SURVEY OF STATISTICAL MODELS FOR OXIDANT AIR QUALITY PREDICTION

Leik N. Myrabo and Kent R. Wilson, Department of Chemistry and Energy Center, University of California, San Diego; and
John C. Trijonis, Transportation and Environmental Operations, TRW, Inc.

Oxidant air pollution is the result of a complex series of chemical reactions stemming from reactive hydrocarbon (RHC) and nitrogen oxide (NO_x) emissions. Transportation planning and other policies affect oxidant air quality by altering the spatial and temporal distribution of RHC and NO_x emissions. To evaluate the oxidant air quality impact of various policies requires that the relation between ambient oxidant levels and precursor emission levels be known. Attempts to determine this relation have followed 3 general approaches: smog chamber modeling, mathematical simulation of physical and chemical processes, and statistical-empirical models. Because of drawbacks in each of these approaches and because of the complexity of the photochemical smog process, considerable uncertainty still surrounds the relation between ambient oxidant concentrations and precursor emission levels.

Smog chamber models are based on the results of laboratory experiments wherein mixtures of RHC and NO_x are irradiated by sunlight to produce oxidant. By altering the amounts of RHC and NO_x in these experiments, one can gain information on the dependence of oxidant on precursor levels. Smog chambers have provided much of the basic understanding we have of photochemical air pollution. However, their use in air quality planning has uncertainties because of questions concerning how representative smog chambers are of real atmospheric conditions. In smog chambers, simulating real meteorology is difficult, and the hydrocarbon mix may differ from that in the real atmosphere. Also, wall effects in smog chambers produce effects that are absent in real atmospheres and conversely do not reproduce real effects of terrestrial surfaces. In addition, smog chamber results do not simulate the spatial distribution of emissions in a region.

The second approach, deterministic or mechanistic modeling of chemical and meteorological processes, involves mathematical simulation of emission patterns, diffusion and mixing, transport, and atmospheric chemistry. This deterministic type of model could in principle be an ideal planning tool since it can explicitly account for changes in the spatial and temporal distribution of emissions as well as changes in overall emission levels and meteorological variability. Much work has been done in developing and testing chemical-meteorological models, but serious questions still exist concerning the accuracy of such models in predicting the impact of future emission changes. These questions stem from the lack of sufficient understanding of turbulent mixing and diffusion, from uncertain knowledge of the reaction rates for atmospheric reactions, and from inadequacies in the available meteorological data. The application of chemical-meteorological models may also be limited by the expense associated with the extensive data base and the computer time required to run them.

The third approach, statistical-empirical modeling, centers on the use of actual atmospheric monitoring data. The relation between oxidant and precursors or oxidant and meteorological variables is derived from aerometric data by statistical analysis and simplifying physical assumptions. Empirical models benefit from the basic advantage that the influences of all the complex atmospheric processes are inherent in the aerometric data base, which forms the foundation of these models. Empirical models

are also relatively inexpensive to develop and simple to apply. A disadvantage of the empirical approach is inaccuracy or sparsity or both of the required aerometric data. Since the models are developed from examining a particular range of real conditions, there is always a danger in extending the conclusions of such a model beyond the range of the data on which it was calibrated. Most (but not all) empirical models, for example, involve the assumption that the spatial distribution of emissions remains fixed and are not geared toward the assessment of the effects of alternative source sitings.

The purpose of this paper is to review the present state of the third approach to oxidant modeling: the statistical-empirical approach. In the interests of brevity, we will restrict our scope mainly to statistical-empirical models that have been used in a predictive capacity, not just to draw correlations. Thus, we will neglect a large part of the important statistical work that has been done on the correlations among meteorological and pollutant variables, for example, in the analysis of past air quality trends. In addition, we will restrict our review only to those models that have been applied to oxidant prediction, even though the methodology used to predict other pollutants can often be applied to oxidant as well.

Two classes of models for oxidant levels are reviewed; both have quite different aims. The first is short-term forecasting, with the goal of episode control. The basic relation is the variation of oxidant level with meteorological parameters. The hope is to predict episodes with enough lead time (hours to days) so that temporary emission controls can be applied to lessen the severity of the episode, or at least to provide health warnings. The second class of models has the goal of predicting the long-term effect of control strategies. The basic relation is that between emission levels and oxidant level. The hope is to evaluate and optimize long-term control strategies before implementation.

Before proceeding to survey statistical-empirical oxidant models, we should note that all models, including smog chamber and chemical-meteorological models, are in part statistical and empirical. Smog chamber and chemical-meteorological models also in reality depend on atmospheric data for calibration before they are used for predictive purposes. As defined here, statistical-empirical models are those that rely on aerometric data to determine the actual form of the oxidant-meteorological or the oxidant-precursor dependence.

SHORT-TERM PREDICTION (EPISODE CONTROL)

If one could predict far enough in advance that an episode of high air pollution level were going to occur, one could then try to reduce the seriousness of the episode by short-term control of emissions, for example, from traffic and particular industries, during the episode period. In addition, health warnings could be issued to advise sensitive individuals of when and where they should take particular precautions. Thus, based on measured pollutant and meteorological variables, several modeling efforts have used statistical-empirical techniques to try to predict air pollution levels hours or days in advance. For example, one might derive a relation statistically linking tomorrow's hourly maximum oxidant level to today's oxidant level, wind speed, inversion height, and temperature.

Three basic types of statistical approaches have been used to try to make predictions on this short-term scale: time series, multiple regression, and pattern recognition. For an introduction to the field (really a preview rather than a review, since this area until now has been characterized more by discussion than by in-depth studies), the reader should review the proceedings of the Conference on Forecasting Air Pollution (6) and, perhaps for background, the proceedings of the Symposium on Statistical Aspects of Air Quality (23).

Time Series

Pollutant monitoring data can be considered as a sequence of observations, a time

series of pollutant measurements $z_1, z_2, \ldots, z_n$ at times 1, 2, ..., n. One wishes, knowing n values, to predict the next ℓ values. One can use just the previous values of the time series itself to predict the future values, a univariate time series (e.g., past oxidant measurements to predict future oxidant measurements), or one can use a multivariate time series and in addition consider other predictor variables such as meteorological measurements.

As explained by Box and Jenkins (3), one can set up a general form of linear stochastic model of a univariate time series as

$$w_t - \varphi_1 w_{t-1} - \ldots - \varphi_p w_{t-p} = a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q} \qquad (1)$$

where $w_t = \nabla^d z_t$, ∇ being the backward difference operator; and $\nabla z_t = z_{t-1}$, which may be repeated d times. Pollutant time series that contain daily, weekly, or yearly seasonality may be handled by using ∇_s, where $\nabla_s z_t = z_t - z_{t-s}$; for example, for daily measurements with a weekly cycle, s = 7. Fitting a model to the time series involves deciding how many times to difference with ∇ (to achieve a stationary time series), choosing how many parameters $\varphi_1, \ldots, \varphi_p$ and $\theta_1, \ldots, \theta_q$ to include, and then adjusting them to obtain optimal forecasts. Diagnostic checking of the model involves examining the residuals $a_t, a_{t-1}, \ldots, a_{t-q}$ that relate w_t to past values $w_{t-1}, \ldots, w_{t-p}$. These residuals represent all factors other than the past values w_{t-1}, w_{t-2}, and so on that actually determine the values of the time series. If the model fits the time series well, the residuals should behave nearly like white noise, being a series of normally distributed random values with mean zero. Since the expected value of the residuals is zero, forecasts of future values of the time series, i.e., future pollutant values, can be made by setting to zero the unknown residuals yet to come.

