Hierarchical Bayes Models for Relating Particulate Matter Exposure Measures

Murali Haran
Bradley P. Carlin
John L. Adgate
Gurumurthy Ramachandran

ABSTRACT Understanding the effects of pollutants on the health of individuals requires consideration of different pollution sources. However, it is well known that there are large discrepancies between exposure measurements taken by indoor pollutant monitors and an individual's actual exposure. In this paper, we study data from the Hazardous Air Pollutants (HAPs) study, which collected reported emissions of particulates smaller than 2.5 microns (PM2.5) in three neighborhoods in the Minneapolis-St.Paul metropolitan area. The data consist of measurements of personal exposure for each individual, concentration inside the individual's home, and ambient concentration in the individual's neighborhood. The three neighborhoods exemplify areas with different numbers of pollutant sources: multiple source, single source, and "no major" source. This data set also features time-activity diary information on the amount of time each individual spent inside the home, outside near the home, and elsewhere. We use a Dirichlet-normal hierarchical structure to model the relationships among the pollutant levels measured by the different monitors. Implemented via a Gibbs-Metropolis algorithm, our model allows for relevant covariates for the various types of exposure, measurement error in the observations, missing data, and differences among exposure levels for different neighborhoods, seasons, and data monitoring sessions. We conclude with a discussion of our results, along with the limitations of (and possible improvements to) both our data set and analytic approach.

1 Introduction

Understanding the effects of pollutants on the health of individuals requires consideration of different kinds of sources of pollution. Since most people spend a significant portion of their time indoors, monitoring both indoor and outdoor air sources and concentrations is important. For a review of studies of particle concentrations and sources in homes and buildings, see Wallace (1996). In addition, it is well known that there are large discrepancies between pollutant concentration measurements taken by indoor pollutant monitors and the actual particulate matter exposure levels for an individual (see, for instance, Rodes et al., 1991 and McBride, 2001).

In this paper, we study data from the Hazardous Air Pollutants (HAPs) study, which collected reported emissions of particulates smaller than 2.5 microns (PM2.5) in three neighborhoods in the Minneapolis-St.Paul metropolitan area. The data consist of measurements of personal exposure for each individual, pollutant concentrations inside the individual's home, and concentration levels at a central site within each community. There is also information on the amount of time spent in different microenvironments.

A major difficulty in assessing pollutant exposure levels for individuals is the presence of large measurement error in the observations based on ambient pollutant monitors. The measurement error can be attributed almost entirely to the fact that the monitors measure the pollutant concentration levels in their immediate vicinity, and therefore their readings are surrogates for the true underlying concentrations. The instrument error is actually negligible in comparison, and can therefore be ignored as a major source of measurement error. We use Bayesian modeling techniques to relate personal, indoor, and outdoor PM2.5 concentrations while accounting for important covariates. We also include data from participants for whom information is partially missing when we draw our inferences. We thus provide a means to assess the relationships among the different exposure levels. This should also give us a sense of how well readings from exposure monitors reflect personal exposure levels.

The remainder of our paper is organized as follows. We begin by describing the data set, along with modeling issues and details of our model in Section 2. In Section 3 we describe our Bayesian computational approach for this model. We summarize our results in Section 4, and end with a discussion of our results and areas for future work in Section 5.

2 Particulate Matter Exposure Modeling

2.1 Description of Data

The Hazardous Air Pollutants (HAPs) study, conducted by researchers in the Division of Environmental and Occupational Health, School of Public Health, University of Minnesota and the Minnesota Pollution Control Agency (MPCA), collected reported PM2.5 emissions in three Minneapolis-St.Paul neighborhoods: Battle Creek (BCK), East St.Paul (ESP), and Phillips (PHI). These neighborhoods were actually selected based on the results of ambient volatile organic compound (VOC) monitoring: ESP and PHI for their relatively high VOC levels, and BCK because of its relatively low VOC levels (Pratt, McCourtney et al. 1998). These neighborhoods could also be thought as representative of different pollution paradigms. ESP has a 3M plant located in the area, making it a community with perhaps only a single major identifiable source of pollutants. On the other hand, PHI has multiple sources of pollution, since it is surrounded by several major roads. BCK is a neighborhood with no recognized major pollutant sources.

