Bayesian Analysis (2007) **2**, Number 3, pp. 611–634

A Spatially-adjusted Bayesian Additive Regression Tree Model to Merge Two Datasets

Song Zhang,* Ya-Chen Tina Shih† and Peter Müller‡

Abstract. Scientific hypotheses of interest often involve variables that are not available in a single survey. This is a common problem for researchers working with survey data. We propose a model-based approach to provide information about the missing variable. We use a spatial extension of the BART (Bayesian additive regression tree) model. The imputation of the missing variables and inference about the relationship between two variables are obtained simultaneously as posterior inference under the proposed model. The uncertainty due to imputation is automatically accounted for. A simulation analysis and an application to data on self-perceived health status and income are presented.

Keywords: BART, CART, Missing variables, Spatial model, Survey

1 Introduction

We consider the problem of inference about the relationship of two variables reported in two different datasets. This is a common problem for researchers working with survey data. Scientific hypotheses of interest often involve variables that are not available in a single survey. Specifically, we are interested in inference on how a variable z is affected by another variable y, when there is no such dataset that collects z and y simultaneously. Instead, z is only reported in dataset D_1 and y is only collected in dataset D_2.

Many model-based methods have been developed to deal with missing data problems, including maximum likelihood (ML) methods, multiple imputation (MI) methods, weighted estimating equations (WEE), and fully Bayesian (FB) methods. See Little (1992), Horton and Laird (1999), Schafer and Graham (2002), Ibrahim et al. (2005) and the references therein for detailed discussions. There are some assumptions associated with each of these methods. Many ML methods assume a large sample size so that the ML estimates are approximately unbiased and normally distributed. The likelihood function is assumed to arise from a parametric model of the complete data. Finally, ML methods usually require the missing at random (MAR) assumption (Rubin 1976). MI methods also rely on large-sample approximation and assume a parametric form for the joint model of the observed and missing data. They require some assumption about the distribution of missingness, although not limited to MAR. WEE methods are extensions of generalized estimating equations (GEE). Two models need to be specifed:

*Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX, mailto:songzhang@mdanderson.org

†Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX, http://gsbs.uth.tmc.edu/tutorial/shih.html

‡Department of Biostatistics, University of Texas MD Anderson Cancer Centeri, Houston, TX, http://odin.mdacc.tmc.edu/~pm/

One regression model for the data, and the other describing the missingness mechanism. WEE methods are considered to be doubly robust because the estimates of the regression parameters remain consistent as long as one of the two models is correctly specified. The MAR assumption and a large sample size are required. FB methods do not require a large sample size. Specifying a joint probability model, however, they require assumptions about the sampling model for the data and about the missingness mechanism. In summary, all the above methods regard missingness as a probabilistic phenomenon. In contrast, in the following discussion, missingness is not random. The variable y is missing for all records in D_1.

The most commonly applied method to borrow information from one dataset (i.e., y in D_2) to provide information not collected in another dataset (i.e., D_1) is the use of census-based socioeconomic status (SES) characteristics to supplement individual-level data, such as medical records, claims or registries (Gornick et al. 1996; Geronimus and Bound 1998; Devesa and Diamond 1983). The census-based approach obtains aggregate statistics of SES variables at certain geographic levels (e.g., census track, county, or zip code) and uses these aggregate numbers as proxy measures of SES in individual-level data. It has been used extensively in studies of health disparities. For more examples, see Mandelblatt et al. (1991), Kraus et al. (1986) and Byrne et al. (1994).

Geronimus and Bound (1998) cautioned that although the census-based approach is easy to execute, these aggregate measures should not be interpreted as if they were micro-level variables. The approach has several limitations. It requires detailed residential information to be collected in D_1. If due to privacy concerns this information is not collected or is not detailed enough (for example, only state code is available), then the method breaks down. The method only makes use of geographic information. Other individual-level covariates are ignored. For example, if we are interested in imputing missing income in D_1, then information such as age, gender, education, occupation could be very informative. Finally, the true value of the missing variable in D_1 may not match the neighborhood profile. This uncertainty is usually ignored.

In this paper we propose to approach the problem in the framework of Bayesian hierarchical modeling. A spatially adjusted Bayesian additive regression tree (SBART) is defined to impute the missing variable in D_1 based on individual-level covariates as well as geographic information. SBART is an extension of the BART model. The idea of BART is to model an unknown function as a mixture of tree models. Each tree is a priori constrained to have a simple structure. It only contributes a small portion to the overall model. Chipman, George and McCulloch (2006a) demonstrated that the sum over all trees provides a sufficiently rich model to incorporate both direct effects and interaction effects of different orders. SBART extends BART by incorporating spatial random effects. Correlation among neighboring areas is utilized to improve inference. Our method implements a full probability model with likelihood and priors. The imputation of the missing variable and the inference about the relationship between the two variables are obtained simultaneously as posterior inference under the model, and the uncertainty due to imputation is accounted for automatically. Unrelated to the problem of merging datasets that we consider here, a similar spatial extension of the

BART model has been developed independently in current work by Chipman, George, McCulloch and Musio (2006b).

The outline of the paper is as follows. Section 2 introduces notation and presents the Bayesian hierarchical model. A simulation study is conducted in Section 3. We illustrate our method with a data analysis example in Section 4. Finally, Section 5 discusses some limitations of our method as well as some possible extension.

2 A Spatial BART Model

Let I be the number of spatial units at the finest level of detail recorded in both datasets. This could be, for example, census tract, zip code area or county.

In dataset D_1, let m_i denote the number of subjects from area i ($i = 1, \cdots, I$). The sample size of D_1 is $m = \sum_{i=1}^{I} m_i$. For the jth subject from area i, we are interested in the relationship between variables z_{ij} and y_{ij}, where z_{ij} but not y_{ij} is recorded in dataset D_1. We use v_{ij} to denote a vector of other individual-level covariates reported in D_1.

The variable y_{ij} that is missing in D_1 is recorded on a different set of individuals in dataset D_2. For notational ease, we use the variable name x_{ij} rather than y_{ij}, to distinguish the fact that these variable values are recorded in D_2 rather than D_1. Similarly, for the vector of variables v_{ij}, we use w_{ij} rather than v_{ij} for those variables recorded in D_2. We assume w_{ij} and v_{ij} to be consistent, i.e., they record the same variables and use the same coding for the values. Because it would be unusual for all covariates recorded in D_1 and D_2 to be consistent, we only assume that after suitable pre-processing a subset of the covariates can be considered consistent across the two datasets. Let n_i be the number of subjects from area i, so $j = 1, \ldots, n_i$ in D_2. Then $n = \sum_{i=1}^{I} n_i$ is the sample size of D_2.

We define $\boldsymbol{Z} = \{z_{ij}, i = 1, \cdots, I, j = 1, \cdots, m_i\}$. Similarly we use $\boldsymbol{Y}$, $\boldsymbol{V}$, $\boldsymbol{X}$ and $\boldsymbol{W}$ to denote the vector of all y_{ij}, v_{ij}, x_{ij} and w_{ij}, respectively.

We describe in words how the proposed approach facilitates learning about the relationship between $\boldsymbol{Z}$ and $\boldsymbol{Y}$ with $\boldsymbol{Y}$ missing. We assume that $(\boldsymbol{Y}, \boldsymbol{V})$ (in D_1) and $(\boldsymbol{X}, \boldsymbol{W})$ (in D_2) arise from the same model M. We use the posterior for the parameters in M, obtained conditional on $(\boldsymbol{X}, \boldsymbol{W})$ to impute the missing $\boldsymbol{Y}$ conditional on $\boldsymbol{V}$. Finally, the regression of $\boldsymbol{Z}$ on the imputed $\boldsymbol{Y}$ approximates the relationship between $\boldsymbol{Z}$ and $\boldsymbol{Y}$. By integrating with respect to $\boldsymbol{Y}$, the marginal posterior distribution of the regression parameter $\boldsymbol{\beta}$ accounts for the variability induced by the imputation. The described learning process is complicated by the need to specify a joint probability model for $(\boldsymbol{Z}, \boldsymbol{Y}, \boldsymbol{X} \mid \boldsymbol{V}, \boldsymbol{W})$. Details are described later.

For the learning process to work we make the key assumption that $(\boldsymbol{X}, \boldsymbol{W})$ and $(\boldsymbol{Y}, \boldsymbol{V})$ are independent samples from the same model. This assumption ensures that we can apply what we have learned from $(\boldsymbol{X}, \boldsymbol{W})$ to $(\boldsymbol{Y}, \boldsymbol{V})$. For example, this assumption is satisfied if both D_1 and D_2 are representative samples from the U.S. population.

