A Bayesian analysis of the effects of particulate matter using a human exposure simulator

By: Brian J Reich[a], Montserrat Fuentes[a], and Janet Burke[b] [1]

Institute of Statistics Mimeo Series #2586

June 14, 2006

[a] *Department of Statistics, North Carolina State University,*
2501 Founders Drive, Box 8203, Raleigh, NC 27695

[b] *US EPA, National Exposure Research Laboratory*
Research Triangle Park, North Carolina 27711

Correspondence Author: Brian J. Reich
E-mail: reich@stat.ncsu.edu
Telephone: (919) 513-7686
Fax: (919) 515-7591

[1] The authors thank Lance McCluney of the U.S. EPA for providing the CO data.

A Bayesian analysis of the effects of particulate matter using a human exposure simulator

Abstract

Particulate air pollution has been associated with mortality in several epidemiological studies. The US EPA currently regulates PM_{10} and $PM_{2.5}$ (mass concentration of particles less than 10 μm and 2.5 μm, respectively), but it is not clear which size of particles are most responsible for adverse heath outcomes. A current hypothesis is that ultrafine particles with diameter less than 0.1μm are particularly harmful because their small size allows them to deeply penetrate the lungs. This paper investigates the effect of exposure to particles of varying diameter on daily mortality. We propose a new dynamic factor analysis model to relate the ambient concentrations of PM_{10}, $PM_{2.5}$, and several sizes of particles with diameters ranging from 0.01 to 0.40 μm with mortality. We introduce a Bayesian model that converts ambient concentrations into simulated personal personal exposure using the EPA's Stochastic Human Exposure and Dose Simulator, and relates simulated exposure with mortality. While no function of ambient PM levels are found to be associated with mortality, our analysis indicates that exposure to particles with diameter 0.20μm is associated with mortality.

Key Words: ecological fallacy, human exposure, latent factor model, SHEDS, ultrafine particles.

1 Introduction

Several epidemiological studies have shown an association between air pollution and adverse health outcomes (Dockerty et al., 1992; Schwartz, 1994; Pope et al., 1995; American Thoracic Society and Bascom 1996a, 1996b). Most of the recent work in this area has focused on PM_{10} and $PM_{2.5}$, the mass concentrations of particles less than 10μm and 2.5μm, respectively. However, it is not clear which sizes of particles are most responsible for adverse heath outcomes. A current hypothesis is that ultrafine particles with diameter less than 0.1μm are particularly harmful because their small size allows them to deeply penetrate the lungs. The literature on ultrafine particles is relatively sparse compared to the literature on PM_{25} and PM_{10}. Pekkanen et al. (2002) demonstrated an association between ultrafine particle levels and cardiovascular symptoms, while de Hartog et al. (2003) and Timonen et al. (2004) failed to find a relationships between ultrafine concentration and cardiorespiratory symptoms. Also, Wichman et al. (2000) showed that ambient ultrafine concentration levels were associated with daily mortality in three European cities.

This paper uses a new data set to investigate the effects of different sizes of particulate matter on mortality. Pollution data is measured at a single monitoring station in downtown Fresno, CA. The ambient concentrations of PM_{10}, $PM_{2.5}$, and several sizes of particles with diameters ranging from 0.01 to 0.40 μm are recorded hourly for 2002. The health outcome is daily natural mortality in downtown Fresno for elderly Caucasians.

We develop a novel dynamic factor model to analyze the multivariate time series of fine particles and to relate the various PM diameters with mortality. Bayesian latent factor models are common in health research (e.g., Wang and Wall, 2003; Biggeri et al., 2005; Lui et al., 2005) and in multivariate time series analysis (Aguilar et al., 1998; West and Harrison, 1997). The dynamic factor model reduces the dimension of the multivariate pollution time series to a small number of temporally-correlated latent time series factors, which are used as predictors of mortality. In our setting, the natural ordering of the diameters suggests an extension of the usual dynamic factor model that makes use of the similarity between adjacent diameters. This extension of the usual dynamic factor model borrows strength across diameters, thereby reducing variability in the latent factors and improving our model for mortality.

A common limitation of observational studies of the effects of air pollution on human health is that ambient concentrations are used as surrogates for personal exposures, and a single value is used to represent the exposure of each individual in a geographic region. However, for a given ambient concentration level, personal exposure can vary widely across individuals with different activity patterns. Assuming a common value of exposure holds for the entire population of individuals leads to the "ecological fallacy" (Selvin, 1958; Wakefield and Shaddick, 2005), and can result in bias.

We propose a new method for mitigating this bias. Although direct measurements of personal exposures are not available, personal exposure is simulated using the Stochastic Human Exposure and Dose Simulator (SHEDS) model, developed by Burke et al. (2001). This stochastic model uses information about human activity patterns, census data, and daily diurnal pollution cycles to estimate the daily population exposure distribution. Meshing the exposure simulator into our Bayesian framework allows us to investigate the association between personal exposure and mortality, and to compare these results to the association between mortality and ambient concentration.

The paper proceeds as follows. Section 2 describes the Fresno data set. The dynamic latent factor model relating ambient concentrations with daily mortality is developed in Section 3. Details of the SHEDS simulator are provided in Section 4, along with a model for relating SHEDS output with mortality via the integrated population relative risk. Section 6 analyzes the effect of ambient concentrations on mortality and Section 7 demonstrates the effects of using simulated exposure, rather than ambient concentrations, as predictors of mortality. Section 8 concludes.

2 Description of the data

The city of Fresno is a located in central California. Its metropolitan area has approximately one million people. Particulate matter was monitored at a single monitoring station in downtown Fresno, located in zip code 93726 about 1km east of Highway 41 (Figure 1), a residual area in central Fresno. There are major highways to the east and west of the station and Fresno Yosemite International Airport is roughly two miles east of the station.

Hourly pollution data for 2001 and 2002 were downloaded from the University of Maryland's Supersites Integrated Relational Database System (`http://supersitesdata.umn.edu`). The sizes of PM we consider are PM_{10}, $PM_{2.5}$, and 17 ranges of fine PM with diameters ranging from 0.01 to 0.40 μm. For the diameters less than 0.40μm, the data are recorded as number per cubic centimeter, rather than mass concentration. The weather covariates temperature, relative humidity, wind speed, and wind direction are recorded daily.

