EPA-460/3-74-024
OCTOBER 1974

AUTOMOBILE EXHAUST EMISSION MODAL ANALYSIS MODEL EXTENSION AND REFINEMENT

U.S. ENVIRONMENTAL PROTECTION AGENCY
Office of Air and Waste Management
Office of Mobile Source Air Pollution Control
Certification and Surveillance Division
Ann Arbor, Michigan 48105

EPA-460/3-74-024

AUTOMOBILE EXHAUST EMISSION MODAL ANALYSIS MODEL EXTENSION AND REFINEMENT

by

H. T. McAdams

Calspan Corporation
P. O. Box 235
4455 Genesee Street
Buffalo, New York 14221

Contract No. 68-03-0435

EPA Project Officer: C. J. Domke

Prepared for

U.S. ENVIRONMENTAL PROTECTION AGENCY
Office of Air and Waste Management
Office of Mobile Source Air Pollution Control
Certification and Surveillance Division
Ann Arbor, Michigan 48105

October 1974

This report is issued by the Environmental Protection Agency to report technical data of interest to a limited number of readers. Copies are available free of charge to Federal employees, current contractors and grantees, and nonprofit organizations - as supplies permit - from the Air Pollution Technical Information Center, Environmental Protection Agency, Research Triangle Park, North Carolina 27711; or, for a fee, from the National Technical Information Service, 5285 Port Royal Road, Springfield, Virginia 22161.

This report was furnished to the Environmental Protection Agency by Calspan Corporation, in fulfillment of Contract No. 68-03-0435.
The contents of this report are reproduced herein as received from Calspan Corporation. The opinions, findings, and conclusions expressed are those of the author and not necessarily those of the Environmental Protection Agency. Mention of company or product names is not to be considered as an endorsement by the Environmental Protection Agency.

Publication No. EPA-460/3-74-024

ABSTRACT

This report on modal analysis of automobile emissions was prepared for the United States Environmental Protection Agency, Division of Certification and Surveillance, Ann Arbor, Michigan, under EPA Contract No. 68-03-0435. The work reported herein constitutes a refinement and extension of a modal analysis exhaust emission model originally developed under EPA Contract No. 68-01-0435. This earlier effort was released as EPA-460/3-74-005, "Automobile Exhaust Emission Modal Analysis Model".

The modal analysis exhaust emission model makes it possible to calculate the amounts of emission products emitted by individual vehicles or groups of vehicles over an arbitrary driving sequence. Refinements to the model permit an improvement in computational efficiency and a reduction in input data requirements. Extensions of the model include a scheme for computation of fuel usage in terms of CO_2, CO and HC output by means of a carbon-balance approach and a procedure for more definitive assessment of the precision of the model in predicting group emissions.

ACKNOWLEDGEMENTS

Support and encouragement by the Environmental Protection Agency is gratefully acknowledged. Particular thanks go to C.J. Domke, M.E. Williams, and L.A. Platte for their guidance and suggestions.

A number of individuals at Calspan Corporation contributed significantly to the work. These included P.E. Yates, who performed much of the computer analysis, and A.C. Keller, who provided valuable suggestions and criticism.

Finally, special recognition goes to Paul Kunselman, formerly of Calspan, who was a prime mover in the development of the original model and whose insight was instrumental in the follow-on refinements and improvements.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Under U.S. Environmental Protection Agency Contract No. 68-01-0435, Calspan Corporation formulated a model for the prediction of motor vehicle exhaust emissions over an arbitrary driving sequence.* The work reported herein was performed under EPA Contract No. 68-03-0435 as a refinement and extension of the original model. Subsequent discussion will assume familiarity with the original model as presented in EPA-460/3-74-005, Automobile Exhaust Modal Analysis Model (January 1974);** however, wherever essential to understanding, details of the model will be repeated in the present report for the sake of clarity.

The impact of motor vehicle exhaust emissions on air quality in a given location depends on a number of factors: the emission characteristics of individual vehicles, the mix of vehicles of different types operating in the location, the numerical density of vehicles per mile or per unit of area, and the driving pattern in which the vehicles are employed. To assess the contribution of motor vehicles to air pollution, therefore, it is necessary to estimate traffic density, composition and flow characteristics, and to have some means for expressing these quantities in terms of pollution burden to the atmosphere.

The required traffic parameters can be estimated in a straightforward way. Traffic in the vicinity can be monitored and classified according to vehicle make, model, age, and other factors known to influence emissions. Moreover, speeds and accelerations prevailing along the traffic way in question can be measured and tabulated. Unless emissions can be expressed as functions of the applicable traffic parameters, however, it is not possible to assess vehicular contributions to air pollution.

---

* Paul Kunselman and H.T. McAdams, Automobile Exhaust Emission Modal Analysis Model, Calspan Report No. NA-5194-D-3 (July 1973).

** Paul Kunselman, H.T. McAdams, C.J. Domke, and Marcia Williams, Automobile Exhaust Emission Modal Analysis Model, Environmental Protection Agency Report No. EPA-460/3-74-005 (January 1974).

The emission tests used for certification of new light duty motor vehicles are based on a prescribed driving sequence by means of which vehicles can be compared according to a standard set of operating conditions. Though this concept of a standard driving sequence makes it possible to implement emission standards and to check compliance with these standards, the concept does not facilitate the prediction of vehicle emissions over an arbitrary driving sequence. By breaking the standard sequence into segments (modes) having specified speeds and accelerations, however, and noting the emissions produced in each segment, it was postulated that these segments might be recombined appropriately to form other driving sequences of interest. Ultimately, it was hoped that this process might lead to a model for defining emissions as continuous functions of vehicle operating conditions and thus make it possible to approximate emissions over any driving sequence of interest.

As developed by Calspan under EPA Contract No. 68-01-0435, the original modal analysis prediction model was based on the concept of an instantaneous emission rate for each of the primary pollutants carbon monoxide (CO), hydrocarbons (HC), and oxides of nitrogen (NOX). In this model, it was assumed that the instantaneous emission rate can be adequately defined as a function $\dot{e} = f(v, a)$ of instantaneous speed, v, and acceleration, a, for each vehicle. Since every point in time over a driving sequence has an associated instantaneous speed and acceleration, the total emission over the driving sequence can be obtained by appropriate integration of the emission rate function. Moreover, by virtue of the mathematical form of the model, it can be advantageously used to predict emissions from either homogeneous or nonhomogeneous groups of vehicles.

Initial experience with the modal analysis prediction model suggested that it be refined and extended with the following objectives in mind:

1) Investigate means to increase the computational efficiency of the model.

2) Determine whether modal testing requirements can be reduced without appreciable loss of information.

3) Define the accuracy and precision with which group emission predictions can be made from modal data.

4) Use the modal analysis approach to predict fuel economy over arbitrary driving sequences.

Each of these areas of investigation will be discussed in turn in subsequent sections of this report.

2. MODEL COMPUTATIONAL EFFICIENCY

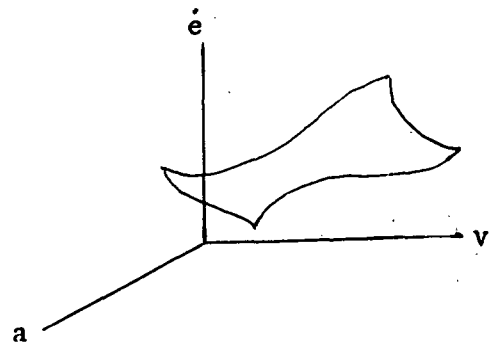
Relative to the original formulation of the modal analysis emission model, a significant increase in computational efficiency can be achieved by a simplification of the method by which the instantaneous emission rate function, $\dot{e}$ (v, a) , is integrated over a driving sequence. As background for this simplification, however, it will be instructive to review the essential features of the model.

2.1 ESSENTIAL FEATURES OF THE BASIC MODEL

Inputs to the model are based on the Surveillance Driving Sequence (SDS), in which emissions are measured over a variety of steady state and transient driving conditions. The acceleration and deceleration modes represented in the SDS consist of all possible combinations of the following five speeds: 0 mph, 15 mph, 30 mph, 45 mph, 60 mph. The average acceleration or deceleration rate observed for each mode in the Los Angeles basin is used during operation of 20 of the transient modes. In addition, 6 of the transient modes are repeated using accel/decel rates higher or lower than the average rate in order to determine the effect of accel/decel rate on emissions.

A difficulty presented by the use of the 37 discrete modes as inputs to a continuous driving sequence model is that, during much of the sequence, the vehicle may be operating at velocities and accelerations not included in the set of five steady state and 32 accel/decel modes. For example, a vehicle traveling at 23 mph is neither in the 15 mph nor 30 mph steady state mode. To arrive at a continuous predictive model, one must be able to interpolate or otherwise estimate the appropriate emission rates for all combinations of speed and acceleration encountered in the driving sequence.

The primary feature of the model is a scheme whereby emissions from the 37 discrete modes can be expanded into a continuous function of time. For this purpose, use is made of a regression function which can, for purposes of visualization, be represented as a "surface" in speed-acceleration space as shown below.



Emission Response Surface

For any point (v, a) in the speed-acceleration plane, there corresponds an instantaneous emission rate $\dot{e}$ (v, a) . The surface can be represented by a mathematical equation of the form: $\dot{e} = f(v, a)$ in which the function f contains a number of adjustable constants. These constants can be selected to represent the emission characteristics of a particular automobile or can be selected to represent the mean emission characteristics of a collection of automobiles.

The mass of a particular pollutant emitted by an automobile is a cumulative, non-decreasing function of time, e(t). The time derivative of this function yields the instantaneous emission rate as a function of time:

$$\dot{e}(t) = \frac{d[e(t)]}{dt} \tag{1}$$

In the modal analysis model, it is assumed that the instantaneous emission rate is a function of vehicle speed and acceleration, both of which are functions of time. Thus,

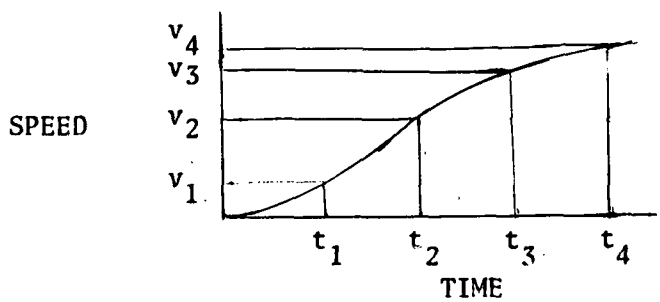
$$\dot{e}(t) = \dot{e}[v(t), a(t)] \tag{2}$$

and

$$e(T) = \int_0^T \dot{e}[v(t), a(t)]\, dt \tag{3}$$

gives the mass of pollutant given off by a vehicle in a driving sequence lasting T seconds. Evaluation of the above integral requires (1) specifying the driving sequence in terms of v(t) and a(t) , and (2) specifying the emission-rate function in terms of speed and acceleration.

In practice, a driving sequence is specified in terms of the speed prevailing at each of n discrete, equally spaced points in time, as shown below.



The integration of equation (3) is then approximated by the summation:

$$e(T) = \sum_{i=1}^{n-1} \dot{e}(\hat{v}_i, \hat{a}_i) \; \Delta t \tag{4}$$

where

$$\hat{v}_i = \frac{v_{i+1} + v_i}{2}$$

$$\hat{a}_i = \frac{v_{i+1} - v_i}{\Delta t}$$

and

$$n \Delta t = T$$

The applicable emission-rate function is developed by application of a generalized version of multiple regression analysis.

As a starting point for development of the multiple regression equation for emission rate as a function of speed and acceleration, it will be instructive to consider first a steady-state emission rate function defined for constant speed (zero acceleration) only. It is assumed that this function can be expanded in the form

$$\dot{e}_S(v) = a_1 f_1(v) + a_2 f_2(v) + \ldots + a_k f_k(v) \tag{5}$$

where $a_1, a_2, \ldots, a_k$ are constants applicable to a specific automobile or group of automobiles, and $f_1(v), f_2(v), \ldots, f_k(v)$ are referred to as basis functions. It is emphasized that these functions can assume any form consistent with the data to be represented, the only requirement being that they be linearly independent and not contain any adjustable constants dependent on the data. The latter requirement assures that, for a given choice of basis functions, the function $\dot{e}_S(v)$ is completely defined by the model coefficients $a_1, a_2, \ldots, a_k$.

In a similar vein, an emission rate function $\dot{e}_A(v, a)$ can be postulated for non-steady-state operation in which $a \neq 0$. It is assumed that this function can be expanded as

$$\dot{e}_A(v, a) = b_1 g_1(v, a) + b_2 g_2(v, a) + \ldots + b_r f_r(v, a) \qquad (6)$$

where $b_1, b_2, \ldots, b_r$ are constants applicable to a specific automobile or group of automobiles and the basis functions $g_1(v, a), g_2(v, a), \ldots, g_r(v, a)$ are, as before, linearly independent and free of any adjustable constants to be determined from the data.

As an extension of equations (5) and (6), it is logical to postulate that, by appropriate definition of basis functions, it should be possible to define an emission rate function $\dot{e}(v, a)$ applicable over the entire (v, a)-plane regardless of whether $a = 0$ or $a \neq 0$. Such a universally applicable equation might assume the form

$$\dot{e}(v, a) = c_1 u_1(v, a) + c_2 u_2(v, a) + \ldots + c_s u_s(v, a) \qquad (7)$$

where $c_1, c_2, \ldots, c_s$ are constants applicable to a specified automobile or group of automobiles and the basis functions $u_1(v, a), u_2(v, a), \ldots, u_s(v, a)$ are linearly independent and contain no constants to be determined from the data. In the original development of the modal analysis model, however, it was found advantageous to develop the instantaneous emission rate function $\dot{e}(v, a)$ as a composite function

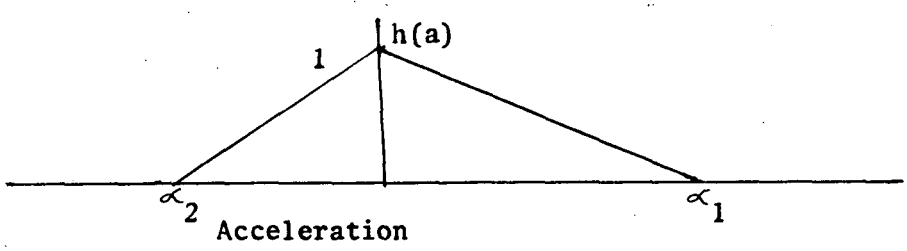
$$\dot{e}(v, a) = h(a)\, \dot{e}_S(v) + [1-h(a)]\, \dot{e}_A(v, a) \qquad (8)$$

where $h(a)$ is a weighting function bounded in the interval $0 \leq h(a) \leq 1$.

As employed in the original form of the model, the function h(a) was defined as follows:

$$h(a) = \begin{cases} -\frac{1}{\alpha_1}\, a+1, & \alpha_1 > a > 0 \\ -\frac{1}{\alpha_2}\, a+1, & \alpha_2 < a < 0 \\ 0, & a \geq \alpha_1 \\ 0, & a \leq \alpha_2 \end{cases}$$

or



By specifying the constants α_1 and α_2 the weightings of the two rate functions will vary between 0 and 1 in a continuous manner when the transition is made between accel/decel and steady state periods of driving.

Once sets of basis functions have been established for equations (5) and (6), the coefficients which define the instantaneous emission rate function could be determined by a straightforward application of least squares theory provided that instantaneous emission rates were known for a sufficient number of (v,a)-positions in the speed-acceleration plane. In reality, however, the data base for vehicle emissions does not contain any instantaneous emission rate observations for accel/decel modes; instead, the observations reported are the total amounts of pollutant collected over each mode and it is possible to calculate only the <u>average</u> emission rate prevailing during the time in mode. In this connection, however, it can be shown that for a postulated form of the emission rate function, it is possible to deduce the applicable model coefficients from the modal average emission rates.

To illustrate this point, consider a situation in which the instantaneous emission rate can be adequately expressed as a linear combination of three basis functions $g_1(v, a)$, $g_2(v, a)$, and $g_3(v, a)$. Then,

$$\dot{e}(v, a) = b_1 g_1(v, a) + b_2 g_2(v, a) + b_3 g_3(v, a) \tag{9}$$

Consider a mode of time duration T. The average emission rate over time T can be computed from equation (9) as

$$\langle \dot{e}(v, a) \rangle_T = \frac{1}{T} \int_0^T \dot{e}[v(t), a(t)]\, dt \tag{10}$$

and from the observed total emission over the mode as $\frac{1}{T} e(T)$, where $e(T)$ is the "bag value" for the mode and $v(t)$ and $a(t)$ are the speed vs time and acceleration vs time profiles for the mode in question. Then,

$$\langle \dot{e}(v, a) \rangle_T = \frac{1}{T} e(T) = \frac{1}{T} \int_0^T \dot{e}[v(t), a(t)]\, dt$$

or

$$\langle \dot{e}(v, a) \rangle_T = \frac{1}{T} \int_0^T \Big\{ b_1 g_1[v(t), a(t)] + b_2 g_2[v(t), a(t)] + b_3 g_3[v(t), a(t)] \Big\}\, dt \tag{11}$$

Termwise integration and removal of the constants b_1, b_2 and b_3 from the integrand yields

$$\begin{aligned} \langle \dot{e}(v, a) \rangle_T = {} & b_1 \left\{ \frac{1}{T} \int_0^T g_1[v(t), a(t)]\, dt \right\} \\ & + b_2 \left\{ \frac{1}{T} \int_0^T g_2[v(t), a(t)]\, dt \right\} \\ & + b_3 \left\{ \frac{1}{T} \int_0^T g_3[v(t), a(t)]\, dt \right\} \end{aligned} \tag{12}$$

Note, however, that the bracketed expressions are just the time averages of the basis functions over the time duration of the mode. Thus, one can write

$$\frac{e(t)}{T} = b_1\bar{g}_1 + b_2\bar{g}_2 + b_3\bar{g}_3$$

where $\bar{g}_1$, $\bar{g}_2$ and $\bar{g}_3$ are, respectively, the time averages for g_1, g_2 and g_3 over the mode in question. Since the total emissions for each mode are known, as well as the corresponding times in mode, the time averages $\bar{g}_1$, $\bar{g}_2$ and $\bar{g}_3$ can be computed for each mode and the coefficients b_1, b_2 and b_3 can be obtained through least squares regression analysis applied to the average emission rates as computed from modal bag values.

In the context of model refinement and extension, the modal analysis model as originally developed under EPA Contract No. 68-01-0435 should be viewed as a family or continuum of models. Though initial application of the modal analysis model concept employed a specific set of basis functions, the model in a broad sense is amenable to infinite variety in the choice of basis. Indeed, choice of basis may itself present an avenue for model simplification and for increased computational efficiency. Every attempt should be made to keep the number of basis functions to a minimum and to employ the simplest basis functions compatible with the data.

In this connection, it is of interest to review the reasoning by which the basis functions for the original model were derived. In the steady-state (zero acceleration) case, the emission rate is a function of speed only. For each of the three pollutants (CO, HC, NOX), steady-state emission rates were averaged over the 1020 vehicles constituting the data base, and these average emission rates were then plotted as a function of speed. These plots suggested that the steady state emission rate function e_s could be expressed as a quadratic function of speed:

$$\dot{e}_s(v) = s_1 + s_2 v + s_3 v^2 \qquad (13)$$

where s_1, s_2 and s_3 are constants.

In the case of non-zero acceleration, it was assumed that the acceleration occurring at a given speed is a perturbation to the steady-state emission rate at this speed. This perturbation can be accounted for by expressing the coefficients s_1, s_2 and s_3 as functions of acceleration. If it is assumed that quadratic functions of acceleration represent good approximations to these coefficients, the coefficients can be expressed as follows:

$$\left.\begin{aligned} s_1 &= s_1(a) = q_{11} + q_{12}a + q_{13}a^2 \\ s_2 &= s_2(a) = q_{21} + q_{22}a + q_{23}a^2 \\ s_3 &= s_3(a) = q_{31} + q_{32}a + q_{33}a^2 \end{aligned}\right\} \tag{14}$$

where the q's are constants. The emission rate function used during times of non-zero acceleration $\dot{e}_A$ can then be written in the form:

$$\begin{aligned} \dot{e}_A(v, a) = b_1 + b_2v + b_3a + b_4av + b_5v^2 + b_6a^2 \\ + b_7v^2a + b_8a^2v + b_9a^2v^2 \end{aligned} \tag{15}$$

where the b's are constants and can be expressed in terms of the q's. It is noted that if $a = 0$ equation, (15) reduces to

$$\dot{e}_A(v) = b_1 + b_2v + b_5v^2 \tag{16}$$

which has the identical form as the equation for $\dot{e}_S$. Thus, in principle, $\dot{e}_A$ could be used to determine emissions for both steady state and non-zero acceleration periods. As noted earlier in the discussion, however, it was found advantageous to express instantaneous emission rate as a composite function

$$\dot{e}(v, a) = h(a)\ \dot{e}_S(v) + \left[1-h(a)\right]\ \dot{e}_A(v, a) \tag{17}$$

in which $\dot{e}_S(v)$ is determined independently of $\dot{e}_A(v, a)$. In this way, the model is provided with greater flexibility, especially in the vicinity of zero acceleration, since it has 12 rather than 9 adjustable coefficients for defining the instantaneous emission rates for each pollutant.

2.2 SIMPLIFICATION OF EMISSION INTEGRATION OVER DRIVING SEQUENCES

In the original version of the modal analysis model, computation of total emissions over a driving sequence of time-duration T was achieved by performing the integration

$$e(t) = \int_0^T \dot{e}\left[v(t),\ a(t)\right]\ dt \tag{18}$$

for each vehicle or group of vehicles of interest. As will become apparent below, however, the integral can be reformulated in such a way that, for a particular driving sequence, a single integration suffices for all vehicles subjected to that sequence.

The composite emission function, as shown in equation (17), can be written in terms of the basis functions $f_i(v)$ and $g_j(v,\ a)$. Noting that

$$\dot{e}_S(v) = \sum_{i=1}^{k} a_i f_i(v) \tag{19}$$

and

$$\dot{e}_A(v,\ a) = \sum_{j=1}^{r} b_j g_j(v,\ a) \tag{20}$$

one can substitute (19) and (20) for $\dot{e}_S(v)$ and $\dot{e}_A(v,\ a)$ in (17) and integrate to obtain

$$e(t) = \int_0^T h\left[a(t)\right] \sum_{i=1}^{k} a_i f_i\left[v(t)\right]\ dt + \int_0^T \left\{1-h\left[a(t)\right]\right\} \sum_{j=1}^{r} b_j g_j\left[v(t),\ a(t)\right]\ dt \tag{21}$$

In view of the fact that a_i $(i = 1,2,\ldots,k)$ and b_j $(j = 1,2,\ldots,r)$ are constants, (21) can be rewritten as

$$e(T) = \sum_{i=1}^{k} a_i \int_0^T h[a(t)]\, f_i[v(t)]\, dt + \sum_{j=1}^{r} b_j \int_0^T \{1-h[a(t)]\}\, g_j[v(t), a(t)]\, dt \qquad (22)$$

Note that (22) contains k integrals of the form

$$\int_0^T h[a(t)]\, f_i[v(t)]\, dt, \qquad i = 1, 2, \ldots, k \qquad (23)$$

and r integrals of the form

$$\int_0^T \{1-h[a(t)]\}\, g_j[v(t), a(t)]\, dt, \qquad j = 1, 2, \ldots, r \qquad (24)$$

The integrands of (23) and (24) are just weighted forms of the basis functions and do not depend on the magnitudes of the coefficients a_i and b_j. Consequently, once these k+r quantities have been computed for a given driving sequence, it is necessary to know only the applicable model coefficients a_i and b_j in order to compute emissions for a particular automobile or group of automobiles negotiating that driving sequence.

For the choice of basis functions employed in the original model, k = 3 and r = 9. Therefore, for each pollutant there are 12 integrals to be evaluated. These 12 quantities can be combined with the coefficients a_i and b_j to compute the mass of pollutant emitted by a particular automobile or group of automobiles in performing the specified driving sequence. Subroutine ESUM of the original model has been revised to integrate the weighted basis functions over a specified driving sequence and return the results to the main program where the total emission is calculated. The revised versions of ESUM and the main program are given in Appendix II.

2.3 HYPSOMETRIC ANALYSIS OF DRIVING SEQUENCES

As shown above, computational efficiency of the model can be improved by noting that, for a given driving sequence, the integrated forms of the basis functions are invariant and do not need to be recalculated unless a different driving sequence is postulated. Further efforts to improve efficiency were aimed at a hypsometric characterization of driving sequences.

