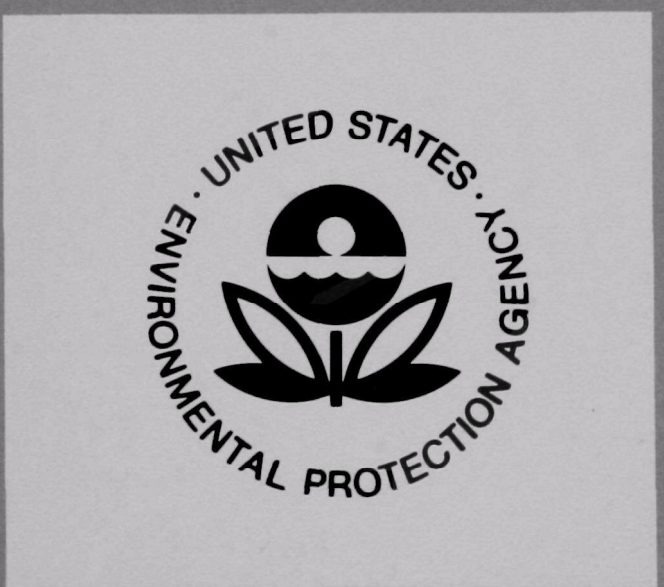
EPA-600/4-76-029a
July 1976

Environmental Monitoring Series

EMPIRICAL TECHNIQUES FOR ANALYZING AIR QUALITY AND METEOROLOGICAL DATA
Part I. The Role of Empirical Methods in Air Quality and Meteorological Analyses



Environmental Sciences Research Laboratory
Office of Research and Development
U.S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

RESEARCH REPORTING SERIES

Research reports of the Office of Research and Development, U.S. Environmental Protection Agency, have been grouped into five series. These five broad categories were established to facilitate further development and application of environmental technology. Elimination of traditional grouping was consciously planned to foster technology transfer and a maximum interface in related fields. The five series are:

1. Environmental Health Effects Research
2. Environmental Protection Technology
3. Ecological Research
4. Environmental Monitoring
5. Socioeconomic Environmental Studies

This report has been assigned to the ENVIRONMENTAL MONITORING series. This series describes research conducted to develop new or improved methods and instrumentation for the identification and quantification of environmental pollutants at the lowest conceivably significant concentrations. It also includes studies to determine the ambient concentrations of pollutants in the environment and/or the variance of pollutants as a function of time or meteorological factors.

This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

EPA-600/4-76-029a
July 1976

EMPIRICAL TECHNIQUES FOR ANALYZING AIR QUALITY AND METEOROLOGICAL DATA
Part I. The Role of Empirical Methods in Air Quality and Meteorological Analyses

by

W. S. Meisel
Technology Service Corporation
2811 Wilshire Boulevard
Santa Monica, California 90403

Contract No. 68-02-1704

Project Officer

Kenneth L. Calder
Meteorology and Assessment Division
Environmental Sciences Research Laboratory
Research Triangle Park, North Carolina 27711

U.S. ENVIRONMENTAL PROTECTION AGENCY
OFFICE OF RESEARCH AND DEVELOPMENT
ENVIRONMENTAL SCIENCES RESEARCH LABORATORY
RESEARCH TRIANGLE PARK, NORTH CAROLINA 27711

DISCLAIMER

This report has been reviewed by the Environmental Sciences Research Laboratory, U.S. Environmental Protection Agency, and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the U.S. Environmental Protection Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

ABSTRACT

Empirical methods have found limited application in air quality and meteorological analyses, largely because of a lack of good data and the large number of variables in most applications. More and better data, along with advances in methodology, have broadened the applicability of empirical approaches. This report illustrates the types of problems for which creative empirical approaches have the potential for significant contributions. The results of two pilot projects are reported in some detail.

PREFACE

This is the first of three reports of work performed under EPA Contract No. 68-02-1704, examining the potential role of state-of-the-art empirical techniques in analyzing air quality and meteorological data. The reports are entitled as follows:

I. The Role of Empirical Methods in Air Quality and Meteorological Analyses

II. Feasibility Study of a Source-Oriented Empirical Air Quality Simulation Model

III. Short-Term Changes in Ground-Level Ozone Concentrations: An Empirical Analysis

The present report is a revision of an interim report prepared in December 1974.

CONTENTS

ABSTRACT iii

PREFACE iv

ACKNOWLEDGMENTS vii

1. INTRODUCTION 1

1.1 EXAMPLE APPLICATIONS 2

1.2 METHODOLOGY 3

2. A SOURCE-ORIENTED EMPIRICAL AIR QUALITY MODEL 5

2.1 MOTIVATION 5

2.2 MATHEMATICAL FORMULATION 7

2.3 TEST DATA 12

2.4 OPTIMIZING PARAMETERS FOR THE GAUSSIAN FORM OF THE SOURCE-RECEPTOR FUNCTION 13

2.5 MORE GENERAL SOURCE-RECEPTOR FUNCTIONS 16

3. EMPIRICAL MODELING OF THE OXIDANT FORMATION PROCESS 17

3.1 MOTIVATION 17

3.2 FORM OF MODEL 18

3.3 THE DATA 20

3.4 THE ANALYSIS 22

3.4.1 Variable Selection 23
3.4.2 Specific Functional Relationship 24

3.5 INTERPRETATION OF MODEL IMPLICATIONS 25

3.6 CONCLUSIONS 32

4. EXTRACTION OF EMISSION TRENDS FROM AIR QUALITY TRENDS . . . 33

4.1 MOTIVATION 33

4.2 REPORT OF A COMPARISON OF EMISSION LEVELS OVER TWO TIME PERIODS 34

4.3 GENERALIZATION AND MATHEMATICAL FORMULATION 37

5. DETECTION OF INCONSISTENCIES IN AIR QUALITY/ METEOROLOGICAL DATA BASES 45

5.1 MOTIVATION 45

CONTENTS (Cont.)

5.2 FORMULATION OF CONSISTENCY MODELS 47
5.3 TYPES OF INCONSISTENCIES 49
5.4 DIFFICULTIES 54

6. REPRO-MODELING: EMPIRICAL APPROACHES TO THE UNDERSTANDING AND EFFICIENT USE OF COMPLEX AIR QUALITY MODELS 57

7. OTHER APPLICATION AREAS 61
7.1 HEALTH EFFECTS OF AIR POLLUTION 61
7.2 SHORT-TERM FORECASTING OF POLLUTANT LEVELS 62

REFERENCES 64

ACKNOWLEDGMENTS

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 in formulating and aiding in the reporting of the studies outlined in Sections 2 and 3. Advice from Leo Breiman and others at Technology Service Corporation further improved the report.

1. INTRODUCTION

The increased availability of appropriate data bases and improvements in methodology have led to increasing use of empirical and statistical approaches for the analysis of air quality and meteorological data [15]. The full application of empirical approaches has, however, been hampered by misconceptions about the nature and intent of such techniques.

One common misconception is that empirical techniques are "black box" techniques, that is, that data is simply processed with little need for adapting the methods to the specific application. The employment of empirical techniques in such a manner simply illustrates that any tool can be abused. As in any other sort of technical work, a great deal of thought and creativity may be required to obtain maximum benefit from an empirical analysis.

A second misconception regarding empirical approaches is a strong distinction between empirical modeling and modeling based upon fundamental physical principles. A good empirical model may involve incorporating a great deal of physical insight. Some empirical analyses do indeed suffer from a lack of physical insight; similarly, some analyses based upon physical insight ignore the presence of measurement data contradicting the major assumptions of those models. In reality, the two approaches are strongly complementary. Physical insight can provide a guide to the appropriate forms of empirical models and to interpretations of the implications of those models; empirical analyses can provide hints as to the key physical mechanisms involved in a process for which measurements are available.

1.1 EXAMPLE APPLICATIONS

It is the intent of this report to illustrate how empirical models might be employed in a number of applications of interest to the Environmental Protection Agency, particularly those with a meteorological aspect. Two of the applications discussed are examined in some detail through limited feasibility studies.

The first feasibility study is on the derivation of a source-oriented empirical air quality model and is discussed in detail in a companion report [7] The study is outlined in Section 2.0. It is often assumed that it is impractical to derive an empirical model which relates the emission source distribution and meteorology to the resulting pollutant concentration distribution. An approach is discussed whereby it is suggested that an empirical meteorological dispersion function can be derived by indirect means.