2.1 The Sampling Model

The proposed approach is model-based. We start the model construction with assumed sampling models for $\boldsymbol{Z}$, $\boldsymbol{X}$ and $\boldsymbol{Y}$. In the following description, we use $N(m, s^2)$ to denote a normal distribution with moments m and s^2. We assume that a sampling model $p(z_{ij} \mid y_{ij}, v_{ij}, \Phi)$ is available for z_{ij}, conditional on v_{ij} and assumed values for y_{ij}, and indexed by a set of parameters Φ. For example, if z_{ij} is continuous, we can assume a linear regression model with z_{ij} being the dependent variable, y_{ij} and v_{ij} defining the design matrix, and Φ including the regression coefficients and variance parameter. If z_{ij} is ordinal, an ordinal probit model may be used. Specific examples of $p(z_{ij} \mid y_{ij}, v_{ij}, \Phi)$ are used in the simulation study and the case study.

The model $p(x_{ij} \mid w_{ij}, f, \boldsymbol{\theta}, \sigma^2)$ describes the relationship between x_{ij} and w_{ij}. Specifically, we assume

$$x_{ij} \mid w_{ij}, f, \boldsymbol{\theta}, \sigma^2 \sim N\left(f(w_{ij}) + \theta_i, \sigma^2\right), \tag{1}$$

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_I)'$ is a vector of random spatial effects, $f(w_{ij})$ is an unknown function associating x_{ij} with w_{ij}, and σ^2 is the residual variance. We represent the mean function $f(w_{ij})$ as a BART model. Since the additional random effects θ_i introduce the desired spatial correlation among neighboring areas, we refer to model (1) as the spatially-adjusted Bayesian additive regression tree (SBART) model.

For reference, and to introduce notation for later use, we give a brief review of the BART model. See Chipman et al. (2006a) for details. We begin with the notation for a single tree model. Let T denote a tree. Its nodes can be divided into two categories, interior nodes and terminal nodes. A splitting rule is defined at each interior node. We limit splitting rules to binary splits. Each rule consists of a splitting variable and a splitting value. The splitting value is a threshold on the splitting variable that defines the splitting rule. Starting from the root, an individual with covariates w_{ij} selects branches in the tree according to the splitting rules until it is assigned to a terminal node. Suppose that there are K terminal nodes. We define $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_K)'$, with μ_k being assigned to the kth terminal node. The tree maps each covariate vector w_{ij} into one element of $\boldsymbol{\mu}$. A single tree model is denoted by the pair $(T, \boldsymbol{\mu})$, and the association between μ_k and w_{ij} through a tree T is written as $\mu_k = g(w_{ij}, T, \boldsymbol{\mu})$.

The BART model defines a summation of such tree models, as

$$f(w_{ij}) = g(w_{ij}, T_1, \boldsymbol{\mu}_1) + g(w_{ij}, T_2, \boldsymbol{\mu}_2) + \cdots + g(w_{ij}, T_L, \boldsymbol{\mu}_L),$$

where L is the total number of trees that form the BART. We usually assign a large value for L (e.g., $L = 200$) to encourage flexibility. On the other hand, to avoid overfitting, the BART model includes a strong prior on each tree to keep its effect small, effectively making each tree into a "weak learner". But overall, the sum of trees provides a sufficiently rich model to fit a variety of functions. For example, μ_k represents an interaction effect if its assignment involves more than one component of w_{ij} (i.e., more than one splitting variable). Furthermore, because $f(w_{ij})$ can be based on trees of different sizes, the BART model can incorporate both direct effects and interaction

effects of different orders. SBART extends BART by incorporating an additional spatial effect into the conditional mean of x_{ij} given w_{ij}.

BART is closely related to ensemble methods that combine a set of tree models. Examples of ensemble methods include boosting, bagging and random forests. Boosting (Freund and Schapire 1997; Friedman 2001) fits a sequence of trees. Each tree is fit conditional on data variation that is not explained by the other trees. Bagging (Breiman 1996; Clyde and Lee 2001) and random forests (Breiman 2001) construct a large number of independent trees through data randomization and stochastic search. The methods then use an average of the trees to improve prediction. Ensemble methods are not derived as coherent inference under a probability model. In contrast, BART is a model-based approach that reports inference as the summary of a full probabilistic description of all relevant uncertainties. Bayesian single tree models have been developed by Chipman et al. (1998) and Denison et al. (1998). Compared with single tree models, the sum-of-trees models provide vastly more flexibility by easily incorporating additive effects. Chipman et al. (2006a) provided a posterior Markov chain Monte Carlo (MCMC) simulation scheme for the BART model. They demonstrated that the proposed MCMC simulation has good mixing properties.

The third part of the top-level sampling model is an assumed model for y_{ij} conditional on the observed covariate vector v_{ij}. We assume the same model as for the regression of x_{ij} on w_{ij}:

$$y_{ij} \mid v_{ij}, f, \boldsymbol{\theta}, \sigma^2 \sim N\left(f(v_{ij}) + \theta_i, \sigma^2\right),$$

with $f(\cdot)$ defined by the SBART model as before.

2.2 The Prior Model

We complete the Bayesian hierarchical model with priors $p(\Phi)$, $p(f)$, $p(\boldsymbol{\theta})$ and $p(\sigma^2)$, for Φ, f, $\boldsymbol{\theta}$ and σ^2, respectively. We assume a priori independence.

The choice of $p(\Phi)$ depends on the particular form of $p(z_{ij} \mid y_{ij}, v_{ij}, \Phi)$. For example, in a linear regression model, conjugate priors are technically convenient choices. That is, normal priors for the regression coefficients and an inverse Gamma prior for the residual variance.

The BART model in (1) is indexed by $\{(T_l, \boldsymbol{\mu}_l), l = 1, \cdots, L\}$. We use

$$p(f) = \prod_{l=1}^{L} p(T_l, \boldsymbol{\mu}_l) = \prod_{l=1}^{L} \Big\{ p(T_l) \cdot p(\boldsymbol{\mu}_l \mid T_l) \Big\}.$$

Following Chipman et al. (2006a), we define $p(T_l)$ by three factors, corresponding to a node being non-terminal, the selection of the splitting variable for a non-terminal node, and the choice of the splitting value conditional on a chosen splitting variable. The probability that a node at depth d is nonterminal, is assumed to be

$$\alpha(1+d)^{-\gamma},$$

where $\alpha \in (0,1)$ and $\gamma \in [0,\infty)$ are two hyper-parameters reflecting our prior belief about the tree. For example, if we believe that the depth of the tree should be small, we can assign a big value for γ, so that the probability decays fast with d. Chipman et al. (2006a) proposed $\alpha = 0.95$ and $\gamma = 2$ as default values, which implies that with prior probability 0.05, 0.55, 0.28, 0.09 and 0.03, the tree has 1, 2, 3, 4, and ≥ 5 terminal nodes, respectively. A natural choice for the selection of the splitting variable, conditional on a node being non-terminal, is a uniform prior over all available variables. A default choice for the distribution of the splitting value is a uniform distribution over the set of available splitting values. Finally, we define a prior for $\boldsymbol{\mu}_l$. Let μ_{lk} be the kth element of $\boldsymbol{\mu}_l$. Conditional on T_l, we assume i.i.d. normal priors for μ_{lk}. The mean and variance of the normal prior are specified in such a way that each tree is constrained to be a weak learner, and it plays a small role in the overall fit. More details can be found in Chipman et al. (2006a), Section 3.2.

