Statistical Methods for Analyzing Complex Spatial and Missing Data

Statistical methods for analyzing complex spatial and missing data

A dissertation presented

by

Joseph Lawrence Antonelli

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

September 2015

Dissertation Advisors: Brent Coull and Francesca Dominici Joseph Lawrence Antonelli

Statistical methods for analyzing complex spatial and missing data

Abstract

In chapter 1, we develop a novel two-dimensional wavelet decomposition to decompose spatial surfaces into different frequencies without imposing any restrictions on the form of the spatial surface. We illustrate the effectiveness of the proposed decomposition on satellite based $PM_{2.5}$ data, which is available on a 1km by 1km grid across Massachusetts. We then apply our proposed decomposition to study how different frequencies of the $PM_{2.5}$ surface adversely impact birth weights in Massachusetts.

In chapter 2, we study the impact of monitor locations on two stage health effect studies in air pollution epidemiology. Typically in these studies, estimates of air pollution exposure are obtained from a first stage model that utilizes monitoring data, and then a second stage outcome model is fit using this estimated exposure. The location of the monitoring sites is usually not random and their locations can drastically impact inference in health effect studies. We take an in-depth look at the specific case where the location of monitors depends on the locations of the subjects in the second stage model and show that inference can be greatly improved in this setting relative to completely random allocation of monitors.

In chapter 3, we introduce a Bayesian data augmentation method to control for confounding in large administrative databases when additional data is available on confounders in a validation study. Large administrative databases are becoming increasingly available, and they have the power to address many questions that we otherwise couldn't answer. Most of these databases, while large in size, do not have sufficient information on confounders to validly estimate causal effects. However, in many cases a smaller, validation data set is available with a richer set of confounders. We propose a method that

uses information from the validation data to impute missing confounders in the main data and select only those confounders which are necessary for confounding adjustment. We illustrate the effectiveness of our method in a simulation study, and analyze the effect of surgical resection on 30 day survival in brain tumor patients from Medicare.

Contents

Acknowledgments

I would like to thank my advisors, Brent Coull and Francesca Dominici, for an incredible experience working with you both. I can't imagine a better and more supportive set of advisors who I consider to not only be great mentors, but more importantly also great friends.

I would also like to thank Cory Zigler and Joel Schwartz for their invaluable advice throughout my time working on these projects. Whether it was at committee meetings or random times that I stopped by your offices I knew that I could come to you guys for help on projects and I would come away with new ideas and a new perspective.

I would also like to thank all my friends that I've made in my time here in Boston, especially those from the Biostatistics department. There are too many people to name here, but they've made my time in Boston special and it's because of them that I now consider Boston home, something I never thought I would say when I arrived here.

I would especially like to thank Georgia Papadogeorgou for being an amazing person and partner. You've been incredibly supportive of me and you always push me to bigger and better things. The last two years have been special in many ways and that is all a credit to how much you mean to me.

Finally, I would like to thank my parents and my brother who have all shaped me into the person that I am today. None of this would be possible without you guys, and I consider myself very lucky to call you my family.

Spatial multiresolution analysis of irregularly spaced grids with application to the effect of $PM_{2.5}$ on birth weights

Joseph Lawrence Antonelli

Department of Biostatistics

Harvard Graduate School of Arts and Sciences

Joel Schwartz

Department of Biostatistics

Harvard Chan School of Public Health

Itai Kloog

The Department of Geography and Environmental Development

Ben-Gurion University of the Negev

Brent Coull

Department of Biostatistics

Harvard Chan School of Public Health

1.1 Introduction

The epidemiologic literature investigating the health effects of air pollution has become vast, as countless studies have found associations between ambient levels of air pollution and a variety of adverse health outcomes (Dockery et al. (1993); Samet et al. (2000); Dominici et al. (2006) and a review of the literature is provided by Dominici et al. (2003); Pope III (2007); Breysse et al. (2013)). Despite the large number of studies investigating the relationship between air pollution exposures and human health, there still exist critical and unanswered questions that need to be addressed for the establishment of new regulations. Currently the EPA regulates total $PM_{2.5}$ levels, however, $PM_{2.5}$ is comprised of many different components and sources of pollution . An important question is the extent to which various sources of pollution adversely affect health, knowledge of which could lead to more effective targeted regulations. Establishing the extent to which fine scale, local pollution or long range regional transport pollution is associated with health effects would be very useful in planning future air regulations on lowering air pollution standards. Furthermore, isolating different sources of air pollution would allow researchers to gain valuable information regarding the chemical composition of $PM_{2.5}$ in different regions, and potentially quantify the health impacts of these separate components.

There have been very few attempts to jointly model the health effects of long range pollution and local pollution sources such as traffic, though it remains a crucial question in air pollution epidemiology. Maynard et al. (2007) used a variety of atmospheric, weather, and land use variables to predict black carbon levels for individuals in the Boston area. Black carbon (BC) is known to be highly associated with local traffic pollution. Sulfates are known to be spatially homogenous and represent long range pollution sources such as coal fired power plants, and the investigators examined the joint health effects of these two exposures. Other articles have decomposed air pollution into local and regional sources without subsequently examining their respective effects on health. Moreno et al. (2009) examined the differences in hourly fluctuations of traffic and urban background components of PM10 in Santander, Spain. Brochu et al. (2011) used quantile regression to estimate the regional and local components of BC in Boston and investigated how the

sources of BC changed both across the year and within a given day.

Due to recent advancements in $PM_{2.5}$ exposure estimation, we no longer need to rely solely on $PM_{2.5}$ monitors, as remote sensing satellite data can now yield reliable $PM_{2.5}$ estimates on a 1km by 1km grid (Kloog et al., 2014). These new estimates of $PM_{2.5}$ are on a scale fine enough to allow novel approaches to spatial decomposition based on image analysis techniques. We show such decompositions of $PM_{2.5}$ can yield insights into the sources of pollution most associated with health effect estimates. A variety of methods have been proposed to decompose surfaces into different spatial scales, two of the most common such techniques are wavelet decompositions and Fourier decompositions. For the remainder of the manuscript we focus discussion on wavelets. Due to the existence of point and line sources of pollution, such as interstates and other roadways, the surface of $PM_{2.5}$ will contain many spikes. Wavelets are well known to be a useful basis function for preserving sharp features of data (Petrosian and Meyer, 2013), and many spike detection algorithms are based on wavelet transforms (Hulata et al., 2002; Nenadic and Burdick, 2005). One of the main goals of the study is to characterize the impact of traffic pollution sources on health and therefore it is important to adequately capture, and not over smooth, these spikes in the data. Moreover, wavelets decompose a spatial surface into multiple spatial scales that are orthogonal, which allows us to avoid multicolinearity and the resulting instability of effect estimates in a health effects model.

Standard two-dimensional Wavelet decompositions typically require that the surface being decomposed is rectangular and the points are uniformly spaced. An additional requirement of standard wavelet analysis is that the points are dyadic, meaning that the number of points on the surface grid is 2^l where l is some positive integer, although "padding" the practice of adding points to the surface can be used to satisfy this requirement. Our interest focuses on decomposing a spatial surface of $PM_{2.5}$, a setting in which none of these conditions are met. The coastline of the U.S is far from rectangular. There is no reason to think our data would be dyadic, and the satellite data yielding the $PM_{2.5}$ estimates is not on a perfectly uniform grid. Previous work has avoided the dyadic assumption as well as the uniform grid assumption through a variety of techniques such as the lifting scheme and interpolation (Sweldens, 1998; Xiong et al., 2006; Pollock and Cas-

cio, 2007; Gupta et al., 2010). Others have generalized Wavelet theory using radial basis functions, which are not constrained to lie on a uniform grid (Buhmann, 1995; Chui et al., 1996). To the best of our knowledge, however, none of these have provided a practical way of decomposing a surface that is not rectangular. In this paper we develop a two-dimensional extension to work originally proposed in Wand et al. (2011), which relaxes these assumptions of a standard wavelet analysis. The advantages of this method include its simple application, its ability to scale to large spatial surfaces, and its ability to avoid restrictive assumptions about the nature of our surface.

In this paper we will use the proposed wavelet based method to decompose daily surfaces of $PM_{2.5}$ across New England and use the components of the resulting decomposed surface as covariates in a health effects model relating birth weight in Massachusetts to scale-specific $PM_{2.5}$. Section 1.2 introduces the pollution data and motivating scientific problem. Section 1.3 introduces the proposed method for performing 2d wavelet decompositions on irregular grids. Section 1.4 illustrates the decomposition in the $PM_{2.5}$ data. Section 1.5 applies the method to analyze the association between scale-specific $PM_{2.5}$ and birthweights in Massachusetts, and Section 1.6 concludes with further discussion.

1.2 $PM_{2.5}$ and Birthweights in Massachusetts

1.2.1 Exposure data

Typically $PM_{2.5}$ is measured at monitoring stations, which are located sporadically across the United States. In early health effect studies, conditional on the monitoring values exposure to $PM_{2.5}$ was assigned to be the value from the nearest monitor or a weighted average of monitors within a pre-defined range. In recent years monitoring data has been augmented with geographical and remote sensing information to yield individual, residence specific estimates of $PM_{2.5}$ levels. Specifically, in previous work we have combined ideas from land use regression and mixed models, and incorporated satellite aerosol optical depth (AOD) measurements to obtain widespread estimates of $PM_{2.5}$ at a 1 x 1 km resolution (Kloog et al., 2014). Satellite AOD is a measure of light attenuation in the atmospheric column that is affected by ambient conditions and can be used to help estimate

$PM_{2.5}$. Satellite estimates of $PM_{2.5}$ on a 1km grid are available daily from 2003 to 2011 for the Northeastern United States and they give an accurate estimate of the surface of $PM_{2.5}$ in this area. Kloog et al. (2014) showed the R^2 values between predictions and true values observed at monitors is around 0.9 indicating high predictive accuracy, although this only applies to areas with monitors and therefore doesn't give much insight into the performance of the predictions in rural areas. Despite potential drawbacks of the surfaces being estimated, the fine scale nature of the exposure surface allows us to apply our proposed decomposition method and examine the effects of air pollution at a wide range of spatial scales.

1.2.2 Birth weights

Many epidemiological studies have established relationships between $PM_{2.5}$ and adverse birth outcomes. Glinianaia et al. (2004); Dadvand et al. (2013) provide a review of the literature. In Massachusetts, Kloog et al. (2012) reported an association between $PM_{2.5}$ and birth weights using Satellite AOD based $PM_{2.5}$ estimates on a 10km by 10km grid. We extend this work by using the finer scale, 1km satellite based $PM_{2.5}$ estimates as well as estimating associations between birthweight and specific spatial-scales of variation of $PM_{2.5}$ exposure. Kloog et al. (2012) provided specific details of the birthweight data. Briefly, the study population includes all singleton live births from the Massachusetts Birth Registry from January 1st, 2000 to December 31st, 2008. We restrict attention to births after October 1st, 2003 as the satellite based $PM_{2.5}$ data is only available from 2003 onwards. The data set contains 332,717 singleton births and the geocoded address of each mother at the time of birth. This data combined with the satellite data described above and our proposed Wavelet decomposition as described in section 1.3 should provide insight into the different health impacts of local and regional sources of $PM_{2.5}$.

1.3 Wavelet decomposition for irregular grids

1.3.1 Standard wavelet analysis

To motivate our approach and establish notation, we begin with a standard one-dimensional Wavelet decomposition. We will then extend the standard Wavelet decompositions to that seen in Wand et al. (2011) which removes the issue of the data lying on a uniform grid. Finally, we will extend this approach to the two dimensional setting as our approach is a two-dimensional extension from that seen in Wand et al. (2011). For now, imagine that we have data, y, which is a function of x, dyadic, and equally spaced on the interval [0,1). In this case we have $R = 2^L$ equally spaced data points, which leads to K = $R - 1$ basis functions. We are trying to represent our data in the following form as the sum of Wavelet basis functions

$$\begin{aligned} y &= f(x) \\ &= \theta_0 + \sum_{k=1}^{K} \theta_k z_k^u(x), \end{aligned} \tag{1.1}$$

where y is our data, and z_k are wavelet basis functions. Wavelet coefficients have a nice interpretation in terms of scale and location of the function they are trying to represent. At level l there are $2^l - 1$ basis functions, which move from left to right in terms of the support of each function. The lower level basis functions represent low frequency changes in our function, while the higher level basis functions represent the higher frequency changes in the function. Going back to our motivating example of $PM_{2.5}$ this means that the lower level functions will capture smooth, regionally varying trends in pollution, while the higher level functions will capture local $PM_{2.5}$ changes such as interstates. Therefore our basis functions are $z_1^u()$ which is the 1 function and represents level 1, $z_2^u()$ and $z_3^u()$ which represent level 2, and so on for levels 3 to L. The above formulation can be rewritten as

$$y = W\theta, \tag{1.2}$$

where row i of the W matrix takes the following form

$$\left[1 \; z_1^u(\frac{i-1}{R}) \; ... \; z_{R-1}^u(\frac{i-1}{R})\right]. \tag{1.3}$$

In this case, determination of θ is trivial since the orthogonality of W leads to

$$\theta = W^T y. \tag{1.4}$$

A variety of standard wavelet basis functions can be used and for this paper we will stick to the family of Debauchies wavelets, however, this choice is not of crucial importance for illustrating the method.

1.3.2 One dimensional penalized wavelets

Now that we have introduced some preliminary notation and the standard setup for this problem we can move to the case where the data is not on a regular grid. What this means is that we now have data, $y_i for i = 1...n$ observed at locations $x_i for i = 1..n$ where we have imposed no restrictions about the location and dimension of y. We can define a new set of basis coefficients as

$$z_k(x) = z_k^u(\frac{x-a}{b-a}), \quad k = 1...K, \tag{1.5}$$

where a and b are the minimum and maximum of x respectively, so we are essentially normalizing our data to lie in the interval [0,1). Now the problem lies in estimating $z_k^u()$ at arbitrary points within the unit interval, since standard Wavelet basis functions using the discrete wavelet transform can only be evaluated via a recursive algorithm on a dyadic grid. The trick to doing this proposed by Wand et al. (2011) is to define a very fine grid of points on the unit interval, with the number of points in our grid being a multiple of 2. We are able to evaluate the basis functions on this grid and we will pick this grid to include a large number of points such as R=16,384 so that any of our data points will lie

very closely between two grid points. We could then calculate the value of $z_k^u()$ at any point in the interval as a linear interpolation of the two nearest grid points as

$$z_k^u(x) \approx \{1 - (xR - \lfloor xR \rfloor)\} z_k^u \left(\frac{\lfloor xR \rfloor}{R} \right) + (xR - \lfloor xR \rfloor) z_k^u \left(\frac{\lfloor xR \rfloor + 1}{R} \right), \tag{1.6}$$

and using this we can define a matrix Z as

$$Z = \begin{pmatrix} 1 & z_1(x_1) & \dots & z_K(x_1) \\ 1 & z_1(x_2) & \dots & z_K(x_2) \\ \vdots & \vdots & & \vdots \\ 1 & z_1(x_n) & \dots & z_K(x_n) \end{pmatrix} \tag{1.7}$$

Now that we have defined wavelet basis functions and how to evaluate them at arbitrary locations, we simply need to estimate the coefficients corresponding to our basis functions. This can be done very easily in any standard regression framework, since we are in the following linear model situation

$$y = Z\theta + \epsilon, \tag{1.8}$$

where ϵ is a mean zero vector of noise. The term 'penalized wavelets' refers to the situation where we find the wavelet basis coefficients using a penalized regression framework such as LASSO or ridge regression. For now we will estimate the parameters using LASSO as this turns out to be a better penalty for our pollution example in section 5. Our estimate of θ is defined by

$$\hat{\theta} = \arg\min \left\{ \sum_{i=1}^{n} (y_i - Z_i\theta)^2 \right\} \text{ subject to } \sum_j |\theta_j| < c, \tag{1.9}$$

where c is a tuning parameter that controls the amount of penalization (Tibshirani, 1996). Now we've shown how to represent a function with wavelet basis functions on any grid and any number of data points. Using this we can decompose a function to investigate how a function changes at different frequencies, by only looking at the elements of θ

that correspond to the scale we are interested in. Our motivating example, however, is the surface of air pollution in New England and therefore requires a two dimensional wavelet decomposition, which we detail in the following sections.

1.3.3 Standard two dimensional wavelets

Again let us assume we are in the standard framework for wavelets where we have dyadic data on a uniform grid, however, now our data lies on a two dimensional surface. Our data, y, now takes the form of a 2^L by 2^L matrix and we wish to decompose it into different frequencies. Performing a two dimensional wavelet decomposition of this data is analogous to performing a one dimensional wavelet transform on the rows of y and then performing another one dimensional wavelet transform on the columns of the resulting matrix. Intuitively in a two dimensional plane this is like doing a wavelet transform in one direction (x_1) and then doing it in the second direction (x_2). This can be written out as

$$y = W^{x_1}\theta^T W^{x_2 T}, \tag{1.10}$$

where in this setting, $W^{x_1} = W^{x_2}$ and both are defined in the same manner as our W matrix from the one dimensional section. We keep the two matrices separated by x_1 and x_2 at the moment for generalizability and to maintain notation in the following section when we introduce two dimensional wavelets for non dyadic and irregular data patterns. One thing to note is any row a and column b of y can be written as

$$y_{ab} = \sum_{i=1}^{2^L}\sum_{j=1}^{2^L} W_{ai}^{x_1}\theta_{ij}W_{bj}^{x_2}, \tag{1.11}$$

and this is important because it shows that the data is linear in θ and quadratic in the W matrices, suggesting that we could potentially fit this in a linear model framework just as in the one dimensional case. What we mean by that is if we write y as a vector instead of a matrix, we can write any element a of the vector y as

$$y_a = \sum_{i=1}^{2^L} \sum_{j=1}^{2^L} \theta_{ij} A_{ij}, \tag{1.12}$$

where A_{ij} is the corresponding quadratic function of elements of the W matrices. This is important because we can now write the data in the usual regression framework

$$y = W^* \theta, \tag{1.13}$$

where W^* is a design matrix with the appropriate elements, A_{ij}, and then we can solve for the wavelet coefficients using regression techniques. We will now show in the following section how we can extend our ideas from the one dimensional setting to the two dimensional setting to perform a Wavelet decomposition for irregular data by the same logic.

1.3.4 Two dimensional penalized wavelets

In the one dimensional setting when we were on an irregular grid we came up with a new basis function by defining

$$z_k(x) = z_k^u(\frac{x-a}{b-a}), \quad k = 1...K, \tag{1.14}$$

and then linearly interpolating on a fine grid to calculate the value of this function for any arbitrary point. Imagine now that we again have n data points, $y_i^*,\ i = 1...n$, which lie on a two dimensional space with the locations defined by x_{1i} and x_{2i}. In the motivating example, y would be $PM_{2.5}$ levels and x_1 and x_2 would be longitude and latitude respectively. We can perform the one dimensional wavelet transform on an irregular grid in both the x_1 and x_2 directions. We can define analogous functions for both directions as

$$z_k^{x_1}(x_1) = z_k^u(\frac{x_1-a_1}{b_1-a_1}), \quad k = 1...K \tag{1.15}$$

$$z_k^{x_2}(x_2) = z_k^u(\frac{x_2-a_2}{b_2-a_2}), \quad k = 1...K, \tag{1.16}$$

where a_1 and b_1 define the range of x_1, and a_2 and b_2 the range of x_2. Using these we can define matrices that are analogous to the Z matrix we defined in the one dimensional setting

$$Z^{x_1} = \begin{pmatrix} 1 & z_1^{x_1}(x_{11}) & \dots & z_K^{x_1}(x_{11}) \\ 1 & z_1^{x_1}(x_{12}) & \dots & z_K^{x_1}(x_{12}) \\ \vdots & \vdots & & \vdots \\ 1 & z_1^{x_1}(x_{1n}) & \dots & z_K^{x_1}(x_{1n}) \end{pmatrix} Z^{x_2} = \begin{pmatrix} 1 & z_1^{x_2}(x_{21}) & \dots & z_K^{x_2}(x_{21}) \\ 1 & z_1^{x_2}(x_{22}) & \dots & z_K^{x_2}(x_{22}) \\ \vdots & \vdots & & \vdots \\ 1 & z_1^{x_2}(x_{2n}) & \dots & z_K^{x_2}(x_{2n}) \end{pmatrix} \tag{1.17}$$

and now that we have defined these matrices as such, we can plug them into 1.12 to obtain the following representation of our data

$$y_a^* = \sum_{i=1}^{K} \sum_{j=1}^{K} Z_{ai}^{x_1} \theta_{ij} Z_{aj}^{x_2} \quad a = 1...n. \tag{1.18}$$

If we think about how we solved for the wavelet coefficients in the one dimensional setting using penalized regression we can start to see how we can solve for the estimated coefficients in this case as well. If we define a design matrix as

$$Z^* = \begin{pmatrix} Z_{11}^{x_1} Z_{11}^{x_2} & Z_{11}^{x_1} Z_{12}^{x_2} & \dots & Z_{1K}^{x_1} Z_{1K}^{x_2} \\ Z_{21}^{x_1} Z_{21}^{x_2} & Z_{21}^{x_1} Z_{22}^{x_2} & \dots & Z_{2K}^{x_1} Z_{2K}^{x_2} \\ \vdots & \vdots & & \vdots \\ Z_{n1}^{x_1} Z_{n1}^{x_2} & Z_{n1}^{x_1} Z_{n2}^{x_2} & \dots & Z_{nK}^{x_1} Z_{nK}^{x_2} \end{pmatrix} \tag{1.19}$$

then we are now essentially in the standard regression setup of

$$y^* = Z^* \theta + \epsilon, \tag{1.20}$$

and as before we can solve for $\hat{\theta}$ using

$$\hat{\theta} = \arg\min \left\{ \sum_{i=1}^{n} (y_i^* - Z_i^* \theta)^2 \right\} \text{ subject to } \sum_j |\theta_j| < c. \tag{1.21}$$

Note that because of the way we defined our wavelet basis functions we have not imposed any restrictions on the dimension or shape of the two dimensional surface we are

representing. Once the Z^* matrix is defined the method is trivial to implement and we can examine particular values of θ to determine how the surface changes at certain frequencies or scales.

1.4 Application to satellite $PM_{2.5}$ data

Now we will apply our method for two dimensional wavelet decompositions on each day of $PM_{2.5}$ data in New England. We will restrict attention to a subset of New England that surrounds Massachusetts as this will be the area of interest in the analysis of birth weights. The goal of the method is to separate different sources of pollution, which vary at different spatial frequencies. Previous work in air pollution exposure assessment has established that pollution separates into three different sources that vary at different spatial distances: Large scale regional pollution, pollution sources contributing to urban background, and sources of localized pollution such as traffic. Within the study region, regional sources of pollution will not vary spatially very much within a given day, though can vary greatly from day to day. This suggests that regional pollution sources will be represented by temporal variation in $PM_{2.5}$, while the urban and local traffic sources will be represented by spatial variability. Urban pollution sources should be represented by lower level Wavelet coefficients as they vary over longer distances. These are represented by differences in pollution between cities and rural areas. Local pollution sources, represented by the higher level Wavelet coefficients, are those that vary quickly across space as they are caused by things such as interstates, and quickly fade away across space.