Daily natural mortality counts for 18 zip codes in the Fresno metropolitan area (Figure

1) for 2001 and 2002 were obtained from the California Center for Health Statistics. We consider only the elderly Caucasian population because the elderly ($>$ 64 years old) are especially susceptible to the effects of PM and Caucasians account for more than 85% of the deaths in the region.

3 A model for estimating the effect of ambient PM levels on mortality

3.1 A latent factor model for ambient PM levels

Let y_{dt} be the observed average daily concentration for diameter d at day t, $d = 1, ..., D$ and $t = 1, ..., T$. The vectors of observations for each diameter are standardized to have mean zero and unit variance. The dynamic Bayesian factor analysis model (Aguilar et al., 1998; West and Harrison, 1997) assumes the mean of y_{dt} is a linear combination of $J \leq D$ independent latent time series, i.e.,

$$\begin{aligned} y_{dt} &= \theta_{dt} + \epsilon_{dt}, & (1) \\ \theta_{dt} &= \mu_d + \sum_{j=1}^{J} w_{dj} f_{jt}, & (2) \end{aligned}$$

where θ_{dt} is the true concentration for diameter d at time t, μ_d is the intercept for diameter d, w_{dj} is the loading of the j^{th} factor for diameter d, f_{jt} is the value of the j^{th} latent factor at time t, and $\epsilon_{dt} \sim N(0, \sigma_d^2)$, independent across d and t.

We model the latent factors $\mathbf{f}_j = (f_{j1}, ..., f_{jT})'$ as independent, stationary time series with mean zero and lag-h covariance functions $\rho_j(h)$. In dynamic factor analysis, vague priors are typically selected for the loadings. However, in our setting, the model can be improved by exploiting the natural ordering of the diameters. Let $\mathbf{w}_j = (w_{1j}, ..., w_{Dj})$, the vector of loadings for the j^{th} factor, have prior mean zero and $cov(w_{d_1 j}, w_{d_2 j}) = \gamma_j(|d_1 - d_2|)$. This prior is used to borrow strength across adjacent diameters.

The induced prior covariance of two true concentrations $\theta_{d_1 t}$ and $\theta_{d_2 t+h}$ is

$$\text{Cov}\,(\theta_{d_1 t}, \theta_{d_2 t+h}) = \sum_{j=1}^{J} \gamma_j(|d_1 - d_2|)\rho_j(h). \tag{3}$$

That is, the covariance between a pair of true concentrations is the sum of the products of the autocovariance functions for time and diameter of the J latent time series. At this level of generality, the factor analysis model results in a non-separable (between diameter and time) covariance function.

In the analysis of Section 5, the latent time series and loading vectors are taken to be independent AR(1) processes. Also, temperature, humidity, carbon monoxide level, wind speed, wind direction, and an indicator of weekday are included as predictors of the factors. That is,

$$f_{jt} \sim N(\rho_j f_{jt-1} + x_t' \mathbf{b}_j, \tau_j^2) \text{ and } w_{jd} \sim N(w_{jd-1}, \delta_j^2) \tag{4}$$

where $\rho_j \in (-1, 1)$, x_t is the vector of explanatory variables on day t, and $\mathbf{b}_j$ are the corresponding regression coefficients for factor j.

Restrictions are necessary to ensure that the factors and loadings are identified. To identify the scale, we fix the conditional variances of the factors to be one, that is $\tau_j^2 \equiv 1$

for all j. Following Aguilar and West (2000), for the first factor, we constrain the loading for the smallest diameter w_{11} to be one. For the second factor, we set the loading for the smallest diameter w_{21} to zero and, to make identification as strong as possible, restrict the loading for the largest diameter w_{2P} to be one. The third loading vector has $w_{31} = w_{3P} = 0$ and $w_{32} = 1$, and so on.

3.2 Relating the latent factors with mortality

Including all $D = 17$ diameters as predictors of mortality leads to substantial multicollinearity and misleading estimates. Clearly, some form of dimension reduction is needed. The factor analysis model of Section 3.1 represents the ambient concentrations as a linear combinations of the latent time series, $\mathbf{f}_1, ..., \mathbf{f}_J$. To circumvent multicollinearity, the latent factors are used as predictors of mortality. This results in supervised factor analysis, in that the loadings and latent factors are chosen not only to provide a reasonable fit to the observed ambient concentrations, but also to help explain the health outcome.

The number of natural deaths on day t for elderly Caucasians, M_t, has a Poisson distribution with the expected number of deaths on day t equal to

$$\eta_t = \exp\Big(\mu + \mathbf{x}_t\boldsymbol{\beta} + \sum C_j(t - l_j)\alpha_j\Big), \tag{5}$$

where μ is the intercept; $\mathbf{x}_t$ is the vector of confounders; $C_j(t - l_j)$ the lag l_j ambient level of pollutant j; and $\boldsymbol{\beta}$ and $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_J)$ are the vectors of regression parameters. We include the pollutants $PM_{2.5}$, PM_{10}, carbon monoxide, and the latent factors $\mathbf{f}_1, ..., \mathbf{f}_J$ and we include the long-term trend, temperature, humidity, and an indicator of weekday as confounding variables in $\mathbf{x}_t$. Following Dominici et al. (2002), we use a natural spline function of time with ten degrees of freedom per year to capture long-term trends in mortality. Temperature and humidity are also smoothed with natural spline function with ten degrees of freedom per year. The effect of the number of degrees of freedom of the spline function on the estimates of the effects of PM on mortality is investigated in Section 6.

To complete the Bayesian model, we specify priors for the hyperparameters. The variance parameters σ_d^2 and δ_{dj}^2 are given independent InvGamma(0.01,0.01) priors (parameterized to have mean 1, variance 100) and the ρ_j are given Uniform(-1,1) priors. The intercepts μ_j and the regression parameters $\mathbf{b}_j$, $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}$ have vague normal priors with mean zero and variance 100.

We also perform a spatial analysis that allows the effect of each pollutant to vary spatially. The daily expected number of deaths in each zip code has the same form as (5), except that each zip code is given its own regression parameters. The number of natural deaths on day t in zip code z, M_{tz}, has a Poisson distribution with the expected number of deaths on day t equal to

$$\eta_{tz} = \exp\Big(\mu_z + \mathbf{x}_t\boldsymbol{\beta}_{\cdot z} + \sum C_j(t-l_j)\alpha_{jz}\Big), \tag{6}$$

where $\boldsymbol{\beta}_{\cdot z}$ and $\alpha_{\cdot z} = (\alpha_{1z}, ..., \alpha_{Fz})'$ are the vectors regression parameters for zip code z.