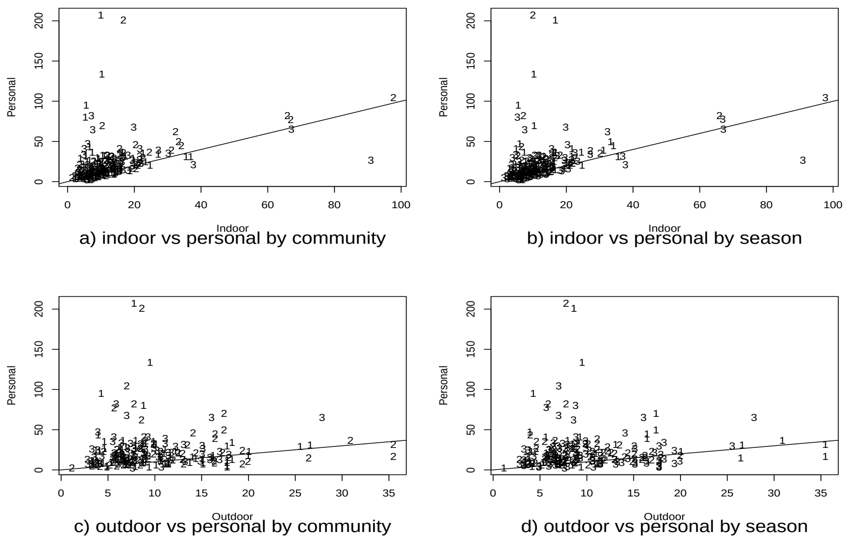
2.2 Data Collection

Healthy nonsmoking working-age adults were recruited within neighborhoods by house-to-house canvassing and direct solicitation. Project staff collected three PM2.5 exposure measurements: individual (personal) exposure, within-home (indoor) exposure, and outside-home (outdoor) exposure. The observations were taken over the three communities for 12, 11 and 9 individuals, respectively (a total of 32 participants), over three seasons: spring, summer, and fall. There were at most 4 data collection sessions for each season, with up to 2 two-day periods within each session. The study commenced on May 24th, 1999 and ended November 11th, 1999. The outdoor exposure measurements were taken at a "core" site in each of the three communities: for ESP, the core site was a fire station (Holman Field), for BCK it was the neighborhood middle school, and for PHI it was the local community center. These act as surrogate measures for concentration levels outside each individual's home.

Individual (personal) exposures were measured using 4 litres per minute (Lpm) pumps. These pumps were put in foam-insulated bags and carried by each participant with a shoulder strap. The indoor pollutant concentrations were measured by 10Lpm pumps that were placed in each home. The stationary indoor inlets were placed at the subject's seated breathing height in the room where he/she reported spending the majority of his/her awake time. The outdoor exposure measures were measured for each community by a 16.67Lpm monitor placed at the core sites, which were at least a block away from major streets and industrial sources.

Personal and indoor exposure measurements were generally taken over a two day period beginning around 7pm on a particular day, and continuing until around 7pm 48 hours later. The outdoor exposure (core site) measurements were also taken over a two day period, though these do not overlap perfectly with the personal and indoor exposure measurements since outdoor measurements start at 12am the first day and end at 12am after 48 hours. Of the 48 hours over which each of the measurements was taken, there are, in practice, 34±4.4 hours of overlap between the two measurements (though up to 43 hours of overlap are possible in theory, had the personal and indoor monitors started and ended at exactly 7pm). The reported observations consist of single numbers representing the average PM2.5 concentration for each 2 day period. For details regarding the methods used to obtain the measurements, see Ramachandran et al. (2000).

Figure 1 plots the raw HAPs data. More specifically, panels (a) and (b) plot indoor versus personal exposure by community (1=BCK, 2=ESP,



a) indoor vs personal by community

b) indoor vs personal by season

c) outdoor vs personal by community

d) outdoor vs personal by season

FIGURE 1. Plots of raw HAPs data; plotting

3=PHI) and season (1=spring, 2=summer, 3=fall), respectively, while panels (c) and (d) repeat the exercise for outdoor versus personal exposure. The plots reveal the expected generally increasing pattern, as well as the presence of a few outliers with unusually high (though not nonsensical) personal exposures. Somewhat more troubling is the presence of so many observations below the "vertical = horizontal" reference lines in the plots, since these correspond to individuals whose indoor or outdoor exposures already exceed the measured total personal exposure for that observation session. This situation can only be the result of measurement error in our data, and accounting for it will be a key focus of our modeling (Subsection 2.5 below).

2.3 Time-Activity and Covariate Data

The time-activity information for each monitored day was obtained by having each of the participants in the study fill out a survey that provided information on the time spent in each of seven environments: inside home, inside school (or work), outside home, outside school (or work), in transit, in other (time spent anywhere else indoors), and out other (time spent anywhere else outdoors). To obtain the fraction of time spent inside the home, the total recorded time was simply divided by 24. We similarly obtained the fraction of time spent outside the home. We set the fraction of time spent elsewhere to 1 minus the sum of these two. Since there were no exposure measurements corresponding to the other time-activity measurements (time spent in transit, time spent elsewhere outdoors and indoors), these time-activity measurements were used only for the purposes of verifying the consistency of the data.

In addition to exposure levels and time-activity data, the study also collected information regarding several other covariates that may have had an effect on personal exposure levels. These covariates include indicators of whether the individual had pumped gasoline, had been in physical contact with gasoline, spent time in an enclosed garage, had cleaned a fireplace, had started or managed a grill, were in the presence of cigarette or tobacco smoke, paint solvents or plastic fumes, among others.

2.4 Missing Data

A complete data set would have contained 768 rows of data (32 individuals × 3 seasons × 4 sessions × 2 session days). Unfortunately, the number of complete rows of data available is just 217. The missing exposure or pollutant concentration data are due to failures of the outdoor, indoor, or personal samplers. Indoor and personal samplers can fail for a variety of reasons: battery failure, mishandling of the sample (dropped, filter punctured, etc.), or equipment breakdown. Similarly, missing outdoor data can be attributed to "equipment failure," which can be anything from a power outage to mechanical problems (a frequent occurrence) to mishandling of the sample (for instance, someone at MPCA dropping the filter in the weigh room). The time-activity diaries, which contain information on the time the subjects spent in different environments and records covariate information, are filled out by the subjects in the study. The missing covariate and time-activity information is therefore due to subjects failing to fill out the information on the monitored days.

2.5 Model Components

Let Y^P_{ijskm} be the observed personal exposure for individual j within community i during season s (spring, summer and fall), session k, and day (within session) m. Y^I_{ijskm} is the corresponding indoor exposure measurement, while Y^O_{ijskm} is the community-level outdoor concentration measured at the same time. Note we retain a subscript j for the outdoor measurement even though it is not a direct measurement for outdoor concentration at the home of a particular individual j.