Figure 1.1 illustrates the average of our Wavelet decompositions across each day in 2006 for New England. We are able to see how the pollution surface breaks into two separate components, representing different spatial frequencies at which pollution changes. To create the lower frequency component seen in figure 1.1 we create a new vector of coefficients $\tilde{\theta}$ which is equal to $\hat{\theta}$ with the exception that the coefficients corresponding to higher frequency basis functions are set to zero. We can then obtain a predicted lower frequency component via $Z^*\tilde{\theta}$. To obtain a high frequency component as seen in figure 1.1 we could apply the same process, where instead we set the coefficients correspond-

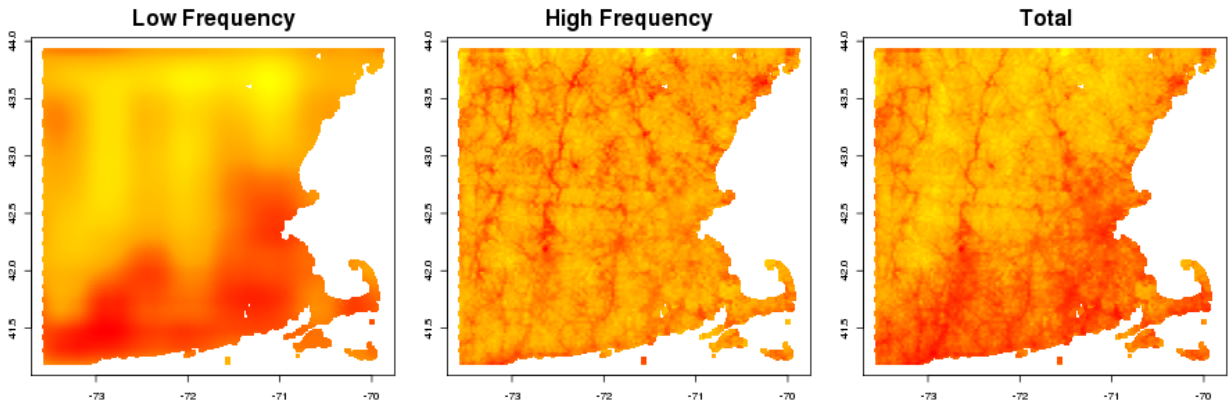

Figure 1.1: Illustration of average wavelet decomposition averaged over the satellite data for 2006. The left panel shows the true surface, the middle panel is the high frequency component from our Wavelet decomposition, and the right panel is the low frequency component from our Wavelet decomposition.

ing to low frequency basis functions to zero. Alternatively, we could just take difference between the true surface and the lower frequency component as the higher frequency component.

One issue with separating the pollution surface into two different components, is the selection of a cutoff for what is deemed to be low frequency. In figure 1.1 we considered basis functions to be low frequency if they represented levels 1 through 3 in either direction (latitude or longitude). Alternatively we could have considered basis functions that represented levels 1 through 2 in each direction as representing low frequency changes, and this would have led to a slightly smoother surface for both the low and high frequency surfaces. For the analysis of birth weights we will use basis function levels 1 through 3 to represent low frequency changes in pollution. We decided upon this threshold both by visual inspection as the decomposed surfaces appear to represent changes of interest to the air pollution epidemiology literature, and by examining the physical distances associated with each wavelet basis functions support. Note that while this decision is subjective, it is not of great importance as we will examine the effect of all spatial scales simultaneously regardless of their status as ”low” or ”high” frequency.

1.5 Analysis of birth weight data

We applied our proposed decomposition to examine the impact of $PM_{2.5}$ at different spatial scales on birth weights in Massachusetts for the period 2003-2008. We perform the Wavelet decomposition for each day in the study period to obtain pollution surfaces representing different spatial scales.

1.5.1 Low vs High frequency components

First we examine the extent to which low and high frequency sources of pollution impact birth weight. We define the low frequency $PM_{2.5}$ component to be represented by scales 1 through 3, and the high frequency component to be scales 4 through 7. When looking at birth weights as an outcome our exposure can be defined either in terms of the mother's full gestation period, a given trimester, or the last 30 days of the gestation period, though for the purposes of this paper we will restrict attention to the individual trimesters. This leads to a temporal component in the exposure as well as a spatial component, because we are looking at exposures averaged across time. Due to this we split our exposure surface for each day into three separate components: A mean component that is simply the mean $PM_{2.5}$ for Massachusetts on the day of interest, a low frequency spatial component, and a high frequency spatial component. For a particular mother in the study, these three exposure components computed for each day are then averaged across the trimester of interest. The idea in separating out the mean component from the surface is that it will capture temporal variation in $PM_{2.5}$ levels and allow the low and high frequency scales to solely represent spatial variability.

We will examine the effects of $PM_{2.5}$ at different scales using one of two models. The first model is defined as follows:

$$BW_{ij} = (\beta_0 + \beta_{0j}) + \beta_1 \text{PM}_{2.5ij} + \boldsymbol{\beta_c}\mathbf{C}_{ij} + \epsilon_{ij}, \tag{1.22}$$

where the subscript ij represents subject i in census tract j. $PM_{2.5ij}$ is the overall $PM_{2.5}$ value for mother i in census tract j. We have included a random intercept for census tract

to control for any correlation among mothers in similar neighborhoods. The vector $\mathbf{C}_{ij}$ represents all potential confounders we have included. ϵ_{ij} is a mean zero, normal error component. Details of the model choice and confounder selection can be found in Kloog et al. (2012). This model does not use any of the spatially decomposed $PM_{2.5}$ exposures, and therefore β_1 represents the combined effect of all $PM_{2.5}$ spatial scales.

To examine how different spatial scales of $PM_{2.5}$ affect birth weight we will also examine the following model:

$$BW_{ij} = (\beta_0^* + \beta_{0j}^*) + \beta_1^* \text{Mean}_{ij} + \beta_2^* \text{Low}_{ij} + \beta_3^* \text{High}_{ij} + \boldsymbol{\beta_c^*} \mathbf{C}_{ij} + \epsilon_{ij}, \tag{1.23}$$

where everything is the same as in the previous model, except now we have split our $PM_{2.5}$ exposure into it's three components. The magnitude and direction of $\beta_1^*, \beta_2^*, \beta_3^*$ should give valuable insight into how the various components of $PM_{2.5}$ are impacting birth weights. Figure 1.2 shows the results from the aforementioned models for the full gestation period, and each of the trimesters.

The results indicate that the effects are fairly similar across trimesters in terms of magnitude, direction, and relationship between sources of $PM_{2.5}$. Both the low and high frequency components of $PM_{2.5}$ have large, significantly negative associations on birth weight suggesting that increased levels of either of these sources adversely affects birth weight. Interestingly the mean component has very small effects, and in the case of trimesters 1 and 2, even slightly positive effects. This would suggest that mothers who gave births during time periods when $PM_{2.5}$ was elevated had healthier babies in terms of birth weight. One potential explanation for this effect is that the mean component is confounded by time. We know that this source of $PM_{2.5}$ represents temporal variation in exposure. We also know that both $PM_{2.5}$ and birth weights are decreasing during the study period, which could explain the slightly positive effect we see. To test this we fit the same model as in 1.23 but included a smooth term for time into the vector of potential confounders, $\mathbf{C}_{ij}$. After applying this model the effect of the mean component drops down to around -2. By separating out this scale from the low and high frequency scales, we have reduced the possibility of temporal confounding influencing our remaining ef-

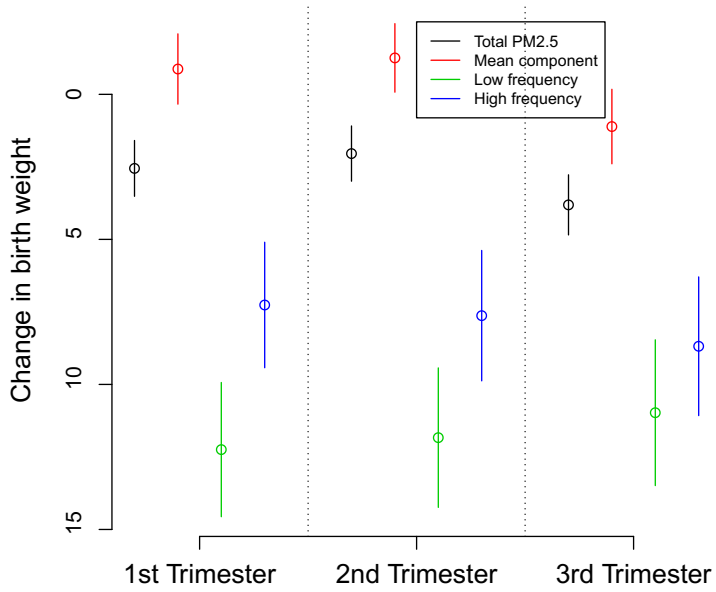

Figure 1.2: Parameter estimates and corresponding 95% confidence intervals from $PM_{2.5}$ models for each time period. Black line is the estimate of β_1 from model 1.22 and the remaining lines are the estimates of $\beta_1^*, \beta_2^*, \beta_3^*$ from model 1.23

fect estimates as they represent only spatial variability in $PM_{2.5}$. The effect sizes that we see for the low and high frequency component are both larger than the overall pollution effect, β_1, and this is likely because we have removed the temporal sources of variation that have smaller effects via the mean component. The effect from the low frequency component seems to be somewhat larger than the high frequency component, though the difference between the two gets smaller across trimesters.

1.5.2 Removing high level information

It is of scientific interest to understand specifically which spatial scales of $PM_{2.5}$ are driving changes in birth weight as this will be important in targeting future environmental regulations. With that in mind, we repeatedly fit the model with overall pollution, but successively removed high frequency scales from the pollution surface to see how the parameter estimates change. We also keep the mean component separated as the results

from the previous section indicate that it could be temporally confounded, and leaving it in the term for $PM_{2.5}$ would dilute the signal that we are seeing from the low and high frequency components. For a given trimester, the model of interest is now

$$BW_{ij} = (\tilde{\beta_0} + \tilde{\beta_{0j}}) + \tilde{\beta_1}\text{Mean}_{ij} + \tilde{\beta_2}\text{PM}^R_{2.5ij} + \tilde{\boldsymbol{\beta}}_{\boldsymbol{c}}\mathbf{C}_{ij} + \epsilon_{ij}, \tag{1.24}$$

noting a couple changes from model 1.22. First we have included the mean component into the model to control for potential sources of temporal confounding. We also have defined a new variable $\text{PM}^R_{2.5ij}$, which represents the total pollution with certain spatial scales removed. $\text{PM}^R_{2.5ij}$ Always will be missing the mean component for the reasons above, and we will further remove the high frequency scales one at at time to see how the effect changes. Figure 1.3 shows the estimates and 95 % confidence intervals from model 1.24 as we remove more and more high frequency information.

The effects show a similar pattern and magnitude among the three different trimesters. We see that removing the highest 1st or 2nd scales from the $PM_{2.5}$ surface actually increases the magnitude of the effect of $PM_{2.5}$ on birth weights as the effect goes from -10.0 to -13.2 for the 3rd trimester, with similar jumps in the other trimesters. This suggests that the effects at these very high frequency scales are much smaller in magnitude than their lower frequency counterparts. It is even plausible that these levels could have no effect on birth weight in which case they would represent measurement error and removing them would be a useful feature of our Wavelet decomposition. Looking at the remaining scales we see that there is a decline in effect size as we remove more and more levels, with a large change occurring at scale 4. For trimester 3 the effect drops in magnitude from -12.2 to -8.5 when we exclude the fourth scale from the overall effect. This is a rather large change compared with the other scales and indicates that a source of pollution occurring at that scale has a large impact on birth weights. Overall though the biggest impact seems to occur simply by including the 1st level into the model. The coefficient when we only include the 1st level into the model is -6.23 for trimester 3, and as high as -9.7 for trimesters 1 and 2, indicating that there is a large effect at this scale.

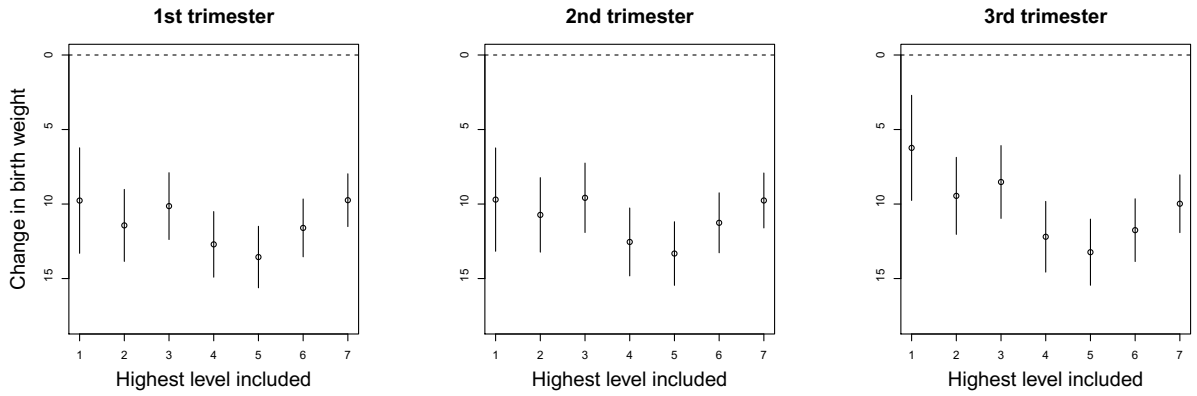

Figure 1.3: Parameter estimates and corresponding 95 % confidence intervals from model 1.24 when we remove high frequency spatial scales for each trimester. Within each panel from right to left we successively remove more and more of the higher frequency scales

1.5.3 Modeling each scale separately

The previous sections lend intuition for which scales are driving the adverse impact of $PM_{2.5}$ on birth weights, but we can also model each individual level separately instead of clustering them together into a joint exposure. We now fit the following model

$$BW_{ij} = (\breve{\beta}_0 + \breve{\beta}_{0j}) + \breve{\beta}_1 \text{Mean}_{ij} + \breve{\beta}_2 \text{Level1} + ... + \breve{\beta}_8 \text{Level7} + \breve{\boldsymbol{\beta}}_{\boldsymbol{c}} \mathbf{C}_{ij} + \epsilon_{ij}, \tag{1.25}$$

where we are now including each scale as a separate predictor in the model. The magnitude and direction of the coefficients from this model should lend insight into exactly which scales are driving the effects of $PM_{2.5}$, conditional on the levels of the other scales. Table 1.1 shows the effect estimates from this model as well as the estimates standardized by their standard errors to show the approximate level of significance for each scale. The results generally agree with those seen in figure 1.3 and show a couple of interesting results.

For any trimester, we again see that there is little to no effect of the highest two frequency scales. This further explains why we saw an increase in the magnitude of the overall effect when we removed these scales. Surprisingly, scale 3 seems to have a very small impact on birth weights despite the fact that levels 1, 2, 4, and 5 all have significant impacts on birth weight. The scales are increasing in terms of spatial frequency or the distance at which

	Trimester 1		Trimester 2		Trimester 3	
	$\hat{\beta}$	$\frac{\hat{\beta}}{SE(\beta)}$	$\hat{\beta}$	$\frac{\hat{\beta}}{SE(\beta)}$	$\hat{\beta}$	$\frac{\hat{\beta}}{SE(\beta)}$
Mean	1.00	1.61	1.49	2.47	-1.03	-1.56
Scale 1	-16.73	-8.81	-16.58	-8.78	-12.86	-6.73
Scale 2	-14.68	-10.29	-13.90	-9.43	-14.16	-9.24
Scale 3	-2.02	-0.74	-2.53	-0.93	-2.57	-0.90
Scale 4	-17.39	-7.55	-18.04	-7.69	-20.38	-8.50
Scale 5	-16.40	-6.03	-15.63	-5.66	-15.82	-5.60
Scale 6	-1.67	-0.80	-0.82	-0.37	-2.89	-1.23
Scale 7	-1.80	-0.99	-2.80	-1.49	-1.80	-0.90

Table 1.1: Effect estimates from model 1.25

$PM_{2.5}$ changes, so it's strange to see a spike at one level, and it merits further investigation. Overall the largest effects of $PM_{2.5}$ are seen in the first two scales, with significant effects at scales 4 and 5 as well.

1.5.4 Examination of scales

Due to some interesting results seen in previous sections regarding specific scales of $PM_{2.5}$ it is of interest to examine these scales further and to be able to translate information from spatial scales to distance, which is more useful to policymakers. We know from sections 1.5.1 - 1.5.3 that scales 1-5 all have significant effects on birth weight with the exception of scale 3, and that the largest such effect sizes seem to come from scales 1, 2, and 4. With this in mind it will be useful to look specifically at these scales and determine what they represent in terms of the $PM_{2.5}$ surface. Figure 1.4 shows the average surface from each spatial scale taken by performing our Wavelet decomposition on each day of data in 2006 and averaging them over the 365 days in the year. Visual inspection of these figures coupled with our knowledge of the Massachusetts area allows us to gain intuition as to which sources of pollution each scale is picking up. The first scale seems to pick up some regional transport pollution that is known move from west to east across lower Massachusetts, while the second scale picks up on some slightly noisier effects including the urban background from Boston. The remaining levels are less obvious from the figure, though it does seem that both the 6th and 7th scales are just random noise as postulated from our modeling results as they don't seem to correlate spatially in any way. We noticed

earlier that the 3rd scale didn't seem to have an effect on birth weight and while it is unclear what sources of pollution are driving this scale, it does seem that most of the signal at scale 3 lies in areas where very few people live. This could be what is keeping this scale from having an effect on birth weight.

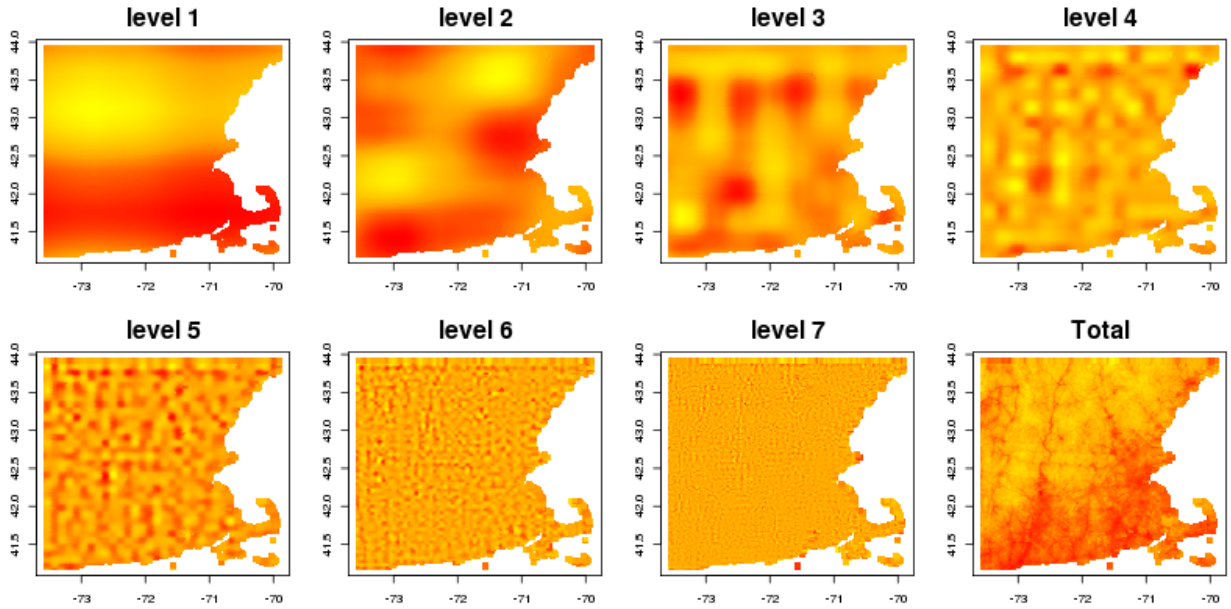

Figure 1.4: Illustration of average wavelet decomposition averaged over satellite data for 2006. From left to right and top to bottom are the individual scales taken from our wavelet decomposition for each day then averaged over the entire year. The final panel is the sum of all the levels.

It is also of interest to assign a physical distance to each spatial scale as this could be more useful when using these results to consider future air quality regulations. Wavelets in their simplest form, the Haar Wavelet, have a very simple interpretation in terms of distance. The support of the Haar Wavelet at scale 1 is the support of the data, at scale 2 the support is half the length of the data, and this continues as each scale represents a fraction of the data that is half the distance as the previous scale. The Haar Wavelet at any given scale will pickup on features of the function that vary at a level around half of the support of that scale, since the Haar Wavelet is positive for half of it's support and negative for the other half. While we use the smoother Debauchies family of Wavelet's that do not have a support that is strictly a fraction of the data, the majority of the support for each scale of the Debauchies Wavelet is splitting up the data into halves. With this in

mind we can assign to each scale a 'rough' measure of distance representing the distances at which $PM_{2.5}$ varies that the particular scale will be accounting for. Since our data is 302km by 302km, we know that scale 1 represents changes that occur at a distance of around 151km, scale 2 changes at around 75.5, etc. This information along with our results on the impact that each scale has on birth weight could be used in conjunction with scientific expertise to determine what sources of pollution are leading to large, adverse effects on birth weight.

1.6 Discussion

In this article we have proposed a two-dimensional wavelet decomposition that is flexible, easy to implement, and scalable to large spatial surfaces. By extending ideas from Wand et al. (2011) we have created a decomposition that does not rely on many of the assumptions of standard wavelet theory that are overly restrictive for many analyses, and places the decomposition in a regression framework which simplifies it's implementation and estimation of wavelet coefficients. Our proposed method will allow researchers to perform multiresolution analyses on spatial data regardless of the structure and scale of their data. Much of the wavelet literature relies on complicated algorithms to perform analyses, but in this paper we simplified wavelet analyses by showing that they are simply basis functions and therefore can be applied in much the same way as any other basis function used to represent a function. We showed in section 1.3 that once we have defined and evaluated the wavelet basis functions and placed them in their appropriate design matrix, then estimation is trivial and no different than any other basis function placed in a regression framework. We have found that the proposed method scales quite well as the surfaces of $PM_{2.5}$ we examined contained approximately 70,000 grid points. Using 7 wavelet levels we are left with $2^{14} = 16,348$ basis functions and therefore the only computational challenge is in fitting a regression model with a design matrix whose dimensions are 70,000 by 16,384. If the dimensions get too large, we suggest doing an approximation to our method that is based on the fact that the wavelet levels are orthogonal. One could fit the same regression model, but only include the first 4 to 5 levels

instead of 7 levels, making the model much faster to calculate. Then the residuals from this model can be regressed against the remaining levels to obtain an approximation to the full wavelet decomposition.