Work along these lines has been carried out by Box and Tiao and coworkers at the University of Wisconsin, by McCollister and Wilson at the University of California, San Diego (27), and by Chock and Terrell at General Motors Research Laboratories (10).

Some results of the work by McCollister and Wilson are shown in Figure 1 for univariate oxidant prediction both of daily instantaneous oxidant maxima and of hourly oxidant values. All forecasts are made by using data before 10 a.m. the day before the day being forecast. Persistence assumes the day being forecast will be the same as the previous days; LAAPCD is the prediction of the Los Angeles Air Pollution Control District from pollutant and meteorological data, and model is a Box-Jenkins univariate time series with 2-step ahead prediction. Such a univariate time series model that predicts future oxidant values solely on the basis of past oxidant values does a bit better than persistence (which assumes the future mimics the past exactly) and, surprisingly, even a bit better than predictions made by trained meteorologists using both meteorological and pollutant data. All 3 methods, however, have fairly large average errors, in the 35 to 50 percent range, perhaps too large for actual health warning or short-term emission control usage. McCollister and Wilson have also developed multivariate time series oxidant forecasts using multivariate time series of meteorological variables.

Chock and Terrell (10) have applied both univariate and multivariate time series techniques to weekly average daily maximum oxidant data. Figure 2 shows a comparison of their various time series studies as applied to weekly 1970 oxidant data. The results shown are for long-term prediction, using the previous years' oxidant data, and could be improved by using oxidant data up to the week to be predicted. Little predictive information was found in week-old meteorological data, as might be expected from weather forecasting experience (6). Thus, their multivariate predictions use future and not past meteorological data (radiation intensity, wind speed, and dry bulb temperature). Although the weekly time span used is probably too long for episode control or short-term health warnings, these same techniques demonstrated by Chock and Terrell could also be applied to daily or hourly data.

Work is in progress at the San Francisco Bay Area Air Pollution Control District on both univariate and multivariate time series techniques as applied to oxidant data.

Figure 1. Comparison of forecasts of daily instantaneous oxidant maxima for 1972 by 3 methods for Los Angeles County monitoring stations.

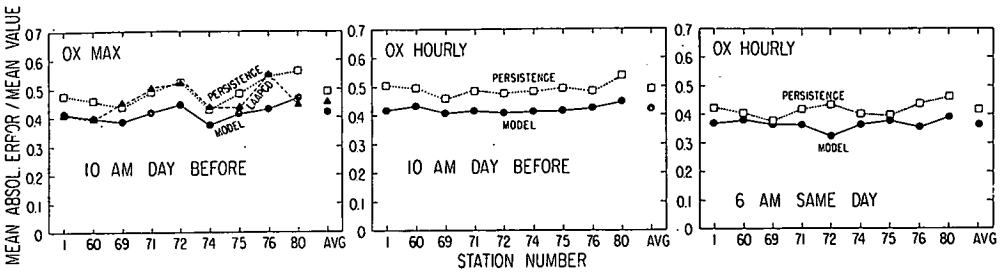
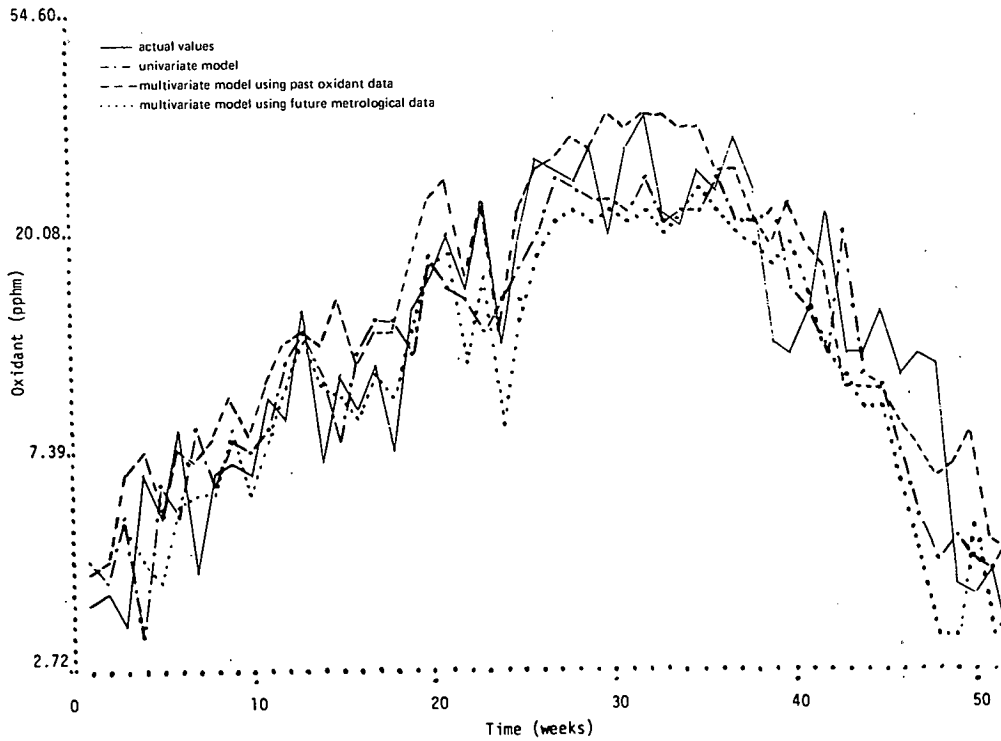


Figure 2. Various time series predictions of 1970 weekly average daily hourly maximum oxidant concentrations (log scale) at Riverside, California.



In addition, studies have been carried out there in which a canonical correlation method (19) is used, which is essentially a regression on the univariate oxidant time series and thus is somewhat related both to Box-Jenkins time series analysis and to multiple regression.

There are also other studies that have aims different from short-term forecasting. These studies, in which time series techniques have been applied to oxidant data, include work by Merz, Painter, and Ryason (29); Lee, Sarin, and Wang (24); Tiao, Box, and Hamming (36); and Box and Tiao (4, 5). Much of the mathematical approach used in these studies is also applicable to short-term oxidant forecasting.

Multiple Regression

Many different measured variables might be used as predictors to forecast oxidant values far enough ahead for short-term controls on emissions to be implemented or for health warnings to be issued: pollutant concentrations (CO, SO_2, RHC, NO, NO_2, oxidant), local meteorological variables (wind vectors, inversion heights, temperatures, solar radiation, relative humidity, precipitation, cloud cover, barometric pressure), and more distant meteorological variables (upper air data, surface meteorology outside the air basin). Multiple regression (15) has been an approach taken by several authors—the linear combination of these predictor variables X_i that best forecasts oxidant levels,

$$Ox = \alpha_0 + \Sigma_i \alpha_i X_i + \epsilon \qquad (2)$$

in which α_0, α_1, ... are constants to be discovered and ϵ is an error term.

