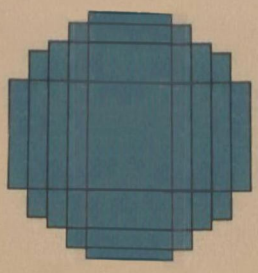
EMPIRICAL METHODS IN THE ANALYSIS OF AIR QUALITY AND METEOROLOGICAL PROBLEMS

William S. Meisel

Interim Report

December 1974



Technology Service Corporation

Technology Service Corporation

2811 Wilshire Boulevard
Santa Monica, California 90403
(213)829-7411

DRAFT: For Internal EPA Review

EMPIRICAL METHODS IN THE ANALYSIS OF AIR QUALITY AND METEOROLOGICAL PROBLEMS

William S. Meisel

Interim Report

December 1974

Contract No. 68-02-1704

EPA Project Officer: Ken Calder

Meteorology Laboratory
National Environmental Research Center
Research Triangle Park, North Carolina 27711

Prepared for

OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
WASHINGTON, D.C. 20460

91851

PREFACE

This interim report serves two functions: it is (1) an outline of the proposed phase II projects, for comment, and (2) a draft of the first volume of the three-volume final report:

I. Empirical Methods in the Analysis of Air Quality and Meteorological Problems

II. A Source-Oriented Empirical Model of the Dispersion of Air Pollutants.

III. The Oxidant Formation Process in the Los Angeles Basin: An Empirical Analysis.

This volume will be revised to serve as the introductory volume indicated; the revised version will be delivered with the final report.

Discussions with EPA personnel led to inclusion of many of the subjects dealt with in this report. The project monitor, Ken Calder, took a particularly active and constructive role. Advice from Leo Breiman and Alan Horowitz at Technology Service Corporation further improved the report.

TABLE OF CONTENTS

TABLE OF CONTENTS (CONT'D)

1.0 INTRODUCTION

The increased availability of appropriate data bases and improvements in methodology have led to the increasing use of empirical and statistical approaches to the analysis of air quality and meteorological problems [1]. This volume suggests how these approaches might be applied to a number of problems of interest to the Environmental Protection Agency, particularly those with a meteorological aspect.

The objectives of this report are limited. The applications discussed at most length are those where either empirical approaches have not been fully exploited and/or the problem can be formulated in an innovative manner. The discussions are intended to highlight opportunities rather than to provide detailed plans for problem solution. As part of the present project, two problem areas will be explored as pilot studies to more fully demonstrate modern empirical techniques; these projects will be reported in separate volumes.

The subjects discussed in this volume are the following:

A source-oriented empirical model: It has often been assumed that it is impractical to derive an empirical model relating emission source distribution and meteorology to the resulting pollutant concentration distribution. The basic argument against an empirical approach has been the dearth of detailed emission inventories in comparison to the relative abundance of air quality data. An approach is postulated whereby it is suggested that an empirical meteorological dispersion function can be derived by indirect means.

Empirical analysis of the oxidant formation processess: Empirical approaches to the problem of determining the relationship between oxidant precursor (HC and NO_x) concentrations and resulting ambient oxidant levels are discussed.

Extraction of emission trends from air quality trends: The estimation of air quality trends from air quality measurements is complicated by the effect of meteorology. We discuss the determination of a "meteorologically adjusted" trend, i.e., a trend more nearly related to the emissions trend.

Detection of inconsistencies in air quality/meteorological data bases: In any data collection or data analysis effort, a major concern is the integrity of the data. It is important to detect problems with monitoring equipment or monitoring methods and to note any important changes in the system monitored so that such errors or changes do not distort the analysis of the data or invalidate a portion of the data collected. We discuss automatic procedures for detecting inconsistencies.

Empirical approaches to the understanding and efficient use of complex air quality models: Computer-based models derived from physical principles are tools which often should be analyzed themselves for the sake of extracting their implications, for modeling aspects of their behavior to reduce input data requirements and running time, for validation, or to suggest further areas for model improvement. Model-generated input/output data can be so analyzed by empirical techniques.

Spatial interpolation of meteorological and air quality measurements: Interpolation of variables such as wind field or pollutant concentration is of interest in several applications. We discuss some general aspects of this problem.

Health effects of air pollution: We comment on this area in which empirical approaches are at present heavily employed.

Short-term pollutant level forecasting: Short-term forecasting for health warning systems or to invoke temporary controls can be approached empirically. Several pitfalls are highlighted.

In discussions of the above subjects, the attempt is to formulate a data-analytic approach which reduces the problem to a straightforward data-analytic technique. The techniques of empirical analysis which are referenced include the following:

1. Hypothesis testing, statistical modeling, and other "classical" statistical approaches: These "classical" approaches are by no means without their subtleties or potential for misapplication, but are the subject of many textbooks and conventional statistics courses.
2. Linear and nonlinear regression: These techniques fit a function to data to model the relationship between independent variables and an ordered, many-valued independent variable. Linear regression, in general, and nonlinear regression in a single independent variable are well-understood and often used. Nonlinear regression with multiple independent variables, particularly for the small-sample case, is more difficult, but significant technical progress has been made in the last few years.

3. Time-series analysis: Time-series analysis takes advantage of the serial nature of the data and presumably of the underlying model. The subject has been studied for many years (sometimes as "signal processing"), but in recent years new developments have arisen and the subject has been treated more systematically. The linear case is much more highly developed than the nonlinear case; however, not all problems involving time series are best treated by techniques designed specifically for time series.
4. Classification analysis ("pattern recognition"): These techniques use data to relate independent variables to a class label (i.e., a possibly unordered, few-valued dependent variable). Because earlier work in this field was oriented toward developing hardware devices rather than analyzing data, its power as a data-analytic tool has only been fully realized in the last few years.
5. Cluster analysis: Cluster analysis does not require a dependent variable but analyzes the distribution of multivariate data, i.e., the joint distribution of the independent variables, to determine distinct groupings of data points in multivariate space. Much work on the subject has been done recently, and it will become better known when several textbooks in press are published. Discussions of cluster analysis tend at present to appear as chapters in books on pattern recognition.

Each of these data-analytic subjects is difficult and tends to have its own language and proponents. Further, few universities currently encourage students to become broadly based experts in data

analysis. Hence, tradeoffs among techniques are not always made with obtaining the best problem solution as the only criterion. In the present report, a sincere attempt has been made to formulate the problem in the most general terms, pointing out the class of techniques applicable but seldom specific algorithms.

2.0 A SOURCE-ORIENTED EMPIRICAL MODEL

2.1 Motivation

Multiple-source simulation models for urban air quality based on meteorological dispersion functions are in broad use [2]; for example, the Gaussian plume formulation is used in many models including the RAM model presently in development at the Environmental Protection Agency [3].

The particular form of relationship between source and receptor used in this formulation was originally developed to describe dispersion from isolated sources and has been adapted to the urban environment. Because a source-oriented model is extremely useful in examining the impact of proposed emission controls, it is of interest to determine if a source-receptor relationship which provides an alternative to the Gaussian plume formulation can be determined empirically.

Since one cannot, in general, isolate the effects of single sources in an urban area to determine the source-receptor relationship, we propose that the relationship be extracted indirectly by determining a formulation which will best predict the pollutant concentration distribution, given the emission distribution and meteorological conditions.

2.2 Formulation

The basic Gaussian plume equation predicts the concentration at a point (x,y,z) from a source of unit strength at (ξ,η,ζ) as

$$R(x,y,z;\xi,\eta,\zeta) = \frac{1}{2\pi\bar{u}\sigma_y(x-\xi)\sigma_z(x-\xi)} \cdot \left[\exp\left\{ -\frac{1}{2}\left[\frac{(y-\eta)^2}{\sigma_y^2(x-\xi)} + \frac{(z-\zeta)^2}{\sigma_z^2(x-\xi)} \right] \right\} + \exp\left\{ -\frac{1}{2}\left[\frac{(y-\eta)^2}{\sigma_y^2(x-\xi)} + \frac{(z+\zeta)^2}{\sigma_z^2(x-\xi)} \right] \right\} \right] \qquad (2\text{-}1)$$

where

$\bar{u}$ = mean wind speed,

ζ = effective source height, and

$\sigma_y(d), \sigma_z(d)$ = horizontal and vertical diffusion functions a distance d downwind from the source.

The first term within the brackets of Eq. (2-1) denotes the dispersion of the pollutant in the lateral direction; the second term, in the vertical direction; and the third term represents the perfect reflection of the pollutant bearing diffusive eddies from the surface of the earth, i.e., there is neither deposition nor reaction at the surface. The coordinates are aligned such that x is along-wind and y is crosswind. In fact, Eq. (2-1) holds only when $x-\xi$ is positive; the concentration is assumed to be zero if the source is downwind. The diffusion functions depend on meteorological parameters, usually mixing layer depth and stability condition.

The concentration from multiple sources in a region V with a source strength distribution $Q(\xi,\eta,\zeta)$ is given by superposition:

$$\chi(x,y,z) = \int_V R(x,y,z;\xi,\eta,\zeta)Q(\xi,\eta,\zeta)d\xi d\eta d\zeta \qquad (2\text{-}2)$$

where the integral over the volume V can be abstractly considered to include both point and area sources.

Following Calder [4], we can assume horizontal homogeneity, as in the Gaussian formulation:

$$R(x,y,z;\xi,\eta,\zeta) = K(x-\xi,y-\eta,z,\zeta) \quad , \qquad (2\text{-}3)$$

yielding an equivalent to 2-2):

$$\chi(x,y,z) = \int_V K(x',y',z,\zeta)Q(x-x',y-y',\zeta)dV' \quad , \tag{2-4}$$

where we have made the change of variable

$$x' = x-\xi$$

$$y' = y-\eta$$

and $dV' = dx'dy'd\zeta$.