For the spatial random effects $\boldsymbol{\theta}$ we use a conditionally autoregressive (CAR) prior. The key idea of the CAR model is simple. It formalizes the notion that each area is similar to its neighbors. Specifically, we define $p(\boldsymbol{\theta})$ by the set of conditional distributions

$$p(\theta_i \mid \theta_{(-i)}, \rho, \delta^2) = N\Big(\frac{\rho}{h_i}\sum_{j\neq i} c_{ij}\theta_j,\ \frac{1}{h_i}\delta^2\Big),\ i = 1, \cdots, I, \tag{2}$$

where $\theta_{(-i)}$ denotes all the elements of $\boldsymbol{\theta}$ except θ_i; ρ is a parameter with range $(-1,1)$; δ^2 is the variance component; $c_{ij} = 1$ $(i \neq j)$ if area i and area j are neighbors, and $c_{ij} = 0$ otherwise, including $c_{ii} = 0$; and $h_i = \sum_{j=1}^{I} c_{ij}$ is the total number of neighbors for area i. The joint distribution $p(\boldsymbol{\theta})$ implied by (2) is

$$p(\boldsymbol{\theta} \mid \rho, \delta^2) = N\Big(0,\ \delta^2(\boldsymbol{H} - \rho\boldsymbol{C})^{-1}\Big), \tag{3}$$

where $\boldsymbol{C} = (c_{ij})$ is an $I \times I$ adjacency matrix, and $\boldsymbol{H}$ is an $I \times I$ diagonal matrix with h_i being the diagonal elements. Model (2) specifies that given random effects from all the other areas, the distribution of θ_i only depends on its neighbors. When $\rho = 0$, the variance matrix in (3) is diagonal, implying that θ_i are independent. When $\rho = 1$, the conditional mean of θ_i in (2) equals the average of its neighbors. However, $\rho = 1$ implies that $\boldsymbol{H} - \rho\boldsymbol{C}$ is singular. That is, the covariance matrix of $\boldsymbol{\theta}$ does not exist. Sun et al. (1999) specified $-1 < \rho < 1$ as a smoothing or spatial correlation parameter. It can be thought of as a measure of spatial association. For more discussion of CAR models, see Cressie (1993) page 407, Besag et al. (1991), Clayton and Kaldor (1987) and Whittle (1954).

We complete the prior model with probability models for the hyper-parameters σ^2, ρ and δ^2. Chipman et al. (2006a) assumed $p(\sigma^2)$ to be an inverse chi-square distribution $\sigma^2 \sim \nu\lambda/\chi^2_\nu$, where ν is the degree of freedom. This is a special case of the inverse Gamma distribution. The key idea to specify the hyper-parameters ν and λ is to first obtain a preliminary estimate $\hat{\sigma}^2$ by exploratory data analysis (for example, through linear regression of x_{ij} and w_{ij}), and then specify ν and λ such that $\hat{\sigma}^2$ matches the qth quantile of $p(\sigma^2)$. The default setting recommended by Chipman et al. (2006a) is $(\nu, q) = (3, 0.90)$. Finally, we define prior distributions for the parameters ρ and δ^2 in

the CAR model. It is natural to assume that the spatial effects are positively correlated. We therefore assume ρ to be uniform between 0 and 1, i.e., $U(0,1)$. We assume $p(\delta^2)$ to be an inverse Gamma distribution, denoted by $IG(a_\delta, b_\delta)$, with density function

$$p(\delta^2) \propto \frac{1}{(\delta^2)^{a_\delta+1}} \exp(-\frac{b_\delta}{\delta^2}).$$

Here a_δ and b_δ are fixed hyperparameters.

For reference, we state the joint probability model on the data $\boldsymbol{Z}$, $\boldsymbol{Y}$, $\boldsymbol{X}$ and the parameters:

$$p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{V}, \Phi) \cdot p(\boldsymbol{X} \mid \boldsymbol{W}, f, \boldsymbol{\theta}, \sigma^2) \cdot p(\boldsymbol{Y} \mid \boldsymbol{V}, f, \boldsymbol{\theta}, \sigma^2) \cdot p(\Phi) \cdot p(f) \cdot p(\boldsymbol{\theta} \mid \rho, \delta^2) \cdot p(\sigma^2) \cdot p(\rho) \cdot p(\delta^2), \quad (4)$$

where

$$\begin{aligned} p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{V}, \Phi) &= \prod_{i=1}^{I} \prod_{j=1}^{m_i} p(z_{ij} \mid y_{ij}, v_{ij}, \Phi), \\ p(\boldsymbol{X} \mid \boldsymbol{W}, f, \boldsymbol{\theta}, \sigma^2) &= \prod_{i=1}^{I} \prod_{j=1}^{n_i} p(x_{ij} \mid w_{ij}, f, \boldsymbol{\theta}, \sigma^2), \\ p(\boldsymbol{Y} \mid \boldsymbol{V}, f, \boldsymbol{\theta}, \sigma^2) &= \prod_{i=1}^{I} \prod_{j=1}^{m_i} p(y_{ij} \mid v_{ij}, f, \boldsymbol{\theta}, \sigma^2). \end{aligned}$$

We are interested in the inference on Φ given all observations, namely $p(\Phi \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{W})$. Carrying out the desired inference requires integration with respect to $\boldsymbol{Y}$ and the other parameters. This integration does not have a closed form solution. We set up MCMC simulation and obtain inference based on random samples from the posterior distribution of Φ. Details of the sampling scheme can be found in the Appendix. By integrating out $\boldsymbol{Y}$, $p(\Phi \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{W})$ automatically accounts for the variability induced by the imputation. A byproduct of this process is the imputation of the missing variable $\boldsymbol{Y}$, which can be obtained as random samples from $p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{W})$.

3 A Simulation Study

We conduct a simulation study to examine the performance of the proposed approach. We define $I = 99$ spatial areas, with an assumed spatial structure (adjacency matrix $\boldsymbol{C}$) equal to that of the 99 counties in the state of Iowa. We also assume $n_i = 4$ and $m_i = 2$ for $i = 1, \cdots, I$. Thus we have sample size $n = 396$ and $m = 198$.

The simulated data are generated as follows. We assume covariate vectors w_{ij} and v_{ij} to be of dimension 10. Each of the 10 elements is generated from independent $U(0,1)$ distribution. We generate the simulation truth for the spatial random effects $\boldsymbol{\theta}$ from a

$N(\mathbf{0}, \delta^2(\boldsymbol{H} - \rho\boldsymbol{C})^{-1})$ distribution, using $\rho = 0.3$ and $\delta = 1$. The mean function $f(u)$ is evaluated as

$$f(u) = 10\sin(\pi u_1 u_2) + 20(u_3 - 0.5)^2 + 10u_4 + 5u_5, \tag{5}$$

where u_i is the ith element of $u = (u_1, \cdots, u_{10})'$. The same function was used in simulation in Friedman (1991) and Chipman et al. (2006a). The added variables together with the interactions and nonlinearities make it difficult to fit the model by standard parametric methods. Conditional on the covariates w_{ij}, we generate x_{ij} by

$$x_{ij} \mid w_{ij}, f, \boldsymbol{\theta}, \sigma^2 \quad \sim \quad N(f(w_{ij}) + \theta_i, \sigma^2),$$

using $\sigma = 0.2$. Similarly, we generate y_{ij} conditional on v_{ij},

$$y_{ij} \mid v_{ij}, f, \boldsymbol{\theta}, \sigma^2 \quad \sim \quad N(f(v_{ij}) + \theta_i, \sigma^2).$$

Thus x_{ij} and y_{ij} only depend on the first 5 elements of w_{ij} and v_{ij}, respectively.

Finally, z_{ij} is generated by

$$z_{ij} \mid y_{ij}, v_{ij}, \boldsymbol{\beta}, \tau^2 \sim N(h(v_{ij}, y_{ij}, \boldsymbol{\beta}), \tau^2), \tag{6}$$

where we assume $\tau = 0.2$, $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_6)' = (3, -3, -2.5, -1, 1.5, 2, 1)'$, and

$$h(v_{ij}, y_{ij}, \boldsymbol{\beta}) = \beta_0 + v_{ij4}\beta_1 + v_{ij5}\beta_2 + v_{ij6}\beta_3 + v_{ij7}\beta_4 + v_{ij8}\beta_5 + y_{ij}\beta_6.$$

Here v_{ijk} denotes the kth element of v_{ij}. The simulation model for z_{ij} is a linear regression model. We assume that part of the covariates (v_{ij4}, v_{ij5}) are involved in the generation of y_{ij} and others $(v_{ij6}, v_{ij7}, v_{ij8})$ are not. Matching the earlier notation $p(z_{ij} \mid y_{ij}, v_{ij}, \Phi)$, we have $\Phi = (\boldsymbol{\beta}, \tau^2)$, where $\boldsymbol{\beta}$ is the vector of regression coefficients and τ^2 is the variance parameter.

Conditional on the simulated data $(\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{V})$, but pretending that $\boldsymbol{Y}$ is missing, we generate a Monte Carlo sample from the posterior distribution $p(\boldsymbol{\beta} \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{W})$ under model (4). See the Appendix for details of the posterior simulation.

