MICRO-ENVIRONMENT TYPES:
A MODEL FOR HUMAN EXPOSURE TO AIR POLLUTION

NAIHUA DUAN

TECHNICAL REPORT NO. 47
MAY 1981

PREPARED UNDER THE AUSPICES OF
SIAM INSTITUTE FOR MATHEMATICS AND SOCIETY

SIAMS

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
MICRO-ENVIRONMENT TYPES: A MODEL FOR HUMAN EXPOSURE TO AIR POLLUTION

NAIHUA DUAN

Department of Statistics
Stanford University

Economics Department
The Rand Corporation

TECHNICAL REPORT NO. 47
MAY 1981

STUDY ON STATISTICS AND ENVIRONMENTAL FACTORS IN HEALTH (SIMS)

PREPARED UNDER SUPPORT TO SIMS FROM

ENVIRONMENTAL PROTECTION AGENCY (EPA)
DEPARTMENT OF ENERGY (DOE)
SLOAN FOUNDATION
NATIONAL SCIENCE FOUNDATION (NSF)

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
MICRO-ENVIRONMENT TYPES: A MODEL FOR HUMAN EXPOSURE TO AIR POLLUTION*

Naihua Duan
Department of Statistics
Stanford University
Economics Department
The Rand Corporation

Abstract

The micro-environment type model is proposed as a model for human exposure to air pollution. The subject's exposure is decomposed into distinct components, each component corresponding to one distinct micro-environment type, such as "urban office buildings," "commuting on local streets," "outdoor away from traffic," etc. The total exposure is represented as a weighted sum of the pollutant concentrations measured in each micro-environment, weighted by the time the subject spent in each specific micro-environment. The concentration and time can be measured separately. Because the time measurements can be obtained at a low cost with diaries or surveys, it is preferable to take the time measurements over a larger population so as to make the maximal use of the more expensive concentration measurements.

The concentration and time measurements can be used to estimate their respective distributions. The two distributions can then be combined to estimate the distribution of the total exposure. Similar combinations can also be done to estimate summary parameters of the total exposure, such as the expectation and variance. It is shown in Section 7 that such an indirect approach will result in better precision than the direct approach of estimating the exposure distribution directly.

In Section 8 we discuss the merit of refining or merging micro-environment types. It is shown that refinement will generally lead to improvement in precision, with the amount of improvement determined by features of the time and concentration distributions. The amount of improvement should be considered together with the cost of maintaining extra micro-environment types to decide on the appropriateness of refinement or merging.

*Part of the work was done while the author served as research affiliate in the Department of Statistics, Stanford University, supported by a research grant from SIMS; part of the work was done while the author served as associate statistician in the Economics Department, The Rand Corporation, supported by a research grant from NIH.
1. **Fixed-Receptor Model**

Human subjects are usually treated as homogeneous receptors to air pollution in air pollution studies. Instead of measuring the true exposure of individual subjects, measurements of ambient pollutant concentration are taken at fixed monitoring stations, sometimes interpolated spatially to estimate the unmeasured ambient concentration at "receptor locations" (e.g., neighborhoods of residence). Whole segments of the general population (people residing in the same neighborhood) are treated as homogeneous receptors fixed at such locations in the ambient environment.

This representation is not quite satisfactory for the human population. The human subjects do not behave like fixed receptors. They carry on a variety of activities. They spend an appreciable amount of time in non-ambient environments. Their exposure levels can be very different from the levels assigned to their idealized counterparts fixed at the receptor location in the ambient environment.

2. **Micro-Environment Model**

The effect of activity on human exposure has received some consideration in the literature. For example, Fugas (1) calculated the "weighted average exposure, using time spent and average concentrations at various places where the population under observation is likely to circulate." The "various places" included the breathing zone in the streets during business hours, indoors close to the street, indoors away from the traffic, and the countryside. The pollutants under Fugas' study were sulfur dioxide and lead. The average concentration of the pollutants measured at the "various places" varied over a wide range. For several subpopulation groups, she estimated the hours per week spent in each of these "various places," and calculated their estimated weighted weekly exposure. Due to the variation in concentration and the variation in activities, the estimated exposure of the subpopulations vary across an appreciable range, though not as widely as the concentration measurements.

The essence of Fugas' work is to recognize the "various places" and decompose an individual's exposure into several components, each
corresponding to one of the "various places." For the clarity of language, we will denote the "various places" as the "micro-environment types." We will have more discussion on the micro-environment types in Section 3.

The total exposure can be expressed as follows:

$$E_{ij} = \sum_{k=1}^{K} C_{ijk} t_{ijk} \quad i = 1, \ldots, I; \quad j = 1, \ldots, J \quad (2.1)$$

In the above expression, "i" indexes the subjects, "j" indexes the time periods, say, weeks, and k indexes the micro-environment types. $E_{ij}$ denotes the (total) exposure of the ith subject during the jth time period.