Again following Calder [4], we note that (2-4) represents an integral equation for the function K if the concentration distribution χ and emission distribution Q are known. In other words, we might conceive of determining the source-receptor function K empirically by examining observed concentration distributions resulting from observed (or estimated) emission distributions. Calder notes that in the case where (1) we are predicting ground-level concentrations from area sources at ground-level, (2) the integral is approximated by a summation over an M-by-N grid of values, and (3) we have measured concentrations and emissions at each grid point, K can be determined in tabular form by solving a set of linear equations. The table specifying the function K(x',y') in this case yields values for any pair of grid points (x,ξ) and (y,η) and is valid for the meteorological conditions which yielded the particular concentration distribution used. (The table would appear as in Figure 2-1). A tabular formulation of the source-receptor function K

$x' = x - \xi$

$y' = y - \eta$ \ K	.1	.2	.3	.4
.1	5.0	5.5	4.5	3.0
.2	4.0	4.5	3.0	2.5
.3	3.0	2.0	1.5	1.4
.4	2.0	1.6	1.2	1.0

Figure 2-1. A tabular representation of a hypothetical source-reception function for a fixed meteorology. Another table would be required for another set of meteorological conditions.

has the advantage of not restricting the class of functions being investigated, but has the disadvantage of requiring values of the concentration distribution at each grid point and making the dependence upon meteorological factors difficult to extract explicitly.

An alternative approach to solving equation (2-4) is to restrict K to membership in a family of functions $K(x',y',z,\zeta;\underline{\alpha})$, where choosing the parameter vector $\underline{\alpha}$ specifies a particular member of the family. A familiar example is the family of multivariate polynomials where a member is specified by a particular choice of values for the coefficients. In this case, K is specified as a specific functional form rather than as a table.

One approach to determining the "best-fitting" function of the chosen family (or, equivalently, of finding the parameter vector $\underline{\alpha}^*$ which gave the best fit) is to fit, by a classical least-squares method, the values yielded by the set of linear equations suggested by Calder. In the hypothetical example of Figure 2-1, the 16 values of the table could be fit by a function of two variables. Since this is a two-step approach, it does not overcome the problems of the first approach, but amounts largely to smoothing the values yielded by that approach.

The more direct approach is to substitute $K(x',y',z,\zeta;\underline{\alpha})$ directly in equation (2-4). If there were an $\underline{\alpha}$ such that the equation could be satisfied exactly, the concentration distribution could be predicted exactly by the function given by that $\underline{\alpha}$. If a perfect fit is not possible, then one could find the parameters $\underline{\alpha}$ which minimized the mean-square error over a number of measurement points (x_i,y_i,z_i), i=1,2,...,M, perhaps corresponding to monitoring stations:

$$e^2(\underline{\alpha}) = \frac{1}{M}\sum_{i=1}^{M}\left[\chi_i - \int_{V'} K(x',y',z_i,\zeta;\underline{\alpha})Q(x_i-x',y_i-y',\zeta)dV'\right]^2 \qquad (2\text{-}5)$$

where $\chi_i = \chi(x_i,y_i,z_i)$.

Equation (2-5) can be minimized with respect to $\underline{\alpha}$ by any number of optimization techniques if the integral can be calculated; many numerical integration techniques are suitable for that purpose. A key problem is the choice of an appropriate family of parameterized functional forms for K such that the error e^2 will be small, but such that the number of

parameters $\underline{\alpha}$ is small. The number of parameters is related to the number of measurements required to make the problem well-determined and to its computational feasibility. We will discuss this point after further refinement of the problem formulation.

Explicit Dependence on Meteorology

Since the concentration distribution in (2-5) is determined by meteorological parameters as well as the source distribution, K determined by that formulation would be valid only for that particular set of meteorological conditions. If we denote the vector of meteorological parameters as $\underline{m}$ (e.g., wind speed, inversion height, and stability class) and express the dependence of K and χ on meteorology as $K(x',y',z,\tau,\underline{m};\underline{\alpha})$ and $\chi(x,y,z,\underline{m})$, then the error e^2 is a function of the choice of parameters $\underline{\alpha}$ and meteorological conditions $\underline{m}$:

$$e^2 = e^2(\underline{\alpha},\underline{m}) \quad . \tag{2-6}$$

If a set of N meteorological conditions we wish to explore is given by $\underline{m}_1,\underline{m}_2,\ldots,\underline{m}_N$, then we may find $\underline{\alpha}$ to minimize

$$E^2(\underline{\alpha}) = \frac{1}{N}\sum_{j=1}^{N} e^2(\underline{\alpha};\underline{m}_j) \quad , \tag{2-7}$$

where e^2 is defined by (2-5). This last equation gives the mean-square error over all measurement points (x_i,y_i,z_i) and over all the chosen meteorological conditions. The function resulting from optimization, $K(x',y',z,\tau,\underline{m};\underline{\alpha}^*)$, should predict these MN points accurately. If $\underline{m}$ is

three-dimensional, then K is a function of six variables and explicitly contains dependence on the meteorology. Equation (2-1), the Gaussian formulation, is an example of such a functional form (although not obtained by optimization).

Area and Point Sources

We can further refine our formulation by explicit consideration of area and point sources.

Area sources at ground level yield concentrations at ground level given by

$$\chi_A(x,y,\underline{m}) = \chi(x,y,0,\underline{m}) = \int_{V'} K_A(x',y',\underline{m};\underline{\beta})Q_A(x-x',y-y')dx'dy' \quad . \qquad (2\text{-}8)$$

Elevated point sources measured at ground level are given by

$$\chi_P(x,y,\underline{m}) = \chi(x,y,0,\underline{m}) = \sum_{\ell=1}^{J} K_P(x'_\ell,y'_\ell,\zeta_\ell,\underline{m};\underline{\gamma})Q_P(x-x'_\ell,y-y'_\ell,\zeta_\ell) \quad , \qquad (2\text{-}9)$$

where the point sources are at

$$(\xi_\ell,\eta_\ell,\zeta_\ell), \qquad \ell = 1,2,\ldots,J,$$

and

$$x'_\ell = x - \xi_\ell \text{ and } y'_\ell = y - \eta_\ell \quad .$$

The total concentration at (x,y) with meteorological conditions $\underline{m}$ is given by

$$\chi(x,y,\underline{m}) = \chi_A(x,y,\underline{m}) + \chi_p(x,y,\underline{m}) \tag{2-10}$$

As before, the optimal source-receptor function K is determined by finding the parameters $\underline{\beta}$ and $\underline{\gamma}$ which minimize the mean-square error in predicting concentrations over a varying set of meteorological conditions:

$$E^2(\underline{\beta},\underline{\gamma}) = \frac{1}{N}\sum_{j=1}^{N}\left\{\frac{1}{M}\sum_{i=1}^{M}\left[\chi_i - \int_{V'} K_A(x',y',\underline{m}_j;\underline{\beta})Q_A(x_i-x',y_i-y')dx'dy'\right.\right.$$

$$\left.\left. - \sum_{\ell=1}^{J} K_p(x'_\ell,y'_\ell,\zeta_\ell,\underline{m}_j;\underline{\gamma})Q_p(x_i-x'_\ell,y_i-y'_\ell,\zeta_\ell)\right]^2\right\} \tag{2-11}$$

(Note that the effective stack height ζ may also be a function of the meteorology: $\zeta = \zeta(\underline{m})$.)

Equation (2-11) summarizes the basic method proposed. Presuming the integral is approximated using a numerical integration technique, the error E^2 can be calculated for any choice of parameters. A number of optimization techniques can be employed to find the parameters which give the best fit. Given those optimal parameter values, the optimal source-receptor function K_A and K_p are fully specified. It remains to consider the feasibility of this approach.

2.3 Feasibility

Feasibility of the proposed method depends upon two closely related problems:

(1) Does the problem as posed have a well-defined solution?

(2) Even if the solution is theoretically well-defined, is it computationally feasible to obtain?

Both questions are heavily dependent on the number of parameters defining K_A and K_P, the dimensionality of $\underline{\beta}$ and $\underline{\gamma}$. The number of values to be fit (M·N) should be greater than the number of free parameters; if so, the solution will in most cases be well-defined (perhaps within some limited region of parameter space).

Computational feasibility also depends on the number of parameters. The cost of most optimization algorithms will tend to go up as a power of the number of parameters whose values must be determined. A major objective of this approach must hence be to find a parameterized functional form which is sufficiently general to be able to model the source-receptor function, but which does not require a large number of free parameters to achieve this generality. Continuous piecewise linear functions, as used in a recent EPA study [5], provide such a class of functions and are a promising candidate for achieving a feasible solution.

Whatever form of approximating function is used, however, one can simplify the problem by making certain assumptions regarding the source-receptor function; for example:

(1) One might assume specific dependencies on some meteorological parameters, e.g., by assuming that the concentration is inversely proportional to the average wind speed rather than extracting that dependence empirically.

(2) One might assume the Gaussian form and determine the dispersion functions empirically.

(3) One might make the narrow-plume assumption for area sources [6].

The more assumptions made, the less general, but also less difficult, the analysis will be.

2.4 The Inverse Problem

Suppose a good source-receptor function has been determined. Then one might pose the problem: Given meteorological conditions, measurements of the pollutant concentration distribution, and source locations, determine the distribution of source strengths.

Equation (2-11) provides a formulation of this problem if $\underline{\beta}$ and $\underline{\gamma}$ are assumed known and the area and/or point source strengths assumed unknown. For example, suppose the area sources are assumed known, and the point source emission rates to be determined. There are then J unknown values for the J point sources, and E^2, the mean-square error in predicting the measured concentrations, is a function of those J values. The values which minimize E^2 are good estimates of the source emission rates. Since the unknowns appear linearly within the brackets, the minimum of (2-11) can be found by solving a set of J linear equations in J unknowns. The solution will be well-determined (except in degenerate cases) if the number of concentration measurements times the number of meteorological conditions exceeds the number of point sources.