The second project was a study of the feasibility of empirically deriving change equations relating to the formation of ozone. While only the difference equation for changes in one-hour average oxidant, ignoring emissions, is derived and interpreted, the technique can be extended to include emissions and a full set of difference equations for the major species involved in oxidant formation.

A number of potential areas for the application of empirical methods are outlined in more abstract and less complete terms. Some of the subjects touched upon in more or less detail include the following:

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 measurement errors or system 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, to compare them to other models, or to suggest further areas for model improvement. Model-generated input/output data can be so analyzed by empirical techniques.

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.

1.2 METHODOLOGY

It would be impossible and inappropriate to undertake a comprehensive discussion of the broad spectrum of data-analytic techniques in this report.

In practice the methodology utilized is often directed by the problem itself and by the characteristics uncovered in the data as the analysis proceeds. In fact, it is typically detrimental to the quality of the results of the data-analytic study if a preconception as to the best methodology to be utilized is formed and stubbornly adhered to. A common difficulty in empirical analyses is the tendency to fit the problem to the method rather than the method to the problem.

It is intended that the examples in the body of the report yield a feeling for the variety of approaches in typical applications.

2. A SOURCE-ORIENTED EMPIRICAL AIR QUALITY MODEL

2.1 MOTIVATION

In commenting on the lack of acceptance of empirical/statistical models in air quality modeling in 1973, Kenneth L. Calder called attention to "the historical belief that air quality models based on statistical regression type of analysis are not source-oriented and, therefore, are largely useless for control strategy in terms of the contribution of individual sources to the degradation of air quality" [6]. He went on to ask "whether, with an appropriate analysis, a source-oriented statistical-type of air quality model could be developed which did not involve prior specification of meteorological dispersion functions per se and incorporation of these as in present air quality models. My thought here is that for given 'meteorological conditions' these dispersion functions play the role of transfer functions between the air quality distribution and the distribution of pollutant emissions, and if one were smart enough might, therefore, conceivably be obtained empirically by a mathematical inversion technique (as, for example, by numerical solution of sets of integral equations) utilizing accumulated data on the distributions of air quality and emissions. If this could be accomplished then maybe a major shortcoming of the current statistical models could be removed and we should then in effect have an alternative to the customary meteorological-dispersion type of modeling." These comments suggest the motivation for the study outlined here and reported more fully elsewhere [7].

The difficulties in developing a source-oriented empirical model can be stated from a statistical point of view. The spatial distribution

of pollutant concentrations over a region is determined by emissions and meteorological conditions. The number of variables determining the concentration at a given point is very large, particularly since emissions arise from a large number of point sources and area sources. Consequently the number of emission variables alone can easily be in the hundreds. If an empirical model were to be developed in the most obvious manner, there should be an attempt to relate the pollutant concentration at a given point to all the possible emission variables and meteorological variables affecting the concentration at that point. Since the determination of the relationship between emission/meteorological variables and concentration requires examples of that relationshp over a very wide range of emission and meteorological variables, a tremendous amount of data would be required to adequately determine this relationship.

If we could, however, isolate a given emissions source and we had a number of receptor locations scattered about the source, the variation in wind speed and direction would cause a wide variation in measured concentration at the receptor locations. With enough examples of the source-receptor relationship, the dispersion function could be determined empirically.

In the urban environment, of course, individual sources cannot be isolated. Measurements are the result of contributions from a number of sources. However, because of the wide diversity of meteorological conditions, the concentration will vary widely at a given point, and the sources which contribute to the concentration at that point will similarly

vary. One may then ask for a consistent source-receptor relationship which, when summed (or integrated) over all the sources, would explain best on the average the observed concentrations. More specifically, one could choose the source-receptor function which minimized the average squared error in prediction of the measured values. This concept is the core of the ideas tested. When the parameters optimized are those of a Gaussian-form source-receptor function, this methodology can be regarded as a means of calibrating some commonly used models to particular urban environments.

The data used to test these ideas was model-created data. Model data was chosen for three major reasons:

1. With model data, the source-receptor function is known and can be compared with the function extracted from the data. With measurement data, "truth" is unknown.

2. Area sources and point sources can be isolated and studied separately as well as jointly.

3. The cost of verifying and organizing measurement data would have been beyond the scope of the present study.

2.2 MATHEMATICAL FORMULATION

We worked with a rectangular coordinate system with x-axis along the mean horizontal wind direction, with y-axis crosswind, and with the z-axis vertical. Then in urban air quality models it is customary to consider the pollutant emissions in terms of a limited number (say J) of elevated point-sources together with horizontal area-sources, the latter being possibly located at a few distinct heights ζ_s (say, for example, for s=1,2,3). The

total concentration $\chi(x,y,0)$ at ground level at the receptor location $(x,y,0)$ will be the sum of the concentration contribution from the point-source distribution, say $\chi_p(x,y,0)$ and that from the area-source distribution, say $\chi_A(x,y,0)$, i.e.,

$$\chi(x,y,0) = \chi_p(x,y,0) + \chi_A(x,y,0) \tag{2-1}$$

where

$$\chi_p(x,y,0) = \sum_{\ell=1}^{J} Q_p(\ell) K(x-\xi_\ell, y-\eta_\ell;\ 0,\zeta_\ell) \tag{2-2}$$

$$\chi_A(x,y,0) = \sum_{s=1}^{3} \int_A \int Q_A(\xi,\eta,\zeta_s) K(x-\xi, y-\eta;\ 0,\zeta_s) d\xi d\eta \tag{2-3}$$

and $Q_p(\ell)$ = emission rate of ℓ-th elevated point-source, located at position $(\xi_\ell,\eta_\ell,\zeta_\ell)$.

$Q_A(\xi,\eta,\zeta_s)$ = emission rate of horizontal area-source distribution located at height ζ_s, and A denotes the total integration domain of the area-source distributions.

$K(x-\xi, y-\eta;\ 0,\zeta)$ = source-receptor function; it gives the ground level concentration at the receptor location $(x,y,0)$ resulting from a point-source of unit strength at (ξ,η,ζ).

Note that this formulation includes the assumption of horizontal homogeneity, namely, that the impact of a given source upon a given receptor depends only upon their relative and not absolute coordinates. This assumption is true for an urban environment only in an average sense. A single wind direction is similarly valid only in an average sense. Finally, it should be noted that the above formulation assumes steady-state conditions and is thus only applicable for relatively short time periods (of the order of one hour), when this may be an adequate approximation providing the emissions and meteorological conditions are not rapidly changing.

In Equations (2-2) and (2-3) above it is convenient to use "source-oriented" position coordinates, and to consider a typical ground-level receptor location as (x_i,y_i), i=1,2,....

Let

$$x' = x_i-\xi \quad , \quad dx' = -d\xi \quad , \quad x'_{i\ell} = x_i-\xi_\ell \tag{2-4}$$

$$y' = y_i-\eta \quad , \quad dy' = -d\eta \quad , \quad y'_{i\ell} = y_i-\eta_\ell$$

Then

$$X_p(x_i,y_i,0) = \sum_{\ell=1}^{J} Q_p(\ell)K(x'_{i\ell},y'_{i\ell};\, 0,\zeta_\ell) \tag{2-5}$$

$$X_A(x_i,y_i,0) = \sum_{s=1}^{3} \int\!\!\int_A Q_A(x_i-x',y_i-y',\zeta_s)K(x',y';\, 0,\zeta_s)dx'dy' \tag{2-6}$$

In the following several different source-receptor functions [$K(x',y';\ 0,\zeta)$] will be considered, including the classical Gaussian form that is the basis for the RAM model [14]. For the latter, and with the meteorological condition of infinite mixing depth

$$K(x',y';\ 0,\zeta) = \frac{\exp\left\{-\frac{y^2}{2\sigma_y^2(x')}\right\}\exp\left\{-\frac{\zeta^2}{2\sigma_z^2(x')}\right\}}{\pi U\ \sigma_y(x')\sigma_z(x')} \tag{2-7a}$$

where U denotes the mean wind speed, and we assume simple power-law dependencies for the standard deviation functions, say

$$\sigma_y(x') = a_y(x')^{b_y} \tag{2-7b}$$

$$\sigma_z(x') = a_z(x')^{b_z} \tag{2-7c}$$

Also, as in the RAM-model we will assume that the narrow-plume hypothesis may be employed in order to reduce the double integral of equation (2-6) to a one-dimensional integral. Thus, under this hypothesis, if

$$\int_{-\infty}^{\infty} K(x',y';\ 0,\zeta_s)dy' = G(x',\zeta_s) \tag{2-8}$$

then in place of equation (2-6) we have

$$\chi_A(x_i,y_i,0) = \sum_{s=1}^{3} \int_{x'} Q_A(x_i-x',y_i,\zeta_s)G(x',\zeta_s)dx' \qquad (2\text{-}9)$$

which only involves values of the area-source emission rates in the vertical plane through the wind direction and the receptor location.