Hypsometry is a term used in geodesy to characterize the measurement of surface elevation. In particular, the hypsometric integral is a function used to quantify that fraction of a geographic area which exceeds a given threshold level, where the threshold level can be regarded as a continuous variable. As applied to the modal analysis model, the hypsometric integral would provide a characterization of the relative frequency of occurrence of various speed and acceleration levels.

In the original form of the model, the driving sequence is described by specifying the speed for each time increment (generally one second) in the sequence. This specification, in turn, establishes the acceleration during each increment of time. It should be noted, however, that the computed contribution to emissions during a particular time increment depends only on the speed and acceleration prevailing during that time interval and is independent of the speed-time history of the vehicle. In short, a particular combination of speed and acceleration is regarded as making the same contribution to the pollutant output of a vehicle regardless of whether that speed-acceleration combination occurs early or late in the overall driving sequence. In view of this fact, it appeared feasible to describe the speed-time history of a driving sequence in terms of the joint frequency distribution of speed and acceleration. It was further postulated that, for "typical" driving sequences--e.g., urban or rural--it might be possible to express the distribution functions in terms of a few adjustable parameters. For example, if speeds and accelerations for a particular sequence essentially were to obey a bivariate normal distribution, then specifying the means, variances and covariance of speed and acceleration would suffice to describe the distribution. A useful application of such parametric description of driving sequences might be in

characterizing the various branches of a road network hypsometrically, so that pollution abatement studies aimed at optimizing routes in a network might be more amenable to analysis.

As far as computation of the total emissions e(T) over a driving sequence of T seconds is concerned, implications of the hypsometric analysis of speed and acceleration would be felt through the functions v(t) and a(t) in equation (22). In view of this fact, and in view of the readiness with which the weighted basis functions can be computed, no further development of the hypsometric description of driving sequences was pursued. In reality, the values of the k+r integrals in equations (23) and (24) constitute a complete description of the driving sequence so far as the model is concerned, and, within the limits of validity of the model, completely characterize the effect of the driving sequence or "route" on emissions. Similarly, the values of the k coefficients a_i (i = 1,2,... k) and the r coefficients b_j (j = 1,2,...,r) completely characterize, again within the limits of model validity, the "vehicle effect" on emissions for that particular route.

2.4 VEHICLE AND DRIVING SEQUENCE AS VECTORS AFFECTING EMISSIONS

The relation between vehicle and driving sequence (route) effects on total emissions e(T) can be expressed succinctly in vector notation.

Let the values of the k+r integrals in equations (23) and (24) be considered as components of a (k+r)-dimensional driving sequence vector

$$\underline{S} = (s_1, s_2, \ldots, s_k, s_{k+1}, \ldots, s_{k+r}) \tag{25}$$

Similarly, let the k coefficients a_i (i = 1,2,..., k) and the r coefficients b_j (j= 1,2,..., r) be considered as the components of a (k+r)-dimensional vehicle vector

$$\underline{V} = (a_1, a_2, \ldots, a_k, b_1, b_2, \ldots, b_r) \tag{26}$$

Then, the total emissions e(T) for a particular vehicle operating according to a specified driving sequence can be expressed as the vector inner (dot) product

$$e(T) = \underline{S} \cdot \underline{V} \tag{27}$$

It is to be observed that when the driving sequence consists of a single mode, either steady-state or accel/decel, the dot product in equation (27) reduces to a computed estimate of the bag value for that mode. Also, it should be noted that the vehicle vector (26) can represent a group of vehicles rather than a single automobile.

The vector form of the model, as elucidated above, can be further systematized to consider the effects on emissions of various mixes of vehicles and various driving sequences (routes). Let $\underline{s}_1, \underline{s}_2, \ldots, \underline{s}_p$ denote the sequence vectors for p alternative driving sequences, and let $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_q$ denote the the vehicle vectors for q alternative mixes of vehicles. Then, if the vectors $\underline{s}_1, \underline{s}_2, \ldots, \underline{s}_p$ are considered as columns of a matrix S and if the vectors $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_q$ are considered as columns of a matrix V, then one can write

$$E = S' V \tag{28}$$

where S is a matrix of order p x (k+r), V is a matrix of order (k+r) x q, and E is a matrix of order p x q. The matrix E consists of elements e_{ij} (i = 1, 2,..., p; j = 1, 2,..., q) which provide estimates of the total emissions generated by the jth mix of vehicles operating according to the ith driving sequence.

3. MODAL TESTING REQUIREMENTS

As originally implemented, the modal analysis model employed three basis functions in the steady-state portion of the emission-rate equation and nine basis functions in the accel/decel portion of the equation. The fact that the resulting 12 regression coefficients are considerably fewer than the 37 modes used as data inputs to the model suggests that there is a certain amount of redundancy in the modal data. On the other hand, there are regions of the speed-acceleration plane not adequately represented by modal data, and this fact could occasion unwarranted imprecision in the performance of the model, particularly in those regions of the speed-acceleration plane where modes are sparse. A revised allocation of speeds and accelerations by modes, as well as a possible reduction in the number of modes, is therefore suggested, provided this reallocation and/or reduction does not adversely affect other aspects of the emission-measurement protocol.

Several techniques were employed to examine the implications of reallocation of modal test points in the speed-acceleration plane. These included visual examination of the modal-distribution pattern, the computation of variance maps indicative of error propagation over the (v, a)-plane, and principal component analysis of modal contributions to the model performance.

Figure 1 is a plot of the average speeds and accelerations for the 32 acceleration/deceleration modes and the 5 steady-state modes which constitute inputs to the emission model as originally formulated. The sparse or empty regions of the speed-acceleration plane are clearly evident, particularly that portion of the plane between accelerations of -1 mph/sec and +1 mph/sec. As will become apparent later, the lack of information in this region of the plane tends to exaggerate the uncertainty of prediction in that region and is, at least in part, the reason that the steady-state and accel/decel portions of the model must be bridged in a rather arbitrary way in the original model.

In view of the fact that for each mode a speed is specified for each second of time in mode, it is possible to estimate the corresponding accelerations on a second-by-second basis and to plot acceleration versus speed profiles for each mode.

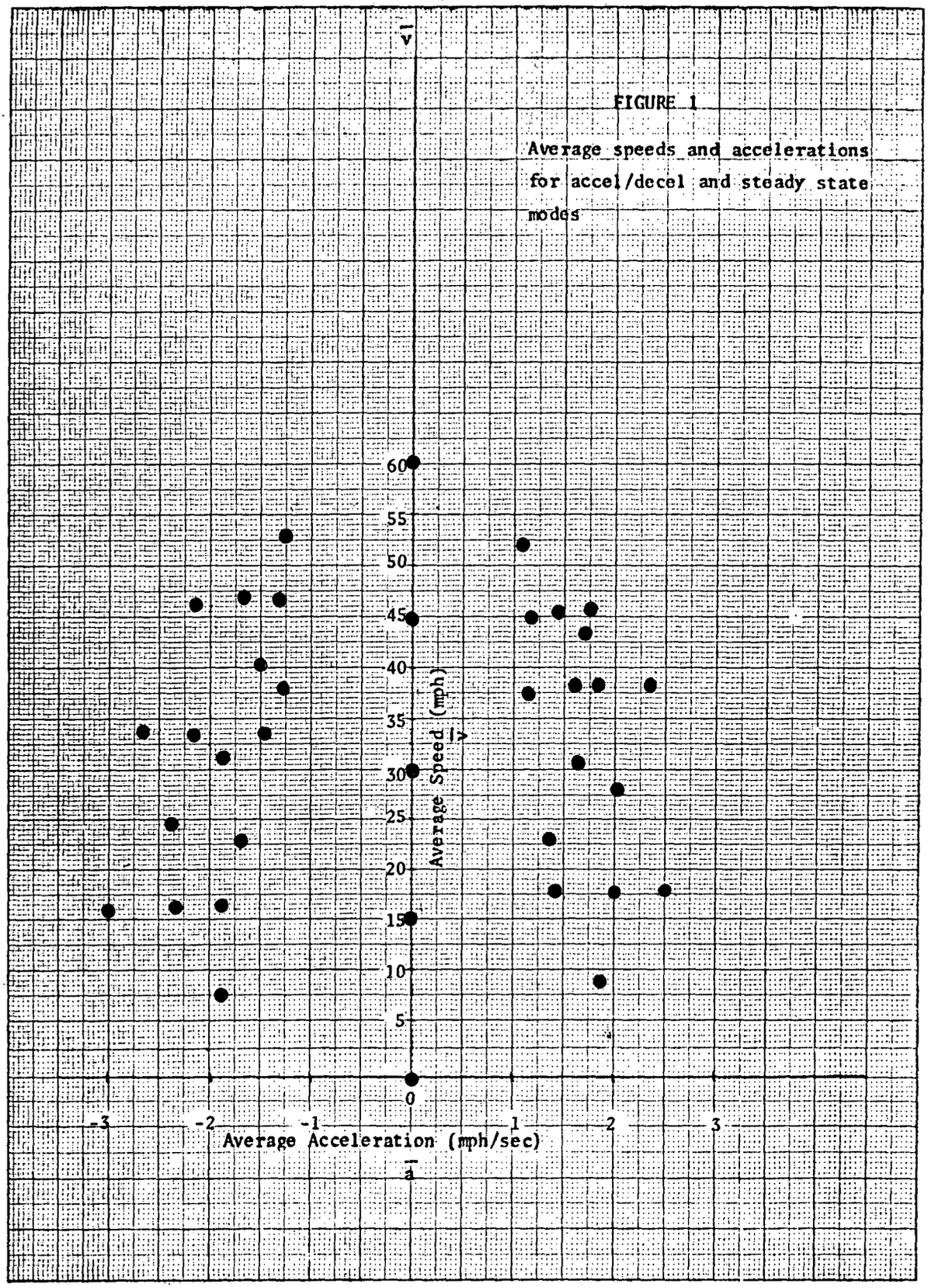


FIGURE 1

Average speeds and accelerations for accel/decel and steady state modes

Let us consider, however, the second-by-second schedules maintained in the various modes and plot acceleration/speed profiles on this basis. Figure 2 is such a plot for mode 23, an acceleration mode, and Figure 3 is such a plot for mode 26, a deceleration mode. As is to be expected, these plots show that actual speeds and accelerations realized over short time increments in these modes span regions of the acceleration-speed plane not represented if only the modal averages are considered. This fact is made clear by Figure 4, which is a composite plot of second-by-second accelerations and speeds achieved when results of all 32 accel/decel modes are combined. The plot suggests that many of the gaps shown in Figure 1 might be filled in if appropriate speeds and accelerations in Figure 4 can be regrouped and averaged to present a revised set of modes more advantageous as model inputs.

Consider the time plot of mode 23, as shown in Figure 5. A noticeable degree of asymmetry in this plot is evident. For example, the early part of the mode exhibits greater accelerations than the latter part of the mode, and this fact suggests that the mode might be divided into two parts so as to provide a model input to fill part of the gap presently existing in the low-acceleration region.

A scheme for examining this concept is as follows. Compute average speed and acceleration for the first n seconds in mode and for the remaining N-n seconds in mode, where N is the total number of seconds in mode. Plot these two results as functions of n over the region 0 n N, as shown in Figure 6. This plot provides a set of options for redefinition of the mode so as to more adequately span the speed-acceleration plane. By electing various options for redefinition of the modal inputs to the model, one can examine the consequences of this redefinition by means of the variance-function concept explained below.

3.2 VARIANCE-FUNCTION ASSESSMENT OF ERROR PROPAGATION

For a particular pollutant and for a particular vehicle or group of vehicles, the emission measured for each of the 5 steady-state and 32 accel/decel modes can be regarded as a random variable. In other words,

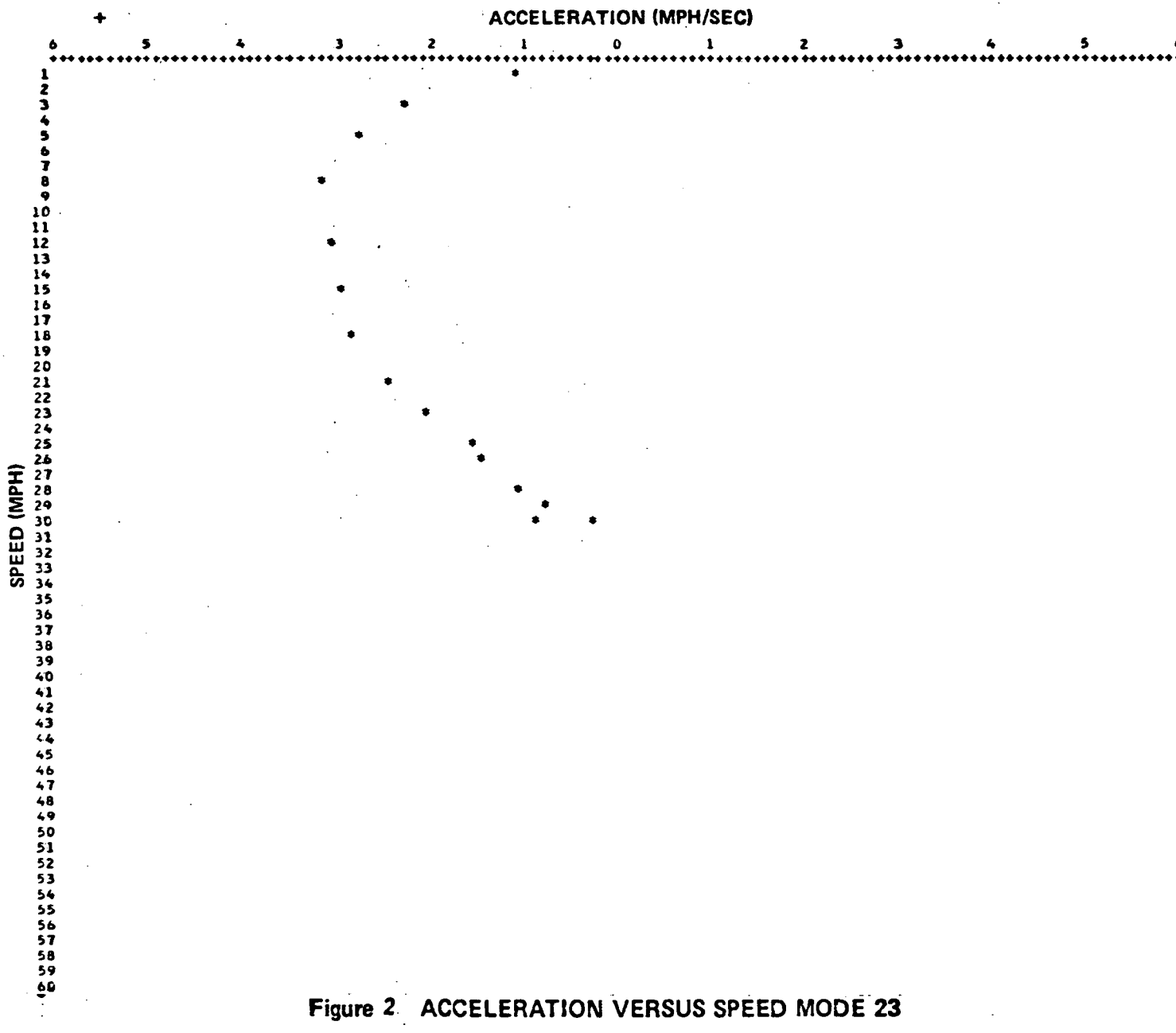


Figure 2 ACCELERATION VERSUS SPEED MODE 23

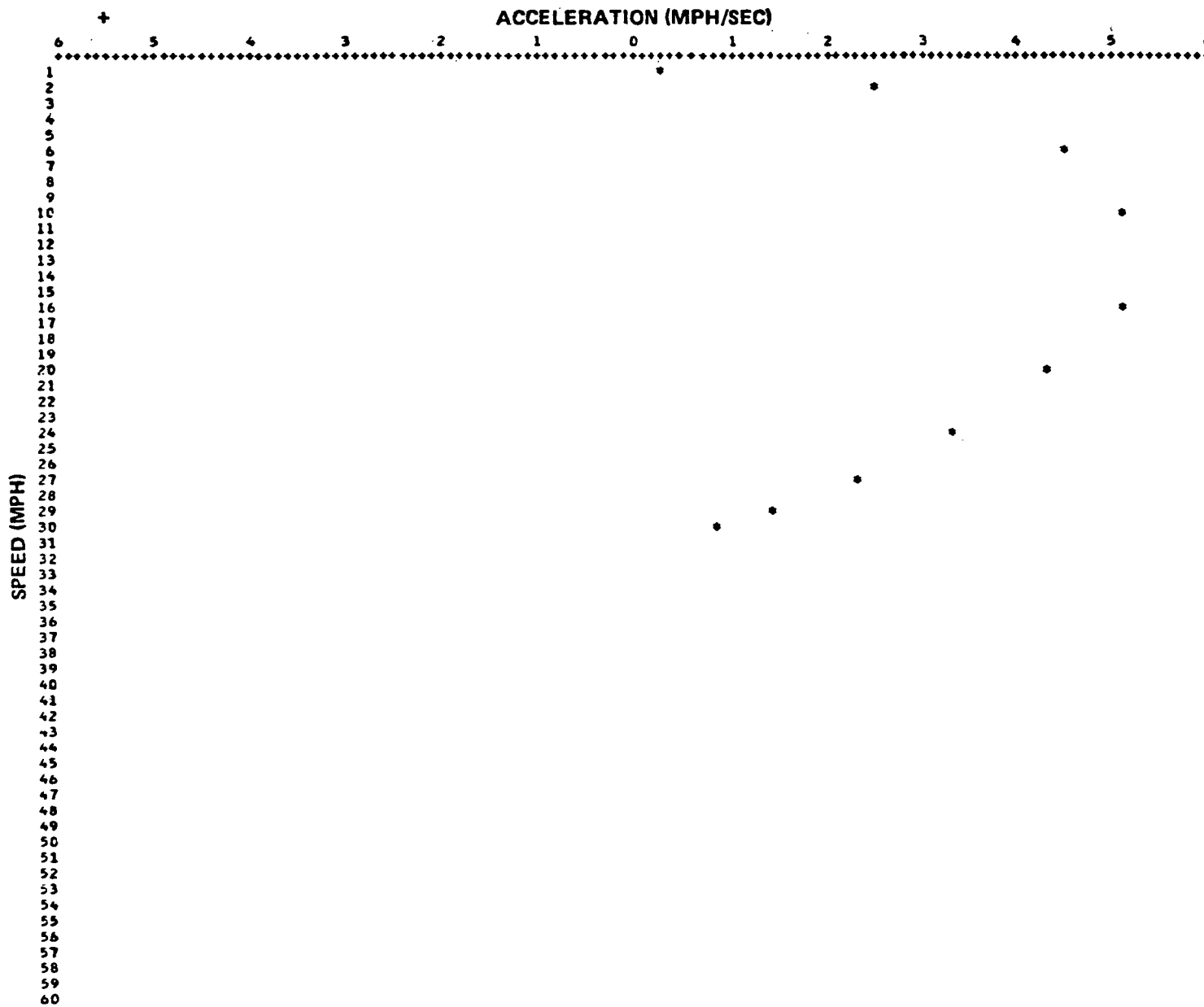


Figure 3 ACCELERATION VERSUS SPEED MODE 26

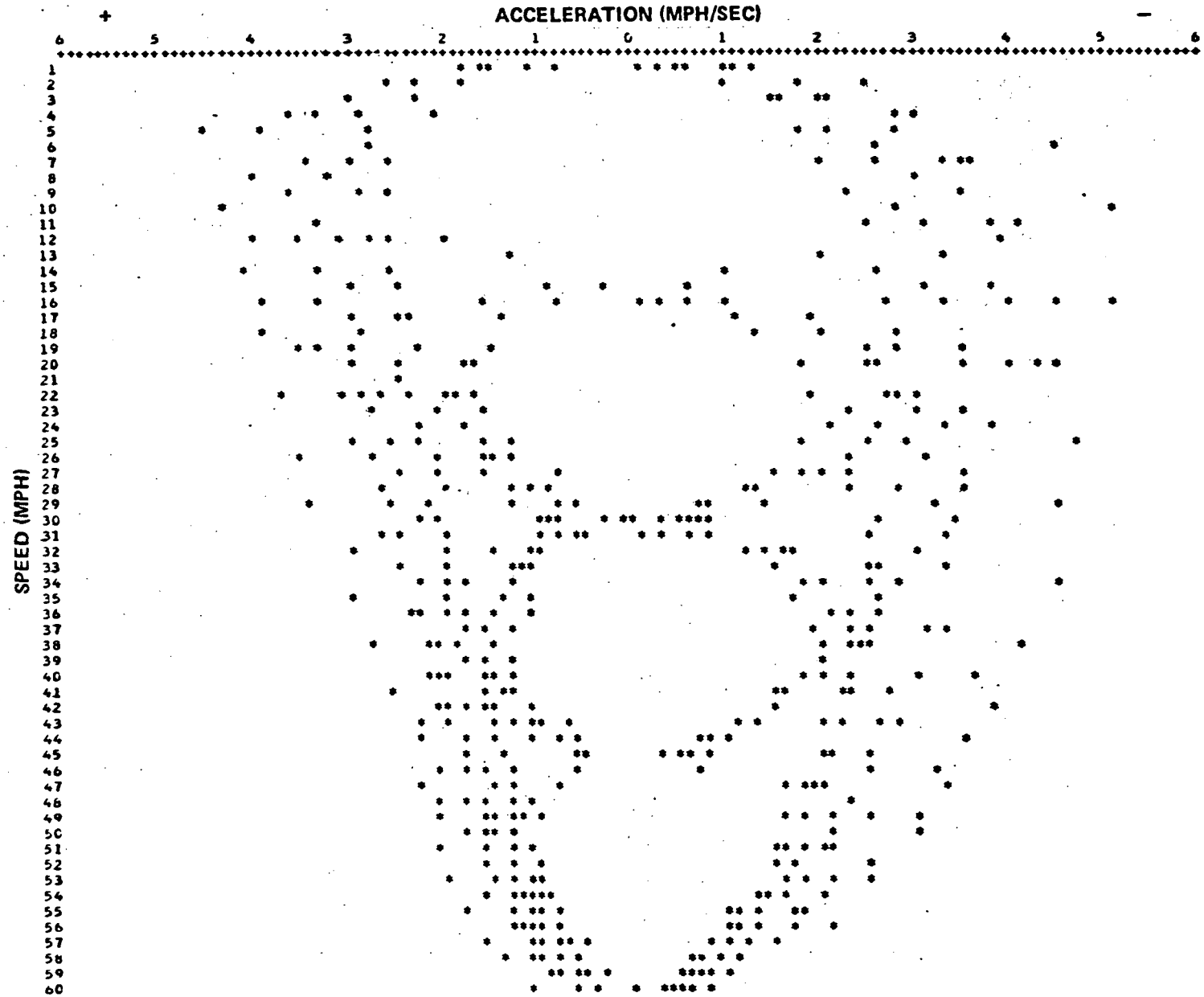


Figure 4 ACCELERATION VERSUS SPEED COMPOSITE FOR ALL ACCEL/DECEL MODES

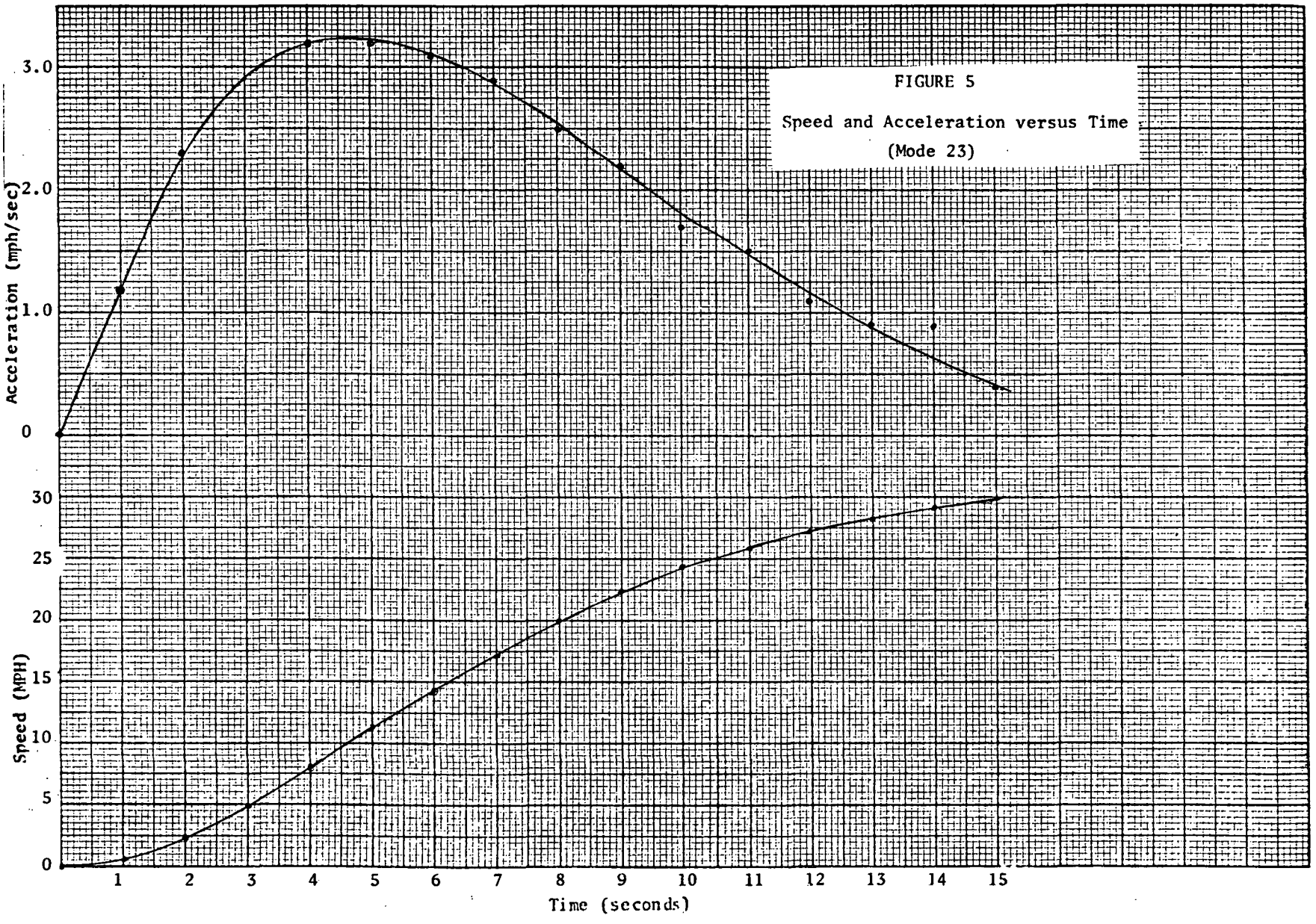


FIGURE 5

Speed and Acceleration versus Time

(Mode 23)

