Title
Stochastic Microenvironment Models for Air Pollution Exposure

Permalink
https://escholarship.org/uc/item/7z23w74b

Author
Duan, Naihua

Publication Date
1991-10-29

Peer reviewed

A RAND NOTE

Stochastic Microenvironment Models for Air Pollution Exposure

Naihua Duan

RAND

The research described in this report was supported by a grant from the Societal Institute of the Mathematical Sciences and by RAND using its own research funds.

This Note contains an offprint of RAND research originally published in a journal or book. The text is reproduced here, with permission of the original publisher.

Published 1991 by RAND
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138

A RAND NOTE

N-3415-SIMS/RC

Stochastic Microenvironment Models for Air Pollution Exposure

Naihua Duan

Supported by the
Societal Institute of the Mathematical Sciences

RAND

STOCHASTIC MICROENVIRONMENT MODELS FOR AIR POLLUTION EXPOSURE

NAIHUA DUAN
RAND Corporation

Exposure assessment is a crucial link in air pollution risk assessment and management. With the recent advances in instrumentation, it has become possible to measure air pollution exposure in the vicinity of the individual human subjects, using either personal monitoring or microenvironment monitoring. For many important pollutants such as CO, NO_2, and VOC, the air pollution exposure depends crucially on the location and activity of the individual: indoor versus outdoor, smoking versus not smoking, etc. The stochastic microenvironment models were developed to relate air pollution exposure to the location and activity. We review the two major existing models, the Cartesianization method (Duan, 1980, 1982, 1987) and SHAPE (Ott, 1981, 1982, 1984), and compare their assumptions and implications. We also propose a new model, the variance components model, which includes both Cartesianization and SHAPE as special cases. The variance components model considers both long-term average concentrations and short-term fluctuations. The Cartesianization focuses on long-term averages, while SHAPE focuses on short-term fluctuations. We propose to choose among the three models by examining the variance function which relates variability to averaging time.

The theory is applied to the data collected from U.S. EPA's Washington CO Study, with the variance function estimated using Carroll and Ruppert's (1984) transform-both-sides regression model and Duan's (1983) smearing estimate. For the microenvironment in transit, *both long-term averages and short-term fluctuations are important.*

[1] Address all correspondence to: Naihua Duan, RAND Corporation, 1700 Main Street, Santa Monica, CA 90407.

[2] Research was supported in part by a cooperative agreement between RAND, SIMS, and U.S. EPA, in part by RAND corporate funds. Part of this paper was presented at the Workshop on Modeling Commuter Exposure, Research Triangle Park, North Carolina, August 1988. I appreciate useful discussions with Joe Behar, Ted Johnson, Bill Nelson, Wayne Ott, Paul Switzer, and Jacob Thomas during the course of this research, and helpful comments from an anonymous referee.

Reprinted by permission from *Journal of Exposure Analysis and Environmental Epidemiology,* Vol. 1, No. 2, April 1991, pp. 235-257. Copyright © 1991 Princeton Scientific Publishing Co., Inc.

INTRODUCTION

Air pollution exposure has a variety of meanings; see, for instance, Duan, Dobbs, and Ott (1989). Given the appropriate instruments for continuous personal monitoring, we might consider each individual's *exposure profile*, i.e., the individual's instantaneous exposure taken as a function of time. We usually summarize the exposure profiles into summary measures such as integrated exposures,

$$Y_i = \int_a^b \mathrm{c_i(t)\ dt}, \tag{1}$$

where Y_i denotes the i-th individual's integrated exposure over a given time period (a,b), say, a twenty-four hour period, and $c_i(t)$ denotes the i-th individual's exposure profile as a function of time t. It is possible to consider other summary measures such as the maximum exposure, $M_i = \max_{a \leq t \leq b} c_i(t)$. We focus on the integrated exposure in this paper, although most of the principles are generalizable to other summary measures.

The integrated exposure usually varies from individual to individual. For air quality management, we need to consider the *distribution* of the integrated exposure in a target population:

$$F_Y(y) = P(Y \leq y). \tag{2}$$

We might, for example, base the regulatory decisions on the proportion of individuals with integrated exposures exceeding a safety threshold y_0. We will refer to the distribution F_Y as the *exposure distribution.*

For many pollutants such as CO, NO_2, and VOC, we can conduct *personal monitoring*: we take a random sample of n individuals from the target population, equip them with personal monitors, and observe their integrated exposures, $\{ Y_{1,\ldots,} Y_n\}$. We can then estimate the exposure distribution by the empirical distribution

$$\hat{F}_Y(y) = n^{-1} \sum_{i=1}^{n} 1(Y_i \leq y), \tag{3}$$

where $1(Y_i \leq y)$ is the indicator function for the event "$Y_i \leq y$." This has been known as the *direct approach.*

In many situations we also observe covariates, such as location and activity, which are predictive of instantaneous exposures. We can stratify each individual's instantaneous exposures into strata using those covariates. Each stratum will be called a *microenvironment.* For example, we might stratify by the location of the individual and define five microenvironments: *indoor at home, indoor at school, indoor at other locations, outdoor in transit,* and *outdoor not in transit.* Duan (1980) developed a criterion for comparing the merits of candidate stratification schemes.

Given a stratification of instantaneous exposures into microenvironments, we can decompse the integrated exposure as follows:

$$Y_i = \sum_{k=1}^{K} T_{ik} C_{ik}, \tag{4}$$

where T_{ik} denotes the amount of time the i-th individual spent in the k-th microenvironment during the time period (a,b), and C_{ik} denotes the average pollutant concentration for the i-th individual in the k-th microenvironment:

$$C_{ik} = \int_a^b c_i(t) 1_{ik}(t)\, dt \;/\; T_{ik}, \tag{5}$$

where $1_{ik}(t)$ is the indicator for the event "the i-th individual was in the k-th microenvironment at time t." The decomposition (4) has been known as the *microenvironment decomposition.*

We will refer to the row vector $\mathbf{T}_i = (T_{i1},...,T_{iK})$ as the *activity pattern* for the i-th individual, and the column vector $C_i = (C_{i1},...,C_{iK})'$ as the *microenvironment concentrations* for the i-th individual. Both **C** and **T** vary from individual to individual.

According to the microenvironment decomposition, we can represent the exposure distribution as follows:

$$F_Y(y) = \iint 1(\mathbf{tc} \leq y)\, dF_{\mathbf{C}|\mathbf{T}}(\mathbf{c} \mid \mathbf{t}) dF_{\mathbf{T}}(\mathbf{t}), \tag{6}$$

where $F_{\mathbf{C}|\mathbf{T}}$ denotes the conditional distribution of **C** given **T**, and will be called the *conditional microenvironment concentration distribution*; $F_{\mathbf{T}}$ denotes the marginal distribution of **T**, and will be called the *activity pattern distribution.*

As an alternative to the direct approach, we can also estimate the exposure distribution *indirectly* from the microenvironment decomposition: we estimate the conditional microenvironment concentration distribution $F_{\mathbf{C}|\mathbf{T}}$ and the activity pattern distribution $F_{\mathbf{T}}$, then use (6) to estimate the exposure distribution. This has been known as the *indirect approach.*