The total exposure is decomposed into K components, each corresponding to one specific micro-environment type. The contribution of the kth micro-environment type is determined by the concentration and exposure time. In the expression (2.1), the concentration coefficient $C_{ijk}$ denotes the average concentration for the ith subject during the jth time period in the kth micro-environment type, while the time coefficient $t_{ijk}$ denotes the amount of time spent by the ith subject during the jth time period in the kth micro-environment type. The product $C_{ijk} t_{ijk}$, therefore, is the exposure of the ith subject during the jth time period coming from the kth micro-environment type, which makes up the kth component of the total exposure $E_{ij}$.

3. Micro-Environment Types

The advantage of considering micro-environment types is to separate the variability in concentration into differences among distinct micro-environment types. It will be shown in Section 7 that such separation gives a more precise estimate of the total exposure as compared to estimating the total exposure directly. The amount of improvement depends on the difference among the micro-environment types. If we separate into

---

*For ease of notation, we are assuming a balanced design with the same subjects observed over the same time periods. An actual study need not be structured this way. Due to the limitation on instrumentation, an unbalanced design using different groups of subjects over different time periods will allow more subjects to be measured, each with less repetitions.
two micro-environment types that are distinctly different (having large differences in the means and/or low correlations), the improvement will be large. If the two micro-environment types have similar means and are highly correlated, the improvement will be small.

The identification of micro-environment types is being worked on actively by researchers, e.g., Ott (2). In the opinion of the present author, an appropriate identification procedure should consist of two stages: a classification stage and an evaluation stage. For the classification stage, the researcher should identify a profile of relevant dimensions which would have appreciable effect on the pollutant concentration, e.g., outdoor-indoor, urban-suburban-rural, proximity to traffic, indoor ventilation, indoor source of pollutant, type of commuting, etc. Based on those dimensions, we can construct a preliminary list of micro-environment types. For example, some potentially interesting micro-environment types could be "indoor, urban, near traffic, air conditioned," such as urban office buildings, "commuting on local streets," "outdoor away from traffic," such as suburban recreation area, etc.

Such a classification might lead to a preliminary list of twenty or thirty micro-environment types. Some of those micro-environment types might behave very similar, and could be merged. Some of them might be too broad, thus should be refined. The second stage of the identification, the evaluation stage, is therefore to evaluate the merits of such merging and/or refinement, which will be discussed in more detail in Section 8.

4. Concentration and Time Distributions

The distribution of the total exposure—the total exposure distribution $F_E$, is determined by the distribution of the time coefficients—the time distribution $F_T$, and the distribution of the concentration coefficients—the concentration distribution $F_C$. For a sampling study, we can estimate the total exposure distribution
either directly—measure the total exposure \( E_{ij} \), or indirectly—
measure the time coefficients \( t_{ijk} \) and the concentration coefficients
\( C_{ijk} \). Estimate the corresponding distributions \( F_T \) and \( F_C \) and combine
the two estimated distributions to estimate the total exposure distribution
\( F_{E'} \).

One important feature of the indirect approach is that the time
coefficients can be measured at relatively low cost. The measurement
of either \( E_{ij} \) or \( C_{ijk} \) would require the use of costly measuring devices,
such as personal dosimeters. The time coefficients, on the other hand,
can be measured with pencil and paper, such as activity diaries or
questionnaires.

Because of the difference in measurement cost, we should measure
a larger sample of time coefficients. This can be implemented in either
of two ways. We can measure the time coefficients on a larger sample
of subjects, and measure the concentration coefficients on a smaller
subsample of subjects. Or, we can measure both sets of coefficients
on the same sample of subjects, using more repetitions for time
coefficient measurements than concentration coefficient measurements:
for any single subject, measure his time coefficients for a long sequence
of time periods, and measure his concentration coefficients for a few
time periods only.

More generally, we can describe the indirect measurements as
follows:

\[
(1) \quad t_{ijk}, \ i = 1, \ldots, I_T; \ j = 1, \ldots, J_T; \ k = 1, \ldots, K
\]

\[
(2) \quad C_{ijk}, \ i = 1, \ldots, I_C; \ j = 1, \ldots, J_C; \ k = 1, \ldots, K
\]

where \( I_C \leq I_T, J_C \leq J_T \).
From the measured time coefficients, we can derive an estimate of the time distribution \( F_T \); from the measured concentration coefficients, we can derive an estimate for the concentration distribution \( F_C \). The two estimated distributions, \( \hat{F}_T \) and \( \hat{F}_C \), are to be combined to yield an estimate of the total exposure distribution \( \hat{F}_E \). The estimated time coefficient distribution \( \hat{F}_T \) is based on a larger sample; the measurement error \( \hat{F}_T - F_T \), compared to the measurement error \( \hat{F}_C - F_C \) of the concentration distribution based on a smaller sample, can be regarded as negligible. Thus the measurement error in the total exposure distribution \( \hat{F}_E \) is dominated by the measurement error in the concentration distribution. This allows us to achieve better precision based upon the same amount of concentration measurements. A more detailed comparison between the direct and indirect approaches will be given in Section 6.