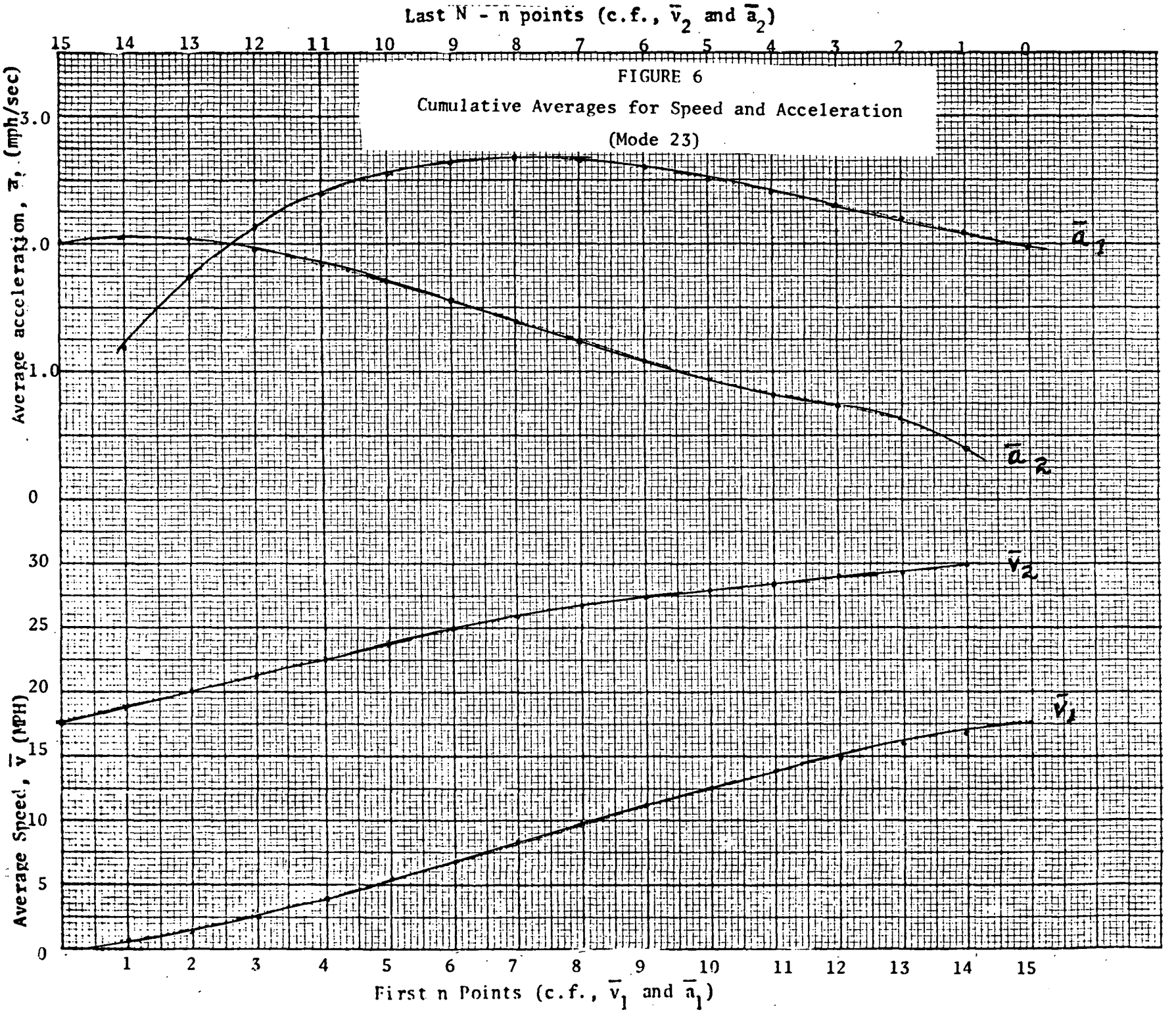


FIGURE 6

Cumulative Averages for Speed and Acceleration

(Mode 23)

the measurement of the modal bag values is subject to error, a fact well demonstrated by the inability to obtain the same emission mass measurements on repeated or replicate tests. Each measurement, therefore, can be regarded as being subject to a certain variance. This variance can be expected to propagate through the regression model to induce uncertainty in the emission estimates computed at every point in the (v, a)-plane. The magnitude of this uncertainty varies as a function of position in the (v, a)-plane and can thus be regarded as a variance function of speed and acceleration. Conceptually, this function can be viewed as a variance "surface" and can be graphically portrayed by means of variance contours. The variance function can be computed if the basis functions of the emission-rate function are specified,if the locations of the modal input points in the (v, a)-plane are known,and if there is available an estimate of the error variance for each of the input-mode bag values.

The functional representation of the emission rate function used in the automobile emission model is given by the weighted composite of the accel/decel and steady-state instantaneous emission rate functions, $\dot{e}_A$ and $\dot{e}_S$ respectively:

$$\dot{e}(v, a) = \omega\ \dot{e}_S(v) + (1-\omega)\ \dot{e}_A(v, a) \qquad (29)$$

where ω is a weighting function dependent upon acceleration.

The accel/decel and steady state instantaneous emission rate functions are expressed as linear combinations of basis functions of speed and acceleration. In general, the linear model which gives the true response, y , of a vehicle or group of vehicles is given by:

$$y = \beta_1 f_1 + \beta_2 f_2 + \beta_3 f_3 + \ldots + e \qquad (30)$$

where β_i , $i = 1,2,3 \ldots$ are constants, f_1 , f_2 , f_3 ... are the basis functions of velocity and acceleration, and e is the random error.

Since e is a random variable, the responses observed at each (v, a) point also constitute a random variable. As a result, it is only possible to obtain from the observations an equation of the form:

$$\hat{y} = b_1 f_1 + b_2 f_2 + b_3 f_3 + \ldots \quad (31)$$

where $\hat{y}$ is an estimate of y and b_i, i = 1,2,3 ... is an estimate of β_i. The estimated responses predicted by the model are currently based on measurements of bag values at 32 accel/decel and 5 steady state average velocity/acceleration points of the Surveillance Driving Sequence.

A detailed explanation of the method of computing the variance function for regression estimates is given in Appendix II and will not be duplicated here. Suffice it to say that the variance function is controlled by three considerations:

1) The type of basis functions employed in the regression model.

2) The positions in the (v, a)-plane, called design points, at which modal emission measurements are taken, and

3) The magnitude of the error variance σ^2 at each design point.

For purposes of this analysis, σ^2 is regarded as constant over all design points.

The estimated emission response $\hat{y}$ as computed by the modal analysis model is a weighted combination of the estimates obtained from the steady-state estimate $\hat{y}_S$ and the accel/decel estimate $\hat{y}_A$:

$$\hat{y} = \omega \hat{y}_S + (1-\omega)\, \hat{y}_A \quad (32)$$

Therefore, on the assumption that the errors involved in the two components of the estimate are statistically independent,

$$\mathrm{Var}(\hat{y}) = \omega^2\, \mathrm{Var}(\hat{y}_S) + (1-\omega)^2\, \mathrm{Var}(\hat{y}_A) \quad (33)$$

In the following discussion, the variance function has been computed using this weighted combination of the steady-state and accel/decel portions.

Var $(\hat{y})$ varies at different coordinates in (v, a)-space. At some points the response can be estimated with relatively little error; at other positions the error can be quite large. As shown in Appendix I, the variance in the estimated response at a point P in the (v, a)-plane is given by

$$\mathrm{Var}(\hat{y}) = \underline{x}\,(X'X)^{-1}\,\underline{x}'\,\sigma^2 \qquad (34)$$

where $\underline{x}$ is a vector obtained by evaluating each of the basis functions at the particular point P and $X'X$ is matrix of the least squares normal equations. Therefore, for every point P, (34) is actually a variance function. By dividing both sides by σ^2, one can obtain the function in normalized form:

$$\mathrm{Var}(\hat{y})/\sigma^2 = \underline{x}\,(X'X)^{-1}\,\underline{x}' \qquad (35)$$

This emission-rate variance function can be viewed as a response surface generated by evaluating the function at given increments over any region of interest. The propagation of error over the (v, a)-space can be considered relative to the basis functions used and the design points chosen by examination of (35), the variance function in normalized form. The actual magnitude of the variance at any point can be examined by evaluating (34), which includes a scalar multiplication by the error variance σ^2.

The reduction in the number of modes or alteration thereof without loss of information was to be investigated. To this end, the change in the variability of the emission rate function as a result of changing the modal design points was examined. Variance surfaces were generated using the normalized variance function so as to isolate the error introduced by changing the design points without introducing the actual error variance σ^2.

As a base for purposes of comparison, the variance surface using the average velocities and accelerations of the 32 accel/decel and 5 steady state modes of the Surveillance Driving Sequence was generated. Figure 7 shows the

locations of these initial design points as "dots" which have been labeled with their modal numbers. The resulting variances are contoured at various thresholds in Figures 8 and 9.

Although average velocities ranging from 0 mph to 60 mph and average accelerations ranging from approximately -3 mph/sec to +2.5 mph/sec are included, examination of Figure 7 reveals that the actual ($\bar{v}$, $\bar{a}$) points are quite randomly located and do not appear to adequately represent the entire region. In particular, the region of $-1.2 \leq a \leq 1.0$ is not well covered except for the steady-state modes. (It was due to this lack of information and the associated uncertainty involved in predicting emissions that the accel/decel and steady-state functions were weighted in the model.) Also not well represented are the regions in which velocity approaches 0 mph or 60 mph and the absolute value of the acceleration rate is large. By dividing each of the accel/decel modes into subsets, it was possible to "fill in" regions which were poorly represented.

The following strategies for decomposing the modes were investigated.

1. 0 – t/2, t/2 – t (t is the mode duration)
2. 0 – t/3, t/3 – t
3. 0 – 2t/3, 2t/3 – t
4. 0 – t/3, t/3 – t for decel modes
   0 – 2t/3, 2t/3 – t for accel modes
5. v_1 – $(v_1+v_2)/2$, $(v_1+v_2)/2$ – v_2 (v_1 = mode initial velocity)
6. v_1 – $\bar{v}$, $\bar{v}$ – v_2 (v_2 = mode final velocity) ($\bar{v}$ = mode average velocity)
7. 0 – 4 sec., 4 sec. – t
8. 0 – (t-4) sec., (t-4) sec. – t

In each case, each segment was arbitrarily constrained to cover at least 4 seconds to allow for adequate data collection.

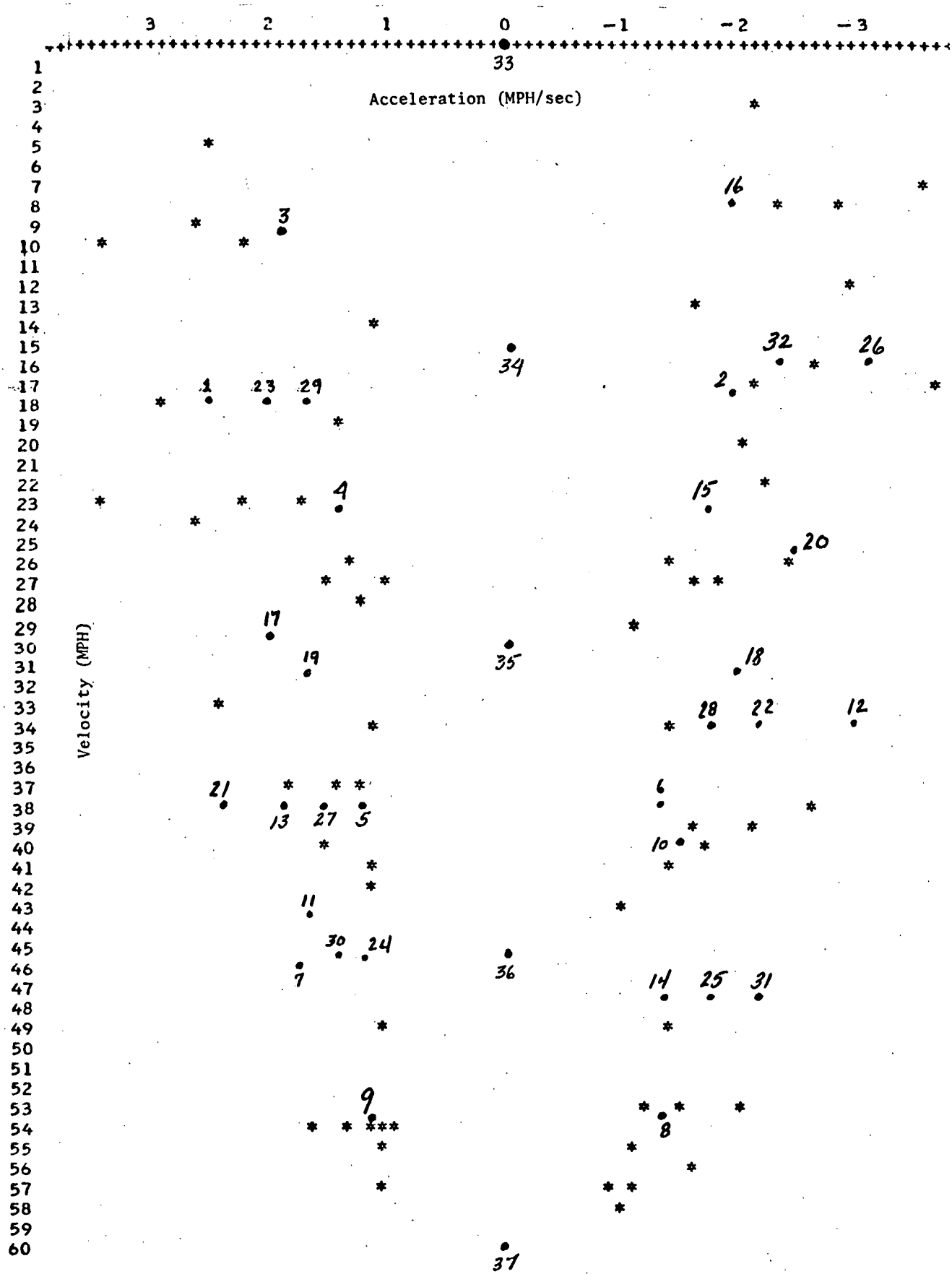


FIGURE 7 SPEED-ACCELERATION TEST DESIGN POINTS

FIGURE 8 Normalized Variance Surface Based on 37 $(\bar{v},\bar{a})$ Design Points

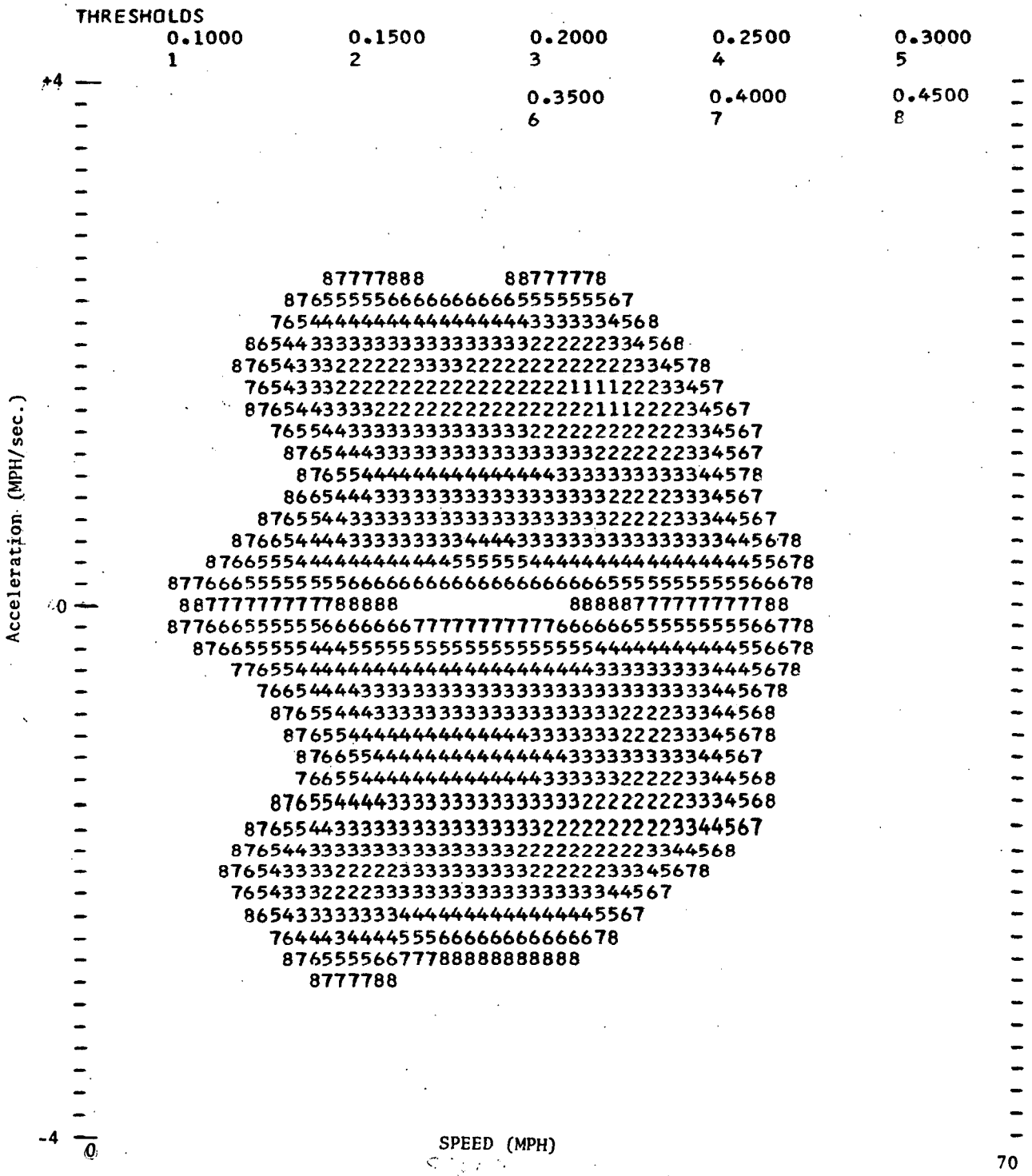
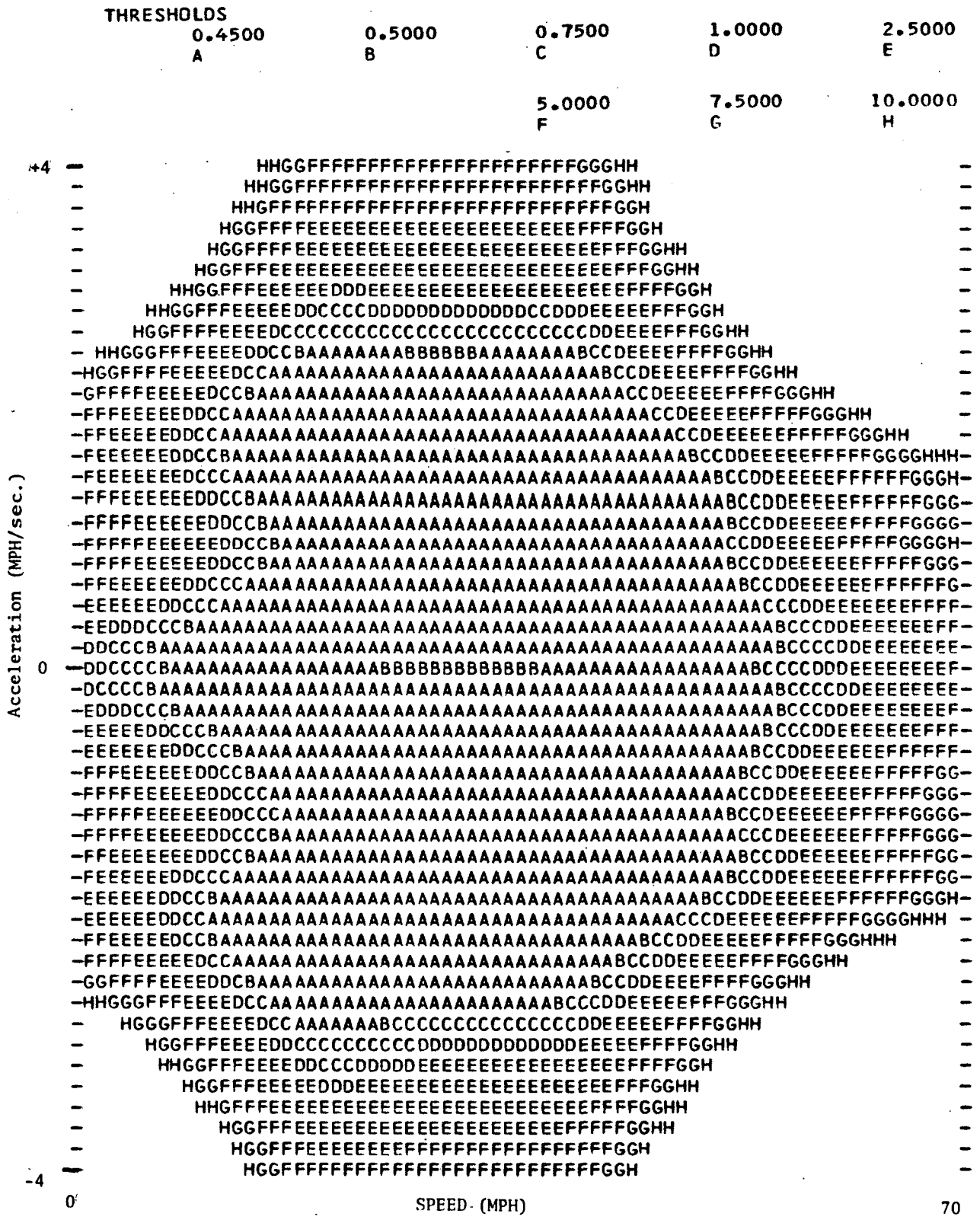


FIGURE 9 Normalized Variance Surface Based on 37 ($\bar{v}$,$\bar{a}$) Design Points



The "stars" in Figure 7 are the design points resulting from using strategy #1, which simply divides each accel/decel mode into two subsets based on t/2; this procedure results in using 69 design points, 64 obtained from accel/decel modes and 5 from the steady-state modes. Figures 10 and 11 are threshold maps of the variance surface resulting from using these design points. It is obvious that the entire level of the variance was lowered as a result.

