12-15-2008

Bayesian Model Averaging for Clustered Data: Imputing Missing Daily Air Pollution Concentration

Howard H. Chang
Johns Hopkins University, hhchang@jhsph.edu

Francesca Dominici
The Johns Hopkins Bloomberg School of Public Health, fdominic@jhsph.edu

Roger D. Peng
Johns Hopkins University, rpeng@jhsph.edu

Bayesian Model Averaging for Clustered Data: Imputing Missing Daily Air Pollution Concentrations

Howard H. Chang[1], Francesca Dominici[1], and Roger D. Peng[1]

Abstract

The presence of missing observations is a challenge in statistical analysis especially when data are clustered. In this paper, we develop a Bayesian model averaging (BMA) approach for imputing missing observations in clustered data. Our approach extends BMA by allowing the weights of competing regression models for missing data imputation to vary between clusters while borrowing information across clusters in estimating model parameters. Through simulation and cross-validation studies, we demonstrate that our approach outperforms the standard BMA imputation approach where model weights are assumed to be the same for all clusters. We then apply our proposed method to a national dataset of daily ambient coarse particulate matter ($PM_{10-2.5}$) concentration between 2003 and 2005. We impute missing daily monitor-level $PM_{10-2.5}$ measurements and estimate the posterior probability of $PM_{10-2.5}$ nonattainment status for 95 US counties based on the Environmental Protection Agency's proposed 24-hour standard.

KEY WORDS: Bayesian model averaging; Missing data; Imputation; Air pollution; Particulate matter

[1] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health.
516 North Wolfe Street Baltimore MD 21205 USA
Phone: 410-955-3067; Fax: 410-955-0958

1 INTRODUCTION

The presence of missing observations is a challenge in statistical analysis especially when data are clustered. Examples of clustered data include repeated measurements of subject-specific outcomes in longitudinal analysis, geographical locations in multi-site air pollution studies, and strata in sample survey. Excluding clusters with incomplete data often results in considerable loss in sample size. When the missing data do not arise completely at random (Little and Rubin 1987), appropriate interpretations of analysis results also become difficult.

In this paper, we consider a missing data imputation problem for clustered data where the observations are partially or completely missing in some clusters. After imputing the missing data, standard complete-data techniques can be used in subsequent analysis and final parameter estimation can also reflect uncertainty in the imputation through data augmentation (Tanner and Wong 1987) or multiple imputation techniques (Rubin 1996).

However, the success of any imputation method relies on specifying a model that best describes the conditional distribution of the missing data given the observed data. Often several plausible imputation models are available for prediction and missing data imputation. Bayesian model averaging (BMA) (Raftery et al. 1997; Hoeting et al. 1999) can be used as a powerful prediction tool that accounts for model uncertainty. BMA assigns different weights to each competing model and prediction is obtained by taking a weighted average of predictions from the competing models. Even though BMA leads to larger prediction variance, it avoids the potential bias associated with choosing only a single model for prediction.

The typical application of BMA to clustered data determines model weights by comparing how data from *all* clusters fit each competing model. Therefore model uncertainty reflects model fit "globally" across clusters and missing data occurring in different clusters are imputed by averaging competing models using the same set of weights. However different

clusters may favor different sets of models possibly due to some unmeasured cluster-specific characteristics. Since the model optimal for imputation can differ between clusters, it may be beneficial to impute missing data within a cluster using only models that are "locally" optimal.

In this paper we develop a BMA-based approach for imputing missing data in a clustered design. Our approach allows the weights of competing models for missing data imputation to differ between clusters while borrowing information across clusters in estimating model parameters. To accomplish Bayesian predictive inference and posterior simulation we consider implementations using Markov chain Monte Carlo (MCMC) and a more computationally efficient approach.

The methodological development of this paper is motivated by the problem of imputing missing daily concentration of ambient particulate matter (PM). Toxicological and epidemiological studies have consistently found that increased level of ambient PM is associated with increased risks of adverse health outcomes (Pope and Dockery 2006; Schwarze et al. 2006). Ambient PM can be characterized into two size fractions, fine and coarse, that represent distinct pollutant mixtures of different sources and properties (Wilson and Suh 1997). Protecting public health from exposure to inhalable coarse particles ($PM_{10-2.5}$) of size between 2.5 and 10 μm aerodynamic diameter has endured considerable controversy and understanding the toxicity of coarse PM remains a top PM research priority (Committee on Research Priorities for Airborne Particulate Matter 2004).

Because of the lack of a national monitoring network for $PM_{10-2.5}$, the Environmental Protection Agency (EPA) measures $PM_{10-2.5}$ by taking the difference between daily PM_{10} (10 μm or less in aerodynamic diameter) and daily $PM_{2.5}$ (2.5 μm or less in aerodynamic diameter) concentrations at monitors that are physically located in the same place (collocated monitor pairs). This leads to a significant loss in sample size because we can only calculate

$PM_{10-2.5}$ (1) at collocated monitor pairs of PM_{10} and $PM_{2.5}$ and (2) on days when both PM_{10} and $PM_{2.5}$ were measured at the collocated monitor pairs. Under this approach, daily measurements of $PM_{2.5}$ and PM_{10} at monitors that are not collocated are excluded.

In our application, we define a cluster as the time series of PM measurements at a particular monitor. We apply the proposed method to impute the missing daily PM_{10} or $PM_{2.5}$ concentrations needed to calculate $PM_{10-2.5}$ measurements whenever only one of PM_{10} or $PM_{2.5}$ value is available. Specifically we substantially increase the number of $PM_{10-2.5}$ measurements in two ways. First, we increase the number of collocated monitor pairs by imputing the entire time series of $PM_{2.5}$ when a daily time series of PM_{10} is available at any given monitor that is not collocated with a $PM_{2.5}$ monitor and vice versa. Second, we increase the number of daily measurements at each collocated monitor pair by imputing each missing daily PM_{10} measurement at a given day when $PM_{2.5}$ is available and vice versa.

While a national network for coarse PM is currently planned in 2011, the data are unlikely to be available for large-scale scientific studies in the near future. Increasing the number of days and locations having $PM_{10-2.5}$ concentration has important implications. First, the imputed concentrations can help us determine nonattainment status for $PM_{10-2.5}$ in counties without collocated monitor pairs. Also, the level of $PM_{10-2.5}$ exhibits higher spatial heterogeneity compared to $PM_{2.5}$ and PM_{10}. Therefore, population average exposure to $PM_{10-2.5}$ may not be well-characterized by distant monitors. A more extensive $PM_{10-2.5}$ monitoring network allow us to better estimate the health effects of $PM_{10-2.5}$ with smaller exposure measurement error (Zeger et al. 2000; Sheppard 2005).

The remainder of the article is organized as follows. Section 2 describes the proposed method and in Section 3 we evaluate its performance by simulation studies. In Section 4 we describe the PM data and create a national dataset of daily $PM_{10-2.5}$ concentrations where

the missing data are imputed accounting for all the sources of uncertainty. From the imputed national data we estimate the nonattainment status of 95 US counties for the period 2003 to 2005. Finally, discussion and future work appear in Section 5.

2 METHODS

2.1 Bayesian Model Averaging for Clustered Data

Let $\boldsymbol{y}^m$ be the vector of n_m observations from cluster m for $m = 1, \ldots, N$. For example, $\boldsymbol{y}^m$ may represent the vector of n_m daily levels of PM at monitor m. Within each cluster, assume we can predict $\boldsymbol{y}^m$ using K competing linear regression models each having X_k^m as the $n_m \times p_k$ design matrix, where $k = 1, \ldots, K$. For example, X_k^m may include columns of time series data of temperature or other pollutants that are potential predictors of $\boldsymbol{y}^m$.