We repeat the described simulation $K = 100$ times. For the kth simulation, we save the simulation truth $\boldsymbol{Y}^{(k)}$ and $\boldsymbol{\beta}$, the imputed values $\hat{\boldsymbol{Y}}^{(k)}$, and the estimated effects $\hat{\boldsymbol{\beta}}^{(k)}$. We obtain $\hat{\boldsymbol{Y}}^{(k)}$ and $\hat{\boldsymbol{\beta}}^{(k)}$ as marginal posterior expectations under $p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{W})$ and $p(\boldsymbol{\beta} \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{W})$, respectively. The mean squared error (MSE) for $\boldsymbol{Y}$ is defined as

$$MSE_Y = \frac{1}{Km} \sum_{k=1}^{K} \left\{ \sum_{i,j} (\hat{y}_{ij}^{(k)} - y_{ij}^{(k)})^2 \right\}.$$

Similarly, for $\boldsymbol{\beta}$ we define

$$MSE_{\beta_p} = \frac{1}{K} \sum_{k=1}^{K} \left\{ (\hat{\beta}_p^{(k)} - \beta_p)^2 \right\}, \; p = 0, 1, \cdots, 6.$$

For comparison we record results under two different models.

Table 1: MSE from Simulation to Compare SBART and BART

	(a)	(b)
β_0	0.0055	0.0062
β_1	0.0078	0.0154
β_2	0.0017	0.0045
β_3	0.0029	0.0030
β_4	0.0048	0.0048
β_5	0.0079	0.0090
β_6	0.0136	0.0321
$\boldsymbol{Y}$	0.596	3.864

Column (a) under SBART; Column (b) under BART.

M1: Model (4) with a $U(0,1)$ prior for ρ, an $IG(0.001, 0.001)$ prior for δ^2, and a CAR prior for $\boldsymbol{\theta}$. This is the proposed SBART model.

M0: Model (4) with $\boldsymbol{\theta} = 0$. This is a BART model without spatial adjustment. Under the BART model, the priors $p(\boldsymbol{\theta} \mid \rho, \delta^2)$, $p(\rho)$ and $p(\delta^2)$ are not needed.

The remaining prior choices include a normal prior for $\boldsymbol{\beta}$, $p(\boldsymbol{\beta}) = N(\mathbf{0}, 100\boldsymbol{I}_6)$, and an inverse Gamma prior for τ^2, $p(\tau^2) = IG(0.001, 0.001)$. Here $\mathbf{0}$ is a vector of $0's$ and $\boldsymbol{I}_6$ is an identity matrix of dimension 6. For the hyper-parameters in $p(f)$ and $p(\sigma^2)$, we use the default setting recommended by Chipman et al. (2006a).

Table 1 compares the MSE from models **M1** and **M0**. The results suggest that when spatial correlation is present, incorporating spatial effects improves the estimation of regression coefficients. This is particularly true for β_6, the coefficient of the missing variable, which is of primary interest. In the simulation, the MSE of β_6 is reduced from 0.0321 to 0.0136. A byproduct of the proposed approach is the inference about the missing variable, which might be of interest to researchers by itself. Monte Carlo sample averages evaluate posterior means and provide point estimates of the missing variables. Other summaries characterize the uncertainty of the imputation. Table 1 shows that incorporating spatial effects greatly improves the imputation of the missing variable. The MSE for $\boldsymbol{Y}$ is reduced from 3.864 to 0.596. This improvement can also be seen in Figure 1, where we plot $\boldsymbol{Y}^{(k)}$ versus $\hat{\boldsymbol{Y}}^{(k)}$ from one simulation.

The estimated spatial correlation parameter $\hat{\rho}^{(k)}$ has a mean 0.414 and a standard deviation 0.091, suggesting a slight overestimation of ρ. The histogram of $\hat{\rho}^{(k)}$ is plotted in Figure 2. We also plot $\boldsymbol{\theta}^{(k)}$ against $\hat{\boldsymbol{\theta}}^{(k)}$, the true and estimated values of $\boldsymbol{\theta}$, respectively, from one simulation in Figure 3. The fact that the points fall around the 45 degree line suggests that the method successfully recovers the spatial pattern.

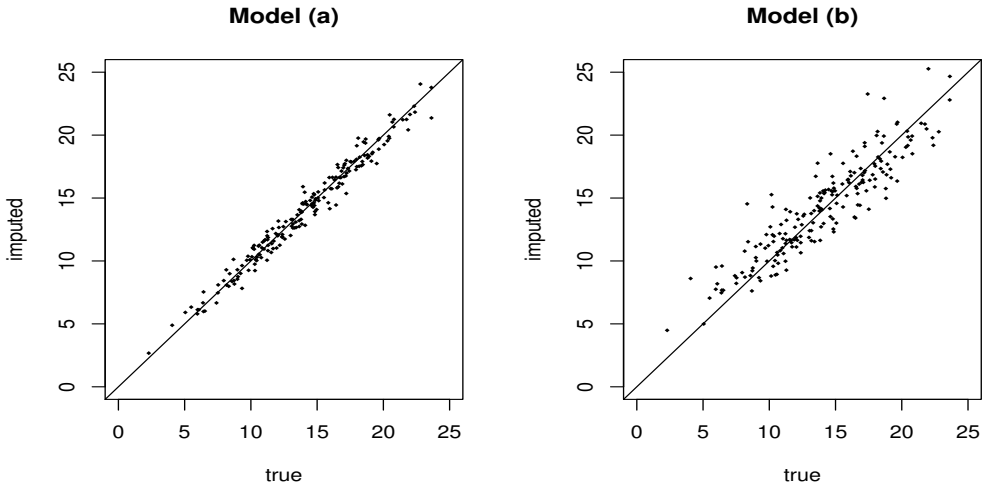

Figure 1: Simulation example. The imputation of $\boldsymbol{Y}$ under **M1** and **M0** (under one simulation). **M1** uses the SBART model. **M0** uses the BART model without spatial random-effects.

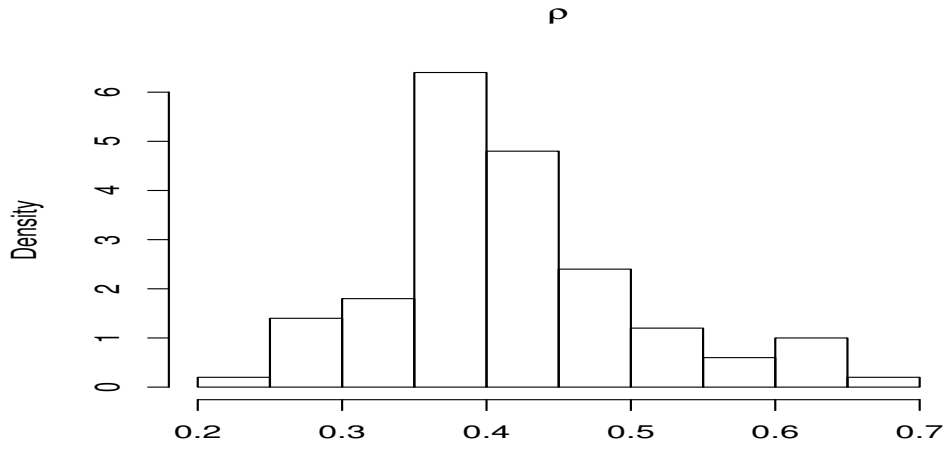

Figure 2: Simulation example. Histogram of $p(\hat{\rho}^{(k)} \mid data)$.

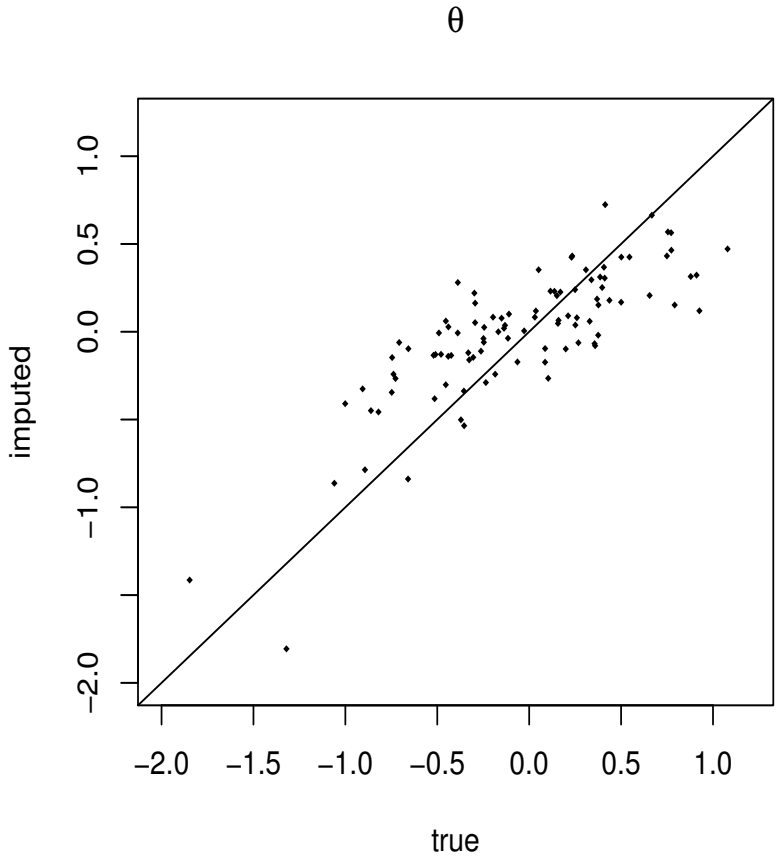

Figure 3: Simulation example. Simulation truth and imputed values of $\boldsymbol{\theta}$.