In addition to better precision, the indirect approach produces valuable information on the concentration distribution in various micro-environment types. Such information can be used to identify high-risk subpopulations, as well as devising control strategies.

5. **Distributional Assumptions**

Under the model (2.1), we adopt the following notations and assumptions:

1. The vector of concentration coefficients

   \[ C_{ij} = (c_{ij1}, \ldots, c_{ijK})^\top \]

   are independently and identically distributed according to the multivariate distribution \( F_C \) to be estimated. For most purposes, we will concentrate on the mean vector and covariance matrix of the multivariate distribution \( F_C \):

   \[
   E(C_{ij}) = \mu^C
   \]

   \[
   \text{Cov}(C_{ij}) = E((C_{ij} - \mu^C)(C_{ij} - \mu^C)^\top) = \Sigma^C
   \]

   *Depending on the purpose of our investigation, we might be interested in the distribution of individual exposure in a given population, either pooled over different time periods, (and thus pooled over different meteorological conditions), or conditioned on a fixed time period (and therefore specific to fixed meteorological conditions).

   The comment applies to later sections as well. We will derive the remaining parts of this paper using the unconditioned approach, averaging over subjects and time periods. For an alternative conditioned approach, our theory remains valid if we average over subjects only.
(2) The vector of time coefficients

\[ t_{ij} = (t_{ij1}, \ldots, t_{ijk})' \]

are distributed independently and identically according to the multivariate distribution \( F_T \), with mean and covariance matrix

\[ E(t_{ij}) = \mu_T \]
\[ \text{Cov}(t_{ij}) = \tau_T \]

(3) The concentration coefficients and the time coefficients are stochastically independent of each other.*

Given a sample of concentration coefficients \( C_{ijk} \), we can estimate the concentration distribution by the empirical distribution:

\[ \hat{F}_C(g) = \frac{\#_C(g)}{I_CJ_C} \]

where \( \#_C(g) \) = number of \( C_{ijk} \)’s for which \( C_{ij} \leq g \), i.e., \( C_{ij1} \leq g_1, \ldots, C_{ijk} \leq g_K \)

The mean vector and covariance matrix can be estimated by the sample mean vector and sample covariance matrix:

\[ \hat{\mu}_C = \frac{1}{I_CJ_C} \sum_{ij} C_{ij} \]
\[ \hat{\Sigma}_C = \frac{1}{I_CJ_C} \sum_{ij} C_{ijk} \]

*The assumption is that the human subjects do not allocate their time after looking at the concentration, which is justified to the extent that a lot of activities are rigid and cannot be reallocated very easily. (For example, the amount of time an office worker has to spend at work.) However, there are also possibilities like people escape to the beach on a smoggy day, which will likely violate this assumption. The overall validity of this assumption remains to be evaluated with real data.
\[ \hat{\gamma}^C = \frac{1}{1 - C^c C^{-1}} \sum_{ij} (c_{ij} - \hat{\mu}^C) (c_{ij} - \hat{\mu}^C) \]

\[ \hat{\gamma}_{k'}^C = \frac{1}{1 - C^c C^{-1}} \sum_{ijk} (c_{ijk} - \hat{\mu}_{k'}^C) (c_{ijk} - \hat{\mu}_{k'}^C) \]

The corresponding estimates for the time distribution can be constructed similarly.
6. Combination of the Concentration and the Time Distributions

Based on the independence assumption (3) in the previous section, we can estimate the joint distribution of the time and concentration coefficients

\[
F_{C,T}(c,t) = P(C \leq c, T \leq t) = P(C \leq c)P(T \leq t) = F_C(c)F_T(t)
\]

by the empirical estimate

\[
\hat{F}_{C,T}(c,t) = \frac{\hat{F}_C(c)\hat{F}_T(t)}{I_C^J_C \cdot I_T^J_T}
\]

(6.1)

The estimate (6.1) can be interpreted as follows. Given the measured concentration coefficients \( C_{ij}, i = 1, \ldots, I_C; j = 1, \ldots, J_C \), and measured time coefficients \( t_{ij}, i = 1, \ldots, I_T; j = 1, \ldots, J_T \), we can construct a total of \( I_C^J_C \cdot I_T^J_T \) time-concentration combinations

\[
(C^*, t^*)_{ij, i'-j'} = (C_{ij}, t_{i'-j'}) \quad i = 1, \ldots, I_C; j = 1, \ldots, J_C; \quad i' = 1, \ldots, I_T; j' = 1, \ldots, J_T
\]

Some of the combinations might have actually occurred together; some are artificial combinations. As the time and concentration coefficients...
are assumed to be independent, all such combinations are equally likely.