Let the true personal, home indoor, home outdoor, and remainder (all but indoor home and outdoor home) exposure levels be P_{ijskm}, I_{ijskm}, O_{ijskm} and R_{ijskm}. Note that these quantities are not observable; we do not even model R_{ijskm} here since it is a term merely used to represent the concentration each individual is exposed to that is not accounted for by home indoor or outdoor concentration levels. Also let the proportion of time spent home indoors, home outdoors and "elsewhere" be a_{ijskm}, b_{ijskm}, and c_{ijskm}, respectively. These are observed data, though we may be interested in imputing values for those that are missing.

Another feature of the data is that there are instances where the physical constraint

$$P_{ijskm} = a_{ijskm} I_{ijskm} + b_{ijskm} O_{ijskm} + c_{ijskm} R_{ijskm} \tag{1}$$

is not satisfied at the data level. That is, there are cases where

$$Y^P_{ijskm} < a_{ijskm} Y^I_{ijskm} + b_{ijskm} Y^O_{ijskm}. \tag{2}$$

for some (i, j, s, k, m). This happens in 25 of the complete rows of data available. Thus, we have to rely on our prior distributions and the borrowing of strength from the remaining data to get the posterior to obey the physical constraint (1).

2.6 Hierarchical Model for PM2.5 Exposures

There are two main challenges when modeling and analyzing the HAPs data set: the presence of measurement error in all three exposure measurements, and the missing observations.

We begin by modeling the measurement error in the observations. We begin with an error model in which we assume that the observed exposures are simply imperfect but unbiased measures of the true exposure levels, i.e.,

$$\begin{aligned} Y^P_{ijskm} | P_{ijskm} &\sim N(P_{ijskm}, 1/\tau_{yP}) \\ Y^I_{ijskm} | I_{ijskm} &\sim N(I_{ijskm}, 1/\tau_{yI}) \\ Y^O_{ijskm} | O_{ijskm} &\sim N(O_{ijskm}, 1/\tau_{yO}) \end{aligned}$$

Recall we also impose constraint (1) on the true exposures P_{ijskm}, I_{ijskm}, and O_{ijskm}.

Next, we model how certain covariates recorded in the time-activity diaries can affect exposure. Many covariates were indicators for events that almost never occurred, and covariate information was often missing. Our criterion for including an indicator covariate in our analysis was (a) the covariate was considered by our subject matter experts (JLA and GR) to be relevant to our model, and (b) that there were at least 10 instances where the event corresponding to the covariate occurred. Four covariates were finally selected for inclusion: dsmoke (was tobacco smoked inside the individual's home?), dgrill (did the individual use a grill, or burn leaves or trash?), dgasoline (did the individual have contact with diesel fumes?), and dcigarette (what was the number of cigarettes smoked in the individual's presence?). Covariates were associated with the appropriate pollutant concentration measurement in our model: dgrill, dcigarette and dgasoline (denoted by $X^{(1)}_{ijskm}, X^{(2)}_{ijskm}$, and $X^{(3)}_{ijskm}$, respectively) were related to personal exposure, while dsmoke ($X^{(4)}_{ijskm}$) was related to indoor pollutant concentration. Thus we have

$$\begin{aligned} P_{ijskm}|\nu_{ijs}, \tau_P &\sim N(\nu_{ijs} + \sum_{l=1}^{3} \beta_l X^{(l)}_{ijskm} \ , \ 1/\tau_P), \\ \text{and } I_{ijskm}|\eta_{ijs}, \tau_I &\sim N(\eta_{ijs} + \beta_4 X^{(4)}_{ijskm} \ , \ 1/\tau_I). \end{aligned}$$

We now complete the hierarchical specification of the distributions of the pollutant levels, starting with outdoor pollutant concentrations:

$$O_{ijskm}|\mu_i, \tau_O \sim N(\mu_i, 1/\tau_O), \text{ and}$$

$$\nu_{ijs}|u, \tau_\nu \sim N(u, 1/\tau_\nu) \ , \ \eta_{ijs}|v, \tau_\eta \sim N(v, 1/\tau_\eta) \ , \ \mu_i|w, \tau_\mu \sim N(w, 1/\tau_\mu) \ .$$

Finally, we place gamma priors on the overall exposure parameters, since we wish these highest-level mean parameters to remain positive. Thus we assume

$$u \sim G(\alpha_u, \beta_u), v \sim G(\alpha_v, \beta_v), w \sim G(\alpha_w, \beta_w),$$

while for the precision components we let

$$\begin{aligned} \tau_{yP} \sim G(\alpha_{yP}, \beta_{yP}), \tau_{yI} \sim G(\alpha_{yI}, \beta_{yI}), \tau_{yO} \sim G(\alpha_{yO}, \beta_{yO}), \\ \tau_P \sim G(\alpha_P, \beta_P), \tau_I \sim G(\alpha_I, \beta_I), \tau_O \sim G(\alpha_O, \beta_O), \\ \text{and } \tau_\nu \sim G(\alpha_\nu, \beta_\nu), \tau_\eta \sim G(\alpha_\eta, \beta_\eta), \tau_\mu \sim G(\alpha_\mu, \beta_\mu) \ . \end{aligned}$$

Note we do not constrain the true exposures $(P_{ijskm}, I_{ijskm}, O_{ijskm})$ to be non-negative, but for now simply rely on the constraint (1) to ensure that they result in sensible posterior distributions for the true exposure levels.