For the special case of a Gaussian plume

$$G(x',\zeta_s) = \sqrt{\frac{2}{\pi}} \frac{\exp\left\{-\frac{\zeta_s^2}{2\sigma_z^2(x')}\right\}}{U\sigma_z(x')} \qquad (2\text{-}10)$$

The basic equations (2-5) and (2-6) (or (2-5) and (2-9)), with the Gaussian forms for $K(x',y';0,\zeta)$ and $G(x',\zeta)$ involve four unspecified parameters through the equations (2-7b) and (2-7c), namely a_y,b_y,a_z and b_z. More generally, any functional form chosen for K (and therefore G) may have unspecified parameters, denoted by the vector $\underline{\alpha}$. Thus for the special Gaussian form

$$\underline{\alpha} = (a_y,b_y,a_z,b_z) \quad . \qquad (2\text{-}11)$$

The explicit dependence of the calculated concentration values on these parameters could be indicated by the notation $\chi(x_i,y_i,0;\underline{\alpha})$.

The basic method employed in this study is that of choosing $\underline{\alpha}$ to minimize the error between calculated and observed values of concentrations.* In order

*"Observed" in the present case is model-created test data; the technique is, of course, intended for practical use on measured data.

to express this statement formally, we must elaborate our notation to indicate explicitly the dependence on wind direction; thus $\chi(x_i,y_i,0;\theta;\underline{\alpha})$. For each wind direction θ_j(j=1,2...R) there is a concentration observation for each receptor location (monitoring station). The receptor locations are denoted (x_i,y_i) for i=1,2...N, and are assumed to be at ground level so that we may omit the symbol 0 in the χ notation. Then the mean square error over all observations is

$$e^2(\underline{\alpha}) = \frac{1}{RN}\sum_{i=1}^{N}\sum_{J=1}^{R}\left[\chi_{obs}(x_i,y_i;\theta_j)-\chi_{calc}(x_i,y_i;\theta_j;\underline{\alpha})\right]^2$$

$$= \frac{1}{RN}\sum_{i=1}^{N}\sum_{j=1}^{R}\left[\chi_{obs}(x_i,y_i;\theta_j)-\chi_p(x_i,y_i;\theta_j;\underline{\alpha})-\chi_A(x_i,y_i;\theta_j;\underline{\alpha})\right]^2 \quad (2\text{-}12)$$

where χ_p and χ_A are given by Eqs. (2-5) and (2-6) (or (2-5) and (2-9)).

The problem of minimizing e^2 with respect to $\underline{\alpha}$ is a standard optimization problem. We will not discuss the particular method used here.

2.3 TEST DATA

For a realistic distribution of point-sources, area-sources and receptor locations, use was made of unpublished information from a 1968 air pollution study conducted in St. Louis, Missouri. The area sources were gridded into over 600 square regions; there were 60 point sources and errors were calculated at 40 receptors for the 16 wind directions. (See Reference [7] for more details.) The corresponding concentration data were generated by the EPA-developed RAM algorithm [14], which is a specific implementation of the classical Gaussian

plume formulation, that considers both point- and area-sources, with three possible heights for the latter, and which uses the "narrow-plume" hypothesis (i.e., Eq. (2-9)) to calculate the area-source concentration contribution χ_A. A constant wind speed U of 5 meters per second was employed, and sixteen wind directions at the points of the compass were simulated. Infinite mixing depth and a neutral atmospheric stability category were assumed. For the latter, in Eqs. (2-7b) and (2-7c), we have

$$\begin{aligned} a_y &= 0.072 \quad , \quad b_y = 0.90 \\ a_z &= 0.038 \quad , \quad b_z = 0.76 \quad . \end{aligned} \tag{2-13}$$

For this data, these values and the indicated equations are optimal and would produce zero mean-square error. It is this result we hope to be able to recover from the data by the optimization procedure.

2.4 OPTIMIZING PARAMETERS FOR THE GAUSSIAN FORM OF THE SOURCE-RECEPTOR FUNCTION

The data base described earlier contains concentration values at forty receptors and sixteen wind direcitons, a total of 640 values (referred to as "actual" values). The contribution to the concentration from point and area sources was available separately, as well as in toto.

Equations (2-5) and (2-7a) provide a prediction of the point-source pollutant concentration at any given receptor location once the four parameters are specified. A comparison of values predicted by these equations versus actual values allows calculation of the root-mean-square value of the error with a given choice of parameter values. (See Eq. (2-12), with area sources at zero.)

With initial guesses of $a_y=a_z=0.1$ and $b_y=b_z=1.0$, the search routine described arrived at values of

$$a_y = 0.74,\ b_y = 0.92,\ a_z = 0.039,\ b_z = 0.77$$

when the "true" values (those used to create the data) were

$$a_y = 0.72,\ b_y = 0.90,\ a_z = 0.38,\ b_z = 0.76.$$

The root-mean-square (RMS) error initially was 157 $\mu g/m^3$ and the maximum error over the 640 values was 1205 $\mu g/m^3$; the parameter values after 100 iterations yielded an RMS error of 14 $\mu g/m^3$ and a maximum error of 175 $\mu g/m^3$, To place the size of the final error in perspective, we note that the actual values (due to point sources alone) were as high as 1545 $\mu g/m^3$.

Employing Eq. (2-9) for area sources and using only the area-source contribution in the "actual" data, we get the values (.037, 0.79) for a_z and b_z versus the true values (.038, 0.76).

The results of treating point and area sources simultaneously, representative of the case which would be encountered with measurement data, are listed in Table 2-1; the algorithm once again closely approaches the optimum values in 100 iterations. Actual values of the total concentrations from both point and area sources go to above 1600 $\mu g/m^3$.

While the initial parameter values we chose in these cases converged toward the values used in creating the data, experimentation indicated that this was not always the case. Small RMS errors could be achieved with combinations of parameters significantly different in value from those used in

Table 2-1: Point and Area Sources Together; Parameter Values at Initial, Mid, and Final Iteration During Search

					Both Point and Area Sources	
Iteration	a_y	b_y	a_z	b_z	RMS Error ($\mu g/m^3$)	Max. Error ($\mu g/m^3$)
0 (initial)	0.100	1.00	0.100	1.00	157	1205
50 (mid)	0.055	0.79	0.044	0.67	79	583
100 (final)	0.074	0.89	0.036	0.74	24	194
ACTUAL VALUES:	(0.072)	(0.90)	(0.038)	(0.76)	(0)	(0)

creating the data. It is easy to find rather different combinations of a and b which yield very similar values of ax^b over the range of x in which we are interested. It is clear that an essentially equivalent combination of values should not be deemed erroneous, since they yield an accurate empirical model. We regard this as a characteristic of the formulation chosen for calculating σ and do not regard it a difficulty of the methdology proposed. Further, in practice, initial values for the parameters would be chosen from the literature, and the solution obtained would be a set of values similar to the initial values, but which minimized the prediction error.

This aspect of implementation also suggests that a good initial guess would be employed and, thus, that convergence to an "optimum" solution would be rapid.

2.5 MORE GENERAL SOURCE-RECEPTOR FUNCTIONS

More complex source-receptor functions (such as multivariate polynomials and piecewise quadratic functions) were tested with success [7], but broad conclusions about alternative forms will not be forthcoming through the analysis of the present test data. Analysis of measurement data may allow meaningful comparison of the Gaussian and more general parameterized forms.