2.5 Testing the Approach

A common problem in testing meteorological and air quality models is that the data base required for the models is subject to errors which may be of the same size as errors introduced by the models. Emission inventories and estimates of diurnal variations in emissions, for example, may tend to be correct on the average, yet be considerably in error in any given hour. A least-squares formulation such as that proposed tends to average out errors and will tend to produce good models. In order to gain confidence in the approach and to explore alternative levels of assumptions, however, it is desirable to use a case where the source of errors will arise from the model formulation rather than measurement error. One means to this end is to use data generated by a model such as the Gaussian-plume RAM model referenced earlier. If the source-receptor function derived by the proposed approach closely approximates the Gaussian form used in generating the data, one would have increased confidence in the applicability of the proposed technique to measured data. Further, alternative versions of the methodology could be analyzed in a controlled environment. The proposed Phase II study thus suggests this approach and follows a work plan suggested by Calder [4].

2.6 Research Plan

Task 1

Select a real urban location (e.g., St. Louis, New York, or Chicago) for which ground-level area- and point-source, short-term emissions distributions are available for SO_2. For a typical one-hour emissions distribution, use a multiple-source Gaussian dispersion model (probably the RAM model of

the EPA Meteorology Laboratory)--for one wind speed (5 m/sec), one stability class (neutral) and infinite mixing depth--to calculate total one-hour concentrations $\chi(P_i,\theta_j)$ at ground level at a number of receptor locations P_i (e.g., as for St. Louis RAPS network) and for various wind directions θ_j. Apply the methodology proposed to attempt to recover the meteorological dispersion function K. Determine

(a) the degree of error in predicting concentrations, for the receptor locations and wind directions actually used to derive the empirical dispersion function,

(b) the degree of error in predicting concentrations at receptor locations and for wind directions not used in the derivation (a measure of interpolation accuracy),

(c) the degree of error in predicting results for a somewhat different emissions distribution (a test of extrapolation accuracy), and

(d) a comparison of the empirical dispersion function with the Gaussian form used to compute input concentrations for the analysis.

Task 2

Test the sensitivity of the method to the number of "observed" concentrations used and to random errors in the emissions inventory.

Task 3

Extend the preceding to a range of wind speeds, atmospheric stability classes, and to several different emissions distributions.

3.0 EMPIRICAL ANALYSIS OF THE OXIDANT FORMATION PROCESSES IN THE LOS ANGELES BASIN

3.1 Motivation

Oxidant is a difficult pollutant to deal with in terms of understanding the effect of particular controls on the resultant level of its concentration. The principal reason for this difficulty is that oxidant is largely an end product of a chemical process rather than being directly emitted from pollutant sources. Oxidant is related to emissions not only through transport and diffusion, but by a complex chemical reaction in which meteorology can play a significant part. The principal pollutants leading to the formation of oxidant are reactive hydrocarbons (HC) and oxides of nitrogen (NO_x). (We shall refer to these "raw" pollutants as "oxidant precursors.") Since emission control policies can affect not only the overall level of emissions but the ratio of emissions of NO_x to hydrocarbons, it becomes important to understand the effect of this ratio as well as of the absolute level of emissions upon the end concentrations of oxidant. These effects are by no means fully agreed upon.

One approach to understanding this problem is a very detailed inspection of all the physical processes involved, including meteorological effects and chemistry. One then obtains a model which, if successful, relates a detailed emissions inventory and detailed meteorological conditions to a resulting time and spatial distribution of oxidant values over the area modeled.

An alternative approach is an empirical analysis of the relationship between observed concentrations of oxidant precursors and meteorological

variables and the resulting observed distribution of oxidant concentration. The objectives of such an analysis would generally be more limited than in the development of detailed models arising from fundamental physical and chemical principles; however, an insight into the relationships which are observed, even for a limited range of conditions, can provide both guidance for setting control policy and guidance as to the dominant effects which should be considered in a chemical/physical model.

3.2 The Data

Any empirical analysis must proceed from a data base. For the analysis proposed, a great deal of data is available on the Los Angeles basin, particularly from the Los Angeles Air Pollution Control District (LAAPCD) and the California Air Resources Board (ARB).

There are at present approximately 30 stations monitoring air quality in the South Coast (Los Angeles area) Basin. Almost all of the stations monitor oxidant and NO_x. Several monitor HC.

The earliest station records date back to 1955, but few stations have such long histories. The early records are of somewhat doubtful value in some cases, due to changes in monitoring technology and standards. There are, however, more than 20 stations with histories of several years.

There are continuing questions about the comparability of data taken by different agencies. While such comparability (and indeed absolute accuracy) is critical for use with deterministic models, it is less important for statistical models as long as a consistent basis is used for each station reporting. The more recent data is generally characterized by such consistency, although data from the ARB may have to be

adjusted downward 20 to 25 percent to be consistent with LAAPCD data due to differing calibration techniques [7,8].

Mesometerological data, such as wind speed and direction, surface temperature, the vertical temperature profile, pressure, humidity, precipitation, visibility, and insolation, is collected at airports and other Weather Service stations and at meteorological stations run by other organizations, for example, Air Pollution Control Districts and the Armed Forces. Data from the Weather Service and Armed Forces stations are available from NOAA. Data from a sizable number of other meteorological stations in Los Angeles County is available from the LAAPCD.

An initial statistical analysis of LAAPCD data has been performed by Tiao, Box, et al. [9,10]. At present, only preliminary results have been reported.

3.3 The Problem Formulation

The general empirical approach is to postulate the possible independent variables which affect the dependent variable to be predicted. In the present case, the independent variables are measures of the precursor pollutant concentrations and of meteorological variables and the dependent variable is the oxidant concentration at a given location (or an aggregate measure such as peak oxidant concentration throughout the basin). (We refer for the sake of conciseness to variables such as averages or peaks which remove either a spatial or time variation from a given independent or dependent variable as "aggregate" variables.) This analysis has two major steps:

(1) Find the independent variables which best explain the behavior of the dependent variable; and

(2) Model that relationship mathematically.

While the second step generally receives the most attention in empirical analyses, the first step is the more difficult and often the more revealing. One can apply both linear and nonlinear models in both steps. If linear models are applied in the first step, the question answered will be whether the variables predict the dependent variable linearly. If general nonlinear forms are allowed, the question answered will be whether the independent variables predict dependent variables in either a linear or nonlinear manner.

An example of a good analysis of the linear dependence of ambient ozone on meteorological parameters was performed in research at Bell Laboratories [11]. There the logarithms of the meteorological variables were used to predict the logarithm of the oxidant concentration by a linear equation; this equation produced a correlation between predicted and actual concentration of 0.84. A single location in New York was analyzed, and the only variables used in attempting to predict the oxidant concentration were solar radiation, wind speed, and temperature. Mixing height was determined to offer no additional information beyond that of the three independent variables indicated. This analysis did not contain any dependence upon precursor pollutants since the oxidant concentration was specific to a certain location and the data was collected over a relatively short period of time; hence, the emissions might be expected to be relatively constant. The results, although highly encouraging in terms of the potential for empirical analysis, should be qualified:

(1) The model correlation was based upon a limited time period and one location and, while a very careful and credible statistical analysis of the prediction errors was made, no test was performed upon independent data not used in creating the linear equation.

(2) Since the errors in predicting the logarithm of the ozone value were found to be normally distributed, the error in predicting the ozone concentration itself tended to be largest at the higher values of ozone concentration. It is at the higher values where difficulty in forecasting is generally encountered but where accuracy of the model is most critical.

We note briefly another study as an example of the possibility of creativity in the definition of potential independent variables. Smith and Jeffrey attempted to predict, by a simple formula, the high concentration of sulfer dioxide in London and Manchester air[12]. One variable they found quite useful was the number of hours when the wind speed was less than three knots. This variable apparently summarized the key aspect of the temporal variation of wind speed as a single number. It is the intent of the proposed project to attempt to exercise limited creativity in potential predictors of oxidant concentration.

A key characteristic of the problem is that the level of ozone at a given time may be the result of precursor concentrations at a different point at an earlier time. (There is even some tentative empirical evidence that one day's hydrocarbons may be important in producing high levels of

oxidant on the following day [13]. The particular location and relevant time delay will be a function of meteorological parameters such as the wind field, temperature, solar radiation, and mixing height. There are several possible ways of handling this key problem:

(1) Stratify the data by general meteorological or wind-field classes, e.g., "a light wind from the ocean." Data for each class of wind field could then be analyzed to discover the location and time delay which best explain oxidant concentration at a given location. One could thus determine empirically the precursor location and time delays for the specific meteorological class.

(2) Aggregate precursor and oxidant values spatially and/or temporally. If the average hydrocarbon and NO_x concentrations across the basin for a given hour are considered independent variables and the peak oxidant reading throughout the basin for the day is considered a dependent variable, then one may seek a direct relationship between those variables. This is much the sort of aggregation attempted successfully in an earlier analysis of data produced by a photochemical smog model [5].

(3) Perform a trajectory analysis. Let the precursors of oxidant at a given location be the hydrocarbon and NO_x concentration in the parcel of air at its location at an earlier time as obtained through analysis of the trajectory of the

parcel. The independent variables in this case would be the precursor concentrations in the parcel three hours earlier, four hours earlier, etc. This approach has the advantage of making it unnecessary to specifically include the wind field as an independent variable but, instead, to use it in defining a more complex independent variable.

