Bayesian Predictive Inference for a Study Variable Without Specifying a Link to the Covariates

by
Ashley Lockwood

A Dissertation
Submitted to the Faculty
of the
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements for the
Degree of Doctor of Philosophy
in
Statistics
May 2023

APPROVED:

Professor Balgobin Nandram, Advisor
Department of Mathematical Sciences
Worcester Polytechnic Institute

Professor Fangfang Wang
Department of Mathematical Sciences
Worcester Polytechnic Institute

Professor Andrea Arnold
Department of Mathematical Sciences
Worcester Polytechnic Institute

Professor Buddika Peiris
Department of Mathematical Sciences
Worcester Polytechnic Institute

Dr. Jai Won Choi
Statistical Consultant
Meho Inc., Maryland

Dr. Myron Katzoff
Former Senior Mathematical Statistician
National Center for Health Statistics

Abstract

We perform Bayesian predictive inference of a finite population mean for a study variable without specifying a link between the study variable and covariates, consequently overcoming some limitations of traditional regression analysis. Not specifying the relationship between the response variable and the covariates adds flexibility and robustness to our models and allows for more applications. For real applications, we take care of three effects (spatial, heterogeneity, and clustering) simultaneously. We have explored several multinomial-Dirichlet models with stick-breaking representation on the mean vector to address a polychotomous regression problem. We also present a continuous regression problem addressed by including a spatial component in our Bayesian hierarchical model. Finally, we demonstrate a solution to the binary predictive inference problem while also incorporating a clustering stick-breaking prior. We illustrate all the aforementioned models using an application with BMI data.

First, to avoid defining the relationship between the study variable and covariates, we use a hierarchical Bayesian multinomial-Dirichlet model with stick-breaking representation on the mean vector to make inference about the characteristics of a finite population. Using this type of model provides us with an approach to make inference about the study variable without having to estimate regression coefficients, unlike the logistic regression model, which is used as a baseline for comparison. The multinomial-Dirichlet model uses polychotomous data from a contingency table, instead of binary data used by logistic regression. Several versions of multinomial-Dirichlet model are explored: an unordered and ungrouped model, an ordered and grouped model, a pooled area model, and a model with survey weights included. All versions of this model use the same general setup, but each is edited slightly to incorporate new techniques. The unordered and ungrouped model shows a good performance compared to the logistic regression model, resulting in a tighter population prediction interval, and requiring less assumptions. We also show how survey weights can be included in our unordered and ungrouped model. The ordered and grouped model is introduced to reduce the number of parameters drawn in the Gibbs sampler. Here ordering refers to the cells of the multinomial table with adjustments to the Dirichlet prior. While the ordered and grouped multinomial-Dirichlet model does speed up computation by requiring less parameters to be drawn in the Gibbs sampler, this model depends almost entirely on the data resulting in a tighter prediction interval. Lastly, the pooled area model is used to introduce small area estimation techniques.

Second, while we avoid specifying the parametric relationship between the study variable and covariates, we illustrate the advantage of including a spatial component to better account for the covariates in our models to make Bayesian predictive inference. We treat each unique covariate combination as an individual stratum, then we use small area estimation techniques

to make inference about the finite population mean of the continuous response variable. The two spatial models used are the conditional autoregressive (CAR) and simple conditional autoregressive (SCAR) models. We include the spatial effects by creating the incidence matrix via the Mahalanobis distance between covariates. We also show how to incorporate survey weights into the spatial models when dealing with probability survey data. We compare the results of two non-spatial models including the Scott-Smith model and the Battese, Harter, and Fuller model to the spatial models. Our goal is to have neighboring strata yield similar predictions, and to increase the difference between strata that are not neighbors. Ultimately, using the spatial models shows less global pooling compared to the non-spatial models, which was the desired outcome.

Third, in order to gain robustness, we combine spatial, heterogeneity, and clustering components. After finding success with the CAR model, we use this spatial model joined with a clustering stick-breaking prior to gain more information from the covariates. The main advantage of including a stick-breaking component in our model is that the number of clusters of the strata is determined by the algorithm and is subject to change at every iteration of the blocked Gibbs sampler. This is unlike the spatial component that manually defines neighborhood relationships before the Gibbs sampler. Allowing the number of clusters of the strata to fluctuate gives the model and the data the opportunity to select more relevant clusters. We compare our spatial stick-breaking model to the Fay-Herriot model that contains covariates directly in the model. The results show that the spatial stick-breaking model outperforms the Fay-Herriot model in both accuracy and precision.

We address concerns with traditional regression analysis by providing multiple Bayesian hierarchical models that allow for inference to be made about a study variable including covariates and without the need for estimating regression coefficients. We have successful results that pave the way for further extensions of alternate regression models. The progression from the multinomial-Dirichlet model to the spatial CAR and SCAR models reduces the amount of global pooling seen in our predictions. Then by adding a clustering component via the stick-breaking prior in our binary spatial model, we are able to extract even more information from the covariates without directly including them in the model. Our models expand the scope of applications we can explore with minimal assumptions, when compared to traditional regression models.

Acknowledgements

First and foremost I wish to thank my esteemed advisor, Professor Balgobin Nandram, for his invaluable mentorship and continuous support during my PhD study. His extensive knowledge and plentiful experience have encouraged me throughout my academic research and daily life.

My gratitude extends to my Dissertation Committee members, Professor Fangfang Wang, Professor Andrea Arnold, Professor Buddika Peiris, Dr. Jai Won Choi, and Dr. Myron Katzoff, who generously contributed their feedback and expertise. I genuinely value the insights and guidance they provided to improve the quality of my work.

I am extremely grateful to the United States Department of Agriculture National Agricultural Statistics Service (NASS) for their generous support in funding my studies. My interaction with the Research and Development Division of NASS provided me with valuable practical experience throughout my Dissertation period of studies as a Research Assistant. Special thanks to Dr. Linda Young for providing financial support.

This endeavor would not have been possible without my classmates and all of the late nights and early mornings spent studying together. It is their friendships that have made my experience at WPI a wonderful time. I must thank the Department of Mathematical Sciences at WPI for support as a Teaching Assistant during the early part of my PhD studies.

Finally, I would like to express my deepest appreciation to my mom, my dad, and my brother for their unconditional support and encouragement. Their belief in me has kept my motivation and spirits high during this process.

Contents

Chapter 1

Introduction

We perform Bayesian predictive inference for a study variable without specifying a link between the study variable and covariates, consequently overcoming some limitations of traditional regression analysis. By not specifying the relationship between the response variable and the covariates, we improve the robustness of our models and allow for more applications. Traditional regression models make strong assumptions about the distribution of the response variable, and we remove the need for such assumptions in the models presented in this dissertation.

Several multinomial-Dirichlet models with stick-breaking representation on the mean vector are explored to address a polychotomous regression problem. The variations of the multinomial-Dirichlet model include an unordered and ungrouped model, an ordered and grouped model, a pooled area estimation model, and we show how to include survey weights in the unordered and ungrouped model. We compare the results from our multinomial-Dirichlet models to the well-known logistic regression model.

We address the continuous regression problem by including a spatial component in our Bayesian hierarchical model. The conditional autoregressive (CAR) and simple conditional autoregressive (SCAR) models are introduced to create a neighborhood relationship between strata with the intent of reducing the amount of global pooling in our predictions. The results from the spatial models are compared to the results from two non-spatial models, the Scott-Smith model and the Battese, Harter, and Fuller (BHF) model.

Finally, the CAR model is joined with a clustering stick-breaking prior to make inference about a finite population proportion. This model includes spatial, heterogeneous, and cluster components to further accommodate the covariates. The spatial relationships are defined by the covariates before sampling, and the clustering component is determined by the data within the algorithm. We compare our spatial stick-breaking model to the Fay-Herriot model that does contain covariates directly in the model.

The remainder of Chapter 1 introduces the models we implemented to avoid defining the relationship between the response variable and the covariates, as well as provides the tools and techniques used throughout the dissertation. Chapter 2 presents the use of stick-breaking weights within the multinomial-Dirichlet model, and the several variations of the multinomial-Dirichlet model. In Chapter 3 the spatial CAR and SCAR models are introduced to make predictive inference about a finite population mean. Chapter 4 contains our binary response model that includes spatial, heterogeneous, and cluster components. Chapter 5 provides a summary and discusses future work.

1.1 Avoiding Defining the Relationship Between the Study Variable and Covariates

When we make inference about the characteristics of a finite population, we do not assume a relationship between the covariates and the response such as the case in a regression model. We avoid making the strong assumptions of regression models, and therefore increase the number of situations our models can be applied to. Traditional regression models suggest that the response variable, y, is a function of the covariates, $\boldsymbol{x}$, and the regression coefficients, $\boldsymbol{\beta}$,

$$y_i = f\left(\boldsymbol{x_i}, \boldsymbol{\beta}\right) + e_i, \tag{1}$$

where e_i are the error terms. In selecting a link function f, the regression model is specifying the relationship between y and $\boldsymbol{x}$ and therefore, assuming the distribution of the response variable.

To allow for our models to be more robust, we do not specify the relationship between the response variable and the covariates. By not specifying this relationship, we also do not have to estimate the coefficients $\boldsymbol{\beta}$.

While we avoid specifying a functional relation between the study variable and the covariates, it is true that when such a relation holds, it will be more efficient than our approach. For example, when a parametric model holds, a nonparametric model will be less efficient if it is used instead. However, as we avoid defining this parametric relationship our models are more general and flexible.

1.2 Multinomial-Dirichlet Model

We first use a hierarchical Bayesian multinomial-Dirichlet model with stick-breaking weights in the prior that avoids defining this relationship between the study variable and the covariates. Then, we use this model to perform Bayesian predictive inference for the finite population mean. The general setup of the multinomial-Dirichlet model is,

$$\begin{gathered} \boldsymbol{n} \mid \boldsymbol{p} \sim \text{Multinomial}\left(T, \boldsymbol{p}\right), \\ \boldsymbol{p} \mid \boldsymbol{\mu}, \tau \sim \text{Dirichlet}\left(\boldsymbol{\mu}\tau\right), \end{gathered} \tag{2}$$

with reasonable priors for $\boldsymbol{\mu}$ and τ, where $n_i \in \{0, \ldots, T\}$ are the counts being modeled with probability p_i such that $\sum_i p_i = 1$ for $T = \sum_i n_i$ independent trials. The data used as input in this model come from a contingency table, and note that any contingency table can be listed as an array. After aggregating all possible variable combinations, we obtain a table containing the count of observations in each cell where each cell has its own unique set of variable values. The cell counts in the contingency table are then vectorized by rows to give us our input data, $\boldsymbol{n}$. We are interested in making finite population predictions for $\boldsymbol{P}$, then we have a predicted proportion for each cell in the table.

Several versions of multinomial-Dirichlet model are explored: an unordered and ungrouped model, the inclusion of survey weights in the unordered and ungrouped model, an ordered and grouped model, and a pooled area model. All versions of this model use the same general setup, but each is edited slightly to incorporate new techniques. The unordered and ungrouped model uses the cell counts of the contingency table in the order they are created, which means that neighboring cells are related. Since there are multiple variables being used to construct the table, then adjacent cells likely have at least one variable value in common. For example, vectorizing Table 1 by rows yields the $\boldsymbol{n}$ used as input for the unordered and ungrouped model. Also in the unordered and ungrouped model, $\boldsymbol{n}$ has dimension $c \times 1$, therefore the sampler contains c cells. We also show the inclusion of survey weights in the unordered and ungrouped model.

The ordered and grouped model is introduced to reduce the number of parameters drawn in the Gibbs sampler. The ordered grouped method first orders the cells in the original contingency table from least to greatest based on cell count. Then, we group cells together by summing sequential ordered cells until their total is greater than some threshold, thus reducing the number of cells. Finally, the ordered and grouped cells are used as input in the multinomial-Dirichlet model resulting in a quicker computation time compared to the unordered ungrouped method since there are fewer parameters to be drawn in the Gibbs sampler. Lastly, we extend the unordered ungrouped model to make inference about many small areas, instead of a single finite population, in the pooled area model (Rao and Molina 2015).

The logistic regression model is a traditional approach for modeling categorical response variables (Lukman, Abdullah, and Rachman 2021). Even if we did not specifically consider the logistic model, and instead considered some general link function to define the relationship between y and $\boldsymbol{x}$ such as:

$$y_i \mid \boldsymbol{\beta} \sim \text{Bernoulli}\left(F(\boldsymbol{x_i}'\boldsymbol{\beta})\right).$$

Then, using this general link function we still have to estimate some $\boldsymbol{\beta}$, which we avoid altogether in the multinomial-Dirichlet model. For this reason, we proceed using the logistic regression model as our baseline model for comparison.

There are other models that represent the relationship between study variables and covariates without explicitly defining the relationship, including the classification and regression trees (CART) and the Bayesian additive regression trees (BART) presented in Chipman, George, and McCulloch (1998, 2010) and Chipman (2010), respectively. The Bayesian CART models presented in Chipman et al. (1998) specify the conditional distribution of a variable y, given the predictor values $\boldsymbol{x}$ using a binary tree to partition the predictor space into subsets where the distribution of y is consecutively more homogeneous. In Chipman et al. (2010) the BART method presented is similar to the machine learning gradient tree boosting methods, since both methods sum the contribution of sequential weak learners. However, instead of multiplying each sequential tree by a small constant (the learning rate) as used in gradient tree boosting, the Bayesian approach is to use a prior to control the size and shape of the tree. While CART and BART are both models that do not assume a distribution of y, both models still require that one variable of interest, y, is selected to make inference. The remainder of the variables in the CART and BART models are used to determine the splitting rules in the decision trees. By nature of training classification and regression trees, it is possible that not all of the variables are used in the model. Since deciding what feature to split on while constructing the trees is based on information gain, then certain variables may not be included if they do not add value to the predictions. The multinomial-Dirichlet model does not require one variable of interest to be specified and ensures that all variables, including all combinations of variables, are included in the model. There are also less tuning parameters in the multinomial-Dirichlet model compared to the CART and BART models.

Another modeling approach that can be used with contingency table data input is the log-linear model. Log-linear models for contingency tables are generalized linear models (GLMs) that treat the cell counts as independent observations of a Poisson random component (Agresti 2012). Conditional on the sum of the cell counts, T, Poisson log-linear models become multinomial models for cell probabilities. Similar to our multinomial-Dirichlet model, the log-linear model does not differentiate between the response variable and the covariates, and instead treats all variables the same. However, the log-linear model assumes the log of all the variables are linearly related. In our multinomial-Dirichlet model, there is no linearity assumption. Additionally, the log-linear model does not consider all interaction terms between variables by default, where our multinomial-Dirichlet model does include all interaction terms as mentioned previously. While the log-linear model is a substitute for the logistic regression model, our multinomial-Dirichlet model requires less assumptions and does not require manual insertion of the interaction terms.