In a standard BMA setting, we introduce *global* model indicators M_k where $M_k = 1$ denotes that model k is chosen for prediction in all clusters. Each model k has a corresponding vector of parameters $\boldsymbol{\theta}_k$ that contains both the vector of regression coefficients $\boldsymbol{\beta}_k$ and the residual variance σ_k^2 such that $[\boldsymbol{y}^m \,|\, \boldsymbol{\theta}_k, M_k] \sim \text{MVN}(\, X_k^m \boldsymbol{\beta}_k,\, \sigma_k^2 \, \text{I}_{n_m})$. Given a prior distribution $[\boldsymbol{\theta}_k \,|\, M_k]\,[M_k]$, the posterior is $[\boldsymbol{\theta}_k \,|\, \boldsymbol{y}] \propto \prod_{m=1}^{N} [\boldsymbol{y}^m \,|\, \boldsymbol{\theta}_k, M_k] \times [\boldsymbol{\theta}_k \,|\, M_k]$ where $\boldsymbol{y} = (\boldsymbol{y}^{m=1}, .., \boldsymbol{y}^{m=N})$. Hence for any given model M_k, we use data from *all* clusters to estimate $\boldsymbol{\theta}_k$. To impute a missing observation on day t^* in cluster m, we calculate the posterior predictive distribution $[y_{t^*}^m \,|\, \boldsymbol{y}] = \sum_{k=1}^{K} [y_{t^*}^m \,|\, M_k, \boldsymbol{y}]\,[M_k \,|\, \boldsymbol{y}]$ where $[y_{t^*}^m \,|\, M_k, \boldsymbol{y}] = \int [y_{t^*}^m \,|\, \boldsymbol{\theta}_k, M_k]\,[\boldsymbol{\theta}_k \,|\, M_k, \boldsymbol{y}]\, d\boldsymbol{\theta}_k$. Under this approach, we impute missing observations by assuming each model M_k has a posterior weight, $[M_k|\boldsymbol{y}]$, which is assumed to be the same across all clusters.

Our main contribution is to introduce a *local* BMA-based approach that relaxes this

assumption by allowing the posterior model weights to differ among clusters. Let M_k^m be the cluster-specific model indicator where $M_k^m = 1$ denotes that model k is chosen in cluster m. For N clusters and K competing models, there are $1/N^K$ different model combinations. To facilitate notation, we introduce a reparameterization of the cluster-specific model indicators. Let S_k denote the subset of cluster indices, $\{1, \ldots, N\}$, that choose model k. If no cluster chooses model k, then S_k is empty. Unlike the *global* BMA approach, here the posterior distribution of $\boldsymbol{\theta}_k$ can be defined as $[\boldsymbol{\theta}_k \,|\, \boldsymbol{y}] = \int [\boldsymbol{\theta}_k \,|\, S_k, \boldsymbol{y}]\, [S_k \,|\, \boldsymbol{y}]\, dS_k$. More specifically, we estimate $[\boldsymbol{\theta}_k \,|\, \boldsymbol{y}]$ accounting for cluster-level model uncertainty by marginalizing over all combinations of S_k. Therefore in estimating $\boldsymbol{\theta}_k$ we borrow information only across clusters that have a posterior probabilities for model k that is larger than zero.

Figure 1 illustrates our notation with 3 clusters and 3 competing models where clusters 1 and 2 choose Model 1 ($S_1 = \{1, 2\}$), no cluster chooses Model 2 ($S_2 = \emptyset$), and cluster 3 chooses Model 3 ($S_3 = \{3\}$). Therefore data from clusters 1 and 2 contribute to the estimation of $\boldsymbol{\theta}_1$, no data will contribute to the estimation of $\boldsymbol{\theta}_2$, and data from cluster 3 will contribute to the estimation of $\boldsymbol{\theta}_3$.

Using the *local* BMA for missing data imputation, the posterior predictive distribution of $y_{t^*}^m$ is given by $[y_{t^*}^m \,|\, \boldsymbol{y}] = \sum_{k=1}^K [y_{t^*}^m \,|\, M_k^m, \boldsymbol{y}]\, [M_k^m \,|\, \boldsymbol{y}]$. Imputation is carried out by sampling repeatedly from the posterior predictive distribution. Given J samples from $[\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, S_1, \ldots, S_K \,|\, \boldsymbol{y}]$, to obtain a particular j^{th} sample of $y_{t^*}^m$, we first determine which $S_k^{(j)}$ does cluster m belong to. Since $S_k^{(j)}$ specifies which imputation model and the corresponding parameter vector to use, we can then draw a sample of $y_{t^*}^m$ according to the data likelihood specified by imputation model k.

2.2 Markov Chain Monte Carlo Computation

The unknown parameters in our local BMA approach include $\boldsymbol{\theta}_k$, $k = 1, \ldots, K$, and $P(M_k^m = 1) = \delta_k^m$, $k = 1, .., K$, subject to $\sum_{k=1}^K \delta_k^m = 1$ for all m. Parameter estimation is accomplished within a Bayesian framework and the complete-data likelihood is given by

$$L(\boldsymbol{\theta}_k, M_k^m) \propto \prod_{m=1}^{N} \left[\boldsymbol{y}^m \,|\, \boldsymbol{\theta}_k, M_k^m\right] \times \prod_{k=1}^{K} \left[\boldsymbol{\theta}_k \,|\, S_k\right] \times [S_1, \ldots, S_K]. \tag{1}$$

More specifically, $[S_1, \ldots, S_K]$ represents the joint prior distribution on the groups of clusters that follow each model k. To reflect a lack of prior knowledge in model choice, we assume a priori that $P(M_k^m = 1) = \delta_k^m = 1/K$ for all m and k.

The prior distribution $[\boldsymbol{\theta}_k|S_k]$ represents the prior for the parameters in model k. We assume that the model parameters are independent between models. Since $\boldsymbol{\theta}_k$ may have different lengths, we adopt the product-space method of Carlin and Chib (1995) to avoid the change-of-dimension problem in posterior sampling. This "parameter saturation" approach requires sample of $\boldsymbol{\theta}_k$ for every k at every iteration. Then the fixed-length parameter vector $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ can be used to update each cluster-specific model indicator. For each model k, we specify a proper and non-informative prior for $\boldsymbol{\theta}_k$ if S_k is not empty and an informative "pseudo-prior" for $\boldsymbol{\theta}_k$ if S_k is empty. Each $\boldsymbol{\theta}_k$ given S_k requires a unique pseudo-prior from which to sample whenever its posterior distribution cannot be learned from the data. Since the sampled value will not be used for imputation at any particular Gibbs iteration, the choice of pseudo-prior distribution is arbitrary. The pseudo-prior can be viewed as a proposal distribution to fill $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ when no data is available for some $\boldsymbol{\theta}_k$'s given S_k. In fact several connections between pseudo-priors and the dimension-matching proposal distribution for reversible jump MCMC have been noted (Godsill 1997). However, pseudo-priors should be dispersed enough such that all potential values in the parameter space are

explored and should propose sensible values to ensure mobility between models.

For linear regression without clustering, Carlin and Chib (1995) suggest using Normal distribution that mimics the marginal posterior distribution of $\boldsymbol{\theta}_k$ as the pseudo-prior for $\boldsymbol{\theta}_k$. Intuitively, this distribution describes the best potential values of $\boldsymbol{\theta}_k$ based on the observed data. Allowing model uncertainty for each cluster presents further complexity since the best and the most likely combination of S_k to estimate $\boldsymbol{\theta}_k$ is not known. In Section 2.3 we describe a faster algorithm to sample from the approximate posterior distributions of all the unknown parameters. We then construct pseudo-priors using samples from the approximate marginal distribution of $\boldsymbol{\theta}_k$.