In order to investigate the changes in the variance as a result of the reduction of the number of modes, a normalized variance surface was generated after certain modes had first been excluded. It was decided to drop 1/4 of the modes simply by excluding points in regions where there seemed to be ($\bar{v}$, $\bar{a}$) redundancy. The modes excluded were 13, 22, 23, 25, 27, 28, 30, and 31. The variance map of the depleted design worsened as expected. However, the t/2 expansion of the 24 modes used (53 design points including 48 accel/decel and 5 steady state) actually showed improvement over the full modal t/2 expansion in some regions. Figure 12 shows the resultant variance surface for thresholds less than 0.45. This surface is a definite improvement over using the initial modal ($\bar{v}$, $\bar{a}$) design points.

In a second strategy, one-half of the modes were dropped. Excluded were modes 4, 5, 6, 9, 13, 14, 15, 17, 19, 22, 23, 24, 27, 28, 30, and 31. The choice of design points in this instance was guided by the results of principal component analysis, to be discussed later in this report. Figure 13 shows the resultant variance surfaces using the t/2 expansion of the remaining 16 accel/decel modes and 5 steady-state modes for thresholds less than 0.45. This surface was generated using 37 points as was the surface based on the original modal points. Comparison of Figures 8 and 13 clearly shows that an improvement in the normalized variability can be realized by appropriate choice of design points.

3.3 FACTOR ANALYSIS OF MODAL DATA

The test data that comprise the input to the original modal emissions model are measurements of individual vehicle emissions given off in time

FIGURE 10 Normalized Variance Surface Based on 69 $(\bar{v},\bar{a})$ t/2 Design Points

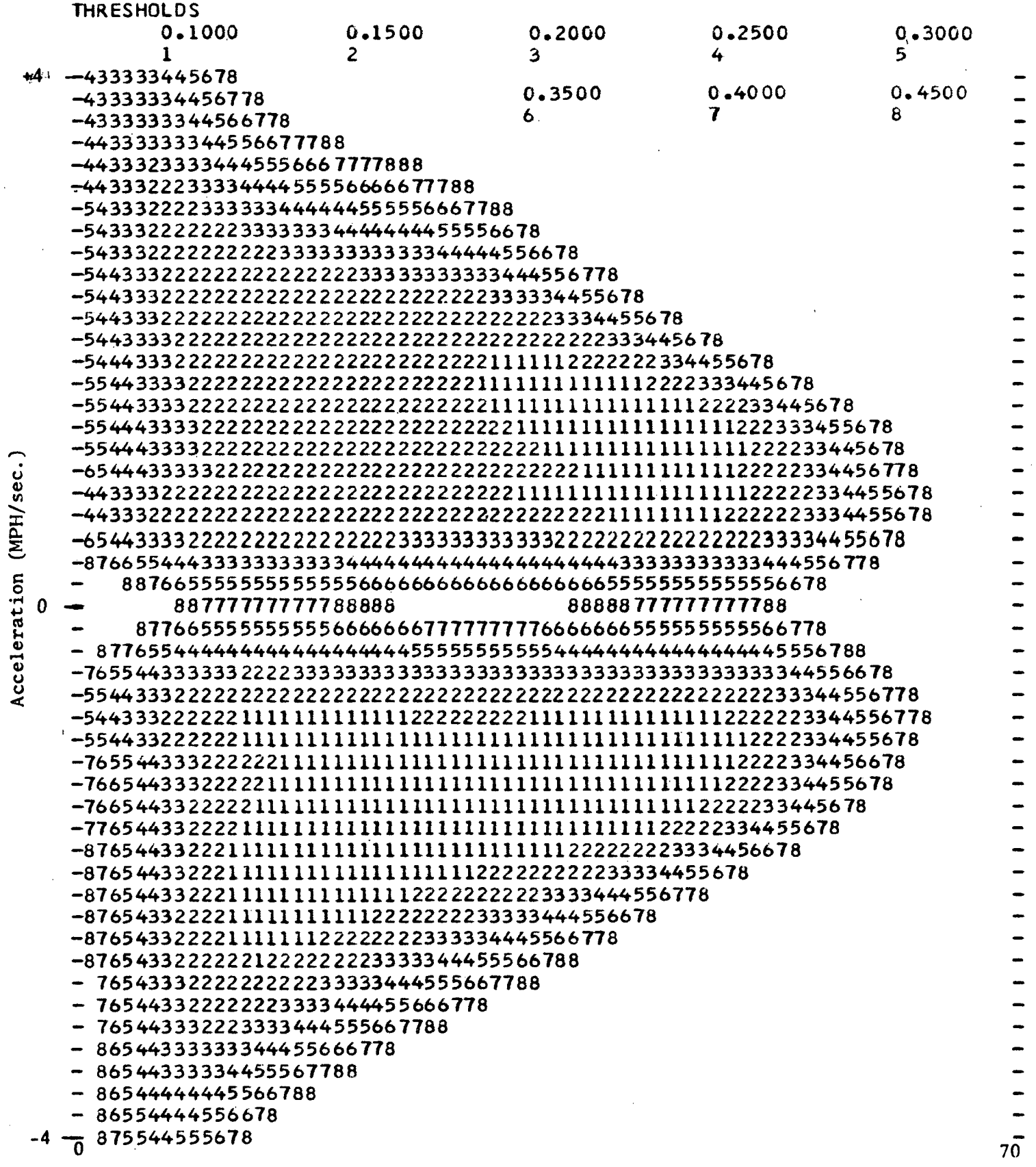


FIGURE 11 Normalized Variance Surface Based on 67 $(\tilde{v},\tilde{a})$ t/2 Design Points

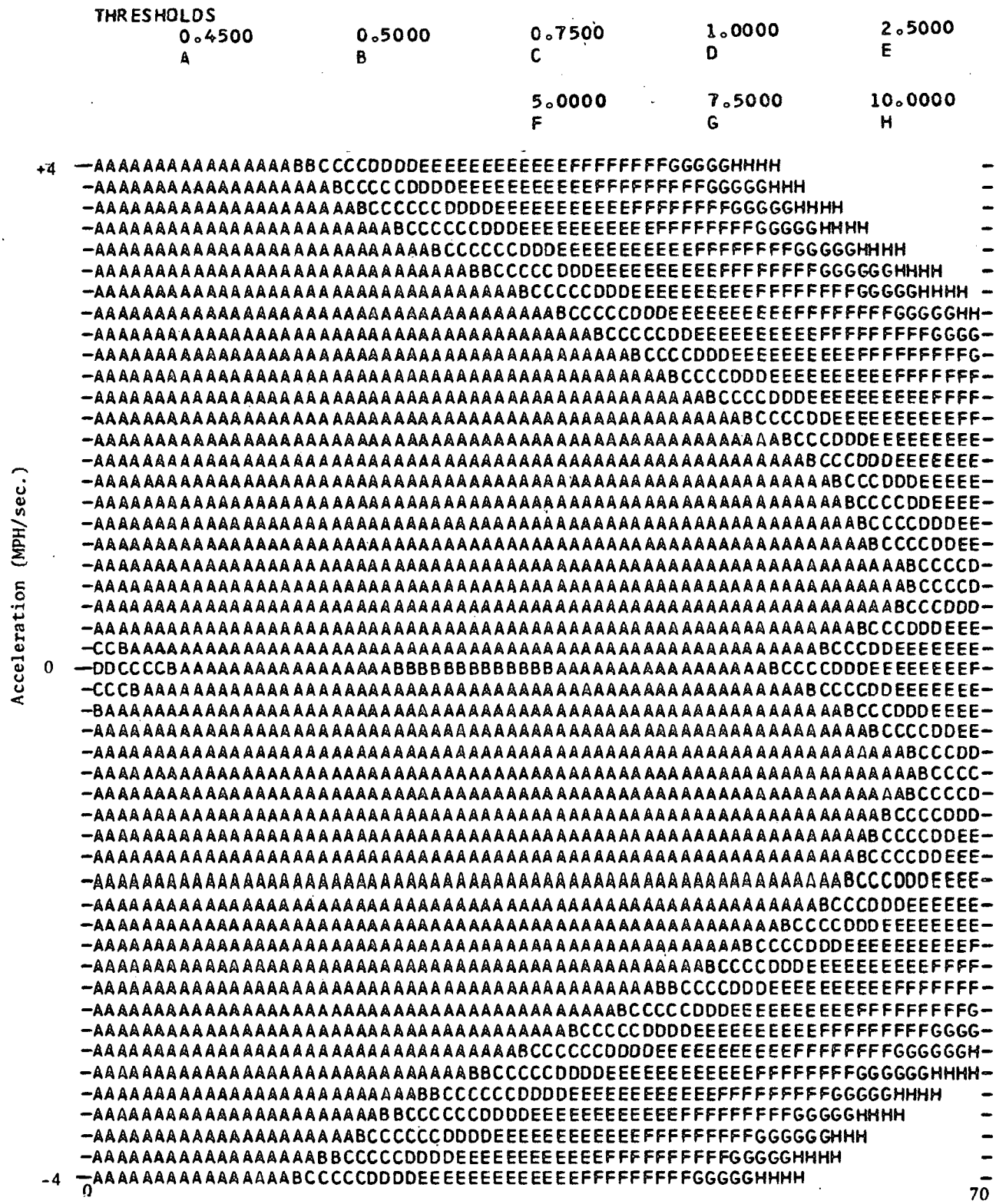


FIGURE 12 Normalized Variance Surface Based on 53 $(\bar{v},\bar{a})$ t/2 Design Points

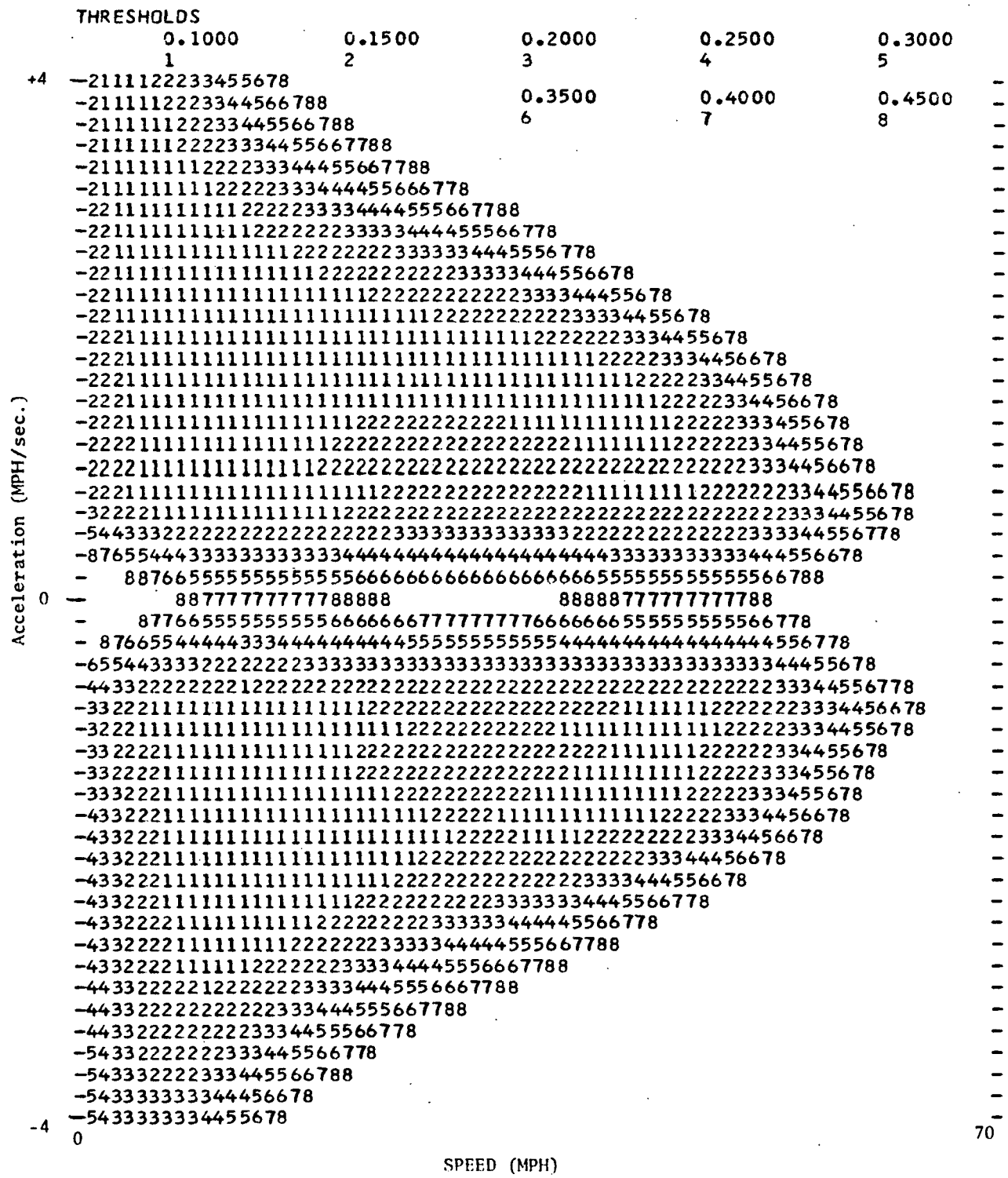
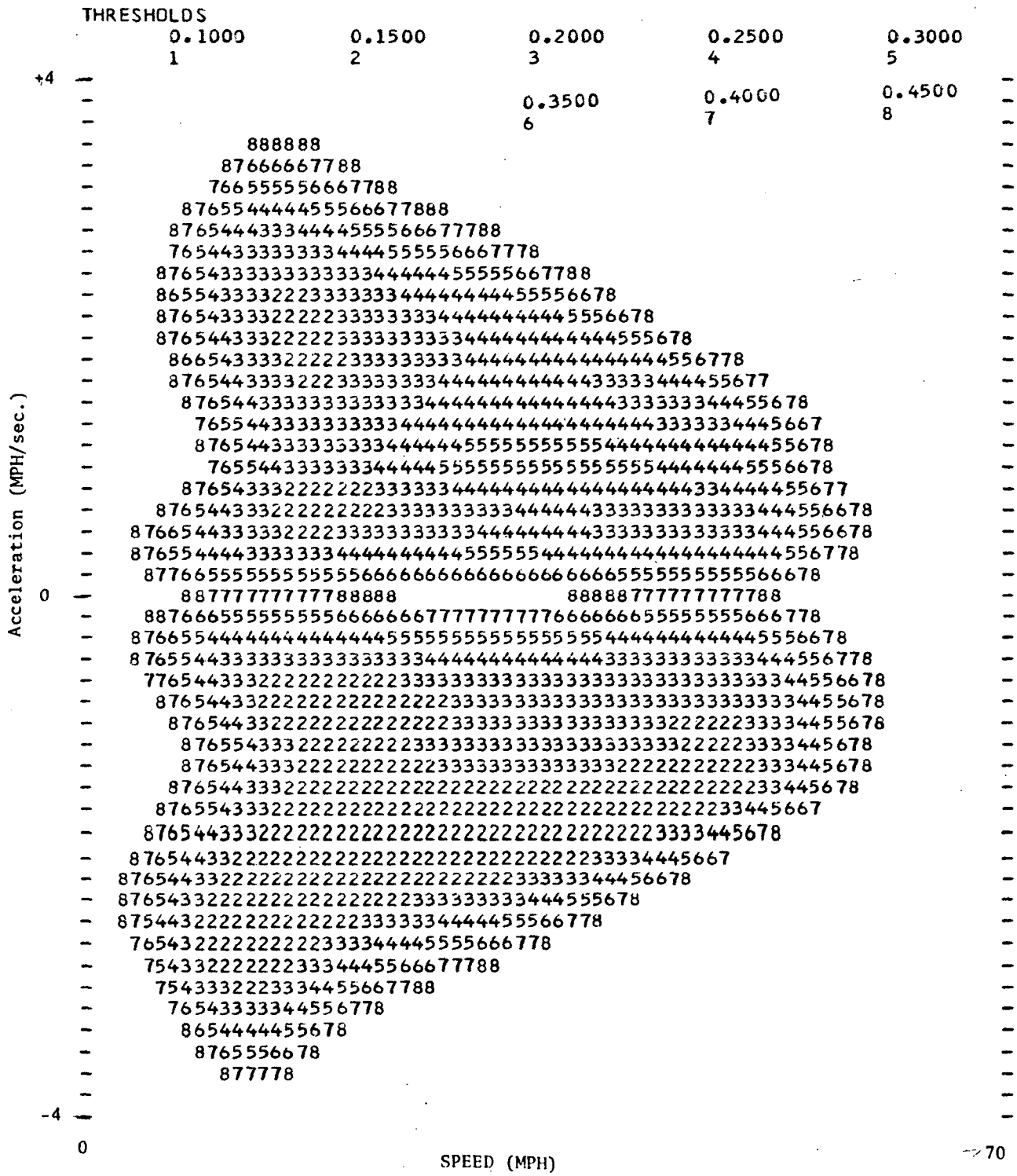


FIGURE 13 Normalized Variance Surface Based on 37 $(\bar{v},\bar{a})$ t/2 Design Points



periods called modes during which the vehicle follows a given speed-time profile. In order to determine whether or not there is any degeneracy in the information being supplied by the various modes, this test data was examined using methods of factor analysis.

Factor analysis is useful in analyzing the intercorrelations within a set of variables in order to identify fundamental and meaningful dimensions in the multivariate domain. This "task of factor analysis is most frequently accomplished by first conducting a principal-components analysis and by then using the resulting principal factors as a set of reference axes for determining the simplest structure, or most easily interpretable set of factors for the domain in question."*

Principal-components analysis is generally useful in determining the minimum number of independent dimensions needed to account for most of the variance in the original set of variables. In the present instance, this statement can be interpreted to mean that the variance among the 1020 vehicles in the data base, so far as emissions is concerned, can be explained by the car-to-car variability observed in the values of a certain number of linear combinations of the modal contributions. The number of these combinations required to account for some specified fraction of the total variance -- say, 90% -- is often referred to as the dimensionality of the space. The essential thrust of the analysis is to take cognizance of the fact that if two variables, such as two modal contributions to emissions, tend to vary in some related way as one goes from vehicle to vehicle, then there is essentially only one variable at work rather than the apparent two.

To achieve such insight, it is heuristically logical to examine the correlations among all pairs of modes for the 1020 vehicles in the data base. The result is a correlation matrix for each of the pollutants under consideration. The correlation matrices based on these 37 modes were determined for each of the three pollutants, HC, CO, and NOX (as well as for CO_2, in connection with fuel-use studies to be discussed later in this report.) These correlation matrices were then subjected to a principal-components analysis in order to

*Cooley, W. and Lohnes, P., Multivariate Data Analysis, Wiley, New York, 1971, p. 131.

determine the eigenvalues (λ) and associated normalized eigenvectors ($\underline{v}$). The factor coefficients or loadings were then derived by:

$$\underline{a}_j = \sqrt{\lambda_j}\, \underline{v}_j, \quad j = 1, 2, \ldots, 37$$

where $\underline{a}_j$ and $\underline{v}_j$ are of dimension 37. The numbers of dimensions or modes needed to account for 90% and 95% of the variance for each pollutant are indicated below:

	90% of Variance	95% of Variance
HC	7	13
CO	9	15
CO_2	9	18
NOX	14	21

For purposes of illustration, Table 1 gives the factor loadings for the first seven principal components derived from the correlation matrix for HC, together with their associated eigenvalues and the percent variance accounted for by these factors.

Besides using the principal-components solution to identify the dimensions of the domain, an attempt could be made to interpret the results. In general, the principal-components solution produces one general factor and p-1 bipolar factors (p is the number of common factors). The general factor is usually all positive (or negative) when the solution is based on a matrix of positive correlations. It could be argued that the first factor in Table 1 is perhaps a "speed" factor. The second factor is a bipolar factor and (except for the five steady-state modes) the modes of acceleration have negative loadings and those of deceleration have positive loadings; this factor could be considered to be an "accel/decel" factor.

In order to improve on the solution offered by the principal-components technique, factors were rotated to positions in which the factor pattern comes closer to criteria of simple structure. The purpose of analytic rotation schemes is to transform the principal components so as to obtain new variables

Table 1 Principal Components (7) of the Correlation Matrix for HC for 37 Modes

Eigenvalue	27.04	2.437	1.438	0.9233	0.6617	0.5611	0.5135
% of Variance	73.08	6.59	3.89	2.50	1.79	1.52	1.39

FACTOR LOADINGS

Variable							
1	-0.8762E+00	-0.8822E-01	0.2557E+00	0.4106E-01	-0.1240E-01	-0.1291E-01	0.1882E+00
2	-0.5340E+00	0.2314E+00	0.1431E+00	-0.2286E+00	0.7646E+00	0.4192E-01	-0.1979E-01
3	-0.7764E+00	0.1518E+00	0.3774E+00	0.5887E-01	-0.4711E-01	-0.3542E-02	0.1961E+00
4	-0.9313E+00	-0.1033E+00	0.1968E+00	-0.7513E-01	-0.1638E-01	-0.1154E-01	0.3493E-01
5	-0.9188E+00	-0.2801E+00	0.2738E-01	-0.1088E+00	-0.3183E-02	-0.9056E-02	-0.2226E-01
6	-0.8445E+00	0.1106E+00	-0.1831E+00	-0.1264E+00	-0.4754E-01	-0.1219E+00	-0.1852E+00
7	-0.8887E+00	-0.3189E+00	-0.5325E-01	-0.9734E-01	-0.5350E-02	-0.9311E-01	0.5413E-01
8	-0.8606E+00	-0.1366E+00	-0.2348E+00	-0.1263E+00	0.5006E-02	-0.1069E+00	-0.8305E-01
9	-0.8793E+00	-0.3205E+00	-0.7850E-01	-0.1064E+00	0.2087E-03	-0.3372E-01	-0.2494E-02
10	-0.8226E+00	0.2774E+00	-0.1920E+00	-0.3623E-01	-0.3155E-02	-0.1609E+00	0.7553E-01
11	-0.9240E+00	-0.2911E+00	0.1317E-01	-0.5269E-01	-0.2383E-02	-0.7655E-01	0.8585E-01
12	-0.8980E+00	0.2137E+00	-0.2272E+00	-0.5106E-01	0.7054E-02	-0.9558E-01	0.1049E+00
13	-0.9282E+00	-0.2896E+00	0.5636E-01	-0.3918E-01	-0.2947E-03	-0.7089E-01	0.6371E-01
14	-0.8564E+00	0.1801E+00	-0.2268E+00	-0.8066E-01	-0.1084E-01	-0.1455E+00	0.5473E-01
15	-0.8939E+00	0.2671E+00	-0.7528E-02	-0.1261E-01	-0.3085E-01	-0.1011E+00	-0.1352E+00
16	-0.8081E+00	0.4034E+00	0.2000E+00	-0.8747E-02	-0.7939E-01	-0.6159E-01	0.1169E-01
17	-0.9365E+00	-0.2345E+00	0.1305E+00	0.4235E-02	-0.2867E-02	-0.6688E-01	0.8413E-01
18	-0.8762E+00	0.2918E+00	-0.1485E+00	-0.3612E-01	-0.3373E-01	-0.1757E+00	-0.5196E-01
19	-0.9287E+00	-0.2204E+00	0.9097E-01	-0.3057E-01	-0.1351E-01	-0.6139E-01	0.6378E-01
20	-0.8892E+00	0.3200E+00	-0.1372E+00	-0.6485E-01	-0.2363E-01	-0.1295E+00	-0.2696E-01
21	-0.8965E+00	-0.2793E+00	0.5654E-01	0.1878E-01	-0.4598E-02	-0.9258E-01	0.1144E+00
22	-0.8086E+00	0.2093E+00	-0.3352E+00	0.1766E+00	0.3427E-01	0.2216E+00	0.2598E+00
23	-0.8590E+00	-0.1016E+00	0.2227E+00	0.3583E+00	0.4799E-01	0.8082E-01	-0.8893E-01
24	-0.9214E+00	-0.3141E+00	0.1551E-01	0.9377E-01	0.2559E-01	0.2871E-02	-0.4575E-01
25	-0.7696E+00	0.1101E+00	-0.4387E+00	0.1675E+00	0.3728E-01	0.2695E+00	0.2218E+00
26	-0.8185E+00	0.2719E+00	0.7664E-01	0.2371E+00	0.2063E-01	0.4481E-01	-0.2560E+00
27	-0.9048E+00	-0.3068E+00	0.5815E-01	0.2100E+00	0.4526E-01	0.2076E-01	-0.5688E-01
28	-0.8954E+00	0.2512E+00	-0.1197E+00	0.1902E+00	0.8484E-02	0.7903E-01	-0.2644E-01
29	-0.8715E+00	-0.1178E+00	0.2264E+00	0.3399E+00	0.5622E-01	0.4783E-01	-0.1018E+00
30	-0.8429E+00	-0.3661E+00	-0.2214E-01	0.9119E-01	0.3661E-01	-0.5161E-01	-0.5309E-01
31	-0.8658E+00	0.8397E-01	-0.3185E+00	0.1126E+00	-0.3037E-02	0.3461E-01	-0.8585E-01
32	-0.8325E+00	0.3644E+00	0.8707E-01	0.1910E+00	-0.4140E-01	-0.8869E-02	-0.1584E+00
33	-0.6557E+00	0.5049E+00	0.3972E+00	-0.1338E+00	-0.1191E+00	0.7551E-01	0.1031E+00
34	-0.7847E+00	0.3135E+00	0.3077E+00	-0.2199E+00	-0.1236E+00	0.1247E+00	0.6197E-01
35	-0.8674E+00	0.1611E-01	0.8615E-01	-0.2929E+00	-0.7677E-01	0.2088E+00	-0.1124E+00
36	-0.7810E+00	-0.1889E+00	-0.4654E-01	-0.2778E+00	-0.1046E+00	0.2824E+00	-0.2151E+00
37	-0.8518E+00	-0.1688E+00	-0.1384E+00	-0.1968E+00	-0.4556E-01	0.2990E+00	-0.3148E-01

which might be more readily interpreted and named.* This rotation was performed on the matrix consisting of the 15 principal components of 37 variables or modes for each pollutant (15 factors accounted for at least 90% of the variance in the case of all pollutants). The "normal" varimax criterion was used for the orthogonal rotation of factors.