Various adaptations of the multinomial-Dirichlet model also exist in Nandram (1998) and Nandram, Kim, and Zhou (2019). Nandram (1998) discusses a three-stage hierarchical multinomial model and Nandram (2021) compares the multinomial-Dirichlet–Dirichlet model and the multinomial-Dirichlet model via two projection methods. Nandram, Kim, and Zhou

(2019) presents both two-stage and three-stage models that use techniques of small area estimation to borrow strength across areas. The novelty of the multinomial-Dirichlet model presented in this work is the utilization of stick-breaking weights in the Dirichlet prior, which does not currently exist in the literature. The use of stick-breaking weights reduces the restrictions needed in the model, and allows for independence between the prior parameters.

Alternate versions of the logistic regression model can be found in Tibshirani and Manning (2013) and Ding and Vishwanathan (2010). Tibshirani and Manning (2013) presents a robust extension of logistic regression that incorporates the possibility of mislabeling directly into the objective by using shift parameters. Ding and Vishwanathan (2010) generalize logistic regression to t-logistic regression by using the t-exponential family, creating a new algorithm that is more robust to label noise. There are many alternative Bayesian regression models that can be used instead of logistic regression, including the various models presented in Nandram and Choi (2010). The concern with any model with regression coefficients is the specification of the relationship between the response variable and the covariates, and we avoid these relationship assumptions altogether in the multinomial-Dirichlet model.

1.2.1 Comparison of Logistic Regression and Multinomial-Dirichlet Models

Our multinomial-Dirichlet model is comparable to logistic regression, however the multinomial-Dirichlet model has many advantages including:

a. The logistic model specifies a relationship between the response and the covariates; the multinomial-Dirichlet model does not define this relationship;
b. The logistic model assumes the response variable, y, has a logistic distribution; the multinomial-Dirichlet model does not assume any distribution on the variables;
c. In the logistic model we must choose one response variable, y; the multinomial-Dirichlet model makes inference about the proportions of all variables;
d. In the logistic model interaction terms must be manually included; in the multinomial-Dirichlet model interaction terms are naturally included by default;
e. In the logistic model the covariates are held fixed; in the multinomial-Dirichlet model they are not held fixed;
f. The logistic model analyzes binary data, while the multinomial-Dirichlet model analyzes polychotomous data.

The multinomial-Dirichlet model is more advanced and requires less assumptions. For (a), our goal is to avoid specifying the relationship between y and $\boldsymbol{x}$, and therefore avoid sampling $\boldsymbol{\beta}$. In the multinomial-Dirichlet model, we do not have to specify the relationship between y and $\boldsymbol{x}$, meaning we do not have to choose $\boldsymbol{\beta}$. In the logistic model we are using the Metropolis-Hastings algorithm to sample the $\boldsymbol{\beta}$ that define this relationship between y and $\boldsymbol{x}$.

For (b), in the multinomial-Dirichlet model, we do not have to declare that y is logistic, because it may not meet the requirements of this distribution - this is a strong assumption we make in the logistic regression model. By not assuming any distribution for y we increase the number of situations available for application using the multinomial-Dirichlet model.

For (c), an advantage of the multinomial-Dirichlet model is that we make inference about the proportions of all possible combinations of variables, so there is not strictly one variable of interest. This means we can use the same multinomial-Dirichlet model to make inference about any of the variables' presence in the population, as well as any combination of those variables in the population. This multinomial-Dirichlet model is an example of a saturated model, because of the fact that all combinations of the covariates are considered by default in this model. However, in the logistic model we have to select a single variable of interest (i.e. the response variable, y). In the BMI example shown in Section 2.3 the response variable is the obesity indicator with three independent variables being age, race, and sex. If we wanted to change the dependent variable in logistic regression we would have to refit the model accordingly.

For (d) in order for interaction terms to be included in the logistic regression model they must be manually inserted into the model and the number of parameters to be sampled increases. However, in the multinomial-Dirichlet model the interaction terms are included by default. Interaction terms are important to be considered, because a model without interactions assumes that the effect of each covariate on the response is independent of other covariates in the model. We do not want to assume independence between covariates, and we want to gain as much information from the covariates as possible. Using the variables from the BMI example in Section 2.3, a two-way interaction term would be age × race; similarly a three-way interaction is age × race × sex and so on. One regression coefficient, β, is predicted for each interaction term in the logistic regression model.

For (e), another advantage of the multinomial-Dirichlet model is that the covariates are not held fixed, and therefore, are not interpreted in the same restricted way as seen in the logistic model. In the logistic model the covariates are held fixed meaning we interpret each $\boldsymbol{\beta}$ value (excluding β_0) by saying: "an increase of one unit in $\boldsymbol{x_i}$ increases/decreases the odds of y by a factor of β_i assuming all other variables remain constant". The multinomial-Dirichlet model does not hold variables constant when fitting the model.

For (f), the multinomial-Dirichlet model allows count data to be considered instead of simply binary data when compared to the logistic model. Using polychotomous data instead of binary data allows for an increased number of potential applications.

1.3 Spatial Models CAR and SCAR

Next, we introduce two versions of spatial models, a conditional autoregressive (CAR) model and a simple conditional autoregressive (SCAR) model (Chung and Datta 2022). We accommodate the covariates by using the spatial model instead of a regression model. For the spatial models, we include the spatial effects by creating the incidence matrix (or adjacency matrix) via the Mahalanobis distance between the covariates for each stratum. We use these spatial models to create a neighborhood relationship between similar strata and allow for less global pooling to the overall sample mean. By enabling strata to have neighbors, we expect neighborhoods to pool together without remote pooling together. Using a spatial model versus a non-spatial model should provide posterior predictions with a larger variation between predicted stratum means.

The data used as input in the spatial models follows a similar idea to the contingency table used in the multinomial-Dirichlet model, except now we do have to select one response variable, y. We aggregate the data for the spatial models into a finite number of possible covariate combinations to use in our design matrix $\boldsymbol{X}$. The difference between $\boldsymbol{X}$ and the contingency table used in the multinomial-Dirichlet model, is that the contingency table is created using all variables, including the response variable, y, and now $\boldsymbol{X}$ is separate from y. This is due to the fact that in the multinomial-Dirichlet model we do not have to specify a particular response variable, but in the spatial models we do have to specify y. Therefore, each row of $\boldsymbol{X}$, denoted $\boldsymbol{x_i}'$ is a unique covariate combination that has responses recorded y_{ij}, $j = 1, \ldots, n_i$. The responses in these spatial models are continuous variables, previously the multinomial-Dirichlet model used polychotomous data.

We include the spatial effects by creating the symmetric incidence matrix, $\boldsymbol{W}$ of size $\ell \times \ell$, via the Mahalanobis distance between $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$, $i = 1, \ldots, \ell$, $j = 1, \ldots, \ell$, and $i \neq j$. The Mahalanobis distance is defined as:

$$d_{ij} = \sqrt{\left(\boldsymbol{x_i} - \boldsymbol{x_j}\right)' \boldsymbol{S}^{-1} \left(\boldsymbol{x_i} - \boldsymbol{x_j}\right)}, \tag{3}$$

where $\boldsymbol{S}$ is the covariance matrix of $\boldsymbol{X}$, and $d\left(\boldsymbol{x_i}, \boldsymbol{x_i}\right) = 0$. We define $\boldsymbol{W}$ by letting $w_{ij} = 1$ if $d_{ij} \leq d_0$ and $w_{ij} = 0$ if $d_{ij} > d_0$ with zeroes on the diagonal (i.e., $w_{ii} = 0$). Here d_0 is the value yielding the $\boldsymbol{W}$ matrix that maximizes Moran's I, which is defined as:

$$I = \frac{\ell}{w_{..}} \frac{\sum_i \sum_j w_{ij} \left(\bar{y}_i - \bar{y}\right) \left(\bar{y}_j - \bar{y}\right)}{\sum_i \left(\bar{y}_i - \bar{y}\right)^2}, \tag{4}$$

where $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ is the mean response for stratum i; $\bar{y} = \sum_{i=1}^{\ell} \bar{y}_i/\ell$ is the overall mean response; w_{ij} corresponds to the elements of $\boldsymbol{W}$; and $w_{..} = \sum_i \sum_j w_{ij}$. Since $\boldsymbol{W}$ is symmetric, this means if $\boldsymbol{x_i}$ is related to $\boldsymbol{x_j}$, then $\boldsymbol{x_j}$ is also related to $\boldsymbol{x_i}$ $i = 1, \ldots, \ell$, $j = 1, \ldots, \ell$, and $i \neq j$. The neighbor relationship goes both ways, and there are no one-sided neighbor relations. The goal of creating neighborhoods is to have neighbors borrow strength from each other to predict closer to the neighborhood mean rather than the global population mean.

The setup of our CAR model is,

$$\begin{gathered}
y_{ij} \mid \boldsymbol{\mu}, \sigma^2 \sim \text{Normal}\left(\mu_i, \sigma^2\right), \\
\boldsymbol{\mu} \mid \theta, \rho, \sigma^2, \gamma \sim \text{Normal}\left(\theta \mathbf{1}, \frac{\rho}{1-\rho}\sigma^2 \left(\boldsymbol{R} - \gamma \boldsymbol{W}\right)^{-1}\right), \\
\pi\left(\theta, \sigma^2, \rho\right) \propto \frac{1}{\sigma^2}, \\
\gamma \sim \text{Uniform}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_\ell}\right), \\
\frac{1}{\lambda_1} \leq \gamma \leq \frac{1}{\lambda_\ell}, \quad -\infty < \theta < \infty, \quad 0 < \rho < 1, \quad \sigma^2 > 0, \\
j = 1, \ldots, n_i \quad i = 1, \ldots, \ell,
\end{gathered} \tag{5}$$

where ℓ in this case represents the total number of possible covariate combinations, which are considered to be the individual stratum. Also λ_1 is the minimum eigenvalue of $\boldsymbol{R}^{-1}\boldsymbol{W}$ and λ_ℓ is the maximum eigenvalue of $\boldsymbol{R}^{-1}\boldsymbol{W}$, and since $\sum_{i=1}^{\ell} w_{ii} = 0$ this results in $\lambda_1 < 0 < \lambda_\ell$ (Chung and Datta 2022). Also $(\boldsymbol{R} - \gamma\boldsymbol{W})$ is guaranteed to be positive definite as long as γ is in the range $\frac{1}{\lambda_1} \leq \gamma \leq \frac{1}{\lambda_\ell}$. To obtain samples from the joint posterior density of this model, we integrate out $\boldsymbol{\mu}$, θ, and σ^2, and then we only need to draw γ and ρ using a griddy Gibbs sampler. The remaining parameters $\boldsymbol{\mu}$, θ, and σ^2 have standard form conditional posterior distributions and we can sample these parameters directly. The only small difference between the CAR model and the SCAR model, is the SCAR model replaces $\boldsymbol{R}$ with the identity matrix, $\boldsymbol{I}$. In making this replacement, the eigenvalues, λ, will also change, thus altering the range of γ.

There are many traditional regression models that make inference about a characteristic of a population, including logistic regression, general linear models, general multivariate normal models, and classification and regression tree (CART). See Lindley and Smith (1972), Ghosh et al. (1998), Albert and Chib (1993), Box and Tiao (1973), and Chipman, George, and McCulloch (1998) for detail about each model. While these models have been widely used throughout history, the strong distribution assumptions made to properly use these models limits the types of data and situations available for application.

There also exists other models without regression coefficients that answer a similar question, including Dirichlet processes, Polya urn scheme, and Bayesian additive regression trees

(BART). See Blackwell and MacQueen (1973), Antoniak (1974), Yin and Nandram (2020), Teh et al. (2006) and Chipman, George, and McCulloch (2010) for information about these alternative models. Dirichlet processes and Polya urn schemes are popular in Bayesian modeling, however these complicated computations can lead to poor mixing in the Markov chain Monte Carlo (MCMC) algorithm. BART is a newer approach, but this method violates traditional Bayesian logic by double use of the data. The data are used in the likelihood of the BART model, and then again in a data-informed prior for two hyperparameters (Hill, Linero, and Murray 2020). We improve the computation of models without regression coefficients while maintaining the coherence of the Bayesian paradigm.

1.4 Stick-breaking Prior Included in Spatial Model

Our final model includes spatial, heterogeneous, and cluster components used to make Bayesian predictive inference for a finite population proportion. We begin with a spatial model having a binary response variable. Next, we include a heterogeneity parameter in the binary spatial model to account for the differences between strata. Ultimately, we end with the model we are most interested in that contains spatial, heterogeneous, and cluster components by incorporating the stick-breaking prior. By considering each unique combination of the covariates in the population as an individual stratum, we avoid directly including the covariates in the model. Then we use small area estimation techniques to make inference about each subset of the population based on its underlying covariates. Finally, we can estimate the overall finite population proportion by pooling predictions of the strata together.

We use spatial modeling techniques via the conditional autoregressive (CAR) model (Chung and Datta 2022). We include the spatial effects by creating the incidence matrix (or adjacency matrix) via the Mahalanobis distance between the covariates for each stratum. By enabling strata to have neighbors, we have seen neighborhoods pool together without all strata pooling together.

We incorporate a clustering component in our model using the stick-breaking prior to cluster the strata (Ishwaran and James 2001). This stick-breaking prior is a finite approximation of the Dirichlet-process prior, which allows the data to determine the number of appropriate clusters. This stick-breaking prior is ideal for when we do not know the distinct number of clusters present in the data before sampling. The spatial relationships are defined by the covariates before sampling, and the clustering component is determined by the data in the algorithm. While strata in the same spatial neighborhood will gain strength from each other, the strata in the same cluster will share cluster-specific parameters.

All of our models begin with an adapted Fay-Herriot model (Rao and Molina 2015;

Nandram, Erciulescu, and Cruze 2019). The Bayesian Fay-Herriot model without covariates is,

$$\begin{aligned}
\hat{\theta}_i \mid \boldsymbol{\mu} &\overset{\text{ind}}{\sim} \text{Normal}\left(\mu_i, \hat{\sigma}_i^2\right), \quad i = 1, \ldots, \ell, \\
\mu_i \mid \theta, \delta^2 &\overset{\text{ind}}{\sim} \text{Normal}\left(\theta, \delta^2\right), \quad i = 1, \ldots, \ell, \\
\pi\left(\theta, \delta^2\right) &\propto \frac{1}{(1+\delta^2)^2},
\end{aligned} \tag{6}$$

where $\pi\left(\delta^2\right)$ is a proper shrinkage prior. We do not include covariates directly in our models to improve the robustness and allow for more applications. Typically, in this model it is assumed that $\hat{\sigma}_i^2$, $i = 1, \ldots, \ell$ are fixed. We construct our model by letting $G = \left\{\prod_{i=1}^{\ell} \left(\hat{\sigma}_i^2\right)^{n_i}\right\}^{1/n}$, the weighted geometric mean where n_i is the sample size of the i^{th} stratum and $n = \sum_{i=1}^{\ell} n_i$ is the total sample size. Also let $\kappa_i = \frac{G}{\hat{\sigma}_i^2}$, $i = 1, \ldots, \ell$, which are also fixed.