2.2.1 Computation Details

It follows from the complete-data likelihood in (1) that the full conditional distribution of $\boldsymbol{\theta}_k$ is

$$[\boldsymbol{\theta}_k \mid S_k, \boldsymbol{y}] \propto \begin{cases} \prod_{m \in S_k} [\boldsymbol{y}^m \mid \boldsymbol{\theta}_k, M_k^m] \cdot [\boldsymbol{\theta}_k \mid S_k] & \text{if } S_k \text{ is not empty} \\ [\boldsymbol{\theta}_k \mid S_k] & \text{if } S_k \text{ is empty.} \end{cases} \tag{2}$$

The full conditional distributions for $\boldsymbol{\theta}_k$ follow the standard form (Gelman et al. 1995) when S_k is not empty. At each Gibbs iteration, when at least one cluster chooses model k, we update $\boldsymbol{\theta}_k$ using all data from clusters that choose model k and the non-informative priors. When no cluster chooses model k (S_k empty), we draw $\boldsymbol{\theta}_k$ from a proper pseudo-prior distribution. Therefore the pseudo-priors ensure that every $\boldsymbol{\theta}_k$ is updated at every Gibbs iteration.

Next we update the model indicator of each cluster. For example, if cluster m chooses model $k = 1$ at the j^{th} iteration, the probability of cluster m being updated to model k is

then

$$[\, M_k^m \,|\, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, S_1 \setminus \{m\}, \ldots, S_K \,] = A_k / \sum_{i=1}^{K} A_i,$$

where A_i is the complete data likelihood in (1) evaluated at the current values of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ with model indicators $S_1 = S_1^{(j)} \setminus \{m\}, S_2 = S_2^{(j)}, \ldots, S_i = \{S_i^{(j)}, m\}, \ldots, S_K = S_K^{(j)}$. Since each conditional prior for $\boldsymbol{\theta}_k$ depends on whether or not S_k is empty, the model indicators must be updated sequentially at each iteration.

2.3 Approximate Posterior Inference

We can decompose the posterior distribution as

$$[\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, S_1, \ldots, S_K \,|\, \boldsymbol{y}] = \prod_{k=1}^{K} [\boldsymbol{\theta}_k \,|\, S_k, \boldsymbol{y}] \; [S_1, \ldots, S_K \,|\, \boldsymbol{y}].$$

In this section we describe a faster MCMC approach for sampling from the posterior distribution of all the unknown parameters. We assume that for any cluster, data from all other clusters is negligible in determining its model preference. More specifically we assume that $[M_k^m|\boldsymbol{y}] \approx [M_k^m|\boldsymbol{y}^m]$. For linear regression models with appropriate conjugate priors, $[M_k^m|\boldsymbol{y}^m]$ can be calculated directly (Raftery et al. 1997). Alternately, for clusters with large sample size we can approximate $[M_k^m|\boldsymbol{y}^m]$ using the Bayesian information criterion (BIC) (Schwarz 1978).

Approximate posterior samples of $[\boldsymbol{\theta}_k \,|\, S_k, \boldsymbol{y}]$ are obtained as follows: (1) estimate the marginal cluster-specific model posterior probabilities, $\hat{\delta}_1^m, \ldots, \hat{\delta}_K^m$ by using BIC or by calculating the Bayes factors; (2) for each cluster m, draw J samples of the model indicator from a multinomial distribution with the estimated model probabilities that are assumed to be independent between clusters; (3) for each j and each k determine $S_k^{(j)}$; (4) draw $\boldsymbol{\theta}_k^{(j)}$ from the conditional distribution $[\boldsymbol{\theta}_k \,|\, \{\boldsymbol{y}^{m=1}, \ldots, \boldsymbol{y}^{m=N} : m \in S_k^{(j)}\}]$, using only data from

clusters that choose model k.

Note that this procedure does not consider posterior dependence between model indicators and does not borrow information across clusters in estimating the cluster-specific model probabilities. Unlike a full MCMC approach, once these model probabilities are estimated, they are assumed fixed and not updated.

3 SIMULATION STUDY

We performed simulation studies to assess the performance of the proposed method. Let $\boldsymbol{X}_t^m = (X_{1t}^m\,,\ X_{2t}^m\,,\ X_{3t}^m\,)$ denote the predictors for the t^{th} missing observation in cluster m. We generate the vector $\boldsymbol{X}_t^m$ from a multivariate Normal distribution with mean zero, variances 1, and pairwise correlations ρ. We assume that there are 50 clusters and n observations within each cluster. We then generate $\boldsymbol{y}^m$ from the following two models:

$$\text{Model A:} \quad y_t^m \sim \text{N}\big(\,X_{1t}^m\,,\ 2\big) \quad m = 1, \ldots, 25 \quad t = 1, \ldots, n$$

$$\text{Model B:} \quad y_t^m \sim \text{N}\big(\,X_{2t}^m\,,\ 2\big) \quad m = 26, \ldots, 50 \quad t = 1, \ldots, n$$

We considered four simulation scenarios: (1) $n = 10$, $\rho = 0$; (2) $n = 25$, $\rho = 0.5$; (3) $n = 25$, $\rho = 0.8$; and (4) $n = 50$, $\rho = 0.8$. Three datasets were generated under each scenario. For each simulated dataset, we fitted the data under four competing models with different predictors. We denote the four linear regression models by A, B, C, and D. Each model has an intercept and model A has covariate X_1; model B has covariate X_2; model C has covariate X_3; and model D has covariates X_1 and X_2.

We compared the predictive power of the following three approaches: (1) using only one model at a time (model A, B, C, or D); (2) model averaging using identical weights across

clusters (global BMA) with competing model spaces: E={A, B}, F={A, B, C}, or G={A, B, D} and (3) model averaging using cluster-specific weights (local BMA) with competing model spaces: E={A, B}, F={A, B, C}, or G={A, B, D}.

For model k, let $\boldsymbol{\beta}_k$ be the corresponding vector of regression coefficients and σ_k^2 be the residual variance. Under the local BMA approach, we use the following prior:

$$\left[\boldsymbol{\beta}_k \,|\, \sigma_k^2, S_k\right] \times \left[\sigma_k^2 \,|\, S_k\right] = \begin{cases} \text{MVN}\left(\, 0 \,,\, \sigma_k^2 V_{0k} \,\right) \times \text{IG}\left(\, v_0/2 \,,\, v_0 s_0^2/2 \,\right) & \text{if } S_k \text{ is not empty} \\ \text{MVN}\left(\, \mu_k \,,\, V_k \,\right) \times \text{IG}\left(\, v_k/2 \,,\, v_k s_k^2/2 \,\right) & \text{if } S_k \text{ is empty} \end{cases} \tag{3}$$

where $V_{0k} = \text{diag}\,(100^2, \ldots, 100^2)$ and $v_0 = s_0^2 = 2$. For cluster m, we first estimate δ_k^m by

$$\hat{\delta}_k^m = p(M_k^m \,|\, \boldsymbol{y}^m) = p(\boldsymbol{y}^m \,|\, M_k^m) / \sum_{i=1}^{K} p(\boldsymbol{y}^m \,|\, M_i^m) \quad \text{where}$$

$$p(\boldsymbol{y}^m \,|\, M_k^m) = \frac{\Gamma(\frac{v_0+n}{2})(v_0 s_0^2)^{\frac{v_0}{2}}}{\pi^{\frac{n}{2}}\Gamma(\frac{v_0}{2})|I_n + X_k^m V_{0k} X_k^{m,T}|^{\frac{1}{2}}} \left(v_0 s_0^2 + \boldsymbol{y}^{m,T}(I_n + X_k^m V_{0k} X_k^{m,T})^{-1} \boldsymbol{y}^m \right)^{\frac{-(v_0+n)}{2}} \tag{4}$$

Pseudo-prior parameters, μ_k, V_k, v_k, s_k^2 in equation (3), were estimated by maximum likelihood using 5,000 approximate posterior samples of $\boldsymbol{\theta}_k$ obtained by first generating model indicator independently across clusters as described in Section 2.3.