The intercepts μ_z are given flat priors and the vectors of parameters for each covariate $\boldsymbol{\beta}_{j\cdot} = (\beta_{j1}, ..., \beta_{j18})$ and $\boldsymbol{\alpha}_{j\cdot} = (\alpha_{j1}, ..., \alpha_{j18})$ are spatially smoothed using the conditional autoregressive (CAR) prior of Besag et al. (2001). The CAR prior for the vector $\boldsymbol{\theta}$ can be defined through the conditional distributions of θ_z given the value of $\boldsymbol{\theta}$ at the remaining zip codes. If $\boldsymbol{\theta} \sim CAR(\gamma^2)$, $\theta_z|\theta_{(z)} \sim N(\bar{\theta}_z, \gamma^2/m_z)$, where $\bar{\theta}_z$ is the mean of $\boldsymbol{\theta}$ at region z's m_z

neighbors, and γ^2 is a variance parameter that controls the amount of spatial smoothing of $\boldsymbol{\theta}$.

4 A model for estimating the effect of individual exposure on mortality

4.1 Simulating exposure using SHEDS

A full description of the SHEDS model can be found in Burke et al. (2001); a brief summary is given below. The SHEDS model simulates personal exposure for a set of I hypothetical individuals chosen to represent the study population in terms of age, gender, employment, housing type, and smoking status. Each day, the activities of the hypothetical individuals are generated by randomly selecting a diary from EPA's Consolidated Human Activity Database (CHAD). CHAD contains personal diaries of over 22,000 individuals from exposure studies conducted around the US. The diaries describe the activity pattern of the individual throughout the day and are selected to match the hypothetical individual based on personal characteristics, housing type, season, day of the week, and average daily temperature.

SHEDS considers nine microenvironments: outdoors, vehicles, residences, offices, schools, stores, restaurants, bars, and other indoor environments. The average exposure for individual i on day t, $E_i(t)$, is the sum of the exposures accumulated in the nine microenvironments. Let $C_{mh}(t)$ and $T_{imh}(t)$ be the PM concentration and time spent, respectively, in microenvironment m for individual i during hour h. Then, the average daily exposure is

$$E_i(t) = \frac{1}{24}\sum_{h=1}^{24}\sum_{m=1}^{9} E_{imh}(t) = \frac{1}{24}\sum_{h=1}^{24}\sum_{m=1}^{9} C_{mh}(t)T_{imh}(t). \tag{7}$$

The PM concentration for microenvironment m is assumed to be a linear function of the ambient concentration, i.e., $C_{mh}(t) = a_m + b_m C_{amb,h}(t)$ where $C_{amb,h}(t)$ is the known ambient PM level for hour h on day t. The coefficients for the residential microenvironment are modelled using a mass balance equation and have the form

$$a_{res} = \frac{E_{smk}N_{cig} + E_{cook}t_{cook} + E_{other}}{(ach + k)V} \quad \text{and} \quad b_{res} = \frac{P \times ach}{ach + k}, \tag{8}$$

where P = penetration factor; k = deposition rate; ach = air exchange rate; E_{smk} = emission rate for smoking; N_{cig} = number cigarettes smoked; E_{cook} = emission rate for cooking; t_{cook} = time spent cooking; E_{other} = emission rate for other sources; and V = residential volume.

Exposure simulation via SHEDS requires reliable prior information for the parameters in the mass balance equation for residential concentration and the linear equations for non-residential concentrations. The priors for several parameters for residential concentration are based on exposure studies conducted in California and are given in Table 1. The priors for the remaining parameters are taken from Burke et al. (2001). Since no data are available for non-ambient source exposure (e.g., smoking and cooking) for diameters other than PM_{25}, we only consider exposure from ambient sources.

The two-stage priors for the SHEDS parameters (e.g., in Table 1) reflect both the inherent variability from person-to-person and day-to-day, and our uncertainty about the hyperparameters that control the variability distributions. To include both types of randomness in our simulation, each day we simulate the exposure of M independent populations of size I. The parameters for all individuals within the same simulated population have the same draw from the uncertainty distribution, but vary from person-to-person based on the variability distribution.

The model described above could theoretically be incorporated into a fully-Bayesian analysis. However, exploratory analysis suggests that the daily exposure distributions can be approximated by normal distributions; the level 0.05 Kolmogorov-Smirnov test of normality rejects the hypothesis that the exposure distribution follows a normal distribution for less than 1% of the simulated distributions for each of the PM diameter analyzed with SHEDS in Section 7. Therefore, we assume the model

$$E_i(t) \sim Normal\left(m(t), v(t)\right), \tag{9}$$

Uncertainly in the exposure distribution on day t is captured by the priors for mean $m(t)$ and variance $v(t)$. Let $\{\bar{x}_1(t), ..., \bar{x}_M(t)\}$ and $\{s_1^2(t), ..., s_M^2(t)\}$ be the sample means and variances, respectively, of the M simulated exposure distributions for day t. Then $m(t)$ is given a normal prior with mean and variance matching the sample mean and sample variance of $\{\bar{x}_1(t), ..., \bar{x}_M(t)\}$, and $v(t)$ is given a gamma prior with mean and variance matching the sample mean and sample variance of $\{s_1^2(t), ..., s_M^2(t)\}$. Combining the distributions of human activity, hourly PM levels, and priors for SHEDS parameters into priors for $m(t)$ and $v(t)$ dramatically reduces the computational burden while still reflecting uncertainly in exposure distribution and allowing the exposure distribution to be updated by the mortality data.

4.2 Relating exposure to mortality

Each day, the exposure distribution is estimated using SHEDS for $PM_{2.5}$ and several diameters of ultrafine particles suggested by the dynamic factor analysis in Section 5. Let $E_{fi}(t)$ be the exposure to pollutant f for individual i on day t. Since mortality is rare, the distribution of the event of individual i dying on day t can be approximated with Poisson distribution with expected value

$$\exp\left(\mu + \mathbf{x}_t\boldsymbol{\beta} + \sum_{f=1}^{F} E_{fi}(t-l)\tilde{\alpha}_f\right), \tag{10}$$

where $\tilde{\alpha}_1$, ..., $\tilde{\alpha}_F$ are the regression parameters associated with the simulated exposures.