The only manner in which the data corresponding to time spent in each environment (a_{ijskm}, b_{ijskm}, c_{ijskm}) appear in this model is to help impose the constraint (1) on the true exposures $(P_{ijskm}, I_{ijskm}, O_{ijskm})$. One way to acknowledge the variability in these measurements (as well as incorporate data rows with missing time-activity data) is by assuming the (a_{ijskm}, b_{ijskm}, c_{ijskm}) follow a Dirichlet distribution:

$$\begin{pmatrix} a_{ijskm} \\ b_{ijskm} \\ c_{ijskm} \end{pmatrix} \overset{ind}{\sim} Dir((\alpha_a)_{ijs}, (\alpha_b)_{ijs}, (\alpha_c)_{ijs}) \ . \tag{3}$$

To complete this portion of the model, we further assume that

$$\begin{aligned} (\alpha_a)_{ijs} \sim G(\gamma_a, \delta_a), (\alpha_b)_{ijs} \sim G(\gamma_b, \delta_b), (\alpha_c)_{ijs} \sim G(\gamma_c, \delta_c), \\ \text{where } \gamma_a \sim G(\alpha_{\gamma a}, \beta_{\gamma a}), \gamma_b \sim G(\alpha_{\gamma b}, \beta_{\gamma b}), \gamma_c \sim G(\alpha_{\gamma c}, \beta_{\gamma c}), \\ \text{and } \delta_a \sim G(\alpha_{\delta a}, \beta_{\delta a}), \delta_b \sim G(\alpha_{\delta b}, \beta_{\delta b}), \delta_c \sim G(\alpha_{\delta c}, \beta_{\delta c}) \ , \end{aligned}$$

where the prior parameters for the gamma distributions are set to 1 and 100, respectively. That is, the γ and δ parameters have prior mean 100 but prior variance 10,000, allowing substantial flexibility.

It is worth noting how we deal with estimating parameters for which several corresponding data points are missing. At the lowest level of the hierarchy, we only estimate P_{ijskm}, I_{ijskm}, or O_{ijskm} when Y^P_{ijskm}, Y^I_{ijskm}, or Y^O_{ijskm} respectively is available. At the next level of the hierarchy, we only estimate ν_{ijs} when at least one Y^P_{ijskm} is available for that (i, j, s). We determine similar criteria for estimating η_{ijs} and μ_i. Details of the full conditional distributions for these parameters are given in Haran (2001).

3 Computing

Due to the high dimensionality and complexity of the model, it is natural to adopt a Bayesian analytic approach, with computer implementation via Markov chain Monte Carlo (MCMC) methods. The prior-likelihood conjugacy for several parameters in our model means that, in the absence of constraint (1), ordinary Gibbs updates would be possible for these parameters. However, there are many parameters for which no closed form full conditional distributions are available, and hence require Metropolis-Hastings updating. We log-transform all such parameters that have strictly nonnegative support, and use random walk Metropolis steps with Gaussian proposals to sample them.

To impose the constraint described in (1), we note that we cannot simply implement the constraint "on the fly," as often done with sum-to-zero constraints in conditionally autoregressive (CAR) priors for spatial lattice models. This is because, since we are not explicitly modeling any R_{ijskm} terms, (1) restricts the $(P_{ijskm}, I_{ijskm}, O_{ijskm})$ triples to a generalized half-plane (not to a set of measure zero). Writing the vector of third and higher stage hyperparameters as $\boldsymbol{\xi}$, the standardization to this half-plane adds a normalizing constant $c(\boldsymbol{\xi})$ to the joint Bayesian model specification. The difficulty in estimating $c(\boldsymbol{\xi})$ would seem to preclude ordinary Gibbs sampling, since it is unknown yet required for the $\boldsymbol{\xi}$ full conditional at every iteration of the sampler.

To circumvent this difficulty, we could instead run a modified Gibbs sampler over the *unconstrained* space, iterating until convergence as usual, and subsequently delete from the final sample any iterates that fail to satisfy the constraint. This approach is the Monte Carlo analog of first finding the unconstrained posterior, and then standardizing it to the constraint space "at the very end." However, since each hyperparameter $\boldsymbol{\xi}$ is associated with many (individual-level) $(P_{ijskm}, I_{ijskm}, O_{ijskm})$ triples, application of this approach to a standard Gibbs sampler would delete essentially every $\boldsymbol{\xi}$ from the sample. But most of these deletions would be caused by violations of (1) by those (i,j,s,k,m) observations corresponding to "data violations" of the form (2). To avoid this problem, we instead *oversample* the $\boldsymbol{\xi}$'s, drawing a new one for each (i,j,s,k,m) just prior to the generation of each $(P_{ijskm}, I_{ijskm}, O_{ijskm})$ triple. Following convergence, we then apply constraint (1) for each (i,j,s,k,m) sequentially, deleting any offending $(P_{ijskm}, I_{ijskm}, O_{ijskm})$ triple (and corresponding oversampled $\boldsymbol{\xi}$).

We ran this MCMC sampler over the unconstrained space for 1000 iterations, meaning we obtained 219,000 $\boldsymbol{\xi}$ values since we have 219 distinct observations (i,j,s,k,m) (217 complete and 2 partially missing but imputed). An examination of sample traces and autocorrelations suggested virtually immediate convergence to the stationary distribution of the (unconstrained) posterior. Application of the constraint then "thinned" our sample by roughly 22% overall (less for some observations, more for others). While our oversampling approach is still somewhat approximate, the relatively low proportion $(25/219 = 11.4\%)$ of data violations of the form (2) suggests it will perform well for those parameters for which thinning is not too severe.