Charles Bennett (2) used stepwise multiple regression to relate maximum hourly average oxidant at several Los Angeles area stations to (a) yesterday's maximum hourly oxidant at the given station, (b) highest hourly NO_2 concentration observed between 6 and 9 a.m. today at the downtown Los Angeles station, (c) the 24-hour height change at 500 mb (5 MPa) based on the 4 a.m. atmospheric soundings above Vandenberg Air Force Base, and (d) the 850-mb (85-MPa) temperature based on the 4 a.m. sounding above Los Angeles Airport. For 4 stations, the correlation coefficients between this afternoon's predicted and measured maximum hourly concentrations ranged from 0.69 to 0.77 during the June-September 1973 and 1974 periods. Stratifying the data by weekends and weekdays further improved the fit, and the tendency to underpredict higher values was reduced by weighting high oxidant days. A comparison of observed and predicted values (with the above refinements) is shown in Figure 3.

Chock and Terrell (10) applied multiple regression to predict weekly averages of maximum daily oxidant from concurrent weekly average weather parameters, first using a regression analysis to screen out the best subset of independent predictors, which were found to be radiation intensity, wind speed, and dry bulb temperature. (These weather parameters were then also used for their multivariate weekly time series analysis.) Regression in terms of logarithms of the variables was found to be useful. The coefficients of determination, R^2, the fraction of the total variation about the mean explained by the regression, ranged from 0.79 to 0.84 for various models with and without previous oxidant data as a predictor. These R^2 values are to be compared with the 1-week ahead forecast values obtained by Chock and Terrell of 0.71 for univariate weekly oxidant time series and 0.82 for multivariate weekly time series using concurrent meteorological data.

Tiao, Phadke, and Box (37) used a regression model on logarithmic data from Los Angeles to derive a forecasting relation for daily maximum hourly oxidant based on the previous day's oxidant value, month of the year, 4 a.m. NO_2 level, 4 a.m. inversion base height and its square, difference between the inversion breaking temperature and the 4 a.m. surface temperature, and the average 1 to 4 a.m. wind speed. The variance of the error was significantly reduced by introducing the early morning NO_2 and meteorological data.

Bruntz, Cleveland, Kleiner, and Warner (7) fitted a regression for New York data relating log of ozone plus a constant from 1 to 4 p.m. to logs of 4 predictor variables: morning average wind speed, solar radiation, maximum daily temperature, and mixing height. The meteorological and ozone data are in part concurrent. The correlation coefficient between predicted and fitted $\log_{10}$ (ozone + 5 ppb) is 0.84 and is little affected by omitting mixing height from the regression equation.

An interesting application of an approximate nonlinear regression, with aspects of pattern recognition, is the point classification system for ozone prediction developed by Zeldin and Thomas (39) of the San Bernardino County Air Pollution Control District in California. Six classification categories are defined: stability, 950-mb (95-MPa)

temperature, inversion base height, coast to desert pressure gradient, day of the week, and month of the year. Points are assigned separately to each of the 10 classes into which each category is divided, and the sum of the number of points over all categories is equated to the predicted peak ozone level. The model, once calibrated, is thus based entirely on meteorological predictors and not on previous ozone levels. For this reason, it can be used to correct monthly or yearly ozone data for the effect of meteorological variability, to allow an approximate removal of meteorological effects from the trends of pollution levels with changing emissions, and, it is hoped, to provide more reliable reflection of the real effects of long-term control strategies.

Pattern Recognition

There is no a priori reason to suppose that the relation of oxidant to meteorological and pollutant predictors should be best fitted by any particular mathematical scheme. More general methods for forecasting exist than either the usual time series or multiple regression techniques discussed above. Groups at Technology Service Corporation (6), at Environmental Research and Technology (6), and at the University of Washington and the University of California, San Diego, have done explanatory work in applying formal pattern recognition techniques (28) to air pollution; to our knowledge, no full treatment has yet been published. Several other studies, however, involve stratification and classification techniques that at least share something of the viewpoint of pattern recognition, and Pollack (32) has discussed some of the possibilities.

Two basic tasks are involved. First, out of the large set of possible predictors, a smaller number of significant features must be selected, the aim being to find those that singly or in combination can best be used to forecast future oxidant levels. These features may themselves be functions of several members of the original predictor set and may include the output of time series and multiple regression forecasts. Various formal techniques exist for such feature extraction and ranking (28); but, since the possible number of features is infinite, good judgment is also helpful. The object is to find a small set of features that contain most of the information of the larger predictor data set. Too many features degrade performance by introducing additional noise and by adding complication and expense to the computations. Too few features result in loss of information and prediction accuracy.

The second task is to find an optimal forecasting method linking these features to oxidant values. Again many formal techniques exist (28). Multivariate piecewise linear regression could be used to approximate global nonlinearities in the "real" feature-oxidant relation yet still give continual predictions. If prediction into categories is desired, for example, whether a given day will or will not exceed a particular oxidant standard, then many pattern classification methods (28) are available.

An example of an approach that is in the spirit of pattern recognition, but does not use its formal mathematical methodology, is the objective ozone forecast system developed by Davidson (13) to predict the occurrence of days from July through October with ozone levels equal to or greater than 0.35 ppm. The forecasts are based on meteorological data available by 9 a.m. The forecaster first classifies days into 3 patterns, as shown in Figure 4, by 2 features: the 24-hour 500-mb (50-MPa) height change at Vandenberg Air Force Base and the 6 a.m. 2,500-ft (750-m) temperature at Los Angeles Airport. If the feature values for that day lie in area A of Figure 4, the forecast is <0.35 ppm oxidant; if they lie in area C, the forecast is ≥ 0.35 ppm. If the feature values lie in the intermediate area B, then other features are called into play for discrimination: first, the differences in 7 a.m. pressure and temperature between the Los Angeles (coastal) and Palmdale (inland) airports and, if necessary, the 8 a.m. visibility at Los Angeles Airport and the surface temperature change during an hour period at Palmdale. The accuracy of the forecasts is shown by the following tabulation of the results of 815 forecasts from July to October, 1964 to 1971.

	Actual	
Predicted	<0.35	≥0.35
<0.35	Correct 484	Incorrect 29
≥0.35	Incorrect 94	Correct 208

The skill score of the 815 trials was 0.66, according to Davidson's definition of a skill score:

$$S = \frac{R - E}{T - E} \tag{3}$$

where

R = number of correct forecasts,
T = total number of forecasts, and
E = number of forecasts expected to be correct due to chance.

In addition to Davidson's work, others have also developed stratification schemes that can be used in oxidant forecasting. For example, in recent work Bruntz et al. (8) related oxidant levels in New York and New Jersey to wind speed and solar radiation and Tiao, Phadke, and Box (37) studied the meteorological conditions when the daily instantaneous peak oxidant exceeded the 0.50-ppm alert level in Los Angeles County.