We illustrated our method on a 1km by 1km grid of $PM_{2.5}$ data and then examined how these different spatial scales impacted birth weight in Massachusetts. We noticed that the temporal component of $PM_{2.5}$ was positive or close to zero for each trimester, which is unexpected as we expect $PM_{2.5}$ to have a negative effect on birth weight. One potential explanation for this is confounding by time as both $PM_{2.5}$ and birth weights are decreasing over time, and therefore leaving time out of the model could lead to misleading effects. We ran further models that included a smooth function of time to eliminate any potential confounding by time and found that the effects of the mean component decreased to negative levels that were expected a priori. We also saw the effect of the low frequency pollution component is larger than the high frequency, though this difference decreases later on in the pregnancy. It is clear though that both have large, negative impacts on birth weight. We also examined the effect of each spatial scale by removing one scale at a time from $PM_{2.5}$ and seeing how the effect estimate changed. We noticed that the very high frequency component, which was represented by the 6th and 7th scales, seems to be just noise and is actually attenuating the effect towards zero. It is believed that there is some noise in the AOD measurements that are used to model $PM_{2.5}$ on the 1km grid and it is possible that the top levels of the wavelet decomposition are picking up solely on this noise. Wavelets have been used to de-noise signals and this de-noising property might be increasing the magnitude of our effects because it is eliminating measurement error which would attenuate the effect to zero. We also noticed that the majority of the effect of $PM_{2.5}$ is driven by scales 1 through 5 and with the largest effects coming from scales 1 and 4. These results could be very important for future regulations as they can be used to target sources of pollution that operate on those spatial scales.

One limitation of the results from the study of birth weights is that we are ignoring the fact that the $PM_{2.5}$ measurements are estimated and therefore come with some uncertainty. The confidence intervals placed on our model estimates are under the assumption that $PM_{2.5}$ is a fixed, known quantity and therefore these intervals are likely to be slightly

anti-conservative. While we are not attempting to make causal statements or statements of significance, rather simply trying to learn about the general magnitude and direction of effects, it would still be ideal to be able to account for this increase in uncertainty. Resampling methods could in theory be used to solve this problem, however, that would require resampling and re-fitting of the models used to estimate the exposure and this would not be feasible in this study. A related limitation is that because we are using estimates of $PM_{2.5}$ there might be measurement error biasing the results of our study. While we hypothesized that we were removing some of the effects of measurement error when we removed the highest two frequency wavelet scales, it's possible that measurement error is still systematically degrading our model estimates. Future work could focus on applying well developed measurement error correction techniques to examine if the effect estimates are drastically impacted.

As this is one of the first papers trying to separate the effects of different sources of air pollution, there are a vast number of possibilities for future research. One such idea is to use the wavelet decompositions of $PM_{2.5}$ to learn more about specific chemicals that make up $PM_{2.5}$. Data is available at monitoring sites about the specific components that $PM_{2.5}$ is comprised of, which means that we can perform canonical correlation analysis between our 1km decompositions and the component monitoring data to learn about what type of sources each component comes from. This could also lead to information about which components of pollution are negatively impacting health outcomes. It would also be of interest to apply these decompositions to a wide variety of health outcomes to identify if there are any outcomes in which only the local or regional sources of pollution have an effect. It is also of interest to apply more meaningful definitions to each scale. Wavelet scales themselves are not very interpretable, but we can potentially assign to each wavelet level a distance corresponding to the frequency of that level. This would be a much more meaningful interpretation to researchers in environmental health looking to relate these results back to their analyses.

The positive effects of population based preferential sampling in environmental epidemiology

Joseph Lawrence Antonelli

Department of Biostatistics

Harvard Graduate School of Arts and Sciences

Luke Bornn

Department of Statistics

Harvard University

Matthew Cefalu

RAND Corporation

2.1 Introduction

In the past few decades, numerous epidemiological studies have investigated the health effects of air pollution. Many studies have found statistically significant associations between ambient levels of air pollution and a variety of adverse health outcomes. Examples of such studies can be found in Dockery et al. (1993); Samet et al. (2000); Dominici et al. (2006) and a review of the literature can be seen in Dominici et al. (2003); Pope III (2007); Breysse et al. (2013). Difficulty in these studies arises due to spatial misalignment of the data as the locations of the subjects does not coincide with the locations at which we can observe air pollution levels. Most studies rely on monitoring data such as the IMPROVE network or the EPA's Air Quality System, which are only available at a fixed set of locations. Then, conditional on the monitors, investigators predict values of air pollution using nearest neighbor, kriging, or land use regression approaches (Oliver and Webster, 1990; Madsen et al., 2008; Kloog et al., 2012). In many instances the locations of these monitors are chosen for a specific reason such as measuring areas of high pollution levels or areas of high pollution density. Chow et al. (2002) discuss different designs by which monitoring sites can be chosen and potential objectives for monitor selection, many of which depend on the nature of the pollution itself. Kanaroglou et al. (2005) develop a formal method for selecting monitor locations that takes into account the spatial variability of the surface to be estimated, and the population being exposed to pollution. Matte et al. (2013) discuss the how monitor placement in New York City was designed to capture intra-urban spatial variability of air pollution. These are among the many of instances just in exposure estimation where monitors were placed in areas with regard to the levels of pollution and therefore is a relevant issue when utilizing these networks in health effect studies.

Numerous papers have been published regarding two stage analyses in environmental applications. The first stage consists of estimating parameters of an exposure model using monitoring data. The second stage then conditions on the exposure estimates from the first stage and uses this estimated exposure to investigate the association between exposure and an outcome. This leads to a complex form of measurement error, which does not

fall specifically into the category of either classical or berkson measurement error. Kim et al. (2009) looked at the impact of various predicted exposures on health effect estimation in air pollution studies. A variety of methods have been proposed to correct for this measurement error. Gryparis et al. (2009) examined the effectiveness of a variety of standard correction methods via simulation and gave intuition for when these measurement error corrections will work. More recently, Szpiro et al. (2011b) show that the measurement error can be decomposed into two components: A classical like and a berkson like component. They further came up with a computationally efficient form of the parametric bootstrap to correct for measurement error in two stage analyses. Szpiro and Paciorek (2013) take an in depth look at the impact of these two components of measurement error, and derive asymptotic results about the bias and variance of health effect estimates.

In this paper our focus will be the preferential sampling of monitors in two stage analyses, and the subsequent impact on measurement error and health effect estimation. Preferential sampling as defined by Diggle et al. (2010) is the scenario where the location of the monitors is dependent on the values of the spatial process they are measuring. In air pollution studies this would amount to monitors being placed in locations due to the amount of air pollution in those locations. Diggle et al. (2010) show that variogram estimates are biased under preferential sampling and come up with a method to control for preferential sampling in geostatistical inference. Gelfand et al. (2012) showed that preferential sampling can perform drastically worse with respect to estimating a spatial surface than sampling under complete spatial randomness (CSR). Interestingly, Szpiro et al. (2011a) show that better exposure prediction doesn't always lead to improved health effect inference, though in general we expect that improving exposure prediction will lead to better inference overall. Lee et al. (2015) examine the impacts of preferential sampling on health effect estimation in environmental epidemiology and find that the locations of monitors can drastically impact inference in second stage analyses. They also illustrate in a simulation study how inference in the second stage is improved under CSR compared with preferential sampling.

It is clear from the literature that from a strictly geostatistical perspective, preferential sampling can lead to poorer inference and must be accounted for in the exposure model-

ing process. The aim of this paper, however, is to show that if interest lies in the second stage health effect estimates then preferential sampling can lead to improved inference in many scenarios. The main reason for this and the key difference in our paper with previous studies regarding preferential sampling is that we will be emphasizing the relationship that population density plays when thinking about potential locations of monitors. Our results will illustrate the claim made in Szpiro and Paciorek (2013) that the densities governing the location of the subjects and the monitors should be the same. We will show that when taking population density into account, preferential sampling can lead to drastically improved inference in air pollution studies. The outline of our paper is as follows: Section 2.2 will introduce the motivating example of $PM_{2.5}$ in New England, Section 2.3 introduces notation and the modeling framework, in section 2.4 we derive some mathematical results regarding inference under two stage sampling designs, section 2.5 presents an illuminating simulation study to shed light on different aspects of the estimation procedure, section 2.6 highlights these results in the context of $PM_{2.5}$ monitoring locations in New England, and section 2.7 concludes with a discussion.

2.2 Motivating example

The majority of the preceding discussion and previous work on measurement error in environmental two stage analyses has been motivated by studying the adverse health effects of $PM_{2.5}$. $PM_{2.5}$ is a pollutant defined as the combination of all fine particles less than 2.5 micrometers in diameter. Nearly all research to date on the associations between $PM_{2.5}$ and health outcomes has been contingent on monitors to estimate exposure, with the lone exception being recent studies that have used aerosol optical depth (AOD) to estimate $PM_{2.5}$ on a finer grid (Kloog et al., 2012). Generally exposure is estimated conditional on monitoring data, and little attention is paid to the location of the monitors. The left panel in figure 2.1 shows a map of the EPA AQS monitor locations over New England as well as a map of estimated $PM_{2.5}$ across New England. The estimated $PM_{2.5}$ for New England is taken from the aforementioned models that use AOD to estimate exposure on a fine grid, which in this case is 1km by 1km. Looking at figure 2.1 helps to show the mo-

tivation for this work as it seems from the figure that there are more monitors in areas of higher pollution. It should also be noted that these areas of higher pollution correspond to higher population densities, for instance the right side of the map represents elevated pollution in the greater Boston metropolitan area. The right panel of figure 2.1 further illustrates that the monitor locations are in fact concentrated in more populated areas. As a measure of population density we use the number of census tracts within 0.3 degrees latitude or longitude of a location, and the monitor locations generally have more census tracts nearby than New England as a whole.

Although not definitive, it seems plausible that the location of monitors in New England follows a non-random sampling scheme, which meets our criteria for preferential sampling. If we are able to gain intuition about the impact of preferential sampling, we should also gain knowledge on how inference is impacted in studies of $PM_{2.5}$ that use the EPA AQS monitoring system.

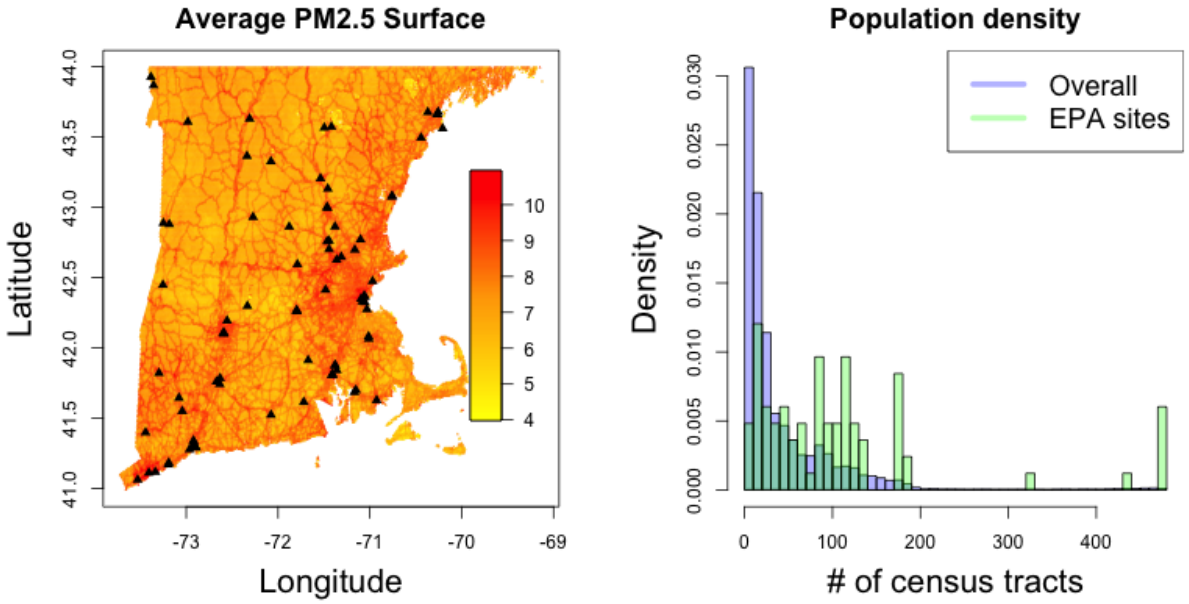

Figure 2.1: The left hand panel shows the average predicted $PM_{2.5}$ surface in New England for the year 2003, with the black points representing the location of the EPA AQS monitoring system. The right hand panel shows a histogram of the number of census tracts within 0.3 degrees of each grid point in New England, along with the histogram of the number of census tracts within 0.3 degrees of the EPA monitoring sites

2.3 General setup

2.3.1 Notation and model

For convenience, we adopt similar notation to Gryparis et al. (2009); Szpiro et al. (2011b); Lee et al. (2015). Throughout we will have n subjects in the study (second stage of analysis), and n^* monitors at which we observe exposure (first stage of analysis). We define X to be the true exposure at the n subject locations, and X^* to be the exposure at the n^* monitor locations. In general we will allow the following to hold:

$$\begin{pmatrix} X \\ X^* \end{pmatrix} \sim N\left\{ \begin{pmatrix} \mu_X(\alpha) \\ \mu_{X^*}(\alpha) \end{pmatrix}, \begin{pmatrix} \Sigma_{X,X}(\phi) & \Sigma_{X,X^*}(\phi) \\ \Sigma_{X^*,X}(\phi) & \Sigma_{X^*,X^*}(\phi) \end{pmatrix} \right\} \tag{2.1}$$

Where $\mu_X(\alpha)$ represents the mean of the exposure surface, and is a linear function of a set of covariates dictated by a parameter vector, α. The covariance matrix of this Normal distribution is dictated by a parameter vector, ϕ, which includes standard parameters such as the range, smoothness parameter, etc. This framework is general and allows for a broad class of exposure models for predicting exposure at new locations such as kriging and land use regression. We now define our outcome model as

$$Y = \beta_0 + X\beta_1 + \epsilon \tag{2.2}$$

Where ϵ is a vector of i.i.d random noise. In general this model can include a vector of covariates, though we keep it simple here to simplify results about the parameter of interest, β_1. In any realistic scenario, we will not observe X and therefore must estimate X with W defined as

$$W = E(X|X^*, \hat{\alpha}, \hat{\phi}) \tag{2.3}$$

In this situation above, this expectation is easily written out using properties of Normal distributions as

$$W = \mu_X(\hat{\alpha}) + \Sigma_{X,X^*}(\hat{\phi})\Sigma_{X^*,X^*}(\hat{\phi})^{-1}(X^* - \mu_{X^*}(\hat{\alpha})) \tag{2.4}$$

and the remainder of this paper will examine the impact of using W in place of X in the outcome model, more specifically the impact of that measurement error under different sampling schemes for monitoring locations.

2.3.2 Definition of preferential sampling

In Diggle et al. (2010) preferential sampling is defined as any dependence between the values of the underlying spatial process (X^* in our setting) and the locations at which we observe the process (the monitors in our setting). Mathematically this can be written as

$$p(X^*, S^*) \neq p(X^*)p(S^*), \tag{2.5}$$

where S^* is a random variable to denote the locations at which we observe X^*. An example of this would be a network of monitors that are placed to measure high levels of a given process. Most geostatistical procedures assume independence between these two quantities and therefore likelihood based inference in the presence of preferential sampling can lead to bias as the likelihood is misspecified and estimates are no longer assured to be consistent. In their paper they introduce preferential sampling by allowing their locations to be drawn from an inhomogeneous Poisson process of the following form

$$\lambda(S^*) = \exp(\gamma_0 + \gamma_1 X^*), \tag{2.6}$$

where preferential sampling is the scenario when $\gamma_1 \neq 0$. Lee et al. (2015) also introduce preferential sampling of monitors in their simulations by drawing locations from an inhomogeneous Poisson process whose intensity function depends on observed covariates and unobserved spatial features of the process. We will define preferential sampling in a related, though slightly different way, which we feel is illuminating for studies involving predicted air pollution from monitors. We will investigate scenarios in which sampling depends on the population density with which subjects in the second stage model are drawn from. This can be written as

$$P(S, S^*) \neq P(S)P(S^*) \tag{2.7}$$

where S now denotes the locations at which we observe subjects from the second stage analysis. This does not strictly imply preferential sampling as defined in equation 2.5, though it will meet that criteria for preferential sampling if the population density is associated with the exposure surface. We saw in section 2.2 that the location of monitors appeared non-random and it seemed that there were more monitors in areas that have both a higher population density and higher pollution levels. The reason for defining preferential sampling in this paper as being related to population density is that this scenario commonly occurs in air pollution epidemiology and the extent to which it affects inference is unclear. Previous studies have shown the negative impact that preferential sampling can have on estimation of variograms or other features of the process, however, the main interest in air pollution epidemiology is the second stage outcome model that uses predictions of the process from the first stage modeling.

2.4 Understanding bias and variance of $\hat{\beta}_1$

We now take a step back to provide mathematical justification for our claim that preferential sampling can improve estimation in two stage analyses of air pollution. We will focus on results regarding β_1 as this is the parameter of interest in most environmental epidemiology studies examining the effect of $PM_{2.5}$ on a health outcome. We will use the same notation as before and define C_i to be the vector of covariates for subject i, and C_j^* to be the vector of covariates for monitor j. As a simplifying assumption we assume that the joint distribution of all the necessary quantities (Y, X, X^*, C, C^*) follows a multivariate normal distribution, as this will simplify some of the algebraic operations. We further impose the following models:

$$Y = \beta_0 + X\beta_1 + \epsilon \tag{2.8}$$

$$X = C\alpha + \epsilon_x \tag{2.9}$$

$$X^* = C^*\alpha + \epsilon_x^* \tag{2.10}$$

Where ϵ is a mean zero vector of i.i.d noise, and ϵ_x and ϵ_{x^*} are mean zero vectors of noise. Notice that we do not impose any independence assumptions about ϵ_x and ϵ_{x^*} allowing for spatial structure in the residuals. Conditional on the estimates $\hat{\alpha}$ and $\hat{\phi}$ from the first stage analysis, we estimate exposure via equation 2.4. Our interest lies in the distribution of $\hat{\beta}_1$, the estimate of β_1 we get in the second stage of the model when we use W instead of X.

2.4.1 Bias of $\hat{\beta}_1$

To examine the bias we can look at the conditional distribution of Y given W. Since we defined everything to be jointly normal, the joint distribution of Y and W can we written as

$$\begin{pmatrix} Y \\ W \end{pmatrix} \sim N\left\{ \begin{pmatrix} \mu_y \\ \mu_w \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \sigma_{yw} \\ \sigma_{yw} & \sigma_w^2 \end{pmatrix} \right\} \tag{2.11}$$

Which leads to

$$Y|W \sim N\left(\mu_y + \frac{\sigma_{yw}}{\sigma_w^2}(W - \mu_w), \sigma_y^2 - \frac{\sigma_{yw}}{\sigma_w^2}\right) \tag{2.12}$$

The coefficient of interest is the one that lies in front of W in the mean component of the above conditional distribution. Using this fact and defining $\theta = [\alpha, \phi]$ we can say that

$$\begin{aligned} E(\hat{\beta}_1) &= \frac{\sigma_{yw}}{\sigma_w^2} \\ &= \beta_1 \frac{cov(X, W)}{Var(W)} \\ &= f(\hat{\theta}) \end{aligned} \tag{2.13}$$

and details of this derivation as well as the exact expression for $f(\hat{\theta})$ can be found in the appendix. One important thing to note is that when $\hat{\theta} = \theta$ then $f(\hat{\theta}) = \beta_1$ and there exists no bias. This shows that the small sample bias in estimating β_1 is a function of $\hat{\alpha}$ and $\hat{\phi}$, so

if we knew the true parameters from the exposure model then we would get an unbiased estimate of β_1 in the outcome model. To gain more intuition into this bias we can perform a taylor series expansion of $f(\hat{\theta})$ around $f(\theta)$.

$$f(\hat{\theta}) - f(\theta) \approx \frac{\partial f(\theta)}{\partial \theta}(\hat{\theta} - \theta) + \frac{1}{2}(\hat{\theta} - \theta)^T \frac{\partial^2 f(\theta)}{\partial \theta \partial \theta^T}(\hat{\theta} - \theta) \tag{2.14}$$

and now we can take the expectation on both sides with respect to the distribution governing the monitoring locations. Denoting these expectations by $E_{S^*}()$ we see that

$$\begin{aligned} E_{S^*}\left(f(\hat{\theta}) - f(\theta)\right) &= E_{S^*}(\hat{\beta}_1 - \beta_1) \\ &\approx \frac{\partial f(\theta)}{\partial \theta} E_{S^*}(\hat{\theta} - \theta) + \frac{1}{2} Tr\left(\frac{\partial^2 f(\theta)}{\partial \theta \partial \theta^T} Var_{S^*}(\hat{\theta} - \theta)\right) \\ &+ \frac{1}{2} E_{S^*}(\hat{\theta} - \theta)^T \frac{\partial^2 f(\theta)}{\partial \theta \partial \theta^T} E_{S^*}(\hat{\theta} - \theta) \end{aligned} \tag{2.15}$$

So we've shown that the unconditional bias (no longer conditional on an estimate of θ) is a function of the bias and variance of $\hat{\theta}$.

2.4.2 Variance of $\hat{\beta}_1$

To gain intuition into the variance of $\hat{\beta}_1$ we can look at var$(X - W)$, the variance of the measurement error. While this does not translate directly into the variance of $\hat{\beta}_1$, it is well understood that increasing the amount of measurement error in an exposure will lead to increased variance in the effect of that exposure on the outcome, regardless of the form of the measurement error. To better understand this quantity we can assume that we know the true model parameters and without loss of generality that the mean of the exposure is zero. We will also write our estimated exposure in a somewhat more general way as $W = \sum_{i=1}^{n^*} w_i X_i^*$, where the weights w_i are a function of the distance between X and X_i^* and sum to one. We can write the variance as

$$\text{var}(X - W) = \text{var}(X - \sum_{i=1}^{n^*} w_i X_i^*)$$

$$
\begin{aligned}
&= \text{var}(X) + \sum_{i=1}^{n^*}\sum_{j=1}^{n^*} w_j w_k cov(X_j^*, X_k^*) - 2\sum_{i=1}^{n^*} w_i cov(X, X_i^*) \\
&= \text{var}(X) + \sum_{i=1}^{n^*}\sum_{j=1}^{n^*} w_j w_k cov(X_j^*, X_k^*) - 2\sum_{i=1}^{n^*}\sum_{j=1}^{n^*} w_i w_j cov(X, X_i^*) \\
&= \text{var}(X) + \sum_{i=1}^{n^*}\sum_{j=1}^{n^*} w_j w_k \left(cov(X_j^*, X_k^*) - 2cov(X, X_i^*)\right) \qquad (2.16)
\end{aligned}
$$

and it is of interest for us to examine how this quantity changes across different sampling schemes. We're interested in the behavior of $E_{pref}(\text{var}(X - W))$ compared with $E_{unif}(\text{var}(X - W))$, where the two expectations are taken with respect the to the distribution of the monitors under a preferential and uniform sampling scheme respectively. If we take the expectation on both sides of equation 2.17 we see that

$$
E_{S^*}(\text{var}(X - W)) = E_{S^*}(\text{var}(X)) + \sum_{i=1}^{n^*}\sum_{j=1}^{n^*} E_{S^*}\left[w_j w_k \left(cov(X_j^*, X_k^*) - 2cov(X, X_i^*)\right)\right] \qquad (2.17)
$$

The first term on the right hand side of the equation does not change across monitoring schemes so it is not of interest to us when comparing uniform and preferential sampling. The other terms are where we see changes under preferential sampling. The term involving $E_{S^*}(w_j w_k cov(X, X_i^*))$ goes up under preferential sampling since we place monitors closer on average to the location where we are trying to estimate X. This term going up leads the overall measurement error variance to go down. However, The term $E_{S^*}(w_j w_k cov(X_j^*, X_k^*))$ also goes up under preferential sampling, since monitors are now closer together on average and this leads the overall measurement error variance to rise. This illustrates the trade-off that comes with preferential sampling. On one hand we should preferentially sample monitors so that their exposure value is more correlated with the subject specific exposure values, while on the other hand we can't preferentially sample too much as the monitors will be too close to each other. We show how this trade-off also manifests for $var(\hat{\beta}_1)$ under some simplifying assumptions in the appendix.