4 Joint Inference with the CPS and SIPP Surveys

We evaluate the proposed approach with real survey data. In this evaluation, we apply our method to explore the relationship between self-perceived health status and income using two different surveys. One survey includes data on health status, income, and other variables $(\boldsymbol{Z}, \boldsymbol{Y}, \boldsymbol{V})$. The second survey reports income and other variables $(\boldsymbol{X}, \boldsymbol{W})$.

We implement inference through the proposed approach *without* using the observed values of income $\boldsymbol{Y}$ in the first survey. That is, we carry out the analysis pretending that we did not have income $(\boldsymbol{Y})$ information in the first survey.

For comparison, we also implement inference *with* the observed $\boldsymbol{Y}$ values. Using data from the first survey only, we implement posterior simulation in the model

$$p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{V}, \Phi) \cdot p(\Phi), \tag{7}$$

and summmarize $p(\Phi \mid \boldsymbol{Z}, \boldsymbol{Y}, \boldsymbol{V})$. By comparing the inference with $\boldsymbol{Y}$ missing versus inference conditional on $\boldsymbol{Y}$, we will validate the proposed model.

4.1 The Datasets

We let D_1 be a dataset extracted from the 2001 Current Population Survey (CPS), March Supplement. The variable $\boldsymbol{Z}$ is self-perceived health status with values 1 to 5, where 1 denotes the best health status and 5 denotes the poorest health status. The variable $\boldsymbol{Y}$ is defined to be total personal income. We are interested in the relationship between $\boldsymbol{Z}$ and $\boldsymbol{Y}$. The set of individual-level covariates are denoted by $\boldsymbol{V}$, which include age, race, gender, education, health insurance coverage, marital status, employment, industry and occupation. The dataset D_2 comes from the 2001 Survey of Income and Program Participation(SIPP), where total personal income $\boldsymbol{X}$ and the other covariates $\boldsymbol{W}$ are collected. Both CPS and SIPP report income, denoted as $\boldsymbol{Y}$ in CPS and $\boldsymbol{X}$ in SIPP. We pretend, however that $\boldsymbol{Y}$ is missing in D_1 to illustrate and validate the proposed method. CPS and SIPP are two independent surveys that each collects information from a representative sample of the U.S. civilian noninstitutional population. It is therefore reasonable to assume that $(\boldsymbol{Y}, \boldsymbol{V})$ and $(\boldsymbol{X}, \boldsymbol{W})$ arise from the same model.

CPS reports annual income while SIPP collects the information of monthly income. To make the income variables consistent between two datasets, we scale them to a common range of 0 to 1. Furthermore, personal income is known to be heavily skewed to the right, which makes the normal assumption in (1) inappropriate. We carry out a square root transformation to mitigate the problem. Thus eventually $\boldsymbol{Y}$ and $\boldsymbol{X}$ denote the square root of the scaled personal income.

The finest available spatial area in both datasets is metropolitan statistical area (MSA), which is defined as a core area that contains a substantial population nucleus, together with adjacent communities having a high degree of social and economic integration with that core. MSAs comprise one or more entire counties. In D_1 and D_2 there

are altogether $I = 239$ MSAs. The original datasets from CPS and SIPP have more than 90,000 and 260,000 records, respectively. For this illustrative analysis, we obtain D_1 and D_2 by randomly sampling 10,000 observations from each of the two original datasets.

4.2 Model Specification

Health status $\boldsymbol{Z}$ is an ordinal categorical variable. We construct an ordinal probit model $p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{V}, \Phi)$. We define the probit model by introducing a latent normal random variable

$$\eta_{ij} \mid \boldsymbol{\beta}, \tau^2 \sim N(\beta_0 + v_{ij1}\beta_1 + v_{ij2}\beta_2 + v_{ij3}\beta_3 + y_{ij}\beta_4, \tau^2).$$

For given values of η_{ij} and a set of cut points $c_1, \cdots, c_4$, we set

$$z_{ij} \mid \eta_{ij} \quad = \quad \begin{cases} 1, & \text{if } \eta_{ij} \leq c_1, \\ r, & \text{if } c_{r-1} < \eta_{ij} \leq c_r \text{ for } r = 2, 3, 4, \\ 5, & \text{if } \eta_{ij} > c_4, \end{cases} \tag{8}$$

where $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_4)'$. See, for example, Johnson and Albert (1999) for a discussion of Bayesian inference in ordinal regression models, including the latent variable construction used here. The latent variable η_{ij} is assumed to arise from a linear regression model with covariates being personal income y_{ij}, health insurance coverage v_{ij1}, gender v_{ij2}, and age v_{ij3}, and $\boldsymbol{\beta}$ is the corresponding coefficient vector. Income and age are continuous; age ranges from 18 to 84; gender is binary with 0 indicating male and 1 indicating female; health insurance coverage is binary with 0 indicating covered and 1 indicating not covered. We define $\boldsymbol{\eta}$ to be the collection of η_{ij}, and $\Phi = (\boldsymbol{\eta}, \boldsymbol{\beta}, \tau^2)$. The cutpoints $(c_1, \cdots, c_4)$ are specified as fixed. Random cutpoints would provide more flexibility. For example, Johnson and Albert (1999) jointly update the cutpoints and the latent probit variable. However, the choice of the sampling model for $\boldsymbol{Z} \mid \boldsymbol{Y}$ is not directly related to the missing data problem. We assume fixed cutpoints to keep the model simple and keep the discussion focused.

The models $p(y_{ij} \mid v_{ij}, f, \boldsymbol{\theta}, \sigma^2)$ and $p(x_{ij} \mid w_{ij}, f, \boldsymbol{\theta}, \sigma^2)$ are defined in (1). We complete the model with priors for $(\rho, \delta^2, \boldsymbol{\beta}, \tau^2)$. We assume diffuse priors, a uniform prior for ρ, $p(\rho) = U(0, 1)$, an inverse Gamma prior for δ^2, $p(\delta^2) = IG(0.001, 0.001)$, independent normal priors for β_p, $p(\beta_p) = N(0, 100)$, $p = 0, \cdots, 4$, and an inverse Gamma prior for τ^2, $p(\tau^2) = IG(0.001, 0.001)$. We use default values recommended in Chipman et al. (2006a) for the hyper-parameters of $p(f)$ and $p(\sigma^2)$.

4.3 Implementation Details

Some practical issues arise in the application to real data. First, in fitting the model $p(y_{ij} \mid v_{ij}, f, \boldsymbol{\theta}, \sigma^2)$, we can use the entire vector of v_{ij}. There is no need for formal variable selection. As pointed out by Chipman et al. (2006a), the BART model is a nonparametric Bayesian regression approach which uses dynamic random basis elements

that are dimensionally adaptive. Variable selection is already part of the model. In contrast, $p(z_{ij} \mid y_{ij}, v_{ij}, \Phi)$ is a generalized linear model and inference can be sensitive to correlation among the covariates (y_{ij}, v_{ij}). Like any other regression analysis, the specification of $p(z_{ij} \mid y_{ij}, v_{ij}, \Phi)$ requires a good understanding of the research questions to identify the relevant covariates. Importantly, high linear correlation among (y_{ij}, v_{ij}) complicates interpretation and should be avoided. With y_{ij} missing, we use (x_{ij}, w_{ij}) instead to check for linear correlation among the covariates.

Another issue concerns a bias in the inference on Φ induced by the imputation of y_{ij}. Figure 4 clearly shows a shrinkage effect. An ideal imputation would have a scatter plot falling around the 45 degree line. In Figure 4 the range of the imputed values is much narrower compared with that of the true values. Chipman et al. (2006a) observed similar shrinkage in a simulation study, which they attributed to extreme extrapolation. That is, when we make prediction outside the observed data, because of lack of information, the prior takes over and the imputed values are shrunk towards the center. We believe, however, that the cause of shrinkage in Figure 4 is more than extreme extrapolation. If the shrinkage arises from extrapolation alone, then it should have equal effect on both extremes. In Figure 4, we see more shrinkage on the higher incomes than on the lower incomes. From this observation, we hypothesize that the shrinkage is caused by a violation of the normality assumption in model (1). If personal income is heavily skewed to the right, then the square root transformation does not suffice to achieve normality, and extremely high incomes are not correctly imputed.