LONG-TERM MODELS (LONG-TERM STRATEGY)

The purpose of long-term air quality models is to predict the air quality impact of long-term changes in emission levels. These emission changes result from the growth or attrition of present sources, from control strategies, and from new developments (such as highways). Long-term oxidant models are specifically concerned with the effects of the level and spatial-temporal distribution of reactive hydrocarbon and nitrogen oxide emissions.

Four types of long-term statistical-empirical models are reviewed below. The first, linear rollback, is the most simplistic. Slightly more complex are modified rollback models based on empirical relations observed between maximum oxidant levels and hydrocarbons in the atmosphere. The third type of model is based on empirical relations of oxidant to both HC and NO_x. Each of the first 3 types neglects the spatial-temporal distribution of emissions; accordingly, each is restricted to analysis of the impacts of regionwide changes in the total level of emissions. The fourth type of empirical model illustrates how the spatial distribution of emissions can be incorporated into the statistical approach.

Linear Rollback

The linear rollback model for oxidant is based on the rather arbitrary assumption that maximum hourly oxidant levels in a region are directly proportional to total reactive hydrocarbon emissions in that region (12, 14). The model is calibrated by using aerometric data for oxidant as well as emission estimates for reactive hydrocarbons in some "base" year. Stated mathematically, the linear rollback relation between maximum hourly oxidant OX and regionwide reactive hydrocarbon emissions is

$$OX = \frac{RHC}{RHC^\circ} OX^\circ \quad (4)$$

where

OX° = maximum hourly oxidant measured in the base year, and
RHC° = total reactive hudrocarbon emissions in the base year.

Actually, the linear rollback model is not a statistical-empirical oxidant model because the relation is based on an arbitrary assumption rather than on an analysis of aerometric data. As with all types of models, linear rollback is calibrated with atmospheric measurements; however, the relation itself is not based on atmospheric observations. We include the linear rollback model in this discussion because it is similar to most statistical-empirical models in its simplicity of application and because it has been so widely used in air quality planning.

The defects of the linear rollback model are many. Linear rollback relates oxidant to hydrocarbons only; yet, it is known that oxidant levels depend significantly on both RHC and NO_x emissions. Linear rollback neglects important nonlinearities in the oxidant-hydrocarbon relation; the existence of these nonlinearities has been demonstrated by smog chamber, aerometric, and theoretical studies. Linear rollback also neglects background oxidant levels that may be significant, especially if the model is applied to determine oxidant concentrations near the federal standard.

In the form presented above, the linear rollback model also suffers from errors introduced by meteorological variance in the aerometric calibration data. Rather than calibrate the model with the actually measured maximum hourly oxidant level, it would be more appropriate to perform a statistical analysis of the base-year aerometric data. A statistical analysis of the data can determine the "expected" maximum hourly oxidant, which may differ significantly from the actually measured value (30). By determining the "expected value" from a statistical analysis, the model can at least be calibrated to predict a statistically well-defined parameter.

Aerometric Relation of Oxidant to HC

A second approach that relates maximum oxidant to reactive hydrocarbon emissions is modified rollback. Unlike linear rollback, modified rollback is based on a statistical relation between oxidant and hydrocarbons. This relation is usually in the form of upper limit curves determined from atmospheric data. Like linear rollback, the modified rollback approach can be applied only to regionwide problems since the spatial distribution of emissions is neglected.

The most widely used modified rollback model is the EPA Appendix J approach (17, 18, 34). The curve shown in Figure 5 represents the upper limit of maximum hourly oxidant levels that are associated with various concentrations of 6 to 9 a.m. nonmethane hydrocarbons (NMHC). The maximum daily oxidant levels and early morning hydrocarbon levels have been measured at the same location. Data have been used from 5 cities for the period from 1966 to 1968.

The modified rollback analysis is as follows: For a given base year, the measured maximum hourly oxidant level is plotted at point A°. The curve is then used to characterize a base year NMHC level at B°. For a new reactive hydrocarbon emission level, a new NMHC level is characterized at B, where the ratio of B to B° is in direct proportion to the ratio of the new and base-year emission levels. The curve is then used to predict an oxidant level, A, corresponding to the new emission level.

The EPA Appendix J modified rollback model has several serious limitations. The following are the main sources of error.

1. The model is subject to inaccuracies in the aerometric data base for oxidant and hydrocarbons.

Figure 3. Comparison of observed oxidant levels with Bennett's predictions using stepwise multiple regression.

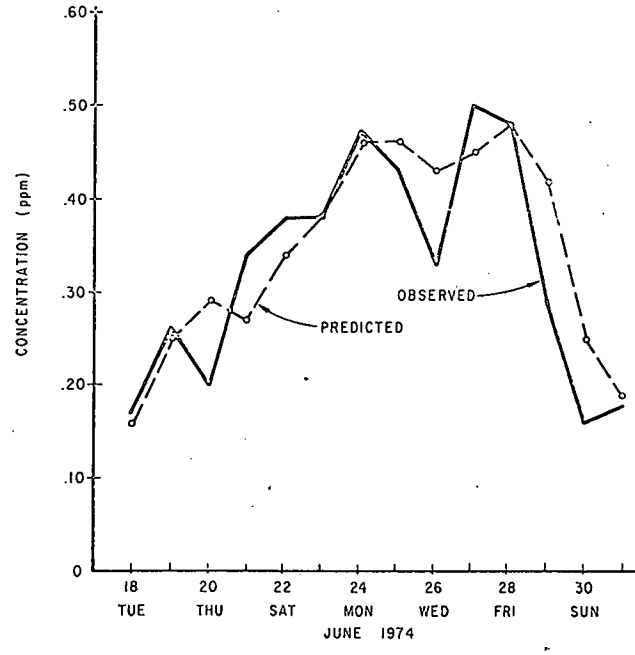


Figure 4. Davidson's initial decision rule for determining whether oxidant will reach the 0.35-ppm level.

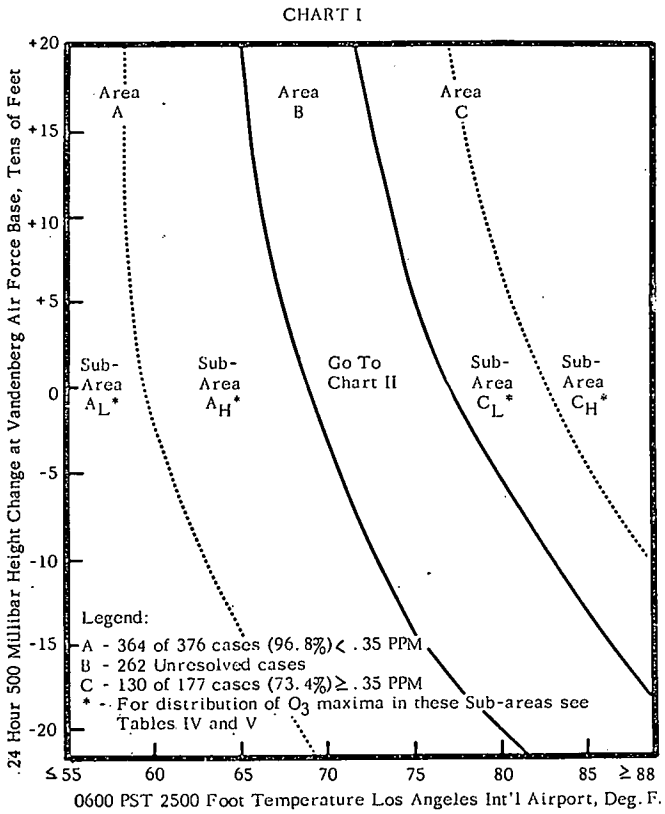
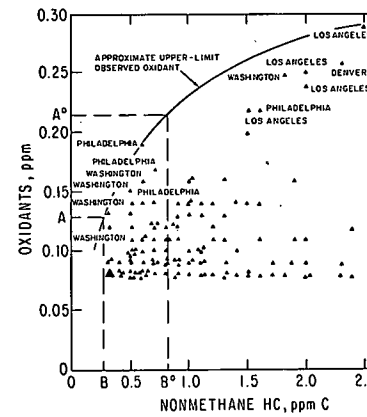