This new set of rotated axes might be preferred for purposes of interpreting the basic dimensions of the domain measured by the 37 modes. This is because the new coefficients are more "simple" in the sense that a given variable tends to have a high coefficient for only one new axis and each factor has zero, or near zero, coefficients for at least some of the variables.* Table 2 gives the derived rotated factors for the first 10 (HC) factors. The general factor has been destroyed and group factors have been produced. In the first factor, high negative weights are given to variables 5, 7, 9, 11, 13, 17, 19, 21, 24, 27, and 30. These modes, which are all highly correlated, are characterized by accelerations between 1 mph/sec and 2.5 mph/sec and by velocities ranging from about 28 mph to 53 mph. When variances based on the rate of emission (grams/sec) were calculated, these modes all showed relatively high variances. These observations suggest that using any one of these modes could provide as much information as using all of them.

Table 3 gives the factor number for any mode which is weighted heavily in that factor. If more than one mode has high loadings within a factor, the factor number is listed for each mode with the mode which is weighted most heavily being "starred." Examination of this table reveals that the eleven variables which had high loadings in the first factor for HC also have high loadings for the other two pollutants and for CO_2. Again, this fact would suggest that these modes provide redundant information. It should also be noted that for CO, CO_2, and NOX, modes 6, 8, 10, 12, 14, 22, 25, and 28 all have high factor loadings in the second factor. These modes are characterized by average accelerations ranging from -1 mph/sec to -3 mph/sec and by average velocities from 24 mph to 47 mph. They also all have relatively low emission-rate variances. Most factors have high coefficients for

*Cooley and Lohnes, op. cit.

TABLE 2 TEN* ROTATED FACTORS OF THE CORRELATION MATRIX FOR HC FOR 37 MODES

	FACTOR									
	1	2	3	4	5	6	7	8	9	10
1	-.68	.37	-.24	.20	.12	.08	.18	-.39	.00	-.10
2	-.22	.24	-.22	.10	.91	.06	.10	-.06	-.01	-.02
3	-.43	.49	-.24	.24	.12	.06	.11	-.61	-.01	-.02
4	-.72	.39	-.30	.19	.14	.14	.14	-.21	.02	-.10
5	-.82	.20	-.30	.13	.12	.20	.16	-.13	-.02	-.13
6	-.47	.24	-.60	.14	.11	.18	.18	-.07	-.02	-.49
7	-.86	.17	-.32	.05	.09	.10	.16	-.03	-.12	-.01
8	-.65	.12	-.49	.06	.11	.17	.21	-.10	-.06	-.12
9	-.83	.12	-.32	.08	.09	.16	.17	-.06	-.15	-.03
10	-.37	.30	-.67	.13	.13	.09	.25	-.10	-.01	-.01
11	-.87	.23	-.31	.10	.10	.08	.18	-.05	-.06	-.00
12	-.47	.31	-.69	.10	.14	.09	.33	-.08	-.02	.06
13	-.86	.23	-.29	.13	.10	.11	.16	-.12	-.07	-.03
14	-.45	.25	-.74	.11	.12	.10	.10	.23	-.10	-.07
15	-.42	.39	-.58	.27	.15	.13	.18	-.14	-.15	-.15
16	-.31	.62	-.46	.24	.13	.05	.16	-.15	-.00	.04
17	-.83	.28	-.29	.18	.11	.09	.16	-.19	.02	-.06
18	-.40	.34	-.73	.19	.13	.11	.20	-.07	.03	-.15
19	-.81	.27	-.33	.16	.11	.12	.16	-.13	.03	-.04
20	-.39	.39	-.71	.16	.15	.13	.23	-.05	.06	-.05
21	-.85	.24	-.29	.13	.08	.06	.18	-.07	.02	.00
22	-.36	.23	-.47	.18	.10	.08	.72	-.09	.00	.00
23	-.63	.27	-.21	.59	.10	.08	.18	-.19	.12	-.03
24	-.85	.18	-.27	.29	.09	.11	.20	-.01	.03	-.06
25	-.38	.12	-.41	.14	.08	.09	.77	-.04	-.02	-.08
26	-.35	.39	-.43	.58	.13	.08	.21	-.00	-.28	-.06
27	-.82	.16	-.23	.40	.08	.07	.21	-.06	.04	-.03
28	-.40	.34	-.56	.39	.12	.12	.38	-.10	.02	.05
29	-.65	.26	-.22	.59	.11	.07	.16	-.20	.06	-.04
30	-.83	.13	-.23	.24	.07	.07	.17	.05	.15	-.00
31	-.48	.14	-.58	.24	.08	.14	.39	-.05	.00	-.07
32	-.31	.47	-.52	.46	.11	.07	.19	-.12	-.04	-.05
33	-.16	.88	-.28	.15	.14	.05	.11	-.12	-.03	-.07
34	-.37	.80	-.30	.11	.13	.14	.14	-.04	.02	-.01
35	-.57	.49	-.31	.09	.15	.36	.16	-.01	-.07	-.07
36	-.58	.19	-.27	.11	.08	.70	.14	-.05	.03	-.04
37	-.64	.23	-.26	.07	.10	.41	.36	-.07	-.23	-.10

*Rotation done on 15 factors.

Table 3 Highly Loaded Modes by Factor Number

	HC	CO	CO_2	NO_x
MODE		FACTOR NUMBER		
1		12	14	15
2	5	11	15	11
3	8	6	10	3
4			1	
5	1	12*, 1	1	
6	10	2	2	
7	1	9*, 1	1	12*, 1
8	11	14*, 2	12*, 2	2
9	1	1	1	1
10	12*, 3	2*	2*	13*, 2
11	1*	1*	1	1
12	3	13*, 2	2	2
13	1	1	1	1
14	3*	2	2	2
15		5		
16	13			5
17	1	1	7*, 1	1
18	3	10	11	
19	1	1	1	1
20	3	2, 15	4	
21	1	1	1	8*, 1
22		2	2	2
23	4			
24	1	1	1	1
25	7	15*, 2	13*, 2	2*
26	9*, 4	7	4	
27	1	1	1*	1
28		2	2	7, 2
29	4			
30	1	1	1	1*
31	15	2	9*, 2	2
32		8	8	7*
33	2*	3*	5	4
34	2	3	6	9
35	14			6
36	6	4*	3	
37	9	4	3*	14

only one mode. Some variables such as 4, 15, and 29 have not been weighted heavily in any factor.

The results of the principal-components analysis indicate that for all pollutants 14 modes are sufficient to account for 90% of the total variance. These 14 modes are not the same for all of the pollutants. However, eleven of the modes seem to provide the same information for all pollutants and eight modes provide the same information for three of the pollutants.

In conclusion, it appears that test procedures could be modified so as to avoid running a vehicle through all 37 of the defined modes and still obtain the same amount of information about its emission response.

4. GROUP EMISSION PREDICTIONS

Individual vehicles represent a wide variation in model year, make, model, engine and drive train equipment, accumulated mileage, state of maintenance, attached pollution abatement devices, and geographic location. Inasmuch as it is a mix of these diverse vehicles which determines the vehicular contribution to air pollution in a given vicinity, however, it is appropriate to aggregate vehicles into groups and to view the group as a composite emission source for various purposes of analysis. Accordingly, considerable interest centers on the accuracy and precision with which the modal analysis emission model can predict group emissions in a given driving sequence.

The characterization of a group of vehicles can be achieved by defining the emission rate function for the average vehicle within the group. Let,

b_{ijk} = k'th coefficient in the emission rate function for the j'th vehicle within the group and i'th kind of pollutant.

N_g = number vehicles in the group.

$\bar{b}_{ik}$ = k'th coefficient in the emission rate function describing the average vehicle's i'th kind of pollutant response.

Then,

$$\bar{b}_{ik} = \frac{1}{N_g} \sum_{j=1}^{N_g} b_{ijk}$$

Thus, the group emission rate functions are determined by averaging the coefficients which make up the emission rate functions of each vehicle in the group. In this way, the group is viewed as consisting of N_g "average" vehicles, each having identical emission characteristics. The emission response of the group over any driving sequence can accordingly be determined by multiplying the response of this average vehicle by the number of vehicles

in the group. Note that, once the emission response of the average vehicle has been characterized in terms of average regression coefficients, its total emission over any specified driving sequence can be obtained by appropriate integration of the emission rate function in exactly the same manner as for any other vehicle.

As was shown in section 3.2 of this report, error propagation in the modal analysis emission model causes the emissions estimated for some regions of the (v , a)-plane to have lower variance than for certain other regions of the plane. A consequence of this fact is that the estimation capability of the model over an arbitrary driving sequence will depend on the relative amounts of time which that sequence devotes to regions of high or low variance. This fact is true for both individual vehicles and for groups of vehicles.

Our approach to an evaluation of the model for group emission prediction was as follows. First, a study was made of the extent of agreement between observed and computed emissions for the Surveillance Driving Sequence (SDS). Then, with this comparison in view as a "base case," a procedure was developed for relating the base case to arbitrary driving sequences which, as a result of differences in their distribution of velocities and acceleration, exhibit different degrees of variance in the emissions computed by the model.

4.1 MODEL PERFORMANCE FOR THE SDS BASE CASE

Two general questions are of interest in connection with the prediction of group emissions: accuracy and precision. Lack of accuracy is reflected as a bias or systematic error in the predicted results. Lack of precision is the consequence of random errors in the prediction and is manifested in terms of variance in the predicted group emissions under repeated sampling and testing of the group. These two aspects of the group prediction question will be addressed below in connection with the performance of the model for the Surveillance Driving Sequence.

4.1.1 Accuracy of Group Emission Prediction

Because of the fact that the "true" or population value for the mass of a pollutant emitted during a particular driving cycle can never be known, the question of accuracy can be resolved only in a relative sense. One possible approach to evaluating the accuracy of model prediction is to compare, for a particular driving sequence, bag values as computed by the model and bag values as actually observed in test.

The approach indicated above was employed in the initial implementation of the original model by comparing computed and observed bag values for the Surveillance Driving Sequence. These results were originally presented in Calspan Report No. NA-5194-D-3 and in EPA Report No. EPA-460/3-74-005. Relevant portions of these results are repeated herein as Table 4 for purposes of reference, because it is here proposed to view these results in a new light.

Table 4

BAG VALUE ERROR STATISTICS
SURVEILLANCE DRIVING SEQUENCE
1020 VEHICLES

POLLUTANT	OBSERVED BAG VALUE (gms) $\bar{O}$	MEAN ERROR (gms) $\bar{R}$	VARIANCE $(gms)^2$ σ_R^2	STANDARD DEVIATION (gms) σ_R	STANDARD ERROR $\sigma_{\bar{R}}$	$u = \frac{\bar{R}}{\sigma_{\bar{R}}}$
HC	53.5	7.2	143.3	12.0	0.38	19.9
CO	625.0	43.1	20420.8	143.0	4.51	9.6
NOX	48.2	-2.7	163.0	12.8	0.40	6.7

Consider the computed and observed bag values as paired samples in an experiment aimed at determining if a significant difference can be shown between two experimental treatments. A rather classical example of such an experiment is one in which two types of pipe are buried in various types of soil for various lengths of time with a view toward observing differences in degree of corrosion. By always burying paired samples, one of each type, the

experimenter can observe the difference in degree of corrosion for each pair and, as far as overall generalization is concerned, can circumvent the variability introduced by inhomogeneity of exposure conditions. If he wants to group the paired samples into classes according to soil type -- clay, loam, cinders -- he can restrict his inferences to these strata, again with the advantage of balanced comparisons within the strata. It is proposed to examine the performance of the modal analysis emission model in this vein. In this analysis, individual vehicles will play the role of exposure conditions, and homogeneous classes of vehicles will play the role of soil strata.

First, let us examine the hypothesis that there is no significant difference between the mean bag value as observed and as computed -- that is, let us examine the hypothesis:

$$H_o : \bar{R} = 0$$

Because of the large sample size, we can use the u-test to test the hypothesis. The standard error of $\bar{R}$ is

$$\sigma_{\bar{R}} = \frac{\sigma_R}{\sqrt{N}}$$

and u is defined as

$$u = \frac{\bar{R}}{\sigma/\sqrt{N}}$$

As shown in Table 4, the hypothesis is rejected at the 0.01 level for all three pollutants. In this connection, however, a word of warning is in order. By pooling a sufficient quantity of data, it is possible to label as statistically significant an effect which may be of negligible engineering magnitude. More germane is the consideration that if the difference between two means is no greater than -- say, 10% -- of their pooled mean, it may be of small consequence that this difference is declared to be statistically significant. The importance of the effect depends on its probable magnitude, and the mere act of declaring it to be statistically significant in no way augments its practical magnitude.

4.1.2 Precision of Group Emission Prediction

For each of the pollutants HC, CO and NOX, the relative importance of statistical and practical views of model performance can be considered in terms of confidence intervals. Let μ_R denote the expected or population mean value of the difference between calculated and observed bag values for a pollutant. The width of a confidence interval for μ_R depends on the dispersion of estimates for individual vehicles comprising the group and on the "size" of the confidence interval. In statistical terminology, the term "size" denotes the probability with which it can be asserted that the population mean falls between two prescribed values. In the following discussion, we shall assume a confidence interval of size 0.95 (95% confidence).

For 95% confidence, the half-width of the confidence interval is approximately 2 $\sigma_{\bar{R}}$ (more exactly 1.958 $\sigma_{\bar{R}}$) and the confidence intervals for the three pollutants are approximately as follows:

HC	6.4	$\leq \mu_R$	$\leq$	8.0
CO	34	$\leq \mu_R$	$\leq$	52
NOX	-3.5	$\leq \mu_R$	$\leq$	-1.9

As a percent of $\bar{O}$, the mean observed value for the pollutant in question, one obtains as extremes:

$$\frac{8.0}{53.5} \times 100\% = 15\% \text{ for HC}$$

$$\frac{52}{525} \times 100\% = 10\% \text{ for CO}$$

and

$$\frac{3.5}{48.2} \times 100\% = 7.3\% \text{ for NOX}$$

In short, for a group of 1020 highly heterogeneous vehicles, the bias for HC would not be expected to be greater than 15% of the mean values as actually observed by direct measurement of these 1020 vehicles. Similar figures of 10% and 7.3% apply for CO and NOX.

4.1.3 Sampling Considerations

It is evident that the dispersion of emissions for individual vehicles within a group depends on the degree of homogeneity of the group as far as such determinants as make, model, mileage, state of maintenance, and other factors are concerned. Because of this fact, the standard errors applicable to the mean emissions computed for the three pollutants for the group also depend on the homogeneity of the group. Consequently, the performance of the model in the estimation of group emissions depends strongly on sampling protocol.

Consider, for example, a population of vehicles having a certain mix of vehicle "types," as specified by make, model, mileage, and other factors which can be rationally employed to differentiate one vehicle from another. A random sample of N vehicles would produce a certain mix of vehicle types within the sample, not necessarily the mix existing in the population. A second sample would most likely produce a different mix of vehicle types and certainly a set of different vehicles than the first sample. One sees, therefore, the influence of two sources of variability as far as the predictions of the model are concerned: vehicle-to-vehicle variability within types and variability in proportionate weighting of types. The result is that a confidence interval based on a random sample from a nonhomogeneous population of vehicles can be expected to be considerably wider than for a case in which some of the sources of variability are controlled.

In this connection, consider the case in which stratified sampling is used to select N vehicles from the population. This procedure is a quite logical one in emission assessment, because it assures that the sample will contain the same relative proportions of different types of vehicles as does the population. Random sampling is then performed within each strata to obtain the desired number of vehicles. In this type of sampling, vehicle-to-vehicle variation will be present but the variation in proportions of the various strata will have been eliminated.

In conclusion, it is not possible to make overall generalizations about the ability of the model to estimate group emissions, unless the nature of the group and the method by which it is sampled is taken into account.

4.2 MODEL PERFORMANCE FOR ARBITRARY DRIVING SEQUENCES

As noted in Section 4.1 of this report, the performance of the modal analysis emission model can be evaluated for the Surveillance Driving Sequence by direct comparison with observed results. No other driving sequence except the FTP permits such a comparison, because bag values are not available for these sequences. To obtain such a comparison for an arbitrary driving sequence, it would be necessary to perform emission tests over that driving sequence as a "validation" of the model. It is possible, however, to compute, for an arbitrary driving sequence, the mean emissions for a group of vehicles and the variance of the emissions exhibited by individual vehicles comprising the group. Thus it is possible to evaluate the precision of model performance for the group for an arbitrary driving sequence, but its accuracy must be judged according to results of the SDS base case.

4.2.1 Theoretical Background

The essence of the approach to precision analysis for the performance of the model in an arbitrary driving sequence resides in the simplification of the emission integration as detailed in Section 2.2 of this report. As background for this approach, however, it will be informative to review the underlying statistical theory.

Consider a set of random variables $X_1, X_2, \ldots, X_p$ and a linear combination of these variables

$$Y = c_1X_1 + c_2X_2 + \ldots + c_pX_p$$

where $c_1, c_2, \ldots, c_p$ are constants. The variance of the random variable Y can be computed as

$$\text{Var } Y = \sum_{i=1}^{p} c_i^2 \text{ Var } X_i + 2 \sum_{\substack{i=1 \\ i \neq j}}^{p} \sum_{j=1}^{p} c_i c_j \text{ Cov } (X_i, X_j)$$

For example, for three variables,

$$\text{Var } Y = c_1^2 \text{ Var } X_1 + c_2^2 \text{ Var } X_2 + c_3^2 \text{ Var } X_3$$
$$+ 2\, c_1 c_2 \text{ Cov } (X_1, X_2) + 2\, c_1 c_3 \text{ Cov } (X_1, X_3)$$
$$+ 2\, c_2 c_3 \text{ Cov } (X_2, X_3)$$

In matrix notation, this result can be written

$$\text{Var } Y = \begin{bmatrix} c_1 c_2 c_3 \end{bmatrix} \begin{bmatrix} \text{Var } (X_1) & \text{Cov } (X_1, X_2) & \text{Cov } (X_1, X_3) \\ \text{Cov } (X_1, X_2) & \text{Var } (X_2) & \text{Cov } (X_2, X_3) \\ \text{Cov } (X_1, X_3) & \text{Cov } (X_2, X_3) & \text{Var } (X_3) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}$$

or, in general,

$$\text{Var } Y = \underline{c}'\, S\, \underline{c}$$

where S is the variance-covariance matrix of the random variables $X_1, X_2, \ldots, X_p$, $\underline{c}$ is a column vector of the weighting coefficients $c_1, c_2, \ldots, c_p$, and $\underline{c}'$ is a row vector, the transpose of $\underline{c}$.

It will be seen that equation (22), pertaining to the integrated basis functions, fits the definition above.

4.2.2 Variance Computations for Arbitrary Driving Sequences

It has been shown that, for an arbitrary driving sequence, the total emission of a pollutant can be written as

$$e(T) = \sum_{i=1}^{3} a_i \int_o^T h\left[a(t)\right] f_i \left[v(t)\right] dt$$

$$+ \sum_{j=1}^{9} b_j \int_o^T \left\{1-h\left[a(t)\right]\right\} g_j \left[v(t), a(t)\right] dt \qquad (36)$$

for the modal analysis emission model as originally developed with 12 basis functions. Moreover, it was shown that, for a given driving sequence, the 12 integrals need be computed only once for any group of vehicles because the values of these integrals are constant for all vehicles in the group and depend only on the nature of the driving sequence. On the other hand, each vehicle in the group gives rise to a different set of a_i and b_j ; consequently, these values can be considered as outcomes of random variables A_i, $i = 1, 2, 3$ and B_j, $j = 1, 2, \ldots, 9$. Thus the A_i and B_j play the role of the X_i in Section 4.2.1 of this report. Similarly, the values of the 12 integrals in equation (36) play the role of the constants c_i in Section 4.2.1. Denoting these integrals $c_1, c_2, c_3, d_1, d_2, \ldots, d_9$, one can then write

$$\text{Var } e(T) = \left[c_1, c_2, c_3, d_1, \ldots, d_9\right] \begin{bmatrix} \text{Var } A_1 & \text{Cov}(A_1,A_2) & \text{Cov}(A_1,A_3) & \text{Cov}(A_1,B_1) & \ldots & \text{Cov}(A_1,B_9) \\ \text{Cov}(A_1,A_2) & \text{Var } A_2 & \text{Cov}(A_2,A_3) & \text{Cov}(A_2B_1) & \ldots & \text{Cov}(A_2,B_9) \\ \text{Cov}(A_1,A_3) & \text{Cov}(A_2,A_3) & \text{Var } A_3 & \text{Cov}(A_3,B_1) & \ldots & \text{Cov}(A_3,B_9) \\ \text{Cov}(A_1,B_1) & \text{Cov}(A_2,B_1) & \text{Cov}(A_3,B_1) & \text{Var } B_1 & \ldots & \text{Cov}(B_1,B_9) \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ \text{Cov}(A_1,B_9) & \text{Cov}(A_2,B_9) & \text{Cov}(B_9) & \text{Cov}(B_1,B_9) & \ldots & \text{Var } B_9 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ d_1 \\ \vdots \\ d_9 \end{bmatrix}$$

or, more succinctly, as

$$\text{Var } e(T) = \underline{c}' \, S \, \underline{c}$$

where S is the variance-covariance matrix of the coefficients of the model for the group of vehicles under consideration and $\underline{c}$ is a column vector of the integrated basis functions as integrated over the driving sequence under consideration. In application, the variance-covariance matrix would be estimated from the N vehicles comprising the group of vehicles under consideration.

To illustrate this principle, four driving sequences were constructed with the intention of accounting for highway and city driving.

Driving Sequence ID	Description
DS1	Highway driving with frequent changes in speed.
DS2	City driving with frequent changes in speed.
DS3	City driving with long periods of constant speed.
DS4	Constant-speed highway driving.

These driving sequences are depicted in Figure 14. Calculations of the total variance over a driving sequence were based on 1050 seconds. Therefore, the sequence shown for DS1 was repeated once and that shown for DS2 was repeated three times. A fifth driving sequence was taken as the first 505 seconds of the Federal Test Procedure. Results for these driving sequences are presented in Table 5 for HC only. The results are based on all 1020 vehicles considered as a group.