It is usually easy to estimate the activity pattern distribution $F_{\mathbf{T}}$. We can take a random sample of individuals from the target population, and conduct an activity survey, say, using an activity diary: we ask each sampled individual to keep a diary, and record the amount of time he spent in each microenvironment during the time period (a,b). From the activity survey, we observe the activity patterns $\{\mathbf{T}_1,...,\mathbf{T}_n\}$ for the sampled individuals. We then estimate $F_{\mathbf{T}}$ by the empirical distribution

$$\hat{F}_T(t) = n^{-1} \sum_{i=1}^{n} 1(\mathbf{T}_i \leq \mathbf{t}). \tag{7}$$

It is usually harder to estimate the conditional microenvironment concentration distribution $F_{C|T}$. Ideally, we like to have an *enhanced personal monitoring study* in which both **C** and **T** are observed on the same individuals. We can then estimate $F_{C|T}$ using an appropriate regression of **C** on **T**, such as Koencker and Bassett's (1978) quantile regression or Efron's (1991) percentile regression. Such an ideal study requires that the personal monitoring study be conducted in conjunction with an activity survey: the same individuals need to fill out activity diaries at the same time their exposures are being monitored. Furthermore, in order to measure **C**, we need to have a personal monitor which provides continuous measurements of the entire exposure profile.[1] We can then match the exposure profile with the activity diary to compute **C**. Such a study design was used in the Washington CO Study (Akland et al., 1985).

An important application for the indirect approach is to re-use the monitoring data from an enhanced personal monitoring study, so as to estimate the exposure distribution for a new target population. We assume the conditional microenvironment concentration distribution $F_{C|T}$ in the new population is the same as that in the previous study. Under this assumption, we need only conduct an activity survey on the new target population; we don't need to conduct personal monitoring in the new study.[2] We can combine the conditional microenvironment concentration distribution estimated from the previous study with the activity pattern distribution estimated from the new study, then use (6) to estimate the exposure distribution for the new target population.

Another important application for the indirect approach is known as a *microenvironment monitoring study*. We don't have any personal monitoring data. Instead, we have monitoring data collected from a random sample of the relevant microenvironments, say, homes, office buildings, etc.: we send technicians to measure the pollutant concentrations in those microenvironments. We assume that we can estimate the conditional microenvironment concentration distribution $F_{C|T}$ from those measurements.[3] Under this assumption, we need only conduct an activity survey on the target population; we don't need to conduct personal monitoring.

[1] For many pollutants such as VOC, there are no personal monitors which can provide continuous measurements. We can only observe Y and **T** on the sampled individuals, not C. Duan (1989) examines several approaches to estimate the conditional microenvironment concentration distribution for such studies.

[2] It is desirable to collect at least some personal monitoring data in the new study to validate of the assumption that the two conditional microenvironment concentration distributions are the same.

[3] This is an important assumption and requires that we have a *representative* sample of the microenvironments. The sampling of some microenvironments such as *homes* might be straightforward, and can be implemented with standard probability sampling techniques. The sampling of some other microenvironments such as *in transient* might be fairly difficult, and remains to be studied.

We can combine the conditional microenvironment concentration distribution estimated from microenvironment monitoring with the activity pattern distribution estimated from the activity survey, then use (6) to estimate the exposure distribution.

When applying the indirect approach to estimate the exposure distribution, it is usually necessary to impose some simplifying assumptions on the conditional microenvironment concentration distribution $F_{\mathbf{C}|\mathbf{T}}$. In other words, we might need to use *models* to implement the indirect approach. We call such models the *stochastic microenvironment models*, and discuss several such models below.

There are two major stochastic microenvironment models for the indirect approach: the Cartesianization method (Duan, 1980, 1982, 1987) and SHAPE (Ott, 1981, 1982, 1984).

The Cartesianization method, also known as the convolution method, assumes **C** is stochastically independent of **T**. We will review this model in the next section. SHAPE decomposes **C** into short term averages such as minute averages, and assumes the minute averages are stochastically independent of **T**. We will review this model in the section "SHAPE." We propose a new model, the variance components model, which includes both the Cartesianization method and SHAPE as special cases. The theory is then applied to the data collected from U.S. EPA's Washington CO Study.

CARTESIANIZATION

Duan (1980, 1982) developed the *convolution method* which can be used to implement the indirect approach. The method was applied to the Washington CO Study in Duan (1985). It was generalized to a broader context in Duan (1987) and renamed as the *Cartesianization method.* The essence of the method is to take the cross product (the Cartesian product) between the microenvironment concentration data and the activity pattern data.

The Cartesianization method assumes that the microenvironment concentrations are stochastically independent of activity patterns. This assumption is equivalent to

$$F_{\mathbf{C}|\mathbf{T}}(\mathbf{c} \mid \mathbf{t}) \equiv F_{\mathbf{C}}(\mathbf{c}), \tag{8}$$

where $F_{\mathbf{C}}$ denotes the marginal distribution of **C**, and will be called the *microenvironment concentration distribution.* Under (8), the joint distribution for **C** and **T** is given by multiplying the activity pattern distribution and the microenvironment concentration distribution. It follows that we can estimate the two distributions separately, then multiply the two estimated distributions to estimate the joint distribution.

It follows from (6) and (8) that the exposure distribution is given as follows:

$$F_Y(y) = \iint 1(\mathbf{tc} \leq y)\, dF_C(\mathbf{c}) dF_T(\mathbf{t}), \tag{9}$$

Given an estimate for F_C and an estimate for F_T, we can use (9) to combine the two estimated distributions, so as to estimate the exposure distribution. We can apply this method to three types of situations.

For the first type of application, we have an enhanced personal monitoring study in which we observe **C** and **T** for a random sample of n individuals from the target population. We estimate F_T by the empirical distribution (7), and estimate F_C by the empirical distribution

$$\hat{F}_C(\mathbf{c}) = n^{-1} \sum_{j=1}^{n} 1(\mathbf{C}_j \leq \mathbf{c}). \tag{10}$$

We then substitute the empirical distributions (7) and (10) into (9), and estimate the exposure distribution by

$$\hat{F}_Y(y) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} 1(\mathbf{T}_i \mathbf{C}_j \leq \mathrm{y}). \tag{11}$$

The estimated exposure distribution (11) can be interpreted as follows. We observe matched data $\{(\mathbf{T}_i, \mathbf{C}_i), i = 1,...,n\}$ from the enhanced personal monitoring study, where each datum **T** is matched with a corresponding datum **C** from the same individual. Under the independence assumption (8), the observed matching is irrelevant: it is equally likely for $\mathbf{T}_1$ to occur with $\mathbf{C}_2$ as it is to occur with $\mathbf{C}_1$. Therefore we neglect the observed matching, and construct a new data set which consists of all possible matches between **T** and **C**: $\{(\mathbf{T}_i, \mathbf{C}_j), i = 1,...,n, j = 1,...,n\}$. The new data set is the cross product (the Cartesian product) of the observed activity pattern data, $\{\mathbf{T}_1,...,\mathbf{T}_n\}$, and the observed microenvironment concentration data, $\{\mathbf{C}_1,...,\mathbf{C}_n\}$. This is the motivation for the terminology *Cartesianization.*

For each matched pair $(\mathbf{T}_i, \mathbf{C}_j)$, the corresponding integrated exposure is given by

$$Y_{ij} = \mathbf{T}_i \mathbf{C}_j = \sum_{k=1}^{K} T_{ik} C_{jk}, \tag{12}$$

We can interprete Y_{ij} as the integrated exposure that would occur if a hypothetical individual had the i-th individual's activity pattern and the j-th individual's microenvironment concentrations. The estimated exposure distribution (11) is the empirical distribution for the Y_{ij}'s.