We considered two implementations of local BMA: (1) full MCMC implementation via the product-space method; (2) approximate posterior sampling where the cluster-specific marginal probabilities are assumed independent. Each chain was initialized with random model indicators, and after 10,000 iterations burn-in, the chains were run for another 10,000 iterations. All Gibbs samplers were implemented in R 2.6.2. Gelman and Rubin convergence diagnostic in the R package CODA was used to evaluate the convergence of our chains.

For each simulated dataset, we then generated a test dataset with identical sample size, clustering, and simulation parameters. We calculated the prediction root mean square error (RMSE) between the true outcome values and their posterior predictive mean using 500

posterior samples of the model parameters.

Figure 3 shows the percent increase in RMSE of different modelling approaches compared to the RMSE if the correct model is used for each cluster. Each connected line represents a replicate dataset and the x-axis indicates model spaces used to make predictions. When considering one model at a time (model spaces A, B, C, D), model D performs the best under all scenarios since it contains both predictors of the two data-generating models. When considering multiple models at a time, overall we do not find the global BMA approach to improve RMSE. When RMSE of individual models differ, the global BMA approach often assigns weight close to 1 to the model with the lowest RMSE due to the large total sample size ($M \times n$).

The lowest relative RMSE was achieved when both true models were included (model space E) using local BMA. The improvement in prediction using local BMA is particularly large when the cluster-size is small (Panel 1) and when the competing models are easily distinguished (Panel 2 where the correlation between covariates is low). Our simulation results also show that including extra models with low predictive power (model space F) or high predictive power (model space G) often slightly increases prediction error. This may be attributed to an increase in model uncertainty and data thinning. Unlike the usual model selection problem, we do not restrict all clusters to follow a particular model. Nonetheless, model averaging at the cluster-level improves prediction considerable compared to using only a single model. We also carried out approximate MCMC sampling algorithm as described in Section 2.3. This approach performs well with only slight increases in prediction error compared to the more computationally intensive MCMC approach.

Figure 4 shows the estimated cluster-specific posterior model probability of model A for one replicate dataset under the four simulation scenarios. Each • denotes the posterior probability estimated via MCMC for a cluster when the competing model space is {A,B}. Each •

denotes the same model probability obtained in the approximate MCMC using equation (4). The correct model has higher posterior probability within each cluster when the correlation between covariates is low and when the sample size is large. There is also evidence that the full MCMC approach assigns more weight to the correct model for imputation compared to the approximate MCMC approach, especially for small cluster size.

4 DATA ANALYSIS

4.1 DATA

Daily average concentrations of PM_{10} and $PM_{2.5}$ for the period 2003 to 2005 were obtained from the EPA's National Air Pollution Monitoring Network. To build our imputation models, we identified 156 collocated $PM_{2.5}$ and PM_{10} monitor pairs in 64 US counties with population greater than 200,000. We restricted our dataset to days with measurements of $PM_{2.5}$ and PM_{10} from at least two monitors in the same county. Table 1 summarizes the median and IQR of $PM_{2.5}$, PM_{10}, and $PM_{10-2.5}$ monitor-level daily concentration across the 156 collocated monitor pairs. We also identified 217 PM_{10} monitors and 226 $PM_{2.5}$ monitors that do not have a collocated monitor counterpart. Figure 2 shows three types of monitoring locations: (1) collocated $PM_{2.5}$ and PM_{10} monitor pair (•), (2) PM_{10} only monitor (∘), and (3) $PM_{2.5}$ only monitor (•) for San Bernardino (CA) and Cook County (IL).

In total, 117,620 PM_{10} and $PM_{2.5}$ daily concentrations are imputed, resulting in an approximate 5-fold increase in the total number of daily monitor-specific $PM_{10-2.5}$ measurements. Table 2 summarizes the number of imputed daily PM measurements. We consider two types of $PM_{10-2.5}$ imputation. First, PM_{10} and $PM_{2.5}$ were often measured on different schedule despite being collocated (Row 2 and 3). Imputing $PM_{2.5}$ values (or PM_{10} values) at a collo-

cated monitor pair, where only PM_{10} (or $PM_{2.5}$) measurement was available, constitutes 18% (23%) of the total imputed values. Second, imputing the entire time-series of $PM_{2.5}$ (PM_{10}) at monitors, where by design only measured PM_{10} (or $PM_{2.5}$), constitutes 26% (33%) of the totall number of PM values imputed (Row 4 and 5). This also provides 443 (217+226) additional imputed collocated monitor pairs to the $PM_{10-2.5}$ network and increases the number of counties with $PM_{10-2.5}$ measurements from 64 to 126.

We modelled values of $PM_{2.5}$ and PM_{10} on the logarithmic scale to account for the strictly positive and right-skewed concentration measurements. Let $PM^m_{2.5,t}$ denote the 24-hour average log $PM_{2.5}$ concentration on day t at the monitor m. Exploratory analyses suggest two predictors for imputing $PM^m_{2.5,t}$: (1) $PM^m_{10,t}$, the log PM_{10} concentration on the same day measured at the same location and (2) $\overline{PM}^m_{2.5,t}$, the averaged log $PM_{2.5}$ concentration from *other* $PM_{2.5}$ monitors within the same county on the same day. Similarly, for imputing $PM^m_{10,}$ we considered $PM^m_{2.5,t}$ and $\overline{PM}^m_{10,t}$ as predictors. Pairwise correlations for the above variables are summarized in Table 3. Daily measurements of PM_{10} and $PM_{2.5}$ are highly correlated when measured at collocated monitor pairs. The median correlation between log-transformed daily PM_{10} and $PM_{2.5}$ concentration at collocated monitors was 0.74. PM_{10} and $PM_{2.5}$ concentrations also showed considerable spatial homogeneity. The median correlation between daily log PM concentration and the average concentration of other monitors in the same county was 0.82 and 0.94 for PM_{10} and $PM_{2.5}$, respectively.

4.2 Cross-validation Analysis

In this section we compare the prediction performance of local BMA using the PM data from the 156 collocated monitor pairs by cross-validation. Three linear regression models

were examined for predicting daily log $PM_{2.5}$ concentration:

$$\begin{aligned}
&\text{Model A:} \quad \mathrm{PM}^m_{2.5,t} \sim \mathrm{N}\left(\beta_{01} + \beta_{11}\mathrm{PM}^m_{10,t},\ \sigma_1^2\right) \\
&\text{Model B:} \quad \mathrm{PM}^m_{2.5,t} \sim \mathrm{N}\left(\beta_{02} + \beta_{12}\overline{\mathrm{PM}}^m_{2.5,t},\ \sigma_2^2\right) \\
&\text{Model C:} \quad \mathrm{PM}^m_{2.5,t} \sim \mathrm{N}\left(\beta_{03} + \beta_{13}\mathrm{PM}^m_{10,t} + \beta_{23}\overline{\mathrm{PM}}^m_{2.5,t},\ \sigma_3^2\right)
\end{aligned} \tag{5}$$

Similarly, for predicting PM_{10} the three models using collocated $PM_{2.5}$ and county-averaged PM_{10} as predictors were examined.

