fourth, and most important, models rely upon emission and meteorological inputs. Often the errors in the determination of these inputs fully justify the disagreements between predictions and observations (e.g., Irwin et al., 1987). In other words, the old computer law "garbage in, garbage out" can happen here, too.

In the last decade, several methods for systematic statistical evaluation of air quality model performance have been proposed (e.g., see the survey by Bornstein and Anderson, 1979, and the methodologies proposed by Venkatram, 1982 and 1983). But the most innovative results came from two workshops organized by the American Meteorological Society. These workshops provided specific guidelines on the use of statistical tools in air quality applications; a summary of their recommendations is provided in two papers by Fox (1981 and 1984).

The most interesting comments and recommendations from the above workshops were

- the concern about the absolute, rather than statistical nature of U.S. air quality standards

- the possibility of computing statistics between measured and computed data, even when these data are not coupled in time and/or in space
- the identification of reducible errors and inherent uncertainties
- the recommendations to decision makers to educate themselves and accept the challenge of decision making with quantified uncertainty

The second point was and is the most controversial. What this means is that apples *can* be compared with oranges, for certain purposes. In fact, a model forecast of the maximum concentration impact c_A at location A at time t_1 can be compared with the measurement of maximum concentration impact c_B at location B at a time t_2, and if the two values are close, we are allowed to conclude that, for practical applications, the model can be considered a "good predictor" of the maximum impact. Scientifically speaking, this is not true; if A is distant from B and t_1 is much different from t_2, the model clearly does not work properly and the similarity between c_A and c_B is only accidental. Scientifically speaking, models should not just predict well, but they should do it for the *right* reason. But for practical regulatory applications, the criterion of "decoupling" concentration data in space and time should not be seen as a complete scientific aberration.

The decoupling in space has some acceptable justifications. As illustrated in Figure 12–10, sometimes plume models work well, but their performance can be spoiled by small (and quite common) errors in the measurement or the estimate of the wind direction. The decoupling in time, however, is hard to swallow.

Several recent studies have continued to investigate the problem of statistically evaluating the performance of air quality models. Interesting new methods were proposed at the DOE Model Validation Workshop, October 23–26, 1984, Charleston, South Carolina, and by Alcamo and Bartnicki (1987) and Hanna (1988). Major operational evaluations of air quality models have been sponsored by EPRI (e.g., Reynolds et al., 1984; Ruff et al., 1984; Moore et al., 1985; and Reynolds et al., 1985).

Some agreements on performance evaluation seem to be well accepted today. Terminology, at least, is more clear. Model *calibration* is the adjustment of empirical model constants, within their physical bounds, to optimize agreement with observations. If properly done, calibration is important and acceptable and should not be referred to as "fudging" or "massaging" results. Model *validity* is the "theoretical" ability of the model, with error-free model inputs. Therefore, a model can be validated against a theory or another model, but not against

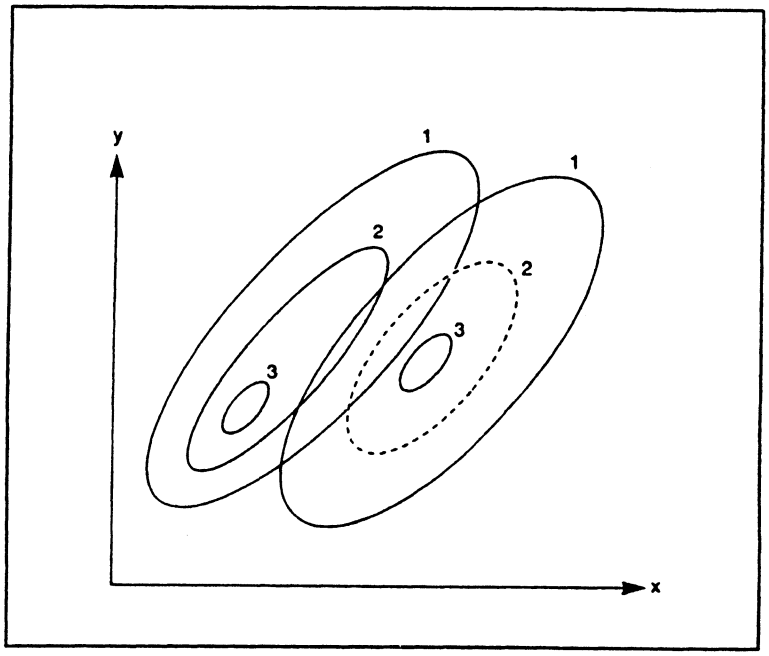

Figure 12–10. Illustration of displacement of observed and predicted ground-level concentration patterns. Isopleths represent points with the same concentration. The point-by-point correlation is poor, but the patterns are clearly similar (from Hanna, 1988). [Reprinted with permission from the Air Pollution Control Association.]

measurements. Model *evaluation* is the quantification of the performance of the model in real cases with real data. Model *verification* is the successful validation and/or evaluation of the model.