Since we do observe the integrated exposures $\{Y_1,..., Y_n\}$ in an enhanced personal monitoring study, we can also use the direct approach and estimate the exposure distribution by the empirical distribution (3) of the observed integrated exposures. We can rewrite the direct approach estimate (3) as follows:

$$\hat{F}_Y(y) = n^{-1} \sum_{i=1}^{n} 1(\mathbf{T}_i\mathbf{C}_i \leq y). \tag{3'}$$

It can be seen from (11) and (3′) that the difference between the direct approach and the Cartesianization method for an enhanced personal monitoring study lies in whether we re-match the observed **C**'s and **T**'s. Duan (1987) showed that the Cartesianization method estimate (11) is more precise than the direct approach estimate (3) when the independence assumption (8) is valid, and gave methods for estimating the amount of improvement in precision when we use the Cartesianization method estimate instead of the direct approach estimate. The improvement can be substantial.

The second type of application is the re-use of monitoring data discussed in the Introduction. We have an activity survey on a random sample from the target population, from which we observe activity patterns $\{\mathbf{T}_1,\ldots,\mathbf{T}_n\}$. We estimate the activity pattern distribution for this target population by (7). We do not have personal monitoring data on this sample. Instead, we have an enhanced personal monitoring study for another population, from which we observe microenvironment concentrations $\{\mathbf{C}_1,\ldots,\mathbf{C}_m\}$. We assume the microenvironment concentration distributions in the two populations are the same, therefore we estimate the microenvironment concentration distribution in the new target population by the empirical distribution for **C**'s from the previous study:

$$\hat{F}_{\mathrm{C}}(\mathbf{c}) = m^{-1} \sum_{j=1}^{m} 1(\mathbf{C}_j \leq \mathbf{c}). \tag{10'}$$

Under the independence assumption (8), we combine (7) and (10′) to estimate the exposure distribution F_Y for the new target population:

$$\hat{F}_Y(y) = (nm)^{-1} \sum_{i=1}^{n}\sum_{j=1}^{m} 1(\mathbf{T}_i\mathbf{C}_j \leq \mathrm{y}). \tag{11'}$$

The estimate (11′) is similar to (11), and can be interpreted similarly.

The third type of application is the microenvironment monitoring study discussed in the Introduction. We assume that we can estimate the microenvironment concentration distribution F_{C} from the microenvironment monitoring data. We can then combine the estimated microenvironment concentration distribution with the activity pattern distribution estimated from the activity survey, then use (9) to estimate the exposure distribution.

For all three types of applications, the Cartesianization method requires the independence assumption (8). This assumption has three major implications. First, it implies that the activity patterns and the microenvironment concentrations are uncorrelated:

$$Cov(T_{ik}, C_{il}) = 0,\ k = 1,...,K,\ l = 1,\ ...,K. \tag{13}$$

Duan (1985) examined this property for the Washington CO Study, and found the correlations to be all small and insignificant. Second, (8) implies that the *mean functions* are constant:

$$E(C_{ik} \mid \mathbf{T}_i) \equiv \mu_k,\ k = 1,...,\ K, \tag{14}$$

where μ_k is the expectation for the microenvironment concentration in the k-th microenvironment. Both (13) and (14) are common to all three stochastic microenvironment models considered in this paper.

Another implication of (8) is that the *variance functions* $Var(C_{ik} | \mathbf{T}_i)$ are constant:

$$Var(C_{ik} \mid \mathbf{T}_i) \equiv \sigma_k^2,\ k = 1,...,\ K, \tag{15}$$

where σ_k^2 is the variance of the microenvironment concentration in the k-th microenvironment. This property does not hold for the other two microenvironment models to be discussed below.

The property (15) might appear counter-intuitive: the microenvironment concentration C_{ik} is the average of the exposure profile over the time spent in the k-th microenvironment, therefore we might expect the variance of C_{ik} would decrease with the averaging time T_{ik}. We will examine the variance function for the Washington CO Study in a later section.

SHAPE

An alternative stochastic microenvironment model, the *Simulation of Human Activity and Pollution Exposure* (SHAPE), was developed in Ott (1981, 1982, 1984). Its validation was studied in Ott, Thomas, Mage, and Wallace (1988) and Ott, Thomas, Wallace, and Hunt (1988).

SHAPE decomposes microenvironment concentrations into short term averages, say, minute averages:

$$C_{ik} = \sum_{s=1}^{T_{ik}} b_{ik}(s) \ / \ T_{ik}, \tag{16}$$

where $b_{ik}(s)$ denotes the average pollutant concentration during the s-th minute spent in the k-th microenvironment by the i-th individual. (We assume from now on that the activity patterns are measured in units of minutes.) It follows from (16) that we can rewrite the microenvironment decomposition (4) as follows:

$$Y_i = \sum_{k=1}^{K} \sum_{s=1}^{T_{ik}} b_{ik}(s). \tag{4'}$$

SHAPE assumes that the minute averages for the same microenvironment have the same distribution, and all minute averages are stochastically independent:

$$b_{ik}(s) \sim G_k(b_k),\ s = 1, ..., T_{ik},\ k = 1, ..., K \tag{17}$$

where G_k denotes the distribution for minute averages in the k-th microenvironment. We will call G_k the *minute average distribution* for the k-th microenvironment.

Given estimated minute average distributions and an estimate for the activity pattern distribution, SHAPE estimates the exposure distribution by a simulation based on the microenvironment decomposition (4′). For each replicate of the simulation, we generate the activity pattern for a hypothetical individual in the target population, for example, by sampling from the estimated activity pattern distribution. We then generate independent random samples for the minute averages. For the k-th microenvironment, we generate T_{ik} minute averages from the corresponding minute average distribution G_k. The microenvironment decomposition (4′) is then used to determine the integrated exposure for this hypothetical individual: we sum over all the minute averages. The procedure is replicated N times to provide the integrated exposures $\{Y_1,...,Y_N\}$ for a hypothetical sample from the target population. The exposure distribution is then estimated by the empirical distribution of the simulated integrated exposures.

SHAPE requires that we have valid estimates for the minute average distributions. This is straightforward if we have an enhanced personal monitoring study in which we observe each individual's exposure profile and his activity pattern. Alternatively, we might conduct a microenvironment monitoring study to obtain minute averages for the relevant microenvironments, from which we can estimate the minute average distributions. SHAPE can be applied to the same three types of situations considered earlier for the Cartesianization method, so long as we can estimate the minute average distributions from the available monitoring data.

The key assumption in SHAPE is that all minute averages are stochastically independent. This assumption leads to the same properties (13) and (14) that were discussed earlier for the Cartesianization method: the microenvironment concentrations and the activity patterns are uncorrelated, the mean functions are constant. On the other hand, SHAPE does lead to a different property for the variance function. Since SHAPE assumes the microenvironment concentration C_{ik} is the average of T_{ik} independent minute averages, the variance of C_{ik} is inversely proportional to the averaging time T_{ik}:

$$Var(C_{ik} \mid \mathbf{T}_i) = \tau_k^2 \ / \ T_{ik}, \ k = 1, ..., K, \qquad (15')$$

where τ_k^2 is the variance for the minute averages in the k-th microenvironment. We will examine the variance function for the Washington CO Study in a later section.