We adjust the standard Fay-Herriot model by,

$$\begin{aligned}
\hat{\theta}_i \mid \boldsymbol{\mu}, \sigma^2 &\overset{\text{ind}}{\sim} \text{Normal}\left(\mu_i, \frac{\sigma^2}{\kappa_i}\right), \quad i = 1, \ldots, \ell, \\
\mu_i \mid \theta, \sigma^2 &\overset{\text{ind}}{\sim} \text{Normal}\left(\theta, \sigma^2\right), \quad i = 1, \ldots, \ell, \\
\pi\left(\theta, \sigma^2\right) &\propto \frac{1}{\sigma^2}.
\end{aligned} \tag{7}$$

By writing the Fay-Herriot model in this way, we no longer require a proper prior for σ^2. Now that σ^2 is in the likelihood and therefore directly connected to the data, it is unlikely that σ^2 would suffer from impropriety. Having σ^2 connected to the data improves identifiability and also allows for σ^2 to have a known conditional posterior density following an inverse-gamma distribution. This general setup is the foundation of the spatial CAR model with the stick-breaking prior.

Our spatial CAR model with the stick-breaking prior is,

$$\begin{gathered}
\hat{\theta}_i \mid \boldsymbol{\mu}, \sigma^2 \sim \text{Normal}\left(\mu_i, \frac{\sigma^2}{\kappa_i}\right), \\
\boldsymbol{\mu} \mid \boldsymbol{\eta}, \sigma^2, \psi \sim \text{Normal}\left(\boldsymbol{\eta}, \sigma^2 \left(\boldsymbol{R} - \psi \boldsymbol{W}\right)^{-1}\right), \\
\eta_i \mid \boldsymbol{z}, \theta, \sigma^2 \sim \sum_{s=1}^{\ell} p_s \cdot \text{Normal}\left(z_s, \sigma^2\right), \\
z_s \mid \rho, \sigma^2 \sim \text{Normal}\left(\theta, \frac{\rho}{1-\rho}\sigma^2\right), \\
\pi\left(\theta, \sigma^2\right) \propto \frac{1}{\sigma^2}, \\
\psi \sim \text{Uniform}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_\ell}\right), \\
\frac{1}{\lambda_1} \leq \psi \leq \frac{1}{\lambda_\ell}, \quad -\infty < \theta < \infty, \quad \sigma^2 > 0, \\
i = 1, \ldots, \ell,
\end{gathered} \tag{8}$$

where ℓ in this case represents the total number of possible covariate combinations, which are considered to be the individual strata. Also, p_s, $s = 1, \ldots, \ell$ are the stick-breaking weights defined in Section 1.6.2.

The stick-breaking weights are used to determine the cluster each stratum belongs in. We use d_i, $i = 1 \ldots, \ell$, as classification variables to identify the z_s, $s = 1, \ldots, m$, associated with each η_i, $i = 1 \ldots, \ell$, where m is the number of unique clusters and $m \leq \ell$. Therefore, the d_i identify the cluster each η_i belongs in, and the η_i in each cluster will share the same z_s. The stick-breaking weights are a component of the probability used to sample each d_i, $i = 1, \ldots, \ell$, as seen in the second block of our algorithm. Once we have the d_i, $1 = 1, \ldots, \ell$, we can fit the model with the blocked Gibbs sampler (Ishwaran and James 2001). By utilizing the blocked Gibbs sampler it is straightforward to obtain the conditional posterior distributions used to obtain samples of this model's parameters.

1.5 Bayesian Predictive Inference with and without Survey Weights

For all models presented in this dissertation, we perform Bayesian predictive inference of the finite population mean both with and without survey weights included in the models. For illustration purposes we use a continuous response, y, in this section, however the logic is similar for other data types.

1.5.1 Bayesian Predictive Inference without Survey Weights

Our goal in all models is to make inference about the finite population mean, say $\bar{Y}_i = \sum\limits_{j=1}^{N_i} y_{ij}/N_i$, based on the observed values of $\boldsymbol{y_i}$. Denote the sampling fraction as $f_i = n_i/N_i$, where n_i represents the sample size and N_i represents the population size for a given area. Therefore, y_{ij}, $j = 1, \ldots, n_i$ represents the sampled portion of the population, and y_{ij}, $j = n_i + 1, \ldots, N_i$ represents the nonsampled portion of the population. Also, $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ represents the sample mean for area i. First, the model parameters, $\boldsymbol{\theta}$ and σ^2, are estimated from the sample model. Then, based on our samples of $\boldsymbol{\theta}$, σ^2, and the observed sample values of $\boldsymbol{y_i}$, we make inference for the finite population mean $\bar{Y}_i$, using the model:

$$\bar{Y}_i \mid \boldsymbol{y_i} \overset{\text{ind}}{\sim} \text{Normal}\left(f_i\bar{y}_i + (1-f_i)\theta_i, (1-f_i)\frac{\sigma^2}{N_i}\right). \tag{9}$$

We use the sample to obtain the model parameters, and then we use the sample mean as a portion of the population prediction. Therefore, we are predicting the nonsampled portion of the population and taking a weighted average of the sample mean and the predicted nonsample mean.

1.5.2 Bayesian Predictive Inference with Survey Weights (Surrogate Sampling Techniques)

When we do not include survey weights in the model, we are not adjusting the sample for any potential bias present. With the adjusted (and trimmed) weights included in the model, we must sample the entire population using surrogate sampling techniques (Nandram and Rao 2021). First, the model parameters are estimated from the sample model with the adjusted (and trimmed) weights included. Then, we use these parameters in the population prediction model to sample N values of the study variable, where N is the finite population size.

Here we use the original survey weights, denoted v_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots \ell$ to calculate the effective sample size and the adjusted weights a_{ij}. First, we calculate the effective sample size, $\hat{n}$:

$$\hat{n} = \frac{\left(\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v_{ij}\right)^2}{\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v_{ij}^2}. \tag{10}$$

The effective sample size, $\hat{n}$, illustrates how severely the variance is increased by the unequal weighting (Nandram and Rao 2021). Then, we calculate the adjusted weights, a_{ij}, which are used to eliminate the bias present in the original survey weights:

$$a_{ij} = \hat{n}\frac{v_{ij}}{\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v_{ij}}. \tag{11}$$

Here $\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v_{ij} = \hat{N}$ is the Horvitz-Thompson estimator of population size, and $\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} a_{ij} = \hat{n}$, is the effective sample size. These adjusted weights a_{ij} are able to be used in a model when the data do not have outliers present. However, in the BMI example in Chapter 3 our original data do have outliers. Therefore, we use Winsorization, which is an effective method to deal with outliers by trimming the survey weights (Yang, Nandram, and Choi 2023). Outliers here are defined as observations greater than $v_0 = Q_3 + 1.5\,(Q_3 - Q_1)$, where Q_1 is the first quartile and Q_3 is the third quartile. Let v^* denote weights after trimming,

$$v_{ij}^* = \begin{cases} v_0, & v_{ij} \geq v_0 \\ r v_{ij}, & v_{ij} < v_0 \end{cases}, \tag{12}$$

where r is a rescaling parameter such that $\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v_{ij}^* = \sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v_{ij} = \hat{N}$. Then we obtain the adjusted trimmed weights a_{ij}^*,

$$\begin{aligned} \hat{n}^* &= \frac{\left(\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v_{ij}^*\right)^2}{\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v_{ij}^{*2}}, \\ a_{ij}^* &= \hat{n}^*\frac{v_{ij}^*}{\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v_{ij}^*}. \end{aligned} \tag{13}$$

These adjusted trimmed weights a_{ij}^* can now be used in the sample model to obtain the model parameters, θ and σ^2.

Now, to make predictive inference about the finite population mean, $\bar{Y}_i = \sum\limits_{j=1}^{N_i} y_{ij}/N_i$, we obtain predictions by,

$$\overline{Y_i} \mid \boldsymbol{\mu}, \sigma^2 \sim \text{Normal}\left(\theta_i, \frac{\sigma^2}{v_{i\cdot}}\right), \quad i = 1, \ldots, \ell, \tag{14}$$

where $v_{i\cdot} = \sum\limits_{j=1}^{n_i} v_{ij}$ represents the sum of the original unadjusted (and untrimmed) survey weights for each area. We no longer need to include the sample means, $\bar{y}_i$, or the sampling fraction, f_i, in this population prediction, since now we are utilizing the survey weights from the probability sample we are able to use surrogate sampling techniques. The adjusted (and trimmed) survey weights, $\boldsymbol{a}$ ($\boldsymbol{a}^*$), are used to simulate an unbiased sample from the population. Then, when we make population predictions we use the unadjusted survey weights for prediction since these weights accurately represent the entire population.

1.6 Statistical Background

In this section, we discuss several statistical concepts that are used throughout the dissertation.

1.6.1 Types of data

Throughout this dissertation we fit models with different types of data, including binary, polychotomous, and continuous data. First, binary data is a categorical type of data that can only take two values, “success” or “failure”. Binary variables are modeled with a Bernoulli distribution, and in binary regression the probability of “success” is modeled. One of the most well-known binary regression models is logistic regression, which is used as a baseline for comparison to our binary models presented. In our BMI application, the binary response variable is considering the proportion of individuals in the population who are obese (i.e. BMI $\geq$ 30).

Polychotomous data is also a categorical type of data, however polychotomous variables can take more than two values. For the purpose of our models, polychotomous data represents the counts of observations in an area. We model the polychotomous variable with the multinomial distribution, which is a generalization of the binomial distribution. Polychotomous data can be used to make inference about proportions, and by using polychotomous we are able to make inference about the proportions of multiple variables simultaneously. When using binary data, we can only make inference about the proportion of the single binary response variable. In our BMI application, the polychotomous variable is also considering the proportion of individuals in the population who are obese (i.e. BMI $\geq$ 30), but is doing so by analyzing count data.

Continuous data is a quantitative type of data that can take any value on the real line. Continuous data can be modeled by many distributions, including the normal distribution, the Cauchy distribution, the student’s t distribution, and more. In our BMI application, we are predicting the finite population mean of continuous BMI with use of the normal distribution.

1.6.2 Stick-breaking Weights

Let $\mu_i \overset{\text{ind}}{\sim} \text{Beta}\,(\alpha, \beta)$, $i = 1, \dots, c{-}1$. The stick-breaking weights used in the multinomial-Dirichlet models, $\boldsymbol{w}$, are defined as:

$$\begin{aligned}
w_1 &= \mu_1, \\
w_j &= \mu_j \prod_{k<j} (1 - \mu_k) \quad j = 2, \dots, c-1, \\
w_c &= \prod_{j=1}^{c-1} (1 - \mu_j)\,.
\end{aligned} \tag{15}$$

It is important to note that the weight of the last cell does not introduce a new μ_c, instead w_c is a product of $(1 - \mu_1), (1 - \mu_2), \dots, (1 - \mu_{c-1})$. This shows for the c cells, we have μ_j, $j = 1, \dots, c-1$, and this stick-breaking definition ensures $\sum_{j=1}^{c} w_j = 1$. The stick-breaking representation on the mean vector allows for smoothness and a strongly mixing Gibbs sampler. The use of the stick-breaking weights in the prior allows for the $\mu_1, \dots, \mu_{c-1}$ to be independent and identically distributed, and by nature the weights sum to 1, so there is no need to manually input an additional restriction. In general the stick-breaking weights tend to 0, since there is less stick to break as each new weight is calculated. For this reason, when using the multinomial-Dirichlet model with stick-breaking weights we would not want c, the number of cells, to be too large or the later weights will be approximately equal to 0.

The stick-breaking weights are also used in a more traditional sense as part of the stick-breaking prior presented in Chapter 4. In Chapter 4, the stick-breaking weights are denoted p_s, $s = 1, \dots \ell$, and are used to define clusters in the model.

1.6.3 Small Area Estimation Techniques

Small area estimation techniques estimate parameters for small sub-populations when the sub-population of interest is included in a larger survey from the finite population (Rao and Molina 2015). Estimation can also be made for the entire finite population by aggregating the results from each of the small areas. A small area can refer to a small geographical area such as a county. Also, a small area may refer to a particular demographic group based off many characteristics including, but not limited to, age, gender, ethnicity, education level, or income. An area is considered “small” if the area’s sample size is not large enough to obtain estimates directly from the sample survey with accuracy and precision. Small area estimation allows for areas with few or zero observations in the sample to borrow strength from surrounding or similar areas in order to make inference. Rao and Molina (2015) provide a detailed review of models using small area estimation techniques.

1.6.4 Global Pooling

When using small area estimation techniques there are two categories of parameters, global and local parameters. Local parameters are unique to each small area, while global parameters are shared across all areas. Especially in the case where we anticipate the small areas to be heterogeneous, we ensure the local parameters adequately capture the differences between areas. Since we are using covariates to define the strata we perform small area estimation with, we make certain that we are limiting the amount of global pooling. Our models are curated to properly express both the area-specific random effects and the large random effects. Tang, Ghosh, Ha, and Sedransk (2018) and Jo, Nandram, and Kim (2021) present additional details about global-local shrinkage priors.

1.6.5 Dirichlet Process

A Dirichlet process is a probability distribution whose range is a set of probability distributions, formally introduced by Ferguson (1973). The Dirichlet process is specified by a base distribution, G_0, and a positive real concentration parameter, α. The Dirichlet process is used as a prior, such as:

$$\begin{aligned} \nu_i \mid G &\overset{\text{iid}}{\sim} G, \\ G \mid \alpha, G_0 &\sim \text{DP}(\alpha, G_0), \end{aligned} \tag{16}$$

$i = 1, \ldots, n$. Here G_0 is a continuous distribution, so the probability that any two samples are equal is zero, however, G is a discrete distribution. If we assume we observe these variables, ν_i in a specific order, then we can obtain ν_n given the previous $n-1$ observations by,

$$\nu_n \mid \nu_1, \ldots, \nu_{n-1} = \begin{cases} \nu_i & \text{with probability } \frac{1}{n-1+\alpha} \\ \nu_n \sim G_0 & \text{with probability } \frac{\alpha}{n-1+\alpha} \end{cases}. \tag{17}$$

As $\alpha \to \infty$ then more samples are obtained from G_0, thus making the realizations less discretized. On the other hand, when $\alpha \to 0$ there are more repeated values, which leads to a more discrete sample. We use this process as a prior when we want to cluster the data.