Following Richardson et al. (1987), the population average risk on day t is

$$\eta_t = \exp\left(\mu + \mathbf{x}_t\boldsymbol{\beta}\right) \prod_{f=1}^{F} \int \exp(E_f(t-l_f)\tilde{\alpha}_f)p(E_f(t-l_f))dE_f(t-l_f), \tag{11}$$

where the exposure distribution on day t for pollutant f has density $p(E_f(t))$. Given η_t, M_t follows a Poisson(η_t) distribution, independent across t.

We assume that $E_f(t)$ follows a normal distribution with mean $m_f(t)$ and variance $v_f(t)$, where

$$\begin{aligned} m_t &\sim N(\bar{m}_f(t), \tau_f^2(t)) && (12) \\ v_t &\sim Gamma(a_f(t), b_f(t)) && \\ & && (13) \end{aligned}$$

Under the normal model for the population exposure distributions, the population average risk conditional on $(\mu_f(t), \tau_f^2(t))$ can be written in closed form as

$$\eta_t | m_f(t), \tau_f^2(t) = \exp\left(\mu + \mathbf{x}_t\boldsymbol{\beta} + \sum_{f=1}^{F} m_f(t - l_f)\tilde{\alpha}_f + \frac{1}{2}\sum_{f=1}^{F} v_f(t - l_f)\tilde{\alpha}_f^2\right). \tag{14}$$

Comparing (14) with the expected number of deaths as a function of ambient pollution levels in (5) shows that the effect of ambient concentration equals the effect of personal exposure if each personal exposure equals the ambient concentration or if $\tilde{\alpha}_f = 0$, i.e., the pollutant has no effect on mortality. Also, the effect of the population mean exposure m_f equals the effect of personal exposure if $v_f = 0$. Therefore, we expect the bias caused by using a single ambient concentration to represent the exposure of each individual in the population to be large if the variation in exposure within the population is large and the pollutant has a large effect on mortality.

When fitting these models to the Fresno data, we choose between models using the deviance information criterion (DIC) of Speigelhalter et al. (2002), defined as $DIC = \bar{D} + P_D$ where $\bar{D}$ is the posterior mean of the deviance, $P_D = \bar{D} - \hat{D}$ is the effective number of parameters, and $\hat{D}$ is the deviance evaluated at the the posterior mean of the parameters in the likelihood. The model's fit is measured by $\bar{D}$, while the model's complexity is captured by P_D. Since modelling mortality is the primary focus, only the likelihood associated with mortality is used in computing DIC, and the likelihood associated with the ambient concentrations is ignored. Models with smaller DIC are preferred.

For each model, three chains of length 50,000 were sampled and the first 10,000 samples were discarded as burn-in. MCMC convergence was monitored for several parameters using the Gelman-Rubin diagnostic (Gelman and Rubin, 1992). All MCMC simulations are carried out in `WinBUGS` (`http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml`)

5 Dynamic factor analysis of fine PM diameters

To understand the relationships between the fine PM diameters less than 0.4μm and their dependencies with meteorological variables, we temporality set aside the mortality data and fit the latent factor model of Section 3.1. A principal components analysis shows that the first three principal components explain 95% of the variance in the daily ambient concentrations, therefore we present results of the three-factor model.

Figure 2 plots the posterior medians of the loadings. The loadings vary smoothly from one diameter to the next, in part due to the prior for the loadings which encourages borrowing strength across nearby diameters. DIC favors the model that smooths the loadings across diameter (DIC = -1030) over the model with vague independent normal priors for the loadings (DIC = -913). The three factors roughly corresponds to diameters less than 0.02μm and greater than 0.30μm (factor 1), diameters between 0.02 and 0.10μm (factor 2), and diameters greater than 0.10μm (factor 3). The breakpoint between the second and third factors at 0.1μm coincides with the definition of an ultrafine particle. These results are similar to the principle components analysis, indicating the identifiability constraints described in Section 3.1 are not affecting the posteriors of the loadings.

Table 2 shows the median and 95% intervals for the predictors of the three factors. The first factor is positively associated with relative humidity and wind speed. Factor 2 is associated with wind speed and carbon monoxide. This suggests that the second factor represents the contribution of mobile sources. Also, of the three factors, the second has the weakest temporal correlation, i.e., smallest posterior median of ρ. The only significant predictor of the third factor is weekday. It is somewhat surprising that carbon monoxide is only associated with the second factor which loads high for diameters greater than 0.10μm, however, low correlation between ultrafine particles and carbon monoxide was also found by Timonen et al. (2004).

6 Analysis of the effect of ambient PM on mortality

6.1 Model building

We begin analyzing mortality by excluding the exposure simulator and directly using the pollutants and the latent factors f_{jt} as predictors, as in Section 3.2. To build our model, we perform a series of simple Poisson regressions to select the appropriate lags for the pollution covariates. Each regression uses only mortality data from zips codes 93710, 93726, and 93703, the zip codes nearest the monitoring station (Figure 1), and included the weather and temporal trend covariates along with the observed daily average ambient concentration for one pollutant at one lag. Separate regressions were performed using all PM diameters and carbon monoxide and each lag from 0 to 14 days. With the exception of the smallest ultrafine particles, the lag most strongly associated with mortality was one day lag. For the smallest ultrafine particles, the lag most strongly associated with mortality was two days. Therefore we use two day lags of the first latent factor and one day lag of all other pollutants as the predictors of mortality.

Using these lags, we fit three models to determine which zip codes to include in our analysis, each with a separate intercept and smooth long-term trend curve for each zip code. The model with constant regression parameters across zip codes ($DIC = 20410$, $p_D = 41.6$) was selected over the spatial model with CAR random effects for each zip code's regression parameters described in Section 3.2 ($DIC = 20460$, $p_D = 86.7$) and the model with different the regression parameters for each zip code with independent vague normal priors ($DIC = 20538$, $p_D = 142.0$). Therefore, we pool all mortality data and allow the effect of each pollutant to be the same in each zip code.