The assumption that the minute averages are stochastically independent over time might be unrealistic, because the adjacent minute averages are likely to be correlated. Switzer (1988) derived methods which can be used when the minute averages in the same microenvironment follow a time series model with autocorrelation. The autocorrelation model also satisfies (13) and (14). Furthermore, the variance function $Var(C_{ik} | \mathbf{T}_i)$ also behaves like (15′): it decreases to zero as T_{ik} becomes large, although it might decrease slower than (15′) because of the autocorrelation.

VARIANCE COMPONENTS MODEL

The Cartesianization method and SHAPE lead to two different variance functions, (15) and (15′), because they make different independence assumptions. Both variance functions might be unrealistic. Because of averaging over time, it is unlikely to have a constant variance function (15). On the other hand, the variance function in (15′) might not allow for heterogeneity of microenvironments, as will be discussed below. We propose a more general model in this section which includes both Cartesianization and SHAPE as special cases.

For an illustration, we consider the microenvironment *home* for two individuals in a target population. The CO concentration in each of their homes might be substantially different at any minute. Assume, however, that we monitor their homes for a very long time period, say, several years, so that the temporal variation would be essentially eliminated. The microenvironment concentrations in those two homes are still likely to be somewhat different: one of the two homes might use gas for heating and cooking, the other might be all electrical; one of the two individuals might be a smoker, the other home might be free from smokers. If the monitoring period is long enough, the microenvironment concentration in those two homes will converge to their long-term averages. However, because the two homes and the two individuals might be different, the two long term averages are likely to be different.

The above scenario suggests that the variance function $Var(C_{ik} | \mathbf{T}_i)$ might decrease as the averaging time T_{ik} becomes longer, although it might never approach zero as (15′) indicates. This suggests that we should consider a variance components model, in which we decompose the minute averages into two components,

$$b_{ik}(s) = a_{ik} + d_{ik}(s), \qquad (18)$$

where a_{ik} denotes the *long-term average* which varies from individual to individual, but does not vary over time, and $d_{ik}(s)$ denotes the *short-term fluctuation* which varies both over individuals and over time. We assume without loss of generality that the mean of the short-term fluctuation is zero:

$$E(d_{ik}(s)) = 0. \tag{19}$$

It follows from (16) and (18) that the microenvironment concentration is given by

$$C_{ik} = a_{ik} + \sum_{s=1}^{T_{ik}} d_{ik}(s) \;/\; T_{ik}. \tag{20}$$

Furthermore, the microenvironment decomposition (4) is given by

$$Y_i = \sum_{k=1}^{K} a_{ik} T_{ik} + \sum_{k=1}^{K} \sum_{s=1}^{T_{ik}} d_{ik}(s). \tag{4″}$$

We assume the long-term averages and the short-term fluctuations are all stochastically independent,[4] and follow the probability distributions

$$a_{ik} \sim F_k(a_k),\ k = 1, ..., K, \tag{21}$$

$$d_{ik}(s) \sim H_k(d_k),\ s = 1, ..., T_{ik},\ K = 1, ..., K. \tag{22}$$

We will call F_k the *long-term average distribution for the* k-th microenvironment, and H_k the *short-term fluctuation distribution* for the k-th microenvironment.

Given estimated long-term average distributions, estimated short-term fluctuation distributions, and an estimate for the activity pattern distribution, we can use a simulation similar to SHAPE to estimate the exposure distribution under the variance components model. For each hypothetical individual from the target population, we generate the activity pattern according to the estimated activity pattern distribution. We then generate the long-term averages from the estimated long-term average distributions, and generate the short-term fluctuations from the estimated short-term fluctuation distributions; for the k-th microenvironment, we generate T_{ik} random samples of $d_{ik}(s)$. We then use the microenvironment decomposition (4″) to determine the integrated exposure for this hypothetical individual. The procedure is replicated N times to provide the integrated exposures $\{Y_1, ..., Y_N\}$ for a hypothetical sample from the target population. The exposure distribution is then estimated by the empirical distribution of the simulated integrated exposures.

The variance components model can be applied to the same types of situations considered earlier for the Cartesianization method and SHAPE, assuming that

[4] We can relax the assumption that the short-term fluctuations in the same microenvironment are stochastically independent over time, and allow for autocorrelation in a way similar to Switzer (1988).

we can estimate both the long-term average distributions and the short-term fluctuation distributions. The estimation of the long-term average distributions and the short-term fluctuation distributions will be studied in a future paper.

The Cartesianization method and SHAPE are both special cases of the variance components model. If the short-term fluctuations are negligible, i.e., $d_{ik}(s) \approx 0$, we have

$$C_{ik} \approx a_{ik},$$

therefore the variance components model simplifies to the Cartesianization method. If the long-term averages are negligible, the variance components model simplifies to SHAPE. If neither component is negligible, the variance components model is different from either the Cartesianization method or SHAPE.

Under the variance components model, the variance function $Var(C_{ik} \mid \mathbf{T}_i)$ is given by

$$Var(C_{ik} \mid \mathbf{T}_i) = \sigma_k^2 + \tau_k^2 \; / \; T_{ik}, \qquad (15'')$$

where σ_k^2 is the variance of the long-term average a_{ik}, and τ_k^2 is the variance of the short-term fluctuation $d_{ik}(s)$. It follows that the variance function decreases as the averaging time T_{ik} becomes large, although its limit is σ_k^2 instead of zero.

Comparing (15), (15′), and (15″), we see that the choice among the three stochastic microenvironment models can be made by examining the variance function. If the variance function is nearly constant, then the short-term fluctuation is negligible, and we can use the Cartesianization method. If the variance function decreases to zero, then the long-term average is negligible, and we can use SHAPE. If the variance function decreases but flattens out at a positive value, then neither component is negligible, and we need to use the variance components model. In the next section we examine the variance function for the Washington CO Study.

WASHINGTON CO STUDY

As was discussed earlier, we can examine the variance function $Var(C_{ik} | \mathbf{T}_i)$ to choose among the three stochastic microenvironment models. We now use the empirical data from the Washington CO Study to illustrate this application. This is an enhanced personal monitoring study in which we observe both **C** and **T** for the same individuals.

The data were collected from a random sample of 705 residents of the Washington, D.C., metropolitan area. Each participant was monitored for approximately twenty-four hours, and each filled out a diary. The study was conducted during the winter of 1982-83. Further details on the study can be found in Akland, et al. (1985), Duan (1985), and Duan, Sauls, and Holland (1985).

TABLE 1
Summary Statistics for *C and T*

	Mean	SD	Min	Max
C	4.45	3.97	0.05	40.68
T	130.23	110.94	8.0	952.0

C: CO concentration in the microenvironment *in transit*, unit = ppm
T: time spent in the microenvironment *in transit*, unit = minute

The major source of CO exposure comes from the exposure to automobile exhaust. We therefore focus this analysis on the microenvironment *in transit*. We use the observed activity diaries to determine when an individual was in transit: driving an automobile, riding a bicycle, walking on the street, etc. We then combine the activity diary data with the personal monitoring data to determine C_{i1} and T_{i1}. (We have labeled the microenvironment *in transit* as the first microenvironment: $k = 1$.)