We propose to address the issue of shrinkage through the following two steps. First we carry out a preliminary analysis using model (4). We compare the distribution of imputed income $\hat{\boldsymbol{Y}}$ based on $p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{W})$ with the observed income distribution from D_2. We use a deterministic adjustment to match some features of these two distributions. For example, in this study we construct a linear transformation of the imputed values, $t(\hat{y}_{ij}) = a\hat{y}_{ij} + b$, such that some selected quantiles (for example, the 10th and 90th quantiles) of $t(\hat{y}_{ij})$ match those of $\boldsymbol{X}$, the incomes observed in D_2. In the second step, we replace $p(z_{ij} \mid y_{ij}, v_{ij}, \Phi)$ in model (4) by

$$p^*(z_{ij} \mid y_{ij}, v_{ij}, \Phi) \equiv p(z_{ij} \mid t(y_{ij}), v_{ij}, \Phi), \tag{9}$$

and proceed with the final analysis. Because $t(y_{ij})$ is a one-to-one transformation of y_{ij}, $p(z_{ij} \mid y_{ij}, v_{ij}, \Phi)$ and $p^*(z_{ij} \mid y_{ij}, v_{ij}, \Phi)$ define the same conditional distribution. But the latter provides a better calibrated estimation of Φ by adjusting for the effect of shrinkage. See Foster and Stine (2004) for more discussion about calibration.

Effectively, the proposed two steps use the SBART model to impute the rank of the missing income variable, and use an observed distribution to set specific values. This approach is valid because both CPS and SIPP are conducted by the US Census Bureau to collect information from representative samples of the US population.

This adjustment can be automated in each MCMC iteration, where we readjust the values of a and b such that the selected quantiles of $t(y_{ij}^{(k)})$ match the corresponding quantiles in the empirical distribution of $\boldsymbol{X}$. We conducted a simple simulation study to assess the performance of the automated adjustment. Because the shrinkage effect

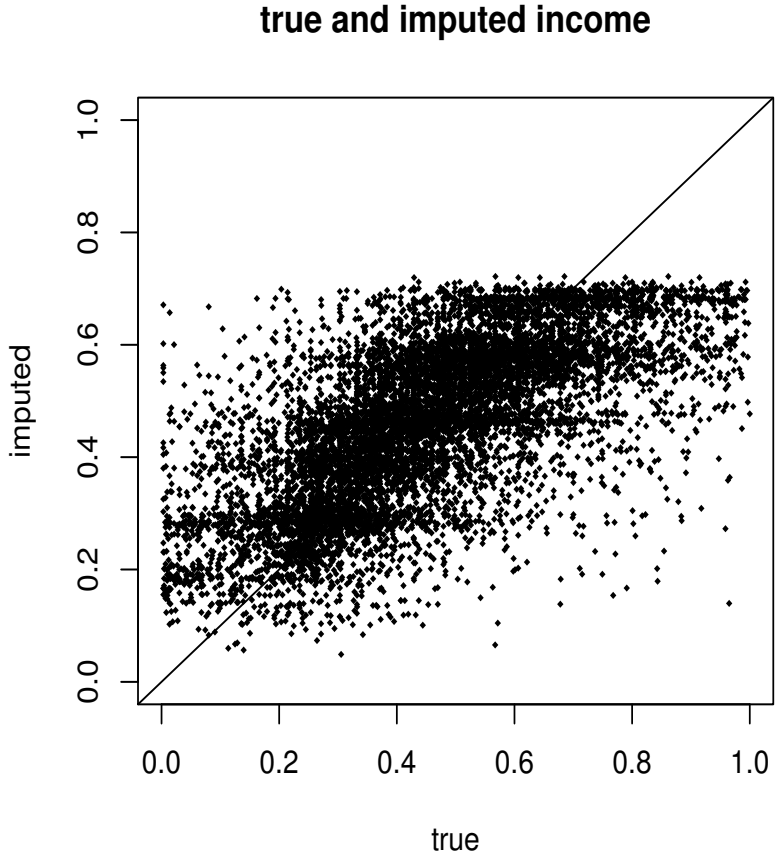

Figure 4: CPS survey: True and imputed income. Income is scaled between 0 and 1. Note the severe shrinkage in the imputed income.

Table 2: MSE from Simulation to Check Adjustment

	(a)	(b)
β_6	0.0078	0.0106
$\boldsymbol{Y}$	0.0238	0.0240

Column (a) with automated adjustment; Column (b) without adjustment.

is more obvious when the sample size is large, we set $n = m = 2000$. The simulation truth is similar to the model assumed in Section 3, except that we drop the spatial component $\boldsymbol{\theta}$ to facilitate computation, and the residual effects in $\boldsymbol{X}$ and $\boldsymbol{Y}$ are assumed to have Student t distribution with 3 degree of freedom. The MSE of the estimated regression coefficients and imputed $\boldsymbol{Y}$ are presented in Table 2. Because Table 2 is based on simulations with a larger sample size and a simpler model, the MSE are much smaller than those in Table 1. Our primary interest is in β_6, the regression coefficient of $\boldsymbol{Y}$. Without adjustment, the shrinkage effect leads to overestimation of β_6. With the automated adjustment, the MSE of β_6 is reduced from 0.0106 to 0.0078.

The simulation indicates that the adjustment can provide better calibrated estimates when there is some shrinkage effect induced by imputation of the missing variable. However, we caution that such an adjustment for shrinkage is ad hoc, and it relies heavily on the assumption that D_1 and D_2 are representative samples of the same population. Researchers should carefully check this assumptions before implementing the approach.

4.4 Results

Table 3 lists the posterior means and standard deviations of the regression coefficients $\boldsymbol{\beta}$ under three inference approaches, which are implemented by MCMC simulation. One set of inference summaries is based on true income and model (7). This serves as the gold standard. The second set of inferences is based on missing income and model (4). The third set is based on missing income and model (9). Both model (4) and model (9) are SBART models, the difference being that model (9) adjusts for the shrinkage effect while model (4) does not. Table 3 shows that if we ignore the shrinkage effect, model (4) will lead to a conclusion that overstates the effect of income. The posterior means based on model (7) and (9) are similar, suggesting that our method successfully merges information from two datasets and provides a good estimate of the relationship between self-perceived health status and income. Due to the uncertainty induced by imputing the missing income, the standard deviations under model (9) are slightly larger. The estimated regression coefficients suggest that subjects with higher income tend to have a better self-perceived health status. Women generally report better self-perceived health. Additionally, younger age and health insurance coverage are associated with better self-perceived health status. We plot the imputed income based on samples from $p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{W})$ versus the true income in Figure 4. The spatial correlation parameter ρ has a posterior mean 0.362 and standard deviation 0.242,

indicating a moderate spatial correlation. A histogram of the samples from its posterior distribution is plotted in Figure 5.

Table 3: Real Data, Posterior Mean (Standard deviation) of $\boldsymbol{\beta}$

	model (7)	model (4)	model (9)
Intercept	-4.157(0.073)	-3.982(0.075)	-4.019(0.075)
Health insurance	0.865(0.059)	0.624(0.069)	0.619(0.069)
Sex	-0.194(0.047)	-0.316(0.054)	-0.320(0.055)
Age	0.057(0.001)	0.052(0.001)	0.052(0.001)
Income	-2.513(0.126)	-3.392(0.216)	-2.677(0.167)

Model (7) uses true income; Model (4) uses SBART to impute "missing" income without adjusting for shrinkage; Model (10) uses SBART to impute "missing" income and adjusts for shrinkage.