3. EMPIRICAL MODELING OF THE OXIDANT FORMATION PROCESS

3.1 MOTIVATION

Typical objectives of a modeling effort are (1) qualitative understanding and (2) quantitative impacts. In air quality modeling, these objectives are aimed at the ultimate objectives of determining the effects of alternative control policies and understanding which policies will be most productive. Ozone air quality modeling efforts have been largely concentrated at extremes of the spectrum of approaches to modeling: (1) simple statistical models with limited applications, or (2) complex models based on the underlying physics and chemistry of the process. The former class of models provides easy-to-use, but rough, guidelines; the latter class of model is capable of detailed temporal and spatial impact analysis, but is costly and difficult to use.

The study outlined in this section illustrates the feasibility of an intermediate class of model which is relatively inexpensive and easy-to-use, but which is capable of providing reasonably detailed temporal and spatial estimates of oxidant concentration. Further, the form of the model makes it possible to understand (with careful inspection) the qualitative implications of the model as a guide to the design of control strategies. This study is reported more fully elsewhere [3].

We hasten to emphasize, however, that a full model in this class was not a result of this study; rather, we present an analysis which we believe indicates the feasibility of the development of such a model. In particular, we develop am empirical difference equation for the production of oxidant from chemical precursors, as effected by meteorological variables. A full

model would involve difference equations for the precursor pollutants as well. Further, data easily available did not include all meteorological variables of possible interest or emission data. (Since ozone is a secondary pollutant, emissions of primary pollutants over a brief interval, e.g., one hour, will not effect the change in ozone levels over that interval to the degree they effect the change in primary pollutant levels. Since we did not derive difference equations for the primary pollutants in this study, not including emissions did not prove serious.) The context in which the reader sould then interpret the results is as the degree to which the change in ozone can be explained <u>despite</u> these limitations. Whatever degree of explanation of the variance in one- or two-hour changes in ozone we can achieve within these limitations can be improved when more of the omitted factors are taken into account. This analysis will thus provide a pessimistic estimate of the degree of success that can be expected in a full-scale implementation of the approach.

3.2 FORM OF MODEL

We consider a "parcel" or air that moves along a trajectory to be determined from the wind field, and we define $O_3(t)$ as the oxidant concentration averaged over the hour preceding time t. We further define $\Delta O_3(t)$ as the change in the hourly average oxidant concentration (in pphm) in the time interval Δt following t; explicitly,

$$\Delta O_3(t) = O_3(t + \Delta t) - O_3(t) \quad . \tag{3-1}$$

(We will consider Δt = 1 hour and Δt = 2 hours.) We seek an equation predicting the change in hourly average concentration after time t from measurements of

of pollutants and meteorology available at time t. Pollutant measurements other than ozone we will consider as possible precursors include the following, all of them in terms of concentration averaged over the hour preceding time t:

$NO(t)$ = NO concentration (pphm)

$NO_2(t)$ = NO_2 concentration (pphm)

$HC(t)$ = Non-methane hydrocarbon concentration (ppm)

$CH_4(t)$ = Methane (ppm).

Meteorological variables considered explicitly include the following, again averaged over the hour preceding time t:

$SR(t)$ = solar radiation (gm-cal/cm^2/hr)

$T(t)$ = temperature (°F).

Mixing height was not used in the present study.

We thus seek a relationship of the form

$$\Delta O_3(t) = F[O_3(t),NO(t),NO_2(t),HC(t),CH_4(t),SR(t),T(t)] , \qquad (3\text{-}2)$$

which accurately reflects observed data. Referring to (3-1), equation (3-2) can be alternatively written as

$$O_3(t+\Delta t) = O_3(t) + F[O_3(t),NO(t),NO_2(t),HC(t),CH_4(T),SR(t),T(t)]. \qquad (3\text{-}3)$$

This form indicates explicitly how such a relationship, if derived, can be used to compute a sequence of hourly or bi-hourly oxidant concentrations.

(Similar equations would be derived for the primary pollutants to provide a complete model.)

We must incorporate transport effects. We have adopted a rather simple model. The model estimates the trajectory of a "parcel" of air from ground-level measurements of the wind field. A parcel arriving at a given location at a given time (e.g., Pasadena at 1600 hours) is estimated, from the current wind direction, to have been at another location upwind one hour earlier. The distance of that location upwind is given by the current wind speed. The trajectory is tracked backwards to give a sequence of hourly locations. The "actual" values of pollutant levels at these points at the given times are obtained by an interpolation procedure from measured data at fixed monitoring stations. The motivation for tracking parcels backwards rather than forwards is to allow choice of parcels which end up at monitoring stations in part, so that the last (and often highest) pollutant concentration need not be interpolated. The air parcel trajectory approach is obviously a simplification of the true physics of the system; this approach is similar to assumptions employed in some physically based air quality models [26]. In the present empirical modeling context, the trajectory approach is a statistical approximation rather than an assumption; that is, the inaccuracy of the approximation will be reflected in the overall error of the final empirical model.

3.3 THE DATA

Data collected by the Los Angeles Air Pollution Control District was employed. Air quality data from the seven monitoring stations indicated in Figure 3-1 was utilized.

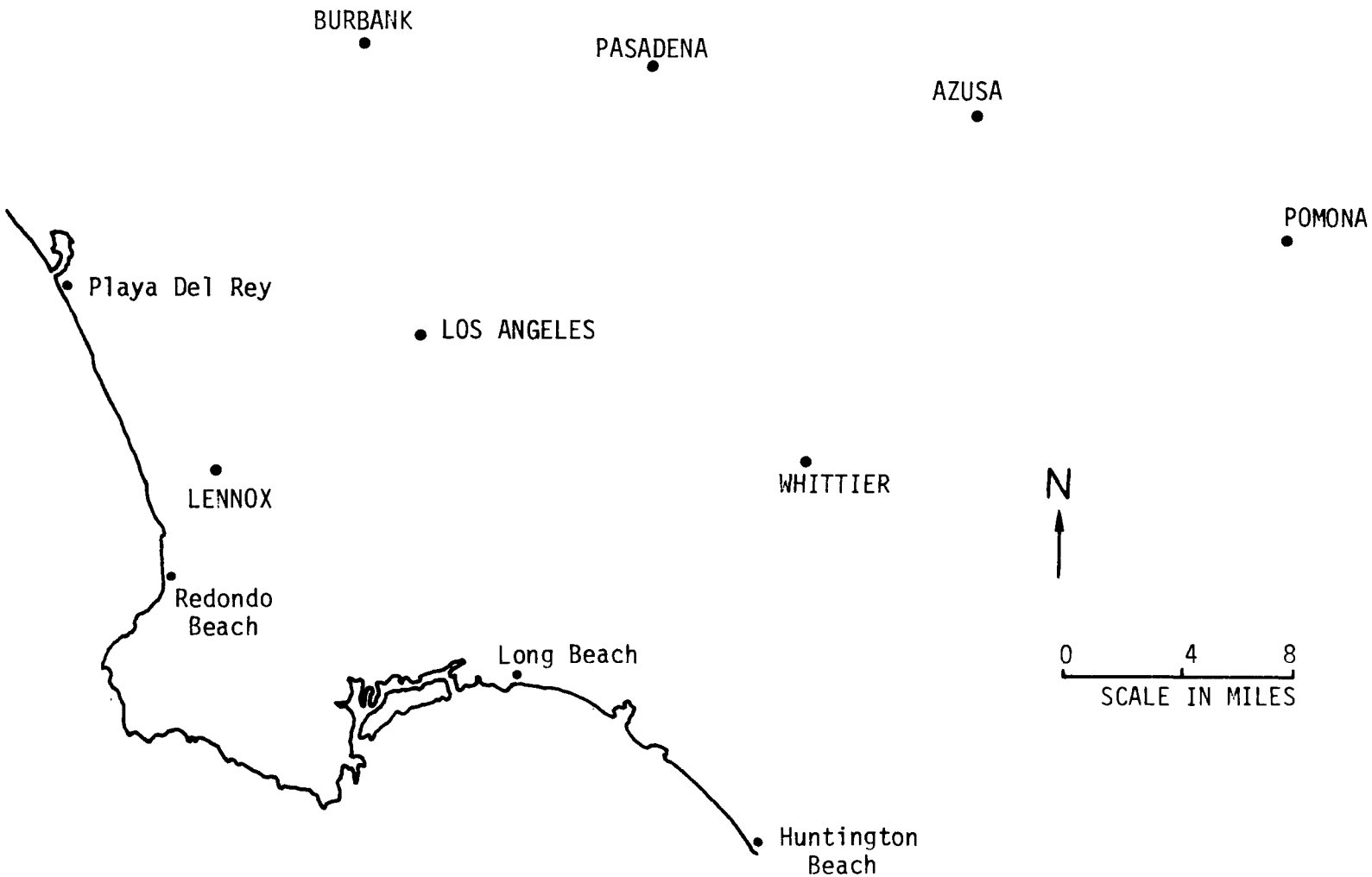


Figure 3-1. The study region (monitoring stations are capitalized).