We repeatedly partition the data into subsets such that analysis is only performed on a single subset (training dataset) and the other subsets (testing datasets) are used to assess prediction performance. We generated training and testing datasets as follows. First we randomly selected 100 monitors from the 156 collocated monitor pairs and half of the observations from each monitor were used to fit the missing data imputation models. The remaining half from the 100 monitors is used to calculate RMSE for within-sample prediction that resembles imputation at collocated monitor pairs. PM values from the 57 unselected collocated monitors is used to calculate RMSE for out-of-sample prediction that resembles imputation at monitors that only measured $PM_{2.5}$ or PM_{10}.

Prediction RMSE was calculated between the true log PM level and the posterior predictive mean based on 50,000 approximate posterior samples. To impute missing data at an out-of-sample monitor, we need to define the relation between the new monitor and the within-sample monitors. We assume that the new monitor is a random sample of the monitors in the training dataset. At each iteration, we impute PM values at the new monitor using Model k ($k = A, B, C$) with probability equals to the proportion of within-sample monitors that choose model k.

Figure 5 shows the prediction RMSE for log $PM_{2.5}$ and log PM_{10} at within-sample moni-

tors and out-of-sample monitors using single or multiple prediction models. Using the average measurements from other monitors within the same county on the same day (Model B) performs better compared to using its collocated counterpart PM (Model A). However, having both model A and B in the model space showed considerable improvement in prediction at both within-sample monitors and out-of-sample monitors. While using both predictors simultaneously (Model C) gives the lowest RMSE, enriching the model space further (eg {A,B,C}) improves prediction for $PM_{2.5}$ at within-sample monitors and does not increase prediction error in out-of-sample monitors.

4.3 Analysis of $PM_{10-2.5}$ Nonattainment Status

We applied our local BMA for imputing missing PM_{10} and $PM_{2.5}$ concentrations. We obtained a national networks of daily $PM_{10-2.5}$ for 599 monitors located in 126 counties during the period 2003 to 2005. We estimated $PM_{10-2.5}$ nonattainment status for 95 US counties based on the 24-hour standard proposed by the EPA. (Environmental Protection Agency 2006b). Posterior distribution of missing $PM_{10-2.5}$ concentrations were imputed using the local BMA approach with the three competing prediction models for PM_{10} and $PM_{2.5}$ described in Section 4.2. Marginal posterior model probabilities of the PM imputation models for the 156 collocated sites were first estimated and 50,000 approximate posterior samples were obtained for each missing PM_{10} and $PM_{2.5}$ concentration. Posterior nonattainment probability for each county was defined as the proportion of time the county has any collocated monitor pair with its annual 98^{th} percentile of daily $PM_{10-2.5}$ concentration averaged over 3 years exceeding 70 $\mu g/m^3$. We followed the EPA's data completeness proposal for $PM_{10-2.5}$ where a monitor is eligible if it contains at least 24 measurements per year.

Posterior nonattainment probability for each county and its location are shown in Figure 6 on a gray scale. We use $\square$ to denote the 52 counties that had at least one collocated

monitor pair before imputation. Among these 52, 6 counties ($\times$) were in nonattainment based on the observed $PM_{10-2.5}$ values. We use $\circ$ to denote the 43 counties without eligible collocated monitor pairs before imputation and among these counties, nonattainment status for 36 counties are based only on imputed $PM_{10-2.5}$ values.

We found evidence of nonattainment for the proposed 24-hour $PM_{10-2.5}$ standards for several counties, particularly in the southwest. Empirical evidence shows that PM_{10} levels are higher in western US (Environmental Protection Agency 2006a). Since coarse particulates arise mainly from mechanical process such as dust suspension, grinding and crushing, a larger proportion of PM_{10} mass may be due to $PM_{10-2.5}$ in western counties. Many of the counties with high $PM_{10-2.5}$ nonattainment posterior probability were also designated nonattainment for PM_{10} by the EPA during same study period (Environmental Protection Agency 2007).

5 DISCUSSION

In this paper, we develop a BMA-based missing data imputation approach for clustered data that accounts for differences in the best fitting models among clusters. There are several advantages to carrying out model selection and Bayesian model averaging within each cluster. Our approach classifies clusters according to their model preference to optimize prediction while borrowing information across clusters in estimating model parameters. The traditional BMA approach assumes that model weights are identical across clusters. Therefore clusters with large sample size can dominate the model selection. While the coarse PM analysis served as the motivating example, the proposed method is widely applicable in many settings with clustered data and multiple prediction models. For example, in longitudinal analysis, missing data can be imputed cross-sectionally, longitudinally, or by ad-hoc approaches (e.g. last-value-carried-forward) (Kristman et al. 2005; Wood et al. 2005).

Allowing model uncertainty at the cluster level requires careful choice of competing models. Particularly, parameter estimation variance may increase due to a loss of sample size from thinning data between similar competing models. To avoid these problems, competing models should be chosen by the two heuristic criteria given by Draper (1995): (1) the model should have high posterior probability given non-zero prior probability, and (2) the model should have predictive consequences that differ substantially from other models being considered.

The methodological development in this paper is easily extended to competing generalized linear models by designing appropriate MCMC samplers for the model parameters and pseudo-priors. Also extension to accommodate competing mixed-effect models is currently being investigated to allow cluster-specific coefficients. Specifically, the fixed-effect coefficients and the mean and variance of the random effects can take similar prior and have the same full conditionals as those in the linear regression models. Only when no cluster chooses a particular model will these parameters be drawn from the pseudo-priors. However, at each Gibbs iteration, only a single model has its random effect updated within each cluster and all cluster-specific random effects from other competing models are drawn from the pseudo-priors.

Introducing a model indicator for each cluster increases the computation burden considerably in posterior simulation. In cases with large number of clusters or competing models, selecting pseudo-priors can also be cumbersome and updating each model indicator sequentially requires significant computation. Our simulation studies demonstrated that approximate posterior sampling approach described in Section 2.3 performs well. However a full MCMC approach can be especially beneficial for clusters of small sample size. To more accurately estimate the posterior predictive mean $E[y_{t*}^m \mid \boldsymbol{y}]$, importance reweighting can be used with the above approximate posterior distribution serving as the importance function.

Often the outcome of interest is completely missing in some clusters. In our application of imputing the complete time-series of PM levels at monitors that do not have a collocated counterpart, we naively assume that the monitor with missing outcome represents a random sample of all the collocated monitor pairs. A more sophisticated approach may examine monitor-level variables and determine the most likely model choice for that monitor. For example, when imputing PM_{10} at a monitor that only measured $PM_{2.5}$, we may use the same set of model weights from the nearest $PM_{2.5}$ monitor with a collocated PM_{10} monitor. It is also possible to utilize characteristics of the monitor's location including landuse (residential, commercial, industrial, agricultural) or economic development (urban, suburban, rural) to determine the best imputation models.