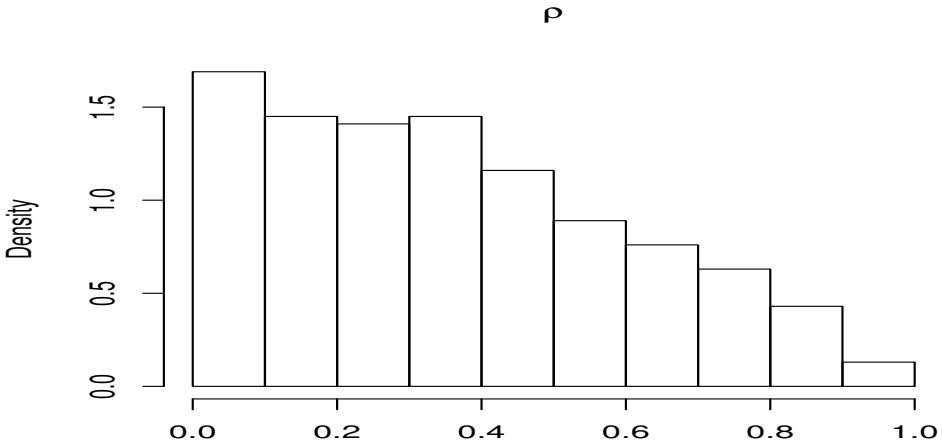

Figure 5: CPS and SIPP surveys. Histogram of $p(\rho \mid data)$.

Besides comparing our results with those based on the complete data, we also compare with results from a census-based approach, which supplements missing individual-level variables with aggregate information based on the neighborhood socioeconomic profile. With MSA being the finest available spatial area, we could supplement missing y_{ij} with average personal income from MSAi. However, compared with the average by census block or census track, the average by MSA is much coarser and would result in a large imputation error. To achieve a fairer comparison with the proposed method we instead proceed as follows. In the CPS dataset, about 41.5% of the records contain county codes. To investigate the performance of census-based methods with finer area

units, we create D_1^* by randomly sampling 10,000 observations from those that have county code in the original CPS dataset. We then replace the missing income with county median income (denoted by $\widetilde{\boldsymbol{Y}}$) from the US census. Conditioning on $(\boldsymbol{Z}, \widetilde{\boldsymbol{Y}}, \boldsymbol{V})$ we report inference on Φ under model (7). This is the result from the census-based method. Table 4 lists the posterior means (standard deviations) of $\boldsymbol{\beta}$ from three procedures: (a) based on model (7) and true income; (b) the proposed method, based on model (9) with missing income; (c) census-based method, based on model (7) and median income at county level. Because Table 3 is based on D_1 while Table 4 is based on D_1^*, the estimates in the two tables do not match exactly. The estimates from the proposed method are close to those based on true incomes. This is not the case for the estimates based on imputation by county median income. The estimated coefficients of health insurance coverage and income are quite different from those based on true incomes. Most strikingly, the estimated coefficient of sex switches the sign. This could lead to very misleading conclusions. In summary, Table 4 shows that our model provides an improvement of the census-based method. This is true even though we have improved the latter by using county median income while keeping our proposed method at the MSA-level, a coarser spatial area.

Table 4: Comparing with Census-Based Method

	(a)	(b)	(c)
Intercept	-4.220(0.073)	-4.129(0.075)	-4.380(0.076)
Health insurance	0.841(0.059)	0.734(0.066)	1.116(0.057)
Sex	-0.088(0.047)	-0.165(0.056)	0.148(0.046)
Age	0.054(0.001)	0.052(0.002)	0.052(0.001)
Income	-2.230(0.127)	-2.335(0.164)	-1.673(0.248)

Column (a) uses true income; column (b) uses SBART to impute missing income and adjusts for shrinkage; column (c) uses county median income as imputation.

5 Discussion

We have developed an approach that allows researchers to borrow information across surveys and investigate hypotheses that cannot be considered using only one dataset alone. The proposed method is flexible and fully model-based. The key assumption is that $(\boldsymbol{Y}, \boldsymbol{V})$ and $(\boldsymbol{X}, \boldsymbol{W})$ are independent samples from the same model. This assumption allows researchers to apply the knowledge learned from $(\boldsymbol{X}, \boldsymbol{W})$ to $(\boldsymbol{Y}, \boldsymbol{V})$. This facilitates imputation of the missing $\boldsymbol{Y}$. By specifying a flexible SBART model, the proposed method does not make restrictive assumptions about the specific model for $(\boldsymbol{X}, \boldsymbol{W})$.

In the simulation study and the data analysis example we have assumed parametric models for the regression of $\boldsymbol{Z}$ and $\boldsymbol{Y}$. This parametric form, however, is not a requirement for the proposed approach. It is unrelated to the missingness of $\boldsymbol{Y}$. Alternatively, a non-parametric regression model could be used. The only caveat is that the increased

uncertainty induced by the imputation of $\boldsymbol{Y}$ might make meaningful data analysis with a non-parametric model difficult.

The proposed imputation of the missing variable is a data-driven procedure. That is, in each MCMC iteration, we have a large number of trees such that each contributes a small portion of the conditional mean. Therefore it is difficult to evaluate the relationship between the missing variable and individual covariates. It is not a critical issue if the primary interest is to explore the relationship between $\boldsymbol{Z}$ and $\boldsymbol{Y}$, instead of $\boldsymbol{Y}$ and $\boldsymbol{V}$. If the researchers are interested in the the marginal effect of a single predictor, partial dependence plots might be a useful tool. See Friedman (2001) and Chipman et al. (2006a) for details.

Appendix: MCMC Sampling Schemes

We use MCMC posterior simulation to implement inference in model (4). See, for example, Gamerman (1997) for a review of MCMC methods. In the following discussion we use $[U \mid \cdots]$ to indicate that the random variable U is updated conditional on the currently imputed values of all other parameters. The transition probability for the implemented MCMC is defined by the following steps.

Step 1. Updating Φ.

$$[\Phi \mid \cdots] \propto p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{V}, \Phi) \cdot p(\Phi).$$

The updating of Φ depends on the specific form of $p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{V}, \Phi)$, which in our example is either a linear regression model or an ordinal probit model. There are well established methods to update parameters in such models. For example, see Gelman et al. (2003) and Albert and Chib (1993).

Step 2. Updating f and σ^2.

$$[f, \sigma^2 \mid \cdots] \propto p(\boldsymbol{X} \mid \boldsymbol{W}, f, \boldsymbol{\theta}, \sigma^2) p(\boldsymbol{Y} \mid \boldsymbol{V}, f, \boldsymbol{\theta}, \sigma^2) p(f) p(\sigma^2). \tag{10}$$

If we define $x^*_{ij} = x_{ij} - \theta_i$ and $y^*_{ij} = y_{ij} - \theta_i$, then (10) is equivalent to

$$[f, \sigma^2 \mid \cdots] \propto \prod \Big\{ p(x^*_{ij} \mid w_{ij}, f, \sigma^2) \Big\} \prod \Big\{ p(y^*_{ij} \mid v_{ij}, f, \sigma^2) \Big\} p(f) p(\sigma^2), \tag{11}$$

with

$$\begin{aligned} p(x^*_{ij} \mid w_{ij}, f, \sigma^2) &= N(f(w_{ij}), \sigma^2), \\ p(y^*_{ij} \mid v_{ij}, f, \sigma^2) &= N(f(v_{ij}), \sigma^2). \end{aligned}$$

Note that (11) is exactly a BART model with x^*_{ij} and y^*_{ij} being the dependent variable, and the updating algorithm can be found in Chipman et al. (2006a) Section 4.

Step 3. Updating $\boldsymbol{\theta}$.

$$[\boldsymbol{\theta} \mid \cdots] \propto p(\boldsymbol{X} \mid \boldsymbol{W}, f, \boldsymbol{\theta}, \sigma^2) p(\boldsymbol{Y} \mid \boldsymbol{V}, f, \boldsymbol{\theta}, \sigma^2) p(\boldsymbol{\theta} \mid \rho, \delta^2).$$

Define $e_{ij} = x_{ij} - f(w_{ij})$ and $s_{ij} = y_{ij} - f(v_{ij})$, and use $\boldsymbol{e}$ and $\boldsymbol{s}$ to denote the collection of e_{ij} and s_{ij}, respectively. We find

$$\begin{aligned}[\boldsymbol{\theta} \mid \cdots] \quad &\propto \quad \exp\Big\{ - \frac{(\boldsymbol{e} - \boldsymbol{U}_x\boldsymbol{\theta})'(\boldsymbol{e} - \boldsymbol{U}_x\boldsymbol{\theta}) + (\boldsymbol{s} - \boldsymbol{U}_y\boldsymbol{\theta})'(\boldsymbol{s} - \boldsymbol{U}_y\boldsymbol{\theta})}{2\sigma^2} \Big\} \\ &\cdot \quad \exp\Big\{ - \frac{1}{2\delta^2} \boldsymbol{\theta}'(\boldsymbol{H} - \rho\boldsymbol{C})\boldsymbol{\theta} \Big\},\end{aligned}$$

where $\boldsymbol{U}_x$ and $\boldsymbol{U}_y$ are the design matrix of $\boldsymbol{\theta}$ corresponding to $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. We can show that $[\boldsymbol{\theta} \mid \cdots]$ is a normal distribution with variance $[(\boldsymbol{U}'_x\boldsymbol{U}_x + \boldsymbol{U}'_y\boldsymbol{U}_y)/\sigma^2 + (\boldsymbol{H} - \rho\boldsymbol{C})/\delta^2]^{-1}$ and mean $[(\boldsymbol{U}'_x\boldsymbol{U}_x + \boldsymbol{U}'_y\boldsymbol{U}_y)/\sigma^2 + (\boldsymbol{H} - \rho\boldsymbol{C})/\delta^2]^{-1}(\boldsymbol{U}'_x\boldsymbol{e} + \boldsymbol{U}'_y\boldsymbol{s})/\sigma^2$.