2.5 Simulation study

We will now illustrate the impact of preferential sampling with a simulation study. To simplify visualizations and interpretation of results we restrict attention to the one-dimensional setting on the unit interval though we expect these results to hold in the two-dimensional setting seen in environmental studies. We can let s represent location and treat values of s between 0.6 and 0.9 as being 'urban' by defining population density to be higher in these locations, with the highest population between 0.7 and 0.8. We simulate exposure using equation 2.1 where the mean component of the model is a linear function of covariates and the covariance is defined by an exponential covariance function. The exponential covariance takes the following form:

$$C(d) = \exp(-d/\phi) \tag{2.18}$$

and we set $\phi = 0.05$. The mean component of the model is defined by $C\alpha$ where C consists of an intercept, $1(0.7 < s < 0.8)$, $(s - 0.6) * 1(0.6 < s \leq 0.7)$, and $(s - 0.9) * 1(0.8 \leq s < 0.9)$. The first covariate is simply an indicator of being in the high density area while the other two are linear functions of s that allow exposure to linearly decrease away from the high density area. We set $\alpha = [5, 3, 3, -3]$. An illustration of the population density and a subsequent realization of the exposure surface can be found in figure 2.2.
We simulate our outcome from

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{2.19}$$

with $\beta_0 = 100$, $\beta_1 = 5$, and ϵ is an i.i.d vector of mean zero Normal errors with variance 9. The only thing left to define is the sampling scheme we use to draw monitoring locations. We use a simple, interpretable scheme where we draw monitors at given locations with probabilities proportional to population density raised to a variety of powers. If we let D represent population density at a location, then we sample monitoring sites proportional to D^p and we vary p across a range of values. p has a nice interpretation in terms of

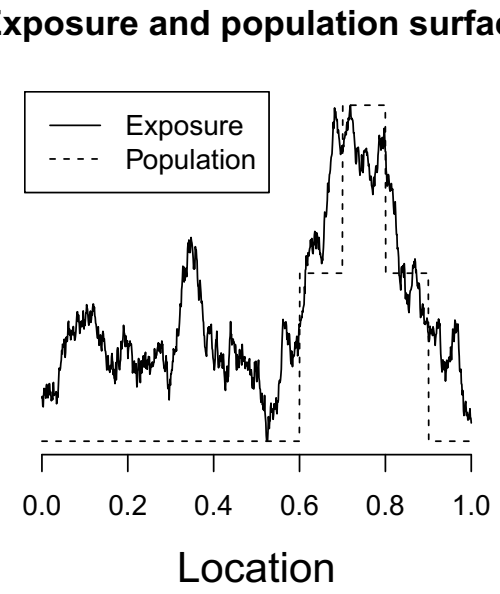

Figure 2.2: Illustration of population and exposure across area of interest

preferential sampling. $p = 0$ represents sampling under CSR, while $p > 0$ represents sampling areas of higher population densities. We sample $n^* = 30, 40, 50$ monitors to see the impact that the number of monitors plays in the performance of various sampling schemes. In all situations we simulate 10000 data sets, and exclude simulations where monitors are not placed in each of the three elevated areas of population density to ensure convergence of parameters.

2.5.1 Impact on exposure estimation

In light of our analytic results that show bias is a function of the bias and variance of the exposure model parameter estimates we can look at the impact that varying p has on estimating θ. Table 2.1 shows the estimates of the exposure model parameters across the simulations for a variety of preferential sampling parameters and number of monitors. Both α and ϕ are estimated with little to no bias under any sampling scheme. The biggest differences seen between values of p is in the standard errors of the parameter estimates. The standard errors for the range parameter do not seem to differ for different values of p, though the standard errors of the mean parameters vary significantly across values of

p. The intercept standard errors grow slightly as we increase p, though the differences are fairly small. The values of the parameters representing differences between high and low population areas have drastically lower standard errors under preferential sampling. For instance under $n^* = 30$ the standard error of $\hat{\alpha_2}$ drops from 0.85 under CSR to 0.69 under p=1. Similar results are seen for $\hat{\alpha_3}$ as the standard error goes from 0.89 to 0.67.

	p	$\hat{\alpha_0}$	$\hat{\alpha_1}$	$\hat{\alpha_2}$	$\hat{\alpha_3}$	$\hat{\phi}$
$n^* = 30$	0.0	5.00 (0.19)	3.00 (0.47)	3.00 (0.85)	- 3.01 (0.89)	0.05 (0.06)
	0.5	5.00 (0.19)	3.00 (0.44)	2.98 (0.76)	-3.00 (0.75)	0.05 (0.06)
	1.0	5.00 (0.20)	3.00 (0.42)	3.01 (0.69)	-3.02 (0.71)	0.05 (0.06)
	1.5	5.00 (0.22)	3.00 (0.44)	3.00 (0.69)	-3.00 (0.67)	0.05 (0.06)
	2.0	5.00 (0.25)	3.01 (0.44)	3.00 (0.68)	-3.00 (0.67)	0.04 (0.07)
$n^* = 40$	0.0	5.00 (0.18)	3.00 (0.45)	2.99 (0.74)	-3.00 (0.75)	0.05 (0.05)
	0.5	5.00 (0.18)	3.00 (0.42)	3.00 (0.65)	-3.00 (0.66)	0.05 (0.05)
	1.0	5.00 (0.19)	3.00 (0.41)	3.00 (0.60)	-3.00 (0.60)	0.04 (0.04)
	1.5	5.00 (0.20)	3.01 (0.42)	3.01 (0.59)	-3.02 (0.57)	0.04 (0.05)
	2.0	5.00 (0.23)	3.00 (0.43)	3.00 (0.59)	-2.99 (0.60)	0.04 (0.05)
$n^* = 50$	0.0	5.00 (0.18)	2.99 (0.43)	2.99 (0.66)	-2.99 (0.68)	0.04 (0.03)
	0.5	5.00 (0.18)	3.00 (0.41)	2.99 (0.60)	-2.99 (0.60)	0.04 (0.03)
	1.0	5.00 (0.19)	3.00 (0.42)	3.00 (0.55)	-2.99 (0.56)	0.04 (0.04)
	1.5	5.00 (0.20)	3.00 (0.41)	3.00 (0.54)	-3.00 (0.54)	0.04 (0.03)
	2.0	5.00 (0.21)	3.01 (0.42)	3.01 (0.54)	-3.01 (0.54)	0.04 (0.04)

Table 2.1: Mean of estimated exposure model parameters across 10000 simulations. Empirical standard errors are in parentheses

We can also look at the variance of the estimated exposure itself. Figure 2.3 shows the variance of $(X - W)$ for a variety of preferential sampling parameters, p, and number of monitors, n^*. This represents the magnitude of the measurement error induced by using monitors to estimate exposure. We discussed in section 2.4 the trade-off that occurs under preferential sampling between placing monitors too close together, and placing them near the locations of the subjects in the second stage of the model. We see that the overall variance is lowest under preferential sampling, around $p = 1$, indicating that the gain from placing monitors near the subjects is outweighing the loss induced by putting monitors close together when monitors are preferentially sampled. To gain further intuition into this trade-off we separate the measurement error variance into a rural and urban component.

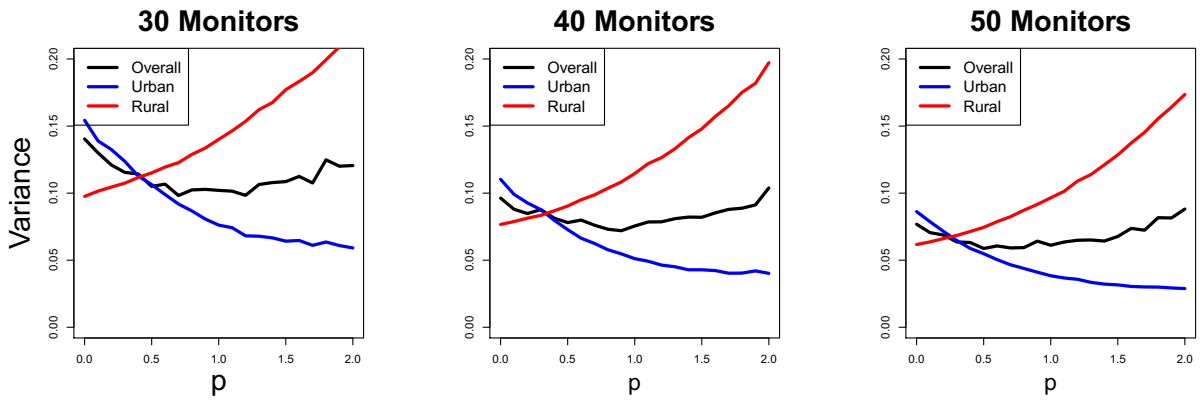

Figure 2.3: The empirical measurement error variance across 10000 simulations for a variety of values for p

The urban components enjoys significantly smaller variation under preferential sampling, which is to be expected as we are placing more monitors in these locations when $p > 0$. The rural component on the other hand suffers from more variation under preferential sampling, which is intuitive because we placed less monitors in those areas. More subjects live in the urban areas, therefore the overall variance is driven by the urban variance, and the ideal trade-off between reducing variance for urban locations and increasing it for rural locations seems to occur around $p = 1$.

2.5.2 Impact on outcome model estimation

To fully understand the impact that preferential sampling plays on the estimates of the second stage outcome model we fit three different outcome models: 1) A model that regresses Y on W; 2) A model that regresses Y on the estimated exposure we would get if we knew the true values of θ; and 3) A model that regresses Y on the exposure we would get if we misspecify the exposure model and do not include population into the set of covariates, C. The first situation is what is done in practice, the second model will show us if and how preferential sampling improves estimation beyond any improvements in exposure model estimation, and the third model is a realistic scenario in which we misspecify how population enters into the exposure model. This is arguably the worst type of misspecification as we are leaving it out completely, but it should lend intuition into what happens when the exposure model is incorrect. We will restrict attention to the estimation

of β_1 as this is the effect of interest.

Figure 2.4 shows the absolute bias of the outcome model parameters under the three models and a variety of values of p.

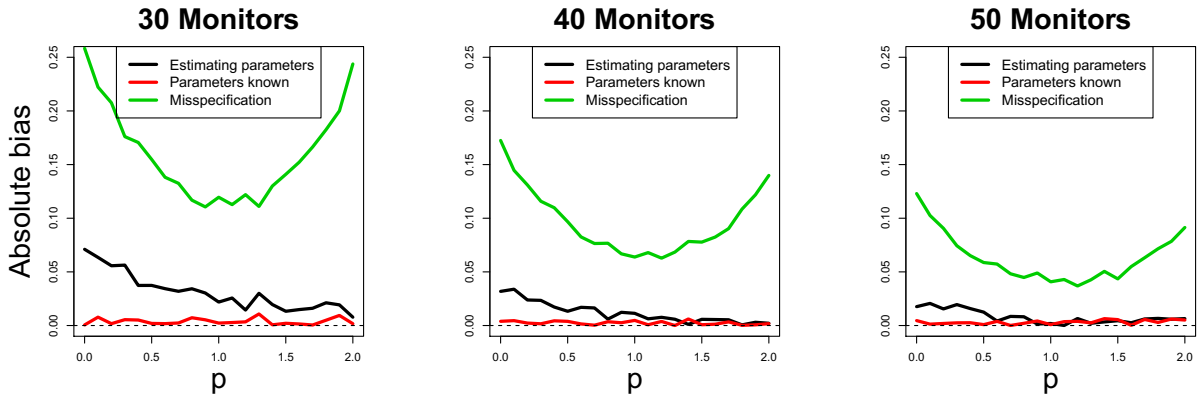

Figure 2.4: The bias of outcome model estimates across 10000 simulations for a variety of values for p

We see that under any of the three models, preferential sampling outperforms CSR with respect to bias, although it should be noted that the magnitude of the bias relative to the true effect size is not substantial under any of the models, with the exception of the misspecified model. The black line in the plots shows that what little bias does exist under $p = 0$ decreases as we increase p. The red line shows that there is no bias for any value of p when we know the true model parameters and this highlights our result from section 2.4 that says the bias is a function of the bias and variance of the exposure model parameters. The most interesting of the lines, is the green line, which shows that when we misspecify the model, $p = 1$ leads to the smallest amount of bias.

Figure 2.5 shows the empirical variance across the 10000 simulations of $\hat{\beta}_1$, and we see a similar U-shape trend in all three models with varying degrees of magnitude. The greatest gains from preferential sampling occur in both the misspecified model and the model in which we estimate the exposure model parameters. For $N^* = 30$ the variance drops from 0.21 at $p = 0$ to 0.13 at $p = 1$ for the model that estimates all parameters, and it drops from 0.19 to 0.12 for the misspecified model. The gains are not as large when we know the true model parameters beforehand, but even then the variance drops from 0.12 to 0.10 when we go from $p = 0$ to $p = 1$. Not surprisingly, the variance tends back upwards after $p = 1$

as we approach $p = 2$ illustrating the trade-off that occurs under preferential sampling. Similar results are found for $n^* = 40, 50$ as the variance drops under $p = 1$, but increases as we preferentially sample too far.

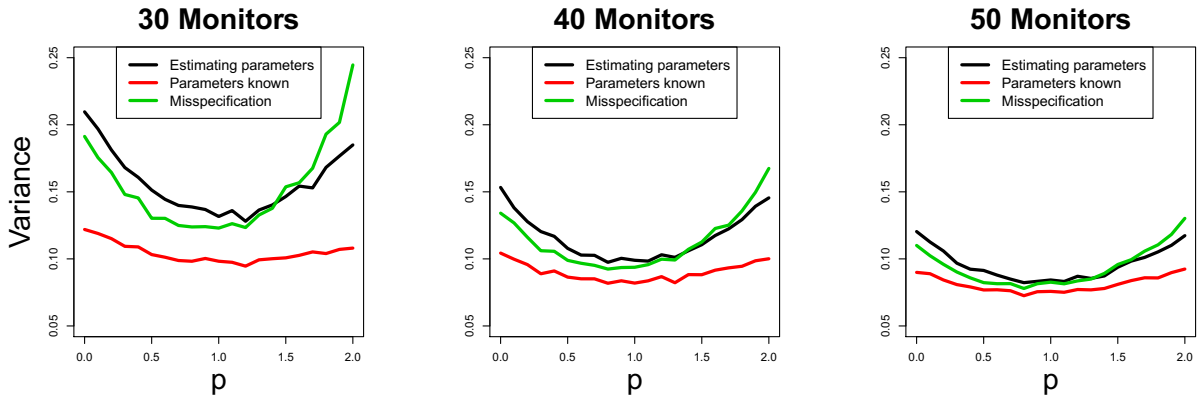

Figure 2.5: The empirical variance of outcome model estimates across 10000 simulations for a variety of values for p

2.6 AQS monitoring system

Now that we have gained insight into the effects of preferential sampling of monitors it is of interest to relate these results back to the motivating example. Figure 2.1 shows the location of the EPA's AQS monitoring system across New England. Data from these monitoring sites is publicly available and has been used in a vast number of environmental studies relating $PM_{2.5}$ to various health outcomes.

2.6.1 One parameter preferential sampling model

To relate the locations of the EPA monitors back to the results we have seen we can impose a simple model for the locations of the monitors. We can split New England into a very fine grid that consists of K grid cells. For each grid cell define D_k to be a measure of population density in grid cell k. We will define D_k to be the number of census tracts within 0.3 degrees of the center of grid k. Now further define

$$Z_k = \begin{cases} 1 & \text{grid cell k has a monitor} \\ 0 & \text{grid cell k does not have a monitor} \end{cases} \tag{2.20}$$

Where $\sum_k Z_k = n^*$, so the number of monitors is fixed. Then the joint distribution of $z = [z_1 ... z_K]$ can be decomposed into successive conditionals and written out as

$$\begin{aligned} P(Z = z) &= P(Z_1)\ P(Z_2|Z_1 = z_1)\ ...\ P(Z_K|Z_1 = z_1 ... Z_{K-1} = z_{K-1}) \\ &= \prod_{i=1}^{n^*} \text{categorical}(w_1, ..., w_K) \end{aligned} \tag{2.21}$$

and we define the vector w as follows:

$$w_k = \begin{cases} 0 & \text{grid cell k already has a monitor} \\ D_k^p & \text{grid cell k does not have a monitor} \end{cases} \tag{2.22}$$

We have now defined a joint probability model for the location of the monitors that depends on a single parameter, p. We can use this model and the location of the EPA monitors to estimate $\hat{p}$ via maximum likelihood, which will provide a measure of how preferentially chosen with respect to population density the monitors are.

We obtain $\hat{p} = 0.64$ for the AQS monitoring system, which tells us that that monitors are in fact preferentially sampled with respect to population density. In light of our previous results this suggests that the manner in which monitors are placed at least in the New England area has led to improved estimation for health effect analyses that rely on an estimated exposure. Obviously the mechanism by which monitor locations are chosen is more complex than the one parameter model we have introduced here, but the results do provide some intuition as to the accuracy of estimates based on this monitoring system. This also provides guidance both for future researchers using monitors to estimate exposure, and for guidance on where to place new monitors. If for instance, in a different study researchers were to estimate $\hat{p}$ for their monitoring system and find that it were near 0 or even negative then additional work would have to be done to correct for bias in the health effect estimate. It's possible that an algorithm could be devised which would take a set of monitors and select which ones to use to achieve a given level of preferential sampling that would be beneficial for inference.

2.6.2 Sampling new monitors in New England

The left side of figure 2.1 gives us a realistic view of the true $PM_{2.5}$ surface across New England, and we can use this to examine the impact that different monitoring schemes would have on health effect estimates in epidemiological studies. We can again discretize New England into a fine grid and then select locations to place monitors under CSR and a variety of values of p. Using the values, D_k, for a grid cell k, we can sample locations proportionally to D_k^p and vary p just as we did in our simulation study. The difference here is that our exposure will be taken from the predicted satellite surface seen in figure 2.1 and will therefore be more realistic to what is actually seen when performing these studies. Once we have a set of monitors we simulate outcomes from the same model as in our simulation study

$$Y = \beta_0 + \beta_1 X + \epsilon, \tag{2.23}$$

where ϵ is a vector of mean zero noise with variance 9 and $\beta = (100, 5)$. Then we can sample $n^* = 83$ monitoring sites without replacement, as this is the number of monitors that exist in the EPA AQS monitoring system. Figure 2.6 shows the corresponding results across 500 simulations. We see that the 95% confidence bands for $\hat{\beta}_1$ decrease in width as p grows larger indicating that the variability in our estimates decreases as we preferentially sample. The confidence interval for CSR goes from 1.4 to 12.1, while the interval under $p = 1$ goes from 3.1 to 7.9. The interval gets even smaller as p grows, although it seems that it comes with some sort of bias. The right hand panel shows the means across the 500 simulations of $\hat{\beta}_1$ and we can see that the estimates are roughly unbiased near $p = 1$ while there is bias under CSR or extreme preferential sampling $p = 2$.

2.7 Discussion

In this paper, we have illustrated the impact that preferential sampling can have on exposure prediction and outcome model estimation in two stage analyses. We defined preferential sampling in the context of environmental studies when the location of monitors

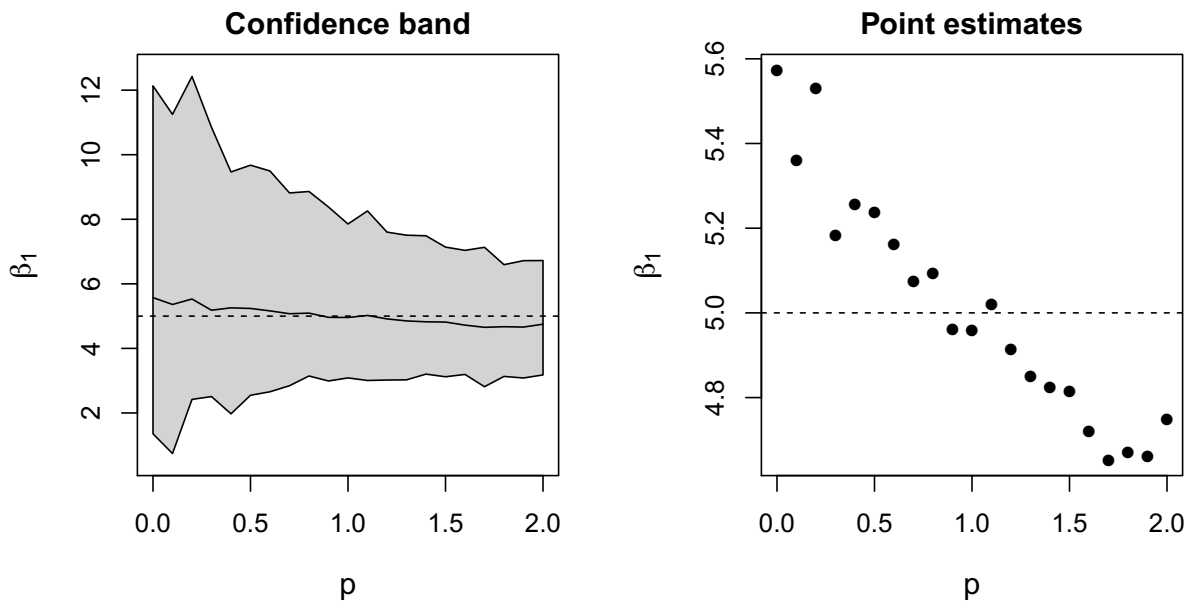

Figure 2.6: Estimates of β_1 after sampling monitors within New England. Left panel shows the confidence band representing the 2.5^{th} and 97.5^{th} quantiles of the 500 simulations. The right panel shows the mean across 500 simulations

is likely to be associated with population density. Through analytic arguments and simulation studies we have shown how inference in two stage analyses that use predicted exposures varies across different scenarios. Finally, combining our previous results with a simple probabilistic model for the EPA monitoring locations we have shown that the EPA monitoring system for $PM_{2.5}$ is in fact preferentially sampled and that this likely is a good thing for estimation of the effects of interest. These results are novel in that they contrast many previous looks into preferential sampling that have shown preferential sampling to lead to poor inference relative to completely random sampling.