The third approach involves the interpolation of the wind field and the tracing of the trajectories. In a later section we will discuss objective interpolation of meteorological parameters and specifically interpolation of the wind field. Given a methodology for interpolating the wind field from a limited number of measurements, trajectories such as those in Figures 3-1 and 3-2 can be estimated. The precursor pollutant concentrations in the parcel of air at an earlier point can be estimated by interpolation of the concentrations of the precursor pollutants between measurement stations. The independent variable can then be the pollutant concentration in the parcel of air at an earlier point in time (or perhaps a weighted average of the earlier pollutant concentrations throughout the trajectory).

The end objective is to obtain functional relationship between a variable measuring the concentration of ozone and a limited number of meteorological and chemical precursor variables. The tool for the ultimate determination of the functional relationship will be nonlinear: continuous piecewise-linear regression, as used in an earlier study [5]. (A comparison between a nonlinear fit and a linear fit will be made to determine the degree of improvement obtained by allowing nonlinearity.)

ESTIMATED TRAJECTORY OF AIR ARRIVING AT PASADENA, EL MONTE, LONG BEACH, AND SANTA ANA AT 0400 SEPTEMBER 29, 1969.

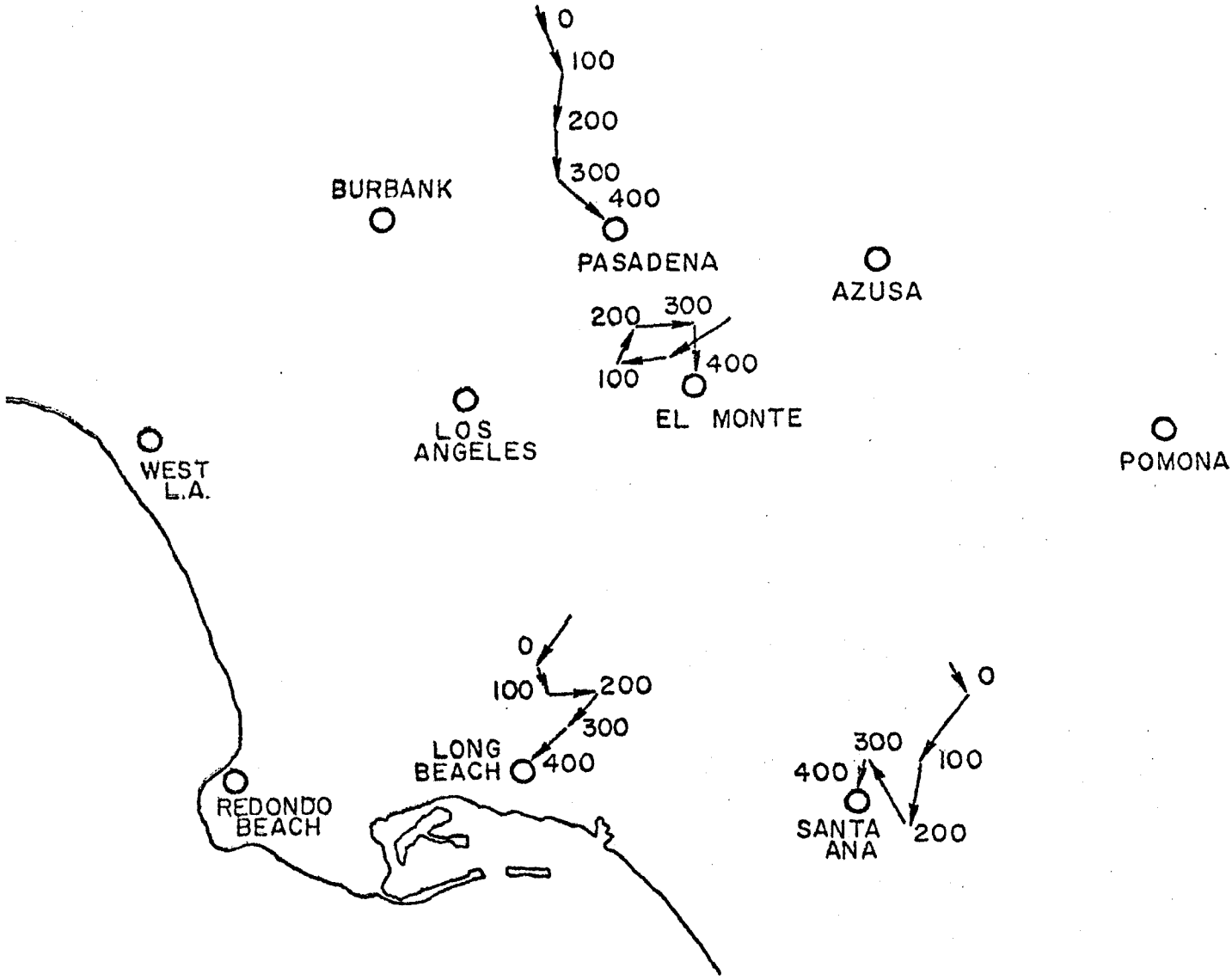


Figure 3-1: Trajectories

The figure shows that an irregular early morning meandering pattern exists at Long Beach, Santa Ana, and El Monte. Pasadena, on the other hand, shows a northerly flow pattern due to nocturnal air drainage down the mountains combined with an offshore wind flow. The lengths of the arrows give an indication of how much the air has moved during an hour interval. None of the stations show more than 4 m.p.h. air movement for the early morning hours.

ESTIMATED TRAJECTORY OF AIR ARRIVING AT PASADENA, EL MONTE, LONG BEACH, SANTA ANA, AND POMONA AT 1600 SEPTEMBER 29, 19

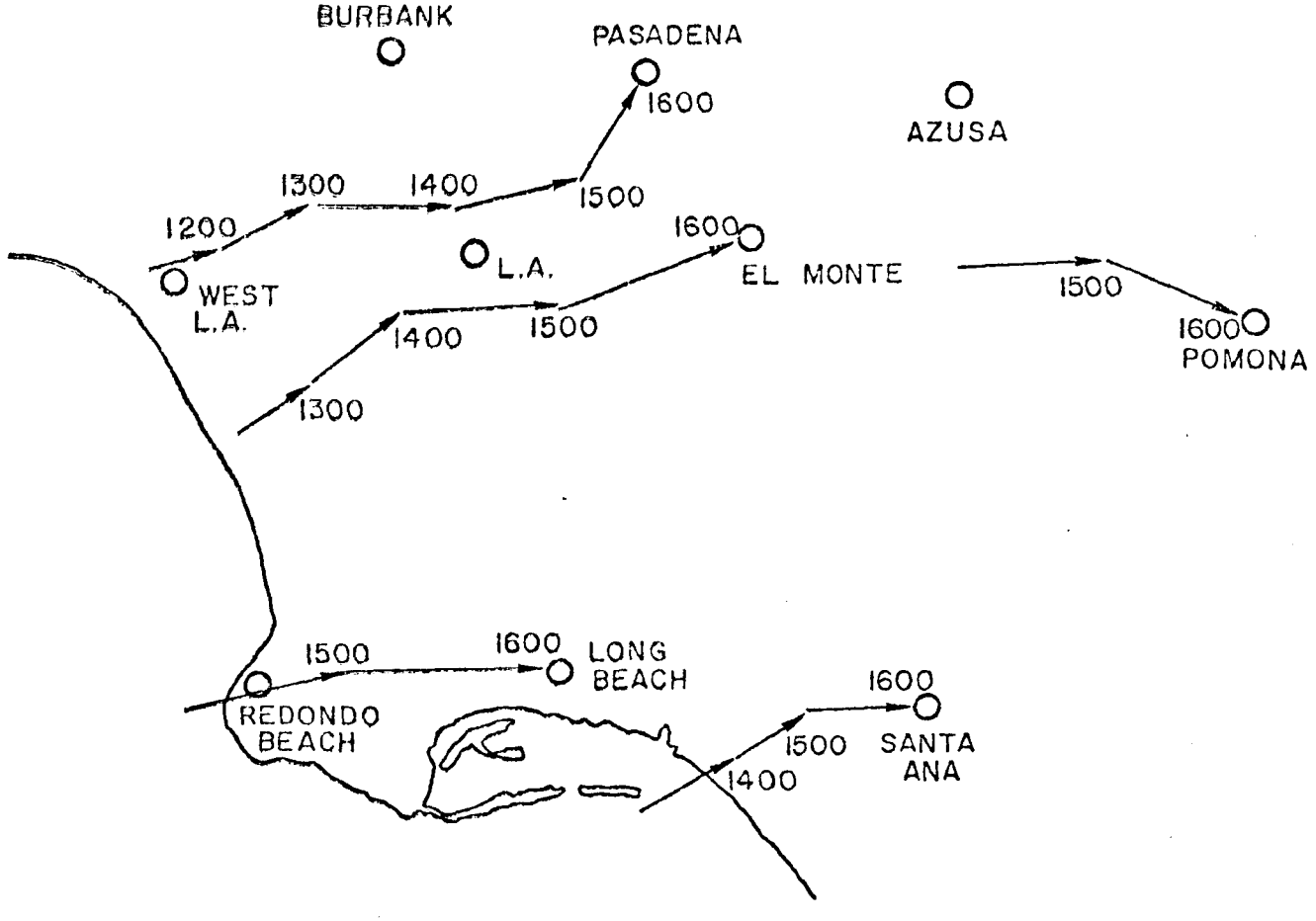


Figure 3-2: Trajectories

The figure shows the estimated air trajectories for the afternoon hours. All of the stations show the dominance of onshore sea breezes with a tendency of higher velocities later in the afternoon. The more regular air trajectories of afternoon also show a greater air movement than early in the morning as shown by the greater lengths of the arrows. The deflecting influence of the Santa Monica mountains causes the air trajectory to curve northward as it approaches Pasadena.

The technique used to determine which variables will be used in the equation will be a combination of linear and nonlinear variable selection techniques [14].

3.4 Research Plan

Task 1

Collect data from the California ARB and LAAPCD (and limited meteorological data from other sources) into a common format. The data will be limited to the South Coast Basin and the years 1968-1974.