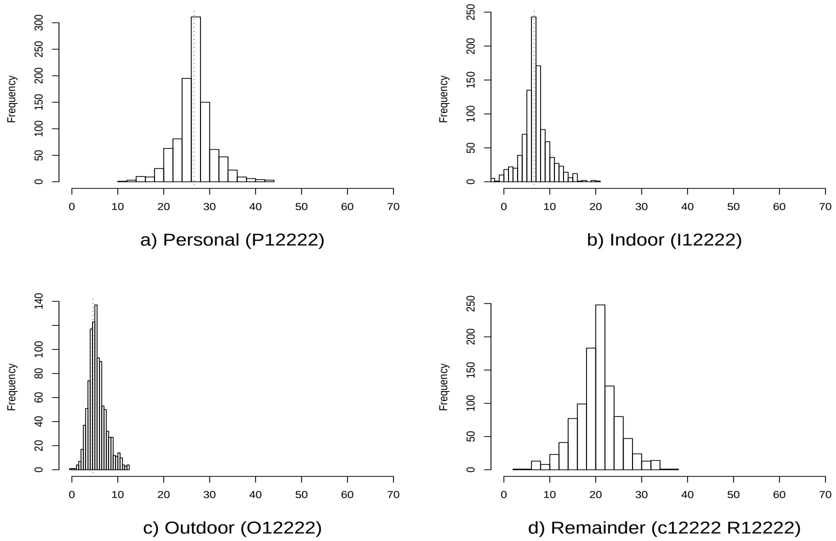
4 Results

We first present posterior summaries for two particular observations, one where the individual spent a lot of time outdoors, and another where the individual spent most of the time indoors. The first set of observations (for subject 2 in community 1, during season 2, session 2 and session day 2) is $Y^P = 26.6, Y^I = 6.6, Y^O = 4.6$ and $a = 0.83, b = 0.125, c = 0.0416$. The second set of observations (for subject 2, in community 3 during season 2, session 2 and session day 2) consists of $Y^P = 35.3, Y^I = 30.2, Y^O = 5.5$, and $a = 0.999, b = 0.0005, c = 0.0005$. (We note these last two values are artificial, used to replace the recorded values $b = c = 0$, which are not permitted by our model).

Table 1 gives estimated posterior means and standard deviations for these two observations, while histograms of the posterior distributions of the true underlying exposure levels are provided in Figures 2 and 3. In general, the posterior distributions estimate true underlying exposures that are fairly close to the observed exposures (indicated by dotted vertical lines in the histograms). In all cases we summarize $c_{ijskm}R_{ijskm}$ (instead of merely R_{ijskm}) since the often very small observed c values in turn lead to overly extreme imputed R values. The high variability in the personal exposure distributions, along with the large estimates for remainder exposure, suggest that the ability of indoor and outdoor exposure concentrations alone to predict personal exposure concentration is poor for this data set. Thus, like previous studies, our study casts some doubt on the ability of indoor and outdoor ambient pollution monitors to accurately assess an individual's

i, j, s, k, m	P_{ijskm}	I_{ijskm}	O_{ijskm}	$c_{ijskm}R_{ijskm}$
1,2,2,2,2	26.7 (4.1)	6.9 (3.1)	5.5 (1.9)	20.3 (4.7)
3,2,2,2,2	35.6 (3.0)	27.9 (4.1)	6.0 (1.9)	7.7 (4.9)

TABLE 1. Posterior means (standard deviations), two individuals.



a) Personal (P12222)

b) Indoor (I12222)

c) Outdoor (O12222)

d) Remainder (c12222 R12222)

FIGURE 2. Estimated posterior distributions, observation (1,2,2,2,2).

personal exposure. This observation is of interest since pollutant exposure levels based on such monitors are routinely used in statistical and other models relating them to disease incidence and mortality (e.g. Dominici et al., 2000).

Figure 4 shows posterior distributions for the missing time-activity diary information for two different instances: individual 9 in community 1, season 3, session 2, session day 2 (first column), and individual 1 in community 3, season 1, session 1, session day 1 (second column). The model borrows information from available time-activity diary information for other sessions for the same individual during the same season and session. The time spent indoors for related sessions was much higher for the first individual, consistent with Figures 4(a) and (d). This individual also had more observed data from which to borrow strength, as indicated by the reduced posterior variability.