To investigate the influence of the smoothness of the long-term trend and weather covariates, Figure 3 plots the relative risks for the pollution covariates for various of degrees of freedom for the spline smoothers. For each fit, mortality is pooled across all 18 zip codes and the factors are fixed at their posterior medians under the 20 degrees of freedom model. The relative risks for all six pollutants remain fairly constant after 20 degrees of freedom. Therefore, our choice of degrees of freedom does not appear to be affecting our results.

6.2 Results

The posterior medians of the factor loadings for the three PM factors are plotted in Figure 4a. The medians are slightly different under this supervised factor analysis that makes use of both PM and mortality data than under the PM-only analysis in Section 5 (Figure 2). For example, the loadings for diameters between 0.05 and 0.10μm for factor 1 are closer to zero and the loadings for diameters between 0.10 and 0.20μm are larger for factor 2. Generally speaking, the three factors divide the 17 diameters into three bins: diameters less than 0.02μm (factor 1), diameters between 0.02 and 0.10μm (factor 3), and diameters greater than 0.10μm (factor 2).

The posteriors of the relative risks for the regression parameters are plotted in Figure 4b. The 95% interval for each relative risk covers one, expect for the 95% interval for carbon monoxide which is (1.002, 1.101). The latent factor with the strongest association with mortality is factor 1, which has an 0.949 posterior probability of being greater than one and loads high for diameters less than 0.02μm. Therefore, there is some evidence that the ambient concentrations of the smallest ultrafine particles are associated with mortality. The second factor which is the most significant contributor for diameters greater than 0.10μm also has median relative risk greater than one, but the posterior probability of this relative risk being greater than one is only 0.775.

7 Analysis of the effect of personal exposure on mortality

As described in Section 1, using a single value of ambient PM levels to represent the entire population's exposure as in Section 6's analysis can lead to bias. In this section, we use the SHEDS simulator to compare the effects of ambient pollution levels and the effects of simulated personal exposure. The exposure distribution is simulated for four PM diameters: 0.02, 0.05, and 0.20 μm, and $PM_{2.5}$. To estimate the exposure distributions, for each day, we simulated the exposure for $M = 20$ populations of $I = 100$ elderly Caucasians in the Fresno area.

Figure 5 illustrates the variability and uncertainty in the exposure distribution for $PM_{2.5}$ on January 1, 2001. For each simulated population, a normal density is fit by matching the first two moments of the sample distribution. For each of the 20 simulated populations, there is substantial variability in personal exposure within the population. The average ambient concentration on this day was 176 $\mu g/m^3$, and personal exposure generally ranges from 50 to 200 $\mu g/m^3$. There is also considerable uncertainly about the true exposure distribution, as evident by the differences in the fitted densities. For the 20 populations, the mean exposure ranges from 91 to 132 $\mu g/m^3$ and the standard deviation of exposure ranges from 20 to 41 $\mu g/m^3$.

The ratio the daily population mean exposure and the average daily ambient concentration varies considerably across diameter. Table 3 shows that the ratio of exposure to ambient concentration is smaller for ultrafine particles than for $PM_{2.5}$. This is due in large part to the small penetration factor and large deposition rate for ultrafine particles (Table 1). Table 3 also shows that the ratio of exposure to ambient concentration depends on the season and the day of the week. For each particle size, people are exposed to the largest proportion of the ambient concentration on summer weekends, times when people are generally more active and spend more time outdoors.

To determine the effect of incorporating the exposure simulator into our analysis, Table 4 gives the results using both ambient levels and simulated exposure as predictors of mortality. Each model includes smooth functions for long-term trend, temperature, and humidity, a weekday indicator, and the two-days lag ambient levels of PM_{10} and carbon monoxide. The first model also includes the daily average ambient level of the three fine diameters and

PM_{25}. As in Section 6, the 95% intervals of the relative risks for the fine diameters all cover one. The diameter with the largest median relative risk is 0.20μm, which has posterior 95% interval (0.997, 1.132).

The second model uses the simulated exposure distributions for the three fine particles and PM_{25}, as described in Section 4.2. Although the model using personal exposure has slightly more effective parameters ($p_D = 13.1$) than the model using ambient PM levels ($P_D = 11.0$), the exposure model is preferred in terms of DIC ($DIC = 2986.96$) to the ambient PM level model ($DIC = 2990.25$). The relatives risk with medians near one under the ambient concentration model also have the relative risks near one under the exposure model. As mentioned in Section 4.2, the bias caused by using ambient concentrations as a surrogate for personal exposure is likely to be small for pollutants with no association with mortality.

The effect of including the exposure simulator is slightly larger for particles with diameter 0.20μm. The posterior probability of the relative risk being greater than one for this diameter increases from 0.969 under the model using ambient PM levels to 0.998 under the model using simulated exposure. Although there is not a dramatic effect on the posterior median of the relative risk, the exposure model provides stronger evidence that particles with diameter 0.20μm are associated with mortality.

8 Discussion

This paper presents a supervised dynamic factor model to relate a multivariate time series of pollutants with daily mortality. The model extends the usual dynamic factor model by borrowing strength across neighboring diameters, which leads to an improvement in DIC. Under this model, none of the latent factors for fine ambient PM levels are significantly associated with mortality. Section 4 analyzes mortality using simulated exposure. This model accounts for both the variability and uncertainty in the population exposure distributions and improves DIC over the model with ambient PM levels. Using this model, exposure to PM with diameter 0.20μm was shown to be significantly associated with natural-cause mortality in elderly Caucasians in the metropolitan Fresno area.

The dynamic factor model proposed in Section 3 could be adapted to model a single pollutant that is repeatedly measured at multiple locations. In this spatiotemporal setting, each site would be assigned a vector of loadings and the loadings for each latent factor would be smoothed with a spatial prior. This would result in a flexible spatiotemporal model that could be fit to non-stationary and non-separable data, as shown in (3).

Section 7 demonstrates the effect of using simulated personal exposure, rather than ambient concentration, to model daily mortality. The benefits of using SHEDS may be more pronounced if applied on the national level because this would account for the different activity patterns and diurnal PM cycles in different regions of the country. Clearly, this would provide computational challenges and may require even further simplification of the SHEDS model developed in Section 4. Also, a multivariate SHEDS model that computes exposures to multiple pollutants simultaneously to capture correlations between exposures may be possible.