Task 2

Analyze data as stratified by classes of wind field (and, perhaps, other meteorological variables). Data from days with meteorological conditions conducive to high oxidant concentration will be emphasized. Perform a linear and nonlinear analysis to determine appropriate independent variables. Examine the utility of using aggregate variables within the stratified classes.

Task 3

Examine the consistency of results obtained in Task 2 with those obtained from a trajectory analysis. It is understood that practical limitations on time and funds available will restrict the extensiveness of this analysis.

Task 4

Summarize the implications of the analysis and of the generality and validity of the models obtained.

The key difference between similar projects and the proposed project is the use of a class of powerful nonlinear techniques and the resulting generality of the conclusions.

4.0 EXTRACTION OF EMISSION TRENDS FROM AIR QUALITY TRENDS

4.1 Motivation

While measured pollutant concentration is the final impact of a given level of emissions, trends in pollutant concentration measurements can be misleading if it is assumed that those trends represent progress (or the lack thereof) in emission control. Since meteorology need not be uniform from time period to time period, the measure of progress should be more directly related to emissions. Emissions come from a large number of diverse sources, however, and are difficult to measure directly. Since air quality has been measured directly for a number of years, it is of significant interest to understand if the effect of meteorology can be removed from air quality trends to more nearly elicit trends in emissions. Such an analysis of trends is the subject of periodic reports both by the Council on Environmental Quality and by the Environmental Protection Agency.

Such a study must implicitly extract information about the influence of meteorological factors on pollution levels for a given level of emissions. This information can be an important subsidiary benefit of an analysis of the sort suggested.

We will discuss this concept by referring to a specific example of a study of the improvement in emissions between the early and late sixties in Oslo, Norway [15]. We will then relate this example to a general formulation to highlight the assumptions involved in such a study, to make the method more specific, and to provide a context for broader application of this approach.

4.2 Report of a Comparison of Emission Levels over Two Time Periods

A study of the changes in emission levels of SO_2 in Oslo, Norway, as deduced from changes in measured SO_2 concentrations, was undertaken to compare the SO_2 emissions of the periods 1959-1963 and 1969-1973. The meteorological conditions during the former period were considerably different from those during the latter period; hence, one could not expect a change in air quality to be directly related to a change in emissions.

Data from the earlier period (1959-1963) was used to do a linear regression analysis. It was discovered that two variables dominated the estimate of SO_2 concentration, a temperature difference between a low altitude and high altitude measuring station and the temperature at the lower station. For example, a typical regression equation for one station was

$$q_{SO_2} = 61.5\ (T_2-T_1) - 11.6T_1 + 472\ , \qquad (4\text{-}1)$$

where

q_{SO_2} = daily mean value of SO_2 concentration in $\mu g/m^3$ at the particular station

T_2 = temperature at higher station at 7 P.M.

T_1 = temperature at lower station at 7 P.M.

This equation explained the observed values of SO_2 concentration with a multiple correlation coefficient of .80; that is, the correlation between values predicted by this equation and observed values for the period indicated was 0.80. Adding other variables did not result in a significantly better predictor equation. It was suggested that the temperature difference term expressed the ventilation in the Oslo area while the temperature term measured the variation in the emission of SO_2 due to space heating. Since the temperature data for the later time period is known, the level of SO_2 expected for the meteorological conditions during that time period can be estimated by equation (4-1). This was done for the days on which data was available in the later time period; the results are indicated and compared with data from the earlier time period in Figure 4-1. The data from the 1959-1963 time period is scattered relatively uniformly about this line of slope 1--as expected, since the regression was performed on that data. However, the data from the later years evinces a much lower observed value of SO_2 concentration than would be expected from the meteorological conditions. The referenced study attributed this to a reduction in emissions.

Figure 4-1 indicates qualitatively the emission reduction (or, if the reader prefers, the "meteorologically normalized" reduction in pollutant levels). A quantitative statement was made in the report that the SO_2 pollution was reduced 50 to 60%. According to a conversation with one of the authors of the report, this latter statement was derived by looking at the ratio of the coefficient on the temperature difference term in the early time period to the ratio of the coefficient of the temperature difference term in a similarly derived equation for the later

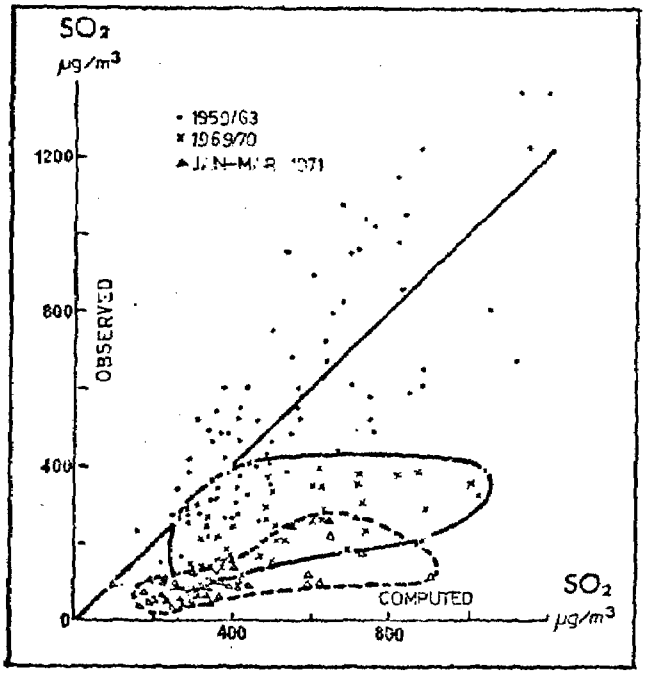


Figure 4-1: Values of daily mean SO_2 concentration computed from temperature measurements at 7 P.M. versus daily mean SO_2 concentration observed. The fact that the values in the later period are much less than would be expected from the meteorology suggests that emissions are less. Ref.[15]

time period. The intuitive justification for such a statement is that the coefficient measures the degree to which a given temperature inversion will be translated into SO_2 concentrations. Thus a 50 or 60% reduction in that coefficient might be thought of as a meteorologically adjusted measure of the trend in air quality. The intent was to obtain a value which can be interpreted as being proportional to the reduction in emissions.

4.3 Generalization and Mathematical Formulation

The purpose of the Oslo study was to compare air quality for two different periods rather than to obtain a continuous estimate of a meteorologically normalized air quality trend. We will formulate the problem in the former terms in order to relate it explicitly to that study; however, this does not at all imply that the approach cannot be modified to yield a continuous estimate of air quality trends. Assume we are given two sets of observations, one set for the first period of time:

$$\begin{matrix} q_1^{(1)} & , & \underline{m}_1^{(1)} \\ q_2^{(1)} & , & \underline{m}_2^{(1)} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ q_{N_1}^{(1)} & , & \underline{m}_{N_1}^{(1)} \end{matrix} \quad , \qquad (4\text{-}2)$$

where

$q_i^{(1)}$ = an air quality measurement during the first period (e.g., a daily mean value of pollutant concentration)

and

$\underline{m}_i^{(1)} = (m_{i1}, m_{i2}, \ldots, m_{in})$

= a vector of meteorological measurements corresponding to the i^{th} air quality measurement $q_i^{(1)}$ (e.g., m_{i1} might be a temperature measurement at a particular station).

There are a similar set of measurements for a later period:

$$\begin{array}{ll} q_1^{(2)} & , \ \underline{m}_1^{(2)} \\ q_2^{(2)} & , \ \underline{m}_2^{(2)} \\ \vdots & \vdots \\ q_{N_2}^{(2)} & , \ \underline{m}_{N_2}^{(2)} \end{array} \quad . \tag{4-3}$$

It is from this information (and without an estimate of emissions during the two periods) that we wish to determine a meteorologically adjusted estimate of the improvement or deterioration of air quality (i.e., to estimate the change in emissions from air quality and meteorological measurements). Suppose there is some "true," but unknown, equation (or model) which relates emissions and meteorological measurements to air quality:

$$q = F(\underline{e},\underline{m}) \quad . \tag{4-4}$$

This equation plus measurement error produced the measurement data of (4-2) and (4-3). We are assuming that the equation does not differ between the two periods, that any changes in air quality are explained either by a change in meteorology or a change in emissions.

For the sake of the present discussion, let us again assume that emissions remain essentially constant over the first time period and over the second time period:

$$\underline{e} = \underline{e}_1 \quad \text{in first period,} \tag{4-5a}$$

$$\underline{e} = \underline{e}_2 \quad \text{in second period .} \tag{4-5b}$$

Now let us suppose that we do a linear or nonlinear regression with the data from the first period, equation (4-2), and obtain a best fit equation to the data:

$$q = f_1(\underline{m}) \quad . \tag{4-6}$$

Equation (4-1) is such an equation.

Since (4-6) was derived with constant emissions $\underline{e}_1$, and since "truth" is assumed to be given by equation (4-4), f_1 represents the relation between meteorological conditions and pollutant concentration for fixed emissions $\underline{e}_1$:

$$f_1(\underline{m}) \approx F(\underline{e}_1,\underline{m}) \quad , \tag{4-7}$$

Now suppose we use the data of the second period, equation (4-3), to obtain a similar empirical model:

$$q = f_2(\underline{m}) \quad . \tag{4-8}$$

Then, as before,

$$f_2(\underline{m}) \approx F(\underline{e}_2, \underline{m}) \quad . \tag{4-9}$$

Let us further assume that F is decomposible:

$$q = F(\underline{e}, \underline{m}) = G(\underline{e})H(\underline{m}) \quad . \tag{4-10}$$

Equation (4-10) implies that the effect of emissions on air quality is essentially independent of the effect of meteorology. The appropriateness of this assumption clearly depends upon the particular definitions of the emission, meteorological, and pollutant variables, as well as the area in question. If the pollutant concentration is location-specific (rather than a spatial average or spatial maximum), then either emissions must be spatially uniform or the direction of the wind field relatively consistent for (4-10) to be reasonable. (The latter seems to be the assumption of the Norwegian study.) If the variables are aggregates (such as spatially averaged SO_2 concentrations, total emissions, and average wind speed), then less severe assumptions need be made for (4-10) to be reasonable.