We have 662 individuals who spent some time in the microenvironment *in transit*, therefore we restrict this analysis to those individuals. Table 1 gives the summary statistics for the observed C_{i1}'s and T_{i1}'s. The microenvironment concentrations range from 0.05 ppm (below detection limit) to 40.68 ppm; the activity times range from 8 minutes to almost 16 hours.

We assume the variance components model, and want to determine whether the model can be simplified to either the Cartesianization method or SHAPE. As was discussed above, we can examine the variance function (15″) to choose among the three models.

In order to estimate the variance function (15″), we define

$$V_i = (C_{i1} - \bar{C}_{\cdot 1})^2,\ i = 1, \ldots, n, \tag{23}$$

where $\bar{C}_{\cdot 1} = n^{-1} \Sigma_{i=1}^{n} C_{i1} = 4.45$ ppm is the average CO concentration in the microenvironment *in transit*. Under (14), V_i has the following expectation:

$$E(V_i) = (1 - \frac{2}{n})\ Var(C_{i1} \mid \mathbf{T}_i) + Var(\bar{C}_{\cdot 1}). \tag{24}$$

Since the sample size $n = 662$ is fairly large, the right hand side of (23) is in essence the variance function $Var(C_{i1} \mid \mathbf{T}_i) = \sigma_1^2 + \tau_1^2 / T_{i1}$. We therefore have a regression model

$$V_i = \sigma_1^2 + \tau_1^2 / T_{i1} + \epsilon_i, \quad E(\epsilon_i) = 0, \tag{25}$$

where the residual ϵ_i denotes the difference between V_i and its expectation. It follows that we can estimate the variances σ_1^2 and τ_1^2 using a regression of V on $1/T$. For the ease of notation, we define

$$R_i = 1 / T_{i1}.$$

We plotted V_i against R_i in Figure 1. Since V_i is fairly skewed, it is difficult to visualize any relationship between V and R. In order to enhance the interpretation for Figure 1, we partitioned the individuals into four groups according to their activity times, each group having approximately the same number of individuals. We then compared their microenvironment concentrations. The results are given in Table 2. For each group, the first row gives the sample size; the second and third rows give the range of activity times; the fourth row gives the average of R; the fifth row gives the average of V. It appears from the table that V increases in R, which is consistent with model (25). However, the relationship is statistically insignificant. We carried out an F test on the difference in V across the four groups; the P-value was 0.57, indicating there is no evidence that the four groups are different.

We also fitted a least squares regression of V on R based on model (25); the results are given in Table 3. The intercept (15.87) is our estimate for σ_1^2; the slope coefficient (−7.53) is our estimate for τ_1^2. The intercept is significantly different from zero, indicating that we cannot neglect the long-term averages. The slope coefficient is insignificant. However, the standard error for the slope coefficient is very large, therefore the data does not rule out the possibility that the short-term fluctuations might not be negligible.

Since V is fairly skewed, the validity of the inference in the above analysis is questionable. Furthermore, the estimates might not be efficient and can be improved upon by taking a suitable transformation on V. Since we have a theoretical model (25) for the relationship between V and R, we apply Carrol and Ruppert's (1984) transform-both-sides model. We choose the logarithmic transformation because it approximately symmetrizes the distribution of V.[5] Our transform-both-sides model is given as follows:

$$W_i = \log(V_i) = \log(\alpha + \beta \ / \ T_{i1}) + \eta_i, \ E(\eta_i) = 0. \tag{26}$$

where α and β are analogous to α_1^2 and τ_1^2 in model (26). In order to see the relationship between the two models, observe that

$$V_i = (\alpha + \beta \ / \ T_{i1}) \cdot \exp(\eta_i),$$

$$E(V_i) = (\alpha + \beta \ / \ T_{i1}) \cdot \phi,$$

where $\phi = E(\exp(\eta))$ is the *retransformation correction.*[6] It follows that

$$\sigma_1^2 = \alpha \cdot \phi, \ \tau_1^2 = \beta \cdot \phi. \tag{27}$$

[5] We can improve upon the symmetry by using the eighth root transformation instead; the analyses based on the two transformations lead to very similar results. The logarithmic transformation has two advantages. First, it is easier to interpret. Second, it simplifies the retransformation problem.

[6] We have assumed implicitly that the residuals η_i in model (26) have a common distribution. This appears to be reasonable, as seen from the residual plot in Figure 3. Duan (1983) gave further discussions on the retransformation problem.

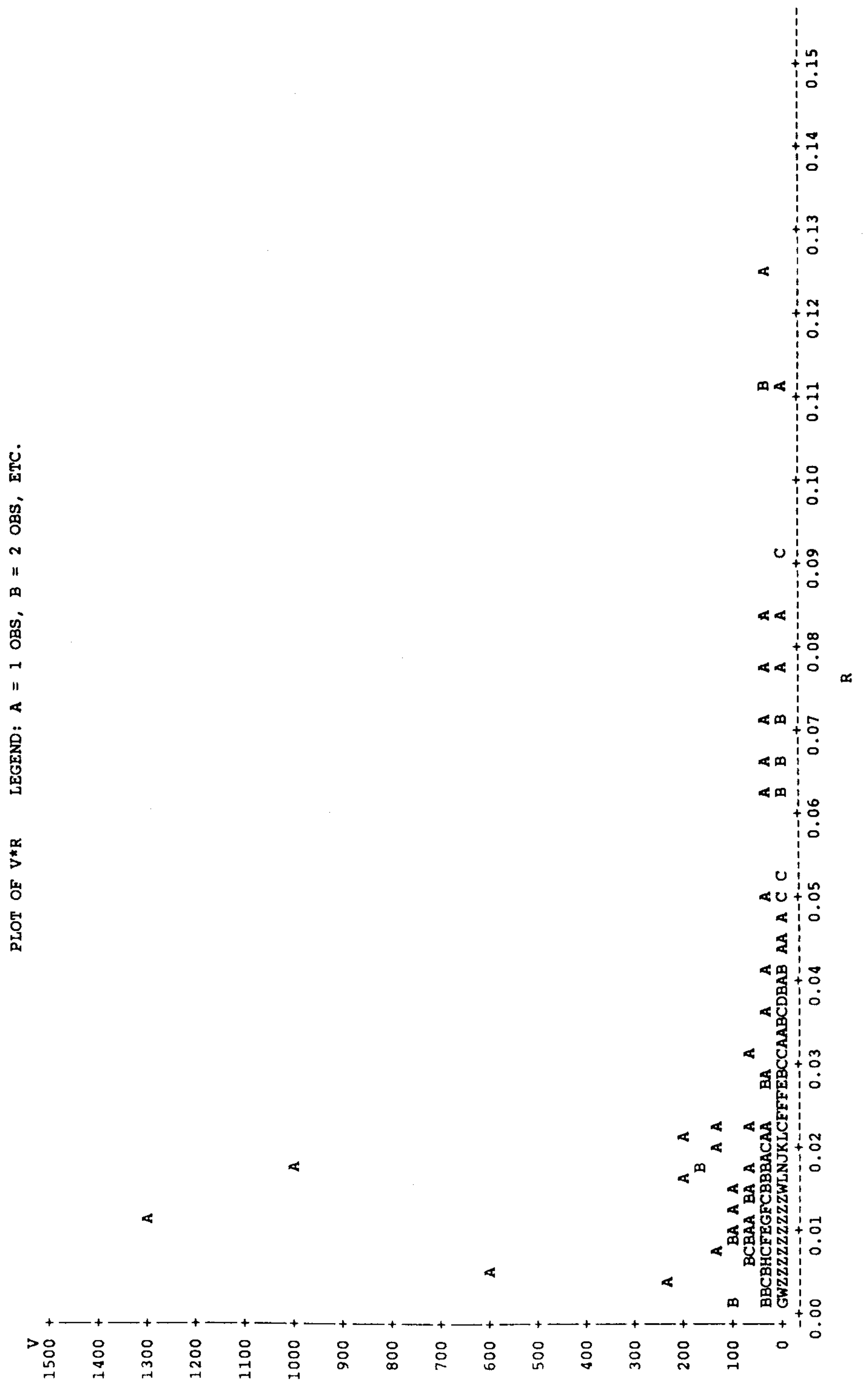

FIGURE 1. Scatterdiagram of V against R = 1/T.