References

Aguilar O, Huerta G, Prado R, West M (1998). Bayesian inference on latent structure in time series. *Bayesian Statistics*, **6**, 1–16.

Aguilar O, West M (2000). Bayesian dynamic factor models and portfolio allocation. *Journal of Business and Economic Statistics*, **18**, 338–357.

American Thoracic Society, and Bascom R (1996a). Health effects of outdoor air pollution, Part 1. *American Journal of Respiratory and Critical Care Medicine*, **153**, 3–50.

American Thoracic Society, and Bascom R (1996b). Health effects of outdoor air pollution, Part 2. *American Journal of Respiratory and Critical Care Medicine*, **153**, 477–498.

Besag J, York JC, Mollié A (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, **43**, 1–59.

Biggeri A, Bonannini M, Catelan, D, Divino F, Dreassi E, Lagazio C (2005). Bayesian ecological regression with latent factors: atmospheric pollutants emissions and mortality for lung cancer. *Environmental and Ecological Statistics*, **12**, 397–409.

Burke JM, Zufall MJ, Ozkaynak H (2001). A population exposure model for particulate matter: case study results for $PM_{2.5}$ in Philadelphia, PA. *Journal of Exposure Analysis and Environmental Epidemiology*, **11**, 470–489.

de Hartog JJ, Hoek G, Peters A, Timonen KL, Ibald-Mulli A. Brunekreff B, Heinrich J, Tiittanen P, van Wijnen JH, Kreyling W, Kulmala M, Pekkanen J (2003). Effects of fine and ultrafine particles on cardiorespiratory symptoms in elderly subjects with coronary heart disease. *Am. J. Epidemiol*, **157**, 613–623.

Dockery DW, Pope CA III, Xu X, Spengler JD, Ware JH, Fay ME, Ferris BG Jr, and Speizer FE. An association between air pollution and mortality in six U.S. cities. New England Journal of Medicine, 1993, 329:1753-1759.

Dominici F, Daniels M, Zeger SL, Samet JM (2002). Air pollution and mortality: estimating regional and national dose-response relationships. *J. Amer. Statist. Assoc.*, **97**, 100–111.

Gelman A, Rubin DB (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, 457–511.

Liu X, Wall MM, Hodges JS (2005). Generalized spatial structural equation modeling. *Biostatistics*, **6**, 539-557.

Murray DM, Burmaster DE (1995). Residential air-exchange rates in the United States: Empirical and estimated parametric distributions by season and climate region. *Risk Analysis*, **15**, 459–465.

Özkaynak H, Xue J, Spengler J, Wallace L, Pellizzari E, Jenkens P (1996a). Personal exposure to airborne particles and metals: Results from the particle TEAM study in Riverside, California. *Journal of Exposure Analysis and Environmental Epidemiology*, **6**, 57–78.

Özkaynak H, Xue J, Weker R, Koutrakis P, Spengler J (1996b). The particle team (PTEAM) study: Analysis of the data. Final Report, Vol. III. EPA/600/R-95/098. US EPA Office of Research and Development, Washington, DC 20460.

Pekkanen J, Peters A, Hoek G, Tiittanen P, Brunekreef B, de Hartog J, Heinrich J, Ibald-Mulli A, Kreyling WG; Lanki T, Timonen KL, Vanninen E (2002). Particulate air pollution and risk of ST-segment depression during repeated submaximal exercise tests among subjects with coronary heart disease: the exposure and risk assessment for fine and ultrafine particles in ambient air (ULTRA) Study. *Circulation*, **106**, 933–938.

Richardson S, Stucker I, Hémon D (1987). Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. *International Journal of Sociology*, **16**, 111–120.

Schwartz J (1994). Air pollution and daily mortality: a review and meta analysis. *Environmental Research*, **64**, 36–52.

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002). Bayesian measures of model complexity and fit (with discussion and rejoinder) *J. Roy. Statist. Soc., Ser. B*, **64**, 583-639.

Thomas D, Stram D, Dwyer J (1993). Exposure measurement error: influence on exposure-disease relationships and methods of corrections. *Annual Review Public Health*, **14**, 69–93.

Timonen KL, Hoek G, Heinrich J, Bernard A, Brunekreef B, de Hartog J,Hmeri K, Ibald-Mulli A, Mirme A, Peters A, Tiittanen P, Kreyling WG, Pekkanen J (2004). Daily variation in fine and ultrafine particulate air pollution and urinary concentrations of lung Clara cell protein CC16. *Occupational and Environmental Medicine*, **61**, 908–914.

Vette AF, Rea AW, Lawless PA, Rodes CE, Evans G, Highsmith VR, Sheldon, L (2001). Characterization of indoor-outdoor aerosol concentration relationships during the Fresno PM exposure studies. *Aerosal Science and Technology*, **34**, 118–126.

Wakefield J, Shaddick G (2005). Health-exposure modelling and the ecological fallacy. *Biostatistics*, **1**, 1–19.

Wang F, Wall MM (2003). Generalized common spatial factor model. *Biostatistics*, **4**, 569–582.

West M, Harrison PJ (1997). *Bayesian Forecasting and Dynamic Models*, 2nd edn. Springer–Verlag: New York.

Wichmann HE, Spix C, Tuch T, Wolke G, Peters A, Heinrich J, Kreyling WG, Heyder J (2000). Daily mortality and fine and ultrafine particles in Erfurt, Germany part I: role of particle number and particle mass. *Res Rep Health Eff Inst*, **98**, 5–86.

Table 1: Prior distributions for selected SHEDS parameters. "Tri(a,b,c)" refers to the triangular density with minimum a, mode b, and maximum c. The references are a=Murray and Burmaster (1995), b=Vette et al. (2001), and c=Özkaynak et al. (1996a,b).