For practical applications, several statistical parameters can be used to evaluate pairs of predicted/observed concentrations. Among them

- The bias, i.e., the average difference of predicted minus observed values
- The gross error, i.e., the average of the absolute differences
- The variance of the differences
- The correlation coefficient between predicted and observed values
- The regression line, which ideally should have slope one and intercept zero

- The normalized fractional bias *FB* (Irwin and Smith, 1984) where

$$FB = 2(\bar{c}_p - \bar{c}_o)/(\bar{c}_p - \bar{c}_o) \quad (12\text{–}17)$$

  and $\bar{c}_p$, $\bar{c}_o$ are the predicted and observed average concentrations, respectively. *FB* varies between –2 and 2 with an optimum value of zero

- The normalized mean square error *NMSE* (Hanna and Heinold, 1985), where

$$NMSE = \overline{(c_p - c_o)^2}/(\bar{c}_p\, \bar{c}_o) \quad (12\text{–}18)$$

  where c_p and c_o are the single concentration values

- Skill scores (e.g., Murphy, 1988)
- Frequency distribution analysis of the differences
- Autocorrelation and spectral analysis of the differences. Often, repetitive or physically meaningful patterns in the differences can be identified and removed, thus improving the practical performance of the model (e.g., a daily cycle in the average difference may indicate emission input errors and can be empirically removed to maximize model performance)

Often data are insufficient for reliable statistical analysis. In this case, resampling procedures, such as "bootstrap" and "jackknife" techniques can be used to generate new "synthetic" data sets from the original data using an empirical set of rules (e.g., Heidam, 1987; Hanna, 1987). Finally, we must emphasize the powerful use of graphical methods for performance evaluation. In many cases, qualitative observations of multiple time plots, isopleths, cumulative frequency distributions, may say more than a thousand skill scores.

12.6 INTERPOLATION METHODS AND GRAPHIC TECHNIQUES

Several techniques that can be labeled as interpolation methods and graphic techniques will be discussed in this section. They are:

- Kriging
- pattern recognition
- cluster analysis
- fractals

12.6.1 Kriging

The Kriging technique was originally developed by Matheron (1971). It is an interpolation technique that possesses three major advantages with respect to other interpolation methods (Venkatram, 1988b):

1. its interpolations are made with weights that do not depend upon data values
2. it provides an estimate of the interpolation error
3. it is an exact interpolation since the interpolation at any observation point is the observation itself

In mathematical notation, observations $z(\mathbf{x}_j)$ of the variable z at locations $\mathbf{x}_j$ allow a Kriging interpolation of $z(\mathbf{x})$ at any point $\mathbf{x}$. Simple Kriging is done by assuming that

$$z(\mathbf{x}) = m + \epsilon(\mathbf{x}) \tag{12-19}$$

where m is a fixed component and ϵ is a stochastic component. Then, the Kriging estimate z'_k of $z(\mathbf{x}_k)$ at a generic point $\mathbf{x}_k$ is assumed to be a linear combination of the observations $z_j = z(\mathbf{x}_j)$, i.e.,

$$z'_k = \sum_j \lambda_j z_j \tag{12-20}$$

where the λ_j are independent of z_j and are calculated by variational calculus, imposing the condition that the ensemble average variance of z'_k be a minimum. This condition allows the calculation of the λ_j terms and the Lagrangian multiplier μ.

The variance of the interpolation error is computed by

$$< (z'_k - z_k)^2 > = \sum_j \lambda_j \gamma_{jk} + \mu \tag{12-21}$$

where the brackets < > indicate ensemble averaging and the semi-variogram γ_{jk} is defined by

$$\gamma_{jk} = < (z_j - z_k)^2 > /2 \tag{12-22}$$

and quantifies the effects of the stochastic term ϵ on the difference between $z(\mathbf{x}_j)$ and $z(\mathbf{x}_k)$. The term γ_{jk} cannot be calculated from observations and its correct determination is the major challenge in the application of the Kriging

technique. Kriging is good only if the assumed model for γ_{jk}, i.e., for the spatial relationships among measurements, is good. Barnes (1980) provides a few choices for γ_{jk}. A common, simplifying assumption is often made by assuming that γ_{jk} depends only upon the distance $|\mathbf{x}_j - \mathbf{x}_k|$.