FIGURE 14 ARBITRARY DRIVING SEQUENCES

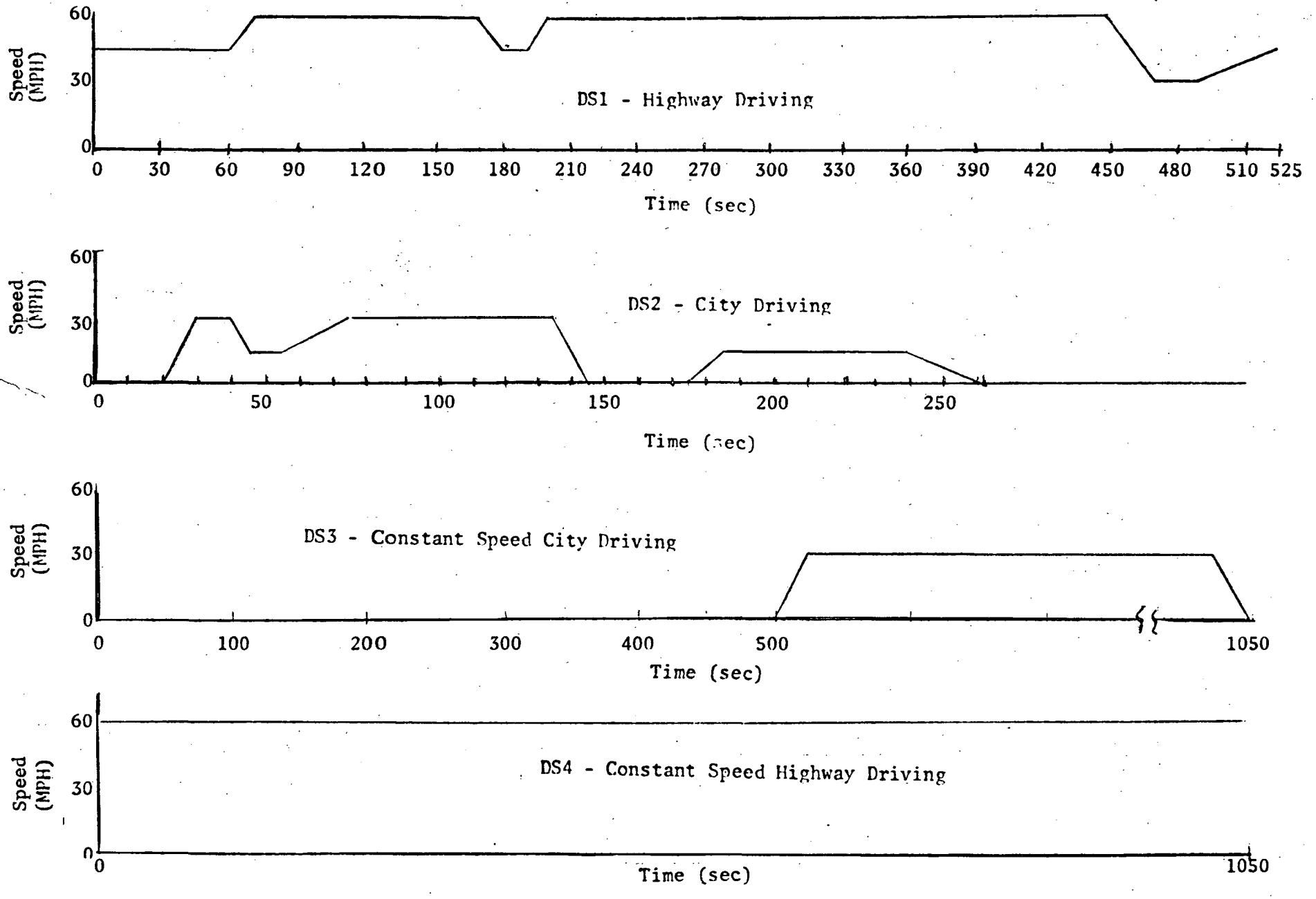


Table 5

TOTAL COMPUTED VARIANCE OF HC
OVER VARIOUS DRIVING SEQUENCES

Driving Sequence	Time Duration (sec)	Variance (gm^2)
SDS	1054	2172.6
FTP	505	336.9
DS1	1050	3992.7
DS2	1050	906.4
DS3	1050	817.5
DS4	1050	5241.9

The table illustrates the fact that the variance of individual vehicle emissions, as computed by the model, depends on the nature of the driving sequence. In the case of the FTP, the low variance reflects, at least in part, the fact that the time duration of the sequence, 505 seconds, is considerably less than the time duration of the other modes.

For a check on the validity of the variances as estimated from the variance-covariance matrix of the model coefficients, compare Table 5 with Table 6 below.

Table 6

MEANS AND VARIANCES OF
CALCULATED AND OBSERVED BAG VALUES (GMS)

	FEDERAL TEST PROCEDURE				SURVEILLANCE DRIVING SEQUENCE			
	MODEL		OBSERVED		MODEL		OBSERVED	
	MEAN	VAR	MEAN	VAR	MEAN	VAR	MEAN	VAR
HC	18.23	336.81	21.05	380.69	46.34	2173.4	53.55	2680.4
CO	214.51	23010.1	223.69	22760.7	582.0	180900.0	625.10	210610.2
NOX	16.72	77.13	17.22	81.62	50.9	699.1	48.17	647.3

In Table 6, the columns labeled "model" were obtained by computing, for each of the 1020 vehicles, the bag value as determined by application of the model emission model. The quantities were then averaged to obtain the mean bag value for the model and their variance was computed by the usual formula

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{1020} x_i^2 - \left(\sum_{i=1}^{1020} x_i\right)^2 \Big/ N}{N-1}$$

where x_i denotes the model-computed bag values for the i^{th} vehicle.

In conclusion, it is noted that the variances of the bag values, as computed by the model, are comparable with the variances of the bag values as actually determined by test. Also, in view of the agreement between Table 5 and Table 6, a method is at hand for estimating the vehicle-to-vehicle variance within a group for any driving sequence. This capability, in turn, makes it practical to estimate the standard error of the group mean.

5. PREDICTION OF FUEL ECONOMY

In view of the fact that the modal analysis emission model provides a means to estimate pollutant emission over any arbitrary driving sequence, it appeared feasible to employ the model to estimate fuel consumption by means of the carbon balance equation. In this connection, reference is made to work by M.E. Williams *et al* with regard to the FY 72 exhaust emission surveillance program.*

The carbon balance equation relates the amount of fuel consumed per mile to the amount of carbon-containing emissions produced per mile. Using the output of the modal emissions model as input into this equation allows one to estimate the fuel consumption over any driving sequence. The carbon-containing emissions that must be inputted are carbon monoxide (CO), carbon dioxide (CO_2), and hydrocarbons (HC).

The carbon balance method of calculating fuel economy in miles per gallon (mpg) is given as:

$$\text{mpg} = \frac{\text{grams of carbon/gallon of fuel}}{\text{grams of carbon in exhaust/mile}}$$

The actual equation incorporated into the model to estimate miles per gallon is

$$\text{mpg} = \frac{2423.0}{0.866\ (\text{HC}) + 0.429\ (\text{CO}) + 0.273\ (\text{CO}_2)}$$

where HC, CO, and CO_2 emissions are estimated in terms of grams/mile.* Implementation of the formula required, first of all, appropriate formulation of the modal analysis emission model to predict CO_2 emissions in addition to CO and HC. Then it was a straightforward matter to substitute these predicted quantities into the carbon-balance equation to obtain predictions of miles per gallon.

---

*M.E. Williams, J.T. White, L.A. Platte, and C.J. Domke, Automobile Exhaust Emission Surveillance - Analysis of the FY 72 Program, Report No. EPA-460/2-74-001, U.S. Environmental Protection Agency, Ann Arbor, Michigan (February 1974)

5.1 PREDICTION OF CO_2

Since the ability of the modal analysis emission model to predict CO_2 was not investigated under Contract Number 68-01-0435, it was necessary to examine the model's effectiveness in predicting CO_2 emissions prior to using these estimates in the carbon balance equation.

In order to determine the form of the emission rate function that should be used to represent CO_2 emissions, the average emission rate of the 1020 vehicles in the data base for each of the steady state modes was plotted versus speed. This curve is shown in Figure 15. On the basis of this figure, the assumption was made that the steady state and accel/decel emission rate functions for CO_2 could be represented by the same weighted quadratic functions of speed and acceleration as those used for HC, CO, and NOX in the original formulation of the model.

By means of the composite emission rate function, the amount of CO_2 emitted for each of the 1020 vehicles was estimated for the Surveillance Driving Sequence (SDS) and for the first 505 seconds of the Federal Test Procedure (FTP) driving sequence. These estimates are reported in Table 7, where they are compared with results as observed in the actual emissions tests. The notation in the table is as follows:

$\overline{O}$ = observed mean bag value for CO_2, in gms/mile, for 1020 vehicles

$\overline{R}$ = difference between mean emissions predicted by the model and the observed mean bag value for CO_2 (gms/mile)

σ_R = standard deviation of errors for individual vehicles

A visual appreciation of the distribution of the errors for individual vehicles is afforded by Figure 16 for the SDS and Figure 17 for the FTP. The occurrence frequencies on which these histograms are based are tabulated in Table 8.

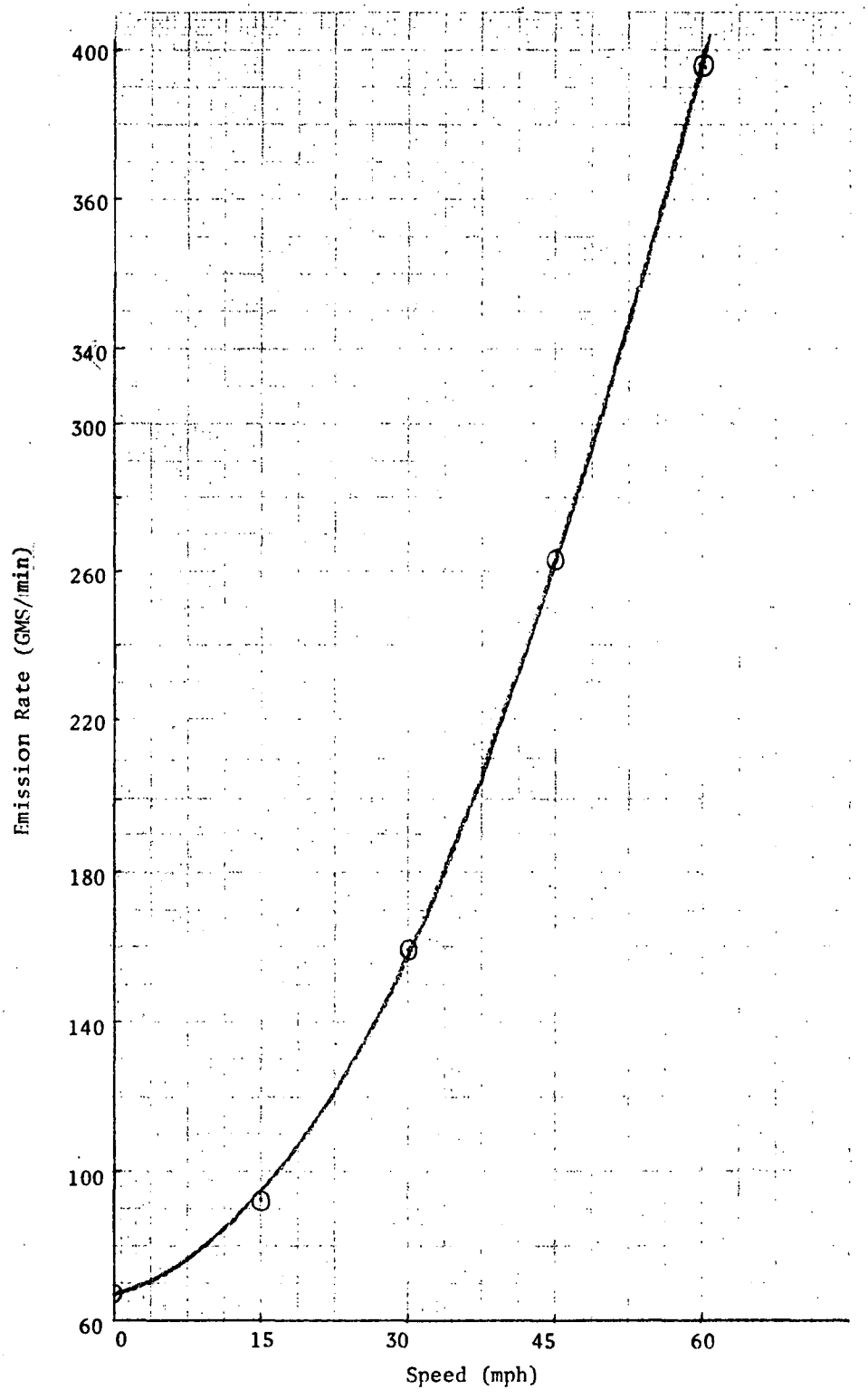


Figure 15 MEAN STEADY STATE CO_2 EMISSION RATES VS. SPEED

Table 7

COMPARATIVE STATISTICS FOR CO_2

Statistic	Surveillance Driving Sequence	FTP (First 505 Sec) Driving Sequence
$\bar{O}$	4347.4	1662.7
$\bar{R}$	270.6	141.5
σ_R^2	356373	108756
σ_R	597.0	329.8
$\sqrt{\bar{R}^2 + \sigma_R^2}$	655.4	358.9
$\frac{\bar{R}}{\bar{O}} \times 100\%$	6.22	8.51
$\frac{\sigma_R}{\bar{O}} \times 100\%$	13.73	19.8
$\frac{\sqrt{\bar{R}^2 + \sigma_R^2}}{\bar{O}} \times 100\%$	15.08	21.6

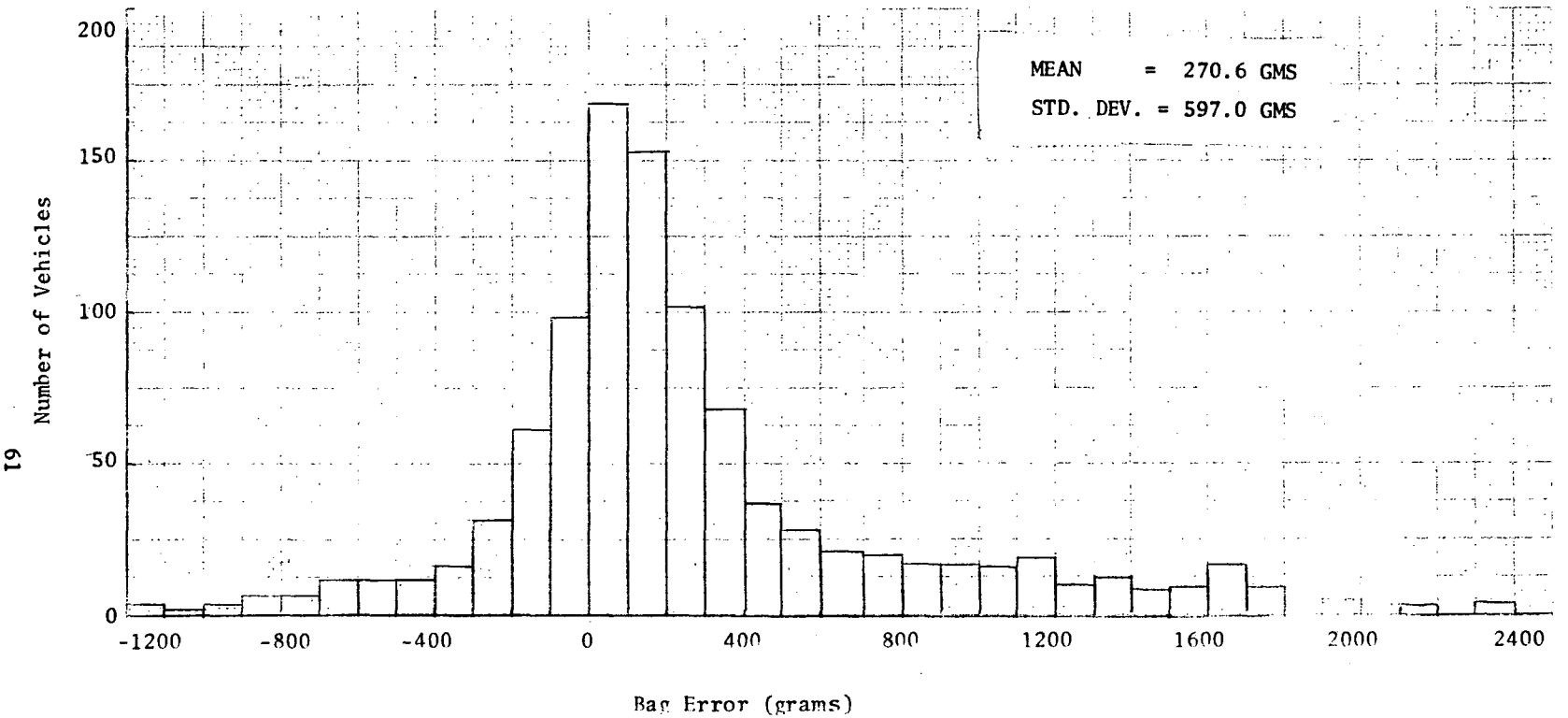


Figure 16 DISTRIBUTION OF CO_2 BAG ERROR FROM THE SURVEILLANCE DRIVING SEQUENCE

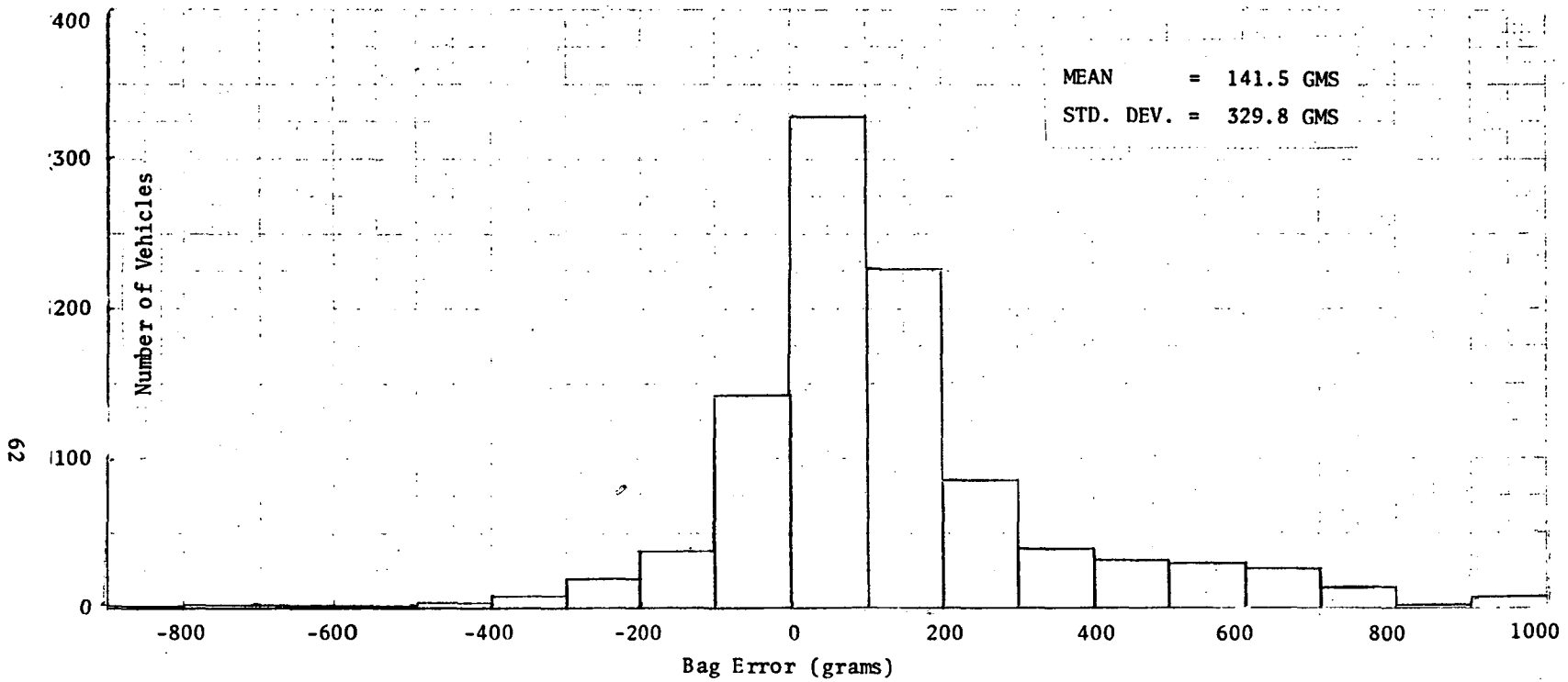


Figure 17 DISTRIBUTION OF CO_2 BAG ERROR (FIRST 505 SECONDS FTP)

Table 8

DISTRIBUTION OF CO_2 BAG VALUE ERROR (OBSERVED-CALCULATED) FROM THE FIRST 505 SEC OF THE FEDERAL TEST PROCEDURE (FTP) AND THE SURVEILLANCE DRIVING SEQUENCES (SDS)

ERROR (GMS)	NUMBER OF VEHICLES FTP	SDS
-2400 to -2300	0	1
-1800 to -1700	0	1
-1700 to -1600	0	2
-1600 to -1500	0	2
-1300 to -1200	1	0
-1200 to -1100	0	4
-1100 to -1000	0	2
-1000 to -900	1	3
-900 to -800	0	7
-800 to -700	3	8
-700 to -600	2	11
-600 to -500	2	11
-500 to -400	3	11
-400 to -300	7	16
-300 to -200	20	32
-200 to -100	38	61
-100 to 0	143	98
0 to 100	328	169
100 to 200	226	153
200 to 300	86	102
300 to 400	40	68
400 to 500	33	36
500 to 600	29	28
600 to 700	26	21
700 to 800	13	20
800 to 900	2	17
900 to 1000	8	17
1000 to 1100	2	14
1100 to 1200	1	19
1200 to 1300	1	10
1300 to 1400	0	12
1400 to 1500	0	8
1500 to 1600	1	9
1600 to 1700	0	16
1700 to 1800	0	9
1800 to 1900	0	3
1900 to 2000	0	5
2000 to 2100	0	3
2100 to 22200	1	3
2300 to 2400	0	4
2700 to 2800	0	2
2900 to 3000	0	1
3400 to 3500	1	1
4200 to 4300	1	0
4700 to 4800	1	0

Several observations can be made from Table 7 and from Figures 16 and 17 which are useful in judging the adequacy of the model. First, the mean or expected difference, $\overline{R}$, between the calculated and the observed values should be zero if the model is to be considered unbiased. The error distributions show that the bag error clusters around the average error $\overline{R}$, and this average for both driving sequences deviates from zero by only a few percentage points of the average measured bag values. Also, the root mean square error which represents the combined systematic and random errors as represented respectively by $\overline{R}$ and σ_R , are largely dominated by the random error component.

In order to further judge what these measures indicate as to the predictive performance of the model, the results for CO_2 were compared with the results obtained by using replicate data. Of the 1020 vehicles in the Surveillance Driving Sequence, 61 had been tested twice each. Thus there were available 61 replicate measurements from which could be obtained a measure of the repeatability of the test measurements themselves. Estimates of the mean, $\overline{X}$, standard deviation, $\hat{\sigma}$, and relative or percent standard deviation, $\hat{\sigma}/\overline{X}$, are given in Table 9 for the SDS, for the FTP, and for individual modes. The percent standard deviation characterizes the repeatability of measurements. These values are 8.34% for the SDS and 9.65% for the FTP driving sequence. For the individual modes, the percent standard deviation ranged from 6% to about 40%. This large variability in the test measurements is reflected as errors in the determination of the regression coefficients which in turn determine the error in estimating the instantaneous emission rate at any point in (v, a)-space. In view of the relatively large errors in the modal input data, the errors obtained for model performance do not appear unreasonable.