Parameter	Category	Variability	Uncertainty dist. of μ	Uncertainty dist. of σ
Air exchange rate[a]	Winter	LogN(μ, σ^2)	N(-0.68, 0.10)	Tri(0.55,0.65,0.75)
	Spring	LogN(μ, σ^2)	N(-0.48, 0.10)	Tri(0.57,0.67,0.77)
	Summer	LogN(μ, σ^2)	N(-0.05, 0.10)	Tri(0.81,0.91,1.01)
	Fall	LogN(μ, σ^2)	N(-0.88, 0.10)	Tri(0.61,0.71,0.81)
Penetration	0.02μm[b]	N(μ, σ^2))	N(0.70, 0.10)	N(0.08, 0.01)
	0.05μm[b]	N(μ, σ^2)	N(0.65, 0.10)	N(0.08, 0.01)
	0.20μm[b]	N(μ, σ^2)	N(0.65, 0.10)	N(0.08, 0.01)
	$PM_{2.5}$[c]	N(μ, σ^2)	N(1.00, 0.10)	N(0.08, 0.01)
Deposition	0.02μm[b]	N(μ, σ^2)	N(2.50, 0.50)	N(0.50, 0.10)
	0.05μm[b]	N(μ, σ^2)	N(0.80, 0.10)	N(0.20, 0.04)
	0.20μm[b]	N(μ, σ^2)	N(0.50, 0.05)	N(0.20, 0.04)
	$PM_{2.5}$[c]	N(μ, σ^2)	N(0.27, 0.07)	N(0.10, 0.02)

Table 2: Median and 95% intervals for the regression parameters of the three factors. "Primary diameters" refers to the diameters with the highest loadings for each factor (Figure 2). $\rho \in (-1, 1)$ measures the strength of temporal association from each factor.

	Factor 1	Factor 2	Factor 3
Primary diameters	$< 0.02\mu$m	$> 0.1\mu$m	> 0.02 and $< 0.1\mu$m
Temperature	0.05 (-0.07, 0.17)	-0.20 (-0.31, -0.08)	0.07 (-0.05, 0.19)
Humidity	0.14 (0.00, 0.27)	-0.09 (-0.22, 0.05)	-0.18 (-0.31, -0.04)
Wind speed	0.05 (0.05, 0.16)	-0.15 (-0.26, -0.04)	-0.04 (-0.15, 0.07)
Carbon monoxide	-0.06 (-0.16, 0.05)	0.39 (0.27, 0.51)	-0.02 (-0.14, 0.09)
Weekday	0.01 (-0.08, 0.09)	-0.04 (-0.12, 0.05)	0.09 (0.00, 0.18)
ρ	0.56 (0.48, 0.64)	0.37 (0.28, 0.47)	0.49 (0.40, 0.59)

Table 3: Mean (sd) of the daily ratios of the population mean exposure (averaged over all uncertainly runs) to daily average ambient concentration by season, weekday, and diameter.

Diameter	0.02μm	0.05μm	0.20μm	$PM_{2.5}$
Winter, weekday	0.27 (0.026)	0.36 (0.078)	0.44 (0.140)	0.65 (0.004)
Winter, weekend	0.23 (0.029)	0.33 (0.024)	0.40 (0.021)	0.66 (0.006)
Spring, weekday	0.30 (0.047)	0.39 (0.032)	0.45 (0.018)	0.64 (0.003)
Spring, weekend	0.27 (0.039)	0.36 (0.022)	0.42 (0.017)	0.65 (0.006)
Summer, weekday	0.34 (0.034)	0.46 (0.020)	0.52 (0.015)	0.76 (0.002)
Summer, weekend	0.38 (0.068)	0.49 (0.042)	0.53 (0.027)	0.79 (0.008)
Fall, weekday	0.23 (0.023)	0.32 (0.016)	0.38 (0.017)	0.62 (0.002)
Fall, weekend	0.27 (0.036)	0.35 (0.024)	0.41 (0.026)	0.64 (0.013)

Table 4: DIC and the median (95% interval) for the relative risks for the pollution covariates for models using the ambient levels of PM_{10} and carbon monoxide along with ambient concentration or summaries of the personal exposure distribution for the fine PM diameters. The relative risks are the relative risk due to a one standard deviation in ambient concentration or population mean exposure.

Diameter	Ambient Concentration	Exposure Distribution
$DIC(p_D)$	2990.25 (11.0)	2986.96 (13.1)
0.02μm	1.01 (0.97, 1.04)	1.00 (0.97, 1.04)
0.05μm	0.99 (0.94, 1.03)	0.99 (0.95, 1.03)
0.20μm	1.07 (1.00, 1.13)	1.08 (1.01, 1.14)
$PM_{2.5}$	0.95 (0.87, 1.04)	0.96 (0.91, 1.05)
PM_{10}	0.98 (0.91, 1.06)	0.97 (0.91, 1.05)
carbon monoxide	1.05 (0.99, 1.11)	1.04 (0.99, 1.10)

Figure 1: Map of Fresno, CA. The monitoring station is located in zip code 93726 about 1km east of Highway 41.

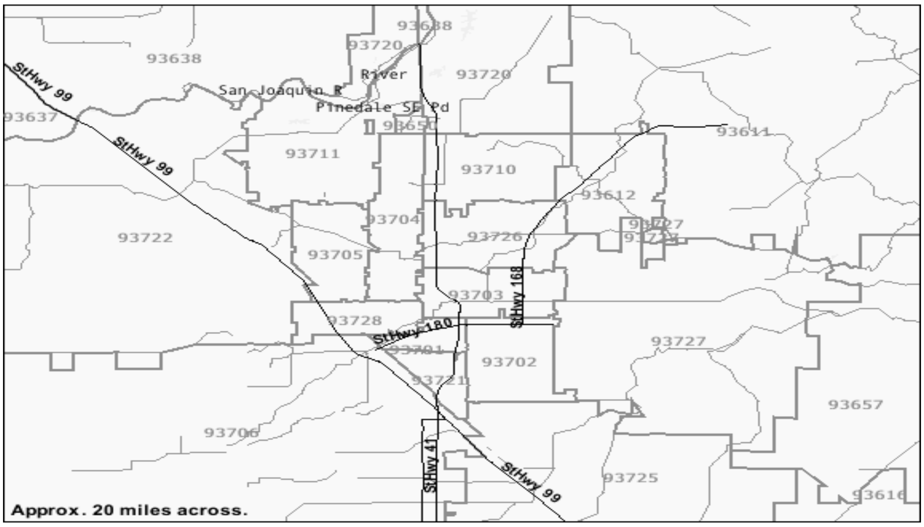


Figure 2: Posterior medians of the loadings of the dynamic factor model for the fine PM diameters.