Figure 5 shows histograms for the posterior distributions for our four

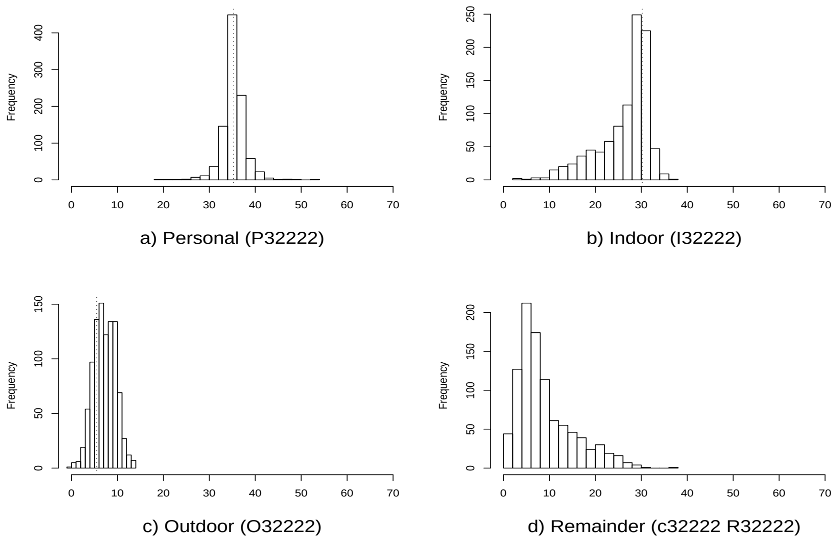
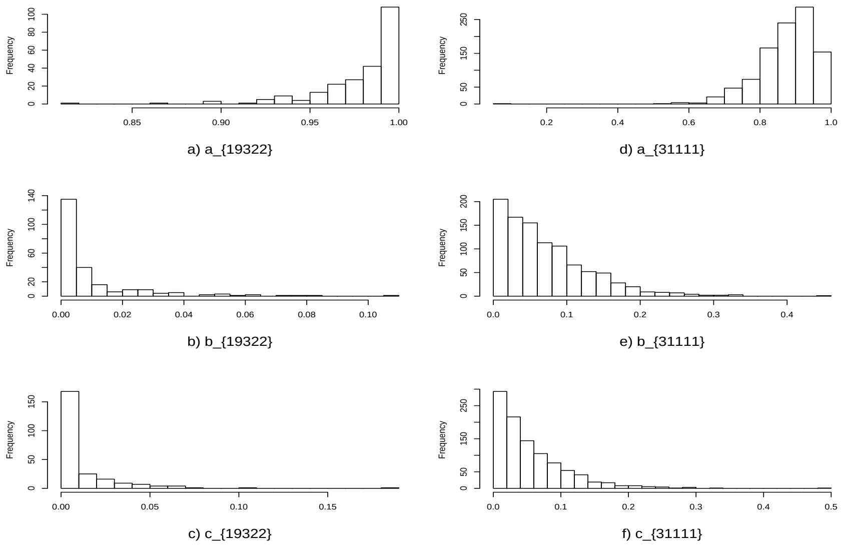


FIGURE 3. Estimated posterior distributions, observation (3,2,2,2,2).

covariate effects. The 95% credible regions for the covariates are (–16.1, 23.9) for dgrill, (0.5 , 2.0) for dcigarette, (3.4, 14.3) for dgasoline, and (7.2, 30.7) for dsmoke. Thus all seem positively associated with higher PM2.5 exposure, although the dgrill covariates is not significant at the 0.05 level. It is not surprising that cigarette smoke has a significant effect on pollutant concentrations: β_2 corresponds to the parameter for the *number* of cigarettes smoked in the individual's presence, while β_4 is the parameter for the indicator of home indoor cigarette smoking. Thus the estimated rise in personal PM2.5 exposure per cigarette smoked is roughly 1.25. It is perhaps not surprising that dgasoline is only marginally significant, since it produces a vapor rather than a fume; personal PM2.5 concentration levels would likely rise while pumping gas only due to fumes from other vehicles idling nearby.

Figure 6 plots indoor versus personal covariate-adjusted posterior mean PM2.5 concentration levels for an individual during a given season, or living in a given community. Based on this data set, there does not seem to be a marked difference in average exposures across different seasons (panel (a)) or communities (panel (b)). Thus the different pollution sources for the different neighborhoods do not appear to have much of an impact on the average PM2.5 exposures.

Turning to the Dirichlet parameters, we first standardize to the unit interval by defining $(\alpha_a)^*_{ijs} = (\alpha_a)_{ijs}/[((\alpha_a)_{ijs} + (\alpha_b)_{ijs} + (\alpha_c)_{ijs})]$, and sim-



a) a_{19322}

d) a_{31111}

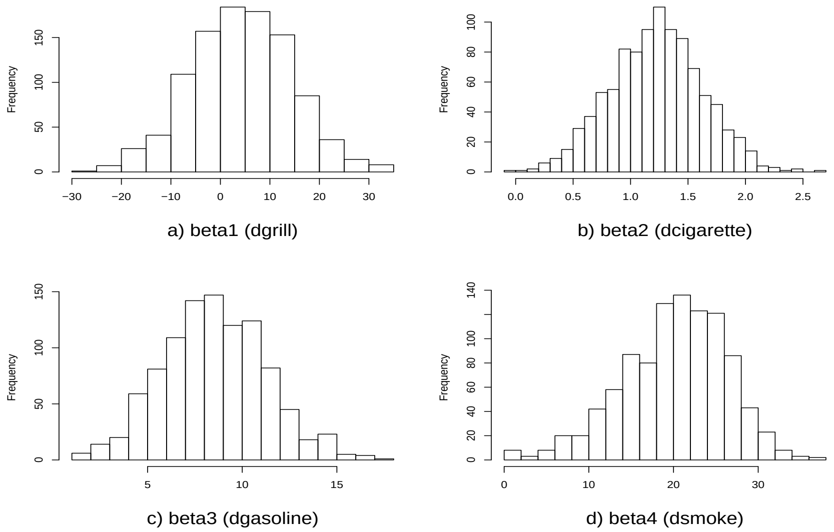
b) b_{19322}

e) b_{31111}

c) c_{19322}

f) c_{31111}

FIGURE 4. Predictive distributions for proportion of time spent indoors, outdoors and elsewhere.