Step 4. Updating ρ and δ^2.

$$[\rho, \delta^2 \mid \cdots] \propto p(\boldsymbol{\theta} \mid \rho, \delta^2) p(\rho) p(\delta^2).$$

CAR is a widely used spatial model and the posterior sampling of ρ and δ^2 has been discussed extensively in literature. For example, see He and Sun (2000).

Step 5. Updating $\boldsymbol{Y}$. We update $\boldsymbol{Y}$ one element at a time, i.e.,

$$[y_{ij} \mid \cdots] \propto p(z_{ij} \mid y_{ij}, v_{ij}, \Phi) p(y_{ij} \mid v_{ij}, f, \boldsymbol{\theta}, \sigma^2).$$

Under model (6), a linear regression model, we have

$$[y_{ij} \mid \cdots] \propto \exp\Big\{ - \frac{1}{2\tau^2}(z_{ij} - h^*_{ij} - y_{ij}\beta_6)^2 \Big\} \exp\Big\{ - \frac{1}{2\sigma^2}(y_{ij} - f(v_{ij}) - \theta_i)^2 \Big\},$$

where $h^*_{ij} = \beta_0 + v_{ij4}\beta_1 + v_{ij5}\beta_2 + v_{ij6}\beta_3 + v_{ij7}\beta_4 + v_{ij8}\beta_5$. We can show that $[y_{ij} \mid \cdots]$ is normal with variance $(\beta_6^2/\tau^2 + 1/\sigma^2)^{-1}$ and mean

$$\Big(\frac{\beta_6^2}{\tau^2} + \frac{1}{\sigma^2}\Big)^{-1} \Big(\frac{1}{\tau^2}\beta_6(z_{ij} - h^*_{ij}) + \frac{1}{\sigma^2}(f(v_{ij}) + \theta_i) \Big).$$

Under model (8), an ordinal probit model, we have

$$[y_{ij} \mid \cdots] \propto \exp\Big\{ - \frac{1}{2\tau^2}(\eta_{ij} - h^{\triangle}_{ij} - y_{ij}\beta_4)^2 \Big\} \exp\Big\{ - \frac{1}{2\sigma^2}(y_{ij} - f(v_{ij}) - \theta_i)^2 \Big\},$$

where $h^{\triangle}_{ij} = \beta_0 + v_{ij1}\beta_1 + v_{ij2}\beta_2 + v_{ij3}\beta_3$. Thus $[y_{ij} \mid \cdots]$ is normal with variance $(\beta_4^2/\tau^2 + 1/\sigma^2)^{-1}$ and mean

$$\Big(\frac{\beta_4^2}{\tau^2} + \frac{1}{\sigma^2}\Big)^{-1} \Big(\frac{1}{\tau^2}\beta_6(z_{ij} - h^{\triangle}_{ij}) + \frac{1}{\sigma^2}(f(v_{ij}) + \theta_i) \Big).$$

References

Albert, J. H. and Chib, S. (1993). "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association*, 88: 669–679. 629

Besag, J., York, J., and Molli, A. (1991). "Bayesian image restoration, with two applications in spatial statistics." *Annals of the Institute of Statistical Mathematics*, 43: 1–20, (Disc: pp21–59). 616

Breiman, L. (1996). "Bagging predictors." *Machine Learning*, 24: 123–140. 615

— (2001). "Random Forests." *Machine Learning*, 45: 5–32. 615

Byrne, C., Nedelman, J., and Luke, R. (1994). "Race, socioeconomic status, and the development of end-stage renal disease." *American Journal of Kidney Diseases*, 23(1): 16–22. 612

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). "Bayesian CART Model Search." *Journal of the American Statistical Association*, 93: 935–948, (C/R: P948–960). 615

— (2006a). "BART: Bayesian Additive Regression Trees." Technical report, Department of Mathematics and Statistics, Acadia University, Canada. 612, 614, 615, 616, 618, 619, 623, 624, 629

Chipman, H. A., George, E. I., McCulloch, R. E., and Musio, M. (2006b). "Spatial BART." *Abstract at the Valencia/ISBA Eighth World Meeting on Bayesian Statistics, Valencia*. 613

Clayton, D. and Kaldor, J. (1987). "Empirical Bayes estimates of age-standardized relative risks for use in disease mapping." *Biometrics*, 43: 671–681. 616

Clyde, M. and Lee, H. (2001). "Bagging and Bayesian Bootstrap." In Richardson, T. and Jaakkola, T. (eds.), *Artificial Intelligence and Statistics 2001*, 169–174. 615

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. John Wiley and Sons. 616

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998). "A Bayesian CART Algorithm." *Biometrika*, 85: 363–377. 615

Devesa, S. and Diamond, E. (1983). "Socioeconomic and racial differences in lung cancer incidence." *American Journal of Epidemiology*, 118(6): 818–831. 612

Foster, D. P. and Stine, R. A. (2004). "Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy." *Journal of the American Statistical Association*, 99(466): 303–313. 624

Freund, Y. and Schapire, R. (1997). "A decision-theoretic generalization of online learning and an application to boosting." *Journal of Computer and System Sciences*, 55: 119–139. 615

Friedman, J. H. (1991). "Multivariate Adaptive Regression Splines." *The Annals of Statistics*, 19: 1–67, (Disc: P67–141). 618

— (2001). "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics*, 29(5): 1189–1232. 615, 629

Gamerman, D. (1997). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall Ltd. 629

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003). *Bayesian Data Analysis*. Chapman & Hall. 629

Geronimus, A. and Bound, J. (1998). "Use of census-based aggregate variables to proxy for socioeconomic group: evidence from national samples." *American Journal of Epidemiology*, 148(5): 475–486. 612

Gornick, M., Eggers, P., Reilly, T., Mentnech, R., Fitterman, L., Kucken, L., and Vladeck, B. (1996). "Effects of race and income on mortality and use of services among Medicare beneficiaries." *New England Journal of Medicine*, 335(11): 791–799. 612

He, Z. and Sun, D. (2000). "Hierarchical Bayes estimation of Hunting success rates with spatial correlations." *Biometrics*, 56(2): 360–367. 630

Horton, N. J. and Laird, N. M. (1999). "Maximum Likelihood Analysis of Generalized Linear Models with Missing Covariates." *Statistical Methods in Medical Research*, 8: 37–50. 611

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., and Herring, A. H. (2005). "Missing-data Methods for Generalized Linear Models: A Comparative Review." *Journal of the American Statistical Association*, 100(469): 332–346. 611

Johnson, V. E. and Albert, J. H. (1999). *Ordinal Data Modeling*. Springer-Verlag Inc. 623

Kraus, J., Fife, D., Cox, P., Ramstein, K., and Conroy, C. (1986). "Incidence, severity, and external causes of pediatric brain injury." *American Journal of Diseases of Children*, 140(7): 687–693. 612

Little, R. J. A. (1992). "Regression with Missing X's: A Review." *Journal of the American Statistical Association*, 87: 1227–1237. 611

Mandelblatt, J., Andrews, H., Kerner, J., Zauber, A., and Burnett, W. (1991). "Determinants of late stage diagnosis of breast and cervical cancer: the impact of age, race, social class, and hospital type." *American Journal of Public Health*, 81(5): 646–649. 612

Rubin, D. B. (1976). "Inference and Missing Data." *Biometrika*, 63: 581–590. 611

Schafer, J. L. and Graham, J. W. (2002). "Missing Data: Our View of the State of the Art." *Psychological Methods*, 7(2): 147–177. 611

Sun, D., Tsutakawa, R. K., and Speckman, P. L. (1999). "Posterior distribution of hierarchical models using CAR(1) distributions." *Biometrika*, 86: 341–350. 616

Whittle, P. (1954). "On stationary processes in the plane." *Biometrika*, 41: 434–449. 616