The empirical estimate (6.1) is actually the empirical c.d.f. of those combinations:

\[ \hat{F}_{C,T}(c,t) = \frac{\#_{C^*,T^*}(c,t)}{I_{C^*} J_{T^*}} \] (6.2)

where \( \#_{C^*,T^*}(c,t) \) = number of combinations for which \( C^* \leq c, \ T^* \leq t \).

The exposure distribution \( F_E \) can be derived from the joint distribution \( F_{C,T} \) as follows:

\[ F_E(e) = P(E \leq e) = P(c^* t \leq e) = \int I\{c^* t \leq e\} dF_{C,T}(c,t) \]

where \( I\{\cdot\} \) denotes the indicator function of the event "\( \cdot \)"

Therefore, we can estimate the exposure distribution \( F_E \) by

\[ \hat{F}_E(e) = \int I\{c^* t \leq e\} \hat{F}_{C,T}(c,t) \]

\[ = \frac{\#\{(c^*, t^*) : c^* t^* \leq e\}}{I_{C^*} J_{T^*}} \] (6.3)

From each combination \( (c^*, t^*) \), we can compute the combined exposure \( e^* \):

\[ e^*_{ij, \tilde{i}, \tilde{j}} = c^*_{ij} t^*_{i, j} \quad i^* = 1, \ldots, I_{C}; \ j^* = 1, \ldots, J_{T}; \]

\[ i^* = 1, \ldots, I_T; \ j^* = 1, \ldots, J_T \]
Again, some of the combined exposures \( e^* \)'s have actually been observed, some have come from artificial combinations. Because of the independence of time and concentration coefficients, all those combined exposures are equally likely. The estimate (6.3) can be interpreted as the empirical c.d.f. of the combined exposures \( e^* \):

\[
F_E(e) = \frac{\#_{E^*}(e)}{\int C_T J_T}
\]  
(6.4)

where \( \#_{E^*}(e) = \) number of \( e^* \)'s for which \( e^* \leq e \).

For many applications, the main feature of the exposure distribution \( F_E \) can be described by its expectation \( \mu(E) \) and variance \( \sigma^2(E) \). Those quantities are related to the corresponding quantities in the concentration and time coefficients distributions in a simple way.

Theorem 6.1

Assuming the independence between the concentration and time coefficients, we have

\[
\mu(E) = (\mu_C) \cdot \mu_T
\]

\[
\sigma^2(E) = (\mu_C) \cdot T_T \mu_C + (\mu_T) \cdot C_T \mu_T + \text{tr}(C_T T_T)
\]

**Proof:**

\[
\mu(E) = E(C \cdot t)
\]

\[
= (E(C)) \cdot (E(t))
\]

\[
= (\mu_C) \cdot (\mu_T)
\]

\[
\sigma^2(E) = \text{Var}(C \cdot t)
\]

\[
= E(C \cdot t \cdot t^T) - E^2(C \cdot t)
\]
\[ -12 - \]

\[
\begin{align*}
= E[E(C^\top t^\top C|C)] - [(\mu^C)^\top \mu^T]^2 \\
= E[C^\top E(t^\top C)] - [(\mu^C)^\top \mu^T]^2 \\
= E[C^\top [\mu^T (\mu^T)^\top + \frac{1}{T} T)]C] - [(\mu^C)^\top \mu^T]^2 \\
= E(C^\top \mu^T)^2 + EC^\top \mu^T - [(\mu^C)^\top \mu^T]^2 \\
= E(\mu^T)^\top CC^\top \mu^T + E tr(\frac{1}{T} \Sigma t^\top t) - [(\mu^C)^\top \mu^T]^2 \\
= (\mu^T)^\top [\mu^C (\mu^C)^\top + \frac{1}{T} \Sigma] \mu^T + tr(\frac{1}{T} \Sigma (\mu^C)^\top + \frac{1}{T} \Sigma) - [(\mu^C)^\top \mu^T]^2 \\
= [(\mu^C)^\top \mu^T]^2 + (\mu^T)^\top \Sigma \mu^T + tr(\frac{1}{T} \Sigma \mu^T (\mu^C)^\top - tr(\frac{1}{T} \Sigma \mu^T) - [(\mu^C)^\top \mu^T]^2 \\
= (\mu^T)^\top \Sigma \mu^T + (\mu^C)^\top \frac{1}{T} \Sigma \mu^T + tr(\frac{1}{T} \Sigma \mu^T) \\
\end{align*}
\]