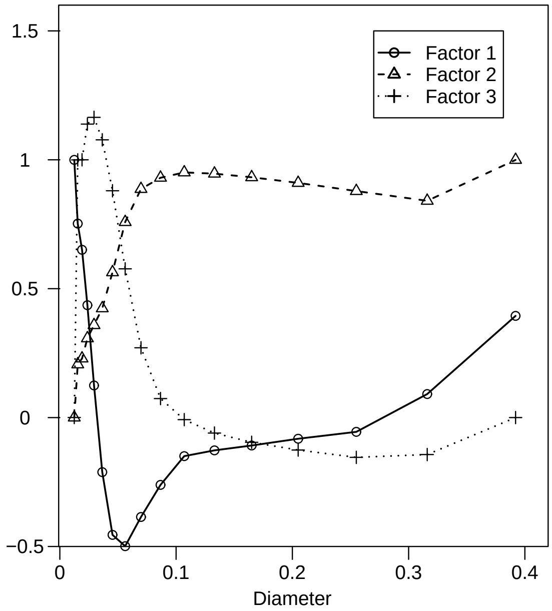


Figure 3: Plots of the median relative risk for the pollutants against the degree of freedom in the spline smooth for the seasonality/weather covariates.

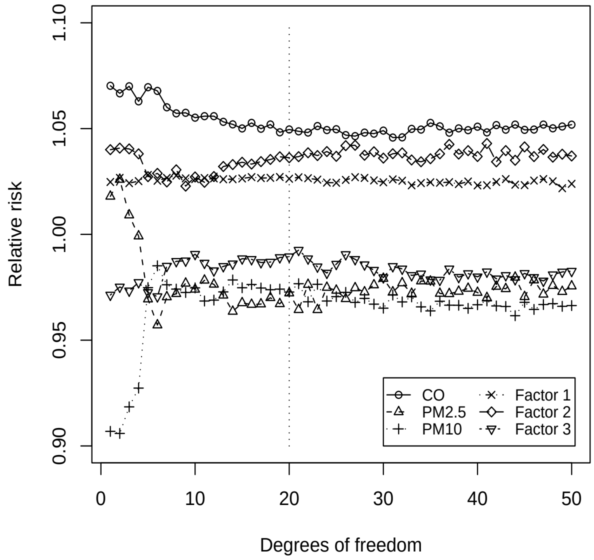
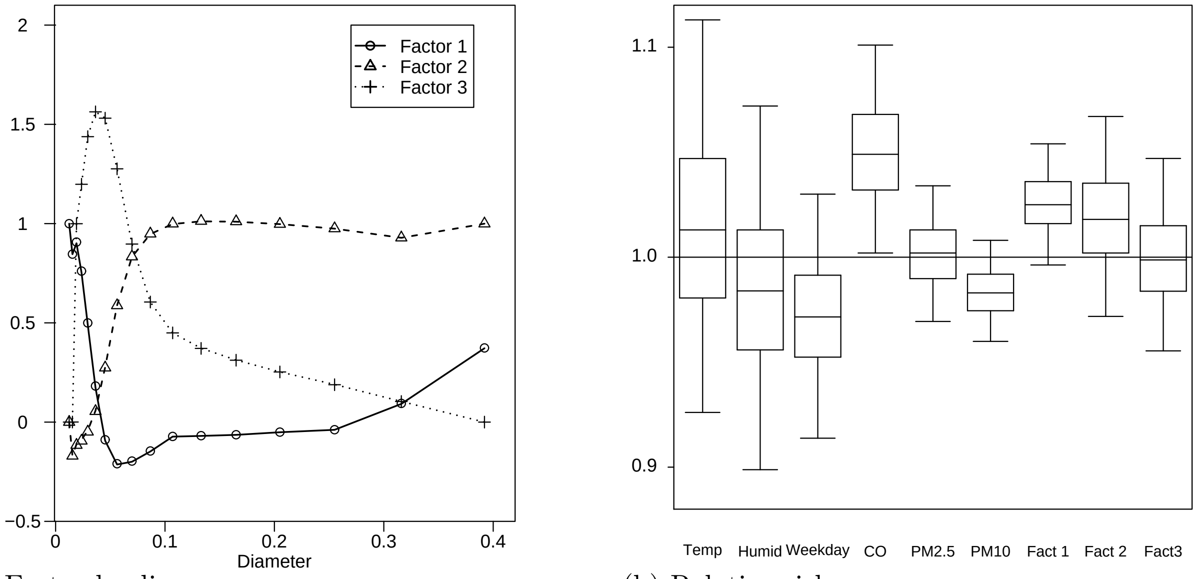


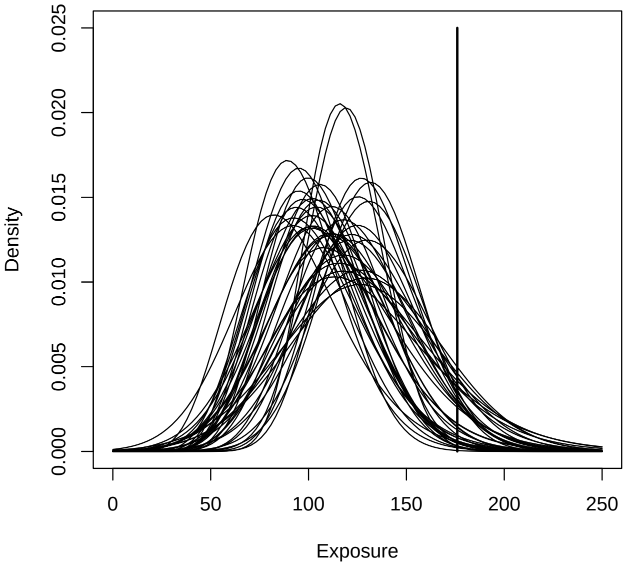
Figure 4: Summary of the analysis of the effects of ambient pollution levels on mortality. Panel (a) shows the posterior medians of the factor loadings. Panel (b) shows the posteriors of the relative risks of the predictors of mortality. The whiskers of the boxplots represent 95% intervals and the relative risks represent a 10 μm increase in $PM_{2.5}$ and PM_{10} and a one standard deviation increase in the other covariates.



(a) Factor loadings

(b) Relative risks

Figure 5: Fitted density curves for 20 simulated $PM_{2.5}$ exposure distributions on January 1, 2001. The vertical line is the ambient concentration.


Density
Exposure