In Chapter 4, we use the stick-breaking prior as a finite approximation to the Dirichlet process prior. The main idea of the Dirichlet process prior remains true in the stick-breaking prior. However, the Dirichlet process prior is fully nonparametric as it allows for an infinite amount of parameters to be sampled. The stick-breaking prior we use has a finite number of parameters, and therefore yields a finite number of clusters.

1.6.6 Logistic Regression Model

The logistic regression model, which we use as a baseline for comparison throughout the dissertation, can be written:

$$y_i \mid \boldsymbol{\beta} \sim \text{Bernoulli}\left(\frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}\right), \qquad i = 1, \dots, n. \tag{18}$$

Using an improper prior, $\pi(\boldsymbol{\beta}) = 1$, for simplicity. Therefore, the conditional posterior density of $\boldsymbol{\beta}$ is:

$$\pi(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}) = \prod_{i=1}^{n}\left[\frac{e^{y_i\boldsymbol{x}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}\right], \qquad y_i = 0, 1, \tag{19}$$

where $\boldsymbol{y}$ represents the vector of binary responses, and $\boldsymbol{X}$ is the design matrix, and $\boldsymbol{x}_i'$ corresponds to the rows of $\boldsymbol{X}$. In the logistic model we are using the Metropolis-Hastings algorithm to sample the $\boldsymbol{\beta}$ that define this parametric relationship between $\boldsymbol{y}$ and $\boldsymbol{x}$. Technical details for obtaining a sample from the logistic regression model can be found in Appendix C.

Alternate versions of the logistic regression model can be found in Tibshirani and Manning (2013) and Ding (2010). Tibshirani and Manning presents a robust extension of logistic regression that incorporates the possibility of mislabeling directly into the objective by using shift parameters. Ding generalizes logistic regression to t-logistic regression by using the t-exponential family, creating a new algorithm that is more robust to label noise. There are many alternative parametric Bayesian models that can be used instead of logistic regression, including the various models presented in Nandram and Choi (2010). The concern with any model with regression coefficients is the specification of the relationship between the response variable and the covariates, and we avoid these relationship assumptions altogether.

1.6.7 Scott-Smith Model

We present a form of the Scott-Smith model (Scott and Smith 1969) as a model comparison for the spatial models with a continuous response variable presented in Chapter 3. The Scott-Smith model does not have a spatial component and also does not include the covariates directly in the model. The original version of the Scott-Smith model was used to model continuous data from two-stage cluster sampling, and there have since been many extensions to allow for a variety of sampling designs. See the references in Nandram, Toto, and Choi (2011) for an extension that accommodates binary data and shows other generalizations of this model.

The adapted Scott-Smith model we use can be written as:

$$\begin{gathered}
y_{ij} \mid \mu_i, \sigma^2 \sim \text{Normal}\left(\mu_i, \sigma^2\right), \\
\mu_i \mid \theta, \rho, \sigma^2 \sim \text{Normal}\left(\theta, \frac{\rho}{1-\rho}\sigma^2\right), \\
\pi\left(\theta, \sigma^2, \rho\right) \propto \frac{1}{\sigma^2}, \\
-\infty < \theta < \infty, \quad 0 < \rho < 1, \quad \sigma^2 > 0, \\
j = 1, \ldots, n_i, \quad i = 1, \ldots, \ell.
\end{gathered} \tag{20}$$

Technical details for obtaining a sample from the Scott-Smith model can be found in Appendix E.

1.6.8 BHF Model

We present a form of the Battese, Harter, and Fuller (BHF) model (Battese, Harter, and Fuller 1988) as a model comparison for the spatial models with a continuous response variable presented in Chapter 3. This model is a more general version of the Scott-Smith model, as the Scott-Smith model does not include the covariates. In general, we avoid defining this relationship between $\boldsymbol{y_i}$ and $\boldsymbol{x_i}$, but the BHF model is used as comparison to our models. The BHF model is:

$$\begin{gathered}
y_{ij} \mid \boldsymbol{\nu}, \boldsymbol{\beta}, \sigma^2 \overset{\text{ind}}{\sim} \text{Normal}\left(\boldsymbol{x_i}'\boldsymbol{\beta} + \nu_i, \sigma^2\right), \\
\boldsymbol{\nu} \mid \rho, \sigma^2 \overset{\text{ind}}{\sim} \text{Normal}\left(0, \frac{\rho}{1-\rho}\sigma^2\right), \\
\pi\left(\boldsymbol{\beta}, \sigma^2, \rho\right) \propto \frac{1}{\sigma^2}, \quad \sigma^2 > 0, \quad 0 < \rho < 1, \quad \boldsymbol{\beta} \in \mathbb{R}^p, \\
j = 1, \ldots, n_i, \quad i = 1, \ldots \ell.
\end{gathered} \tag{21}$$

This non-spatial model introduces covariates and includes the random effects for each stratum. Technical details for obtaining a sample from the BHF model can be found in Appendix F.

1.6.9 Fay-Herriot Model

The Fay-Herriot model is used as comparison to our binary response spatial models with the stick-breaking prior included (Nandram, Erciulescu, and Cruze 2019). This model does not contain a spatial or clustering component. The Fay-Herriot model includes the covariates directly in the model, unlike our models we present. Here $\hat{\theta}_i$ and $\hat{\sigma}_i^2$ are observed data and let $\boldsymbol{x_i}$ be the covariates. The Fay-Herriot model we fit is,

$$\begin{aligned}\hat{\theta}_i \mid \theta_i &\overset{ind}{\sim} \text{Normal}\left(\theta_i, \hat{\sigma}_i^2\right), \\ \theta_i \mid \boldsymbol{\beta}, \delta^2 &\overset{ind}{\sim} \text{Normal}\left(\boldsymbol{x_i}'\boldsymbol{\beta}, \delta^2\right), \\ \pi\left(\boldsymbol{\beta}, \delta^2\right) &= \frac{1}{(1+\delta^2)^2}, \quad i = 1, \ldots, \ell.\end{aligned} \tag{22}$$

Technical details for obtaining a sample from the Fay-Herriot model can be found in Appendix G.

1.6.10 Markov Chain Monte Carlo Methods

Throughout this dissertation, we use various Markov chain Monte Carlo (MCMC) methods to obtain samples from the models we discuss. Our most commonly used MCMC algorithm is the Gibbs sampler. In general, we are able to use the Gibbs sampler when all parameters are either bounded or have a standard posterior density. Aside from the standard Gibbs sampler, we also use the griddy Gibbs sampler and the blocked Gibbs sampler.

The griddy Gibbs sampler can be used when all parameters are contained within a bounded interval (Ritter and Tanner 1992). This allows us to use the grid method to sample every parameter in the Gibbs sampler. Therefore, the griddy Gibbs sampler is a special case of the Gibbs sampler where every parameter is sampled using the grid method.

The blocked Gibbs sampler is used in Bayesian semiparametric models with finite dimension (Ishwaran and James 2001). This sampler updates blocks of parameters at a time, where blocks are arranged such that the parameters can be sampled from simple multivariate distributions. Blocks of parameters are otherwise sampled in the same way as traditionally seen in the Gibbs sampler.

The only time we do not use some version of the Gibbs sampler is when we sample the logistic regression model. We must use the Metropolis-Hastings algorithm to obtain a sample from the logistic regression model, since $\boldsymbol{\beta}$ is unbounded and does not have a standard posterior density. The Metropolis-Hastings algorithm allows us to use a standard proposal density to obtain samples of our nonstandard target density.

In all of our applications, we carefully check the MCMC diagnostics to ensure a strongly mixing sampler. These diagnostics include the trace plots, auto-correlations, Geweke test of stationarity, and the effective sample sizes.

1.7 Novelty of Research

The novelty of the multinomial-Dirichlet model presented in Chapter 2 of this dissertation is the utilization of stick-breaking weights in the Dirichlet prior, which does not currently exist in the literature. The use of stick-breaking weights reduces the restrictions needed in the model, and allows for independence between the prior parameters. The definition and additional benefits of stick-breaking weights are given in Section 1.6.2.

The innovation in Chapters 3 and 4 begins with how we organize our data into strata based on the unique covariate combinations. Then we are able to use small area estimation techniques with these strata to make inference about the finite population. By structuring our data in this way, we are able to avoid directly including the covariates in our models while still gaining information from the variables. We extract information from the covariates by including spatial, heterogeneous, and cluster components. For the spatial component, we use the conditional autoregressive (CAR) model. Then we combine the CAR model with the clustering stick-breaking prior to further accommodate the covariates.

1.8 Dissertation Plan

Chapter 1 introduced the models we implement to avoid defining the relationship between the response variable and the covariates, as well as provided the tools and techniques used throughout the dissertation. For the remainder of the dissertation, Chapter 2 presents the use of stick-breaking weights within the multinomial-Dirichlet model, and variations of this multinomial-Dirichlet model. Several versions of multinomial-Dirichlet model are explored: an unordered and ungrouped model, an ordered and grouped model, a pooled area model, and a model with survey weights included. All versions of this model use the same general setup, but each is edited slightly to incorporate new techniques. See Lockwood and Nandram (2023a).

In Chapter 3, while continuing to avoid defining the parametric relationship between covariates and the response variable, we illustrate the advantage of including a spatial component our models to make Bayesian predictive inference about the finite population mean. The two spatial models used are the conditional autoregressive (CAR) and simple conditional autoregressive (SCAR) models. We include the spatial effects by creating the incidence matrix via the Mahalanobis distance between covariates. We also show how to incorporate survey weights into the spatial models when dealing with probability survey data. See Lockwood and Nandram (2023b).

In Chapter 4, we build a model that contains spatial, heterogeneous, and cluster components used to make Bayesian predictive inference for a finite population proportion. We begin by extending the spatial CAR model to incorporate a binary response variable. Next, we include a heterogeneity parameter in the binary spatial model to account for the differences between strata. Ultimately, we end with the model we are most interested in that contains spatial, heterogeneous, and cluster components by incorporating the stick-breaking prior. See Lockwood (2023).

Chapter 5 presents a summary of the dissertation and future work. For future work, we discuss expanding the spatial CAR model with a stick-breaking prior to handle a polychotomous response variable. The polychotomous data problem is a natural progression from the binary data model. In our BMI application, BMI categorization can have several levels including underweight, healthy, overweight, obese, and extremely obese. Rather than the simple binary case of labeling individuals as either obese or not obese. The polychotomous data problem accounts for extreme values at either end of the BMI spectrum to provide a better understanding of the population's overall health.

Chapter 2

Multinomial-Dirichlet Model with Stick-Breaking Weights in the Prior

2.1 Introduction

When making inference about the characteristics of a finite population, it may not be reasonable to assume a relationship between the covariates and the response. We propose a hierarchical Bayesian multinomial-Dirichlet model with stick-breaking weights in the prior that avoids defining this relationship.

We describe both the multinomial-Dirichlet and logistic model and show how we fit both models using a Markov chain Monte Carlo (MCMC) sampler. For the multinomial-Dirichlet model we use the Gibbs sampler, and for the logistic model we use the Metropolis–Hastings sampler. We must use the Metropolis-Hastings algorithm to obtain a sample from the logistic regression model, since $\boldsymbol{\beta}$ is unbounded and does not have a standard posterior density. We are able to use the Gibbs sampler when all parameters are either bounded or have a standard posterior density. After generating samples from these models, then each model will be used to make population predictions of the proportion of interest.

For the remainder of this chapter, we discuss the methodology of all techniques of the multinomial-Dirichlet model with stick-breaking weights in Section 2.2. Then an application using BMI data is given in Section 2.3, followed by a conclusion in Section 2.4. The appendices contain technical details including the description of the logistic regression model for comparison.

2.2 Methodology for the Multinomial-Dirichlet Model

In this section, we show four techniques of using the multinomial-Dirichlet model. In Section 2.2.1, we use the unordered and ungrouped data cells in the sampler. In Section 2.2.2, we discuss Bayesian predictive inference of a finite population proportion. In Section 2.2.3, we show the inclusion of survey weights in the unordered and ungrouped model. In Section 2.2.4, we order and group cells to reduce the number of parameters drawn in the Gibbs sampler. In Section 2.2.5, we introduce small area estimation techniques with the pooled area model. Appendix C contains the methodology of the logistic regression model we use for comparison.

Let the cell counts for the contingency table be n_i, $i = 1, \dots, c$, where c is the number of cells. Also let $T = \sum_{i=1}^{c} n_i$ represent the total number of observations.

2.2.1 Unordered and Ungrouped Model

The unordered and ungrouped multinomial-Dirichlet hierarchical model is:

$$
\begin{gathered}
\boldsymbol{n} \mid \boldsymbol{p} \sim \text{Multinomial}(T, \boldsymbol{p}), \\
\boldsymbol{p} \mid \boldsymbol{\mu}, \rho \sim \text{Dir}\left(\mu_1 \frac{1-\rho}{\rho}, \mu_2(1-\mu_1)\frac{1-\rho}{\rho}, \mu_3(1-\mu_2)(1-\mu_1)\frac{1-\rho}{\rho}, \ldots, \prod_{j=1}^{c-1}(1-\mu_j)\frac{1-\rho}{\rho}\right), \\
0 < \mu_1, \mu_2, \ldots, \mu_{c-1}, \rho < 1, \\
\mu_1, \mu_2, \ldots, \mu_{c-1} \overset{\text{ind}}{\sim} \text{Beta}\left(\phi \frac{1-\gamma}{\gamma}, (1-\phi)\frac{1-\gamma}{\gamma}\right), \\
0 < \phi, \gamma < 1.
\end{gathered}
\tag{23}
$$

Let the Dirichlet concentration parameters be denoted for simplicity:

$$
\alpha_i = w_i \frac{1-\rho}{\rho}, \qquad i = 1, \ldots, c, \tag{24}
$$

as the stick-breaking weights, $\boldsymbol{w}$, are defined as:

$$
\begin{aligned}
w_1 &= \mu_1, \\
w_j &= \mu_j \prod_{k<j} (1-\mu_k), \quad j = 2, \ldots, c-1, \\
w_c &= \prod_{j=1}^{c-1} (1-\mu_j).
\end{aligned}
\tag{25}
$$

This definition (25) is referred to as a stick-breaking procedure. At each iteration we randomly break the remainder of a stick of unit length and assign the length of this break to the current w_j value. It is important to note that the weight of the last cell does not introduce a new μ_c, instead w_c is a product of $(1-\mu_1), (1-\mu_2), \ldots, (1-\mu_{c-1})$. This shows for the c cells, we have μ_j, $j = 1, \ldots, c-1$, and this stick-breaking definition ensures $\sum_{j=1}^{c} w_j = 1$. The stick-breaking representation on the mean vector allows for smoothness and a strongly mixing Gibbs sampler (Ishwaran and James 2001). The use of the stick-breaking weights in the prior allows for the $\mu_1, \ldots, \mu_{c-1}$ to be independent and identically distributed, and by nature the weights sum to 1, so there is no need to manually input an additional restriction (Sethuraman 1994). The stick-breaking weights are conventionally used in the prior for Dirichlet processes. In general the stick-breaking weights tend to 0, since there is less stick to break as each new weight is calculated. For this reason, when using the multinomial-Dirichlet model with stick-breaking weights we would not want c, the number of cells, to be too large or the later weights will be approximately equal to 0. In our BMI example given in Section 2.3 we have $c = 56$, and found this number of cells to be reasonable.