In this paper, we choose linear models to impute log PM_{10} and log $PM_{2.5}$ without modelling the temporal trends and spatial correlation explicitly. A multivariate space-time prediction model for PM_{10} and $PM_{2.5}$ (Kibria et al. 2002; Fuentes et al. 2006) offers an alternative approach for imputing $PM_{10-2.5}$. This approach is especially attractive in applications such as estimating health effects when it is necessary to characterize the spatial gradient of $PM_{10-2.5}$ to minimize exposure measurement error. However, this increases model complexity considerably when considering PM monitors of a national coverage and incorporating model uncertainty for competing prediction models becomes challenging. PM_{10} and $PM_{2.5}$ time series exhibit complex missing data structure and temporal trends are likelily to be monitor-specific. Moreover, most $PM_{10-2.5}$ measured at collocated monitor pairs were only available every sixth day and we currently do not plan to impute PM_{10} and $PM_{2.5}$ on days without any PM measurements. Therefore we expect the county-averaged PM_{10} and $PM_{2.5}$ predictor to capture most of the spatial variability of the pollutants for same-day prediction purpose. However, predictive power of additional covariates including day of the week, weather variables, and monitor location characteristics should be explored further.

Acknowledgments

The research described in this article has been funded by grants RD-83241701 from the United States Environmental Protection Agency, ES012054-03 from the National Institute for Environmental Health Sciences, and P30 ES 03819 from the National Institute for Environmental Health Sciences Center in Urban Environmental Health. It has not been subjected to the funding agencies peer and policy review and therefore does not necessarily reflect the views of the agencies and no official endorsement should be inferred.

References

Carlin, B. P. and Chib, S. (1995), "Bayesian Model Choice via Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society. Series B*, 57, 473–282.

Committee on Research Priorities for Airborne Particulate Matter (2004), *Research Priorities for Airborne Particulate Matter: IV. Continuing Research Progress*, National Academies Press.

Draper, D. (1995), "Assessment and propagation of model uncertainty," *Journal of the Royal Statistical Society. Series B*, 57, 45–97.

Environmental Protection Agency (2006a), "Air Quality Criteria for Particulate Matter," .

— (2006b), *National ambient air quality standards for particulate matter: proposed rule.*

— (2007), "Report on Air Quality in Nonattainment Areas for 2003-2005 Covering Ozone, Particulate Matter, Carbon Monoxide, Sulfur Dioxide, Nitrogen Dioxide, and Lead," .

Fuentes, M., Song, H.-R., Ghosh, S. K., Holland, D. M., and Davis, J. M. (2006), "Spatial Association between Speciated Fine Particles and Mortality," *Biometrics*, 62, 855–863.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysis*, Chapman and Hall.

Godsill, S. (1997), "On the relationship between MCMC model uncertainty methods," *J. Comput. Graph. Statist*, 10, 1–19.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–417.

Kibria, B. M. G., Sun, L., Zidek, J. V., and Le, N. D. (2002), "Bayesian Spatial Prediction of Random Space-time Fields With Application to Mapping PM2.5 Exposure," *Journal of the American Statistical Association*, 97, 112–124.

Kristman, V., Manno, M., and Cote, P. (2005), "Methods to account for attrition in longitudinal data: Do they work? A simulation study," *European Jouranl of Epidemiology*, 20, 657–662.

Little, R. J. A. and Rubin, D. B. (1987), *Statistical analysis with missing data*, John Wiley.

Pope, C. A. and Dockery, D. W. (2006), "Health effects of fine particulate air pollution: lines that connect." *J Air Waste Manag Assoc*, 56, 709–742.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179–191.

Rubin, D. (1996), "Multiple imputation after 18+ years (with discussion)," *Journal of the American Statistical Association*, 91, 473–489.

Schwarz, G. (1978), "Estimating the dimension of a model," *The Annals of Statistics*, 6, 461–464.

Schwarze, P. E., Ovrevik, J., Lg, M., Refsnes, M., Nafstad, P., Hetland, R. B., and Dybing, E. (2006), "Particulate matter properties and health effects: consistency of epidemiological and toxicological studies." *Hum Exp Toxicol*, 25, 559–579.

Sheppard, L. (2005), "Acute air pollution effects: consequences of exposure distribution and measurements." *J Toxicol Environ Health A*, 68, 1127–1135.

Tanner, M. A. and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 398, 528–540.

Wilson, W. E. and Suh, H. H. (1997), "Fine particles and coarse particles: concentration relationships relevant to epidemiologic studies." *J Air Waste Manag Assoc*, 47, 1238–1249.

Wood, A. M., White, I. R., Hillsdon, M., and Carpenter, J. (2005), "Comparison of imputation and modelling methods in the analysis of a physical activity trial with missing outcomes," *International Journal of Epidemiology*, 34, 89–99.

Zeger, S. L., Thomas, D., Dominici, F., Samet, J. M., Schwartz, J., Dockery, D., and Cohen, A. (2000), "Exposure measurement error in time-series studies of air pollution: concepts and consequences." *Environ Health Perspect*, 108, 419–426.

Table 1: Median and IQR of PM_{10}, $PM_{2.5}$, and $PM_{10-2.5}$ daily concentration ($\mu g/m^3$) for the period 2003 to 2005 across 156 collocated monitor pairs.

	Median (IQR)
PM_{10}	22.0 (18.0, 28.0)
$PM_{2.5}$	11.0 (8.6, 13.6)
$PM_{10-2.5}$	10.5 (7.4, 14.5)

Table 2: Number of daily PM levels ($\mu g/m^3$) at (1) collocated $PM_{2.5}$ and PM_{10} monitor pair;(2) PM_{10} only monitor; and (3) $PM_{2.5}$ only monitor between 2003 to 2005 (total of 1096 days).

	Monitor type	Number of monitors	PM Measured		Percentage of total imputed PM values[†]
			PM_{10}	$PM_{2.5}$	
1	PM_{10} and $PM_{2.5}$	156	Yes	Yes	0
2	collocated		Yes	No	18
3	monitor pair		No	Yes	23
4	PM_{10} only	217	Yes	No	26
5	$PM_{2.5}$ only	226	No	Yes	33

[†]Total number of daily $PM_{2.5}$ or PM_{10} values imputed is 117,620.
Row 2 and 3 denote situations where PM_{10} and $PM_{2.5}$ were measured on different schedule despite the monitor pairs being collocated.
Row 4 and 5 denote situations where by design all PM_{10} observations were measured without a collocated $PM_{2.5}$ measurement and vice versa.

Table 3: Median and IQR of correlations between daily monitor-level and county-averaged $PM_{2.5}$ and PM_{10} concentrations across 156 collocated monitor pairs. The correlations are calculated on the log-transformed PM concentration. $PM^m_{10,t}$ denotes the log PM_{10} concentration on the day t measured at the monitor m and $\overline{PM}^m_{10,t}$ denotes the average log PM_{10} concentration on day t from $PM_{2.5}$ monitors within county, excluding monitor m. Similar notations apply to $PM^m_{2.5,}$ and $\overline{PM}^m_{2.5,t}$.