Given (4-10), the ratio of the empirical equations for the two time periods is

$$\frac{f_2(\underline{m})}{f_1(\underline{m})} = \frac{F(\underline{e}_2,\underline{m})}{F(\underline{e}_1,\underline{m})} = \frac{G(\underline{e}_2)}{G(\underline{e}_1)} \quad , \tag{4-11}$$

using (4-6), (4-7), and (4-8). Thus, the ratio of the two equations should be very nearly constant if (4-10) is valid, and that constant will be a measure of the change in emissions between the two periods. (The function $G(\underline{e})$ can be, for example, total emissions in tons.)

If (4-11) is not nearly constant, it can be interpreted as implying that the improvement is a function of the meteorology. This might easily be the case. For example, if there is substantial reduction in industrial emissions but no improvement in emissions from space heating, the improvement in emissions will be less when the temperature is lower. If the improvement is a function of wind direction, the location of major emission sources may be the cause. In the Oslo study, the ratio of the temperature difference terms alone was taken and is exactly constant. Since the full Oslo model, (4-1), contains other terms, however, the ratio suggested by this discussion is not constant. Since the equation for the later time period was not explicitly reported, we cannot calculate the ratio. Let us examine, however, an analysis which is consistent with Figure 4-1 and which provides an alternative approach.

Suppose we create a model f_1 for the first time period only and apply it to the meteorological conditions for the second time period:

$$\hat{q}_1^{(2)} = f_1(\underline{m}_1^{(2)})$$
$$\hat{q}_2^{(2)} = f_1(\underline{m}_2^{(2)})$$
$$\vdots$$
$$\hat{q}_{N_2}^{(2)} = f_1(\underline{m}_{N_2}^{(2)}) \quad . \tag{4-12}$$

We obtain estimates for the air quality $\hat{q}_i^{(2)}$ to be expected if the emissions have not changed; these calculated values can be compared with observed values. These are the values plotted in Figure 4-1. If we now perform a linear regression of observed versus estimated values, i.e., $q_i^{(2)}$ versus $\hat{q}_i^{(2)}$ for $i=1,2,\ldots,N_2$, we obtain a regression equation:

$$q = a\,\hat{q} + b \quad , \tag{4-13}$$

with specific values of a and b. Suppose we then assume that the "true" equation is of the form

$$q = F(\underline{e},\underline{m}) = G(e)H(\underline{m}) + q_o \quad , \tag{4-14}$$

where

q_o = a "background" air quality level not related to local emissions.

Then (4-13) is consistent with (4-14) if

$$a = \frac{G(\underline{e}_2)}{G(\underline{e}_1)} \tag{4-15a}$$

and

$$b = q_o^{(2)} - a\, q_o^{(1)} \quad . \tag{4-15b}$$

Then "a" can be interpreted as the increase in emissions and, more controversially, "b" can be related to the change in "background" level (where the background level may contain contributions from sources outside the emissions inventory included in e--for example, long-range transport from other cities).

Estimating the best-linear-fit equations graphically, from Figure 4-1, we find that the equation for the 1969/70 data is approximately

$$q = 0.25\, \hat{q} + 120 \tag{4-16a}$$

and for the 1971 data

$$q = 0.25\, \hat{q} \quad . \tag{4-16b}$$

Thus, the reduction in emissions is about 75% by this analysis for both periods. The 1969/70 period had higher "background" than the 1959/63 period by 120 $\mu g/m^3$, but the 1971 period had about the same background as 1959/63. Thus, the improvement between 1969/70 and 1971 could be attributed to improvements in areas other than Oslo.

Note that this latter approach requires that only one model be created. Since the approach is symmetrical, the model can be created for the period in which the most data is available and applied to the other period.

5.0 DETECTION OF INCONSISTENCIES IN AIR QUALITY/METEOROLOGICAL DATA BASES

5.1 Motivation

Air quality and meteorological data bases are collected for many purposes (and often used for purposes not intended when collected). An important objective either during collection or after the fact is the detection of inconsistencies in the data. In most data collection efforts, an attempt is made to study the data for strange behavior or to employ intuition and problem knowledge to uncover sources of system changes causing data problems, such as changes or discrepancies in monitoring techniques. A recent example is the detection of a significant discrepancy in certain calibration techniques used by the California Air Resources Board and the Los Angeles Air Pollution Control District, making oxidant measurements of the agencies inconsistent without a correction factor [8]. The frequent occurrence of detected inconsistencies in data bases leads one to expect the possibility of undetected inconsistencies. An automatic technique for flagging potential inconsistencies using the data itself would be an important tool. Such a technique would take an existing data base and detect potential problems for closer inspection or detect problems occurring in an ongoing data collection effort before a substantial amount of data was irretrievably lost.

In this section, we will indicate how data-analytic/statistical techniques can be employed to achieve this objective, we will distinguish the types of inconsistencies for which one might search, the appropriate approaches to detecting these various types of inconsistencies, and the

potential difficulties in this formal approach to the detection of inconsistencies.

The key concept will be that of using the data collected to form a model of the relationship between selected sets of measurements and to automatically detect the measurements or points in time when (1) the model changes or (2) the data is least consistent with the model. Note that the model need not be a prediction model or relate independent to dependent variables. Any consistent relationships in the data can be employed in detecting inconsistencies.

It is important to distinguish inconsistencies from extremes. An extreme value of air pollution is not necessarily inconsistent--it may be consistent with extreme meteorological conditions. If the model adequately incorporates the extreme conditions, the extreme values would be indicated as being consistent and not flagged. If, however, the extreme conditions were not previously observed in the data base or not otherwise represented by a similar condition in the data base, the extreme conditions may not be incorporated in the model and may be flagged as possible inconsistencies. We bring up these points to emphasize two key concepts: (1) the intent of a consistency analysis is not to flag simple extreme values but to flag values which are inconsistent, i.e., extreme and inconsistent values are not equivalent; (2) the intent of a consistency analysis is to flag <u>potential</u> inconsistencies for inspection. An inconsistency analysis will be successful if it does not miss key inconsistencies that could seriously damage an empirical analysis or data

collection effort. It will not have failed if it also flags potential inconsistencies which upon further examination are more accurately categorized as extremes or unusual occurrences.

Let us structure these ideas more formally.

5.2 Formulation of Consistency Models

We imagine the basic situation of the simultaneous collection of air quality and meteorological data, as well as possible adjunct data depending upon the application (e.g., health effects data, emissions data, etc.). Suppose the basic data is a sequence of measurements over time of a number of variables:

$$\begin{aligned}
&\text{Measurement 1:} \quad x_1(t_1),\ x_1(t_2),\ldots,\ x_1(t_N) \ , \\
&\text{Measurement 2:} \quad x_2(t_1),\ x_2(t_2),\ldots,\ x_2(t_N) \ , \\
&\qquad\qquad \vdots \\
&\text{Measurement n:} \quad x_n(t_1),\ x_n(t_2),\ldots,\ x_n(t_N) \ . \qquad (5\text{-}1)
\end{aligned}$$

There are three basic formulations of consistency models available.

Time Sequence Inconsistencies

The consistency of individual time series can be examined. The model constructed can be a model which predicts the value at a given point in time from past and future values of itself. An inconsistency will then be detected as a significant discrepancy between the forecast and observed value. That is, the model could be of the form

$$\hat{x}_i(t_j) = F[x_i(t_1),\ldots,\ x_i(t_{j-1}),\ x_i(t_{j+1}),\ldots,\ x_i(t_N)] \ , \qquad (5\text{-}2)$$

where $\hat{x}_i(t_j)$ is the value of $x_i(t_j)$ predicted by the model. We emphasize

that since we are testing consistency rather than predicting behavior, values occurring after the particular value tested can be used in the model when available. While many time series techniques employ recursively expressed predictor models, they imply a general dependence of the form indicated.

An inconsistency would be a sufficiently large deviation between predicted and measured values, i.e., a large value of

$$|x_i(t_j) - \hat{x}_i(t_j)| \quad . \tag{5-3}$$

Cross Measurement Inconsistencies

This type of model is constructed by modeling the relationships between measurements at a given point in time. An example is a derived relationship between a vertical temperature difference and average wind speed at the same time. Formally, such a model is of the form

$$\hat{x}_i(t_j) = G[x_1(t_j),\ldots, x_{i-1}(t_j), x_{i+1}(t_j),\ldots, x_n(t_j)] \quad . \tag{5-4}$$

An inconsistency would be detected by large values of (5-3), as before.

Combined Model

In general, measurements will depend upon both past history and concurrent measurements. A full model would then be a technique which used data both at other times and from other variables:

$$\begin{aligned} x_i(t_j) = H[&x_1(t_1),\ldots, x_1(t_N);\cdots; x_i(t_1),\ldots, \\ &x_i(t_{j-1}), x_i(t_{j+1}),\ldots, x_i(t_N);\cdots; \\ &x_n(t_1),\ldots, x_n(t_N)] \quad . \end{aligned} \tag{5-5}$$

Note that in many cases it is neither easy nor important to categorize the type of modeling being employed. It might be unclear for example what category one should place a model where the time slice was fairly broad, for example, where monthly averages of daily values were compared to one another. If the daily values are considered the basic data, then the model is a combined model; if the monthly averages are considered the basic data, then the model is a cross-measurement model. It is clearly less important to categorize a model than to create and use it appropriately.