That our results differ from previous literature on preferential sampling is not surprising as we take a different approach to defining preferential sampling. Defining preferential sampling with respect to population density is a more specific scenario, and is in fact a subset of the scenarios proposed in Diggle et al. (2010); Lee et al. (2015). We think that this is a very important scenario to look at, however, as it is a scenario that we feel is very likely to occur in practice and we even showed that it does occur when using the EPA's AQS monitoring system. The main intuition behind the idea that preferential sampling is

good for estimation has to do with the fact that while estimation of the exposure surface might be biased under preferential sampling (Diggle et al., 2010), it is more accurately estimated in areas where the majority of the population of interest resides. Possibly more importantly, the standard errors of our exposure estimates are smaller in areas where the majority of the population lives, which effectively reduces the amount of measurement error induced by using estimates of exposure. This reduction in measurement error leads to substantial reductions in the variance of estimates from second stage outcome models, which are usually of interest in air pollution epidemiology.

Our results can be used to help interpret past and future studies that use monitors to predict exposure in environmental studies. Our results agree with the claim made by Szpiro and Paciorek (2013) that the density of the monitor locations should agree with the density of the subject locations in a two stage analysis. While it is difficult to say whether this holds exactly in any study, we have proposed a simple method to check for preferential sampling with respect to population density that can be used to gain intuition to whether the two densities are close enough to lead to valid inference. This among with our analytic and simulation results should help to guide further researchers on investigating the health impacts of air pollutants.

2.8 Appendix

2.8.1 Details of bias calculation from section 2.4

We define the vector (Y, X, X^*, C, C^*) to be jointly normal, and define this distribution as

$$\begin{pmatrix} Y \\ X \\ X^* \\ C \\ C^* \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_y \\ \mu_x \\ \mu_{x^*} \\ \mu_c \\ \mu_{c^*} \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \sigma_{yx} & \Sigma_{yx^*} & \Sigma_{yc} & \Sigma_{yc^*} \\ & \sigma_x^2 & \Sigma_{xx^*} & \Sigma_{xc} & \Sigma_{xc^*} \\ & & \Sigma_{x^*} & \Sigma_{x^*c} & \Sigma_{x^*c^*} \\ & & & \Sigma_c & \Sigma_{cc^*} \\ & & & & \Sigma_{c^*} \end{pmatrix} \right)$$

Recall that we also defined our exposure, W as

$$\begin{aligned} W &= \mu_X(\hat{\alpha}) + \Sigma_{X,X^*}(\hat{\phi})\Sigma_{X^*,X^*}(\hat{\phi})^{-1}(X^* - \mu_{X^*}(\hat{\alpha})) \\ &= C\hat{\alpha} + \Sigma_{X,X^*}(\hat{\phi})\Sigma_{X^*,X^*}(\hat{\phi})^{-1}(X^* - C^*\hat{\alpha}) \end{aligned}$$

Where the only random variables, C, C^*, and X^* are normally distributed and therefore W is normally distributed. We are interested in the coefficients of the model that regresses Y on W, i.e the conditional distribution of Y given W, which can now be written as

$$Y|W \sim N\left(\mu_y + \frac{\sigma_{yw}}{\sigma_w^2}(W - \mu_w), \sigma_y^2 - \frac{\sigma_{yw}}{\sigma_w^2}\right)$$

and the coefficient of interest is the one that lies in front of W in the mean component of the above conditional distribution. Using this we can say that

$$\begin{aligned}
E(\hat{\beta}_1) &= \frac{\sigma_{yw}}{\sigma_w^2} \\
&= \frac{cov(Y, W)}{cov(W, W)} \\
&= \frac{cov(X\beta + \epsilon, W)}{cov(W, W)} \\
&= \beta_1 \frac{cov(X, W)}{cov(W, W)} \\
&= \beta_1 \left\{\frac{A}{B}\right\}
\end{aligned}$$

Where

$$\begin{aligned}
A &= \alpha\Sigma_c\hat{\alpha} + \alpha\hat{\Sigma}_{xx^*}\hat{\Sigma}_{x^*}^{-1}\Sigma_{c^*c}\alpha - \alpha\hat{\Sigma}_{xx^*}\hat{\Sigma}_{x^*}^{-1}\Sigma_{c^*c}\hat{\alpha} + \hat{\Sigma}_{xx^*}\hat{\Sigma}_{x^*}^{-1}\Sigma_{x^*x} \\
B &= \hat{\alpha}\Sigma_c\hat{\alpha} + \alpha\hat{\Sigma}_{xx^*}\hat{\Sigma}_{x^*}^{-1}\Sigma_{c^*}\hat{\Sigma}_{x^*}^{-1}\hat{\Sigma}_{x^*x}\alpha + \hat{\Sigma}_{xx^*}\hat{\Sigma}_{x^*}^{-1}\Sigma_{x^*}\hat{\Sigma}_{x^*}^{-1}\hat{\Sigma}_{x^*x} \\
&\quad + \hat{\alpha}\hat{\Sigma}_{xx^*}\hat{\Sigma}_{x^*}^{-1}\Sigma_{c^*}\hat{\Sigma}_{x^*}^{-1}\hat{\Sigma}_{x^*x}\hat{\alpha} + 2\hat{\alpha}\hat{\Sigma}_{xx^*}\hat{\Sigma}_{x^*}^{-1}\Sigma_{c^*c}\alpha - 2\hat{\alpha}\hat{\Sigma}_{xx^*}\hat{\Sigma}_{x^*}^{-1}\Sigma_{c^*c}\hat{\alpha} \\
&\quad - 2\alpha\hat{\Sigma}_{xx^*}\hat{\Sigma}_{x^*}^{-1}\Sigma_{c^*}\hat{\Sigma}_{x^*}^{-1}\hat{\Sigma}_{x^*x}\hat{\alpha}
\end{aligned}$$

2.8.2 Trade-off for variance of $\hat{\beta}_1$

Before we illustrated the trade-off that comes with preferential sampling for $var(X - W)$, the measurement error variance. We used this to show how preferential sampling could lead to less measurement error variance and therefore less variance in estimating β_1. Here we illustrate directly how this trade-off manifests in the estimation of $\hat{\beta}_1$ by making a couple simplifying assumptions and approximations. Let's assume that our exposure surface follows equation 2.10 and that we estimate exposure W via

$$W_i = C_i\hat{\alpha}$$

where C_i represents a single covariate and there is no intercept, because it is centered. The exposure model parameter, $\hat{\alpha}$ is estimated using least squares as

$$\hat{\alpha} = \frac{\sum_{j=1}^{n^*} C_j^* X_j^*}{\sum_{j=1}^{n^*} C_j^{*2}}$$

Then conditional on our estimates, W, we estimate the parameter of our outcome model, which again for simplification we assume is centered with no intercept and we estimate via least squares

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} W_i Y_i}{\sum_{i=1}^{n} W_i^2} \\
&= \frac{\sum_{i=1}^{n} \hat{\alpha} C_i Y_i}{\sum_{i=1}^{n} \hat{\alpha}^2 C_i^2} \\
&= \frac{1}{\hat{\alpha}} \frac{\sum_{i=1}^{n} C_i Y_i}{\sum_{i=1}^{n} C_i^2} \\
&= \frac{\hat{\eta}}{\hat{\alpha}}
\end{aligned}$$

where now we have written the estimate of β_1 as a ratio of two random variables, one of which involves the monitor locations and the other involving the subject locations. Now we take the variance of this ratio and apply a taylor series approximation to the variance of a ratio

$$\begin{aligned}
var(\hat{\beta}_1) &= var(\frac{\hat{\eta}}{\hat{\alpha}}) \\
&\approx \left(\frac{E(\hat{\eta})}{E(\hat{\alpha})}\right)^2 \left[\frac{var(\hat{\eta})}{E(\hat{\eta})^2} + \frac{var(\hat{\alpha})}{E(\hat{\alpha})^2} - 2\frac{cov(\hat{\eta}, \hat{\alpha})}{E(\hat{\alpha})E(\hat{\eta})}\right]
\end{aligned}$$

We also assume that $E(\hat{\alpha}) = \alpha$ regardless of the sampling scheme for the location of the monitors. Since $\hat{\eta}$ is not dependent on the monitor locations we can now see that only two terms in the above equation for the variance of $\hat{\beta}_1$ depend on the locations of the

monitors. $cov(\hat{\eta}, \hat{\alpha})$ and $var(\hat{\alpha})$ will both change as a function of the monitors. Writing these terms out we see that

$$\begin{aligned} cov(\hat{\eta}, \hat{\alpha}) &= cov\left(\frac{\sum_{i=1}^{n} C_i Y_i}{\sum_{i=1}^{n} C_i^2}, \frac{\sum_{j=1}^{n^*} C_j^* X_j^*}{\sum_{j=1}^{n^*} C_j^{*2}}\right) \\ &= \frac{\beta_1}{(\sum_{i=1}^{n} C_i^2)(\sum_{j=1}^{n^*} C_j^{*2})} \sum_{i=1}^{n} \sum_{j=1}^{n^*} C_i C_j^* cov\left(\epsilon_{x_i}, \epsilon_{x_j^*}\right) \end{aligned}$$

Which will go up under preferential sampling because the locations of the monitors will be closer to the locations of the subjects. Now we can look at

$$\begin{aligned} var(\hat{\alpha}) &= cov(\hat{\alpha}, \hat{\alpha}) \\ &= cov\left(\frac{\sum_{j=1}^{n^*} C_j^* X_j^*}{\sum_{j=1}^{n^*} C_j^{*2}}, \frac{\sum_{j=1}^{n^*} C_j^{*2} X_j^*}{\sum_{j=1}^{n^*} C_j^{*2}}\right) \\ &= \frac{1}{(\sum_{j=1}^{n^*} C_j^{*2})^2} \sum_{j=1}^{n^*} \sum_{k=1}^{n^*} C_j^* C_k^* cov\left(\epsilon_{x_j^*}, \epsilon_{x_k^*}\right) \end{aligned}$$

Which will also go up under preferential sampling because the monitors will be located more closely to each other. Now we have illustrated the trade-off that comes with preferential sampling. On one hand the variance of $\hat{\beta}_1$ will go down under preferential sampling, since the monitors are closer to the subjects and $cov(\hat{\eta}, \hat{\alpha})$ goes up leading the overall variance to go down. On the other hand preferential sampling makes $var(\hat{\alpha})$ go up as well, which increases the variance of $\hat{\beta}_1$ as monitors get closer together.

Utilizing validation data: A Bayesian variable selection approach to adjust for confounding

Joseph Lawrence Antonelli

Department of Biostatistics

Harvard Graduate School of Arts and Sciences

Francesca Dominici

Department of Biostatistics

Harvard Chan School of Public Health

3.1 Introduction

The analysis of large administrative databases is an increasingly important topic as these databases become more widespread and methods to handle such large data sources become available. In comparative effectiveness research and many other related fields, these data sources are frequently used to estimate the causal effect of an exposure on a particular health outcome. While these data sources are very large allowing us to identify a wide range of potential effects with high statistical power, they typically do not contain a rich enough set of covariates to properly adjust for confounding bias. In some instances, without additional data we may not be able to accurately estimate the effects of interest from administrative health databases. However, in many instances there exists additional "validation" data, which is either a subset of the original data or very similarly structured data set from a different population, that contains far richer covariate information on a much smaller sample. In these instances, researchers can leverage information from the validation data source to fully adjust for confounding in the main data.

There exist two previous approaches in the literature to control for confounding, when validation data is available. The first of which is referred to as propensity score calibration (Stürmer et al., 2005), which builds a "gold standard" propensity score in the validation data and relates it to an "error prone" propensity score that only includes variables in the main data set. Propensity score calibration is useful in settings where outcome information is not available in the validation data as it only uses propensity scores to adjust for confounding. This approach has been shown to work well in some settings (Stürmer et al., 2007), though it relies on a surrogacy assumption, which states that the error prone propensity score is independent of the outcome given the gold standard propensity score and exposure. Another limitation of this method is that it is only applicable to binary exposures as it relies on propensity score models to adjust for confounding. A second approach introduces the idea of conditional propensity scores (McCandless et al., 2012), and takes an approximately Bayesian approach to controlling for confounding using validation data. The main idea behind this approach is to approximate the likelihood of the data accounting for the missing covariates by integrating over the distribution of the so

called conditional propensity score. They show theoretically that the conditional propensity score can eliminate confounding bias, and show via simulation that their method outperforms propensity score calibration. However, this method relies on strong distributional assumptions for the missing conditional propensity score and similarly to propensity score calibration only works for binary exposures.

In this paper, we develop a new approach, which we will refer to as Guided Bayesian Adjustment for Confounding (guided BAC), that overcomes the limitations described above and that is applicable to both binary and continuous exposures. Specifically, our approach is rooted on the idea of Bayesian data augmentation and extends the work by Wang et al. (2012) called Bayesian Adjustment for Confounding (BAC) to scenarios with missing data in the potential confounders. The main idea in Wang et al. (2012) is to apply ideas of Bayesian model averaging (Raftery et al., 1997; Hoeting et al., 1999) to the scenario where interest is in the estimation of an effect of an exposure rather than prediction of an outcome. Our proposed approach will treat the main data and validation data as one combined data set and utilize model averaging to identify which missing confounders are in fact required for valid estimation of the average causal effect, then using the validation data impute these values within a fully Bayesian framework. The proposed approach is appealing for a number of reasons. Typically in similar settings, a particular set of confounders is chosen and then the remaining analysis is performed conditional on this given set of confounders without accounting for the uncertainty in the confounder selection, while our procedure will incorporate uncertainty into the selection of confounders into the causal effect estimates. Another advantage in the scenario of missing data is that estimation relies on choosing a model for each potential confounder that is missing in the main study (Little and Rubin, 2014). It has been noted in some cases that imputation can be robust to model misspecification (Rubin, 1996). However, in many cases it can lead to misleading inference, especially if careful consideration isn't taken as to what enters the imputation model (Schenker and Taylor, 1996; Meng, 1994). It is well known that choosing a correct imputation model is a crucially important aspect of any missing data analysis. Our proposed approach, has the nice feature that relies on variable selection to identify the key confounders (both in the main study and in the

validation). This leads to more robust inference with respect to misspecification of the missing data imputation model, since we are including in the outcome model only the imputed variables that are necessary for confounding adjustment. This proposed method also has advantages over the aforementioned existing approaches as it is valid for continuous and binary exposures, whereas previous methods relied on propensity score models and therefore only can be applied to binary exposures. Finally, performing this analysis within a fully Bayesian framework is useful as we will be able to account for the uncertainty in confounder selection and variable imputation to obtain valid standard errors for the average causal effect of the exposure in a straightforward manner.

The remainder of the paper is structured as follows: Section 3.2 will provide details of the model and prior specification, section 3.3 will highlight the assumptions made in previous approaches as well as in guided BAC, section 3.4 will introduce a simulation illustrating the proposed method and compare it to other approaches in the literature, section 3.5 will analyze the effect of surgical resection on cancer patients in Medicare, and we will conclude in section 3.6 with further discussion.

3.2 Model formulation

3.2.1 With no missing data

We will first present the ideas of model averaging when the goal is valid causal effect estimation. This will effectively be a review of work from Wang et al. (2012); Lefebvre et al. (2014); Wang et al. (2015), but is important in understanding how the proposed approach will be useful in the setting of validation data to control for confounding and how it can be extended to missing data. We define a model for an exposure X, as well as a model for an outcome Y with the goal of estimating the causal effect of X on Y. We also define a set of P fully observed covariates $C = (C_1...C_P)$, and we would like to identify which of these are necessary for controlling for confounding and include those into the outcome model.

In general, our interest will lie in the average causal effect (ACE), which we will define as $\Delta(x_1, x_2)$. We will not utilize potential outcomes notation, however, this quantity is

the difference in potential outcomes under exposure levels x_1 and x_2. We will always be assuming that each subject has a positive probability of receiving each level of exposure and that the potential outcomes are independent of the level of exposure given covariates, C. Formally the ACE is estimated as

$$E_P\left[E(Y|X = x_1, C) - E(Y|X = x_2, C)\right], \tag{3.1}$$

provided that there is no unmeasured confounding and that all necessary confounders are included in C. Our goal is to average over the uncertainty in selecting confounders to include in the analysis while assigning as much weight as possible to those sets that include all necessary confounders. With this in mind we introduce two parameter vectors, $\alpha^x \in \{0,1\}^P$ and $\alpha^y \in \{0,1\}^P$ which dictate which covariates are included in the exposure and outcome models respectively. For example, $\alpha_p^x = 1$ indicates that the p^{th} covariate should be included in the exposure model and vice versa. Conditional on regression parameters and other unknown parameters, we can write out the models for exposure and outcome as follows:

$$f(E(X_i)) = \theta_{x0} + \sum_{p=1}^{P} \alpha_p^x \theta_{xp} C_{ip} \tag{3.2}$$

$$g(E(Y_i)) = \theta_{y0} + \beta X_i + \sum_{p=1}^{P} \alpha_p^y \theta_{yp} C_{ip} \tag{3.3}$$

where i indexes the sampling unit for i=1,...,N, f() and g() are arbitrary link functions, and θ_x and θ_y are regression parameters for the exposure and outcome models respectively. We've also introduced the parameter of interest, β, the effect of the exposure on the outcome conditional on covariates, C. In the setting of a continuous outcome we can define the ACE as

$$\Delta(x_1, x_2) = \beta(x_1 - x_2) \tag{3.4}$$

with β defined as in 3.3. In the setting of non continuous outcomes there is no clear

expression for the ACE as it depends on the marginal distribution of the covariates, C. We can write out the posterior distribution of Δ as

$$P(\Delta|X,Y,C) = \sum_{\alpha^y} P(\Delta|X,Y,C,\alpha^y)P(\alpha^y|X,Y,C), \tag{3.5}$$

where both portions of the right hand side of the equation can be estimated from the data. We define by $\boldsymbol{\alpha^{y*}}$ the minimal model, that is, a model that includes as covariates the minimal set of confounders. then we can rewrite the above as

$$\begin{aligned} P(\Delta|X,Y,C) = &\sum_{\alpha^y \in \boldsymbol{\alpha^{y*}}} P(\Delta|X,Y,C,\alpha^y)P(\alpha^y|X,Y,C) + \\ &\sum_{\alpha^y \notin \boldsymbol{\alpha^{y*}}} P(\Delta|X,Y,C,\alpha^y)P(\alpha^y|X,Y,C) \end{aligned} \tag{3.6}$$

Where the second part of the sum includes models that do not contain the necessary confounders and therefore we will be inducing bias in our estimate of Δ. We want to specify a prior distribution on (α^x, α^y), so that a posteriori we assign high weight to models α^Y that include the minimal model ($\boldsymbol{\alpha^{y*}}$) and small or no weight to models that do not include the minimal model. In Wang et al. (2012, 2015) the authors demonstrated that Δ has a causal interpretation and therefore does not change for all models that include the minimal model.

3.2.2 Prior formulation

Wang et al. (2012) introduces a prior distribution to ensure that the posterior distribution, $P(\alpha^y|X,Y,C)$, assigns most of the posterior mass to models α^Y that include $\boldsymbol{\alpha^{y*}}$. The prior ensures that any covariate that is associated with X, will receive higher weight a priori for entering into the model for Y. The idea behind this is that a covariate could be weakly associated with Y, but strongly associated with X and would therefore introduce confounding bias into our results. This covariate is only weakly associated with Y and therefore may not enter into the outcome model, but our prior will place larger weight to this variable making it more likely to be included into the outcome model. The prior for $\alpha^y|\alpha^x$ is defined as follows:

$$\frac{P(\alpha_p^y = 1|\alpha_p^x = 1)}{P(\alpha_p^y = 0|\alpha_p^x = 1)} = \omega \tag{3.7}$$

$$\frac{P(\alpha_p^y = 1|\alpha_p^x = 0)}{P(\alpha_p^y = 0|\alpha_p^x = 0)} = 1 \text{ for p=1...P} \tag{3.8}$$

where ω is a tuning parameter, which dictates how much prior weight we place for including variables into the outcome model given they are included in the exposure model. Setting $\omega = \infty$ provides the greatest protection against missing important confounders, as it forces any variable included in the exposure model into the outcome model. Setting $\omega = 1$ is analagous to implementing BMA on the outcome model as it places a flat prior on all inclusion probabilities into the outcome model and only utilizes a variable's association with Y when deciding if it should be included into the model.

One potential pitfall of this prior specification is in the context of instrumental variables. Instrumental variables are only associated with X and should not be included into the outcome model as they will introduce bias in the estimation of the causal effect of X on Y. One way to approach this issue is to set $\omega = \omega_0$ where ω_0 is some value between 1 and ∞ that balances our desire to include confounders into the model and our apprehension for including potential instrumental variables into the model. When ω is finite we can implement a conditional prior on $\alpha^x|\alpha^y$ in the following manner:

$$\frac{P(\alpha_p^x = 1|\alpha_p^y = 1)}{P(\alpha_p^x = 0|\alpha_p^y = 1)} = 1 \tag{3.9}$$

$$\frac{P(\alpha_p^x = 1|\alpha_p^y = 0)}{P(\alpha_p^x = 0|\alpha_p^y = 0)} = \frac{1}{\omega} \text{ for p=1...P} \tag{3.10}$$

For $\omega < \infty$ we have the possibility to assign lower probabilities on including variables into the exposure model that are not associated with the outcome. This should reduce the possibility of including instrumental variables into the outcome model, while still increasing the probabilities of including important confounders.

Assuming there is no unmeasured confounding it can be shown that our joint prior on (α^x, α^y) leads to an increase in the posterior probability of including all necessary confounders into the outcome model. If we let $P_1(\alpha^y \in \boldsymbol{\alpha^{y*}}|D)$ be the marginal posterior

probability that α^y contains the minimal model under our joint prior and $P_2(\alpha^y \in \boldsymbol{\alpha^{y*}}|D)$ be the marginal posterior probability under a BMA prior then one can show that

$$P_1(\alpha^y \in \boldsymbol{\alpha^{y*}}|D) \geq P_2(\alpha^y \in \boldsymbol{\alpha^{y*}}|D) \tag{3.11}$$

Details of the proof are in the appendix. This shows how our conditional prior specification assigns more posterior mass to models that contain the true confounder set and therefore is more likely to adjust for confounding bias.