TABLE 2
By-group Analyses Results

Variable	Group 1	Group 2	Group 3	Group 4
n	164	168	161	169
T_{min}	8	65	100	160
T_{max}	64	99	159	952
$\bar{R}$	0.0322	0.0125	0.0080	0.0043
$\bar{V}$	20.79	18.09	10.96	13.13
$\bar{W}$	1.492	0.996	0.987	0.900
$\hat{W}$	1.479	1.087	0.995	0.912

n: number of individuals
T_{min}: lowest T
T_{max}: largest T
$\bar{R}$: average of $R = 1/T$
$\bar{V}$: average of V
$\bar{W}$: average of $W = \log(V)$
W: average of $\hat{W}$ based on model (5.4)

We can estimate α and β using a nonlinear regression of $W = \log(V)$ on $R = 1/T$. We can estimate ϕ using the smearing estimate in Duan (1983). We can then combine those estimates to estimate σ_1^2 and τ_1^2 using (27).

We plotted W against R in Figure 2. It is difficult to visualize any obvious relationship between W and R. In order to enhance the interpretation of Figure 2, we carried out the same by-group analysis on W; for each group, the sixth row in Table 2 gives the average of W. It appears that W is increasing in R, which is consistent with model (26). We carried out an F test on the difference in W across the four groups; the P-value was 0.07. Although the F test is not significant at the conventional 5% level, it is close.

We fitted a nonlinear least squares regression of W on R based on model (26); the results are given in Table 4. The estimated intercept (A = 2.24) is significantly different from zero, indicating that we cannot neglect the long-term averages. The estimated slope (B = 58.27) is also significantly different from zero at the conventional 5% level, indicating that we cannot neglect the short-term fluctuations either.

The data appears to fit model (26) reasonably well. The seventh row in Table 2 gives the average predicted values for W for each group of individuals; they were very close to the corresponding values in the sixth row. We also plotted the fitted residuals against the predicted values in Figure 3. In order to avoid the illusion of heteroscedasticity due to the uneven density on the horizontal axis, we transformed the predicted values into their ranks. There does not appear to be any obvious anomalies in this scatterdiagram. We tested for the presence of nonlinearity using the by-group analysis; the P-value was 0.86. We also tested

TABLE 3
Least Squares Regression of V on R = 1/T
General Linear Models Procedure

Dependent Variable: V

Source	DF	Sum of Squares	Mean Square	F Value	PR > F	R-Square	C.V.
Model	1	8.75871941	8.75871941	0.00	0.9670	0.000003	453.2036
Error	660	3366433.14099422	5100.65627423				
Corrected					Root MSE		V Mean
Total	661	3366441.89971363			71.41887898		15.75867554

Source	DF	Type I SS	F Value	PR > F	DF	Type III SS	F Value	PR < F
R	1	8.75871941	0.00	0.9670	1	8.75871941	0.00	0.9670

		T for H0:	PR > |T|	STD Error of
Parameter	Estimate	Parameter = 0		Estimate
Intercept	15.86553286	4.19	0.0001	3.78873275
R	−7.52846122	−0.04	0.9670	181.67657745

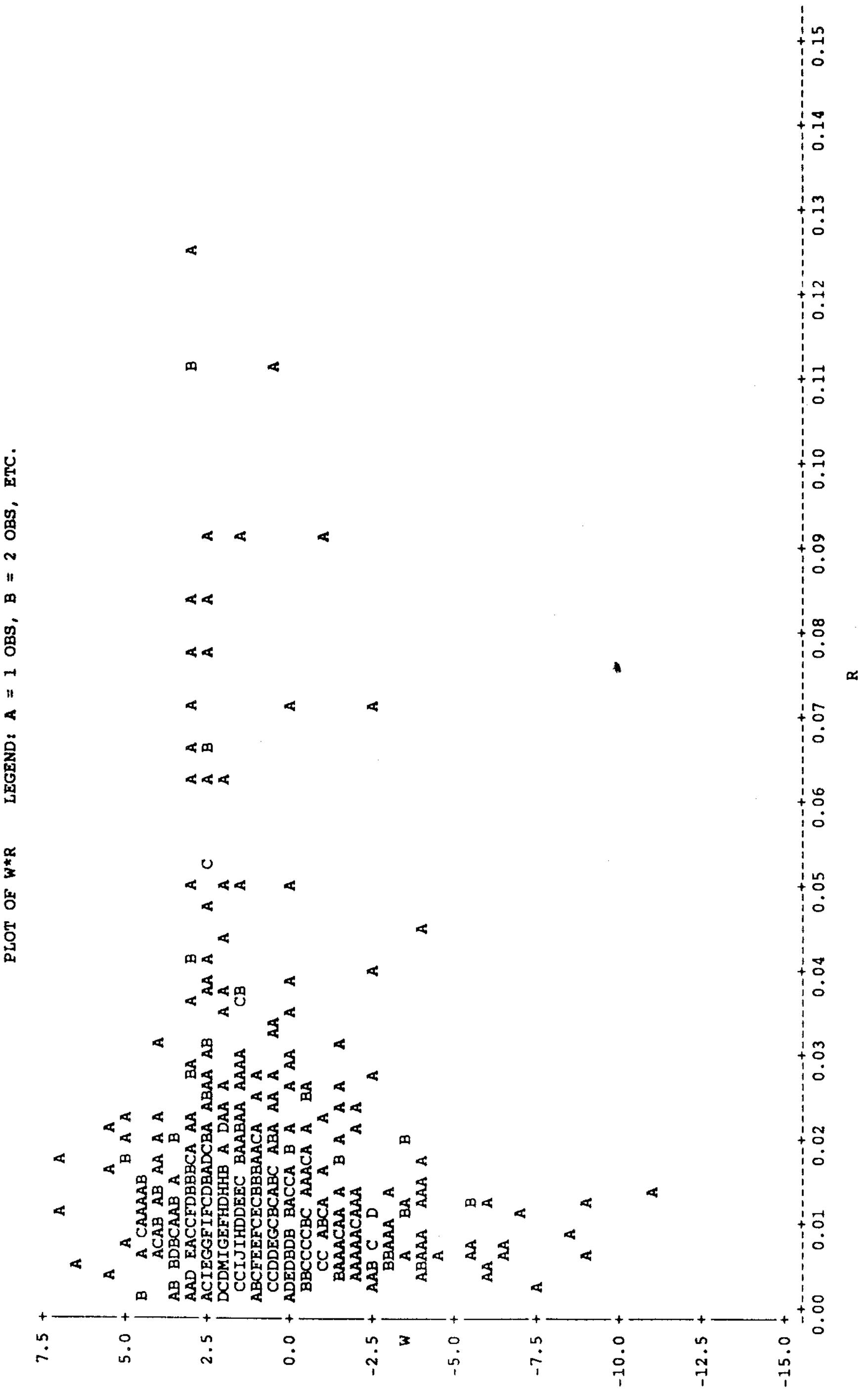

FIGURE 2. Scatterdiagram of W = log(V) against R = 1/T.