5.3 Types of Inconsistencies

There are several types of inconsistencies one might be interested in detecting in the data:

1. Abrupt, but persistent, changes;
2. Slow nonstationarities; and
3. Anomalous data (abrupt, nonpersistent changes).

Let us discuss these categories of problem and formulation of models for their solution.

Abrupt, Persistent Changes

The change in the data may occur suddenly in time, i.e., at an identifiable point in time.

There are generally two types of abrupt, persistent changes of interest:

1. Malfunctioning measurement or recording devices - If a measuring device suddenly begins to malfunction, it will generally continue to malfunction until repaired or replaced. The motivation for detecting such a problem is obvious. In the present categorization, we intend to mean by an abrupt, persistent change a change

in the underlying model which occurs over a relatively short period of time. This is as distinguished from slow changes or short-term changes.

2. Changes in the system - We refer to major changes in the system which occur over a short period of time such as the opening of a new freeway or the opening of a major indirect source. As well as permanent changes, there may be temporary but significant changes, such as if a city were to host the Olympic Games. Without specific attention to such events, the conclusions of an analysis could be misleading. The analysis of this type of abrupt change has been called "intervention analysis" by Box and Tiao [16].

There is also clearly a matter of degree. An event can have a relatively mild effect, as might the closing of several on-ramps to a freeway. One output of a consistency analysis should be a measurement of the degree of inconsistency.

This category of inconsistency has the basic character of having a significantly different relationship between variables in the time periods before the event and after the event. The point in time separating the two periods is assumed unknown (since the purpose of a consistency analysis is to discover such points).

The first of two basic technical approaches to this problem consists creating a series of models and searching for a statistically significant change in model structure or parameters. One may create a model over

the interval $[t_1,\ldots,t_k]$ and predict $x_j(t_{k+1})$. If the prediction is consistent with observation, then a model over $[t_1,\ldots,t_{k+1}]$ is created to predict $x_j(t_{k+2})$, and so on, until a discrepancy occurs. A simple modeling technique or recursive procedure is probably a requirement if a high computing cost is to be avoided.

The second approach does not require as abrupt a change as the first but may be more computational. Here, one can create two models, one for the period $[t_1,t_k]$ and one for the period $[t_k,t_N]$. One can calculate an appropriate measure of the difference in the models, say D_k. Repeating this for varying breakpoints t_k, one can determine the value at which the difference D_k is maximized, presumably the point when the change occurred.

Slow Nonstationarities

Many types of change will occur gradually over a period of time. For example, the retrofitting of emission control devices in automobiles in California was mandated by law to occur in a month-by-month fashion depending upon the digit of the car owner's license plate. The slow introduction of the retrofitting might affect the time sequence of air quality measurements. Another example is a slow but significant drift in a measuring instrument. Such an inconsistency would be detected as a systematic change in the appropriate model over time as opposed to an abrupt inconsistency.

As with abrupt changes, categorizations of slow nonstationarities are possible. They may be related both to measurement device drift or

to changes in the system, and they may be both temporary and permanent. (An example of a temporary but slow nonstationarity is a slow but definite degradation in the degree of compliance with the 55-miles-per-hour speed limit.)

The most straightforward approach to this problem is to postulate the form of the nonstationarity and test for it. For example, two air-quality monitoring stations near each other might measure the same pollutant, recording $x_1(t)$ and $x_2(t)$, respectively. One could then do a linear regression of day-to-day changes of the stations against one another, i.e., find the best-fit linear relationship between

$$\nabla_1(t_k) = x_1(t_k) - x_1(t_{k-1})$$

and

$$\nabla_2(t_k) = x_2(t_k) - x_2(t_{k-1})$$

for k=2,3,...,N. The result will be of the form

$$\nabla_1 = a\nabla_2 + b \quad .$$

One can then test statistically whether b is significantly different than zero. If it is, the values measured by one station are drifting relative to the other. Unless this can be explained by a constantly increasing (or decreasing) emission source affecting one of the stations selectively, it is an inconsistency.

Another approach is to compare a model created on $[t_1,t_k]$ with a model created on $[t_{k+\delta},t_N]$, where the time gap δ between periods modeled is sufficient to detect a slow drift. This approach requires fewer assumptions regarding the form of a possible nonstationarity.

Anomalous Data

This type of inconsistency might be categorized as a "noisy" measurement. It could be caused by erroneous recording or digitization of the data by a temporarily malfunctioning instrument or by an anomalous occurrence such as might be caused by sidewalk repairs raising dust near a site monitoring suspended particulate levels. Such an occurrence is a short-term abrupt inconsistency in either a time sequence or cross-measurement model. It is a relatively conventional type of problem encountered in data analysis and is often referred to as "outlier analysis."

This problem can be approached in the single variable case by studying extreme values detected by creating a histogram (the empirical distribution) of measured values. The more variables measured, the greater the potential for outliers which are not obvious by looking at individual variables. (The classical example is the existence of a "pregnant male" in a medical data base; neither "pregnant" nor "male" is illegal, only the combination.) In the multivariate case, the most general class of techniques for detecting outliers is "cluster analysis"[17]. Very small clusters of points or single-point clusters in multivariate space are inconsistencies which should be examined.

5.4 Difficulties

The major technical difficulties in consistency analysis are, first, nonlinearities and secondly, lack of data relative to the number of variables the relation of which is to be modeled. Most air quality and meteorological parameters are nonlinearly related. Further, it often takes a large number of variables to determine with accuracy other meteorological or air-quality variables. This means that the diversity of joint observations of values of a large number of variables that one can expect in a given data base or at the start of a measurement program is limited. Compounding the problem, nonlinear models will, in general, require more parameters than linear models and, hence, require more data for accurate model determination.

These problems can be alleviated by both technical and operational solutions. A technical consideration is that an efficient (low-parameter) nonlinear form will require less data for the determination of the model than an inefficient (overparameterized) nonlinear form; hence, efficient functional forms, such as continuous piecewise linear functions, can help alleviate this problem. A second technical point is that a set of models of relatively simple form can be created with subsets of the relevant variables.

The operational consideration is the fact that one may operationally be able to tolerate a high level of "false alarms" in detecting inconsistencies at the beginning of a data collection project or in analyzing a data base in the initial stages. It is at this early point in most data

collection or data analysis efforts that most of the problems are encountered. As more data is collected, the model will become more refined and flag fewer potential inconsistencies.

Another possible problem is the inclusion of inconsistencies into the model. Without care, the data can be modeled including inconsistencies in such a way that the inconsistencies are fitted and do not become apparent as a discrepancy in the model. This pitfall can be avoided by simply employing good data-analytic practices to avoid overfitting.

For many projects in data collection and analysis, the use of conventional tools in a careful manner can provide a systematic analysis of consistency which may avoid erroneous analyses and a great deal of wasted effort.

6.0 REPRO-MODELING: EMPIRICAL APPROACHES TO THE UNDERSTANDING AND EFFICIENT USE OF COMPLEX AIR QUALITY MODELS

Several computer-based mathematical models derived from basic physical principles have been constructed to model air pollution and meteorological phenomena. The diversity of inputs to such models and the typically long running times often make it difficult to understand the full implications of the models or to use the models in certain planning applications where large numbers of alternatives must be rapidly evaluated. The concept of "repro-modeling" is to treat a model as a source of data for an empirical analysis [18]. Such an analysis will, in general, have two major objectives:

1. To understand the implication of the model by discovering which variables most affect the outputs of interest and in what way they affect the outputs of interest; and
2. To construct as a simple functional form a model of the relationship between key independent variables and key model outputs.

Since this approach has been a subject of a previous EPA contract, in which the technique of repro-modeling was applied to a reactive dispersive model of photochemical pollutant behavior in the Los Angeles basin [5], we will not discuss it in further depth in this report. We do wish to emphasize the role of such an analysis in evaluating and validating models, as well as in suggesting to modelers the characteristics which a current version of the model implies which might bear further investigation.

One point in earlier discussions of repro-modeling which has not been emphasized is its use in model validation and sensitivity analysis. Often sensitivity analysis is performed on models in order to determine which parameters of the model are most critical in determining the model output [19]. The change in model output with a small change in a given parameter or input value is the sensitivity of the model to that parameter. Since the sensitivity of a model to a particular parameter will, in general, depend upon the values of the other parameters, classical sensitivity analysis is usually performed in one of two ways:

1. One set of typical values for the parameters and inputs is chosen and the effect of small changes in the parameters about that nominal condition are made in order to examine sensitivity. This obviously indicates only the sensitivity at the particular nominal condition chosen.
2. A "factorial" analysis is performed, where a number of diverse nominal values are chosen and the above analysis repeated for this large number of diverse conditions. This exercises the full range of potential operation of the model, but creates the problem of commensurating the implications of what are often thousands of model runs. It also has the obvious disadvantage of requiring a large number of model runs.

If one is willing to perform a given number of model runs to get a number of nominal points for a sensitivity analysis, it is more efficient, rather than to do a sensitivity analysis at each point, to fit the points with an appropriate functional form such as a continuous

piecewise linear form [5]. As demonstrated in the referenced report, this results in regimes in which the model output is a linear function of the model inputs and/or parameters and the sensitivity to those parameters and inputs is quite clearly displayed. This approach automatically determines those regimes in which the sensitivity is relatively constant over a large area of parameter/input variations. This "global" sensitivity analysis approach can be more easily interpreted and more efficient than a "local" sensitivity analysis approach.

7.0 OTHER APPLICATION AREAS

Three additional topics are treated briefly here. The brevity is not related to a judgment of importance, but simply to the limited nature of the remarks.