The Kriging technique has recently been applied to environmental problems. Venkatram (1988b) used it with annual averages of sulfur wet deposition in the eastern United States, and Eynon (1988) applied Kriging to perform a statistical analysis of chemical measurements in about 10,000 precipitation samples collected during the period 1979 through 1983 in the eastern United States. Results seem encouraging, even though Fedorov (1989) claimed that other estimators, such as the generalized least squares (GLS) and the moving least squares (MLS) can successfully compete with the Kriging technique.

An example of Kriging is given in Figures 12–11 through 12–13. Figure 12–11 shows locations and values of annual wet deposition sulfur measured in 1980 in the eastern United States. Figure 12–12 illustrates the simple Kriging applied to these data, while Figure 12–13 shows a more realistic interpolation in which the pattern is estimated by a simple statistical long range model. Clearly, interpolation features improve when some deterministic information is added.

12.6.2 Pattern Recognition

Pattern recognition techniques have been applied to a large number of fields. These techniques can categorize sets of observations, by graphical methods, and perform forecasting. The theory of pattern recognition is found in Nilsson (1965), Arkadev and Braverman (1967), Fu (1968; 1974) and Fukanaga (1976).

Pattern recognition methods have been applied in atmospheric studies for air pollution control (Tauber, 1978), to characterize local sources (Edgerton and Holdren, 1987), and to automatically compute the mixing height from LIDAR measurements (Endlich et al., 1979).

12.6.3 Cluster Analysis

Methods of hierarchical cluster analysis are frequently applied in many research fields. A clear and detailed introduction to cluster analysis is given by Romesburg (1984). This method covers a variety of techniques that can be used to find out which objects in a set are similar. Cluster analysis is useful for classification purposes, even though it is used for several other purposes. Cluster analysis techniques have been applied, for example, to identify sources of

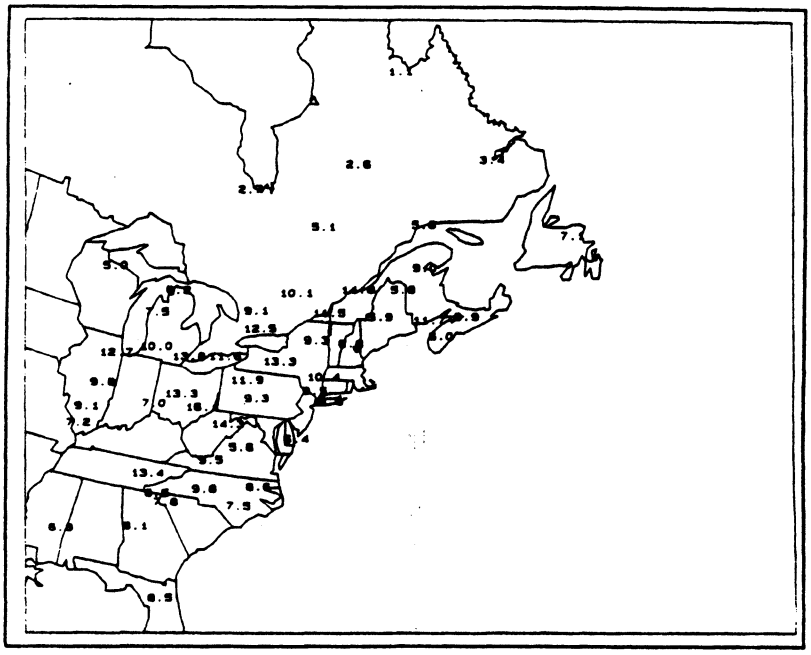

Figure 12–11. Locations and values of annual wet deposition of sulfur measured during 1980. Units are kg ha^{-1} (from Venkatram, 1988). [Reprinted with permission from Pergamon Press.]