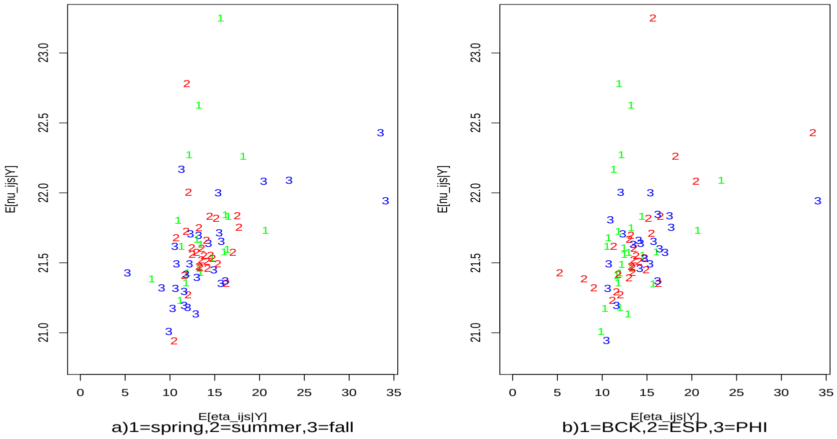
a) beta1 (dgrill)

b) beta2 (dcigarette)

c) beta3 (dgasoline)

d) beta4 (dsmoke)

FIGURE 5. Posterior distributions for covariate parameters.



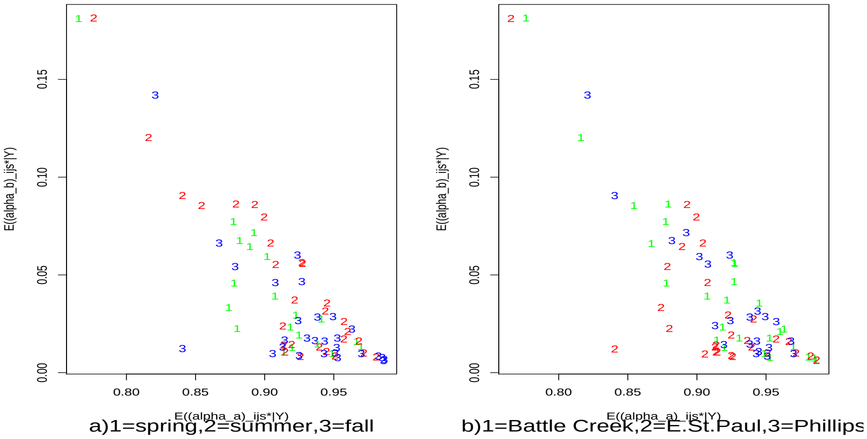
a)1=spring,2=summer,3=fall

b)1=BCK,2=ESP,3=PHI

FIGURE 6. $E(\nu_{ijs}|Y)$ vs. $E(\eta_{ijs}|Y)$

ilarly for $(\alpha_b)^*_{ijs}$ and $(\alpha_c)^*_{ijs}$. The triple $((\alpha_a)^*_{ijs}, (\alpha_b)^*_{ijs}, (\alpha_c)^*_{ijs})$ then gives the expected proportions of time spent by individual j in community i and season s in the three microenvironments: indoors, outdoors, and elsewhere, respectively. Figure 7(a) shows the average $(\alpha_a)^*_{ijs}$ versus $(\alpha_b)^*_{ijs}$ marked by season, while Figure 7(b) gives the same plot marked by community. As with Figure 6, there is no apparent separation of time-activity information by season or community. This indicates to us that there is nothing intrinsically different about the habits of these individuals over different seasons or communities.

Finally, we do some standard model checking by examining model residuals (see e.g. Carlin and Louis, 2000, Chapter 6). Figures 8(a-c) plot standardized residuals versus fitted values for personal, indoor, and outdoor concentrations, respectively. That is, in Figure 8(a) the predicted values (horizontal axis) are $\overline{P}_{ijskm}$, the averages of the $P^{(g)}_{ijskm}$ over the (appropriately thinned) MCMC samples, while the residuals (vertical axis) are $\overline{r}_{ijskm}$, the averages of $r^{(g)}_{ijskm} \equiv (Y^P_{ijskm} - P^{(g)}_{ijskm})\sqrt{(\tau_{yP})^{(g)}}$ over the same samples. These plots facilitate the identification of outliers, though many of these were already apparent from the raw data plots in Figure 1. The increasing trend in Figure 8(a) is the result of shrinkage of the fitted values away from the data points towards their grand mean, and is unavoidable in plots of this type for models inducing shrinkage (see e.g. Hodges, 1998, Sec. 4.5). None of the three plots indicate increasing variability as the fitted values increase (the classic "megaphone opening to the right" pattern), but the increasing sparseness of the plots as we move up and to the right does suggest a better fit would be obtained if the exposure measurements were



a)1=spring,2=summer,3=fall

b)1=Battle Creek,2=E.St.Paul,3=Phillips

FIGURE 7. $E((\alpha_b)^*_{ijs}|Y)$ vs. $E((\alpha_a)^*_{ijs}|Y)$

log-transformed before the normality assumptions were imposed. Unfortunately, there is no scientific justification for the additive model (1) for the exposures on the log scale, and as such we have not pursued the lognormal distributional option.