Figure 5. Maximum daily hourly average oxidant as a function of 6 to 9 a.m. average NMHC at stations throughout the country.



2. The role of NO_x in oxidant formation is neglected. The upper limit curves may no longer be appropriate if the HC-NO_x emission ratio is altered.
3. Relating oxidant concentrations to 6 to 9 a.m. precursor concentrations neglects the role of post 9 a.m. emission in oxidant production.
4. The approach does not account for transport. Early morning precursor and afternoon oxidant measurements at one location are likely to be associated with 2 different air masses.
5. The effect of meteorological variables is not taken into account. The observed relation of maximum oxidant to hydrocarbons may be spurious in the sense that it may be due to a mutual correlation with meteorological variables.
6. The upper limit curves are not defined in a statistically meaningful manner. Likewise, the calculation of degree of control required neglects statistical considerations.
7. Background HC and background OX contributions are neglected.

In addition, since the oxidant values shown in Figure 5 reach only 0.3 ppm, the EPA Appendix J approach cannot be used for regions that currently experience maximum oxidant levels greater than 0.3 ppm.

Schuck and Papetti of EPA have specialized and improved the modified rollback approach for the Los Angeles area (35). Their model is based on Figure 6, which gives the upper limit of maximum daily oxidant levels measured anywhere in the Los Angeles basin as a function of 6 to 9 a.m. hydrocarbons averaged over 8 stations in the basin. Figure 6 is based on data for 1971. The application of the Schuck and Papetti curve directly parallels the procedure for the EPA Appendix J curve.

The Schuck and Papetti model has the advantage of being specific to the region in which it is applied. Also, it accounts for transport in an approximate way by including all the monitoring stations in the air basin simultaneously. However, the other limitations of the EPA Appendix J approach (e.g., errors in the aerometric data, neglect of NO_x, neglect of post 9 a.m. emissions, neglect of meteorology, and the lack of statistical treatment) are shared by the Los Angeles upper limit curve.

Aerometric Relations of Oxidant to HC and NO_x

Several investigators have formulated statistical-empirical models that relate oxidant to both HC and NO_x. Nearly all applications have been restricted to the Los Angeles area because of the relative abundance of aerometric data in that area. As with the previously discussed models, these empirical models are restricted to the analysis of regionwide emission changes.

Merz, Painter, and Ryason (29) of Chevron Research Corporation used regression analysis to examine the relation between oxidant and early morning precursor levels at downtown Los Angeles. They regressed maximum daily 1-hour oxidant against 6 to 9 a.m. concentrations of NO_x and total hydrocarbons (THC). To minimize meteorological variations and, therefore, to minimize spurious oxidant-precursor dependencies due to mutual interrelations with meteorological variables, they entered data only for August, September, and October. The results of their regression analysis are shown in Figure 7.

The simple log-linear regression used by Merz, Painter, and Ryason indicated that NO_x reductions would have a slight but beneficial impact on oxidant air quality. This is in contrast to the results of the 2 models that follow in this discussion. The next 2 models indicate that NO_x emission reductions may have an adverse effect on oxidant air quality.

The Chevron research model can be used to predict the impact of regionwide changes in emission levels by proportioning the atmospheric concentrations of HC and NO_x to the changes in the respective emission levels. The improvements of the Chevron research model over the modified rollback models based on upper limit relations of OX to HC are the inclusion of NO_x and a better statistical treatment. However, the Chevron research model shares many limitations with the previous models: inaccuracies in the data base,

neglect of post 9 a.m. emissions, neglect of transport, and neglect of background contributions.

Kinosian and Paskind (22) of the California Air Resources Board examined the relation between oxidant and precursors at 4 locations in the metropolitan Los Angeles air quality control region. They used ambient data for 6 to 9 a.m. THC and NO_x concentrations and for maximum hourly oxidant concentrations measured at the same station. The data base consisted of measurements for July through September from 1969 to 1972. THC measurements were converted to NMHC estimates by using correlations established between THC and NMHC at 2 Los Angeles monitoring sites.

At each location, the data were grouped according to various early morning HC concentrations. For each HC level, a regression was run between oxidant levels and NO_x concentrations. The resulting curves, giving expected oxidant levels as functions of early morning HC and NO_x concentrations, are shown in Figure 8.

The Kinosian and Paskind results can be used to predict the impact of emission level changes in the same way as the Chevron research model. The limitations in using the Kinosian and Paskind results are the same as in using the Chevron research results.

Trijonis (38) used a stochastic model to examine the relation of oxidant levels in central Los Angeles to HC and NO_x emission levels. For given HC and NO_x emission levels, he determined the joint distribution of morning HC and NO_x concentrations (7:30 to 9:30 a.m. averages) at downtown Los Angeles from 5 years of Los Angeles APCD monitoring data (1966 to 1970). He also determined the probability that midday oxidant would violate the state standard (0.10 ppm for 1 hour) as a function of the morning concentrations. For oxidant, an average was taken of maximum hourly values between 11 a.m. and 1 p.m. at downtown Los Angeles, Pasadena, and Burbank, weighted according to wind speed and direction, so that the maximum oxidant would correspond as closely as possible to that in the air mass that had been over downtown in the morning. The joint morning HC-NO_x distribution and the probability of a standard violation as a function of morning precursor levels were determined separately for summer and winter.

By assuming that the joint HC-NO_x distribution responds linearly to emissions and that the oxidant standard violation function remains constant as emission levels change, Trijonis calculated the expected number of days per year that midday oxidant in central Los Angeles would exceed the state standard as a function of HC and NO_x emission levels. Figure 9 shows the results.

The model used by Trijonis involves several limitations similar to those of the upper limit models: inaccuracies in the aerometric data, neglect of post 9 a.m. emissions, and neglect of background contributions. The improvements in the approach are the inclusion of NO_x, a better statistical treatment, and the allowances for pollutant transport.