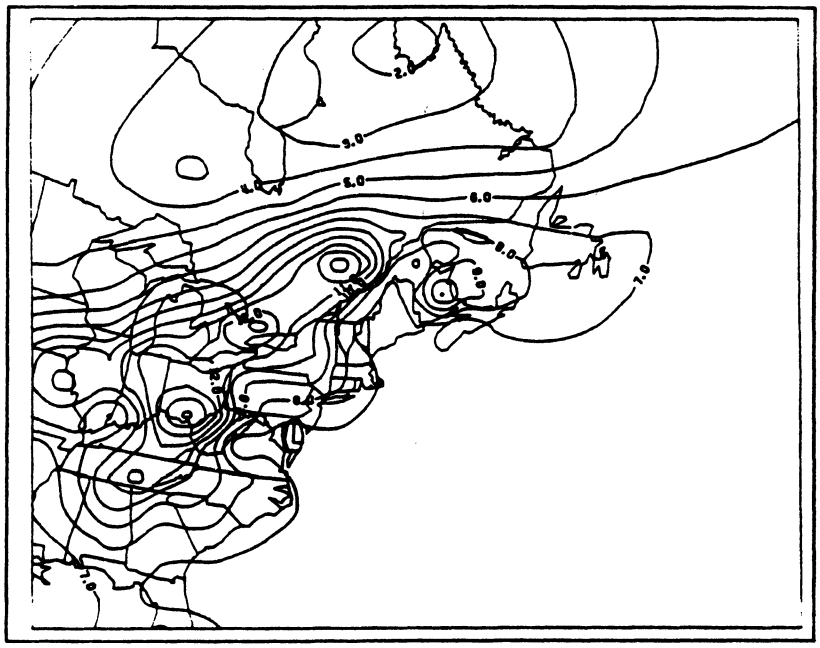

Figure 12–12. Pattern of annual sulfur wet deposition derived by applying simple Kriging to observations shown in Figure 12–11 (from Venkatram, 1988). [Reprinted with permission from Pergamon Press.]

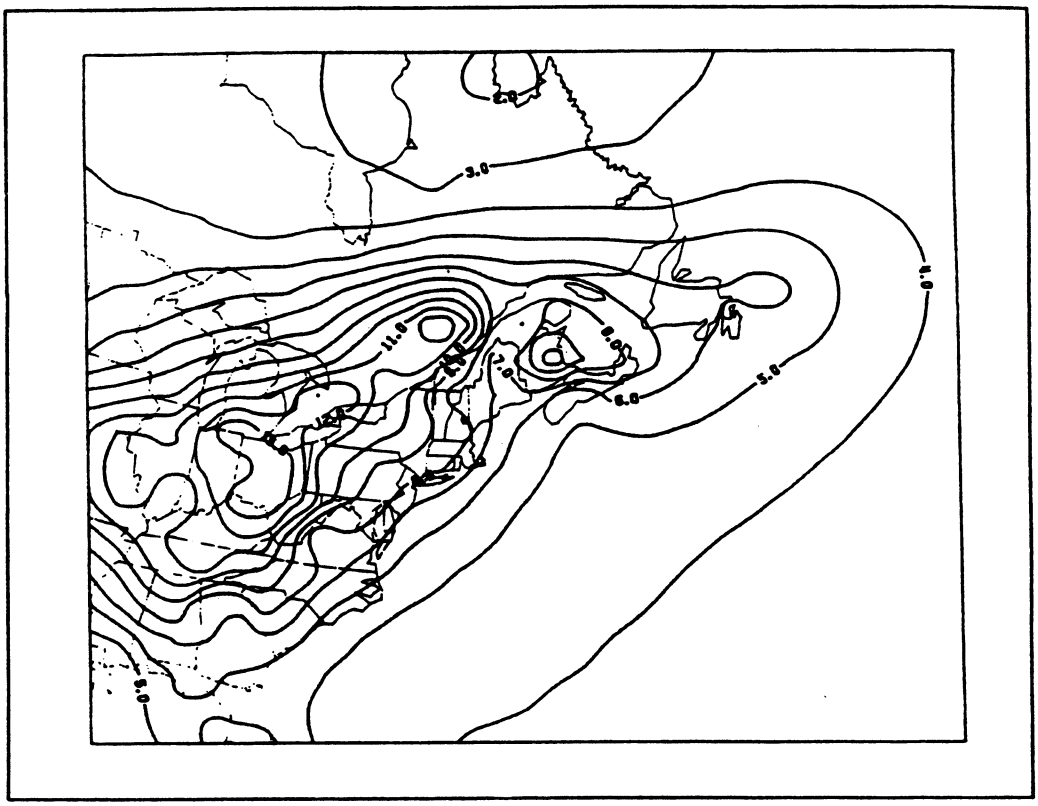

Figure 12–13. Same as Figure 12–12, except that pattern is estimated from statistical long–range transport model (from Venkatram, 1988). [Reprinted with permission from Pergamon Press.]

particulate matter (Gomez and Martin, 1987) and to perform source apportionment of atmospheric aerosols (Van Borm and Adams, 1988).

12.6.4 Fractals

Fractals are geometric shapes with roughness characteristics that are qualitatively similar at all scales. Techniques based on the fractals concept were introduced by Mandelbrot (1975) and have become very popular. In fact, many lines and surfaces in nature are well depicted by fractals, which, therefore, allow the production of synthetic, but realistic looking, landscapes. Fractals are useful for qualitative reproduction of natural phenomena, such as turbulent motion, and for image compression techniques. Their most interesting application is probably in conjunction with chaos theories, which were briefly discussed at the beginning of this chapter.