For the remainder of the priors, we assume $\rho \sim \text{Uniform}(0, 1)$, and the ranges of ϕ and γ are constrained to ensure that the prior beta distribution parameters are greater than 1, this provides log-concavity of the prior. Details for these restrictions are given in Appendix A. Therefore, $\phi \sim \text{Uniform}(\frac{\gamma}{1-\gamma}, \frac{1-2\gamma}{1-\gamma})$ and $\gamma \sim \text{Uniform}(0, \frac{1}{3})$.

The joint posterior density of $\boldsymbol{\mu}, \boldsymbol{p}, \rho, \phi, \gamma$ is:

$$\begin{aligned}\pi\left(\boldsymbol{\mu}, \boldsymbol{p}, \rho, \phi, \gamma \mid \boldsymbol{n}\right) \propto & \frac{\Gamma\left(\frac{1-\rho}{\rho}\right)}{\prod_{i=1}^{c}\Gamma(\alpha_i)} \cdot \prod_{i=1}^{c} p_i^{\alpha_i+n_i-1} \cdot \left[\frac{\Gamma\left(\frac{1-\gamma}{\gamma}\right)}{\Gamma\left(\phi\frac{1-\gamma}{\gamma}\right)\Gamma\left((1-\phi)\frac{1-\gamma}{\gamma}\right)}\right]^{c-1} \\ & \times \left(\prod_{j=1}^{c-1}\mu_j\right)^{\left(\phi\frac{1-\gamma}{\gamma}-1\right)}\left(\prod_{j=1}^{c-1}(1-\mu_j)\right)^{\left((1-\phi)\frac{1-\gamma}{\gamma}-1\right)}\frac{1-\gamma}{1-3\gamma}.\end{aligned} \tag{26}$$

Therefore, after integrating out $\boldsymbol{p}$ we obtain:

$$\begin{aligned}\pi\left(\boldsymbol{\mu}, \rho, \phi, \gamma \mid \boldsymbol{n}\right) \propto & \frac{\Gamma\left(\frac{1-\rho}{\rho}\right)}{\prod_{i=1}^{c}\Gamma(\alpha_i)}\frac{\prod_{i=1}^{c}\Gamma(\alpha_i+n_i)}{\Gamma\left(\frac{1-\rho}{\rho}+T\right)} \\ & \times \left[\frac{\Gamma\left(\frac{1-\gamma}{\gamma}\right)}{\Gamma\left(\phi\frac{1-\gamma}{\gamma}\right)\Gamma\left((1-\phi)\frac{1-\gamma}{\gamma}\right)}\right]^{c-1}\left(\prod_{j=1}^{c-1}\mu_j\right)^{\left(\phi\frac{1-\gamma}{\gamma}-1\right)}\left(\prod_{j=1}^{c-1}(1-\mu_j)\right)^{\left((1-\phi)\frac{1-\gamma}{\gamma}-1\right)}\frac{1-\gamma}{1-3\gamma} \\ = & \frac{\prod_{i=1}^{c}\prod_{m=1}^{n_i}(\alpha_i+m-1)}{\prod_{k=1}^{T}\left(\frac{1-\rho}{\rho}+k-1\right)}\left[\frac{\Gamma\left(\frac{1-\gamma}{\gamma}\right)}{\Gamma\left(\phi\frac{1-\gamma}{\gamma}\right)\Gamma\left((1-\phi)\frac{1-\gamma}{\gamma}\right)}\right]^{c-1} \\ & \times \left(\prod_{j=1}^{c-1}\mu_j\right)^{\left(\phi\frac{1-\gamma}{\gamma}-1\right)}\left(\prod_{j=1}^{c-1}(1-\mu_j)\right)^{\left((1-\phi)\frac{1-\gamma}{\gamma}-1\right)}\frac{1-\gamma}{1-3\gamma}.\end{aligned} \tag{27}$$

2.2.2 Bayesian Predictive Inference of a Finite Population Proportion

Next, we use the griddy Gibbs sampler to obtain samples of $\boldsymbol{\mu}, \rho, \phi, \gamma$ of size M from the joint posterior density (27) (Ritter and Tanner 1992). In the BMI example illustrated in Section 2.3, we let $M = 1,000$. In this unordered and ungrouped version of the multinomial-Dirichlet model $\boldsymbol{\mu}$ has dimension $(c-1) \times 1$. Details about the griddy Gibbs sampler and the conditional posterior densities used can be found in Appendix D. We obtain Rao-Blackwellized density estimators of $\boldsymbol{p}$ after obtaining a sample of $(\boldsymbol{\mu}, \rho)$ from the Gibbs sampler. The conditional posterior density of $\boldsymbol{p} \mid \boldsymbol{n}, \boldsymbol{\mu}, \rho$, which we obtain a sample from to use for finite population predictions is:

$$\boldsymbol{p}^{(h)} \mid \boldsymbol{n}, \boldsymbol{\mu}^{(h)}, \rho^{(h)} \sim \text{Dir}\left(\alpha_1^{(h)} + n_1, \alpha_2^{(h)} + n_2, \alpha_3^{(h)} + n_3, \ldots, \alpha_c^{(h)} + n_c\right). \tag{28}$$

Therefore, we use $\boldsymbol{\mu}^{(h)}$ and $\rho^{(h)}$ in (28) to obtain $\boldsymbol{p}^{(h)}$, $h = 1, \ldots, M$. This conditional posterior density of $\boldsymbol{p}$ depends only on $\boldsymbol{\mu}$ and ρ, not on ϕ and γ. Since ϕ and γ are not directly connected to the data counts, $\boldsymbol{n}$, they are considered weakly identifiable parameters.

We make predictions for the desired characteristic of a finite population, call this p^*. We first discuss how to make finite population predictions of the study variable in the unordered and ungrouped model, then we show how to make predictions in the logistic regression model.

To begin making predictions in the unordered and ungrouped model, let $\boldsymbol{N} = (N_1, \ldots, N_c)'$ denote the population cell counts. Here, the population cell counts, N_i, $i = 1, \ldots, c$, are unknown and we estimate them using inverse probability weighting. The survey weights are denoted v_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, c$, therefore we calculate $N_i = \sum_{j=1}^{n_i} v_{ij}$ as estimates of the population cell counts. Also, let $N_{pop} = \sum\limits_{i=1}^{c} N_i$ denote the Horvitz-Thompson estimator of the finite population total, which is a design-based estimate since random sampling was used to obtain the survey weights. Therefore, we predict:

$$\boldsymbol{N}^{*(h)} \mid \boldsymbol{p}^{(h)}, \boldsymbol{n} \sim \text{Multinomial}\left(N_{pop}, \boldsymbol{p}^{(h)}\right), \tag{29}$$

where $\boldsymbol{p}^{(h)}$ are the sampled values we obtained in (28) after the griddy Gibbs sampler, and we use $\boldsymbol{p}^{(h)}$ in (29) to predict $\boldsymbol{N}^{*(h)}$, $h = 1, \ldots, M$. Finally, we end with a sample of $\boldsymbol{N}^*$ of size M.

To get an estimate for p^* we calculate the proportion of the values in $\boldsymbol{N}^*$ that correspond to cells containing the characteristic we are interested in making inference about. That is,

$$p^{*(h)} = \sum_{s \in S} \frac{N_s^{*(h)}}{N_{pop}}, \tag{30}$$

where S is the set of cells in the contingency table that contain the variable of interest. For each $\boldsymbol{N}^{*(h)}$ we obtain $p^{*(h)}$, $h = 1, \ldots, M$. In the BMI example given in Section 2.3, one of the variables we are interested in predicting is the obesity rate, so S is the set of cells corresponding to where the individuals are obese. In this obesity rate example, S would be all of the cells in columns 5-8 in Table 1, which contains 28 total cells.

Next, to make predictions in the logistic regression model we have a sample of $\boldsymbol{\beta}$ of size M from the Metropolis-Hastings sampler (see Appendix C for the technical details about sampling the logistic regression model). Since our data for the multinomial-Dirichlet model has been aggregated into a finite number of cells in a contingency table, then we use a similar data setup for logistic regression. We aggregate the data for the logistic model into a finite number of possible covariate combinations to use in our design matrix $\boldsymbol{X}$. The difference between $\boldsymbol{X}$ and the contingency table used previously, is that the table for the multinomial-Dirichlet model is created using all variables, including the response variable, $\boldsymbol{y}$. This is due to the fact that in the multinomial-Dirichlet model we do not have to specify a particular response variable, but in the logistic regression model we do have to specify a response variable. Therefore, if there are c cells in the contingency table for the multinomial-Dirichlet model, then after excluding the binary response, $\boldsymbol{y}$, from the list of covariates there are exactly $c/2$ rows in the logistic regression design matrix, $\boldsymbol{X}$. The rows of $\boldsymbol{X}$ correspond to unique covariate combinations.

Let $\boldsymbol{x_i}'$ denote the rows of $\boldsymbol{X}$, $i = 1, \ldots, c/2$. Then, we can calculate the probability of the desired characteristic occurring for every $\boldsymbol{x_i}'$ and every sampled value of $\boldsymbol{\beta}^{(h)}$, $h = 1, \ldots, M$, by:

$$p_i^{*(h)} = \frac{1}{1 + e^{-\boldsymbol{x_i}'\boldsymbol{\beta}^{(h)}}}. \tag{31}$$

Then let $\boldsymbol{N}^L = \left(N_1^L, \ldots, N_{c/2}^L\right)'$ denote the population count of each row of $\boldsymbol{X}$, $i = 1, \ldots, c/2$. Also, let $N_{pop} = \sum\limits_{i=1}^{c} N_i = \sum\limits_{i=1}^{c/2} N_i^L$ therefore the estimator of the finite population total is the same as used previously in the multinomial-Dirichlet model. Therefore, we then predict

$$N_i^{*(h)} \mid \boldsymbol{X}, \boldsymbol{\beta} \sim \text{Binomial}\left(N_i^L, p_i^{*(h)}\right), \tag{32}$$

$i = 1, \ldots, c/2$, $h = 1, \ldots, M$, which are the predicted finite population counts of the desired response for each covariate combination. Finally, we can calculate the predicted finite population proportion of the response variable by

$$p^{*(h)} = \sum_{i=1}^{c/2} \frac{N_i^{*(h)}}{N_{pop}}, \tag{33}$$

$h = 1, \ldots, M$, to end with a sample of p^* of size M, which we can use to interpret the finite population proportion of the response variable.

2.2.3 Survey Weights in Unordered and Ungrouped Model

Survey weights can be included in any of the multinomial-Dirichlet models, and here we illustrate the inclusion of survey weights in the unordered and ungrouped model described in Section 2.2.1. The same logic can be applied to include survey weights in any of the models discussed in this chapter. Here we use the original survey weights, denoted v_{ij}, $j = 1, \ldots, n_i$,

$i = 1, \dots, c$, to calculate the adjusted weights a_{ij} used in the models. Previously in Section 2.2.1, we used the unadjusted cell counts, $\boldsymbol{n}$, in our model. We simply counted to determine how many observations were categorized into each cell based on each observation's covariates. Now, each observation is assigned an adjusted weight a_{ij} and the adjusted cell counts are the sum of the adjusted weights for all of the observations that belong to a given cell.

We calculate $\hat{n}$, the effective sample size,

$$\hat{n} = \frac{\left(\sum\limits_{i=1}^{c} \sum\limits_{j=1}^{n_i} v_{ij}\right)^2}{\sum\limits_{i=1}^{c} \sum\limits_{j=1}^{n_i} v_{ij}^2}, \tag{34}$$

which illustrates how severely the variance is increased by the unequal weighting (Nandram and Rao 2021). Next, we calculate a_{ij}, the adjusted weights, that remove the bias present in the original survey weights:

$$a_{ij} = \hat{n} \frac{v_{ij}}{\sum\limits_{i=1}^{c} \sum\limits_{j=1}^{n_i} v_{ij}}, \tag{35}$$

where $\sum\limits_{i=1}^{c} \sum\limits_{j=1}^{n_i} v_{ij} = N$, the population size, and $\sum\limits_{i=1}^{c} \sum\limits_{j=1}^{n_i} a_{ij} = \hat{n}$, the effective sample size. These adjusted weights $a_{i\cdot}$ are used instead of the previous cell counts, n_i. It is important to note that we are not adjusting the number of cells, we are only adjusting the cell counts based on the survey weights. The model with the adjusted weights is:

$$\begin{gathered}
\boldsymbol{a} \mid \boldsymbol{p} \sim \text{Multinomial}\left(\sum_{i=1}^{c} a_{i\cdot}, \boldsymbol{p}\right), \\
\boldsymbol{p} \mid \boldsymbol{\mu}, \rho \sim \text{Dir}\left(\mu_1 \frac{1-\rho}{\rho}, \mu_2(1-\mu_1)\frac{1-\rho}{\rho}, \mu_3(1-\mu_2)(1-\mu_1)\frac{1-\rho}{\rho}, \dots, \prod_{j=1}^{c-1}(1-\mu_j)\frac{1-\rho}{\rho}\right), \\
0 < \mu_1, \mu_2, \dots, \mu_{c-1}, \rho < 1, \\
\mu_1, \mu_2, \dots, \mu_{c-1} \overset{\text{ind}}{\sim} \text{Beta}\left(\phi\frac{1-\gamma}{\gamma}, (1-\phi)\frac{1-\gamma}{\gamma}\right), \\
0 < \phi, \gamma < 1.
\end{gathered} \tag{36}$$

We use the same priors as Section 2.2.1 with $\rho \sim \text{Uniform}(0,1)$; $\phi \sim \text{Uniform}(\frac{\gamma}{1-\gamma}, \frac{1-2\gamma}{1-\gamma})$; and $\gamma \sim \text{Uniform}(0, \frac{1}{3})$. We also are able to draw samples of $\boldsymbol{\mu}, \rho, \phi, \gamma$ from this model using the same methods and similar conditional posterior densities as described in detail in Section 2.2.1. The only difference in the conditional posterior densities would be replacing $\boldsymbol{n}$ with $\boldsymbol{a}$. Similarly, the updated condition posterior density of $\boldsymbol{p} \mid \boldsymbol{a}, \boldsymbol{\mu}, \rho$, which we obtain a sample from to use for finite population predictions is:

$$\boldsymbol{p} \mid \boldsymbol{a}, \boldsymbol{\mu}, \rho \sim \text{Dir}\left(\alpha_1 + a_1, \alpha_2 + a_2, \alpha_3 + a_3, \ldots, \alpha_c + a_c\right). \tag{37}$$

We are now able to get Rao-Blackwellized estimators of $\boldsymbol{p}$ and make finite population predictions using the same methods described in Section 2.2.2. We examine the performance of this unordered and ungrouped model with the inclusion of survey weights in Section 2.3 with an application using BMI data.