5.2 PREDICTION OF MILES PER GALLON

Prediction of fuel consumption in terms of miles per gallon is achieved by substituting the computed emissions of CO_2, CO and HC into the carbon balance equation. Though direct measurements of miles per gallon were not available, it was possible to compute "observed" values by

Table 9 REPLICATE MODAL ANALYSIS OF CO_2 FOR 61 VEHICLES

MODE	TIME (sec)	$\bar{X}$ (gm/min)	$\hat{\sigma}^2$ (gm/min)^2	$\hat{\sigma}$ (gm/min)	$\hat{\sigma}/\bar{X} \cdot 100\%$
1	12	1004.87	39566.29	198.91	19.79
2	16	329.28	7501.51	86.61	26.30
3	8	1129.19	192329.25	438.55	38.84
4	11	693.08	14943.85	122.24	17.64
5	13	580.97	8654.38	93.03	16.01
6	12	194.23	1909.12	43.69	22.50
7	17	673.03	15976.41	126.40	18.78
8	12	217.70	2632.25	51.31	23.57
9	14	569.71	9127.54	95.54	16.77
10	30	202.64	2134.50	46.20	22.80
11	26	671.98	12228.70	110.58	16.46
12	21	229.06	4525.77	67.27	29.37
13	32	726.05	11218.22	105.92	14.59
14	23	190.92	1952.54	44.19	23.14
15	9	236.69	2042.98	45.20	19.10
16	8	569.40	18156.47	134.75	23.66
17	22	731.97	12361.45	111.18	15.19
18	16	201.06	3033.29	55.08	27.39
19	18	671.54	17670.24	132.93	19.79
20	19	245.55	6892.80	83.03	33.81
21	25	748.31	11888.24	109.03	14.57
22	28	223.31	2553.20	50.53	22.63
23	15	882.50	19518.84	139.71	15.83
24	25	582.99	7184.77	84.76	14.54
25	18	191.30	1578.71	39.73	20.77
26	10	333.03	7648.20	87.45	26.26
27	38	644.71	12949.74	113.80	17.65
28	35	216.89	2629.09	51.27	23.64
29	18	777.47	19903.79	141.09	18.15
30	21	601.46	16428.74	128.17	21.31
31	14	191.20	2318.55	48.15	25.18
32	13	316.58	4451.17	66.72	21.07
33	60	73.43	170.64	13.06	17.79
34	60	387.26	3180.27	56.39	14.56
35	60	330.37	560.51	23.68	7.17
36	60	357.10	439.12	20.96	5.87
37	60	402.84	1397.43	37.38	9.28
FTP (gm)		483.86	2178.55	46.67	9.65
SDS (gm)		457.40	1453.77	38.13	8.34

incorporating the observed values of CO_2, CO and HC into the carbon balance equation. Miles per gallon figures, as based on the model outputs and as based on the observed bag values for the three emission products, could then be compared.

The results of this comparison are given in Table 10. The applicable notation is as follows:

$\bar{O}$ = mean value of miles per gallon for 1020 vehicles, as computed from bag values for CO_2, CO and HC emissions

$\bar{R}$ = mean difference between "observed" miles per gallon and miles per gallon as determined from model outputs for CO_2, CO and HC emissions

σ_R = standard deviation of errors for individual vehicles

As was the case for CO_2, the quantity $\bar{R}$ denotes a systematic error, whereas σ_R denotes a random error. Visual appreciation of the distribution of errors is afforded by Figures 18 and 19. The frequencies on which these histograms are based are given in Table 11.

In view of the errors in the input modal data, as revealed by replicate analyses, the predictive ability of the model is considered to be good. In this connection, it is remarked that errors in modal measurements of CO_2, CO and HC are compounded in the computation of miles per gallon.

Table 10

Statistics for Miles/Gallon Error Based on Bag Values

Statistic	Surveillance Driving Sequence	FTP (First 505 Sec) Driving Sequence
$\bar{O}$	17.07	16.44
$\bar{R}$	-1.00	-1.30
σ_R^2	12.49	6.63
σ_R	3.53	2.58
$\sqrt{\bar{R}^2 + \sigma_R^2}$	3.67	2.88
$\frac{\bar{R}}{\bar{O}} \times 100\%$	-5.88	-7.89
$\frac{\sigma_R}{\bar{O}} \times 100\%$	20.70	15.67
$\frac{\sqrt{\bar{R}^2 + \sigma_R^2}}{\bar{O}} \times 100\%$	21.52	17.54

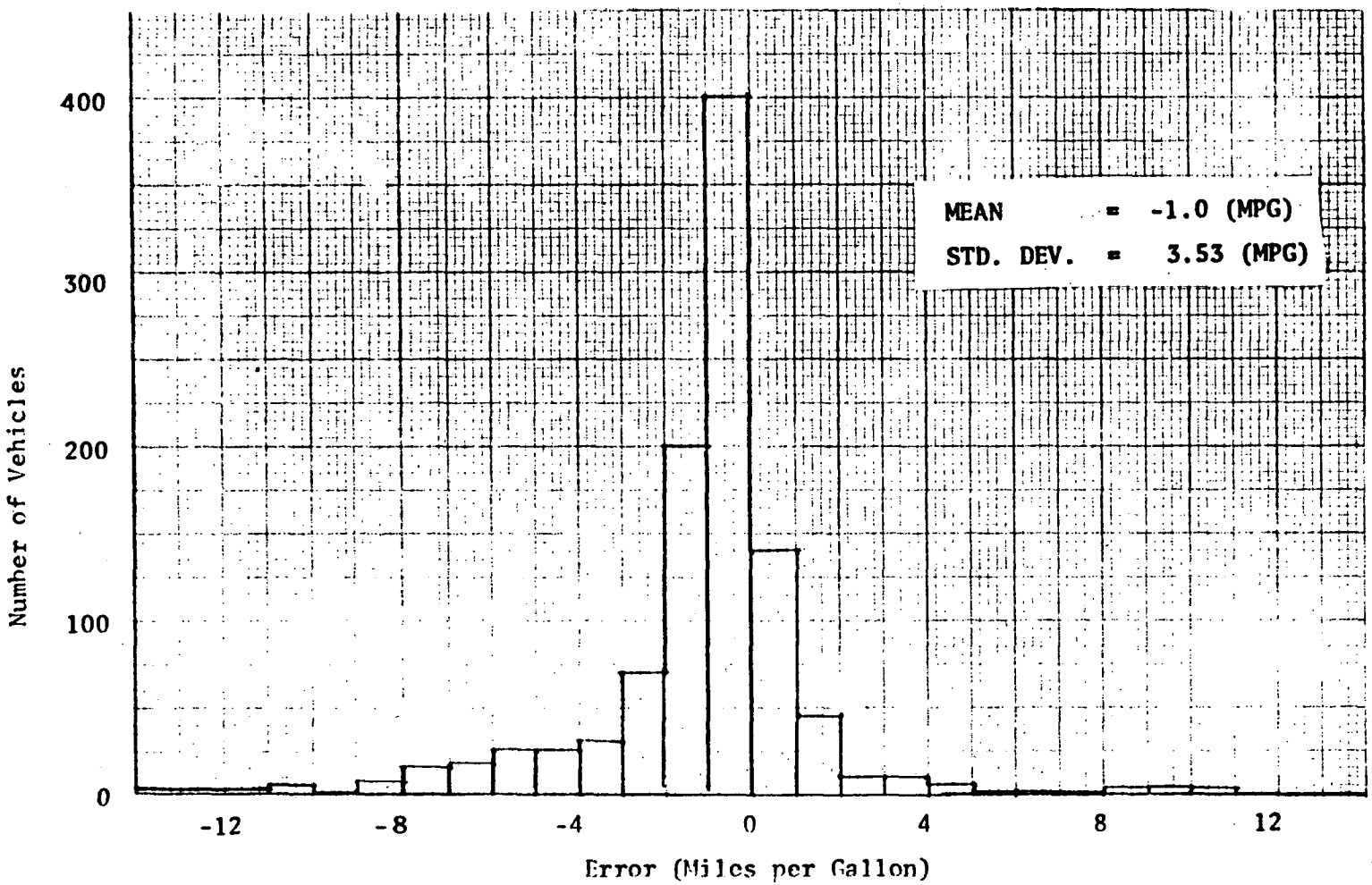


Figure 18 DISTRIBUTION OF ERROR FOR MILES PER GALLON BASED ON BAG VALUES FROM THE SURVEILLANCE DRIVING SEQUENCE

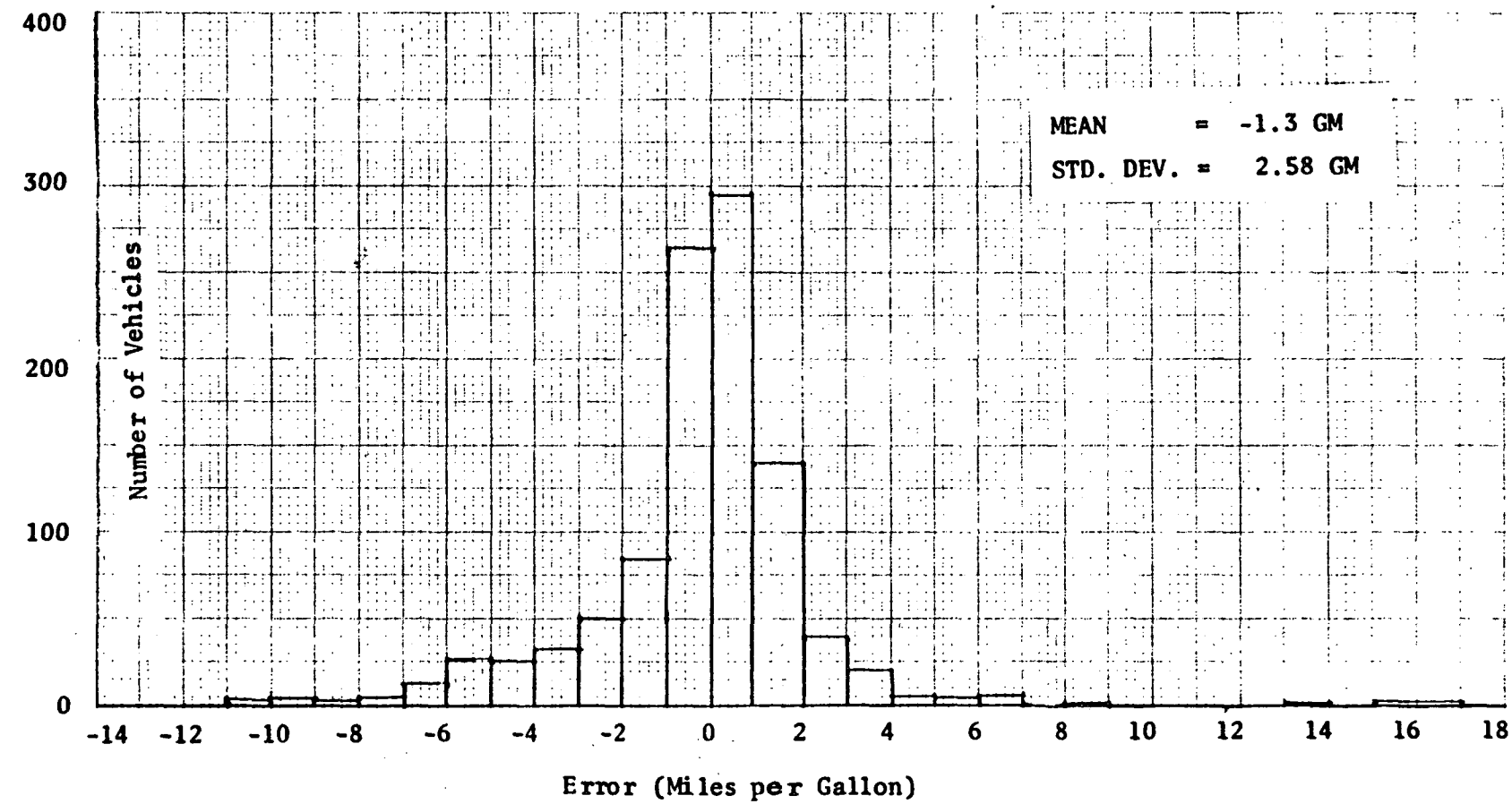


Figure 19 DISTRIBUTION OF ERROR FOR MILES PER GALLON BASED ON OBSERVED BAG VALUE CALCULATIONS (FIRST 505 SECONDS FTP)

Table 11

DISTRIBUTION OF MILES/GALLON ERROR (OBSERVED-CALCULATED) FROM THE FIRST 505 SEC OF THE FEDERAL TEST PROCEDURE (FTP) AND THE SURVEILLANCE DRIVING SEQUENCES (SDS)

ERROR (MPG)	NUMBER OF VEHICLES	
	FTP	SDS
-14 to -13	0	2
-13 -12	0	1
-12 -11	0	1
-11 -10	4	5
-10 -9	3	1
-9 -8	4	6
-8 -7	5	14
-7 -6	13	17
-6 -5	27	26
-5 -4	25	27
-4 -3	32	30
-3 -2	50	71
-2 -1	84	200
-1 0	263	397
0 1	296	140
1 2	139	45
2 3	37	12
3 4	18	8
4 5	5	6
5 6	4	2
6 7	5	0
7 8	0	1
8 9	1	1
9 10	0	0
10 11	1	2
11 12	0	0
12 13	0	1
13 14	1	1
14 15	1	1
15 16	1	0
16 17	1	1
17 18	0	1
22 23	1	0
77 78	0	1

6. SUMMARY AND CONCLUSIONS

Refinements and extensions of the automobile exhaust modal analysis model, as originally reported in EPA-460/3-74-005, have been completed in four important areas:

1) Increased computational efficiency
2) Reduction of modal testing requirements
3) Accuracy and precision of group emission predictions
4) Prediction of fuel economy

These improvements broaden the capability of the model and provide increased opportunities for applying the results of standard emissions tests to new contexts.

Improved computational efficiency derives primarily from a simplification of the method by which the instantaneous emission function $\dot{e}(v, a)$ is integrated over a driving sequence. For a particular driving sequence, it was shown that integration of the basis functions over the sequence need be performed only once for all vehicles subjected to that sequence. Moreover, it was shown that vehicle factors and driving sequence factors affecting emissions can be essentially separated as specific vector quantities, the inner product of which yields the emissions for the particular vehicle and driving sequence combination.

Redefinition of modal testing requirements was examined by means of variance-function analysis and principal-component analysis. It was shown that, although some redundancy exists in the modes as formulated, there are also regions of the speed-acceleration plane not well represented by existing modes. It is indicated that the number of modes could be reduced without serious loss of information but that modes should also be introduced to cover the region of the (v, a)-plane in the vicinity of accelerations between -1 mph/sec and +1 mph/sec.

The accuracy and precision of the model in predicting group emissions was assessed by (1) comparing model predictions with observed test values for the SDS and FTP driving sequences, and (2) evolving a scheme for defining

model output variance for an arbitrary driving sequence. Variances of the model predictions for the SDS and FTP compare favorably with the corresponding variances of observed bag values as actually determined by test. The scheme for computing model output variance for an arbitrary driving sequence stems directly from the basis function integration simplifications evolved in improving model computational efficiency. In particular, it is shown that if the variance-covariance matrix of the model coefficients for the vehicles comprising a group is known, this information can be adapted to defining the variance of total emissions over any driving sequence. It is necessary only to know, in addition to this variance-covariance matrix, the integrated forms of the basis functions for the driving sequence under consideration.

Prediction of fuel economy by means of the model can be achieved by developing an equation for the emission-rate surface for CO_2 in addition to CO and HC. Then, by means of a carbon-balance equation, the total output of carbon-containing emission products can be transformed into an estimate of the amount of fuel which produced these emission products. Accuracy and precision of the miles-per-gallon predictions are considered to be limited primarily by the errors in measuring the modal outputs of CO_2, CO and HC.

APPENDIX I

VARIANCE FUNCTIONS FOR REGRESSION ESTIMATES

Regression analysis is one of the basic tools employed in the formulation of the modal analysis emission model. As used in this context, it is to be understood in a generalized way permitting a relatively wide choice of form for the regression equation. Of particular interest is precision of the regression model over the treatment space; this precision can be evaluated by means of the concept of variance functions as discussed in this appendix.

In the following discussion, the mathematical basis of generalized regression analysis is presented, together with a discussion of variance functions and their visualization as a variance surface by means of variance maps.

1. FOUNDATIONS OF GENERALIZED REGRESSION ANALYSIS

Assume that a variable ζ, which shall be referred to as the response variable, depends on k experimental variables $x_1, x_2, \ldots, x_k$ and that a functional relation

$$\zeta = f(x_1, x_2, \ldots, x_k) \tag{I-1}$$

exists. Under certain conditions, the response equation (I-1) can be expanded as

$$\zeta = \beta_1 f_1(x_1,x_2,\ldots,x_k) + \beta_2 f_2(x_1,x_2,\ldots,x_k) + \beta_3 f_3(x_1,x_2,\ldots,x_k) + \ldots \tag{I-2}$$

where the β_i, $i = 1, 2, 3, \ldots$ are constants to be determined. The functions $f_1, f_2, f_3 \ldots$ are of arbitrary form provided only that they are linearly independent and do not involve the constants β_i. Equation (I-2) is thus a linear function of the β_i, although the f_i may be nonlinear in $x_1, x_2, \ldots, x_k$. Equation (I-2) is said to be a linear model, and the functions f_i may be regarded as basis vectors spanning a vector space comprising a certain class of functions.

Suppose, now, that the function (I-1) is to be estimated from experimental observations. The variables x_1, x_2, ..., x_k constitute a k-dimensional space called the x-space, and one may estimate (I-1) from observations taken at n points in this space. These n points constitute what shall subsequently be referred to as the experimental design. In general, ζ cannot be observed at these points because of error. Rather, it is possible only to observe a variable y related to ζ by

$$y = \zeta + \epsilon \tag{I-3}$$

where ϵ is a random error. Then (I-2) assumes the form

$$y = \beta_1 f_1(x_1,x_2,\ldots,x_k)+\beta_2 f_2(x_1,x_2,\ldots,x_k)+\ldots+\beta_p f_p(x_1,x_2,\ldots,x_k)+\epsilon \tag{I-4}$$

Since ϵ is a random variable, the responses observed at each design point also constitute a random variable. As a result, it is possible only to obtain from the observations an equation of the form

$$\hat{y} = b_1 f_1(x_1,x_2,\ldots,x_k)+b_2 f_2(x_1,x_2,\ldots,x_k)+\ldots+b_p f_p(x_1,x_2,\ldots,x_k) \tag{I-5}$$

where $\hat{y}$ is an estimate of y and b_i, $i = 1, 2, \ldots, p$ is an estimate of β_i.

Clearly, two types of errors can affect the approximation of the function (I-1). First, if (I-1) is to be approximated by the linear model (I-2), then (I-1) must belong to the class of functions spanned by the basis functions $f_1, f_2, \ldots, f_p$. Second, some means must be found for minimizing or controlling the effect of the random errors ϵ, since these will affect the estimation of the β_i. Generally, the form of (I-1) is unknown at the outset, and the experimenter has the option of assuming a set of basis functions according to experience or prior knowledge concerning the system under study. For control of random errors, the theory of least squares is employed.

In view of equation (I-5), each observation y_j obtained in the process of data collecting can be represented as

$$y_j = b_1 f_1(x_{1j}, x_{2j}, \ldots, x_{kj}) + b_2 f_2(x_{1j}, x_{2j}, \ldots, x_{kj}) + \ldots + \epsilon_j \qquad \text{(I-6)}$$

where the b_i are estimates of β_i and the ϵ_j are random errors as computed from

$$\epsilon_j = y_j - \hat{y}_j \qquad \text{(I-7)}$$

At the outset, $\hat{y}$ is not known, and it is the object of the least-squares algorithm to estimate $\hat{y}$ in such a way that

$$\sum_{j=1}^{n} = \text{a minimum.} \qquad \text{(I-8)}$$

We proceed to display the theory of this algorithm.

In matrix notation, equation (I-6) can be written as

$$\underline{y} = X\,\underline{b} + \underline{e} \qquad \text{(I-9)}$$

where $\underline{y}$ and $\underline{e}$ are n-rowed column vectors (or $n \times 1$ matrices), $\underline{b}$ is a p-rowed column vector (or $p \times 1$ matrix), and X is a matrix of dimension $n \times p$. The set of points at which observations are made will be referred to as the design region, and the set of functions $f_1, f_2, \ldots, f_p$ will be referred to as the basis functions or simply the basis.

For a two-variable case, the X- matrix is generated as shown below.

X - MATRIX

	f_1	f_2	...	f_p
x_{11}, x_{21}	$f_1(x_{11}, x_{21})$	$f_2(x_{11}, x_{21})$	...	$f_p(x_{11}, x_{21})$
x_{11}, x_{22}	$f_1(x_{11}, x_{22})$	$f_2(x_{11}, x_{22})$	...	$f_p(x_{11}, x_{22})$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
x_{11}, x_{2M}	$f_1(x_{11}, x_{2M})$	$f_2(x_{11}, x_{2M})$	...	$f_p(x_{11}, x_{2M})$
x_{12}, x_{21}	$f_1(x_{12}, x_{21})$	$f_2(x_{12}, x_{21})$	...	$f_p(x_{12}, x_{21})$
x_{12}, x_{22}	$f_1(x_{12}, x_{22})$	$f_2(x_{12}, x_{22})$	...	$f_p(x_{12}, x_{22})$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
x_{12}, x_{2M}	$f_1(x_{12}, x_{2M})$	$f_2(x_{12}, x_{2M})$	...	$f_p(x_{12}, x_{2M})$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
x_{1K}, x_{21}	$f_1(x_{1K}, x_{21})$	$f_2(x_{1K}, x_{21})$	...	$f_p(x_{1K}, x_{21})$
x_{1K}, x_{22}	$f_1(x_{1K}, x_{22})$	$f_2(x_{1K}, x_{22})$	...	$f_p(x_{1K}, x_{22})$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
x_{1K}, x_{2M}	$f_1(x_{1K}, x_{2M})$	$f_2(x_{1K}, x_{2M})$	...	$f_p(x_{1K}, x_{2M})$

The columns of the matrix are identified with the basis functions, the rows with the design points. In the example shown, x_1 assumes K distinct values and x_2 assumes M distinct values, so that there are $n = KM$ points in the design. Each basis function is evaluated at every point in the design, and the resulting $n \times p$ array constitutes the X-matrix. If the array is rearranged, so that the rows become columns and the columns become rows, the resulting matrix is the transpose of X. The matrix X and its transpose X' will be used extensively in the sequel.

Consider (I-9) and write the error vector $\underline{e}$ as

$$\underline{e} = \underline{y} - X\,\underline{b} \qquad \text{(I-10)}$$

The sum of squares of the components $e_1, e_2, \ldots, e_n$ of the error vector $\underline{e}$ can be written as

$$Q = \underline{e}'\underline{e} = (\underline{y}-X\underline{b})'(\underline{y}-X\underline{b}) \qquad \text{(I-11)}$$

where $\underline{e}'$ is the transpose of $\underline{e}$ and $(\underline{y} - X\underline{b})'$ is the transpose of $(\underline{y} - X\underline{b})$.

In extenso, (I-11) becomes (since $n = KM$)

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{k=1}^{K}\sum_{m=1}^{M} e_{km}^2$$
$$= \sum_{k=1}^{K}\sum_{m=1}^{M} \left[y_{km} - b_1 f_1(x_{1k}, x_{2m}) - b_2 f_2(x_{1k}, x_{2m}) - \ldots - b_p f_p(x_{1k}, x_{2m})\right]^2 \qquad \text{(I-12)}$$

By differentiating (I-12) with respect to the b_j , one obtains a set of p equations of the form

$$\frac{\partial Q}{\partial b_j} = 0 \qquad \text{(I-13)}$$

which can be solved for the b_j to minimize Q . The result can be summarized succinctly in the form

$$X'X\,\underline{b} = X'\,\underline{y} \qquad \text{(I-14)}$$

where $X'X$ is a square matrix of order p . Equation (I-14) provides the so-called normal equations of least squares.
Then

$$\underline{b} = (X'X)^{-1}\,X'\,\underline{y} \qquad \text{(I-15)}$$

is the formal solution minimizing the error sum of squares.

It is of interest to investigate the statistical properties of the least squares solution under certain assumptions. For the error vector we assume that

$$E(\underline{\epsilon}) = 0 \qquad (I\text{-}16)$$
$$E(\underline{\epsilon}\,\underline{\epsilon}') = I\sigma^2$$

where E denotes expectation and I is the identity matrix of order n. Equations (I-16) are equivalent to the assumption that the errors are uncorrelated, with mean zero and constant variance σ^2.

From equation (I-15), it is clear that

$$\underline{b} = C\,X'\,\underline{y} \qquad (I\text{-}17)$$

where $C = (X'X)^{-1}$. Note that, for the design points, (I-4) can be written as

$$\underline{y} = X\underline{\beta} + \underline{\epsilon} \qquad (I\text{-}18)$$

Substituting (I-18) into (I-17) one obtains

$$\underline{b} = C\,X'\,(X\underline{\beta}+\underline{\epsilon}) = C\,X'X\underline{\beta} + C\,X'\underline{\epsilon} = \underline{\beta} + C\,X'\underline{\epsilon} \qquad (I\text{-}19)$$

Since $\underline{\epsilon}$ is a random vector, $\underline{b}$ is also a random vector. Taking expectations in (I-19), one sees that

$$\begin{aligned} E(\underline{b}) &= E(\underline{\beta} + C\,X'\underline{\epsilon}) \\ &= E(\underline{\beta}) + E(C\,X'\underline{\epsilon}) \\ &= \underline{\beta} + C\,X'E(\underline{\epsilon}) = \underline{\beta} \end{aligned} \qquad (I\text{-}20)$$

Thus, the estimates provided by (I-15) are unbiased, provided the postulated form of the function ζ is correct. In the event of an incorrect choice of model, the estimates of the coefficients will be biased to an extent depending on the degree of discrepancy between the postulated and true models.