In atmospheric studies, fractals allow the depiction of turbulent eddies, the reproduction of the similarity theory, and the numerical simulation of de-

tailed and complex characteristics of fluid flows. An overview of the application of fractals to atmospheric sciences is presented by Ludwig (1989).

12.7 OPTIMIZATION METHODS

Optimization needs are often present in air quality studies. For example, emission reductions should always be optimized, to allow the most effective reductions within the allowable budgets. The most common application of optimization methods occurs in the design of a monitoring network, where nonlinear programming techniques are used to determine the number and disposition of ambient air quality stations. This determination can be done under different constraints (e.g., a primary purpose of a network might be the assessment of maximum ground-level concentration impact for compliance with air quality standards). A vast literature exists in this field. Examples of applications are given by Seinfeld (1972), Noll et al. (1977), Nakamori and Sawaragi (1984), Liu et al. (1986), and Langstaff et al., (1987).

REFERENCES

Alcamo, J., and J. Bartnicki (1987): A framework for error analysis of a long-range transport model with emphasis on parameter uncertainty. *Atmos. Environ.*, **21**(10):2121–2131.

Arkadev, A.G., and E.M. Braverman (1967): *Teaching Computers to Recognize Patterns.* London: Academic Press.

Bacci, P., P. Bolzern, and G. Fronza (1981): A stochastic predictor of air pollution based on short-term meteorological forecasts. *J. Appl. Meteor.*, **20**(2):121–129.

Bankoff, S.G., and E.L. Hanzevack (1975): The adaptive filtering transport model for prediction and control of pollutant concentration in an urban airshed. *Atmos. Environ.*, **18**:701–712.

Barnes, M.G. (1980): The use of Kriging for estimating the spatial distribution of radionuclides and other spatial phenomena. Tran-Stat, Battelle Memorial Institute, Pacific Northwest Laboratories., Richland, Washington.

Barone, J.B., T.A. Cahill, R.A. Eldred, R.G. Flocchini, D.J. Shadoan, and T.M. Dietz (1978): A multivariate statistical analysis of visibility degradation at four California cities. *Atmos. Environ.*, **12**:2213–2221.

Berge, P., Y. Pomeau, and C. Vidal (1984): *Order Within Chaos.* New York: John Wiley.

Bornstein, R.D., and S.F. Anderson (1979): A survey of statistical techniques used in validation studies of air pollution prediction models. Technical Report No. 23, Stanford University, Stanford, California.

Box, G.E., and G.M. Jenkins (1976): *Time Series Analysis, Forecasting and Control.* San Francisco: Holden-Day.

Buishand, T.A., G.T. Kempen, A.J. Frantzen, H.F. Reijnders, and A.J. van den Eshof (1988): Trend and seasonal variations of precipitation chemistry data in the Netherlands. *Atmos. Environ.*, **22**(2):339–348.

Cats, G.J., and A.A. Holtslag (1980): Prediction of air pollution frequency distribution. Part I: The lognormal model. *Atmos. Environ.*, **14**:255–258.

Chock, D.P., and P.S. Sluchak (1986): Estimating extreme values of air quality data using different fitted distributions. *Atmos. Environ.*, **20**(5):989–993.

Chock, D.P., T.R. Terrel, and S.B. Levitt (1975): Time-series analysis of Riverside, California air quality data. *Atmos. Environ.*, **9**:978–989.

Chow, J.C., P.W. Severance, and J.D. Spengler (1985): A composite application of source and receptor models to fine particle concentrations in Portage, Wisconsin. *Proceedings,* 78th Annual APCA Meeting. Detroit, Michigan, June.

Cooper, J.A. (1983): Receptor model approach to source apportionment of acid rain precursors. *Proceedings,* VIth World Congress on Air Quality, Paris, France, May 16–20, pp. 223–229.

Desalu, A.A., L.A. Gould, and F.C. Schweppe (1974): Dynamic estimation of air pollution. *IEEE Transactions on Automatic Control AC-19,* pp. 904–910.

Drufuca, G., and M. Giugliano (1978): Relationship between maximum SO_2 concentration, averaging time and average concentration in an urban area. *Atmos. Environ.*, **12**:1901–1905.

Edgerton, S.A., and M.W. Holdren (1987): Use of pattern recognition techniques to characterize local sources of toxic organics in the atmosphere. *Environ. Sci. Technol.*, **21**(11):1102–1107.

Endlich, R.M., F.L. Ludwig, and E.E. Uthe (1979): An automatic method for determining the mixing depth from LIDAR observations. *Atmos. Environ.*, **13**:1051–1056.