TABLE 4
Nonlinear Regression of W = log(V) on R = 1/T
A = Intercept, B = Slope

Non-Linear Least Squares Summary Statistics			Dependent Variable W
Source	DF	Sum of Squares	Mean Square
Regression	2	821.3635044	410.6817522
Residual	660	3264.5364056	4.9462673
Uncorrected Total	662	4085.8999100	
(Corrected Total)	661	3295.9798961	

Parameter	Estimate	Asymptotic Std. Error	Asymptotic 95% Confidence Interval	
			Lower	Upper
A	2.23931654	0.387305719	1.4788043456	2.99982874
B	58.27135457	29.087978064	1.1542995185	115.38840962

Asymptotic Correlation Matric of the Parameters

Corr	A	B
A	1.0000	−0.7665
B	−0.7665	1.0000

for heteroscedasticity, which gave a P-value of 0.18. Neither P-values are close to being significant.

As was noted earlier in (27), we need to correct for retransformation bias in order to relate the estimated intercept and slope from model (26) to σ_1^2 and τ_1^2. We estimate the retransformation correction ϕ with the smearing estimate in Duan (1983), which for model (26) is given by

$$\hat{\phi} = n^{-1} \sum_{i=1}^{n} \exp(\hat{\eta}_i),$$

where $\hat{\eta}_i = W_i - \log(\hat{\alpha} + \hat{\beta} \ / \ T_{i1})$ is the fitted residual for the *i*-th individual. The estimates are given as follows:

$$\hat{\phi} = 5.35, \ \hat{\sigma}_1^2 = 11.97, \ \hat{\tau}_1^2 = 311.52.$$

Since both σ_1^2 and τ_1^2 are significantly different from zero, we cannot use either the Cartesianization method or SHAPE, and need to use the variance components model and consider both the long-term averages and the short-term fluctuations.

Based on the above results, the variance function (15″) is estimated to be

$$Var(C_{i1} \mid \mathbf{T}_i) = 11.97 + 311.52/T_{i1}.$$

If the activity time is short, the short-term fluctuation dominates the long-term average: the former makes a larger contribution to $Var(C_{i1} \mid \mathbf{T}_i)$. If the activity

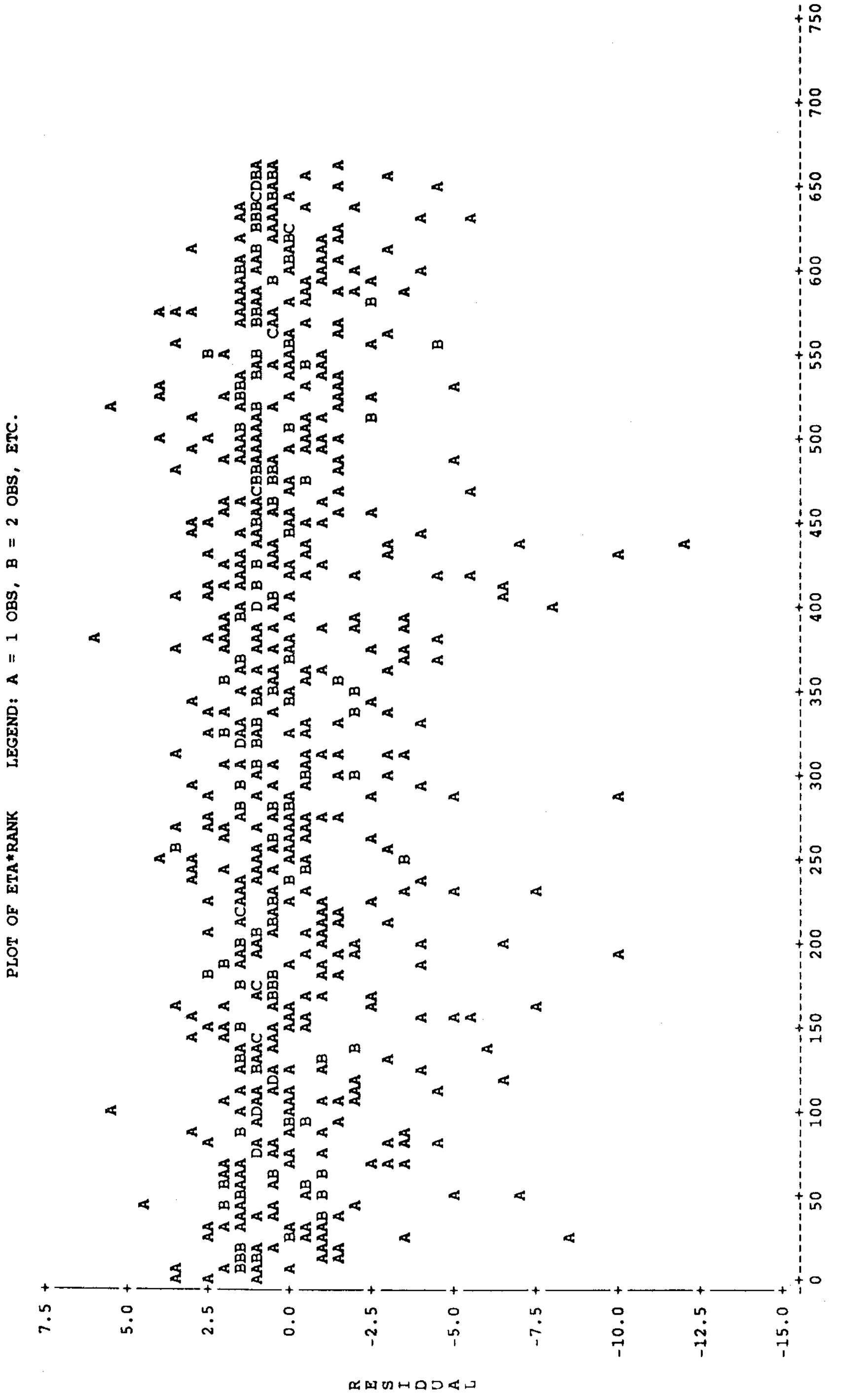

FIGURE 3. Scatterdiagram of residual against prediction.

time is 26 minutes (311.52/11.97), the two components are comparable: they make the same contribution to $Var(C_{i1} \mid \mathbf{T}_i)$. If the activity time is longer than 26 minutes, the long-term average dominates over the short-term fluctuation. Among the 662 individuals in our sample, only 36 spent 26 minutes or less in the microenvironment *in transit*. Therefore the long-term average dominates over the short-term fluctuation for the vast majority of our sample.