Spatial Resolution

Many statistical approaches consider the air basin as a single point. Although greatly simplifying the development of a model, this approach is inadequate for land use and transportation planning, which practically must consider the geographical positioning of alternative sites and the spatial distribution of their effects. A key assumption for the validity of a point air basin model is that the spatial distribution of the emission sources remains constant. In contrast, one of the thrusts of a model for planning should be the evaluation of the effects of changes in this distribution. An additional problem with the use of a point air basin model lies in the practical implementation of an effective environmental review procedure. In reviewing a proposed source, the agency granting approval may not wish to allow a significant air quality deterioration as a result of the project. If, for example, a significant deterioration is defined as a 10 percent change in the index of air quality, a serious problem results. It is unlikely that any single source would result in a 10 percent increase in air quality index for a total air basin. By a lumped air basin evaluation, therefore, a source having a significant environmental impact would rarely be found, and the air quality element of an environmental review would be rendered ineffective. For these reasons, a group at the

Figure 6. Schuck and Papetti's aggregated upper limit curve for the metropolitan Los Angeles air quality control region.

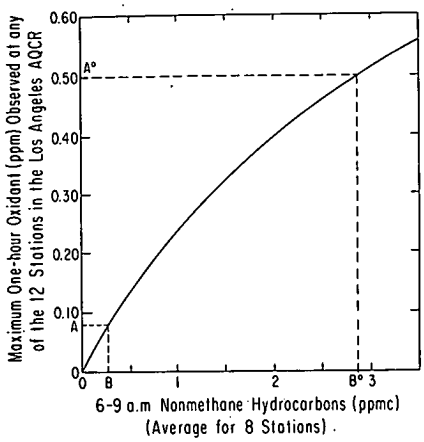


Figure 7. Merz, Painter, and Ryason's relation of NO_x and NMHC assumed as 50 percent of total HC and oxidant for downtown Los Angeles.

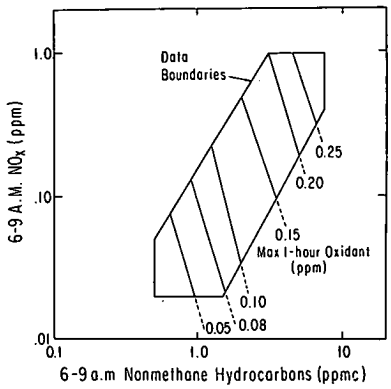


Figure 8. California Air Resources Board aerometric results, relation between 6 to 9 a.m. NO_x, 6 to 9 a.m. HC, and maximum hourly oxidant concentrations at selected sites (individual curves show total and NMHC concentrations in ppm).

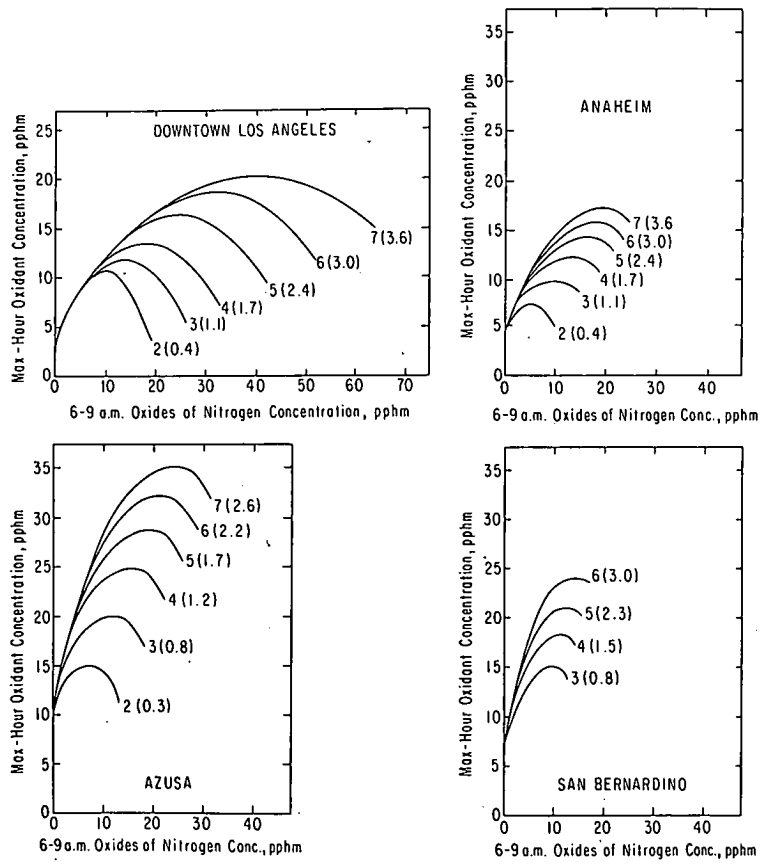


Figure 9. Expected number of days per year exceeding 0.10 ppm versus NO_x and RHC emissions levels for central Los Angeles.

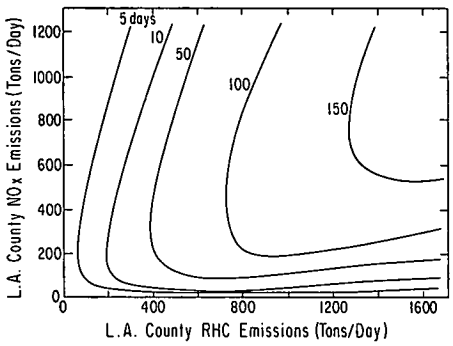
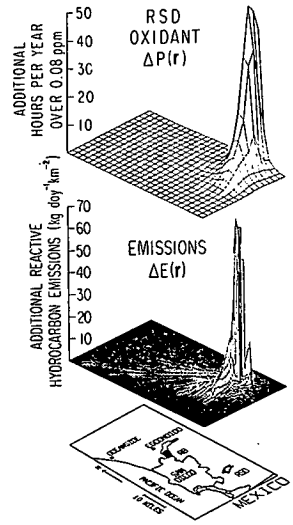


Figure 10. Application of spatially resolved model to predict effect on oxidant levels of a large regional development.



University of California, San Diego, has developed a statistical-empirical oxidant model with spatial resolution (31, 9).

Calibration of the relation between the spatially resolved emission function $E(\underline{r})$ and the resultant spatially resolved pollution function $P(\underline{r})$ is accomplished by using emissions and air quality data. Geocoded average daily RHC emission source functions $E(\underline{r})$ from mobile and stationary sources are calculated for Los Angeles and San Diego. $P(\underline{r})$ is expressed in terms of the number of hours per year higher than the federal 0.08-ppm oxidant standard. Data from 17 air quality monitoring stations in the San Diego and Los Angeles area are used in the calibration. The dependence of $P(\underline{r})$ on $E(\underline{r})$, averaged over meteorology $M(\underline{r})$, is approximated by integrating the RHC emissions picked up by seasonal prevailing windstream corridors from the ocean to each monitoring station and then by averaging over 4 seasons. The corridor width is chosen to be approximately a tenth of the maximum dimension of the Los Angeles and San Diego study areas.