Therefore, we can estimate \( \mu(E) \) and \( \sigma^2(E) \) by

\[
\hat{\mu}(E) = (\hat{\mu}^C)^\top (\hat{\mu}^T) \quad (6.5)
\]

\[
\hat{\sigma}^2(E) = (\hat{\mu}^C)^\top \Sigma^\top \hat{\mu}^C + (\hat{\mu}^T)^\top \Sigma^\top \hat{\mu}^T + tr(\frac{1}{T} \Sigma^\top \Sigma) \quad (6.6)
\]

It should be noted that all the estimates given in this section, namely, (6.2), (6.4), (6.5) and (6.6), do not depend on any assumption on the parametric form of either the concentration or the time distribution.

7. Comparison of the Direct and Indirect Approaches

For simplicity, we will restrict to the problem of estimating \( \mu(E) \).

With the direct approach, we will take \( I C \) measurements of exposure

\[ E_{ij}, \quad i=1, \ldots, I; \quad j=1, \ldots, J_C \]

and estimate \( \mu(E) \) by
\[ \hat{\mu}(E) = \bar{E} = \frac{1}{IJ} \sum_{i,j} E_{ij} \]

which is unbiased and has variance

\[ \text{Var}[\hat{\mu}(E)] = \frac{1}{IJ} \sigma^2(E) \]

where \( \sigma^2(E) \) is given by Theorem 6.1.

With the indirect approach, we will take concentration measurements

\[ C_{ij} \quad i=1, \ldots, I_C; \quad j=1, \ldots, J_C \]

which will incur about the same cost as the exposure measurements in the direct approach. However, we can conceivably take more measurements of the time coefficients such that the variability in \( \hat{\mu}^T \) can be regarded as negligible. (Assuming that the time coefficients can be measured at practically no cost compared to the cost of concentration measurements.) We can estimate \( \mu(E) \) by

\[ \hat{\mu}(E) = (\hat{\mu}^C)^T \hat{\mu}^T \approx (\hat{\mu}^C)^T \mu^T \]

which is also unbiased* and has variance

---
*The unbiasedness does not depend on the assumption that the concentration and time coefficients are independent.
\[
\text{Var}[\hat{\mu}(E)] \approx \text{Var}[(\mu^C)^\top \mu^T] \\
= (\mu^T)^\top \text{Cov}(\mu^C)\mu^T \\
= (\mu^T)^\top \cdot \frac{1}{I^C} \mu^C \cdot \mu^T \\
= \frac{1}{I^C} \cdot (\mu^T)^\top \mu^C \mu^T
\]

**Theorem 7.1**

The efficiency of the indirect estimate \(\hat{\mu}(E)\) relative to the direct estimate \(\hat{\gamma}(E)\) is

\[
\frac{\text{Var}[\hat{\gamma}(E)]}{\text{Var}[\hat{\mu}(E)]} \approx 1 + \frac{(\mu^C)^\top \mu^C + \text{tr}(\mu^C C^T \mu^T)}{(\mu^T)^\top \mu^C \mu^T}
\]

which is strictly larger than one. ||

Theorem 7.1 shows that with about the same sampling cost the indirect approach is likely to improve the precision of the estimated \(\mu(E)\). The magnitude of the improvement depends on \(\mu^C, \mu^C, \mu^T\) and \(\mu^T\).

8. **Evaluation of Micro-Environment Type Classification**

Like in Section 7, we will restrict to the problem of estimating \(\mu(E)\). We will consider a classification scheme which identified \(K\) micro-environment types indexed as types 1, 2, ..., \(K\). We will consider the merit of merging the two micro-environment types 1 and 2.*

*Conceivably one might also be interested in splitting a certain micro-environment type, say, 1, into two finer types, say 1' and 1''. The problem can be regarded as the problem of merging the finer types 1' and 1'' into 1.
In other words, we will compare the sampling properties of the estimate \( \hat{\mu}(E) \) for the full classification compared with the reduced (with types 1 and 2 merged) classification.

We will make an additional assumption on the distribution of the time coefficients which will simplify the mathematical computation to be carried out in this section.