	Median (IQR)
cor ($PM^m_{2.5,\,t}$, $PM^m_{10,\,t}$)	0.74 (0.65, 0.81)
cor ($PM^m_{2.5,\,t}$, $\overline{PM}^m_{2.5,\,t}$)	0.94 (0.87, 0.97)
cor ($PM^m_{10,\,t}$, $\overline{PM}^m_{10,\,t}$)	0.82 (0.71, 0.89)

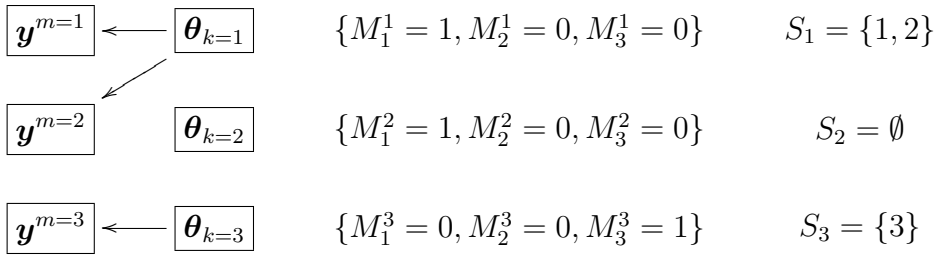
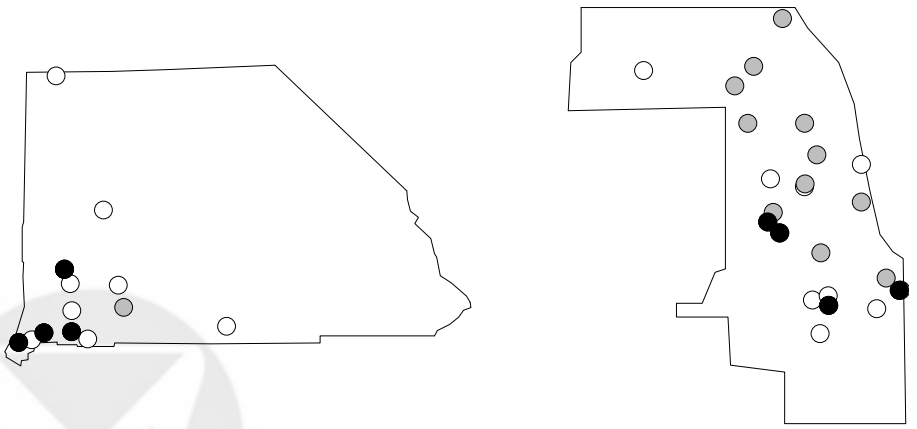


Figure 1: Illustration of notation with N=3 clusters and K=3 competing models. $\boldsymbol{y}^m$ denotes the outcome data for cluster m. $\boldsymbol{\theta}_k$ denotes the model parameter for model k. M_k^m is the model indicator for cluster m and model k. S_k denotes the set of cluster indices for model k. Each arrow denotes model choice by the cluster.



(a) (b)

Figure 2: Locations of PM monitors in (a) San Bernardino County, CA and (b) Cook County, IL: (1) collocated PM_{10} and $PM_{2.5}$ monitor pair (•), (2) PM_{10} only monitor (○), and (3) $PM_{2.5}$ only monitor (•).

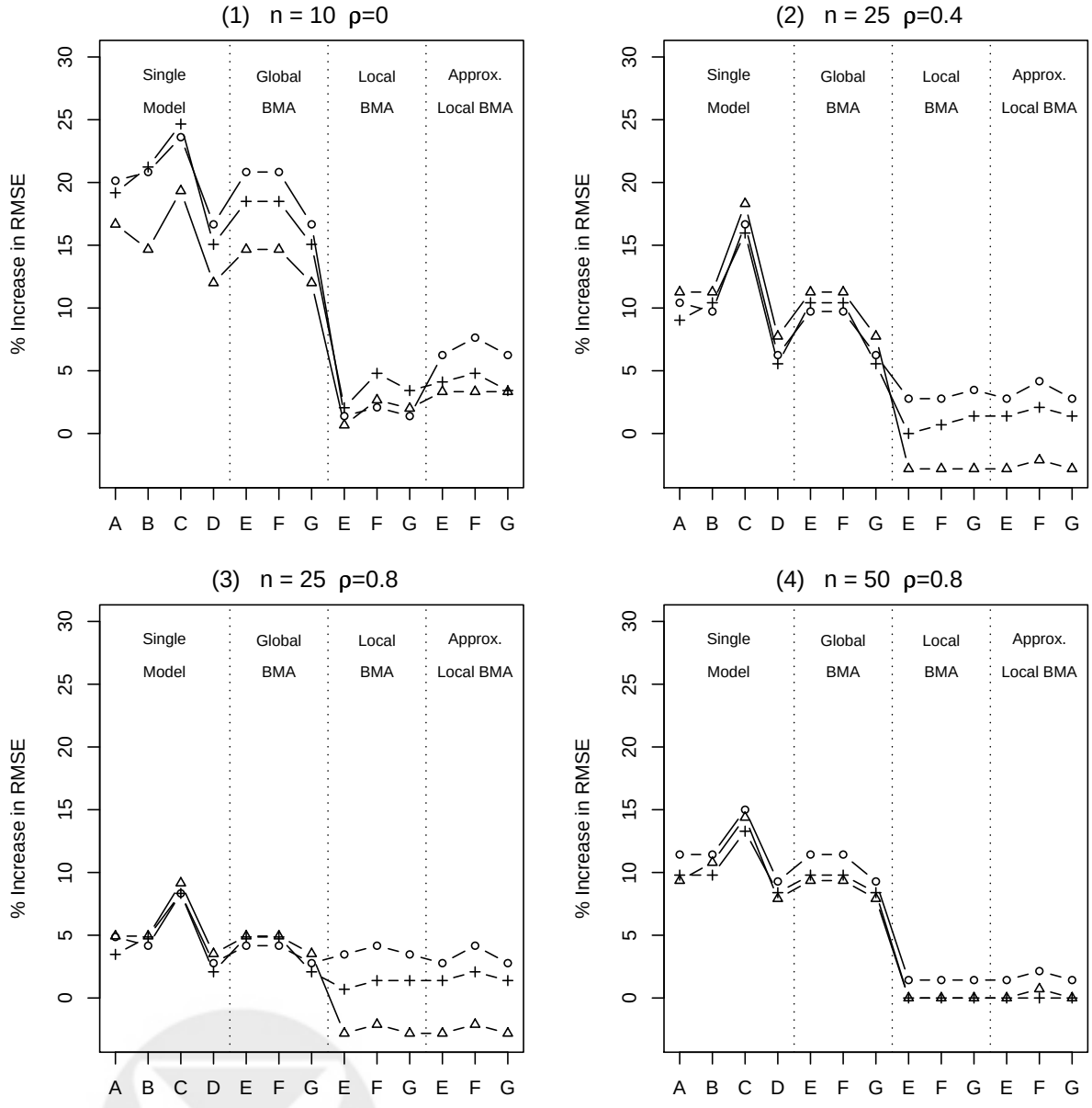


Figure 3: Percent increase in root mean-squared error (RMSE) compared to the RMSE if the true model for each cluster is used. Each connected line represents a replicate dataset. The x-axis indicates model(s) used to make predictions: single model: A, B, C, D and multiple models: E = {A, B}, F = {A, B, C}, G = {A, B, D}. Global BMA: model weights are identical across clusters; Local BMA: cluster-specific model weights estimated via MCMC; Approx. Local BMA: cluster-specific model weights approximated as in Section 2.3