A relation between hours per year higher than the federal standard and RHC picked up by the windstream is then developed by regression on the data from the 17 stations. This statistical relation may then be used to predict the spatial distribution of Ox resulting from a given spatial distribution of RHC emissions. An example of such a calibration is shown in Figure 10. The change in land use function $\Delta L(\underline{r})$ produces a change in the emission function $\Delta E(\underline{r})$, which is added to the base case emission function $E(\underline{r})$. The new total emission function is then used to calculate a new air quality function $P'(\underline{r})$, differing from the base case air quality function $P(\underline{r})$. The change in oxidant air quality resulting from altered land use $\Delta L(\underline{r})$ is $\Delta P(\underline{r}) = P'(\underline{r}) - P(\underline{r})$, in other words, the difference between air quality functions calculated with and without the development.

The model, like all existing oxidant models, has several limitations. Specifically, only RHC and not NO_x is considered. The windstreams used are only a crude representation of meteorological reality. The 17 data points on which the calibrating regression is based are quite noisy, and therefore the regression has a large uncertainty.

USEFULNESS OF STATISTICAL MODELS

Advantages

Statistical-empirical models have 2 major advantages. The first is their close relation to the actual atmospheric data on which they are based. Thus, we can hope to predict successfully even when deterministic understanding of the complex real world is incomplete. We can also hope that relations first inferred statistically may sometimes lead us in the end to a more fundamental understanding of the physical or chemical mechanism underlying the relation.

The second advantage is the relative simplicity and low cost of the development and use of statistical-empirical models. Computation using statistical-empirical models is usually rapid and relatively inexpensive. Thus, such models may be widely and repetitively applied, for example, to predict air quality each day at all monitoring stations in a region or to evaluate the air quality impact of large numbers of proposed land use and transportation projects.

Limitations and Dangers

That statistical-empirical models are derived from real atmospheric data is at once an advantage and a disadvantage. The disadvantage is that one is not assured of reliability in extrapolating the model beyond the range of conditions contained in the data from which it was derived. Since control strategies often are designed to drastically alter the present situation, one should be quite cautious in assessing the accuracy of such statistical predictions. One can make reliable error estimates for predicting tomorrow's oxidant level based on today's pollutant and meteorological measurements because tomorrow's meteorological pattern will probably be a repetition of the past. On the other hand, the probable error to be assigned to a prediction of the change in oxidant level due to an untried control strategy that would drastically change RHC and NO_x emissions is difficult to assess.

Short-term air quality forecasting has intrinsic limitations because air quality depends on meteorological variables that are themselves only imperfectly predictable (6). One should not expect to be able to predict air quality better than the weather. Certainly, unless there are major breakthroughs in weather forecasting, we cannot expect to predict short-term air quality more than a few days into the future any better than the mean value for the area in question at that time of year and time of day (and perhaps day of the week).

Caution should also be exercised in the use made of short-term predictions for episode control. One should not blindly assume that short-term emissions changes will always produce the obvious short-term results. For example, the evidence from both the east and the west coasts indicates that, although average precursor concentrations (RHC and NO_x) drop on weekends with altered emission patterns, the average oxidant level does not necessarily drop, and under some circumstances even rises (8, 11, 16, 20, 21, 25, 26, 33).

Suitable Areas for Application

The field of statistical-empirical oxidant modeling is in an early stage of development. Many obvious possibilities have not yet been tried, and imaginative new applications will certainly emerge. Several possible application areas are, however, already clear.

Short-Term Forecasting

1. Episode control. If one can predict air pollution episodes ahead of time, perhaps one can reduce their severity by altering the emission pattern through short-term

control of traffic and stationary sources.

2. Health warning. Whether one can control episodes, one can warn those particularly susceptible to air pollution effects to take precautions such as avoiding exercise, staying in filtered rooms, or leaving the area.

3. Crop protection. Perhaps with episode warnings, agricultural crops in some instances can be protected by spraying, irrigating, or covering.

Long-Term Prediction

1. Control strategy assessment. If one can assess the air quality impact of control strategies, one can optimize the benefit-cost ratio among control strategy options.

2. Regional planning. One can evaluate the air quality impact of alternative futures implied by different regional plans.

3. Transportation planning. The air quality impact of alternative transportation plans, for example, mode-split variations, can be assessed.

4. Environmental impact reports. Given a statistical model with spatial resolution, one can assess the air quality impact of highways, airports, industrial and commercial centers, and new towns. The impact can in theory at least be predicted as a function of location of the emission source and location of the receptor.

CONCLUSION

Statistical-empirical modeling of oxidant levels is currently proceeding in 2 directions, each dictated by the prediction time span and the envisioned use. The first direction is short-term oxidant forecasting over the range of hours to days, with the goal of episode control by short-term emission control as well as the issuance of health and perhaps agricultural warnings. The major tools being used are Box-Jenkins time series analysis, multiple regression, and aspects of pattern recognition.

The second direction is long-term prediction over the scale of years of the expected changes in oxidant levels due to changes in RHC and NO_x emissions, either aggregated for an air basin or more rarely spatially resolved. The goal of such prediction is the assessment of the probable effects on oxidant air quality of various emission control strategies. Possible applications include optimization of control strategies, regional planning, transportation planning, and environmental impact reports. The major categories of models are rollback, prediction from HC, prediction from HC and NO_x, and spatial resolution. The advantages of using statistical-empirical models include (a) their close relation to the actual air measurement data from which they are derived, allowing the possibility of correct prediction even in the absence of complete understanding of the underlying phenomenon, and (b) their usual simplicity and low cost of development and use.

Their disadvantages and dangers include uncertain reliability outside the range of the data used in calibration, inability to predict short-term levels better than the associated weather features, and danger of assuming that short-term emission changes will always produce the expected result.

Oxidant modeling and prediction of all varieties are still an uncertain art that is in the early stages of development, and for this reason it is often wise to evaluate the same question by using a variety of modeling approaches. No aspect of the field is yet mature or free from difficulties. Chemical modeling in a smog chamber is on too small a scale and has too great differences from conditions of the real atmosphere. Mechanistic modeling of chemistry and meteorology has too many unknowns for the state of our chemical and meteorological understanding as well as for our computational capacity. Statistical-empirical modeling is most often called on to perform tasks that carry it into statistically shaky ground: (a) short-term prediction of extreme values, episodes on the tail of the distribution that are not representative of the data set as a whole on which the model is based and (b) long-term prediction of what would happen under control strategies that would alter the emission pattern far beyond the situation that pro-

duced the data used to derive the model.

Yet, uncertainty is no excuse for inaction. We can set reasonable bounds with a variety of prediction schemes. Although the numerical uncertainties may be large, the direction in which to proceed in controlling air pollution is usually clear. We can climb up a hill without knowing precisely how far it is to the top, and, as we proceed, the view of the top becomes clearer.