7.1 Spatial Interpolation of Meteorological and Air Quality Measurements

Several recent studies have adopted a simple interpolation formula to construct continuous wind fields. (The approach is applicable to the interpolation of other quantities as well.) This formula includes every monitoring station with the weight of each measurement inversely proportional to the distance to the monitoring station location raised to a power. More explicitly, the interpolation formula is

$$V_j = \frac{\sum_{i=1}^{n} \frac{v_i}{R_{ij}^{\alpha}}}{\sum_{i=1}^{n} \frac{1}{R_{ij}^{\alpha}}} \qquad (7\text{-}1)$$

where there are n measurements within a prespecified distance of the point and where v_i is the measurement at the i^{th} location; R_{ij} is the distance between point i and j, and α is the exponent. This formula is applied separately to each vector component of the wind vector and the two resulting estimates are combined to recover an interpolated wind speed and direction. The value of α has been chosen to be either 1 or 2 in previous studies.

This approach is closely related to some recent Russian work [20,21,22] and work by M. Rosenblatt [23]. In these papers the concept of nonlinear regression is explored by means of kernel functions and density estimates. The use of these methods in the wind field problem would involve estimates of the type

$$\underline{V}(\underline{x}) = \sum_{i=1}^{N} \underline{v}_i \; K_i(\underline{x}-\underline{x}_i) \qquad (7\text{-}2)$$

where $\underline{x}_i$ is the location of the i^{th} station and $\underline{v}_i$, the measured wind vector at the i^{th} station. The kernel functions $K_i(\underline{y})$ have generally been taken, in the statistical literature, to be the same for all i and generally to be a smooth Gaussian type function with the sharpness of its peak determined by a shape parameter σ. Equation (7-1) fits the formulation using instead an inverse-distance kernel function with shape parameter α. The referenced papers and on-going research in probability density estimation are thus relevant to a deeper understanding of the implications of using (7-1) and to the development of alternative approaches.

7.2 Health Effects of Air Pollution

Empirical approaches (in particular, linear and nonlinear regression techniques) have been employed in estimating the effects of air pollution levels on health. The main difficulty encountered in this type of analysis is that of determining an incremental effect on respiratory health

measurements which are often dominated by vagaries of general health problems such as flu epidemics or of individual differences such as the habit of smoking or occupational environment. Yet, very strong relations must be derived if causal effects are implied. In such conditions, the best hope for improvement is in more highly controlled data collection efforts (which are, however, very expensive).

This situation highlights an important aspect of data analysis projects: A legitimate result of the analysis is a negative conclusion, a conclusion that the data does not admit of reliable results. A negative result is constructive to the degree that it makes the strong statement that the information desired is not present in the data; this settles the matter unless the data base is augmented. A less conclusive culmination of a data analysis effort is a limited negative statement, for example, a conclusion that no linear function of the independent variables predicts the desired variable with statistically significant accuracy.

We note, however, that a negative conclusion does not necessarily imply a faulty data collection effort; it may instead imply that the relationship of interest is less pronounced than initially expected relative to the effect of uncontrolled (or unmeasured) variables. Unfortunately, a well-conceived data analysis or collection effort is often labeled a failure when only negative results are produced--a charge which implies that the knowledge which the study was designed to elicit should have been obvious before the data was collected.

7.3 Short-term Forecasting of Pollutant Levels

The forecasting of pollutant levels the next day is of importance for health warning systems and/or to initiate short-term control procedures.

Forecasting pollution levels and forecasting the weather are closely related problems; it is not clear which is the most difficult, but certainly neither is easy. The empirical approach attempts to model directly the relation implicit in measured meteorological and air-quality data.

Persistence (i.e., assuming tomorrow's peak pollutant concentration equals today's peak concentration) usually proves a reliable forecast at lower pollution levels, but not necessarily at high levels when accuracy is most critical [13]. Certainly persistence will not predict a high pollutant level on a day following a low-pollutant level. Regression or time-series approaches tend to exploit persistence and may not be best suited to a situation where the determinants of the future pollution level can be considerably different depending on the level. Further, the performance estimate can be misleadingly high due to the number of low or intermediate pollution days usually included in the analysis.

Classification analysis is probably a more natural approach to the problem. The joint distribution of attributes (i.e., descriptive variables) of high-pollution days can be derived by looking at high-pollution days alone and can be compared to the joint distribution of attributes of intermediate days and to the joint distribution of attributes of low-pollution days. The variables of importance in distinguishing the 3 classes can be determined, and an algorithm to predict the classes can be derived.

8.0 REFERENCES

1. Meisel, W. S., "Empirical Approaches to Air Quality and Meteorological Modeling," Proc. of Expert Panel on Air Pollution Modeling, NATO Committee on Crises in Modern Society, Risø, Denmark, June 6, 1974. (This document may be obtained from the Air Pollution Technical Information Center, Office of Air and Water Programs, Environmental Protection Agency, Research Triangle Park, North Carolina 27711.)

2. Calder, K. L., "Some Miscellaneous Aspects of Current Urban Pollution Models," Proc. Symp. on Multiple Source Urban Diffusion Models, EPA, Research Triangle Park, North Carolina, 1970.

3. Hrenko, J. M., and D. B. Turner, "RAM: Real-Time Air-Quality Simulation Model," EPA, Research Triangle Park, North Carolina (Preliminary draft, July 12, 1974).

4. Calder, K. L., "The Feasibility of Formulation of a Source-Oriented Air Quality Simulation Model that Uses Atmospheric Dispersion Functions Empirically Derived from Joint Historical Data for Air Quality and Pollutant Emissions," EPA, Research Triangle Park, North Carolina (draft, August 1974).

5. Horowitz, Alan, and W. S. Meisel, "The Application of Repro-Modeling to the Analysis of a Photochemical Air Pollution Model," EPA Report No. EPA-650-74-001, NERC, Research Triangle Park, North Carolina, December 1973.

6. Calder, K. L., "A Narrow Plume Simplification for Multiple Source Urban Pollution Models" (informal unpublished note), December 31, 1969.

7. "ARB Oxidant Readings to Be Adjusted Downward," Calif. ARB Bulletin, Vol. 5, No. 8 (September 1974), pp 1-2.

8. "Calibration Report: LAAPCD Method More Accurate; ARB More Precise," Calif. Air Resources Board Bulletin, Vol. 5, No. 11 (December 1974), pp 1-2.

9. Tiao, G. C., G. E. P. Box, and W. J. Hamming, "Analysis of Los Angeles Photochemical Smog Data: A Statistical Overview," Technical Rept. No. 331, Dept. of Statistics, U. of Wisconsin, April 1973.

10. Tiao, G. C., et al., "Los Angeles Aerometric Ozone Data 1955-1972," Technical Rept. No. 346, Dept. of Statistics, U. of Wisconsin, October 1973.

REFERENCES (CONT'D)

11. Bruntz, S. M., W. S. Cleveland, B. Kleiner and J. L. Warner, "The Dependence of Ambient Ozone on Solar Radiation, Wind, Temperature, and Mixing Height," Proc. Symp. on Atmospheric Diffusion and Air Pollution, Santa Barbara, Calif., September 9-13, 1974, American Meteorological Society, Boston, Mass.

12. Smith, F. B., and G. H. Jeffrey, "The Prediction of High Concentrations of Sulfphur Dioxide in London and Manchester Air," Proc. 3rd Meeting of NATO/CCMS Expert Panel on Air Pollution Modeling, Paris, October 2-3, 1972.

13. Horowitz, A. J., and W. S. Meisel, "On-time Series Models in the Short-term Forecasting of Air Pollution Concentrations," Technology Service Corporation Report No. TSC-74-DS-101, Santa Monica, Calif., August 22, 1974.

14. Breiman, Leo, and W. S. Meisel, "General Estimates of the Intrinsic Variability of Data in Nonlinear Regression Models," TSC Report, Technology Service Corp., Santa Monica, Calif., October 1974.

15. Gronskei, K. E., E. Joranger and F. Gram, "Assessment of Air Quality in Oslo, Norway," Published as Appendix D to the NATO/CCMS Air Pollution Document "Guidelines to Assessment of Air Quality (Revised) SO_x, TSP, CO, HC, NO_x Oxidants," Norwegian Institute for Air Research, Kjeller, Norway, February 1973. (This document may be obtained from the Air Pollution Technical Information Center, Office of Air and Water Programs, Environmental Protection Agency, Research Triangle Park, North Carolina.)

16. Box, G.E.P., and G. C. Tiao, "Intervention Analysis with Applications to Economic and Environmental Problems," Technical Report NO. 335, Department of Statistics, University of Wisconsin, Madison, Oct. 1973.

17. "Cluster Analysis," Chapter VIII of W. S. Meisel, Computer-Oriented Approaches to Pattern Recognition, Academic Press, 1972.

18. Meisel, William S., and D. C. Collins, "Repro-Modeling: An Approach to Efficient Model Utilization and Interpretation," IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-3, No. 4, July 1973, pp 349-358.

19. Thayer, S.D., and R.C. Koch, "Sensitivity Analysis of the Multiple-Source Gaussian Plume Urban Diffusion Model," Preprint volume, Conference on Urban Environment, October 31-Nov. 2, 1972, Philadelphia, Pennsylvania (published by American Meteorological Society, Boston, Mass.).

REFERENCES (CONT'D)

20. Nadaraya, E.A., "On Estimating Regression," Theor. Probability Appl., Vol. 4, pp 141-142, 1965.

21. Nadaraya, E.A., "On Non-parametric Estimates of Density Functions and Regression Curves," Theor. Probability Appl., Vol. 5, pp 186-190, 1965.

22. Nadaraya, E.A., "Remarks on Non-parametric Estimates for Density Functions and Regression Curves," Theor. Probability Appl., Vol. 15, pp 134-137, 1970.

23. Rosenblatt, M., "Conditional Probability Density and Regression Estimators," Multivariate Analysis, Vol. II, pp 25-31, Academic Press, New York, 1969.