2.2.4 Ordered and Grouped Model

For the ordered and grouped version of the multinomial-Dirichlet model we use the the same initial hierarchical Bayesian model as given in Section 2.2.1, except we reduce the total number of cells with μ's that need to be drawn in the Gibbs sampler. We make this reduction by first ordering the cells in the contingency table containing values $\boldsymbol{n}$ so that they range from least to greatest in value, namely $n^{(1)}, n^{(2)}, \ldots, n^{(c)}$ where $n^{(1)} \leq n^{(2)} \leq \cdots \leq n^{(c)}$. Next, we sum sequential ordered cells until their total is greater than some threshold, call this threshold h_0. This means we add $n^{(1)} + n^{(2)} + \cdots + n^{(j)}$ for some j until we first satisfy $\sum\limits_{i=1}^{j} n^{(i)} \geq h_0$. Once this inequality is satisfied we set $\sum\limits_{i=1}^{j} n^{(i)} = n_1^{og}$ (here the superscript "og" represents that these n's have been ordered and grouped, and the same logic applies to other parameters) and we continue grouping in the same way until we get a new set of $\boldsymbol{n}^{og}$. Now $\boldsymbol{n}^{og} = (n_1^{og}, n_2^{og}, \ldots, n_{c^*}^{og})$ where c^* is the reduced number of ordered and grouped cells.

The reasoning for reducing the number of cells is to also reduce the parameters needed to be drawn in the Gibbs sampler. In this version of the multinomial-Dirichlet model, $\boldsymbol{\mu}^{og}$ has dimension $(c^* - 1) \times 1$, and previously in Section 2.2.1 $\boldsymbol{\mu}$ had dimension $(c - 1) \times 1$ with $c \geq c^*$ (typically we would select h_0 such that c is strictly greater than c^*). Since the model presented in Section 2.2.1 is over-parameterized, we investigate a version of this model with less parameters in order to speed up computation and observe the effect on the finite population prediction.

The model used in the ordered and grouped model is similar to (23) with the exception of the reduced number of parameters:

$$
\begin{gathered}
\boldsymbol{n}^{og} \mid \boldsymbol{p}^{og} \sim \text{Multinomial}(T, \boldsymbol{p}^{og}),\\
\boldsymbol{p}^{og} \mid \boldsymbol{\mu}^{og}, \rho \sim \text{Dir}\left(\mu_1^{og}\frac{1-\rho}{\rho}, \mu_2^{og}(1-\mu_1^{og})\frac{1-\rho}{\rho}, \dots, \prod_{j=1}^{c^*-1}(1-\mu_j^{og})\frac{1-\rho}{\rho}\right),\\
0 < \mu_1^{og}, \mu_2^{og}, \dots, \mu_{c^*-1}^{og}, \rho < 1,\\
\mu_1^{og}, \mu_2^{og}, \dots, \mu_{c^*-1}^{og} \overset{\text{ind}}{\sim} \text{Beta}\left(\phi\frac{1-\gamma}{\gamma}, (1-\phi)\frac{1-\gamma}{\gamma}\right),\\
0 < \phi, \gamma < 1.
\end{gathered}
\tag{38}
$$

Within the Gibbs sampler the conditional posterior densities we are sampling from are the same as used in Section 2.2.1. However, we still need to make inference about the original unordered and ungrouped cells since the grouped cells were created regardless of their covariates. Therefore, we need to take the $\boldsymbol{\mu}^{og}$ with dimension $(c^*-1) \times 1$ and expand it to $\boldsymbol{\mu}$ with dimension $(c-1) \times 1$. We draw the $\boldsymbol{\mu}^{og}$ in the griddy Gibbs sampler exactly as we did in Section 2.2.1, then after the sampler we proceed by drawing:

$$
\left(\frac{\mu_{11}}{\mu_1^{og}}, \frac{\mu_{12}}{\mu_1^{og}}, \dots, \frac{\mu_{1j}}{\mu_1^{og}}\right) \sim \text{Dir}(1, 1, \dots, 1). \tag{39}
$$

Since we already have the values of $\boldsymbol{\mu}^{og}$ from the Gibbs sampler, now we can calculate $\mu_{11}, \mu_{12}, \dots, \mu_{1j}$ from the values of $\left(\frac{\mu_{11}}{\mu_1^{og}}, \frac{\mu_{12}}{\mu_1^{og}}, \dots, \frac{\mu_{1j}}{\mu_1^{og}}\right)$, thus effectively ungrouping the cells we previously combined together in μ_1^{og} to be $\mu_{11}, \mu_{12}, \dots, \mu_{1j}$. Here, j is the number of cells we previously summed together to obtain n_1^{og}. We continue in a similar method by sampling:

$$
\left(\frac{\mu_{21}}{\mu_2^{og}(1-\mu_1^{og})}, \frac{\mu_{22}}{\mu_2^{og}(1-\mu_1^{og})}, \dots, \frac{\mu_{2j}}{\mu_2^{og}(1-\mu_1^{og})}\right) \sim \text{Dir}(1, 1, \dots, 1). \tag{40}
$$

Then calculating $\mu_{21}, \mu_{22}, \dots, \mu_{2j}$. We continue this process until all of the grouped cells are ungrouped and we are back to having $\boldsymbol{p}$ with dimension $c \times 1$.

The conditional posterior density of $\boldsymbol{p} \mid \boldsymbol{n}, \boldsymbol{\mu}, \rho$ is:

$$
\boldsymbol{p} \mid \boldsymbol{n}, \boldsymbol{\mu}, \rho \sim \text{Dir}\left(\mu_{11}\frac{1-\rho}{\rho} + n_1, \mu_{12}\frac{1-\rho}{\rho} + n_2, \dots, \mu_{(c^*-1)j}\frac{1-\rho}{\rho} + n_c\right). \tag{41}
$$

Now we are able to get Rao-Blackwellized estimators of $\boldsymbol{p}$ and make finite population predictions using the same methods described in Section 2.2.2.

One drawback to the ordered and grouped approach is that we lose any relationship between neighboring cells when we order the cells from least to greatest based on their counts. Originally, cells located near each other in the contingency table are likely to have some covariates in common as seen in the example in Section 2.3 in Table 1. Results are shown in Section 2.3 where we illustrate the ordered and grouped model using the BMI data application.

2.2.5 Pooled Area Model

Using a model similar to the unordered and ungrouped model presented in Section 2.2.1, we extend the model here to cover multiple small areas simultaneously (Rao and Molina 2015). This model can be referred to as the pooled area model, since some parameters are shared between areas. The pooled area model is:

$$\begin{gathered}
\boldsymbol{n_i} \mid \boldsymbol{p_i} \sim \text{Multinomial}(n_{i\cdot}, \boldsymbol{p_i}), \\
\boldsymbol{p_i} \mid \boldsymbol{\mu}, \rho \sim \text{Dir}\left(\mu_1 \frac{1-\rho}{\rho}, \mu_2(1-\mu_1)\frac{1-\rho}{\rho}, \mu_3(1-\mu_2)(1-\mu_1)\frac{1-\rho}{\rho}, \ldots, \prod_{k=1}^{c-1}(1-\mu_k)\frac{1-\rho}{\rho}\right), \\
0 < \mu_1, \mu_2, \ldots, \mu_{c-1}, \rho < 1, \\
\mu_1, \mu_2, \ldots, \mu_{c-1} \overset{\text{ind}}{\sim} \text{Beta}\left(\phi\frac{1-\gamma}{\gamma}, (1-\phi)\frac{1-\gamma}{\gamma}\right), \\
0 < \phi, \gamma < 1,
\end{gathered} \tag{42}$$

for n_{ij} and p_{ij}, $i = 1, \ldots, \ell$, $j = 1, \ldots, c$, where ℓ represents the number of areas and c represents the number of cells in each contingency table (now we have ℓ contingency tables). Let $n_{i\cdot} = \sum_{j=1}^{c} n_{ij}$ represent the total number of observations for area i, and let the concentration parameters, $\boldsymbol{\alpha}$, be the same as defined in (24). We also use the same priors as seen in Section 2.2.1 including: $\rho \sim \text{Uniform}(0,1)$, $\phi \sim \text{Uniform}(\frac{\gamma}{1-\gamma}, \frac{1-2\gamma}{1-\gamma})$, and $\gamma \sim \text{Uniform}(0, \frac{1}{3})$. In this model, the parameters $\boldsymbol{\mu}, \rho, \phi$, and γ are shared between areas. However, each area now has its own $\boldsymbol{p_i}$ we obtain a sample of to use for small area population predictions. Although we increased the number of parameters we are sampling in this model, we are simultaneously increased the number of data points used in the model. By increasing the number of data points and by sharing the parameters $\boldsymbol{\mu}, \rho, \phi$, and γ between areas, we prevent over-parameterization in this model. The conditional posterior densities we sample from in the griddy Gibbs sampler to obtain samples of $\boldsymbol{\mu}, \rho, \phi, \gamma$ of size M in Section 2.2.1 are the same for this pooled area model, with the exception of now having ℓ number of areas and therefore using $\boldsymbol{n_i}$ instead of $\boldsymbol{n}$.

The conditional posterior density of $\boldsymbol{p_i} \mid \boldsymbol{n_i}, \boldsymbol{\mu}, \rho$ that we obtain a sample from is:

$$\boldsymbol{p_i} \mid \boldsymbol{n_i}, \boldsymbol{\mu}, \rho \sim \text{Dir}\left(\alpha_1 + n_{i1}, \alpha_2 + n_{i2}, \alpha_3 + n_{i3}, \ldots, \alpha_c + n_{ic}\right). \tag{43}$$

Similar to Section 2.2.1, we can obtain Rao-Blackwellized density estimators of $\boldsymbol{p_i}$ after obtaining a sample of $(\boldsymbol{\mu}, \rho)$ from their joint posterior density. This conditional posterior density of $\boldsymbol{p_i}$ depends only on $\boldsymbol{\mu}$ and ρ, not on ϕ and γ. We also make finite population predictions similar to previous methods, but instead we are predicting each of the ℓ small area's population. We make predictions for the desired characteristic of each small area's population, call this p_i^*. Let $\boldsymbol{N_i} = (N_{i1}, \ldots, N_{ic})'$, $i = 1 \ldots, \ell$, $j = 1, \ldots, c$, denote each small area's population cell counts that came from the survey weights. Also, let $N_{pop_i} = \sum_{j=1}^{c} N_{ij}$ denote the Horvitz-Thompson estimator of each small area's population total. Therefore, we predict:

$$\boldsymbol{N_i}^{*(h)} \mid \boldsymbol{p_i}^{(h)}, \boldsymbol{n_i} \sim \text{Multinomial}\left(N_{pop_i}, \boldsymbol{p_i}^{(h)}\right), \tag{44}$$

where $\boldsymbol{p_i}^{(h)}$ are the sampled values we obtained in (43) after the griddy Gibbs sampler. Therefore, we use $\boldsymbol{\mu}^{(h)}$ and $\rho^{(h)}$ in (43) to obtain $\boldsymbol{p_i}^{(h)}$, $h = 1, \ldots, M$. Then we use $\boldsymbol{p_i}^{(h)}$ in (44) to obtain $\boldsymbol{N_i}^{*(h)}$, $h = 1, \ldots, M$, $i = 1, \ldots, \ell$. Finally, we are left with samples of $\boldsymbol{N}_i^*$ of size M. To get an estimate for each $p_i{}^*$ we calculate the weighted average of the values in $\boldsymbol{N}_i^*$ that correspond to cells containing the characteristic we are interested in making inference about. For each $\boldsymbol{N}_i^{*(h)}$ we obtain $p_i^{*(h)}$, $h = 1, \ldots, M$, $i = 1, \ldots, \ell$.

2.3 Application using BMI Data

The proportion of the population determined to have obesity is an important indicator to study when interested in understanding the health of a finite population. We illustrate all variations of our multinomial-Dirichlet model using a probability sample of BMI data of 1,867 individuals within eight counties in California from the Third National Health and Nutrition Examination Survey (NHANES III). The survey weights sum to 12,232,099, therefore our sample accounts for 0.015% of the population.

These data have four covariates including age, race, sex, and an obesity indicator. We are interested in predicting the proportion of the finite population that has obesity. The factors of age, race, and sex have been shown to have an impact on an individual's BMI in numerous studies (e.g., Jackson et al. (2002), Department of Health and Human Services (2018), and the references within), so these variables are anticipated to improve our obesity prediction for the population. The values for sex are "male" or "female", and the values for race are "white" or "non-white". The obesity indicator is defined such that if an individual's BMI is greater than or equal to 30 kg/m^2 then the indicator's value is "obese", otherwise the indicator's value is "not obese". The age values range from 20 years old to 90 years old, and this variable is the only continuous variable included in the data. We bin the age variable by 10 years into groups of 20-29 years old, 30-39 years old, and so on, in order to obtain the counts for each possible covariate combination in a reasonable number of cells. By binning the age variable, we lessen the need to rely on exact accuracy of the data and allow for inference to be made about a broader age group (similar logic could be applied to any continuous variable).

After aggregating over all possible age, race, sex, and obesity combinations we have a table containing 56 cells each with its own unique set of covariate values. The contingency table for the BMI data used in the unordered and ungrouped application is below in Table 1. Vectorizing Table 1 by rows gives us $\boldsymbol{n}$ as described in (23). We vectorize by rows instead of columns intentionally, because we want neighboring consecutive cells to be in the same or similar age group. We expect similar age groups to be the most related in terms of their health, since health typically declines with age.