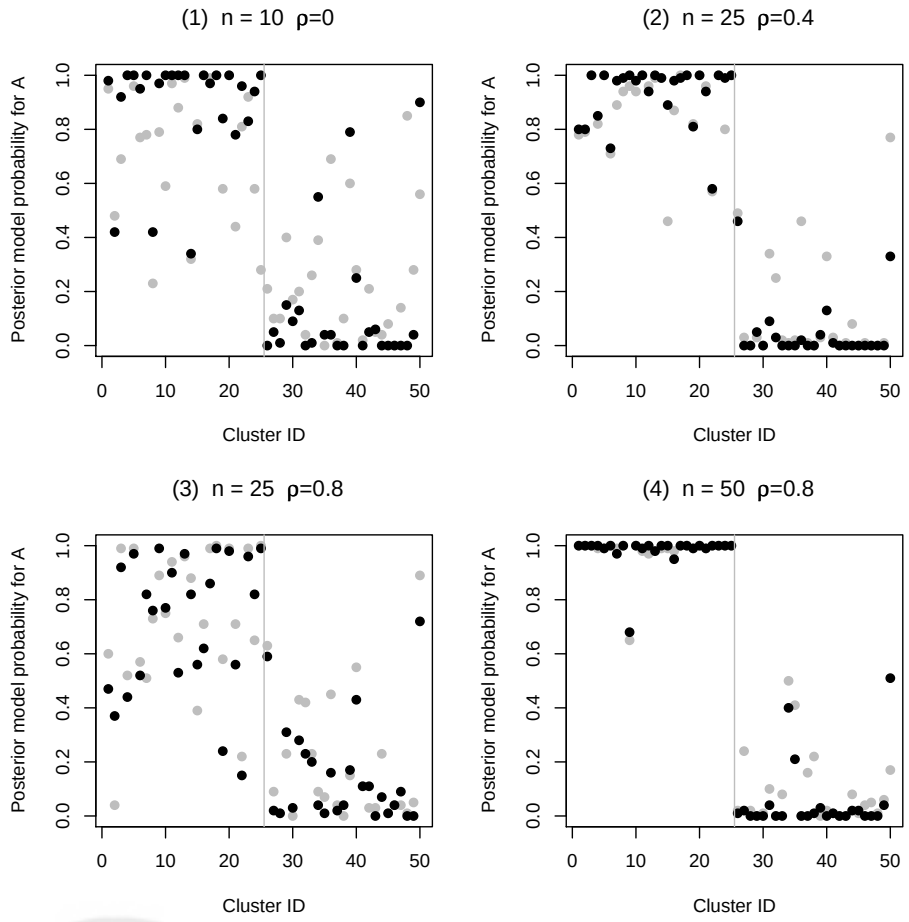


Figure 4: Posterior model probability for choosing model A under local BMA with model space {A, B}. Legend: • = posterior model probability estimated via MCMC and • = posterior model probability estimated by equation (4) for approximate MCMC. For each scenario, data were generated under model A for the first 25 clusters and under model B for the second set of 25 clusters.

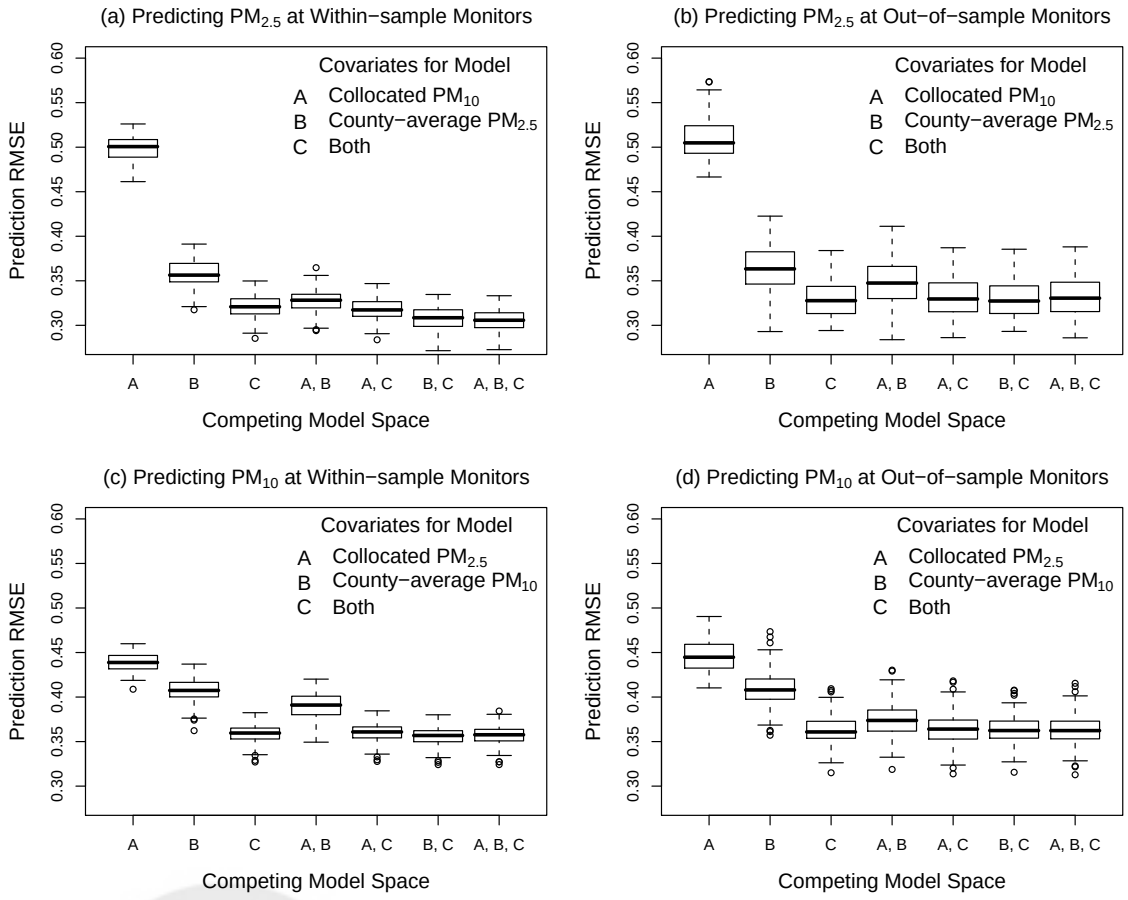


Figure 5: Boxplots of prediction root mean-squared error (RMSE) for log $PM_{2.5}$ and log PM_{10} concentrations at within-sample monitors and out-of-sample monitor for 100 cross-validation simulation.

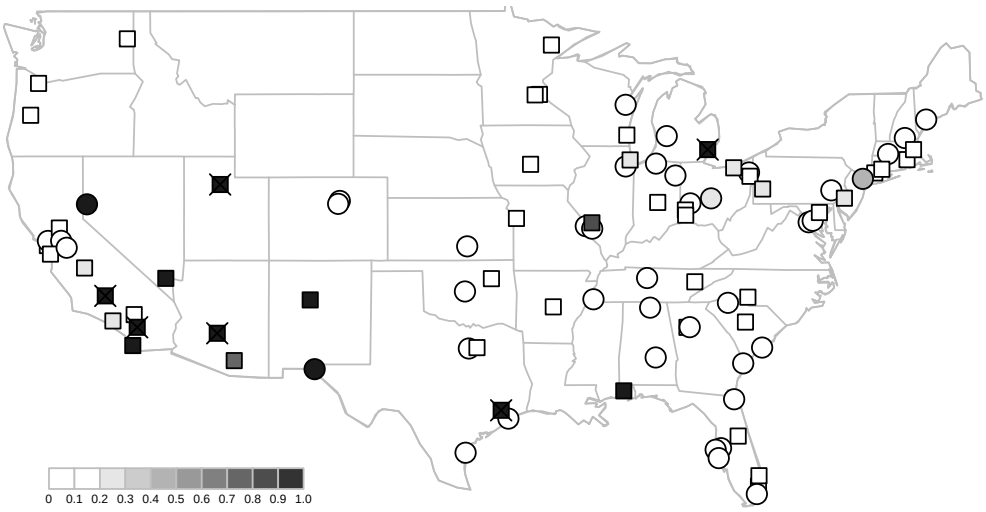


Figure 6: The grey scale represents the county-specific posterior probability of $PM_{10-2.5}$ nonattainment status for the period 2003-2005 for 95 US counties. □ denotes counties with at least one collocated monitor pair before imputation and ∘ denotes counties without eligible collocated monitor pairs before imputation. × denotes that a county was in nonattainment without imputed $PM_{10-2.5}$ values.