Unbiasedness derives from the ability to substitute for $\underline{y}$ its equivalent $X\underline{\beta} + \underline{\epsilon}$. Suppose the model is inadequate and requires additional basis functions so that the true model is

$$\underline{y} = X\underline{\beta} + X_1\underline{\beta}_1 + \underline{\epsilon}$$

where $X_1\underline{\beta}_1$ denotes the additional terms in the expansion. Then

$$\underline{b} = CX'\underline{y} = CX'(X\underline{\beta}+X_1\underline{\beta}_1+\underline{\epsilon})$$
$$= CX'X\underline{\beta} + CX'X_1\underline{\beta}_1 + CX'\underline{\epsilon}$$

and

$$E(\underline{b}) = \underline{\beta} + CX'X_1\underline{\beta}_1$$
$$= \underline{\beta} + A_1\underline{\beta}_1$$

The matrix $A_1 = CX'X_1 = (X'X)^{-1}X'X_1$ is called the alias matrix. Though it is useful in indicating the extent of confounding among various coefficients in the correct model, it is defined only in relation to an alternative hypothesis.

Consider, now, the covariance matrix of $\underline{b}$. Denoted $V(\underline{b})$, the covariance matrix is of order $p \times p$ and is given by

$$V(\underline{b}) = E\left[\underline{b}-E(\underline{b})\right]\left[\underline{b}-E(\underline{b})\right]' \qquad \text{(I-21)}$$

But, from (I-20)

$$\underline{b} - E(\underline{b}) = \underline{b} - \underline{\beta} \qquad \text{(I-22)}$$

and applying the results of (I-19) gives

$$\underline{b} - E(\underline{b}) = \underline{\beta} + CX'\underline{\epsilon} - \underline{\beta} = CX'\underline{\epsilon} \qquad \text{(I-23)}$$

Therefore,

$$V(\underline{b}) = E[CX'\underline{\epsilon}]\,[CX'\underline{\epsilon}]' = E[CX'\underline{\epsilon}\,\underline{\epsilon}'XC'] \tag{I-24}$$

or, since C is symmetric,

$$V(\underline{b}) = E[CX'\underline{\epsilon}\,\underline{\epsilon}'XC] \tag{I-25}$$

But, by the assumption of (I-16),

$$E(\underline{\epsilon}\,\underline{\epsilon}') = I\sigma^2$$

Therefore,

$$V(\underline{b}) = CX'XC\sigma^2 = C\sigma^2 \tag{I-26}$$

The results of (I-26) can be expressed as follows:

(a) The diagonal elements of the matrix $C = (X'X)^{-1}$, when multiplied by σ^2, the variance of the individual observations, provide the variances of the estimated coefficients in the model.

(b) The off-diagonal elements of C similarly provide the covariance between two estimated coefficients, b_i and b_j

2. VARIANCE FUNCTIONS

The variable y is often referred to as the response. We wish to study the variance to which the estimated response $\hat{y}$ is subject as we consider different points in the x-space. The necessary information can be obtained by an extension of the above reasoning.

Consider an arbitrary point (x_1, x_2) in treatment space and define a corresponding vector $\underline{x}$ as

$$\underline{x} = [f_1(x_1,x_2)\ f_2(x_1,x_2)\ \ldots\ f_p(x_1,x_2)] \tag{I-27}$$

Then the estimated response at that point is

$$\hat{y} = \underline{x}\underline{b} = b_1 f_1(x_1,x_2) + b_2 f_2(x_1,x_2) + \ldots + b_p f_p(x_1,x_2) \qquad \text{(I-28)}$$

where $\underline{b}$ is a column vector of coefficients.
Then,

$$E(\hat{y}) = E(\underline{x}\underline{b}) = \underline{x}E(\underline{b}) = x\beta \qquad \text{(I-29)}$$

and

$$\begin{aligned} \text{Var } (\hat{y}) &= E\left[\hat{y}-E(\hat{y})\right]\left[\hat{y}-E(\hat{y})\right]' \\ &= E\left[\underline{x}\underline{b}-\underline{x}\beta\right]\left[\underline{x}\underline{b}-\underline{x}\beta\right]' \\ &= E\left[\underline{x}(\underline{b}-\beta)\right]\left[\underline{x}(\underline{b}-\beta)\right]' \\ &= E\left[\underline{x}(\underline{b}-\beta)(\underline{b}-\beta)'\underline{x}'\right] \\ &= \underline{x}E\left[(\underline{b}-\beta)(\underline{b}-\beta)'\right]x' \end{aligned} \qquad \text{(I-30)}$$

But $E(\underline{b}-\beta)(\underline{b}-\beta)' = V(\underline{b})$. Therefore,

$$\text{Var } (\hat{y}) = \underline{x}(X'X)^{-1}\underline{x}'\sigma^2 \qquad \text{(I-31)}$$

Equation (I-31) gives the variance of the estimated response at an arbitrary point (x_1, x_2) in the sampling plane. Note that this variance depends strongly on the form of the X matrix, which is determined both by the location of the design points and the form of the basis functions.

Equation (I-31) theoretically provides an estimate of the variance of the estimated response at every point in the x-space. In the event that the x-space is two-dimensional, it is possible to display contour maps of this variance. The variance is computed at every point in an array of points in the x-space plane, and these values are then thresholded and displayed as a variance map. Figure I-1 provides a graphic presentation of such a variance

surface for a 3^2 factorial design using the basis functions listed. Similar techniques were applied to the modal analysis data in generating the variance maps exhibiting the effect of reallocation of modes on the precision of estimation of the model. In those applications, the two variables x_1 and x_2 were speed v and acceleration a.

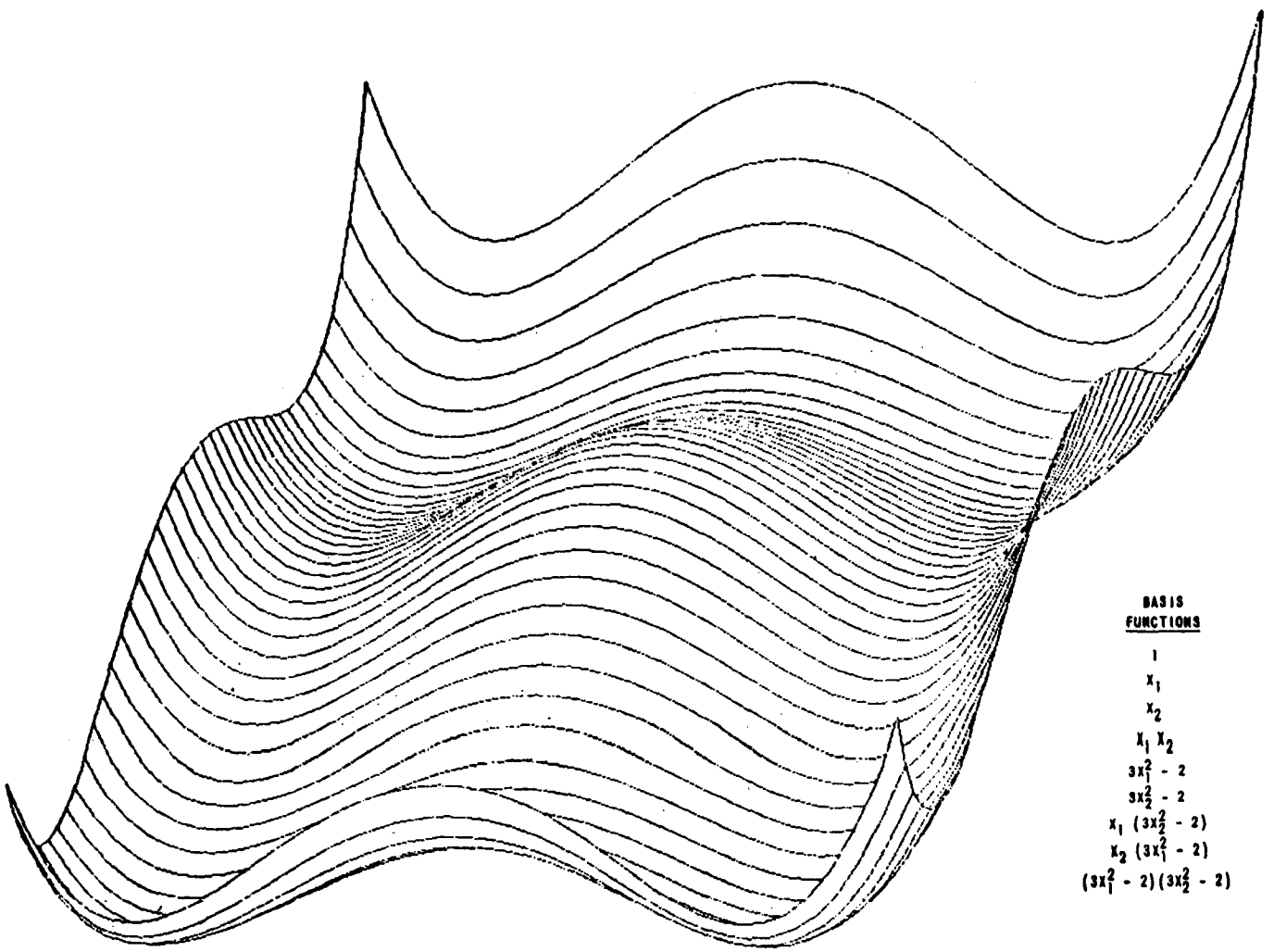


FIGURE I - 1 VARIANCE SURFACE FOR 3^2 FACTORIAL DESIGN

APPENDIX II

COMPUTER PROGRAM REVISIONS FOR INCREASED COMPUTATIONAL EFFICIENCY

MAIN PROGRAM I

```
C
C     MAIN PROGRAM I DETERMINES THE AMOUNT OF EMISSIONS GIVEN OFF BY
C     INDIVIDUAL VEHICLES OVER A DRIVING SEQUENCE SPECIFIED BY ARR'VVT'.
C
C     VTM(I)=>VELOCITY VS. TIME(IN ONE SECOND INTERVALS) OF THE SURVEIL-
C     -LANCE DRIVING SEQUENCE.VTM(I)=VELOCITY(MPH) AT TIME (I-1)SEC
C     (REAL*4)
C
C     VVT(I)=>VELOCITY VS. TIME(IN ONE SECOND INTERVALS) OF ANY DRIVING
C     SEQUENCE OVER WHICH  EMISSIONS ARE TO BE CALCULATED.VVT(I)=VELOC-
C     -ITY AT TIME (I-1) SEC. (REAL*4)
C
C     AMTC(I,J)=> AMOUNT OF I'TH EMMITTANT GIVEN OFF IN J'TH MODE.
C
C     DS(I)=DISTANCE(MILES)TRAVELED IN I'TH MODE.NOTE,STEADY STATE MODES
C     ARE 60 SEC IN DURATION.
C
C     FUNC(I)=> INTEGRATED BASIS FUNCTIONS CHARACTERIZING A
C                DRIVING SEQUENCE (REAL*8)
C
C
C     ***********************************************************************
      DIMENSION ITAB(20,2),IDAT(4,19),RDAT(161,19),DS(37)
      DIMENSION VTM(1055),VVT(2000),AMTC(4,37)
      REAL*8 C(4),FUNC(12)
      REAL*8 AA(9,32),AS(3,5),BAD(4,12),XMPG,HCGPM,COGPM,CO2GPM
      DATA DS           /.0602,.0741,.0201,.0705,.1360,.1268,.2163,.1716
     C,.2043,.3367,.3136,.1973,.3313,.2994,.0579,.0173,.1759,.1392,.1528
     C,.1304,.2654,.2634,.0737,.3134,.2362,.0444,.4009,.3293,.0886,.2599
     C,.1813,.0592,.0000,.2500,.5000,.7500,1.000/
      DEFINE FILE 99(75,3256,U,N1)
C
C      READ IN SURVEILLANCE DRIVING SEQUENCE
C
      PRINT 1003
 1003 FORMAT(1H0,'SURV. DRIVING SEQ.'//)
      DO 3000 I=1,100
      NX1=((I-1)*16)+1
      NX2=NX1+15
      READ(5,100)(VTM(K),K=NX1,NX2)
      PRINT 1002, (VTM(K),K=NX1,NX2)
```

```
 1002 FORMAT(1H0,16F8.0)
  100 FORMAT(16F5.0)
      IF(VTM(NX1).GT.99.0)GOTO3111
 3000 CONTINUE
 3111 CONTINUE
C
C     READ IN DRIVING SEQUENCE OVER WHICH EMMISSIONS ARE TO BE CALCULATE
C
C     IN THIS EXAMPLE VVT=> FIRST 505 SEC. OF FTP
C
      PRINT 1004
 1004 FORMAT(1H0,'FTP DRIVING SEQ.'//)
      NPTS=506
      DO 1500 I=1,100
      NX1=((I-1)*16)+1
      NX2=NX1+15
      READ(5,100)(VVT(K),K=NX1,NX2)
      PRINT 1002, (VVT(K),K=NX1,NX2)
      IF(VVT(NX1).GT.99.0)GOTO1555
 1500 CONTINUE
 1555 CONTINUE
C
C     SET UP BASIS FUNCTION FACTOR ARRAYS AA,AS.
C
      CALL SETUP(VTM,AA,AS)
C
C  INTEGRATE BASIS FUNCTIONS OVER THE DRIVING CYCLE
C
      CALL ESUM(VVT,NPTS,FUNC,DIST)
      PRINT 1006, DIST
 1006 FORMAT(1H0,'DISTANCE (MI) =',F5.3/)
C
C     READ IN INDIVIDUAL VEHICLE MODAL EMISSIONS DATA.
C     PUT MODAL EMISSIONS DATA INTO ARRAY AMTC
C
C      IN THIS EXAMPLE THE MODAL EMISSIONS DATA IS READ OFF A DISK FILE
C
      NCART=0
      PRINT 501
  501 FORMAT(1H1,' CAR ID',7X,'HC (GMS)',7X,'CO (GMS)',6X,'CO2 (GMS)',
     *      7X,'NOX (GMS)',9X,'MPG'//)
      READ(99'75)ITAB
      DO 2000 IY=57,71
      IREC=IY-56
      JSTART=ITAB(IREC,1)
      JEND=ITAB(IREC,2)
      DO 2001 J=JSTART,JEND
      READ(99'J) ((IDAT(L,K),L=1,4),(RDAT(L,K),L=1,161),K=1,19)
      DO 2002 K=1,19
      IF(IDAT(1,K).EQ.-9)GOTO2001
      NCART=NCART+1
```

```
      DO 1000 IR=1,37
      DD=1.0
C
C   FOR A/D MODES CHANGE DATA FROM GRMS/MILE TO GRMS
C
      IF(IR.LE.32)DD=DS(IR)
      DO 1001 IC=1,4
      IW=((IR-1)*4) + 13 + IC
      AMTC(IC,IR)=RDAT(IW,K)*DD
 1001 CONTINUE
 1000 CONTINUE
C
C     DETERMINE INDIVIDUAL VEHICLE EMISSION RATE FUNCTION COEFFICIENTS
C
      CALL EDOT(AMTC,AA,AS,BAD)
C
C     CLEAR ARRAY WHICH CONTAINS RESULTANT EMISSIONS
C
      DO 1005 I=1,4
 1005 C(I)=0.0D0
C
C
C   DETERMINE THIS VEHICLE'S EMISSIONS
C
      DO 4000 IC=1,4
      DO 4001 IE=1,12
 4001 C(IC)=C(IC) + FUNC(IE) * BAD(IC,IE)
 4000 CONTINUE
C
C   CALCULATE GMS/MILE AND USE TO GET MILES PER GALLON
C
      HCGPM=C(1)/DIST
      COGPM=C(2)/DIST
      CO2GPM=C(3)/DIST
      XMPG=2423.0D0/(0.866D0*HCGPM+0.429D0*COGPM+0.273D0*CO2GPM)
C
C     WRITE OUT EMISSION RESULTS
C
      PRINT 502, IDAT(3,K),(C(L),L=1,4),XMPG
  502 FORMAT(1X,I8,5(5X,F10.3))
 2002 CONTINUE
 2001 CONTINUE
 2000 CONTINUE
 1234 STOP
      END
```

```
      SUBROUTINE ESUM(VVT,NT,FUNC,DIST)
C     ******************************************************************
C
C     SUBROUTINE ESUM INTEGRATES THE BASIS FUNCTIONS OVER THE
C      INPUTTED DRIVING SEQUENCE AND DETERMINES THE DISTANCED TRAVELED
C
C
C     VVT(I)=>VELOCITY VS. TIME HISTORY(DRIVING CYCLE) IN ONE SECOND
C             INTERVALS.VVT(I)=VELOCITY(MPH) AT THE I'TH SECOND.REAL*4
C
C     NT=>MAXIMUM NUMBER SECONDS IN DRIVING CYCLE+1 SECOND
C
C     FUNC(I)=> INTEGRATED BASIS FUNCTIONS CHARACTERIZING THE DRIVING
C            SEQUENCE (REAL*8)
C
C     DIST=DISTANCE(MILES)IN SPECIFIED DRIVING CYCLE,REAL*4
C
C     ******************************************************************
      DIMENSION VVT(NT)
      REAL*8 X(12),FUNC(12),DIS,AMIN,AMAX,A1,A2,HOA
      AMAX=1.0D0
      AMIN=-1.20D0
      A1=-1.0D0/AMIN
      A2=-1.0D0/AMAX
C
C  CLEAR FUNC ARRAY
C
      DO 1000 I=1,12
 1000 FUNC(I)=0.0D0
C
C     INTEGRATE AUTO'S EMISSION RATE FUNCTION OVER DRIVING CYCLE
C
      DIS=0.0D0
      NTT=NT-1
      DO 3000 IT=1,NTT
      KT=IT+1
      X(1)=1.0D0
      X(2)=DBLE((VVT(IT)+VVT(KT))/2.0)
      X(3)=DBLE(VVT(KT)-VVT(IT))
      X(4)=X(2)*X(3)
      X(5)=X(2)**2
      X(6)=X(3)**2
      X(7)=(X(2)**2)*X(3)
      X(8)=(X(3)**2)*X(2)
      X(9)=(X(2)**2)*(X(3)**2)
      X(10)=X(1)
      X(11)=X(2)
      X(12)=X(5)
      IF(X(3).GE.AMAX)HOA=0.0D0
      IF(X(3).LE.AMIN)HOA=0.0D0
      IF(X(3).GE.0.0D0.AND.X(3).LT.AMAX)HOA=(A2*X(3))+1.0D0
      IF(X(3).LE.0.0D0.AND.X(3).GT.AMIN)HOA=(A1*X(3))+1.0D0
      DO 2999 IE=1,9
 2999 FUNC(IE)=FUNC(IE) + (1.0D0-HOA)*X(IE)
      DO 2998 IE=10,12
 2998 FUNC(IE)=FUNC(IE) + HOA * X(IE)
      DIS=DIS+X(2)

 3000 CONTINUE
      DIS=DIS/3600.0D0
      DIST=DIS
 4444 RETURN
      END
```

```
      SUBROUTINE EDOT(AMTC,AA,AS,BAD)
C     ***************************************************************
C      SUBROUTINE EDOT COMPUTES THE COEFFICIENTS THAT SPECIFY AN AUTO'S
C     INSTANTANEOUS EMISSION RATE FUNCTIONS FOR HC,CO,NOX(ARRAY 'BAD'),
C     GIVEN THE AMOUNT OF EACH EMITTANT GIVEN OFF BY THE AUTO IN 32 A/D
C     MODES AND 5 STEADY STATE MODES(ARRAY 'AMTC'),AND THE BASIS
C     FUNCTION FACTOR ARRAYS(AA,AS).
C
C*****THIS VERSION CALCULATES COEFFICIENTS FOR CO2 ALSO
C     THE DO 1000 LOOP CHANGED TO I=1,4
C
C      AMTC(I,J)=AMOUNT(GMS) OF THE I'TH EMITTANT GIVEN OFF BY THIS AUTO
C     IN THE J'TH MODE.  I=1=>HC,I=2=>CO,I=3=>CO2,I=4=>NOX,
C     J=1,37 (32 A/D MODES, 5 STEADY STATE MODES). (REAL*4)
C
C     BAD(I,J)=J'TH COEFFICIENT OF THIS AUTO'S INSTANTANEOUS EMISSION
C     RATE FUNCTION FOR THE I|TH KIND OF EMITTANT.I=1=>HC,I=2=>CO,
C     I=3=>CO2,I=4=>NOX. (REAL*8)
C
C     AA=>BASIS FI
C     AA=BASIS FUNCTION FACTOR ARRAY FOR ACEL/DECEL(CALCULATED BY SUBROU
C    -TINE SETUP).
C
C     AS=BASIS FUNCTION FACTOR ARRAY FOR STEADY STATE(CALCULATED BY
C     SUBROUTINE SETUP).
C
C     TM(I)=TIME(SEC) IN I'TH MODE.(REAL*4)
C
C     ***************************************************************
      DIMENSION TM(37),AMTC(4,37)
      REAL*8 AA(9,32),AS(3,5),BAD(4,12),SUM,YA(32),YS(5),B(3),XO,X1,X2
     C,A1,A2
      DATA TM/12.,16.,8.,11.,13.,12.,17.,12.,14.,30.,26.,21.,32.,23.,9.,
     C8.,22.,16.,18.,19.,25.,28.,15.,25.,18.,10.,38.,35.,18.,21.,14.,13.
     C,60.,60.,60.,60.,60./
      NOBSA=32
      NOBSS=5
      NBFA=9
      NBFS=3
C
      DO 1000 IC=1,4
C
C      IC=1=>HC,IC=2=>CO,IC=3=>CO2,IC=4=>NOX
C
C     CALCULATE OBSERVED AVERAGE EMISSION RATES OVER 32 A/D MODES
C
      DO 1100 I=1,32
      A1=AMTC(IC,I)
      A2=TM(I)
      YA(I)=A1/A2
 1100 CONTINUE
```

```
C
C       CALCULATE COEFFICIENTS THAT SPECIFY A/D EMISSION RATE FUNCTIONS
C
      DO 1200 I=1,NBFA
      SUM=0.0D0
      DO 1250 J=1,NOBSA
      SUM=SUM+(AA(I,J)*YA(J))
 1250 CONTINUE
      BAD(IC,I)=SUM
 1200 CONTINUE
C
C     CALCULATE OBSERVED AVERAGE EMISSION RATES OVER 5 SS MODES
C
      DO 2000 I=33,37
      IP=I-32
      A1=AMTC(IC,I)
      A2=TM(I)
      YS(IP)=A1/A2
 2000 CONTINUE
C
C     CALCULATE COEFFICIENTS THAT SPECIFY SS EMISSION RATE FUNCTIONS
C
      DO 2001 I=1,NBFS
      SUM=0.0D0
      DO 2100 J=1,NOBSS
      SUM=SUM+(AS(I,J)*YS(J))
 2100 CONTINUE
      B(I)=SUM
 2001 CONTINUE
C
C     CHECK ON EXISTANCE OF NEGATIVE EMISSION RATES
C
      LOOP=0
      IF(B(3).EQ.0.0D0)GOTO2151
      XO=(B(2)**2)-(4.0D0*B(3)*B(1))
      IF(XO.LT.0.0D0)GOTO2153
      XO=DSQRT((B(2)**2)-(4.0D0*B(3)*B(1)))
      X1=(-B(2)+XO)/(2.0D0*B(3))
      X2=(-B(2)-XO)/(2.0D0*B(3))
      IF((X1.GT.0.0D0.AND.X1.LT.60.0D0).OR.(X2.GT.0.0D0.AND.X2.LT.60.0D0
     C))LOOP=1
      GOTO2153
 2151 XO=-B(1)/B(2)
      IF(XO.GT.0.0D0.AND.XO.LT.60.0D0)LOOP=2
 2153 IF(LOOP.EQ.0)GOTO2154
C
C       IF LOOP=0=>NO NEGATIVE EMISSIONS FOR VELOCITYS BETWEEN 0,60
C
C     IF LOOP=1 OR 2=> NEGATIVE EMISSION RATES BETWEEN 0,60MPH.
C
C     CALL SUBROUTINE PAD TO FIND COEFFICIENTS WHICH DO NOT PRODUCE
C     NEGATIVE EMISSION RATES.
C
      CALL PAD(YS,B)
C
 2154 BAD(IC,10)=B(1)
      BAD(IC,11)=B(2)
      BAD(IC,12)=B(3)
 1000 CONTINUE
      RETURN
      END
```