**Assumption 8.1**

The total duration \( t_{ij1} + t_{ij2} \) and relative ratios

\[
\begin{align*}
\tau_{ij1} &= \frac{t_{ij1}}{t_{ij1} + t_{ij2}} \\
\tau_{ij2} &= \frac{t_{ij2}}{t_{ij1} + t_{ij2}}
\end{align*}
\]

are uncorrelated, therefore

\[
\begin{align*}
E\tau_{ij1} &= \frac{E(t_{ij1})}{E(t_{ij1} + t_{ij2})} = \frac{T}{\mu_1 + \mu_2} \\
E\tau_{ij2} &= \frac{E(t_{ij2})}{E(t_{ij1} + t_{ij2})} = \frac{T}{\mu_1 + \mu_2}
\end{align*}
\]

*One model to justify this assumption is that the time allocated to each micro-environment time are generated from independent Poisson processes, for which the total duration is independent of allocation. The assumption remains to be validated with real data.*
In the full classification, we will estimate \( \hat{\mu}(E) \) by

\[
\hat{\mu}(E) = (\hat{\mu}_C)^\top (\hat{\mu}_T)
\]

\[
= (\bar{C}_{..})^\top (\bar{E}_{..})
\]

\[
\approx (\bar{C}_{..})^\top \mu_T
\]

which has variance

\[
\text{Var}[\hat{\mu}(E)] \approx \text{Var}[(\bar{C}_{..})^\top \mu_T]
\]

\[
= \frac{1}{\text{C}^\top \text{C}} (\mu_T)^\top \hat{t}_C \mu_T
\]

\[
= \frac{1}{\text{C}^\top \text{C}} \left[ \text{Var}(\mu_{11}^{T} C_{1j1} + \mu_{22}^{T} C_{1j2})ight.
\]

\[
+ 2 \text{ Cov}(\mu_{11}^{T} C_{1j1} + \mu_{22}^{T} C_{1j2} , \sum_{k=3}^{K} \mu_{3}^{T} C_{1jk})
\]

\[
\left. + \text{ Var}(\sum_{k=3}^{K} \mu_{3}^{T} C_{1jk}) \right]
\]

where \( \hat{t}_C = \text{ Cov}(C_{ij}) \)

In the reduced classification, we will adopt the following notations:

\[
C_{ij}^* = (C_{ij0}, C_{ij3}, \ldots, C_{ijk})^\top
\]

\[
t_{ij}^* = (t_{ij1} + t_{ij2}, t_{ij3}, \ldots, t_{ijk})^\top
\]

\[
\mu^*_C = (\mu_0^*, \mu_3^*, \ldots, \mu_K^*)^\top
\]
\[ \mu^T = (\mu_1^T, \mu_2^T, \mu_3^T, \ldots, \mu_K^T) \]

\[ \mu^C = \frac{1}{T} \sum_{i,j} C_{ij} \]

\[ \mu^T = \frac{1}{T^T} \sum_{i,j} t^*_{ij} \]

\[ \hat{C}^* = \text{Cov} (C^*_{ij}) \]

\[ \hat{T}^* = \text{Cov} (t^*_{ij}) \]

where \( C^*_{ij0} \) = concentration measurement for the merged type denoted as type \( 0^* \)

\[ = r_{ij1} C_{ij1} + r_{ij2} C_{ij2} \]

\[ \mu^C = EC^*_{ij0} \]

\[ = E_{ij1} \cdot \mu_1^C + E_{ij2} \cdot \mu_2^C \]

\[ = \frac{T \cdot C}{\mu_1^T + \mu_2^T} \]

(Assumption 8.1)

and

\[ r_{ijk} = \frac{t_{ijk}}{t_{ij1} + t_{ij2}} \quad k = 1, 2 \]

*The concentration coefficient for the merged type can be measured in various ways; we are assuming here that it is measured by averaging integrated exposure.*
We will estimate \( \mu(E) \) by

\[
\hat{\mu}(E) = (\hat{\mu}^*_C)^\top (\hat{\mu}^*_T)
\]

\[
= (\hat{C}^*)^\top (\hat{t}^*)
\]

\[
= (\hat{c}^*)^\top \mu^T
\]

which has variance*

\[
\text{Var}[\hat{\mu}(E)] = \text{Var}[(\hat{C}^*)^\top \mu^T]
\]

\[
= \frac{1}{I_C J_C} (\mu^T)^\top \hat{C} \mu^T
\]

\[
= \frac{1}{I_C J_C} \left[ \text{Var}(T + \mu_2^T C_{ij0}^*) \right.
\]

\[
+ 2 \text{Cov}(T + \mu_2^T C_{ij0}^*, \sum_{k=3}^{K} \mu_k^T C_{ijk})
\]

\[
+ \text{Var}(\sum_{k=3}^{K} \mu_k^T C_{ijk})
\]

The variance-of-sum term is the same as the corresponding term in the full classification estimation. We will compute the other two terms as follows.