We interpolated the wind field in a region of the Los Angeles basin so that we were able to track parcels of air as they moved through the basin. The pollutant readings at seven APCD stations were also interpolated so that we could keep hourly records of the pollutants discussed. We also had the hourly solar radiation readings at the Los Angeles Civic Center location of the APCD, and hourly temperature readings at three representative locations in the basin.

Our study was carried out using data from five summer months, June through October 1973. About 7000 trajectories were formed and placed in the primary data base.

From the bank of 7000 trajectories, we extracted a sample of about 1900 data vectors of the form

$$(\Delta O_3,\ O_3,\ NO,\ NO_2,\ HC,\ CH_4,\ SR,\ T) \quad ,$$

where ΔO_3 was a one-hour change and about 1800 vectors where ΔO_3 was a two-hour change.

3.4 THE ANALYSIS

Since time is only implicit in (3-2), we search for a fixed relationship

$$\Delta O_3 = F(O_3,\ NO,\ NO_2,\ HC,\ CH_4,\ SR,\ T) \quad . \tag{3-4}$$

Equally important, we want to determine which of the variables were most significantly related to the change in ozone. Therefore, we really had two objectives in this study:

(1) To find those subgroups of the variables most significantly related to the ozone change.

(2) To find the form of the function F providing the best fit to the data.

3.4.1 Variable Selection

For the variable selection and exploratory phase, we used INVAR, a general nonparametric method for estimating efficiently how much of the variability in the dependent variable can be explained by a subgroup of the independent variables [2]. This technique estimates the limiting value of percent of variance explained (PVE) by a "smooth" nonlinear model.* We first tested all independent variables as individual predictors, then pairs of variables, and then added variables to find the best three, etc. The most significant individual variables (in approximate order of importance) are SR, NO_2, T, and O_3.

Exploring pairs of variables, we found the best pair was (O_3, SR) for both one- and two-hour ΔO_3.

Triplets of variables were then explored with one really significant improvement showing. The best triple by a good margin was (O_3, SR, NO_2), explaining 60% of the variance in one-hour ΔO_3 and 71% of the two-hour ΔO_3.

The final significant increase occurred when we added temperature to O_3, NO_2, SR. But, somewhat strangely, the increase was significant only for the data base of one hour ΔO_3. Here we obtained

*Percent of variance explained equals

$$100 \times \left[1 - \frac{\text{variance of error in prediction}}{\text{variance of dependent variable}}\right]$$

Variables	One-Hour PVE
O_3, NO_2, SR, T	66%

In all of the INVAR runs using HC and CH_4, neither of them significantly increased the PVE. For instance, when HC and CH_4 were individually added to the variables NO_2, NO, O_3, and SR, the maximum increase in PVE was 2.1%.

These results are encouraging; the three variables O_3, NO_2, SR predict about 71% of the variance in two-hour ozone changes, that is, with a correlation between predicted and actual values of 0.84 over 1800 samples.

3.4.2. Specific Functional Relationship

The exploratory analysis provided nonparametric estimates of the degree of predictability of two-hour ΔO_3 as a function of O_3, NO_2, SR. In this subsection we discuss the derivation of a specific simple functional form to make explicit this relationship.

To get a continuous functional form for the relationship of ΔO_3 to O_3, NO_2, and SR, we used continuous piecewise linear regression [16,12]. Since the function generated by this method is smoother and less general than that used in INVAR estimates, we did not achieve the level of PVE obtained by INVAR. The continuous piecewise linear function which minimizes the mean-square error in the fit to the 1800 sample points is given by*

$$\Delta O_3 = -5.125 \cdot \max\{A,B,C\} - 1.167 \cdot \max\{D,E,F\} + 10.48 , \qquad (3\text{-}5)$$

*The notation max {A,B,C} means the largest of the three values computed by equations A, B, and C.

where

$$A = -0.2146 \cdot O_3 - .0701 \cdot NO_2 - .002268 \cdot SR + .9376$$

$$B = .02114 \cdot O_3 - .1013 \cdot NO_2 - .01075 \cdot SR + 2.275$$

$$C = .1638 \cdot O_3 - .09855 \cdot NO_2 - .005938 \cdot SR - .2263$$

$$D = .02709 \cdot O_3 - .3015 \cdot NO_2 + .001298 \cdot SR + 2.304$$

$$E = -.009565 \cdot O_3 + .0005252 \cdot NO_2 - .001079 \cdot SR + .2306$$

$$F = -.0144 \cdot O_3 + .2066 \cdot NO_2 - .003171 \cdot SR - 2.943$$

(The unusual form of the equation has no physical interpretation, but is simply a consequence of the particular methodology employed.) This equation explained 60.7% of the variance, a correlation between predicted and actual values of 0.78. Figure 3-2 illustrates the form of the function.

This equation can be used to calculate a sequence of oxidant concentrations in a parcel of air by using known values of the other pollutant concentrations (since difference equations for these pollutants have not been derived). Figures 3-3, 3-4, and 3-5 illustrate the result for three air parcel trajectories. Numerical values are listed in Table 3-1.

3.5 INTERPRETATION OF MODEL IMPLICATIONS

Let us attempt to interpret the functional form in (3-5). The final fitted surface is fairly simple, consisting of a continuous patching together of eight hyperplane segments. Of the eight regions, there are three small regions that together contain only 1.0% of the total number of points. We will ignore these and restrict our analysis to the information contained in the functional fit to ΔO_3 in the five other regions.