Eynon, B.P. (1988): Statistical analysis of precipitation chemistry measurements over the eastern United States. Part II: Kriging analysis of regional patterns and trends. *J. Appl. Meteor.*, **27**:1334–1343.

Fedorov, V.V. (1989): Kriging and other estimators of spatial field characteristics (with special reference to environmental studies). *Atmos. Environ.*, **23**(1):175–184.

Finzi, G., and G. Tebaldi (1982): A mathematical model for air pollution forecast and alarm in an urban area. *Atmos. Environ.*, **16**(9):2055–2059.

Fox, D.C. (1981): Judging air quality model performance. *J. Climate and Appl. Meteor.*, **62**:599–609.

Fox, D.C. (1984): Uncertainty in air quality modeling. *J. Climate and Appl. Meteor.*, **65**:27–36.

Fronza, G., A. Spirito, and A. Tonielli (1979): Real-time forecast of air pollution episodes in the Venetian region. Part 2: The Kalman predictor. *Appl. Math. Model.*, **3**:409–415

Fu, K.S. (1968): *Sequential Methods in Pattern Recognition.* New York: Academic Press.

Fu, K.S. (1974): *Syntactic Methods in Pattern Recognition.* New York: Academic Press.

Fukunaga, K. (1972): *Introduction to Statistical Pattern Recognition.* New York: Academic Press.

Georgopoulos, P.G., and J.H Seinfeld (1982): Statistical distributions of air pollutant concentrations. *Environ. Sci. & Technol.*, **16**:401A–415A.

Gilbert, R.O. (1987): *Statistical Methods for Environmental Pollution Monitoring.* New York: Van Nostrand Reinhold.

Gomez, M.L., and M.C. Martin (1987): Applications of cluster analysis to identify sources for airborne particles. *Atmos. Environ.*, **21**(7):1521–1527.

Gordon, G.E. (1988): Receptor models. *Environ. Sci. & Tech.*, **22**(10):1132.

Grebogi, C., E. Ott, and J.A. Yorke (1987): Chaos, strange attractors, and fractal basin boundaries in nonlinear dynamics. *Science*, **238**:632–637.

Hanna, S.R., and D.W. Heinold (1985): Development and application of a simple method for evaluating air quality models. American Petroleum Institute Publication 4409, Washington, D.C.

Hanna, S.R. (1988): Air quality model evaluation and uncertainty. *JAPCA*, **38**:406–412.

Hanna, S.R. (1989): Confidence limits for air quality model evaluations, as estimated by bootstrap and jackknife resampling methods. *Atmos. Environ.*, **23**(6):1385–1389.

Heidam, N.Z. (1987): Bootstrap estimates of factor model variability. *Atmos. Environ.*, **21**(5):1203–1217.

Henry, R.C. (1987): Current factor analysis receptor models are ill-posed. *Atmos. Environ.*, **21**(8):1815–1820.

Henry, R.C., and G.M. Hidy (1979): Multivariate analysis of particulate sulfate and other air quality variables by principal components. Part I: Annual data from Los Angeles and New York. *Atmos. Environ.*, **13**:1581–1596.

Henry, R.C., C.W. Lewis, P.K. Hopke, and H.I. Williamson (1984): Review of receptor model fundamentals. *Atmos. Environ.*, **18**:1507–1515.

Hino, M. (1974): Prediction of atmospheric pollution by Kalman filtering. *Proceedings*, Symposium on Modeling for Prediction and Control of Air Pollution.

Hopke, P.K. (1988): Target transformation factor analysis as an aerosol mass apportionment method: A review and sensitivity study. *Atmos. Environ.*, **22**(9):1777–1792.

Horowitz, J., and S. Barakat (1979): Statistical analysis of the maximum concentration of an air pollutant: Effects of autocorrelation and non-stationarity. *Atmos. Environ.*, **13**:811–818.

Irwin, J.S., and M.E. Smith (1984): Potentially useful additions to the rural model performance evaluation. *J. Climate and Appl. Meteor.*, **65**:559.

Irwin, J.S., S.T. Rao, W.B. Peterson, and D.B. Turner (1987): Relating error bounds for maximum concentration estimates to diffusion meteorology uncertainty. *Atmos. Environ.*, **21**(9):1927–1937.

Jazwinski, A.H. (1970): *Stochastic Processes and Filtering Theory*. New York: Academic Press.

Jenkins, G.M., and D.G. Watts (1968): *Spectral Analysis and Its Applications*. San Francisco: Holden-Day.

Kahn, H.D. (1973): Distribution of Air Pollutants. *JAPCA*, **23**:973.