3.2.3 Extension to missing data

We will now extend these ideas to the setting with missing covariates, motivated in particular by the validation data setting where we have a validation data set with information on an additional set of potential confounders that are not measured in the main data. We have a main data set with N_1 subjects and a validation data set with N_2 subjects. We can again define our matrix of covariates as C, though now we introduce U, which represents the subset of the covariates in C that are missing in the main data but observed in the validation data. Covariates 1 through M are fully observed for all $N_1 + N_2 = N$ subjects in the study, while covariates M+1 through P are only available in the N_2 subjects in the validation data. U represents covariates M+1 through P in the N_1 subjects in the main data set. The quantity of interest is the same as before, but is now defined as

$$\begin{aligned} P(\Delta|X,Y,C) &= \int P(\Delta|X,Y,C,U)P(U|X,Y,C)dU \\ &= \int \sum_{\alpha^y} P(\Delta|X,Y,C,U,\alpha^y)P(\alpha^y|X,Y,C,U)P(U|X,Y,C)dU \end{aligned} \tag{3.12}$$

where we are now averaging over both the potential models for Y as well as the missing data U. In practice this will be implemented using MCMC integration where within one MCMC chain we will update the missing data, the choice of outcome model (denoted by α^y) and then the remaining unknown parameters conditional on a given imputed value of the missing data and selected model. To do this we must specify regression models for

variables M+1...P as they are the variables with missing data in this case. Specifically, we regress each missing covariate (p=M,..P+1) to the set of covariates $C_1..C_{p-1}$. We can write out the model for a missing covariate p as follows:

$$h(E(C_{ip})) = \theta_0^p + \sum_{j=1}^{p-1} \theta_j^p C_{ij} \tag{3.13}$$

where i indexes the sampling unit for i=1,...,N_1+N_2, and h() is an arbitrary link function. It is important to note that covariate $C_p, p = M, .., P+1$ (which is missing in the main study but observed in the validation data) is being regressed on the M covariates $(C_1, .., C_M)$ that are available both in the main and in the validation study, but also on a subset of the missing covariates, where the subset of missing covariates included depends on the subscript p.

3.3 MAR, transportability, and other assumptions

Any method attempting to use validation data to control for confounding in a larger data set needs to make assumptions about the data generating mechanism. The existing approaches, such as the ones described in the introduction, implicitly make strong assumptions about the relationship between the main and validation data. In this section we will try to clarify these assumptions and relate them to assumptions our proposed procedure makes.

The propensity score approach of (Stürmer et al., 2005) makes a surrogacy assumption, which states that the error prone propensity score (PS_{ep}) built on only the covariates observed in the main data is independent of the outcome given the exposure and the gold standard propensity score (PS_{gs}) built with the larger set of covariates available in the validation data. Mathematically this assumption is written as

$$PS_{ep} \perp\!\!\!\perp Y | X, PS_{gs} \tag{3.14}$$

Although this assumption seems reasonable, it does place strong restrictions on the direction of confounding bias. The assumption will not hold when the direction of confound-

ing bias from the missing covariates (those only observed in the validation data) differs from the direction of confounding bias from the covariates observed in the main data. This assumption can not be realistically expected to hold in a wide range of scenarios and it has been shown that violations of this assumption can actually lead to larger amounts of bias than simply ignoring the covariates in the validation data (Stürmer et al., 2007).
The conditional propensity score approach of (McCandless et al., 2012) uses similar ideas in the sense that they use propensity score models to adjust for confounding in the main data set. Using the validation data they build the following model

$$\log\left(\frac{p(X_i=1)}{1-p(X_i=1)}\right) = \gamma_1 C_i + \gamma_2 U_i, \tag{3.15}$$

where i indexes the sampling unit for i=1...N_2 as this model is fit in the validation data only. They refer to $\gamma_2 U$ as the conditional propensity score. They estimate the distribution of $\gamma_2 U$ in the validation data and then approximate the full data likelihood by integrating over this distribution in the main data set. Although not explicitly mentioned in the paper, this makes some strong assumptions about the relationships between variables in the main and validation data. The first assumption they make is that the conditional propensity score is normally distributed. They allow the mean of this distribution to depend on the covariates observed in the main data set via

$$\gamma_2 U_i | C_i \sim N(\tilde{\gamma} C_i, \tilde{\tau}^2) \tag{3.16}$$

where i=1...$N_1 + N_2$ and $\tilde{\gamma}$ and $\tilde{\tau}^2$ are estimated from the validation data. This doesn't, however, include information on the exposure or outcome. This is making assumptions about the missingness mechanism of the data as they do not allow the missing quantities to vary due to differences in X or Y. This is a weaker assumption than MCAR, because they allow the missing quantity to depend on C, though stronger than MAR because the missing quantity can not depend on X or Y.
Our approach does not make any assumptions analogous to the surrogacy assumption of propensity score calibration as we are not using any form of regression calibration. We

also do not make as strong of assumptions about the missingness mechanism as the conditional propensity score approach. While we build imputation models conditional only on the observed and missing covariates, information on the exposure and outcome enters into the posterior distribution of the missing data imputations. This means that we are making the assumption that our data is missing at random (MAR), which is a standard assumption made in missing data problems. Both our approach and the other 2 propensity score approaches make the assumption of transportability, which has been discussed in detail in the measurement error literature (Carroll et al., 2006). For our approach this assumption states that

$$P_{main}(Y|X,C,U) = P_{val}(Y|X,C,U) \tag{3.17}$$

$$P_{main}(X|C,U) = P_{val}(X|C,U) \tag{3.18}$$

$$P_{main}(U|C) = P_{val}(U|C) \tag{3.19}$$

The assumptions above are are required because these distributions from the validation data are used to impute potential confounders in the main data set. Misspecification of these distributions would lead to incorrect imputations in the main data set, and our ability to adjust for confounding would be compromised. The extent to which differences in these distributions affects inference in our setting is unknown and a subject of further study.

3.4 Simulation

We simulated data with a binary exposure generated from a probit model. In general our proposed method will work for binary or continuous exposures, but we restrict attention to the former, to be able to compare our method with those relying on propensity scores. We simulate our outcome to be either binary from a probit model or continuous from a normal distribution with $\sigma_y^2 = 5$. We simulated M independent covariates to be observed in the main data set and a subset of these are important confounders. The P-M covariates missing in the main data set were further assumed to be independent of each other and

of the M fully observed covariates, and a subset of these were chosen to be confounders. The coefficients of the models were set to reflect the level of confounding desired in each scenario. Scenarios 1-3 were designed in very similar manners as $N_1 + N_2 = 2000$ in all three of these scenarios, and we varied N_2 the number of subjects in the validation data from 100 to 1000. Scenario 4 is intended to more accurately reflect the data generating mechanism seen in our data analysis in section 3.5 and the sample sizes are chosen to mimic those seen in SEER-Medicare. To study the effectiveness of all the approaches in the simulation study we then varied the number of total covariates (P), the number of missing covariates (P-M), the number of true confounders, the type of covariate missing (binary or continuous), the prevalence of exposure and each of the binary covariates, and the missingness mechanism (MAR or MCAR). Table 3.1 describes the data generating mechanism for each simulation scenario.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
missingness	MCAR	MAR	MCAR	MCAR
# Observed covariates (M)	5	5	5	25
# Missing covariates (P-M)	5	5	5	9
# Observed confounders	2	2	2	6
# Missing confounders	2	2	2	2
# Total covariates	10	10	10	34
# Total confounders	4	4	4	8
Outcome type	Continuous	Continuous	Continuous	Binary
True ACE	5	5	5	0.025
% Confounding bias	30%	30%	20%	30%
Observed covariate type	Continuous	Continuous	Continuous	Binary
Missing Covariate type	Continuous	Continuous	Binary	Binary
Covariate prevalence	NA	NA	11%	Varying
Exposure prevalence	20%	20%	8%	34%

Table 3.1: Description of simulation scenarios

To analyze the data sets we fit a variety of approaches aimed at obtaining the causal effect of X on Y. It is important to note that propensity score approaches are aimed at obtaining an unconfounded estimate, not at capturing the true data generating mechanism. Our approach is fitting the correct model and will therefore be more efficient than corresponding propensity score approaches, which fit misspecified models. In scenarios 1-3 we are deal-

ing with a continuous outcome and binary exposure so the effect of interest, the ACE is defined as

$$\Delta(1,0) = \beta \tag{3.20}$$

where β is the coefficient for the effect of X on Y in all of the outcome regression models that include the minimal model. In scenario 4 we have a binary outcome and the ACE is calculated using the regression coefficients from our outcome model and the empirical distribution of the covariates in the data to obtain the marginal risk difference. We fit the following approaches to estimating $\Delta(1,0)$

1. Naive approach that fits a regression model that only includes all covariates available for the main data set
2. Gold standard approach which pretends we observe U and fits the regression model in the full data set that includes all covariates
3. Validation only approach which fits the regression model with all covariates, but only in the validation data
4. Propensity score calibration approach of Stürmer et al. (2005)
5. Conditional propensity score approach of McCandless et al. (2012)
6. Our proposed approach, which imputes all the covariates missing in the main study, but is "guided" is the sense that only the covariates that are highly likely to be true confounders are included into the outcome model

To examine the performance of the various methods in estimating the ACE we look at three operating characteristics: Bias, mean squared error, and 95% interval coverage. For Bayesian analyses, non informative priors were used for all regression coefficients, and $IG(0.001, 0.001)$ priors were used for all variance parameters.

3.4.1 Scenario 1

Here we present the results when the validation data is a random sample of the full data set (MCAR). Figure 3.1 shows the bias, mean squared error (MSE), and 95% interval coverage of the estimated average causal effect (β from equation 3.3). We see that under this scenario, all of the approaches show little to no bias for any validation sample size, with the exception of propensity score calibration. When the validation data sample size is only 100 our guided BAC approach suffers from a small amount of bias, though it is small relative to the true effect size. This bias likely stems from the fact that the inclusion probabilities (not shown) for the true confounders are not 100% at low sample sizes so we are averaging over some models, which do not include the minimal confounder set. Looking at the MSE we see that guided BAC and conditional propensity score approaches perform the best among all models other than the gold standard, which no model should outperform. The conditional propensity score has a smaller MSE in $N_2 = 100$, and the guided BAC approach has the smallest MSE for the other sample sizes, though the difference is quite small. Importantly, guided BAC outperforms the "validation data only" approach suggesting that imputing the data is improving efficiency of estimates. Looking at the 95% interval coverages we see that as expected, both the validation data only and gold standard approaches always achieve the desired coverage probabilities. Guided BAC achieves the desired coverage at nearly all sample sizes with the exception of $N_2 = 100$, when it only slightly underperforms with a coverage of 0.90. The conditional propensity score approach does not achieve the desired coverage at any sample size though it is attenuating to 0.95 as the validation sample size grows.

3.4.2 Scenario 2

We generated data under the exact same setup as in scenario 1, but the missing covariates in the main study are missing at random and not missing completely at random. Specifically, we selected the subjects in the validation data to be a subset of the main data where they were chosen with probabilities that depended on both the exposure and outcome. If we let I be an indicator of missingness then we use equation 3.21 to select the validation

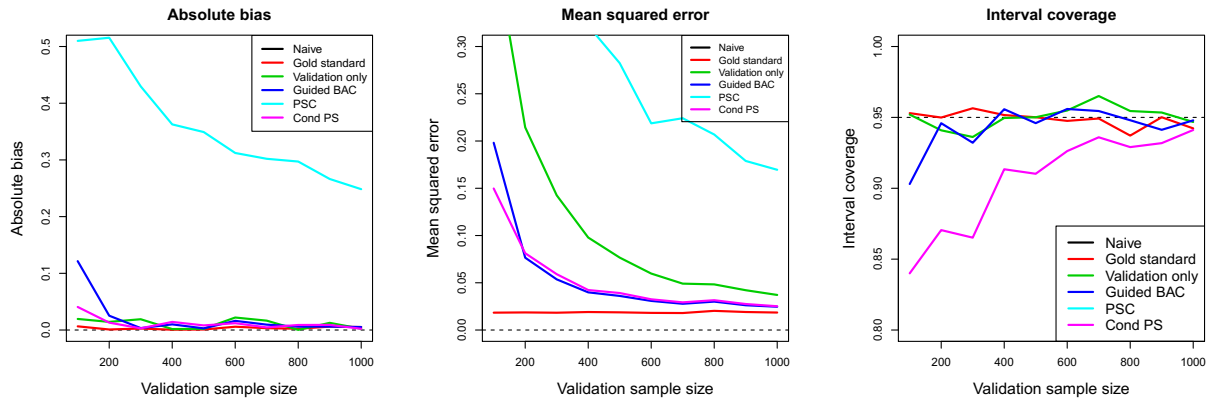

Figure 3.1: Bias, MSE, and 95% interval coverage of estimate of β across 1000 simulations under scenario 1

sample.

$$log\left(\frac{P(I=1)}{1-P(I=1)}\right) \propto 0.25X + 0.5(Y > \bar{Y}) \tag{3.21}$$

We've now introduced missingness into the data in such a manner that subjects with higher exposure and higher outcome values are less likely to be in the validation sample. Note that the probabilities are proportional to and not equal to, as the exact probabilities depend on how large of a validation sample we select. Figure 3.2 shows the results under this simulation setup. One key difference from previous results under MCAR is that using the validation data only will give a biased causal effect, which is expected because ignoring all observations with missing data only works under the MCAR assumption. We hypothesized in section 3.3 that the conditional propensity score approach relied on similar MCAR assumptions and this becomes evident in the simulation as this method suffers from bias even under fairly large validation sample sizes. The guided BAC approach seems to handle the missing mechanism better than the other proposed approaches as it reaches unbiasedness at $N_2 = 400$. While it does tend towards unbiasedness, guided BAC is still biased at small sample sizes, N_2, of the validation data. This suggests that more information is required from the validation data to properly handle the missingness mechanism compared with scenario 1. Results in terms of MSE show very similar results as in scenario 1 as we see that the conditional propensity score has

the lowest MSE of all methods for $N_2 = 100$, while guided BAC has the lowest MSE for all other values of N_2, although this difference is not substantial at most validation sample sizes. We do, however, see about a 20% decrease in MSE for guided BAC compared with the conditional propensity score at $N_2 = 300$. Coverage probabilities are lower for the conditional propensity score than in scenario 1, likely due to the increase in bias. The validation data has coverage probabilities between 0.8 and 0.9, which is less than desired, and this is likely due to the increase in bias seen due to the missingness mechanism. Guided BAC again obtains the desired 0.95 coverage probabilities for all N_2 greater than 200. In the lowest validation sample sizes, guided BAC is somewhat biased leading to small decreases in coverage probabilities as we see a probability of 0.8 for $N_2 = 100$ and 0.9 for $N_2 = 200$.

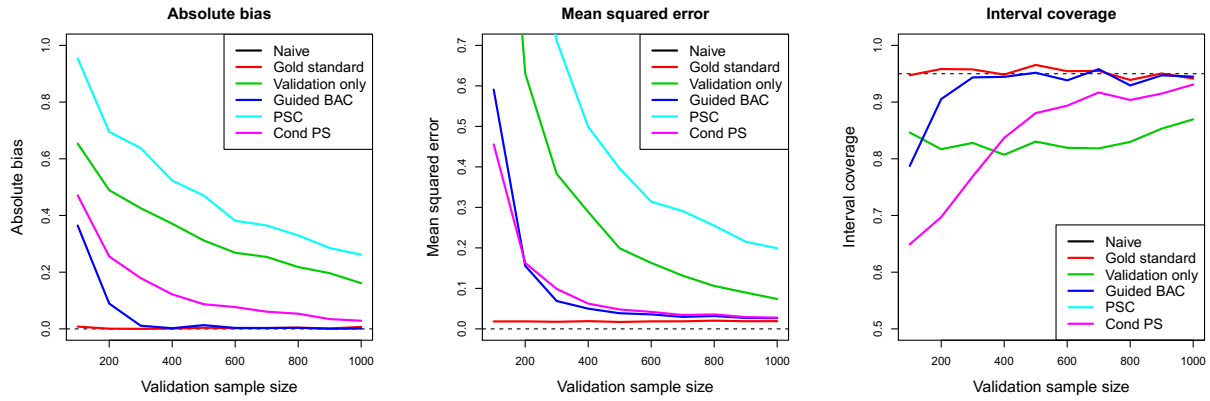

Figure 3.2: Bias, MSE, and 95% interval coverage of estimate of β across 1000 simulations under scenario 2, where the missingness mechanism is MAR

3.4.3 Scenario 3

We now simulate under the scenario where all the missing covariates are binary. We simulate each binary confounder to have a prevalence near 11% and use regression coefficients that induce 20% bias in the effect of interest when all are left out of the model. The exposure in this scenario is also fairly rare with a prevalence near 8% on average. Figure 3.3 shows the performance of the various approaches under this simulation scenario. We see that all of the approaches are biased in estimating β, with the exception of guided BAC and the validation only approaches. The propensity score approaches suffer

from bias even at large validation sample sizes of $N_2 = 1000$. Guided BAC on the other hand handles these types of covariates very well, as the imputation works very well when there exists strong, binary confounders. This is evident when looking at MSE, as we see that guided BAC achieves an MSE almost as small as the gold standard. Propensity score methods on the other hand suffer from substantial amounts of error, as their MSE is larger than the validation only approach. This is likely due to the fact that the prevalence of both the covariates and exposure is quite small and the propensity score models that the other methods rely on are extremely noisy, especially at small validation sample sizes. The 95% interval coverages lead to similar conclusions as the bias from estimating β leads to small coverage probabilities for all of the methods except guided BAC and the validation only approach.

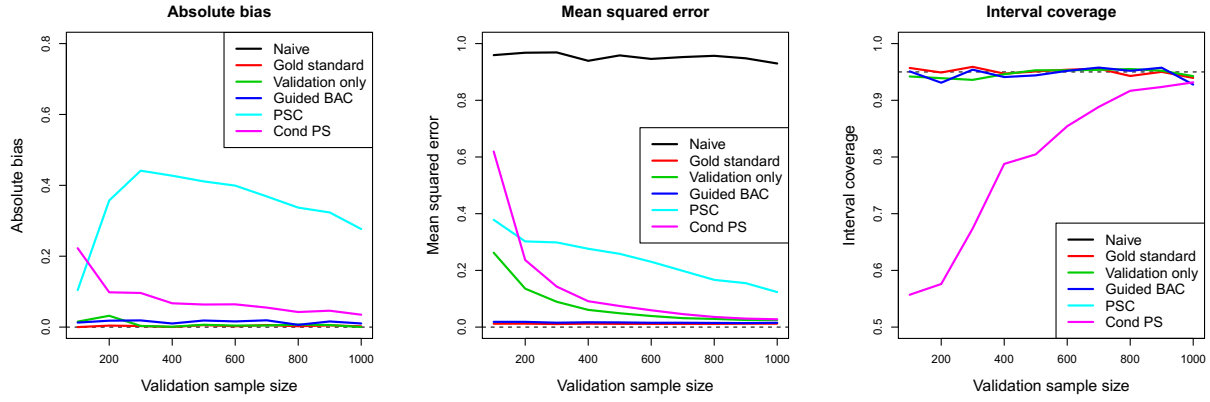

Figure 3.3: Bias, MSE, and 95% interval coverage of estimate of β across 1000 simulations under scenario 3, where covariates are binary and have low prevalences

3.4.4 Scenario 4

We generate data using a similar structure as the available SEER-Medicare. We set M=25 and P=34. Of the 25 fully observed covariates we set 6 of them to be confounders, while 2 of the missing 9 covariates are true confounders. We fit a probit regression model in SEER-Medicare regressing 30 day survival against surgical resection and all available covariates, as well as a probit exposure model regressing surgical resection against all available covarisates. We use these estimated coefficients to generate data under probit regression models. To simulate the potential confounders we calculated the prevalence

of each covariate in SEER-Medicare and simulated independent bernoulli variables with the same prevalences. The main data sample size in SEER-Medicare is 26,559 with a validation sample size of $N_2 = 4,428$. We simulate data under this same sample size, however, we also vary the values of N_1 and N_2 such that the ratio, $\frac{N_2}{N_1+N_2}$ stays the same but the sample size decreases by factors of 2. In this setup N_2 will take the values (277, 544, 1107, 2214, 4428) and $N_1 + N_2$ will take the values (1660, 3320, 6640, 13280, 26559). This gives us a good idea of how the various methods perform under the sample size from SEER-Medicare, but also gives intuition as to whether the overall magnitude of N_2 is what drives our ability to control for confounding, or if it's the ratio of the sample sizes that matters more. We see that at small values of N_2 each method suffers from some bias with the exception of the gold standard. When $N_2 = 1107$, both the guided BAC and validation only approaches attenuate back towards unbiasedness. In terms of MSE, we see that the propensity score approaches suffer from large increases in MSE relative to the gold standard MSE. Guided BAC on the other hand, has a relatively good MSE compared to the gold standard even at low values of N_2. When $N_2 = 277$ the ratio of the guided BAC MSE to the gold standard MSE is only 1.4 while it is 22.7 for the conditional propensity score method. When $N_2 = 4428$, the sample size of the SEER-Medicare validation data set, both the conditional propensity score and guided BAC perform well relative to the gold standard, however, guided BAC is still better than the conditional propensity score in terms of MSE. In terms of interval coverage, as in previous simulations, guided BAC obtains interval coverages near the desired level while the conditional propensity score does poorly except in large validation sample sizes. The results of this simulation suggest that the magnitude of N_2 and not the ratio, $\frac{N_2}{N_1+N_2}$ is more important in determining how well the various methods work.

3.5 Analysis of SEER-Medicare data

We apply the methods proposed in this paper, to estimate the causal effect of resection versus biopsy on 30 day survival for Medicare beneficiaries ages 65 and older, diagnosed with malignant neoplasm of the brain between 1999 and 2007. We use the Medicare

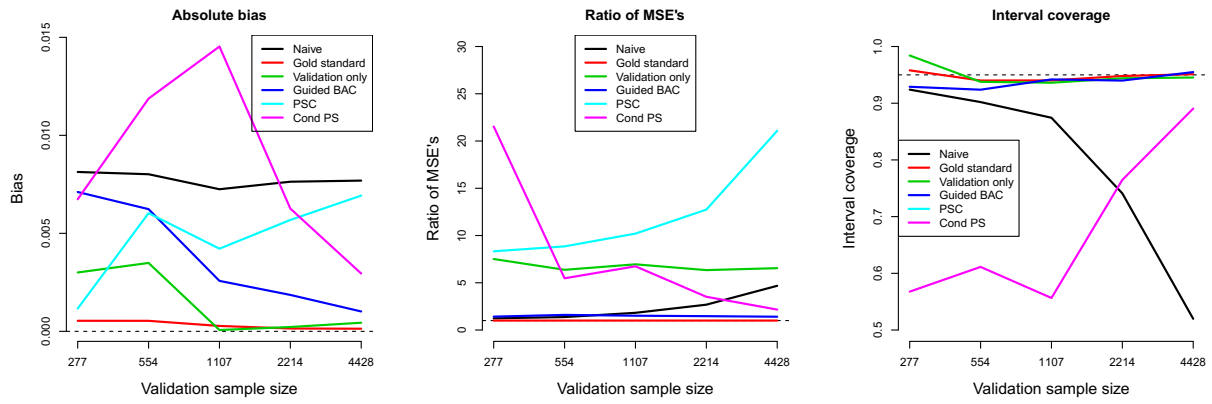

Figure 3.4: Bias, MSE, and 95% interval coverage of estimate of β across 1000 simulations under scenario 4, which emulates the SEER-Medicare data set

dataset as our main study, and the SEER-Medicare dataset as our validation study. We focus on the subset of the Medicare population that had no previous history of cancer that underwent surgical resection or biopsy. The sample size of the SEER-Medicare (e.g. the validation sample) is $N_2 = 4428$ and the sample size of the Medicare sample (e.g. the main study) is $N_1 = 22,131$ leading to an overall sample size of 26559. The SEER-Medicare data set contains subjects from only a subset of the states in the US, while the full Medicare data set is not restricted in this way. It is also known that patients in SEER tend to be healthier than those in the broader Medicare population. This leads us to question the MCAR assumption as it is possible that the SEER-Medicare data set is comprised of people who are different with respect to survival time, exposure, or some observed covariates.