*Also note that \( \hat{\mu}^*(E) \) is still unbiased.

\[
E[\hat{\mu}(E)] = (\hat{E}_C^*)^\top (\hat{E}^*_T)
\]

\[
= \mu^*_0 (\mu_1^T + \mu_2^T) + \sum_{k=3}^{K} \mu_k^T C_k^T
\]

\[
= \mu(E)
\]
\[
\text{Cov}[(\mu_1^T + \mu_2^T)C_{ij0}^*, \Sigma_{k=3}^K \mu_k^T C_{ijk}]
\]
\[
= (\mu_1^T + \mu_2^T) \sum_{k=3}^K \mu_k^T \text{Cov}(C_{ij0}^*, C_{ijk})
\]
\[
= (\mu_1^T + \mu_2^T) \sum_{k=3}^K \mu_k^T [\text{Cov}(r_{ij1} C_{ij1}, C_{ijk}) + \text{Cov}(r_{ij2} C_{ij2}, C_{ijk})]
\]

Note that

\[
\text{Cov}(r_{ij1} C_{ij1}, C_{ijk})
\]
\[
= E r_{ij1} C_{ij1} C_{ijk} - E r_{ij1} C_{ij1} E C_{ijk}
\]
\[
= E r_{ij1} E C_{ij1} C_{ijk} - E r_{ij1} E C_{ij1} E C_{ijk}
\]
\[
= E r_{ij1} \text{Cov}(C_{ij1}, C_{ijk})
\]
\[
= \frac{T}{\mu_1 + \mu_2} \rho_{tk}^{12}
\]

(Assumption 8.1)

A similar result holds for \(\text{Cov}(r_{ij2} C_{ij2}, C_{ijk})\), thus

\[
\text{Cov}[(\mu_1^T + \mu_2^T)C_{ij0}^*, \Sigma_{k=3}^K \mu_k^T C_{ijk}]
\]
\[
= \sum_{k=3}^K \mu_k^T (\rho_{1k}^{T1k} + \rho_{2k}^{T2k})
\]
\[
= \sum_{k=3}^K \text{Cov}(\mu_1^T C_{ij1} + \mu_2^T C_{ij2}, \mu_k^T C_{ijk})
\]
\[
= \text{Cov}(\mu_1^T C_{ij1} + \mu_2^T C_{ij2}, \sum_{k=3}^K \mu_k^T C_{ijk})
\]
Thus the covariance terms in $\text{Var} [\hat{\mu}(E)]$ and $\text{Var}[\mu^*(E)]$ are the same.

\[
\text{Var}(\mu_1^T + \mu_2^T)_{ij0} = (\mu_1^T + \mu_2^T)^2 \cdot \text{Var}(r_{ij1}C_{ij1} + r_{ij2}C_{ij2})
\]

\[
= (\mu_1^T + \mu_2^T)^2 \cdot [E(r_{ij1}C_{ij1} + r_{ij2}C_{ij2})^2 - E^2(r_{ij1}C_{ij1} + r_{ij2}C_{ij2})^2]
\]

\[
= (\mu_1^T + \mu_2^T)^2 \cdot [E r_{ij1}^2 E C_{ij1} + E r_{ij2}^2 E C_{ij2} + 2 E r_{ij1} r_{ij2} E C_{ij1} C_{ij2} - (E r_{ij1} E C_{ij1} + E r_{ij2} E C_{ij2})^2]
\]

\[
= (\mu_1^T + \mu_2^T)^2 \cdot [(E^2 r_{ij1} + \text{Var } r_{ij1}) (E^2 C_{ij1} + \text{Var } C_{ij1}) + (E^2 r_{ij2} + \text{Var } r_{ij2}) (E^2 C_{ij2} + \text{Var } C_{ij2}) + 2(\text{Er}_{ij1} \text{Er}_{ij2} + \text{Cov}[r_{ij1}, r_{ij2}]) - (EC_{ij1} EC_{ij2})] + \text{Var } r_{ij1} \text{ Var } C_{ij1} + E^2 r_{ij2} \text{ Var } C_{ij2} + \text{Var } r_{ij2} E^2 C_{ij2} + \text{Var } r_{ij2} \text{ Var } C_{ij2} + 2 \text{ Er}_{ij1} \text{ Er}_{ij2} \text{ Cov}(C_{ij1}, C_{ij2}) + 2 \text{ Cov}(r_{ij1}, r_{ij2}) EC_{ij1} EC_{ij2} + 2 \text{ Cov}(r_{ij1}, r_{ij2}) \text{ Cov } (C_{ij1}, C_{ij2})]
\]
\[
\begin{align*}
&= (\mu_1^T + \mu_2^T)^2 \cdot [E^2 r_{ij1} \hat{r}_{11} + E^2 r_{ij2} \hat{r}_{22} + 2 Er_{ij1} Er_{ij2} \hat{r}_{12}] \\
&\quad + \text{Var} \ r_{ij1} (\mu_1^C)^2 + \text{Var} \ r_{ij2} (\mu_2^C)^2 + 2 \text{Cov}(r_{ij1}, r_{ij2}) \mu_1^C \mu_2^C \\
&\quad + \text{Var} \ r_{ij1} \hat{r}_{11} + \text{Var} \ r_{ij2} \hat{r}_{22} + 2 \text{Cov}(r_{ij1}, r_{ij2}) \hat{r}_{12}] \\
&= (\mu_1^T + \mu_2^T)^2 \cdot [E^2 r_{ij1} \hat{r}_{11} + E^2 r_{ij2} \hat{r}_{22} + 2Er_{ij1} Er_{ij2} \hat{r}_{12}] \\
&\quad + (\mu_1^T + \mu_2^T)^2 \cdot \text{Var} \ r_{ij1} \cdot [(\mu_1^C - \mu_2^C)^2 + (\hat{r}_{11} + \hat{r}_{22} - 2\hat{r}_{12})]^2 \\
&= (\mu_1^T)^2 \hat{r}_{11} + (\mu_2^T)^2 \hat{r}_{22} + 2\mu_1^T \mu_2^T \hat{r}_{12} \\
&\quad + (\mu_1^T + \mu_2^T)^2 \text{Var} \ r_{ij1} \cdot [(\mu_1^C - \mu_2^C)^2 + (\hat{r}_{11} + \hat{r}_{22} - 2\hat{r}_{12})] \\
&= \text{Var}(\mu_1^T r_{ij1} + \mu_2^T r_{ij2}) \\
&\quad + (\mu_1^T + \mu_2^T)^2 \text{Var} \ r_{ij1} \cdot [(\mu_1^C - \mu_2^C)^2 + (\hat{r}_{11} + \hat{r}_{22} - 2\hat{r}_{12})]
\end{align*}
\]