5 Summary and Future Work

We have described a hierarchical Bayesian model for the HAPs data set on PM2.5 exposure concentration levels in the Twin Cities. Our model not only reflects our uncertainty about the estimates due to measurement error in the observations, but also imposes physical constraints on the true, underlying (unobservable) exposure levels. Our model also allows us to include important covariate information, as well as data from the participants for whom information is partially missing.

Our model allows us to account for several sources of uncertainty in a unified and systematic manner. By sampling from the posterior distribution of the model, we can also easily answer a variety of estimative and predictive questions of interest without too much additional analytical or computational effort. Our results indicate that the ability of indoor and outdoor exposure to predict personal exposure is poor for this data set. We also do not see any clear difference in terms of individual, indoor or outdoor exposure levels among the three different neighborhoods. This data set does not support the hypothesis that varying outdoor sources of pollutants (no major source, single source, and multiple source) lead to significant changes

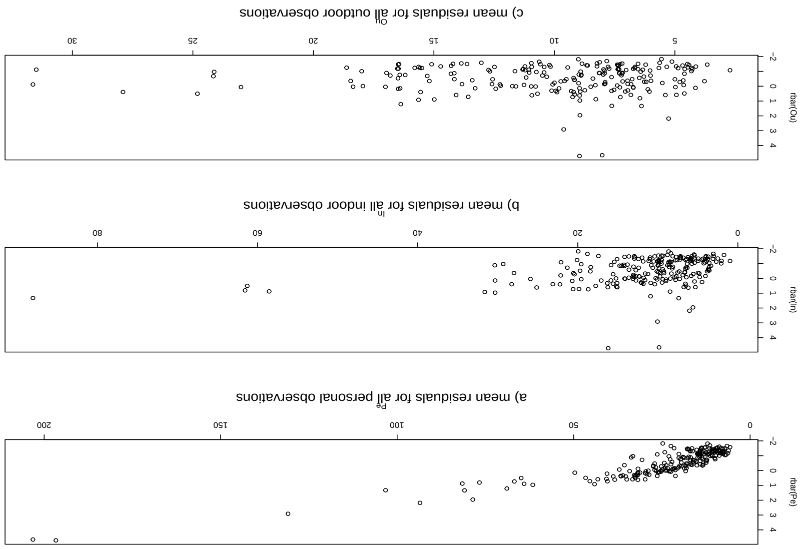


FIGURE 8. Plots of residuals versus fitted values for personal, indoor, and outdoor concentrations.

in personal PM2.5 exposure levels.

The term “personal activity cloud” (see e.g. Rodes et al., 1991) is used to describe the effect of an individual’s activities on the pollutant levels within a small space around the individual. There are very complex models relating personal and indoor exposure levels which try to account for personal clouds, such as the one proposed by McBride (2001). However, the model in this paper relies on controlled conditions and almost continuously available data from all the monitors, neither of which is available in our context.

A future study would ideally include outdoor concentration data that are truly recorded outside each home, rather than a community-wide outdoor measurement that has to be used as a surrogate for the measurement of interest. Exposure monitors in all seven microenvironments would help us estimate the magnitude of the personal cloud. It would also be very useful to have a better time-activity questionnaire, which included information on covariates that might significantly affect the PM2.5 personal exposure (say, a covariate that indicated how often and how long a participant cooked during the period in question).

Since we expect pollutant concentration levels to be different in the winter season, it would be helpful to expand the study to include observations for this season as well. The large amount of missing exposure and time-activity diary information may be, at least partially, the cause for the poor predictive relationship between indoor, outdoor and personal exposure measures. The missing data may also account for the weak distinction in exposure levels for the three different neighborhoods. A more complete study, with fewer missing data may well help answer such questions more conclusively.

References

Carlin, B.P. and Louis, T.A. (2000) *Bayes and empirical bayes methods for data analysis* (2nd ed.), Boca Raton, FL: Chapman & Hall/CRC Press.

Dominici, F. and Zeger, S.L., Samet, J.M. (2000) "A measurement error model for time-series studies of air pollution and mortality," *Biostatistics*, **1**, 157–175.

Haran, M. (2001) "Hierarchical Bayes Modeling and Computation for Environmental Exposure and Health Outcome Assessment", Unpublished Ph.D. dissertation, School of Statistics and Division of Biostatistics, University of Minnesota.

Hodges, J.S. (1998) "Some algebra and geometry for hierarchical models, applied to diagnostics" (with discussion), *J. Roy. Statist. Soc., Series B*, **60**, 497–536.

McBride, S.J. (2001) "A marked point process model for indoor air pollutant concentration time series," *Duke University Technical Report.*

Pratt, G., McCourtney, M., Wu C.Y., Bock D., Sexton K., Adgate J., Ramachandran G. (1998) "Measurement and source apportionment of human exposures to toxic air pollutants in the Minneapolis-St. Paul metropolitan area,"*Measurement of Toxic and Related Air Pollutants.*, Cary, N.C., AWMA.

Ramachandran, G., Adgate, J.L., Hill, N., and Sexton, K. (2000) "Comparison of short-term variations (15 minute averages) in outdoor and indoor PM2.5 concentrations," *Journal of Air and Waste Management Association*, **50**, 1157–1166.

Rodes, C.E., Kamens, R.M., and Wiener, R.W. (1991) "The significance and characteristics of the personal activity cloud on exposure assessment measurements for indoor contaminants," *Indoor Air*, **2**, 123–145.

Wallace, L. (1996) "Indoor particles: a review," *Journal of Air and Waste Management Association*, **46**, 98–126.