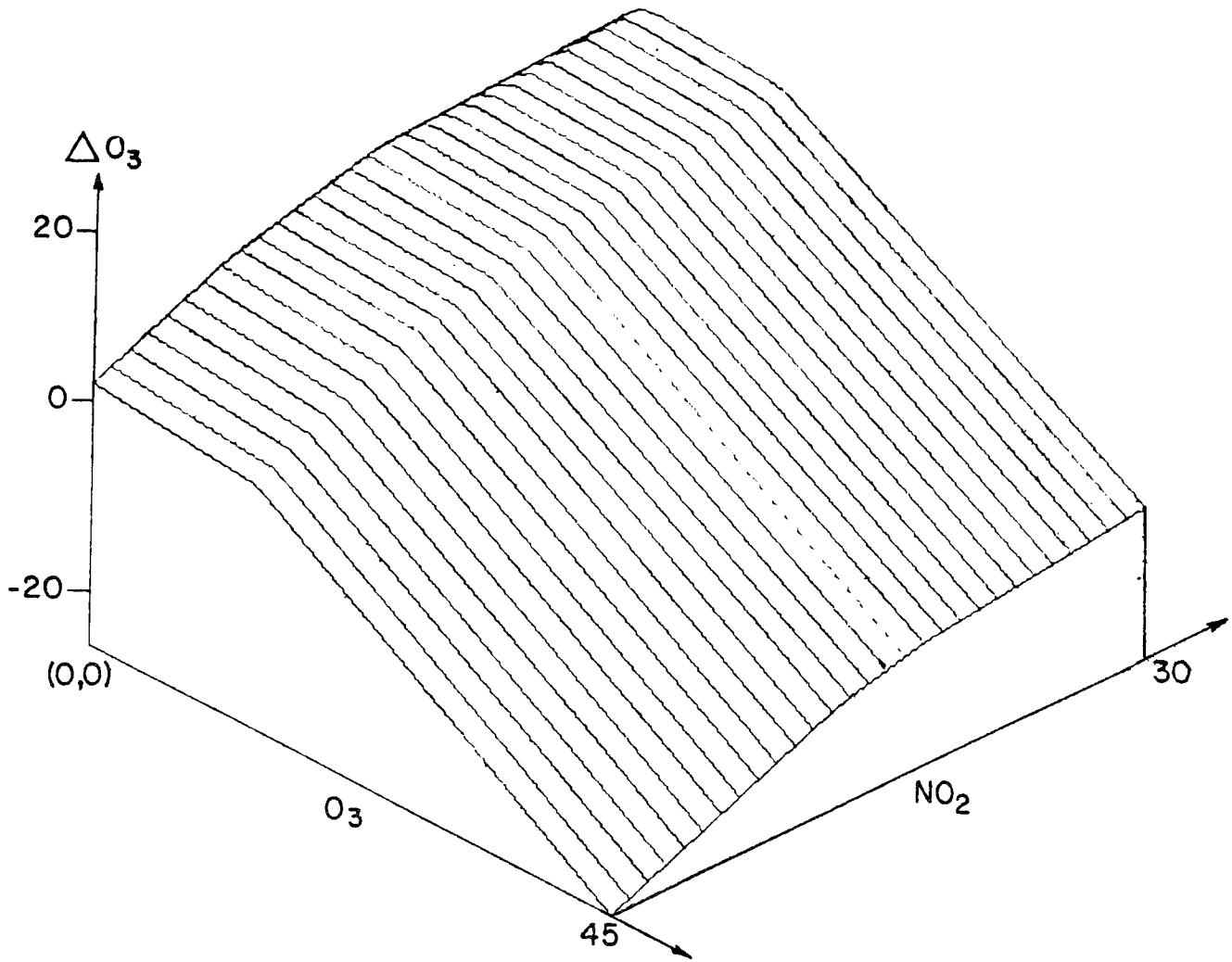


Figure 3-2. Graph of regression surfaces with SR = 100

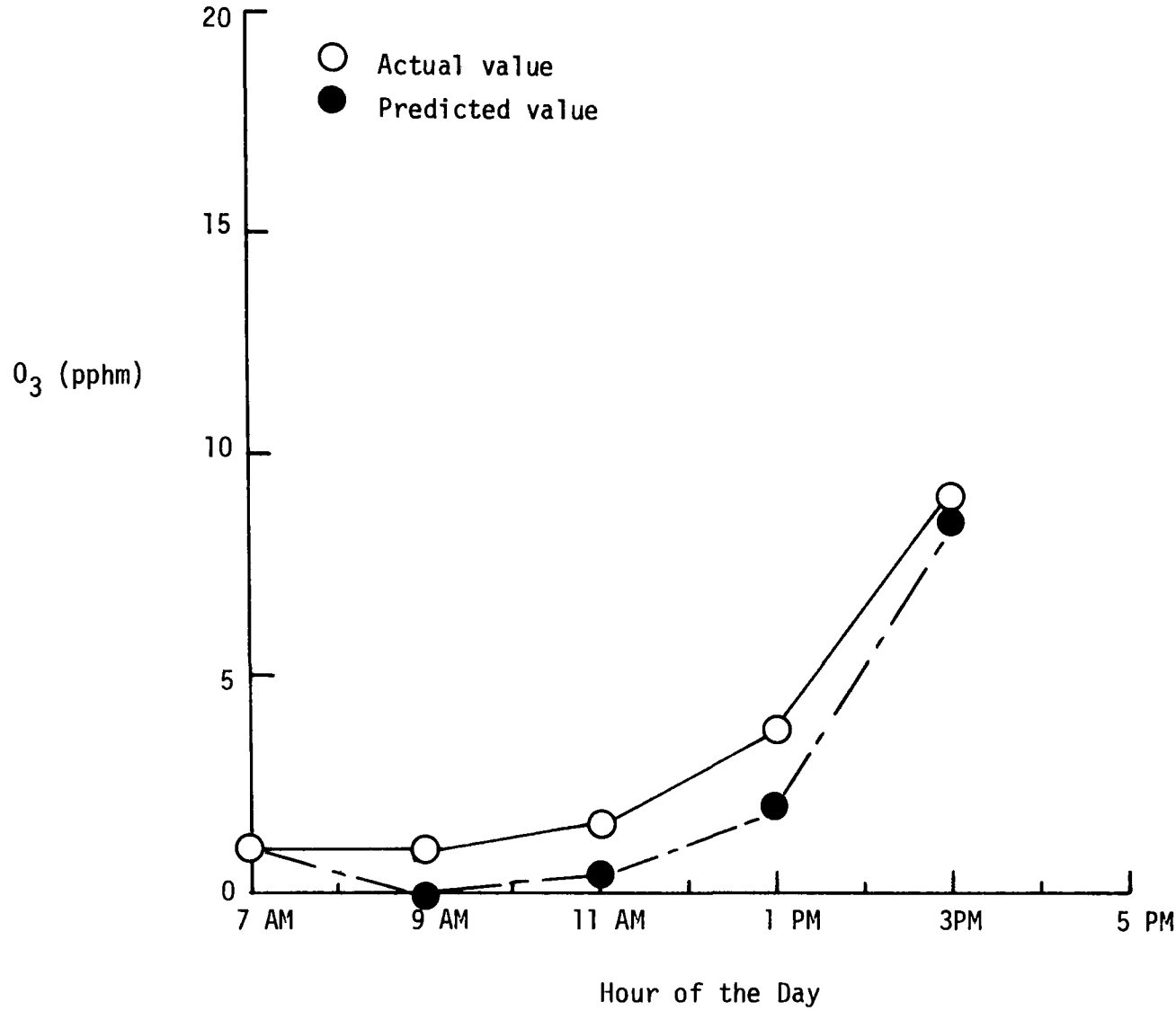


Figure 3-3. Air parcel arriving at the Pomona station at 3 PM.

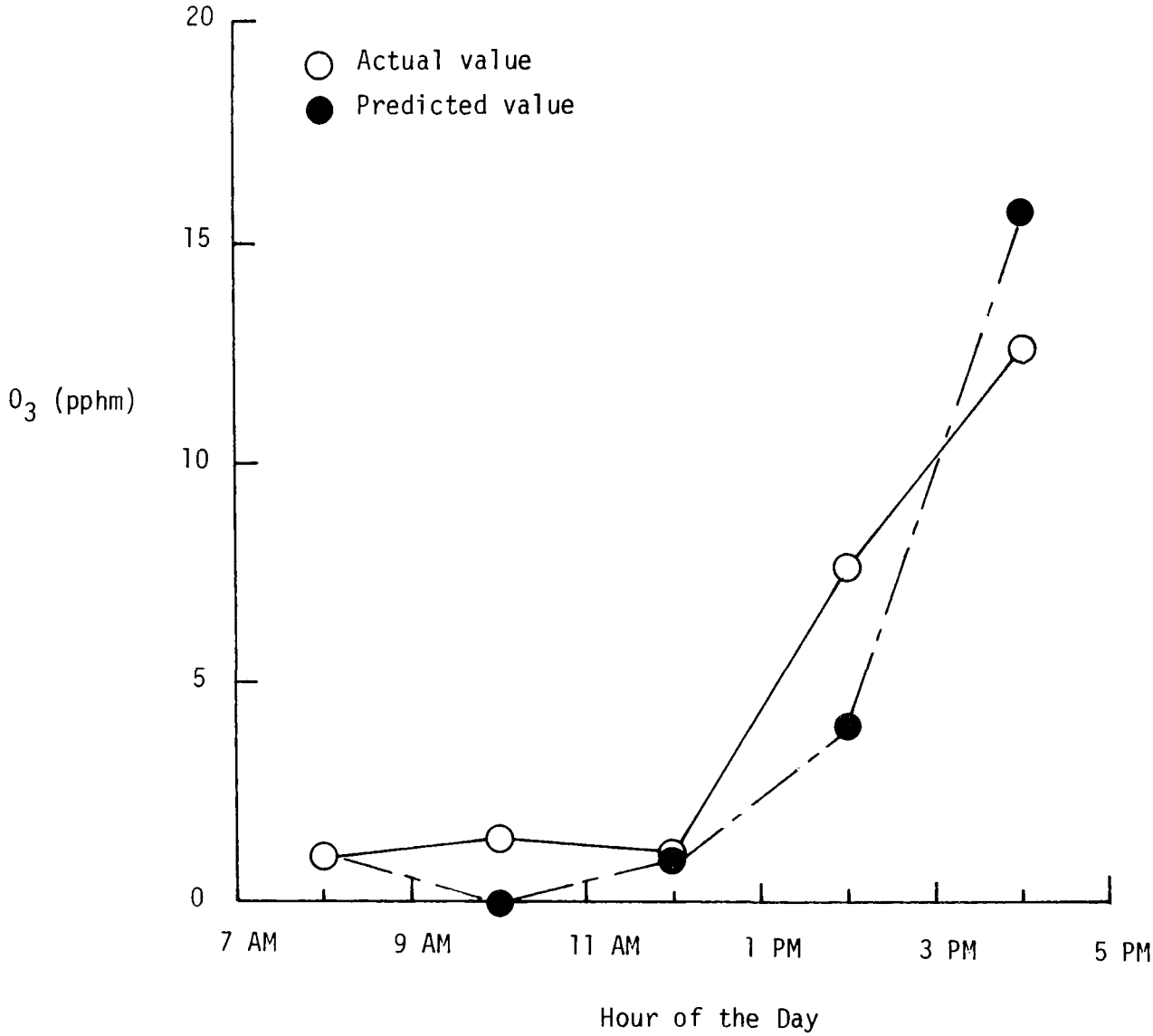


Figure 3-4. Air parcel arriving at the Pomona station at 4 PM.