Table 1: Contingency Table for BMI Data

BMI:	Not Obese				Obese			
Sex:	Male		Female		Male		Female	
	White	Non-white	White	Non-white	White	Non-white	White	Non-white
Age:								
20-29 yrs	191	17	151	22	23	5	27	10
30-39 yrs	137	20	117	23	32	5	55	14
40-49 yrs	84	15	91	14	42	3	43	10
50-59 yrs	50	8	48	7	31	3	35	10
60-69 yrs	87	11	82	10	33	6	40	5
70-79 yrs	49	9	60	8	15	4	18	0
80-90 yrs	42	2	31	4	3	0	5	0

Note that while we are interested in measuring the proportion of obesity, the multinomial-Dirichlet model can be used to make inference about any of the covariates and their interactions without having to refit the model. This is an advantage seen in the multinomial-Dirichlet model that does not exist in logistic regression. We fit both the multinomial-Dirichlet model and the logistic regression model with the same BMI data. However, if we wanted to make inference about a different variable in the logistic model, then we would have to run the model again with the new study variable and covariates.

In this section we make comparisons between the results of various techniques using the multinomial-Dirichlet model and the logistic model, and the complete details for the logistic model are in Appendix C. Section 2.3.1 displays the results from applying the unordered and ungrouped model to the BMI data, both with and without survey weights included. Section 2.3.2 introduces the ordered and grouped model applied to the BMI data, then this model is compared to both the logistic model and the preceding unordered and ungrouped multinomial-Dirichlet model. Section 2.3.3 uses the small area estimation pooled area model to fit the BMI data. All methods are compared to the performance of the logistic regression model and to the performance of the other multinomial-Dirichlet models.

2.3.1 BMI Application using Unordered and Ungrouped Model

When applying the unordered and ungrouped model to the BMI data, there are 58 total parameters being drawn in the griddy Gibbs sampler, including $(\mu_1, \mu_2, \ldots, \mu_{55})$, ρ, ϕ, and γ. We assume $\rho \sim \text{Uniform}(0, 0.1)$ in all versions of our multinomial-Dirichlet model, since we searched for the location of the CPD of ρ beginning with $\rho \sim \text{Uniform}(0, 1)$ and found it exists within the range of $0 < \rho < 0.1$. Details for this search are shown in Appendix B. We ran 25,000 iterations of the sampler, then dropped the first 1,000 sampled values and chose every 24th sampled value to end with a final sample size of 1,000 for each parameter. The griddy Gibbs sampler shows good performance as evident by the trace plots, autocorrelations, Geweke test of stationarity, and the effective sample sizes. For ρ, ϕ, and γ the P-values for stationarity are 0.879, 0.283, and 0.391, respectively, meaning all three parameters pass the Geweke test - results for $(\mu_1, \mu_2, \ldots, \mu_{55})$ are similar. Additionally, for ρ, ϕ, and γ the effective samples sizes are 1000, 1142, and 1000, respectively, which is an accurate representation of our true sample size - results for $\boldsymbol{\mu}$ are similar.

After the Gibbs sampler, we obtain a sample of $\boldsymbol{p} \mid \boldsymbol{n}, \boldsymbol{\mu}, \rho$ of size $1,000$ as described in (28) by drawing samples from this known density directly. We then proceed in making the population cell count predictions, $\boldsymbol{N}^*$, given in (29) with the sample of $\boldsymbol{p}$. With the predicted population cell counts, $\boldsymbol{N}^*$, we now can obtain the proportion of each cell in the population. Finally, we sum the proportion of each cell that corresponds to any desired characteristic of a finite population, p^*. In this application we are mainly interested in the proportion of obesity in the population, and we also analyze p^* that corresponds to the proportion of females in the population and p^* that corresponds to the proportion of non-white individuals in the population. The results for these single variable proportion predictions can be found in Table 3 and are compared to the results from the logistic regression model, as well as the model including survey weights. In Table 4 we include the predicted proportions of two-, three-, and four-way interactions in the population for non-white males, obese white females, and not obese white females age 60-69, respectively. We include the results for predicting alternate variables other than the proportion of obesity to demonstrate the flexibility of the multinomial-Dirichlet model. Since we do not have to refit the model for each proportion of interest we are able to make inference about any variable or any variable combination from the same results of this fitted model. In the logistic regression model that we fit for comparison, we had to manually change the response variable and the corresponding design matrix each time we wanted to make inference about an alternate variable.

2.3.1.1 BMI Application including Survey Weights in Unordered and Ungrouped Model This application uses the same logic and predictions as seen in Section 2.3.1 in the unordered and ungrouped model, but the difference is we use the survey weights to adjust the cell counts instead of simply using the unadjusted cell counts $\boldsymbol{n}$ directly from the sample data. Since our BMI data is a probability sample from 8 counties in California, we are able to accurately adjust these cell counts as $\boldsymbol{a}$. Recall that we use use the original survey weights, denoted v_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, c$, to calculate the adjusted weights, a_{ij}, such that,

$$a_{ij} = \hat{n} \frac{v_{ij}}{\sum\limits_{i=1}^{c} \sum\limits_{j=1}^{n_i} v_{ij}}, \tag{45}$$

where $\sum\limits_{i=1}^{c} \sum\limits_{j=1}^{n_i} v_{ij} = N$, the population size, and $\sum\limits_{i=1}^{c} \sum\limits_{j=1}^{n_i} a_{ij} = \hat{n}$, the effective sample size. Also, we calculate the effective sample size, $\hat{n}$:

$$\hat{n} = \frac{\left(\sum\limits_{i=1}^{c} \sum\limits_{j=1}^{n_i} v_{ij} \right)^2}{\sum\limits_{i=1}^{c} \sum\limits_{j=1}^{n_i} v_{ij}^2}. \tag{46}$$

The table with the adjusted cell counts, $\boldsymbol{a}$, is shown below in Table 2.

Table 2: Survey Weight Adjusted Contingency Table for BMI Data

BMI:	Not Obese				Obese			
Sex:	Male		Female		Male		Female	
	White	Non-white	White	Non-white	White	Non-white	White	Non-white
Age:								
20-29 years	56	3	33	4	2	0	2	1
30-39 years	40	3	42	3	8	0	12	2
40-49 years	27	2	37	2	7	0	13	1
50-59 years	25	1	15	1	12	0	13	1
60-69 years	24	1	22	1	6	0	7	0
70-79 years	10	0	18	1	3	0	3	0
80-90 years	4	0	5	0	0	0	1	0

When using the survey weight adjusted cell counts in the unordered and ungrouped model on the BMI data, there are 58 total parameters being drawn in the griddy Gibbs sampler, including $(\mu_1, \mu_2, \ldots, \mu_{55})$, ρ, ϕ, and γ. The number of parameters does not change when we include the adjusted cell counts compared to the unordered and ungrouped model with the unadjusted cell counts. We ran 25,000 iterations of the sampler, then dropped the first 1,000 sampled values and chose every 24th sampled value to end with a final sample size of 1,000 for each parameter. The griddy Gibbs sampler shows good performance as evident by the trace plots, auto-correlations, Geweke test of stationarity, and the effective sample sizes. For ρ, ϕ, and γ the P-values for stationarity are 0.478, 0.485, and 0.379, respectively, meaning all three parameters pass the Geweke test - results for $(\mu_1, \mu_2, \ldots, \mu_{55})$ are similar. Additionally, for ρ, ϕ, and γ the effective samples sizes are 1000, 1129, and 1000, respectively, which is an accurate representation of our true sample size - results for $\boldsymbol{\mu}$ are similar.

After the Gibbs sampler, we obtain a sample of $\boldsymbol{p} \mid \boldsymbol{n}, \boldsymbol{\mu}, \rho$ of size $1,000$ as described in (37) by drawing samples from this known density directly. We then are able to make finite population proportion predictions using the same method as shown in the unordered and ungrouped model without adjusted cell counts. In this application we compare the proportion of obesity, the proportion of females, and the proportion of non-white individuals in the population for the unordered and ungrouped models with and without survey weights. The results for the comparison of these single variable proportion predictions can be found in Table 3. In Table 4 we also include the comparison of proportions for two-, three-, and four-way interactions in the population for non-white males, obese white females, and not obese white females age 60-69, respectively, for both models.

2.3.1.2 BMI Application Results with and without Survey Weights in Unordered and Ungrouped Model In Table 3 the row group name corresponds to the characteristic of interest (obesity, female, or non-white) that represents the p^* described previously. As seen in Table 3, our unordered and ungrouped multinomial-Dirichlet model predicts a slightly higher proportion of p^* compared to logistic regression for all obesity, female, and non-white population proportion predictions. The posterior standard error of p^* in the multinomial-Dirichlet model is roughly equal to the posterior standard error found in the logistic model. Therefore, performance is similar between the multinomial-Dirichlet model and the logistic regression model, while the multinomial-Dirichlet model requires less assumptions. Also, the multinomial-Dirichlet model only needed to be run one time to get the results for all three p^*, where the logistic regression model needed to be rerun for each change in the response variable - therefore needing to be adjusted and ran three times to get the results in Table 3. Now, looking at the survey weights being included in the model, we see the model with weights included predicts a lower proportion of obesity and non-white population predictions compared to the model without weights. However, the model with weights does predict a slightly higher proportion of females compared to the model without weights. The posterior standard errors for all predicted proportions in the model with weights are larger than the posterior standard errors found in the model without weights. This is because of the fact that the adjusted cell counts yield a smaller sample size, as seen in Table 2, which naturally results in a higher posterior standard error. Therefore, the prediction intervals are going to be wider for the model including the survey weight adjusted cell counts.

Table 3: Unordered and Ungrouped Model Results Using BMI Data

	Predicted p^*	SE of p^*	95% HPDI of p^*
Female			
M-D model	0.503	0.011	(0.483, 0.526)
M-D model with weights	0.504	0.023	(0.465, 0.554)
Logistic Model	0.497	0.012	(0.475, 0.522)
Non-white			
M-D model	0.139	0.008	(0.121, 0.154)
M-D model with weights	0.080	0.013	(0.057, 0.108)
Logistic Model	0.129	0.008	(0.114, 0.144)
Obesity			
M-D model	0.259	0.010	(0.239, 0.280)
M-D model with weights	0.209	0.018	(0.177, 0.244)
Logistic Model	0.252	0.010	(0.233, 0.273)

Similarly, since it is simple to obtain predictions for other covariates and their interactions in the multinomial-Dirichlet model, Table 4 below shows the results for various two-, three-, and four-way interactions using the BMI data with and without survey weights. Predicting interactions with the logistic regression model is not as straight-forward and is not included in the table. Being able to predict multiple covariate proportions and their interactions from the same multinomial-Dirichlet model results discussed in Table 3 is a big advantage for this model.

Table 4: Predicting Interactions in BMI Data Using Unordered and Ungrouped Model

	M-D model			M-D model with weights		
	Pred. p^*	SE p^*	95% HPDI p^*	Pred. p^*	SE p^*	95% HPDI p^*
Obese	0.259	0.010	(0.239, 0.280)	0.209	0.018	(0.177, 0.244)
Non-white males	0.062	0.006	(0.051, 0.073)	0.032	0.008	(0.016, 0.049)
Obese white females	0.119	0.007	(0.105, 0.132)	0.107	0.014	(0.082, 0.136)
Not obese white females age 60-69	0.043	0.005	(0.035, 0.052)	0.042	0.009	(0.025, 0.058)

In Table 4, the various interaction predictions were all made using the same finite population prediction technique described in Section 2.2.2. This means we only need to run the Gibbs sampler and predict $\boldsymbol{N}^*$ one time for each model (with and without survey weights) to make population predictions for any combination of cells presented in the sample data. As the number of the desired cells increases, we see that the posterior standard error also increases. In the last row of Table 4, predicting “not obese white females age 60-69” corresponds to making a prediction for a single cell. Also in Table 4, we see that for all four interaction predictions, the unordered and ungrouped model with the survey weights included predicts a lower population proportion of that interaction compared to the model without survey weights. Also the model with the survey weights included yields a larger posterior standard error in all four predictions, compared to the model without the adjusted cell counts. Therefore, all of the prediction intervals for the model with the survey weights included are wider compared to the model without. By using either model, it is straight forward to obtain predictions for any covariate combination regardless of the adjusted cell counts.

2.3.2 BMI Application using Ordered and Grouped Model

For the ordered and grouped multinomial-Dirichlet model, we set the threshold described in Section 2.2.4 such that $h_0 = 10$. We order the cells from least to greatest then sum sequential cells until their total is greater than or equal to the threshold, h_0. In doing so we collapse the number of μ's that need to be sampled in the Gibbs sampler from 55 to 43 μ's. Therefore, there are now 46 total parameters being drawn in the griddy Gibbs sampler, including $(\mu_1, \mu_2, \ldots, \mu_{43})$, ρ, ϕ, and γ. Previously, in the unordered and ungrouped method we had 58 parameters sampled via the Gibbs sampler, by decreasing the number of parameters drawn in the Gibbs sampler to 46 we decrease computation time while maintaining good results. As seen in Section 2.3.1, we ran 25,000 iterations of the sampler, then dropped the first 1,000 sampled values and chose every 24th sampled value to end with a final sample size of 1,000 for each parameter. The griddy Gibbs sampler shows good performance as evident by the trace plots, auto-correlations, Geweke test of stationarity, and the effective sample sizes. For ρ, ϕ, and γ the P-values for stationarity are 0.580, 0.928, 0.135, respectively, meaning all three parameters pass the Geweke test - results for $(\mu_1, \mu_2, \ldots, \mu_{43})$ are similar. Additionally, for ρ, ϕ, and γ the effective samples sizes are 1000, 893, and 1185, respectively, which is a mostly accurate representation of our true sample size aside from ϕ being slightly less than the sample size of 1000. Effective sample size results for $\boldsymbol{\mu}$ are similar, with their effective sample sizes ranging from 884 to 1486.

Then we ungrouped the μ's drawn in the Gibbs sampler to get Rao-Blackwellized estimates for all 56 original p's as described in Section 2.2.4. Once we have estimates for the 56 original p's, we use the same methods described previously in Section 2.2.2 to achieve finite population predictions for the proportion of obesity. The results in Table 5 compare the results from estimating the finite population obesity proportion in the unordered and ungrouped multinomial-Dirichlet model from Section 2.2.1 (labeled M-D model v1), the ordered and grouped multinomial-Dirichlet model from Section 2.2.4 (labeled M-D model v2), and the logistic regression model.