There are 25 covariates that are fully observed in the main sample including age, sex, race, dual eligibility to Medicaid status, as well as a variety of comorbid conditions. The covariates available only in the validation data include marital status, MRI status, glioblastoma, income, CT scan status, and various covariates regarding the severity and location of the tumors. Many of these covariates are likely to be important confounders of the average causal effect of surgical resection on survival. This shows the need for a method that takes into account the missing confounders in the main data set. For the analysis we used a binary indicator of 30 day survival as our outcome and, our exposure is a binary indicator of surgical resection vs. biopsy. We also set $\omega = 20$, which assigns more weight

to covariates to be included in the outcome model if they're found to be associated with exposure, though still allows for them to be excluded if the data strongly suggests these covariates are only associated with the exposure and therefore should not be included into the outcome model. We use non informative priors for all regression coefficients and $IG(0.001, 0.001)$ priors for all variance parameters. Convergence was assessed via visual inspection of trace plots and calculation of the potential scale reduction factor (Gelman et al., 2014). We fit our proposed method as well as the previous approaches that were included in the simulation study: Propensity score calibration, Conditional propensity scores, and the validation only approach.

3.5.1 Main data analysis

Table 3.2 shows descriptive statistics of all the covariates (the fully observed and the missing ones) as well as posterior inclusion probabilities into the exposure and outcome models under our proposed approach. We see that of the variables which are available in the full data set, very few appear to be confounders. Age is included in both models 100% of the time indicating it is a confounder. Of the comorbidities, pneumonia and protein calorie malnutrition are the only ones that are all included in both models with high probability. Of the missing covariates only CT scan status and stage of tumor (localized vs. other) appear to be confounders as both are included in both the exposure and outcome models at least 90% of the time with CT scan status being included in both models 100% of the time. Other covariates such as glioblastoma, MRI status, tumor size, and tumor location all seem like they could be potential confounders as they are included in both models with some regularity, though much less than 100% of the time indicating uncertainty in the data about whether they are associated with either the exposure or outcome. Overall this suggests that a naive analysis based on the full data alone would likely lead to biased estimates of the causal effect of surgical resection on 30 day survival. Due to the large sample size in the validation data one might be inclined to simply analyze this data alone and ignore the remaining Medicare information. We have no reason to believe, however, that the SEER-Medicare registry is a random sample of the full Medicare database and therefore using the validation data only could also lead to biased estimation.

Figure 3.5 shows the resulting estimates of the average causal effect of surgical resection on the probability of 30 day survival for various models. In all cases we see similar substantive results to those seen in Chaichana et al. (2011), though there is substantial variability in the estimates across models. Ignoring the covariates only available in SEER and analyzing the entire $N_1 + N_2$ subjects leads to a naive estimate of 0.054 (0.046, 0.061), which we expect to be biased as table 3.2 indicates that many of these missing covariates are potentially confounders. Analyzing the validation data alone leads to a causal effect estimate of 0.016 (0.002, 0.030). This is far different than the naive estimator, though this difference should not be solely attributed to confounding as the SEER population could be different than the entire Medicare population. The guided BAC approach, which incorporates both sources of information leads to an estimate of 0.042 (0.031, 0.053). The conditional propensity score gives a similar estimate of 0.045 (0.034, 0.056), while the propensity score calibration gives a much smaller estimate of 0.018 (-0.010, 0.046). Obviously there is no way of knowing which, if any, of these models are close to accurately estimating the effect of surgical resection on survival. Our scientific knowledge combined with table 3.2 leads us to disregard the naive approach. Our guided BAC approach should give correct estimates conditional on specifying the exposure, outcome, and imputation models correctly. As mentioned earlier, the binary nature of all of the missing covariates leads us to worry less about correct specification of the exposure and outcome models, though interaction terms could still lead to misspecification. We also reduced the possibility of model misspecification by performing variable selection to remove unnecessary imputed variables from entering into the models. It is possible that we misspecified the imputation models, but again these models are regressing the missing covariates on the other covariates, which are binary reducing the possibility of model misspecification. The validity of the validation only, conditional PS, and PSC approaches is contingent on stronger assumptions than those made for the guided BAC approach and a breach of these assumptions could explain any differences between the estimates.

		Biopsy	Resection	$P(\alpha_{xj} = 1\|D)$	$P(\alpha_{yj} = 1\|D)$
Main data	$75 <$ Age < 85	3853 (0.4)	5647 (0.33)	1.00	1.00
+ validation	Age > 85	842 (0.09)	882 (0.05)	1.00	1.00
$(\mathbf{N_1 + N_2 = 26,559})$	Female	4689 (0.49)	7817 (0.46)	0.01	0.65
	White	8878 (0.92)	15643 (0.92)	0.00	0.02
	Head CT scan done	793 (0.08)	1011 (0.06)	0.02	0.02
	Brain MRI done	906 (0.09)	1812 (0.11)	0.01	1.00
	Dual eligible	734 (0.08)	1231 (0.07)	0.00	0.01
	Chronic Atherosclerosis	2089 (0.22)	3180 (0.19)	0.01	0.17
	Substance abuse	695 (0.07)	1329 (0.08)	0.00	0.00
	Hypertension	5936 (0.62)	10151 (0.6)	0.03	0.01
	Cerebrovascular disease	386 (0.04)	556 (0.03)	0.00	0.02
	COPD	1029 (0.11)	1938 (0.11)	0.26	1.00
	Pneumonia	280 (0.03)	583 (0.03)	0.99	1.00
	Protein calorie malnutrition	135 (0.01)	294 (0.02)	0.57	1.00
	Dementia	1052 (0.11)	1504 (0.09)	0.01	1.00
	Functional disability	346 (0.04)	460 (0.03)	0.00	0.01
	Trauma in past year	398 (0.04)	618 (0.04)	0.00	0.03
	Parkinson's/Huntington's	110 (0.01)	159 (0.01)	0.03	1.00
	Chronic fibrosis	115 (0.01)	198 (0.01)	0.00	0.01
	Depression	736 (0.08)	1098 (0.06)	0.01	0.02
	Seizure disorder	1999 (0.21)	3487 (0.21)	0.00	0.02
	Asthma	276 (0.03)	484 (0.03)	0.00	0.01
	Hypertensive heart disease	122 (0.01)	211 (0.01)	0.00	0.00
	Valvular and rheumatic heart disease	583 (0.06)	866 (0.05)	0.00	0.01
	Diabetes	1827 (0.19)	3067 (0.18)	0.01	0.11
Validation only	Married	830 (0.6)	1967 (0.65)	0.00	0.02
$(\mathbf{N_2 = 4,428})$	MRI_done	896 (0.65)	2256 (0.74)	1.00	0.28
	GBM	946 (0.69)	2633 (0.86)	1.00	0.31
	One primary tumor	1220 (0.88)	2713 (0.89)	0.00	0.03
	Supratentorial tumor	931 (0.67)	2453 (0.81)	1.00	0.34
	Tumor $>$ 3cm	703 (0.51)	1879 (0.62)	0.98	0.45
	CT_scan_done	1143 (0.83)	2233 (0.73)	1.00	1.00
	Income	507 (0.37)	1009 (0.33)	0.00	0.02
	Localized tumor	926 (0.67)	2549 (0.84)	1.00	0.93

Table 3.2: Patient characteristics and posterior inclusion probabilities for covariates into the exposure and outcome models. Binary variables are reported as number of patients (percentage), and continuous covariates are reported as mean (standard deviation)

3.5.2 Examining effectiveness of guided BAC

To further validate our approach compared to alternatives in the context of the real data analysis we use cross-validation. Within the validation data we can artificially create missingness in a subset of the variables for a subset of the data. Given this new data source in which we know the true missing values we can compare all the existing approaches to estimating the average causal effect to see which one comes nearest to the gold standard approach estimate, which includes all of the covariates. To do this we allow 1500 randomly chosen subjects to maintain their full covariate information while the remaining subjects from the validation data have information on certain covariates removed. To maintain similarity to the main data analysis we induced missingness in the

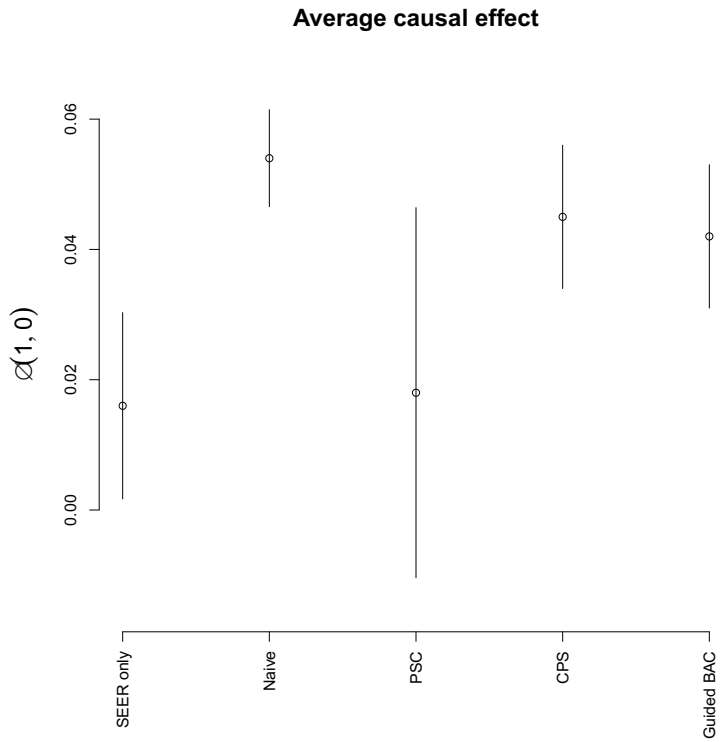

Figure 3.5: Estimates and confidence intervals for the causal effect of surgical resection on the probability of 30 day survival

variables which are truly missing in the main Medicare data as detailed in table 3.2. We repeated this process K times and kept track of the effect estimates for each of the proposed approaches at each iteration. We look at the average absolute relative difference of the estimator and the gold standard estimate to assess how well the various approaches perform when data is missing. Define $\hat{\beta}_k$ to represent the estimate of β from a given method and $\hat{\beta}_g$ to represent the gold standard estimate of β that exists if we have no missing information. Then the quantity we look at is

1. $L_1 = \frac{1}{K} \sum_{k=1}^{K} 100 * \left| \frac{\hat{\beta}_k - \hat{\beta}_g}{\hat{\beta}_g} \right|$

Table 3.3 shows the results from performing cross validation on the validation sample. We see that on average, the guided BAC approach is closer to the gold standard approach with respect to absolute relative difference. This suggests that the guided BAC approach tends to perform the best under the data generating mechanism that dictates the SEER data set. While there is no guarantee that this data generating mechanism and therefore

these results would extend to the larger Medicare population, it provides us with further intuition that our approach is useful in analyzing this data.

Method	L_1
Naive	40.98
PSC	67.20
Guided BAC	19.42
Conditional PS	34.32

Table 3.3: Results from cross validation analyses of validation data

In analyses with a binary exposure, it is of interest to check for covariate balance, and this is typically done via propensity scores. While our approach does not utilize propensity scores to control for confounding, we can build a propensity score model at every iteration of the MCMC using the most recent updates for the missing covariates. Once we have this propensity score model we can create a matched data set via propensity score matching and examine the extent to which this new propensity score balanced the missing covariates by looking at standardized differences of the missing variables between the two matched sets. Since we do this at every iteration of the MCMC we can take the average across posterior draws of the standardized differences to obtain an idea of how well the imputed covariates are able to balance the true missing covariates. To check this balance we need the true missing values, so just as before with cross validation we must do this in the SEER data only and induce missingness in the covariates artificially. One key difference here is we are not repeating this process K times, we are simply doing this for one data set to get an idea of how well the covariates are balanced. Figure 3.6 shows the mean across posterior draws of the balance of the missing covariates as well as the balance when the propensity score is built using only the fully observed covariates. We see that for each variable with a large imbalance in the covariate distribution, the imputed covariate propensity score significantly improves the standardized difference between treated and controls. This suggests that the naive analysis that only looks at the fully observed covariates is less likely to give a valid estimate of the causal effect than the guided BAC analysis, which imputes these covariates. While the imputed covariate propensity score is never quite able to reach the gold standard propensity score balance, it does do well for

the variables that our data are suggesting could be confounders. Looking at the posterior inclusion probabilities on the Y-axis we see that the variables which are closer to the gold standard than the naive balance such as CT scan status and whether a tumor is supratentorial are those that have high inclusion probabilities into both the X and Y models. This indicates that the imputed covariates are good at achieving balance for those covariates for which balance is crucial to estimating a valid causal effect.

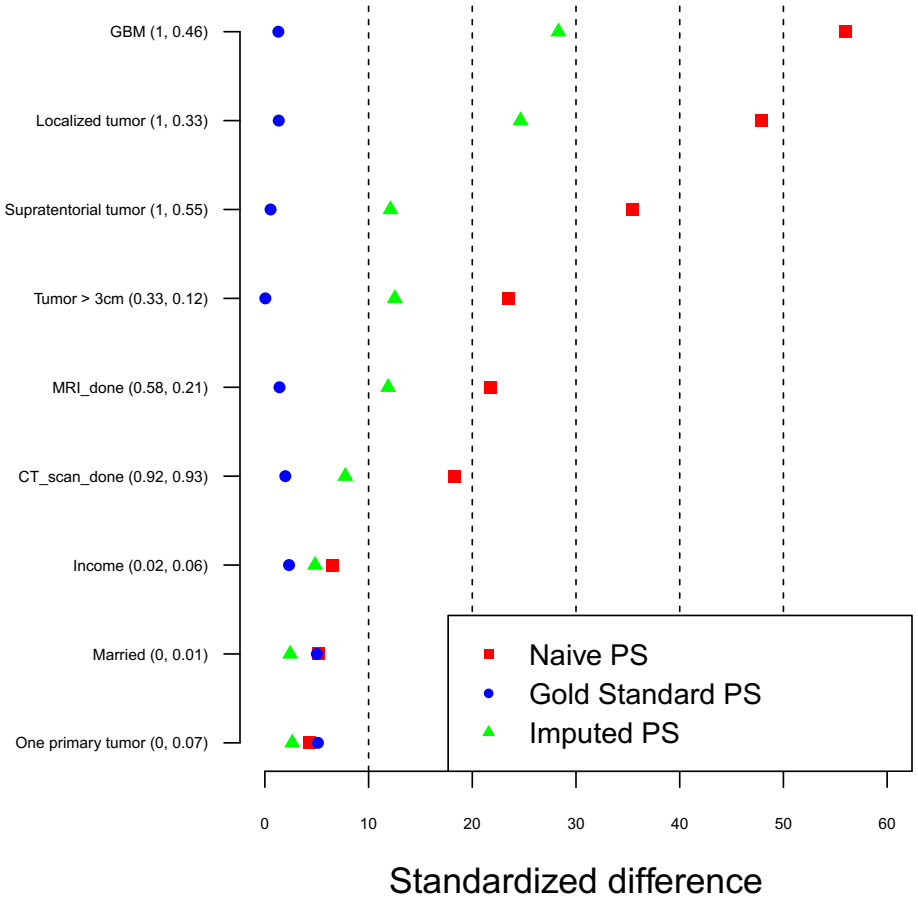

Figure 3.6: Balance of missing covariates using a naive propensity score built only on the fully observed covariates, from a propensity score that utilizes imputed values of missing covariates, and from a propensity score built on all of the covariates' true values. On the X-axis is standardized difference to measure the amount of imbalance between treated and controls. The Y-axis shows the covariate names as well as the posterior inclusion probabilities for the exposure and outcome models respectively in parentheses.

3.6 Discussion

In this article we have combined ideas of Bayesian model averaging and missing data imputation to estimate the effect of an exposure on an outcome when there are missing confounders in the data set, but auxiliary information on the confounders is available through a validation data set. Our proposed method has advantages in that it requires less restrictive assumptions than previous approaches in the literature, can handle a wide variety of data types, can accomodate the situation where the validation data is not a random subset of the main study, and can identify necessary confounders for unbiased effect estimation all while accounting for the uncertainty in confounder selection and variable imputation. Through simulation we showed that under a wide variety of scenarios our proposed approach works well and performs comparatively well to or better than existing approaches to utilizing validation data to control for confounding. Finally, we illustrated our approach in an analysis of surgical resection on 30 day survival in the SEER-Medicare data and found that there was likely unmeasured confounding bias from missing covariates in Medicare. Our approach attenuated the naive estimate of the effect of surgical resection on 30 day survival by 20% indicating that the original estimate may in fact have been suffering from confounding bias. We examined the validation data more closely to examine how well our approach can balance covariates and eliminate bias in SEER, and found that our method performed best in the SEER data relative to other approaches and improved balance of the missing covariates.

While the results of our simulation study and data analysis suggest that our method is very useful in controlling for confounding in the validation data setting, there are some limitations. Our proposed procedure, as well as any variable imputation procedure, relies on correctly specifying the imputation model for a given covariate. Information used to impute a covariate is not limited to simply the imputation model, as both the exposure and outcome models impact the imputations at every stage of our gibbs sampler as well. This means that we must correctly specify the manner in which the covariate we are imputing enters into both the exposure and outcome models as well. For instance, if our imputation model was correct, but we included the covariate into the outcome model

linearly when in truth it should be included as a quadratic term then we could induce bias. Our method, however, is more robust than standard multiple imputation as it only includes variables deemed necessary for controlling for confounding. Nonetheless the extent to which various types of model misspecification affect our ability to control for confounding is unclear and a potential topic of further research. If one does not want to make these assumptions about the parametric form of the models then propensity score approaches might be favorable. Our method also relies on a transportabillty assumption stating that the conditional distributions relating our observed variables must be the same between the main and validation data. While this is not necessarily a drawback of the method, because any method relying on validation data will be forced to make this assumption, it could lead to invalid inference and careful thought should be done before performing an analysis to see if this holds true. We also stress that in any analysis examining a causal effect, scientific knowledge should be used to exclude any potential instrumental variables or collider variables before running any analysis. A data driven method for selecting confounders such as ours will include these variables into the model, which can introduce bias, so careful thought and expertise must be stressed before running our approach.

One advantage of the proposed approach is the ability to not include into the outcome model covariates that are not confounders. In scenarios where there are lots of potential confounders, but only a small subset of them are important for valid effect estimation, variable selection can drastically improve efficiency by eliminating unnecessary variables. In the case of missing data this could be even more important because we will be excluding imputed variables, which could be even noisier than their fully observed counterparts due to the uncertainty in variable imputation. Other variable selection approaches exist in the literature, however, the majority of these approaches are good at including variables to predict the outcome, not control for confounding, which is an important distinction between standard BMA and the prior used in our paper and Wang et al. (2012). Confounder selection is also very important to protect against model misspecification of the imputation models. As discussed previously, model misspecification for our imputation model can lead to bias of the exposure effect we're interested in, but we will be removing vari-

ables from the model and reducing the number of imputations which could potentially be incorrect. To even further reduce model misspecification we could include variable selection into the imputation models as well in a similar manner as Mitra and Dunson (2010). This would reduce the possibility of misspecifying the imputation models and would keep noise out of these models that would increase the mean squared error of our final effect estimates.

One future direction is extending these ideas to data that is neither continuous nor binary. Our method in theory can handle any exposure or outcome variable that falls in the GLM framework, though avoiding Metropolis Hastings updates that require tuning would require more complicated computational tricks using latent variable representations to handle count or categorical data. We do not believe this to be a major hurdle, as these MCMC techniques exist in the literature and simply must be implemented in our setting. These approaches could also be used to impute categorical missing covariates.

In summary, we have proposed a procedure to control for confounding in the presence of missing confounders when validation data is available. The proposed procedure utilizes a fully probabilisitic, Bayesian approach, which accounts for the uncertainty in the selection of confounders and in missing data imputations simultaneously. The procedure extends previous work on the control of confounding in the presence of missing data and validation data by allowing for both continuous and and binary exposures and by alleviating some of the restrictive assumptions necessary for a valid effect estimate.