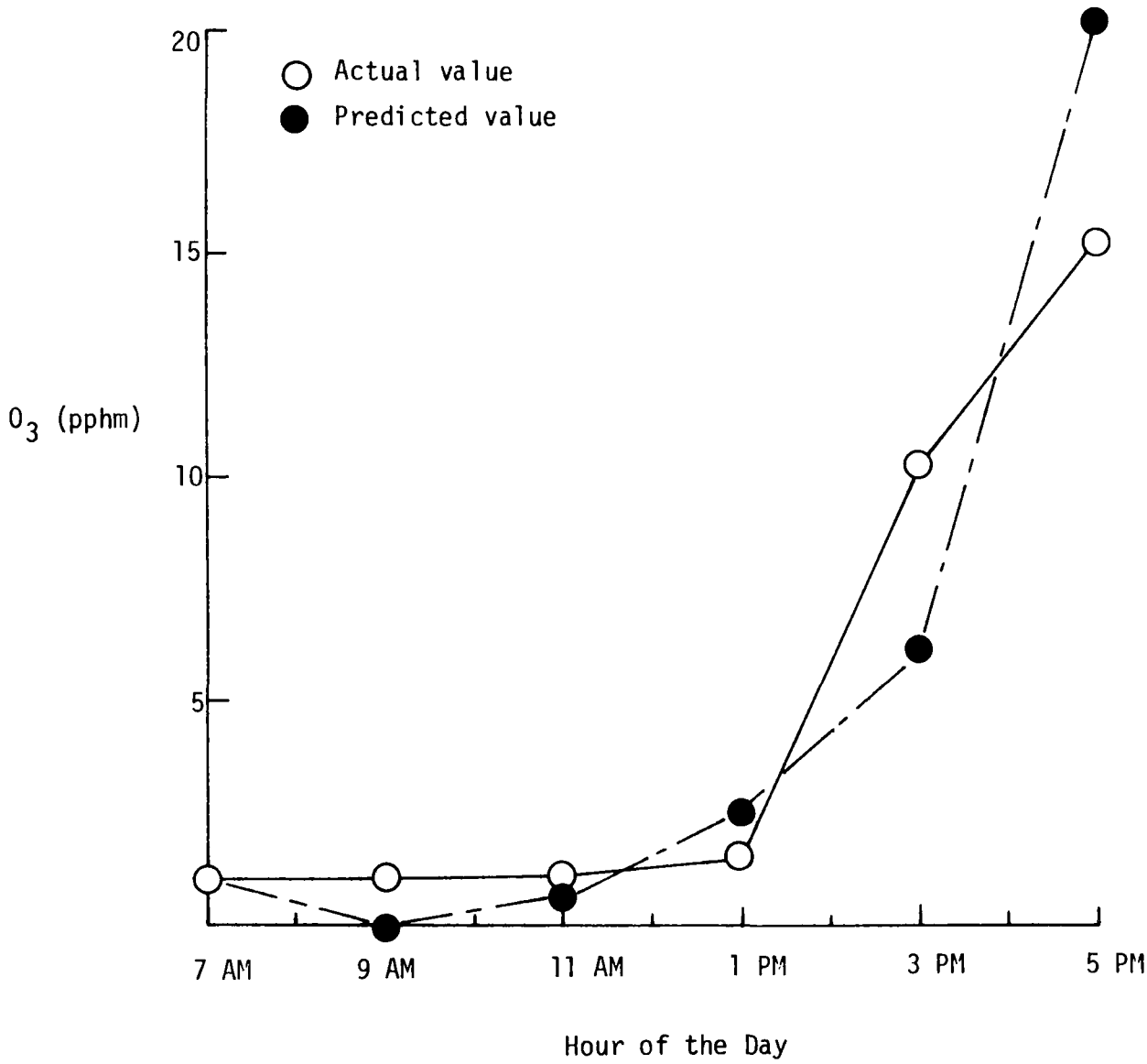


Figure 3-5. Air parcel arriving at the Pomona station at 5 PM.

Table 3-1. Comparison of Model Predictions to Actual Values of O_3

(a) Air parcel arriving at the Pomona station at 3 PM

Hour of the Day	Actual Value (pphm)	Predicted Value (pphm)
7 AM	1.0	
9 AM	1.0	0.0
11 AM	1.1	0.6
1 PM	1.6	2.5
3 PM	10.2	6.1
5 PM	15.1	20.2

(b) Air parcel arriving at the Pomona station at 4 PM

Hour of the Day	Actual Value (pphm)	Predicted Value (pphm)
8 AM	1.0	
10 AM	1.4	0.0
12 N	1.2	1.0
2 PM	7.8	4.0
4 PM	12.7	15.8

(c) Air parcel arriving at the Pomona station at 5 PM

Hour of the Day	Actual Value (pphm)	Predicted (pphm)
7 AM	1.0	
9 AM	1.0	0.0
11 AM	1.6	0.4
1 PM	3.9	2.0
3 PM	9.0	8.7

Region 1, containing almost half of the sample points, is representative of low pollution levels, low O_3 production, and low solar radiation. Region 2, with 33% of the points, contains data with above average mean NO_2 and solar radiation levels, below average O_3 levels, and high positive changes in O_3. The other three regions, with a total of 20% of the sample points, represent more extreme conditions.

Since the size of the coefficients depends on the scaling of the variables, we introduce normalized variables by dividing the original variables by their overall standard deviations; i.e., denoting normalized variables by primes:

$$O_3' = O_3/6.2,\ NO_2' = NO_2/5.2,\ SR' = SR/52.8\ . \qquad (3\text{-}6)$$

The equations are given in terms of the normalized variables, in Table 3-2. These are derived from equation (3-5).

Table 3-2. Normalized Equations for ΔO_3
(ΔO_3 not normalized)

Region	Equations
1	$-0.9\ (O_3')\ +4.5\ (NO_2')\ +2.8\ (SR')\ -\ 3.9$
2	$-0.6\ (O_3')\ +2.6\ (NO_2')\ +2.9\ (SR')\ -\ 1.4$
3	$-5.4\ (O_3')\ +4.4\ (NO_2')\ +1.5\ (SR')\ +\ 9.0$
4	$-0.57\ (O_3')\ +1.4\ (NO_2')\ +3.1\ (SR')\ +\ 2.3$
5	$-5.1\ (O_3')\ +1.4\ (NO_2')\ +1.8\ (SR')\ +15.1$

The major qualitative conclusions that can be inferred from this table (see [3] for fuller discussion) are the following:

(1) At below average O_3 levels, the O_3 change is determined largely by the SR and NO_2 levels, with larger values of these latter two related to larger values of the O_3 change. The largest positive changes in O_3 occur in this regime.

(2) At above average O_3 levels, the O_3 has a strong negative association with O_3 change, and moderate to high levels of NO_2 and SR are associated with low to only moderately above-average changes in O_3. The consistent negative sign on O_3' suggests a possible self-limiting effect.

3.6 CONCLUSIONS

It is possible to derive rather accurate empirical equations predicting the short-term change in oxidant concentration (considering the limitations of the data and the difficulty of the problem). These results are encouraging in terms of the practicality of a full model involving emission variables and all the major reactive pollutants.