REFERENCES

1. A. P. Altshuller. Evaluation of Oxidant Results at CAMP Sites in the United States. Journal of the Air Pollution Control Association, Vol. 25, 1975, p. 19.
2. C. Bennett. California Air Quality Data. California Air Resources Board, Sacramento, Vol. 6, Oct.-Nov.-Dec. 1974.
3. G. E. P. Box and G. M. Jenkins. Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco, 1970.
4. G. E. P. Box and G. C. Tiao. Intervention Analysis With Applications to Economic and Environmental Problems. Journal of the American Statistical Association, Vol. 70, 1975, p. 70.
5. G. E. P. Box and G. C. Tiao. Comparison of Forecasts and Actuality. Department of Statistics, Univ. of Wisconsin, Madison, Technical Rept. 402, May 1975.
6. D. R. Brillinger and E. L. Scott, eds. Forecasting Air Pollution. Proc., Conference on Forecasting Air Pollution, 1974, Department of Statistics, Univ. of California, Berkeley, 1975.
7. S. M. Bruntz, W. S. Cleveland, B. Kleiner, and J. L. Warner. The Dependence of Ambient Ozone on Solar Radiation, Wind, Temperature, and Mixing Height. Symposium on Atmospheric Diffusion and Air Pollution, American Meteorological Society, Preprint Volume, Sept. 1974.
8. S. M. Bruntz, W. S. Cleveland, T. E. Graedel, and B. Kleiner. Ozone Concentrations in New Jersey and New York: Statistical Association With Related Variables. Science, Vol. 186, 1974, p. 257.
9. J. M. Caporaletti, L. N. Myrabo, P. Schleifer, A. Stanonik, and K. R. Wilson. Statistical Oxidant Air Quality Prediction for Land Use and Transportation Planning. Univ. of California, San Diego.
10. D. P. Chock and T. R. Terrell. Time Series Analysis of Riverside, California Air Quality Data. General Motors Research Laboratories, Warren, Mich., Publ. GMR-1591, June 1974.
11. W. S. Cleveland, T. E. Graedel, B. Kleiner, and J. L. Warner. Sunday and Workday Variations in Photochemical Air Pollutants in New Jersey and New York. Science, Vol. 186, 1974, p. 1037.
12. W. A. Daniel and J. M. Heuss. Ambient Air Quality and Automotive Emission Control. Journal of Air Pollution Control Association, Vol. 24, Sept. 1974, p. 849.
13. A. Davidson. An Objective Ozone Forecast System for July Through October in the Los Angeles Basin. Los Angeles Air Pollution Control District, Air Quality Rept., 1974.
14. N. de Nevers and J. R. Morris. Rollback Modelling: Basic and Modified. Journal of Air Pollution Control Association, Vol. 25, Sept. 1975, p. 943.
15. N. R. Draper and H. Smith. Applied Regression Analysis. John Wiley and Sons, New York, 1966.
16. B. Elkus and K. R. Wilson. Photochemical Air Pollution: Weekend-Weekday Differences. Univ. of California, San Diego, 1974.
17. Air Quality Criteria for Nitrogen Oxides. U.S. Environmental Protection Agency, Publ. AP-84, Jan. 1971.
18. Federal Register. Vol. 36, No. 158, Aug. 14, 1971.
19. M. Feldstein, J. Sandberg, L. Robinson, and A. Norris. Relationship of Oxidant Peak, High-Hour and Slope Values as a Guide in Forecasting Health-Effect Days. Technical Services Division, Bay Area Air Pollution Control District, Feb. 1973.
20. A. Haagen-Smit and M. F. Brunelle. International Journal of Air Pollution,

Vol. 1, 1958, p. 51.
21. A. J. Hocker. Los Angeles APCD Air Quality Report. Los Angeles Air Pollution Control District, Rept. 51, Jan. 1963.
22. J. R. Kinosian and J. J. Paskind. Hydrocarbons, Oxides of Nitrogen, and Oxidant Trends in the South Coast Air Basin, 1963-1972. Division of Technical Services, California Air Resources Board, internal working paper, 1973.
23. L. D. Kornreich, ed. Proc., Symposium on Statistical Aspects of Air Quality Data, 1972, U.S. Environmental Protection Agency, Research Triangle Park, N.C., EPA-650/4-74-038, 1974.
24. E. S. Lee, S. C. Sarin, and K. M. Wang. Parametric Time Series Modeling of Stochastic Air Pollution Data. Department of Electrical Engineering, Univ. of California, Los Angeles, Technical Rept. 73-15, April 1973.
25. S. B. Levitt and D. P. Chock. Weekday-Weekend Pollutant and Meteorological Studies of the Los Angeles Basin. General Motors Research Laboratories, Warren, Mich., Publ. GMR-1866, June 1975.
26. W. A. Lonneman, S. L. Kopczynski, P. E. Darley, and F. D. Sutterfield. Hydrocarbon Composition of Urban Air Pollution. Environmental Science and Technology, Vol. 8, 1974, p. 229.
27. G. M. McCollister and K. R. Wilson. Linear Stochastic Models for Forecasting Daily Maxima and Hourly Concentrations of Air Pollutants. Atmospheric Environment, Vol. 9, 1975, p. 417.
28. W. S. Meisel. Computer-Oriented Approaches to Pattern Recognition. Academic Press, New York, 1972.
29. P. H. Merz, L. J. Painter, and P. R. Ryason. Aerometric Data Analysis: Time Series Analysis and Forecast and an Atmospheric Smog Diagram. Atmospheric Environment, Vol. 6, 1972, p. 319.
30. J. L. Mitchner and J. W. Brewer. A Comment on the Method Used by EPA to Calculate Required Reductions in Emissions. Univ. of California, Berkeley, working paper, 1973.
31. L. N. Myrabo, P. Schleifer, and K. R. Wilson. Oxidant Prediction Model for Land Use and Transportation Planning. California Air Environment, Vol. 4, 1974, p. 3.
32. R. I. Pollack. Studies of Pollutant Concentration Frequency Distributions. Polytechnic Institute of Brooklyn, New York, PhD thesis; Lawrence Livermore Laboratory, Univ. of California, Livermore, Rept. UCRL-51459, Oct. 1973.
33. E. A. Schuck, J. N. Pitts, Jr., and J. K. S. Wan. Relationship Between Certain Meteorological Factors and Photochemical Smog. International Journal of Air and Water Pollution, Vol. 10, 1966, p. 689.
34. E. A. Schuck, A. P. Altshuller, D. S. Barth, and G. B. Morgan. Relationship of Hydrocarbons to Oxidants in Ambient Atmospheres. Journal of Air Pollution Control Association, Vol. 20, May 1970, p. 297.
35. E. A. Schuck and R. A. Papetti. Examination of the Photochemical Air Pollution Problem in the Southern California Area. U.S. Environmental Protection Agency, internal working paper, May 1973.
36. G. C. Tiao, G. E. P. Box, and W. J. Hamming. Analysis of Los Angeles Photochemical Smog Data: A Statistical Overview. Journal of Air Pollution Control Association, Vol. 25, 1975, p. 260.
37. G. C. Tiao, M. S. Phadke, and G. E. P. Box. Some Empirical Models for the Los Angeles Photochemical Smog Data. Department of Statistics, Univ. of Wisconsin, Madison, Technical Rept. 412, June 1975.
38. J. C. Trijonis. An Economic Air Pollution Control Model Application: Photochemical Smog in Los Angeles County in 1975. Environmental Science and Technology, Vol. 8, Sept. 1974, p. 811.
39. M. D. Zeldin and D. Thomas. Ozone Trends in the Eastern Los Angeles Basin Corrected for Meteorological Variations. International Conference on Environmental Sensing and Assessment, Las Vegas, Sept. 17, 1975.