Kalman, R.E. (1960): A new approach to linear filtering and prediction problems. *J. Basic Eng.*, pp. 35–108

Langstaff, J.E., C. Seigneur, M.-K. Liu, J.V. Behar, and J.L. McElroy (1987): Design of an optimum air monitoring network for exposure assessments. *Atmos. Environ.*, **21**(6):1393–1410.

Larsen, R.I. (1971): EPA Publication No. AP-89, Research Triangle Park, North Carolina.

Little, R.J., and D.B. Rubin (1987): *Statistical Analysis with Missing Data*. New York: John Wiley.

Lin, G.Y., (1982): Oxidant prediction by discriminant analysis in the south coast air basin of California. *Atmos. Environ.*, **16**(1):135–143.

Longhetto, A., Ed. (1980): *Atmospheric Planetary Boundary Layer Physics*. New York: Elsevier.

Liu, M.–K., J. Arvin, R.I. Pollack, J.V. Behar, and J.L. McElroy (1986): Methodology for designing air quality monitoring networks. I: Theoretical aspects. *Environ. Monitoring and Assessment*, **6**:1–11.

Ludwig, F.L. (1989): Atmospheric fractals — A review. *Environ. Software*, **4**(1):9–16.

Malm, W., K. Gebhart, D.A. Latimer, T.A. Cahill, R. Eldred, R.A. Pielke, R. Stocker, J.G. Watson (1989): Winter Haze Intensive Tracer Experiment. National Park Service draft final report, December.

Mandelbrot, B.B. (1975): On the geometry of homogeneous turbulence with stress on the fractal dimension of the iso–surfaces of scalars. *J. Fluid Mech.*, **72**:401–416.

Marani, A., 1. Lavagnini, C. Buttazzoni (1986): Statistical study of air pollutant concentrations via generalized gamma distributions. *JAPCA*, **36**:1250–1254.

Matheron, G. (1971): The theory of regionalized variables and its applications. Ecole des Mines de Paris, Fontainbleau No. 5.

Melli, P., P. Bolzern, G. Fronza, and A. Spirito (1981): Real–time control of sulphur dioxide emissions from an industrial area. *Atmos. Environ.*, **15**: 653–666.

Moore, G.E., M.–K. Liu, and R.J. Londergan (1985): Diagnostic validation of Gaussian and first–order closure plume models at a moderately complex terrain site. Systems Applications, Inc., Final Report EA–3760, San Rafael, California.

Murphy, A.H. (1988): Skill scores based on the mean square error and their relationships to the correlation coefficient. *J. Climate and Appl. Meteor.*, **116**:2417–2424.

Murray, L.C., and R.J. Farber (1982): Time series analysis of an historical visibility data base. *Atmos. Environ.*, **16**:2299–2308.

Nakamori, Y., and Y. Sawaragi (1984): Interactive design of urban level air quality monitoring network. *Atmos. Environ.*, **18**(4):793–799.

Nilsson, N.J. (1965): *Learning Machines*. McGraw–Hill.

Noll, K.E., T.L. Miller, J.E. Norco, and R.K. Raufer (1977): An objective air monitoring site selection methodology for large point sources. *Atmos. Environ.*, **11**:1051–1059.

Petersen, J.T. (1970): Distribution of sulfur dioxide over metropolitan St. Louis, as described by empirical eigenvectors, and its relation to meteorological parameters. *Atmos. Environ.*, **4**:501–518.

Plackett, R.L. (1950): *Biometrika*, **37**:149.

Reynolds, S.D., C. Seigneur, T.E. Stoeckenius, G.E. Moore, R.G. Johnson, and R.J. Londergan (1984): Operational validation of Gaussian plume models at a plains site. Systems Applications, Inc., Final Report EA–3076, San Rafael, California.

Reynolds, S.D., T.C. Myers, J.E. Langstaff, M.-K. Liu, G.E. Moore, and R.E. Morris(1985): Operational validation of Gaussian and first-order closure plume models at a moderately complex terrain site. Systems Applications, Inc., Final Report EA-3759, San Rafael, California.

Romesburg, H.C. (1984): *Cluster Analysis for Researchers*. Belmont, California: Lifetime Learning Publications.

Roberts, E.M. (1979): Review of statistics of extreme values with applications to air quality data. Part I: Review. *JAPCA*, **29**(6):632–637.

Roberts, E.M. (1979): Review of statistics of extreme values with applications to air quality data. Part II: Applications. *JAPCA*, **29**(7):733–740.

Roy, R., and I. Pellerin (1982): On long term air quality trends and intervention analysis. *Atmos. Environ.*, **16**:161–169.