3.7 Appendix

3.7.1 Details of posterior simulation

Here we show the full posterior and corresponding conditionals for implementing a gibbs sampler. For simplifying notation we will ignore U and let the matrix C represent all covariates in the data, while acknowledging that some of these covariates are missing in a subset of the subjects. First let $\theta_y = [\theta_{y0}, \beta, \theta_{y1}, ..., \theta_{yp}]$. Then Letting X_i^*, Y_i^*, and C_{ij}^* be latent variables for X_i, Y_i, and C_{ij} respectively the posterior can be written as follows:

$$
\begin{aligned}
&P(\theta_y, \theta_x, \theta_p, \sigma_y^2, \sigma_x^2, \sigma_p^2, \alpha^x, \alpha^y, C, C^*, Y^*, X^* | X, Y) \\
&\propto \prod_{i=1}^{N} p(Y_i|Y_i^*) p(Y_i^*|\theta_y, X_i, C_i, \sigma_y^2, \alpha^y) \\
&\times p(X_i|X_i^*) p(X_i^*|\theta_x, C_i, \sigma_x^2, \alpha^x) \\
&\times \prod_{j=M+1}^{P} p(C_{ij}|C_{ij}^*) p(C_{ij}^*|\theta_j, \sigma_j^2, C_i) \\
&\times P(\theta^y) P(\theta^x) P(\theta^p) P(\sigma_y^2) P(\sigma_x^2) P(\sigma_p^2) P(\alpha^y, \alpha^x)
\end{aligned}
$$

$$
\begin{aligned}
&\propto \prod_{i=1}^{N} p(Y_i|Y_i^*) N(Y_i^*; \theta_{y0} + \beta X_i + \sum_{k=1}^{P} \alpha_{yk}\theta_{yk} C_{ik}, \sigma_y^2) \\
&\times \prod_{i=1}^{N} p(X_i|X_i^*) N(X_i^*; \theta_{x0} + \sum_{k=1}^{P} \alpha_{xk}\theta_{xk} C_{ik}, \sigma_y^2) \\
&\times \prod_{j=M+1}^{P} p(C_{ij}|C_{ij}^*) N(C_{ij}^*; \theta_{j0} + \sum_{k=1}^{j-1} \theta_{jk} C_{ik}, \sigma_y^2) \\
&\times P(\theta^y) P(\theta^x) P(\theta^p) P(\sigma_y^2) P(\sigma_x^2) P(\sigma_p^2) P(\alpha^y, \alpha^x)
\end{aligned}
$$

For each regression coefficient in the model we assign independent, non informative $N(0, K)$ priors, where K is set to be very large relative to the magnitude of the coefficients. For each variance parameter in the model we assign an $IG(a, b)$ prior. The prior distribution $P(\alpha^y, \alpha^x)$ is implemented as described in the text. Under these priors the full conditionals take the following form:

$$
\begin{aligned}
P(\theta_x|\bullet) &\sim N\left(\left(W_x^{\alpha^x T} W_x^{\alpha^x} + \frac{\sigma_x^2 I}{k}\right)^{-1} W_x^{\alpha^x T} X^*, \left(W_x^{\alpha^x T} W_x^{\alpha^x}/\sigma_x^2 + I/k\right)^{-1}\right) \\
P(\sigma_x^2|\bullet) &\sim IG\left(N/2 + a, b + \frac{(X - W_x^{\alpha^x}\theta^x)^T (X - W_x^{\alpha^x}\theta^x)}{2}\right) \\
P(\theta_y|\bullet) &\sim N\left(\left(W_y^{\alpha^y T} W_y^{\alpha^y} + \frac{\sigma_y^2 I}{k}\right)^{-1} W_y^{\alpha^y T} Y^*, \left(W_y^{\alpha^y T} W_y^{\alpha^y}/\sigma_y^2 + I/k\right)^{-1}\right) \\
P(\sigma_y^2|\bullet) &\sim IG\left(N/2 + a, b + \frac{(Y - W_y^{\alpha^y}\theta^y)^T (Y - W_y^{\alpha^y}\theta^y)}{2}\right)
\end{aligned}
$$

where $W_y^{\alpha^y}$ represents the design matrix for the outcome model defined by α^y, and $W_x^{\alpha^x}$ represents the design matrix for the exposure model defined by α^x. This means that the dimension of θ_x and θ_y are changing as we run through our gibbs sampler. In practice to implement this algorithm we set $\theta_{yj} = 0$ when $\alpha_j^y = 0$ for all j, and then update the remaining values of θ_y in the manner described above. We also note that if X or Y are binary then $\sigma_x^2 = 1$ or $\sigma_y^2 = 1$ by definition and no updating of those parameters is necessary. The full conditionals for the parameters of the imputation model for covariate j where j=M+1...P are as follows:

$$P(\theta_j|\bullet) \sim N\left(\left(W_j^T W_j + \frac{\sigma_j^2 I}{k}\right)^{-1} W_j^T C_j^*, \left(W_j^T W_j/\sigma_j^2 + I/k\right)^{-1}\right)$$
$$P(\sigma_j^2|\bullet) \sim IG\left(N/2 + a, b + \frac{(C_j - W_j\theta^j)^T(C_j - W_j\theta^j)}{2}\right)$$

Where if covariate j is binary then by definition $\sigma_j^2 = 1$ and no updating of the variance parameter is needed. Now we need to update from the full conditional distribution of the missing covariates. If covariate j is missing and continuous then we will impute from a Normal distribution:

$$N(V_{ij}\mu_{ij}, V_{ij})$$

Where

$$\mu_{ij} = \alpha_{yj}\frac{Y_{i(-j)}\theta_{yj}}{\sigma_y^2} + \alpha_{xj}\frac{X_{i(-j)}\theta_{xj}}{\sigma_x^2} + \frac{\tilde{\mu_{ij}}}{\sigma_j^2} + \sum_{k=j+1}^{P}\frac{\theta_{kj}C_{ik(-j)}^*}{\sigma_k^2}$$
$$V_{ij} = \alpha_{yj}\frac{\theta_{yj}^2}{\sigma_y^2} + \alpha_{xj}\frac{\theta_{xj}^2}{\sigma_x^2} + \frac{1}{\sigma_j^2} + \sum_{k=j+1}^{P}\frac{\theta_{kj}^2}{\sigma_k^2}$$

And

$$Y_{i(-j)} = Y_i^* - \theta_{y0} - \beta X_i - \sum_{l\neq j} C_{il}\theta_{yl}$$

$$X_{i(-j)} = X_i^* - \theta_{x0} - \sum_{l \neq j} C_{il}\theta_{xl}$$

$$\tilde{\mu_{ij}} = \theta_{j0} + \sum_{k=1}^{j-1} \theta_{jk} C_{ik}^*$$

$$C_{ik(-j)}^* = C_{ik}^* - \theta_{k0} - \sum_{l \neq j, l=1}^{k-1}$$

When C_{ij} is binary we will impute it's corresponding latent variable, C_{ij}^*. If C_{ij} is observed then we still update it's full conditional from

$$C_{ij}^* \sim C_{ij} TN_+(V_{ij}\mu_{ij}, V_{ij}) + (1 - C_{ij}) TN_-(V_{ij}\mu_{ij}, V_{ij})$$

Where TN_+ represents a truncated normal distribution that only assigns positive probability to the positive real line, and TN_- the same but only assigning mass to the negative real line.

$$\mu_{ij} = \frac{\tilde{\mu_{ij}}}{\sigma_j^2} + \sum_{k=j+1}^{P} \frac{\theta_{kj} C_{ik(-j)}^*}{\sigma_k^2}$$

$$V_{ij} = \frac{1}{\sigma_j^2} + \sum_{k=j+1}^{P} \frac{\theta_{kj}^2}{\sigma_k^2}$$

Where again

$$\tilde{\mu_{ij}} = \theta_{j0} + \sum_{k=1}^{j-1} \theta_{jk} C_{ik}^*$$

$$C_{ik(-j)}^* = C_{ik}^* - \theta_{k0} - \sum_{l \neq j, l=1}^{k-1}$$

For binary variables that are missing, we again get a mixture of truncated normals, though we replace the binary indicator with the posterior probability that variable is 1 or 0 as follows:

$$C_{ij}^* \sim \pi_{ij} TN_+(V_{ij}\mu_{ij}, V_{ij}) + (1 - \pi_{ij}) TN_-(V_{ij}\mu_{ij}, V_{ij})$$

With the probability defined as

$$\pi_{ij} = \frac{\Phi(\mu_{ij})\phi(Y_{i(-j)} - \theta_{yj})\phi(X_{i(-j)} - \theta_{xj})}{\Phi(\mu_{ij})\phi(Y_{i(-j)} - \theta_{yj})\phi(X_{i(-j)} - \theta_{xj}) + (1 - \Phi(\mu_{ij}))\phi(X_{i(-j)})\phi(Y_{i(-j)})}$$

The only parameters left to sample from are the vector of variable inclusion indicators (α^x, α^y) and to do this we can follow the ideas of Wang et al. (2015) utilizing the MC^3 technique for searching a model space (Madigan et al., 1994, 1995). We will illustrate how to sample from $P(\alpha^y|\alpha^x, D)$, however, the algorithm for sampling from $P(\alpha^x|\alpha^y, D)$ is analagous. We can define a neighborhood of α^y to be the set of all outcome models with one covariate either added or removed from the model defined by α^y. If we are at iteration t of our current Markov chain, and we are currently at the values $(\alpha^y_{(0)}, \alpha^x_{(0)})$, then we randomly draw a model $\alpha^y_{(1)}$ from the neighborhood of $\alpha^y_{(0)}$ and we accept the new model with probability

$$\min\left\{1, \frac{P(\alpha^y_{(1)}|\alpha^x_{(0)}, D)}{P(\alpha^y_{(0)}|\alpha^x_{(0)}, D)} = \frac{P(Y|\alpha^y_{(1)}, X, C)}{P(Y|\alpha^y_{(0)}, X, C)} * \frac{P(\alpha^y_{(1)}|\alpha^x_{(0)})}{P(\alpha^y_{(0)}|\alpha^x_{(0)})}\right\}$$

Otherwise the chain stays at $\alpha^y_{(0)}$. We are easily able to calculate $\frac{P(\alpha^y_{(1)}|\alpha^x_{(0)})}{P(\alpha^y_{(0)}|\alpha^x_{(0)})}$ using our conditional prior specification from section 3.2.2. To calculate the ratio of marginal likelihoods we can use the BIC approximation to the Bayes factor (Raftery, 1995) defined as

$$\frac{P(Y|\alpha^y_{(1)}, X, C)}{P(Y|\alpha^y_{(0)}, X, C)} \approx exp\left\{\frac{1}{2}(BIC_0 - BIC_1)\right\}$$

3.7.2 Proof that prior increases posterior probability of including minimal confounder set

Here we outline why our prior will be useful in controlling for confounding. Let all distributions with a subscript 1 be distributions associated with using our conditional prior specification, and all distributions with a subscript 2 be associated with a flat prior on the

model space similar to BMA. Distributions without a subscript are those that are unaffected by the choice of prior. Then without loss of generality if we assume that the first m covariates are the ones necessary for controlling for confounding then the following holds:

$$
\begin{aligned}
P_1(\alpha^y \in \boldsymbol{\alpha^{y*}}|D) &= P_1(\alpha_1^y = 1, \alpha_2^y = 1, ..., \alpha_m^y = 1|D) \\
&= P_1(\alpha_1^y = 1|D)\ P_1(\alpha_2^y = 1|\alpha_1^y = 1, D)...P_1(\alpha_m^y = 1|\alpha_1^y = 1, ..., \alpha_{m-1}^y = 1, D) \\
&= P_1(\alpha_1^y = 1|D)\ *\ C_1 \\
&= \frac{P(D|\alpha_1^y = 1)P_1(\alpha_1^y = 1)}{P(D)}\ *\ C_1 \\
&= \sum_{\alpha_1^x} \frac{P(D|\alpha_1^y = 1)P_1(\alpha_1^y = 1, \alpha_1^x)}{P(D)}\ *\ C_1 \\
&= \frac{P(D|\alpha_1^y = 1)\frac{2\omega}{3\omega+1}}{P(D)}\ *\ C_1 \\
&\geq \frac{P(D|\alpha_1^y = 1)\frac{1}{2}}{P(D)}\ *\ C_1 \\
&= \sum_{\alpha_1^x} \frac{P(D|\alpha_1^y = 1)P_2(\alpha_1^y = 1, \alpha_1^x)}{P(D)}\ *\ C_1 \\
&= P_2(\alpha_1^y = 1|D)\ *\ C_1 \\
&= P_2(\alpha_1^y = 1|D)P_1(\alpha_2^y = 1|\alpha_1^y = 1, D)...P_1(\alpha_m^y = 1|\alpha_1^y = 1, ..., \alpha_{m-1}^y = 1, D) \\
&= P_1(\alpha_2^y = 1|\alpha_1^y = 1, D)\ *\ C_2 \\
&= \frac{P(D|\alpha_2^y = 1, \alpha_1^y = 1)P_1(\alpha_2^y = 1|\alpha_1^y = 1)}{P(D|\alpha_1^y = 1)}\ *\ C_2 \\
&= \frac{P(D|\alpha_2^y = 1, \alpha_1^y = 1)P_1(\alpha_2^y = 1)}{P(D|\alpha_1^y = 1)}\ *\ C_2 \\
&= \sum_{\alpha_2^x} \frac{P(D|\alpha_2^y = 1, \alpha_1^y = 1)P_1(\alpha_2^y = 1, \alpha_2^x)}{P(D|\alpha_1^y = 1)}\ *\ C_2 \\
&= \frac{P(D|\alpha_2^y = 1, \alpha_1^y = 1)\frac{2\omega}{3\omega+1}}{P(D|\alpha_1^y = 1)}\ *\ C_2 \\
&\geq \frac{P(D|\alpha_2^y = 1, \alpha_1^y = 1)\frac{1}{2}}{P(D|\alpha_1^y = 1)}\ *\ C_2 \\
&= \sum_{\alpha_2^x} \frac{P(D|\alpha_2^y = 1, \alpha_1^y = 1)P_2(\alpha_2^y = 1, \alpha_2^x)}{P(D|\alpha_1^y = 1)}\ *\ C_2
\end{aligned}
$$

$$= P_2(\alpha_2^y = 1|\alpha_1^y = 1, D) \ * \ C_2$$
$$= P_2(\alpha_1^y = 1|D)P_2(\alpha_2^y = 1|\alpha_1^y = 1, D)$$
$$\times P_1(\alpha_3^y = 1|\alpha_1^y = 1, \alpha_2^y = 1, D)...P_1(\alpha_m^y = 1|\alpha_1^y = 1, ..., \alpha_{m-1}^y = 1, D)$$

And the proof concludes by performing analogous algebraic operations for each of the remaining conditional posteriors from j=3,...,m. We have now shown how our conditional prior specification assigns more posterior mass to models that contain the true confounder set.

References

BREYSSE, P. N., DELFINO, R. J., DOMINICI, F., ELDER, A. C., FRAMPTON, M. W., FROINES, J. R., GEYH, A. S., GODLESKI, J. J., GOLD, D. R., HOPKE, P. K. ET AL. (2013). Us epa particulate matter research centers: summary of research results for 2005–2011. *Air Quality, Atmosphere & Health* **6** 333–355.

BROCHU, P. J., KIOUMOURTZOGLOU, M.-A., COULL, B. A., HOPKE, P. K. and SUH, H. H. (2011). Development of a new method to estimate the regional and local contributions to black carbon. *Atmospheric Environment* **45** 7681–7687.

BUHMANN, M. (1995). Multiquadric prewavelets on nonequally spaced knots in one dimension. *Mathematics of computation* **64** 1611–1625.

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. CRC press.

CHAICHANA, K. L., GARZON-MUVDI, T., PARKER, S., WEINGART, J. D., OLIVI, A., BENNETT, R., BREM, H. and QUINONES-HINOJOSA, A. (2011). Supratentorial glioblastoma multiforme: the role of surgical resection versus biopsy among older patients. *Annals of surgical oncology* **18** 239–245.

CHOW, J. C., ENGELBRECHT, J. P., WATSON, J. G., WILSON, W. E., FRANK, N. H. and ZHU, T. (2002). Designing monitoring networks to represent outdoor human exposure. *Chemosphere* **49** 961–978.

CHUI, C. K., WARD, J., JETTER, K. and STOECKLER, J. (1996). Wavelets for analyzing scattered data: An unbounded operator approach. *Applied and Computational Harmonic Analysis* **3** 254–267.

DADVAND, P., PARKER, J., BELL, M. L., BONZINI, M., BRAUER, M., DARROW, L. A., GEHRING, U., GLINIANAIA, S. V., GOUVEIA, N., HA, E.-H. ET AL. (2013). Maternal exposure to particulate air pollution and term birth weight: a multi-country evaluation of effect and heterogeneity .

DIGGLE, P. J., MENEZES, R. and SU, T.-L. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **59** 191–232.

DOCKERY, D. W., POPE, C. A., XU, X., SPENGLER, J. D., WARE, J. H., FAY, M. E., FERRIS JR, B. G. and SPEIZER, F. E. (1993). An association between air pollution and mortality in six us cities. *New England journal of medicine* **329** 1753–1759.

DOMINICI, F., PENG, R. D., BELL, M. L., PHAM, L., MCDERMOTT, A., ZEGER, S. L. and SAMET, J. M. (2006). Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *Jama* **295** 1127–1134.

DOMINICI, F., SHEPPARD, L. and CLYDE, M. (2003). Health effects of air pollution: a statistical review. *International Statistical Review* **71** 243–276.

GELFAND, A. E., SAHU, S. K. and HOLLAND, D. M. (2012). On the effect of preferential sampling in spatial prediction. *Environmetrics* **23** 565–578.

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2014). *Bayesian data analysis*, vol. 2. Taylor & Francis.

GLINIANAIA, S. V., RANKIN, J., BELL, R., PLESS-MULLOLI, T. and HOWEL, D. (2004). Particulate air pollution and fetal health: a systematic review of the epidemiologic evidence. *Epidemiology* **15** 36–45.

GRYPARIS, A., PACIOREK, C. J., ZEKA, A., SCHWARTZ, J. and COULL, B. A. (2009). Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics* **10** 258–274.

GUPTA, C., LAKSHMINARAYAN, C., WANG, S. and MEHTA, A. (2010). Non-dyadic haar wavelets for streaming and sensor data. In *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*. IEEE.

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. and VOLINSKY, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science* 382–401.

HULATA, E., SEGEV, R. and BEN-JACOB, E. (2002). A method for spike sorting and detection based on wavelet packets and shannon's mutual information. *Journal of neuroscience methods* **117** 1–12.

KANAROGLOU, P. S., JERRETT, M., MORRISON, J., BECKERMAN, B., ARAIN, M. A., GILBERT, N. L. and BROOK, J. R. (2005). Establishing an air pollution monitoring network for intra-urban population exposure assessment: A location-allocation approach. *Atmospheric Environment* **39** 2399–2409.

KIM, S.-Y., SHEPPARD, L. and KIM, H. (2009). Health effects of long-term air pollution: influence of exposure prediction methods. *Epidemiology* **20** 442–450.

KLOOG, I., CHUDNOVSKY, A. A., JUST, A. C., NORDIO, F., KOUTRAKIS, P., COULL, B. A., LYAPUSTIN, A., WANG, Y. and SCHWARTZ, J. (2014). A new hybrid spatio-temporal model for estimating daily multi-year pm 2.5 concentrations across northeastern usa using high resolution aerosol optical depth data. *Atmospheric Environment* **95** 581–590.

KLOOG, I., MELLY, S. J., RIDGWAY, W. L., COULL, B. A., SCHWARTZ, J. ET AL. (2012). Using new satellite based exposure methods to study the association between pregnancy pm2. 5 exposure, premature birth and birth weight in massachusetts. *Environmental Health* **11** 1–8.

LEE, A., SZPIRO, A., KIM, S. and SHEPPARD, L. (2015). Impact of preferential sampling on exposure prediction and health effect inference in the context of air pollution epidemiology. *Environmetrics* .

LEFEBVRE, G., DELANEY, J. A. and MCCLELLAND, R. L. (2014). Extending the bayesian adjustment for confounding algorithm to binary treatment covariates to estimate the effect of smoking on carotid intima-media thickness: the multi-ethnic study of atherosclerosis. *Statistics in medicine* **33** 2797–2813.

LITTLE, R. J. and RUBIN, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

MADIGAN, D., RAFTERY, A. E., YORK, J. C., BRADSHAW, J. M. and ALMOND, R. G. (1994). Strategies for graphical model selection. In *Selecting Models from Data*. Springer, 91–100.

MADIGAN, D., YORK, J. and ALLARD, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique* 215–232.

MADSEN, L., RUPPERT, D. and ALTMAN, N. (2008). Regression with spatially misaligned data. *Environmetrics* **19** 453–467.

MATTE, T. D., ROSS, Z., KHEIRBEK, I., EISL, H., JOHNSON, S., GORCZYNSKI, J. E., KASS, D., MARKOWITZ, S., PEZESHKI, G. and CLOUGHERTY, J. E. (2013). Monitoring intraurban spatial patterns of multiple combustion air pollutants in new york city: Design and implementation. *Journal of Exposure Science and Environmental Epidemiology* **23** 223–231.

MAYNARD, D., COULL, B. A., GRYPARIS, A. and SCHWARTZ, J. (2007). Mortality risk associated with short-term exposure to traffic particles and sulfates. *Environmental Health Perspectives* 751–755.

MCCANDLESS, L. C., RICHARDSON, S. and BEST, N. (2012). Adjustment for missing confounders using external validation data and propensity scores. *Journal of the American Statistical Association* **107** 40–51.

MENG, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* 538–558.

MITRA, R. and DUNSON, D. (2010). Two-level stochastic search variable selection in glms with missing predictors. *The international journal of biostatistics* **6**.

MORENO, T., QUEROL, X., ALASTUEY, A., VIANA, M. and GIBBONS, W. (2009). Profiling transient daytime peaks in urban air pollutants: city centre traffic hotspot versus urban background concentrations. *Journal of Environmental Monitoring* **11** 1535–1542.

NENADIC, Z. and BURDICK, J. W. (2005). Spike detection using the continuous wavelet transform. *Biomedical Engineering, IEEE Transactions on* **52** 74–87.

OLIVER, M. A. and WEBSTER, R. (1990). Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information System* **4** 313–332.

PETROSIAN, A. A. and MEYER, F. G. (2013). *Wavelets in signal and image analysis: from theory to practice*, vol. 19. Springer Science & Business Media.

POLLOCK, S. and CASCIO, I. L. (2007). Non-dyadic wavelet analysis. In *Optimisation, Econometric and Financial Analysis*. Springer, 167–203.

POPE III, C. A. (2007). Mortality effects of longer term exposures to fine particulate air pollution: review of recent epidemiological evidence. *Inhalation Toxicology* **19** 33–38.

RAFTERY, A. E. (1995). Bayesian model selection in social research. *Sociological methodology* **25** 111–164.

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. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association* **91** 473–489.

SAMET, J. M., DOMINICI, F., CURRIERO, F. C., COURSAC, I. and ZEGER, S. L. (2000). Fine particulate air pollution and mortality in 20 us cities, 1987–1994. *New England journal of medicine* **343** 1742–1749.

SCHENKER, N. and TAYLOR, J. M. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis* **22** 425–446.

STÜRMER, T., SCHNEEWEISS, S., AVORN, J. and GLYNN, R. J. (2005). Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American journal of epidemiology* **162** 279–289.

STÜRMER, T., SCHNEEWEISS, S., ROTHMAN, K. J., AVORN, J. and GLYNN, R. J. (2007). Performance of propensity score calibrationa simulation study. *American journal of epidemiology* **165** 1110–1118.

SWELDENS, W. (1998). The lifting scheme: A construction of second generation wavelets. *SIAM Journal on Mathematical Analysis* **29** 511–546.

SZPIRO, A. A. and PACIOREK, C. J. (2013). Measurement error in two-stage analyses, with application to air pollution epidemiology. *Environmetrics* **24** 501–517.

SZPIRO, A. A., PACIOREK, C. J. and SHEPPARD, L. (2011a). Does more accurate exposure prediction necessarily improve health effect estimates? *Epidemiology (Cambridge, Mass.)* **22** 680.

SZPIRO, A. A., SHEPPARD, L. and LUMLEY, T. (2011b). Efficient measurement error correction with spatially misaligned data. *Biostatistics* kxq083.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

WAND, M., ORMEROD, J. T. ET AL. (2011). Penalized wavelets: Embedding wavelets into semiparametric regression. *Electronic Journal of Statistics* **5** 1654–1717.

WANG, C., DOMINICI, F., PARMIGIANI, G. and ZIGLER, C. M. (2015). Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. *Biometrics* .

WANG, C., PARMIGIANI, G. and DOMINICI, F. (2012). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics* **68** 661–671.

XIONG, R., XU, J. and WU, F. (2006). A lifting-based wavelet transform supporting non-dyadic spatial scalability. In *Image Processing, 2006 IEEE International Conference o=n*. IEEE.