(Assumption 8.3)

**Theorem 8.2**

The reduced classification results in an increase in variance of the estimated \( \mu(E) \):

\[
\text{Var}[\hat{\mu}(E)] - \text{Var}[\hat{\mu}(E)]
\]

\[
= \frac{1}{1 - \frac{r_{ij1} + r_{ij2} = 1; \text{Var} \ r_{ij2} = \text{Var} \ r_{ij1}, \text{Cov}(r_{ij1}, r_{ij2}) = -\text{Var}(r_{ij1})}{\text{Var}(r_{ij1})}}
\]
The amount of increase in variance is determined by two factors.

The first factor

\[(\mu_1^T + \mu_2^T)^2 \text{ Var } r_{1j1}\]

is determined by the distribution of the time coefficients. If more time \(\mu_1^T + \mu_2^T\) is expected to be spent in the micro-environment types 1 and 2, there is a larger increase in variance. The variability of the ratio \(r_{1j1}\) is also a determining factor.

The second factor

\[(\mu_1^C - \mu_2^C)^2 + (t_{11} + t_{22} - 2t_{12})\]

in the amount of increase in variance is determined by the distribution of the concentrate coefficients. If the two micro-environment types have larger differences in expected concentrations, \([(\mu_1^C - \mu_2^C)^2 \text{ large}]\), the increase will be larger. The other term in this factor can be re-expressed as

\[t_{11} + t_{22} - 2t_{12} = t_{11} + t_{22} - 2\rho(C_{1j1}, C_{1j2})\sqrt{t_{11} t_{22}}\]

which will be larger if \(\rho(C_{1j1}, C_{1j2})\) is smaller, and/or if \(t_{11}\) and \(t_{22}\) have larger difference. If the two micro-environment types have low or negative correlation, the increase in variance will be large.

For an extreme case, the "increase" in variance will be null, i.e., we can merge the two micro-environment types without losing any precision, if either (1) \(\text{Var } r_{1j1} = 0\), or (2) \(\mu_1^C = \mu_2^C, t_{11} = t_{22}\), \(\rho(C_{1j1}, C_{1j2}) = 1\), i.e., \(C_{1j1}\) and \(C_{1j2}\) are identically distributed.
Recommendation

Ideally, we will always achieve gains in precision by refining our micro-environment types. Nevertheless, it is also difficult to keep track of a large number of micro-environment types. (Difficulty in keeping an accurate diary, etc.) Therefore, the improvement in refinement should be evaluated along with the increased cost of maintaining extra micro-environment types. If the improvement is large, the refinement might be worth it. If the improvement is small, it is probably not worth it. Generally speaking, the more "valuable" refinements are the ones for which the expected time $\mu_1^T + \mu_2^T$ is large, the ratio $r_{ij1}$ is variable, the expected concentrate levels, $\mu_1^C$ and $\mu_2^C$, are appreciably different, and the measurements $C_{ij1}$ and $C_{ij2}$ are less correlated.

Acknowledgment

The author would like to express deep appreciation to Dr. Wayne Ott for encouragement and enlightenment on this work.
REFERENCES