DISCUSSION

The stochastic microenvironment models are useful for implementing the indirect approach. The choice among the three models can be made by examining the variance function. If the variance of the microenvironment concentration does not depend on the time spent in the microenvironment, we can use the Cartesianization method. If the variance decreases with time and approaches zero, we can use SHAPE. If the variance decreases with time but does not approach zero, we need to use the variance components model and consider both the long-term averages and the short-term fluctuations.

The transform-both-sides model (26) is useful for estimating the variance function. For the microenvironment *in transit* in the Washington CO Study, we estimate the variance for the long-term average to be $\sigma_1^2 = 11.97$ ppm^2, and the variance for the short-term fluctuation to be $\tau_1^2 = 311.52$ ppm^2/minute. Both variances are significantly different from zero.

All results in this paper require some independence assumptions. This is intrinsic for the indirect approach. Whether those assumptions are realistic remains to be studied empirically more thoroughly. Duan (1985) examined the correlations among activity pattern and microenvironment concentrations for the Washington CO study, and found that all correlations were small and insignificant. Switzer (1988) examined the minute averages from a microenvironment monitoring study conducted on El Camino Real, an arterial route in Palo Alto, California, and found little autocorrelation beyond the first few minutes. More empirical studies of this type still need to be done.

The results in this paper can be extended in many directions. We now discuss one important generalization, the incorporation of covariate information into the variance components model. Assume that we observe covariates $\mathbf{x}_{ik}$ which describe the characteristics about the microenvironments, such as ventilation rate, type of fuel used in cooking and heating, etc. Those covariates can be used to predict part of the difference in the long-term averages. For example, we might use a linear regression model:

$$a_{ik} = \gamma_k + \delta_k \mathbf{x}_{ik} + e_{ik}, \tag{28}$$

where γ_k denote the average of the long-term averages for individuals with $\mathbf{x}_{ik}$

$= \mathbf{O}$, δ_k denotes the residual variation in the long-term averages among individuals with the same covariates. The minute average is then given by

$$b_{ik}(s) = \gamma_k + \delta_k \mathbf{x}_{ik} + e_{ik} + d_{ik}(s). \tag{29}$$

Given the appropriate data, we can estimate the parameters γ_k and δ_k, the distribution for e, and the short-term fluctuation distribution. We can then estimate the exposure distribution in two ways. First, we can restrict to the subpopulation of individuals with the same covariates; we simulate their e's and short-term fluctuations, then add the term $\gamma_k + \delta_k \mathbf{x}_{ik}$ to estimate their integrated exposures and their exposure distribution. Alternatively, if we want to estimate the exposure distribution for a target population consisting of individuals with different covariates, we need to conduct a survey to estimate the distribution of the covariates in the target population, then simulate the covariates along with the other stochastic elements.

We can also use model (28) to incorporate observed ambient conditions as part of the covariates $\mathbf{x}_{\mathrm{ik}}$. For some individuals observed on a windy day with low ambient CO concentration, we might have $\mathbf{x}_{ik}$ being (windy, 0.5 ppm). For some individuals observed on a windless day with high ambient CO concentration, we might have $\mathbf{x}_{ik}$ being (windless, 10.0 ppm). The long term average a_{ik} in (28) should be interpreted as the long term average for the microenvironment concentration under the same ambient condition.

Again, we can estimate the exposure distribution in two ways. First, we can restrict to the exposure distribution under a given ambient condition. We simulate the e's and short term fluctuations, then add the term $\gamma_k + \delta_k \mathbf{x}_{ik}$ corresponding to the given ambient condition. Alternatively, we can estimate the exposure distribution pooled across different ambient conditions. We need to collect data on the ambient conditions, for example, on 10% of the days $\mathbf{x}_{ik}$ is (windy, 0.5 ppm), on 5% of days $\mathbf{x}_{ik}$ is (windless, 10.0 ppm), etc. We simulate the exposure distribution for each ambient condition, then pool across different ambient conditions, by taking the weighted average of those exposure distributions, weighted by the prevalence for each ambient condition.

REFERENCES

AKLAND, G.G., HARTWELL, T.D., JOHNSON, T.R., and WHITMORE, R.W. (1985). "Measuring human exposure to carbon monoxide in Washington, D.C., and Denver, Colorado, during the winter of 1982-1983." *Env. Sci. Technol.* **19**:911–918.

CARROLL, R., and RUPPERT, D. (1984). "Power transformation when fitting theoretical models to data." *Journal of American Statistical Association.* **79**:321–328.

DUAN, N. (1980). "Micro-environment types: a model for human exposure to air pollution." SIMS Technical Report No. 47, Dept. of Statistics, Stanford University, Stanford, CA.

DUAN, N. (1982). "Models for human exposure to air pollution." *Environment International.* **8**:305–309.

DUAN, N. (1983). "Smearing estimate: a nonparametric regression method." *Journal of the American Statistical Association.* **78**:605–610.

DUAN, N. (1985). "Application of the microenvironment monitoring approach to assess human exposure to carbon monoxide." R-3222-EPA, RAND Corporation, Santa Monica, CA.

DUAN, N. (1987). " Cartesianized sample mean: imposing known independence structures on observed data." Manuscript, RAND Corporation, Santa Monica, CA.

DUAN, N. (1989). "Estimation of microenvironment concentration distribution using integrated exposure measurements." In: Proceedings of the Research Planning Conference on Human Activity Patterns, T. H. Starks, ed., EPA/600/4-89/004, Environmental Monitoring Systems Laboratory, U.S. EPA, Las Vegas, NV.

DUAN, N., DOBBS, A., and OTT, W. (1989). "Comprehensive definitions of exposure and dose to environmental pollution." In: Total Exposure Assessment Methodology: A New Horizon, 166–195, Air and Waste Management Association, Pittsburg, PA.

DUAN, N., SAULS, H., and HOLLAND, D. (1985). "Modeling exposures: activity patterns and microenvironments." *Proceedings of the Fifth SGOMSEC Workshop*, in print.

EFRON, B. (1991). "Regression percentiles using asymmetric squared error loss." Statistica Sinica. **1**:93–126.

KOENKER, R., and BASSETT, G. (1978). "Regression quantiles." *Econometrica.* **46**:33–50.

OTT, W. (1981). "Computer simulation of human exposures to carbon monoxide." Paper presented at the 74th Annual Meeting of the Air Pollution Control Association, Philadelphia, PA.

OTT, W. (1982). "Concepts of Human Exposure to Air Pollution." *Env. Int.* **7**:179–186.

OTT, W. (1984). "Exposure Estimates Based on Computer Generated Activity Patterns." *J. Toxicol. - Clin. Toxicol.* **21**:97–128.

OTT, W., THOMAS, J., MAGE, D., and WALLACE, L. (1988). "Validation of the simulation of human activity and pollutant exposure (SHAPE) model using paired days from the Denver, Colorado, carbon monoxide field study." *Atmospheric Environment.* **22**:2101–2113.

OTT, W., THOMAS, J., WALLACE, L., and HUNT, H. (1988). "Validation of the simulation of human activity and pollutant exposure (SHAPE) model using Washington, D.C. carbon monoxide field study." Paper presented at the Workshop on Modeling Commuter Exposure, Research Triangle Park, NC.

SWITZER, P. (1988). "Developing empirical concentration autocorrelation functions and averaging time models." Paper presented at the Workshop on Modeling Commuter Exposure, Research Triangle Park, NC.

received: 1/4/90
accepted: 6/1/90

RAND/N-3415-SIMS/RC