4. 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 [11]. 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 coefficient of the temperature difference term in a similarly derived equation for the later

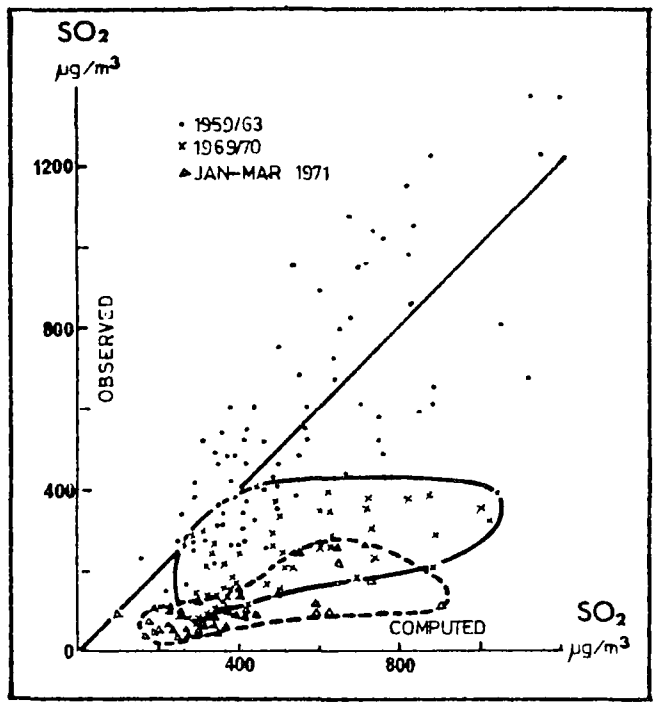


Figure 4-1: Values of daily mean SO_2 concentration computed from temperature measurements at 7 P.M. versus daily mean SO_2 concentration observed. The fact that the values in the later period are much less than would be expected from the meteorology suggests that emissions are less. Ref. [11]

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:

$$\begin{array}{ccc} q_1^{(2)} & , & \underline{m}_1^{(2)} \\ q_2^{(2)} & , & \underline{m}_2^{(2)} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ q_{N_2}^{(2)} & , & \underline{m}_{N_2}^{(2)} \end{array} \quad . \qquad (4\text{-}3)$$

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:

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

where $\underline{e}$ is a vector of emission measurements.

This equation plus measurements error produced the measurement data of (4-2) and (4-3). We are assuming that the equation does not differ between the two periods.

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:

$$q^{(1)} = f_1(\underline{m}) \quad . \tag{4-6}$$

Equation (4-1) is such an equation.

Then obviously

$$q^{(1)} = f_1(\underline{m}) \approx F(\underline{e}_1,\underline{m}) \quad , \tag{4-7}$$

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

$$q^{(2)} = f_2(\underline{m}) \quad . \tag{4-8}$$

Then, as before,

$$q^{(2)} = f_2(\underline{m}) \approx F(\underline{e}_2,\underline{m}) \quad . \tag{4-9}$$

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 on 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 constant 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

$$\frac{q^{(2)}}{q^{(1)}} = \frac{f_2(\underline{m})}{f_1(\underline{m})} = \frac{F(\underline{e}_2,\underline{m})}{F(\underline{e}_1,\underline{m})} = \frac{G(\underline{e}_2)}{G(\underline{e}_1)} \quad , \tag{4-11}$$

using (4-6), (4-7), and (4-8). Thus, the ratio 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 is 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. 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 [9]. 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:

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

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 in 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 [1].

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" [10]. 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. 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 [16]. 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 [12], we will not discuss it in further depth in this report. We do wish to emphasize the role of such an analysis in evaluating, validating and comparing 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 [23]. 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 [16]. 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. OTHER APPLICATION AREAS

Two 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 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 <u>linear</u> 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.

7.2 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 one with 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 evaluating the results can be misleadingly promising 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-pollution 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. Such an approach will build into the methodology the natural discontinuities inherent to the onset and conclusion of a high air pollution episode.

REFERENCES

1. 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.

2. Breiman, Leo, and W.S. Meisel, "General Estimates of the Intrinsic Variability of Data in Nonlinear Regression Modesl," TSC Report, Technology Service Corp., Santa Monica, Calif., October 1974, to be published in J. Amer. Stat. Asso.

3. Breiman, Leo, and W.S. Meisel, "Short-term Changes in Ground-Level Ozone Concentrations: An Empirical Analysis," Part III of "The Role of Empirical Methods in Air Quality and Meteorological Analyses," Final Report for EPA Contract No. 68-02-1704, October 1975.

4. 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., Sept. 9-13, 1974, American Meteorological Society, Boston, Mass.

5. Calder, K.L., "Some Miscellaneous Aspects of Current Urban Pollution Models," Proc. Symp. on Multiple Source Urban Diffusion Models, EPA, Research Triangle Park, No. Carolina, 1970.

6. Calder, K.L., Quoted by Niels Busch in the proceedings of the fourth meeting of the NATO/CCMS Panel on Air Pollution Modeling, from a letter written in March 1973.

7. Calder, K.L., W.S. Meisel, and M.D. Teener, "Feasibility Study of a Source-Oriented Empirical Air Quality Model," (Part II of "Empirical Techniques for Analyzing Air Quality and Meteorological Data"), Final Report on EPA Contract No. 68-02-1704, Dec. 1975.

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

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

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

11. 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, Feb. 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, No. Carolina.)

REFERENCES (Cont.)

12. Horowitz, A. and W.S. Meisel, "The Application of Repro-Modeling to the Analysis of a Photochemical Air Pollution Model," EPA Report No. EPA-6504-74-001, NERC, Research Triangle Park, No. Carolina, Dec. 1973.

13. Horowitz, A. 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, CA, Aug. 22, 1974.

14. Hrenko, J.M. and D.B. Turner, "An Efficient Gaussian-Plume Multiple Source Air Quality Algorithm," Paper 75-04.3, 68th Annual APC Meeting, Boston, June 1975.

15. 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, Riso, 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, No. Carolina 27711.)

16. Meisel, W.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.

17. Meisel, W.S., Computer-Oriented Approaches to Pattern Recognition, Academic Press, New York, 1972.

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

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

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

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

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

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

REFERENCES (Cont.)

24. 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.

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

26. Wayne, Kokin, and Weisburd, "Controlled Evaluation of Reactive Environmental Simulation Model (REM)," Vols. I & II, NTIS PB 220 456/8 and PB 220 457/6, Feb. 1973.

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO.	2.	3. RECIPIENT'S ACCESSION NO.
EPA-600/4-76-029a		

4. TITLE AND SUBTITLE	5. REPORT DATE
EMPIRICAL TECHNIQUES FOR ANALYZING AIR QUALITY AND METEOROLOGICAL DATA. Part I. The Role of Empirical Methods in Air Quality and Meteorological Analyses	July 1976
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S)	8. PERFORMING ORGANIZATION REPORT NO.
W.S. Meisel	TSC-PD-132-2
9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT NO.
Technology Service Corporation 2811 Wilshire Boulevard Santa Monica, California 90403	1AA009
	11. CONTRACT/GRANT NO.
	EPA 68-02-1704
12. SPONSORING AGENCY NAME AND ADDRESS	13. TYPE OF REPORT AND PERIOD COVERED
Environmental Sciences Research Laboratory Office of Research and Development U.S. Environmental Protection Agency Research Triangle Park, North Carolina 27711	Final May 74-Dec 75
	14. SPONSORING AGENCY CODE
	EPA-ORD

15. SUPPLEMENTARY NOTES
This is the first of three reports examining the potential role of state-of-the-art empirical techniques in analyzing air quality and meteorological data.

16. ABSTRACT

Empirical methods have found limited application in air quality and meteorological analyses, largely because of a lack of good data and the large number of variables in most applications. More and better data, along with advances in methodology, have broadened the applicability of empirical approaches. This report illustrates the types of problems for which creative empirical approaches have the potential for significant contributions. The results of two pilot projects are reported in some detail.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
*Air pollution		13B
*Meteorological data		04B
*Atmospheric diffusion		04A
*Mathematical models		12A
*Environment simulation		14B

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS *(This Report)*	21. NO. OF PAGES
RELEASE TO PUBLIC	UNCLASSIFIED	73
	20. SECURITY CLASS *(This page)*	22. PRICE
	UNCLASSIFIED	

EPA Form 2220-1 (9-73)