Ruff, R.E., K.C. Nitz, F.L. Ludwig, C.M. Bhumralkar, J.D. Shannon, C.M. Sheih, I.Y. Lee, R. Kumar, and D.J. McNaughton (1984): Regional air quality model assessment and evaluation. SRI International Final Report EA-3671, Menlo Park, California.

Sage, A.P., and I.L. Melsa (1971): *Estimation Theory with Applications to Communications and Control*. New York: McGraw-Hill.

Sawaragi, Y., T. Soeda, T. Yoshimura, S. Oh, Y. Chujo, and H. Ishihara (1976): The predictions of air pollution levels by nonphysical models based on Kalman filtering method. *J. Dynamic Sys., Measurement and Control*, **98**(12):375–386.

Scheff, P.A., R.A. Wadden, and R.I. Allen (1984): Development and validation of a chemical element mass balance for Chicago. *Environ. Sci. Technol.*, **18**:923–931.

Seinfeld, J.H. (1972): Optimal location of pollutant monitoring stations in an airshed. *Atmos. Environ.*, **6**:847–858.

Seinfeld, J.H. (1986): *Atmospheric Chemistry and Physics of Air Pollution*. New York: John Wiley.

Simpson, R.W., and A.P. Layton (1983): Forecasting peak ozone levels. *Atmos. Environ.*, **17**:1649–1654.

Surman, P.G., J. Bodero, and R.W. Simpson (1987): The prediction of the numbers of violations of standards and the frequency of air pollution episodes using extreme value theory. *Atmos. Environ.*, **21**(8):1843–1848.

Tauber, S. (1978): Pattern recognition methods in air pollution control. *Atmos. Environ.*, **12**:2377–2382.

Tiao, G.C., G.E. Box, and W.J. Hamming (1975): Analysis of Los Angeles photochemical smog data: A statistical overview. *JAPCA*, **25**:260–268.

Tilley, T., and G.A. McBean (1973): An application of spectrum analysis to synoptic-pollution data. *Atmos. Environ.*, **7**:793–801.

Trivikrama, S.R., P.I. Samson, and A.R. Pedadda (1976): Spectral analysis approach to the dynamics of air pollutants. *Atmos. Environ.*, **10**:375–379.

Tsukatami, T., and K. Shigemitsu (1980): *Atmos. Environ.*, **14**:245.

Van Borm, W.A., and F.C. Adams (1988): Cluster analysis of electron microprobe analysis data of individual particles for source apportionment of air particulate matter. *Atmos. Environ.*, **22**(10):2297–2307.

Venkatram, A. (1982): A framework for evaluating air quality models. *Boundary-Layer Meteor.*, **24**:371–385.

Venkatram, A. (1983): Uncertainty in predictions from air quality models. *Boundary-Layer Meteor.*, **27**:185–196.

Venkatram, A. (1988a): Inherent uncertainty in air quality modeling. *Atmos. Environ.*, **22**(6):1221–1227.

Venkatram, A. (1988b): On the use of Kriging in the spatial analysis of acid precipitation data. *Atmos. Environ.*, **22**(9):1963–1979.

Watson, J.G. (1984): Overview of receptor model principles. *JAPCA*, **34**:619–623.

Wiener, N. (1949): *The Extrapolation, Interpolation and Smoothing of Stationary Time Series.* New York: John Wiley.

Williams, P.C. (1984): Data handling, simultaneity, and rare events. *JAPCA*, **34**:945–951.

Young, P. (1974): Recursive approaches to time series analysis. In *The Inst. of Mathematics and its Application*, pp. 209–224.

Zannetti, P. (1978): Short-term real-time control of air pollution episodes in Venice. *Proceedings*, 71st Annual APCA Meeting, Houston, Texas, June.

Zannetti, P., G. Finzi, G. Fronza, and S. Rinaldi (1978): Time series analysis of Venice air quality data. IFAC Symposium on Environmental Systems, Planning, Design, and Control, August 1–5, 1977, Kyoto, Japan.

Zannetti, P., and P. Switzer (1979a): The Kalman filtering method and its application to air pollution episode forecasting. Paper presented at the APCA Specialty Conference on Quality Assurance in Air Pollution Measurement, New Orleans, Louisiana, March.

Zannetti, P., and P. Switzer (1979b): Some problems of validation and testing of numerical air pollution models. *Proceedings*, Fourth Amer. Meteor. Soc. Symp. on Turbulence, Diffusion, and Air Pollution, Reno, Nevada. January, pp. 405–410.

Zinsmeister, A.R., and T.C. Redman (1980): A time series analysis of aerosol composition measurements. *Atmos. Environ.*, **14**:201–215.