Table 5: Comparing Versions of Multinomial-Dirchlet Model to Logistic Model

	Predicted p^*	SE of p^*	95% HPDI of p^*	Expected ρ
M-D model v1	0.259	0.01	(0.239, 0.280)	0.003
M-D model v2	0.257	0.01	(0.237, 0.277)	0.023
Logisitic model	0.252	0.01	(0.233, 0.273)	N/A

The ordered and grouped multinomial-Dirichlet model has a slightly smaller prediction for the proportion of obesity in the population, p^*, when compared to the unordered and ungrouped model, but the prediction is still greater than the logistic regression model prediction. The posterior standard errors for all three models are very similar. We include the values of ρ in Table 5 because, using the ordered and grouped method appears to increase the value of ρ. In the previous multinomial-Dirichlet model $\rho = 0.003$, which is much smaller compared to $\rho = 0.023$ in the ordered and grouped model. As ρ increases, $\frac{1-\rho}{\rho}$ as seen in $\boldsymbol{\alpha}$ in (24) decreases, which means the posterior predictions for $\boldsymbol{p}$ made using (28) and (41) rely on the data, $\boldsymbol{n}$, more than the stick-breaking weights drawn in the Gibbs sampler. This results in the ordered and grouped model predicting the proportion of obesity to be 0.257, which is very close to the sample proportion of obesity at 0.255. A smaller value of ρ allows for the stick-breaking weights to have more of an impact on prediction of the population parameter. Figure 1 below illustrates the multinomial-Dirichlet model has a slightly larger expected value of obese p^* with slightly larger variation compared to the logistic regression model. The larger variation in the multinomial-Dirichlet model can be seen in the heavier tails and lower peak compared to the logistic density. The shapes of the two densities from the multinomial-Dirichlet models are similar, but the unordered and ungrouped model appears to be farther away from the sample obesity proportion and therefore is allowing the stick-breaking weights to have more influence on the population prediction. In general, we do not want the population predictions to be based solely on the sample data as we do want the model parameters to have influence in prediction. The red dashed vertical line represents the sample proportion of obesity at 0.255.

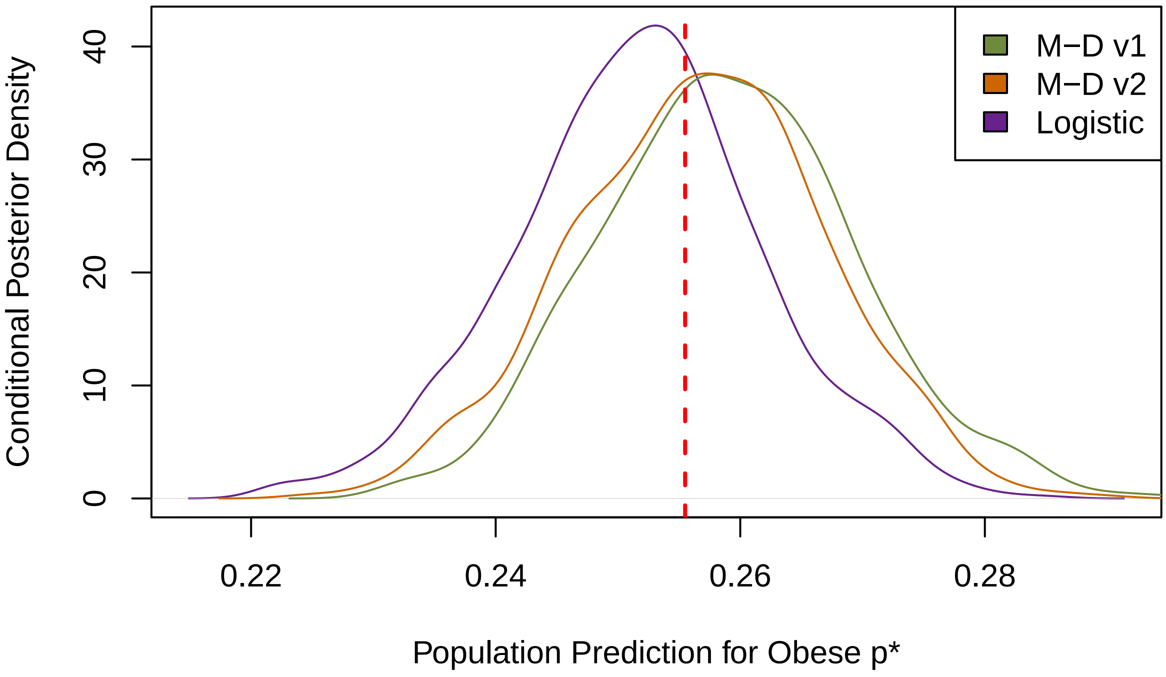

Figure 1: Comparing M-D v1, M-D v2, and Logistic Predictions

2.3.3 BMI Application using Pooled Area Model

The BMI data consist of 8 counties in California, therefore naturally when using the pooled area model we treat each county as an individual area. We predict the county population proportion of 56 cells in each of the 8 areas. When applying the pooled area model to the BMI data, there are 58 total parameters being drawn in the griddy Gibbs sampler, including $(\mu_1, \mu_2, \ldots, \mu_{55})$, ρ, ϕ, and γ - this is the same number of parameters drawn in the Gibbs sampler for the unordered and ungrouped model. These 58 parameters are shared between the 8 areas. We ran 25,000 iterations of the sampler, then dropped the first 1,000 sampled values and chose every 24th sampled value to end with a final sample size of 1,000 for each parameter. The griddy Gibbs sampler shows good performance as evident by the trace plots, auto-correlation plots, Geweke test of stationarity, and the effective sample sizes. For ρ, ϕ, and γ the P-values for stationarity are 0.474, 0.275, and 0.608, respectively, meaning all three parameters pass the Geweke test - results for $(\mu_1, \mu_2, \ldots, \mu_{55})$ are similar in passing the Geweke test. Additionally, for ρ, ϕ, and γ the effective samples sizes are 809, 1323, and 1000, respectively, which is a mostly accurate representation of our true sample size - results for $\boldsymbol{\mu}$ are similar.

After the Gibbs sampler, we obtain samples of $\boldsymbol{p_i} \mid \boldsymbol{n}, \boldsymbol{\mu}, \rho$ of size $1,000$, $i = 1, \ldots, 8$, as described in (43) by drawing samples from this known density directly. Each p_i is unique to the corresponding county, as these parameters are not shared between areas. We then proceed in making the county population cell count predictions, $\boldsymbol{N}_i^*$, given in (44) with the samples of $\boldsymbol{p}_i$, $i = 1, \ldots, 8$. With the predicted county population cell counts, $\boldsymbol{N}_i^*$, we now can obtain the proportion of each cell in the county population. Finally, we sum the proportion of each cell that corresponds to any desired characteristic of a county population, p_i^*. In this application we are interested in the proportion of obesity in each county's population. The pooled area model results for the predicted proportion of obesity in each county's population can be found in Table 6.

Table 6: Pooled Area Model Results Using BMI Data

	Predicted p^*	SE of p^*	95% HPDI of p^*	Sample size
County 1	0.261	0.019	(0.219, 0.297)	164
County 2	0.267	0.020	(0.227, 0.304)	176
County 3	0.260	0.014	(0.235, 0.287)	795
County 4	0.246	0.019	(0.208, 0.281)	162
County 5	0.259	0.021	(0.219, 0.300)	125
County 6	0.262	0.021	(0.220, 0.300)	141
County 7	0.262	0.021	(0.223, 0.300)	128
County 8	0.261	0.019	(0.220, 0.292)	176

From the pooled area model results we can see that the proportion of obesity in each county population varies from 0.246 to 0.267, suggesting that there is some differentiation between the counties. County 6 and 7 have very similar predictions, both sharing obesity proportions and posterior standard errors of 0.262 and 0.021, respectively. Analogously, County 1 and 8 are alike with predicted obesity proportions and posterior standard errors of 0.261 and 0.019, respectively. County 4 has the most distinct predicted proportion of obesity at 0.246, which is noticeably lower compared to the other counties. County 2 has the largest predicted proportion of obesity at 0.267. We observe that most of the county's posterior standard error for the predicted proportion of obesity is between 0.019 to 0.021, with the exception of County 3, which has a significantly lower posterior standard error of 0.014. County 3 has a lower posterior standard error due to the fact that the sample size for this county is 795 individuals compared to the other counties that all have sample sizes of less than 200 observations. All of the individual counties' posterior standard errors are larger compared to the unordered and ungrouped model from Section 2.3.1, because of this decrease in sample size when we look at the county level. This results in County 3 having the tightest prediction interval for the proportion of obesity in the county, but this interval is still wider than the unordered and ungrouped model. Using this pooled area method allows for the characteristics of each county to be analyzed instead of the effects of all the areas pooling together in a grand prediction. This pooled area model is a good option when analysis is desired at a smaller scale.

2.4 Conclusion

We have explored several different versions of the multinomial-Dirichlet model with the goal of avoiding defining the relationship between a response variable and the covariates. Each version of the model addresses a different concern or situation, and therefore gives us flexibility to choose what model best fits our problem. The various situations we illustrated that this model can be applied to include, making inference about any characteristic (or multiple characteristics) for a finite population, for a number of small areas, or for a finite population using survey weight adjusted data.

The unordered and ungrouped multinomial-Dirichlet model shows a good performance compared to the logistic regression model, by providing comparable results and requiring less assumptions. The main issue with the unordered and ungrouped multinomial-Dirichlet model is that the number of parameters drawn in the Gibbs sampler is greater than the number of data points we have, so our model is over-parameterized. We also illustrated the inclusion of survey weights in the unordered and ungrouped model using adjusted cell counts. When using the adjusted cell counts the adjusted sample size decreases, which leads to larger posterior standard errors in every prediction. This is to be expected with any decrease in sample size. The adjusted cell counts are adjusted to remove bias from our sample and represent a more accurate depiction of the presence of these cells in the population. Since our BMI data is a probability survey we are able to calculate these adjusted cell counts.

We attempted to solve the issue of having too many parameters in the unordered and ungrouped model by incorporating the ordered and grouped multinomial-Dirichlet model. However, the population predictions from this model move closer to simply predicting the sample proportion and relying less on the model. In the ordered and grouped model, the stick-breaking weights are used less in the posterior predictions due to the increased value of ρ. We would rather utilize the stick-breaking weights rather than depend more heavily on the sample data, and for that reason the ordered and grouped model is not a good solution to reduce the number of parameters in the unordered and ungrouped model.

Lastly, the pooled area model shares some of the parameters across counties and more data points are introduced when we split the data up by county. We increase the number of data points, but we do not increase the number of parameters at the same rate. This model has a good performance and allows the unique characteristics of each area to be reflected in the predictions. Therefore, proportion predictions in Section 2.3.3 vary by county and there is less pooling to the overall proportion average.

In future work, we aim to revise the over-parameterized model using a different method that does not lead to increased global pooling. One idea is to truncate the stick-breaking weights once their value falls below a certain threshold and immediately stop drawing μ's in the Gibbs sampler. Since the stick-breaking weights tend to zero by nature, we can set the threshold such that the number of parameters being sampled are less than the number of data points present. We also try modeling with spatial relations included to see if borrowing strength from neighbors can help improve prediction with less global pooling.

Appendix A - Prior Beta Parameters

Recall the prior on $\boldsymbol{\mu}$ is:

$$\mu_1, \mu_2, \ldots, \mu_{c-1} \overset{\text{ind}}{\sim} \text{Beta}\left(\phi\frac{1-\gamma}{\gamma}, (1-\phi)\frac{1-\gamma}{\gamma}\right).$$

The ranges of ϕ and γ are constrained to ensure that the prior beta distribution parameters are greater than 1, this provides log-concavity of the prior. To show how we obtain the distributions: $\phi \sim \text{Uniform}(\frac{\gamma}{1-\gamma}, \frac{1-2\gamma}{1-\gamma})$ and $\gamma \sim \text{Uniform}(0, \frac{1}{3})$

We impose the restriction such that $\phi\frac{1-\gamma}{\gamma} > 1$ and $(1-\phi)\frac{1-\gamma}{\gamma} > 1$. Solving the first inequality for ϕ we obtain $\phi > \frac{\gamma}{1-\gamma}$, and solving the second inequality for ϕ we obtain

$\phi < \frac{1-2\gamma}{1-\gamma}$. Resulting in $\phi \sim \text{Uniform}(\frac{\gamma}{1-\gamma}, \frac{1-2\gamma}{1-\gamma})$.

Then, we solve the inequality $\frac{\gamma}{1-\gamma} < \frac{1-2\gamma}{1-\gamma} < 1$ for γ to obtain $0 < \gamma < \frac{1}{3}$. Resulting in $\gamma \sim \text{Uniform}(0, \frac{1}{3})$.

Appendix B - Search for ρ

We need to sample ρ differently, as we have to search for the range where ρ actually exists. In the search for the posterior location of ρ, we began with $\rho \sim \text{Uniform}(0, 1)$. Initially, we ran the multinomial-Dirichlet model with $0 < \rho < 1$ and sampled values using the griddy Gibbs sampler. Then, we noticed the sampler almost always selected the first grid as seen below:

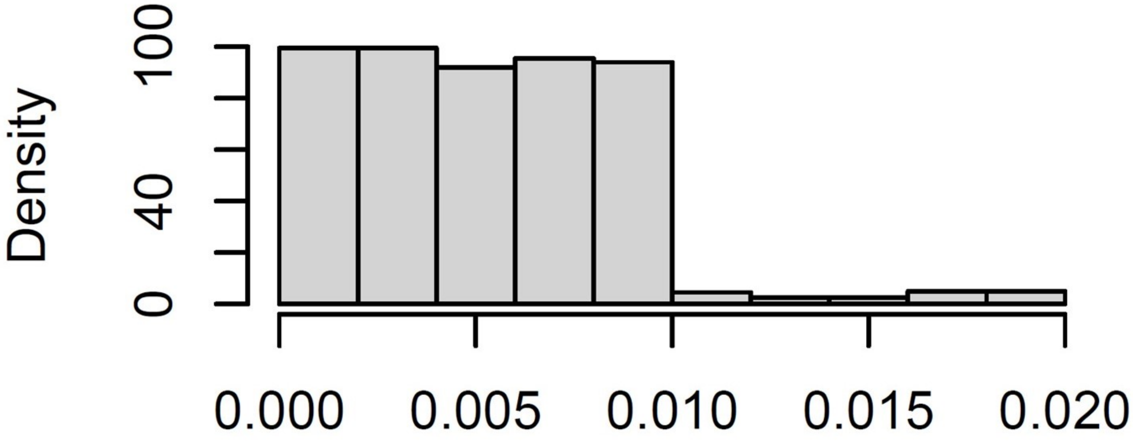

Figure 2: With ρ location between (0,1)

Since ρ was not identifiable in this range, we reduced the scope of ρ to be $0 < \rho < 0.1$, which makes the grids smaller in the sampler. This adjustment provides a more identifiable distribution of ρ, and we use this range for ρ in Section 2.3. Updated histogram shown below:

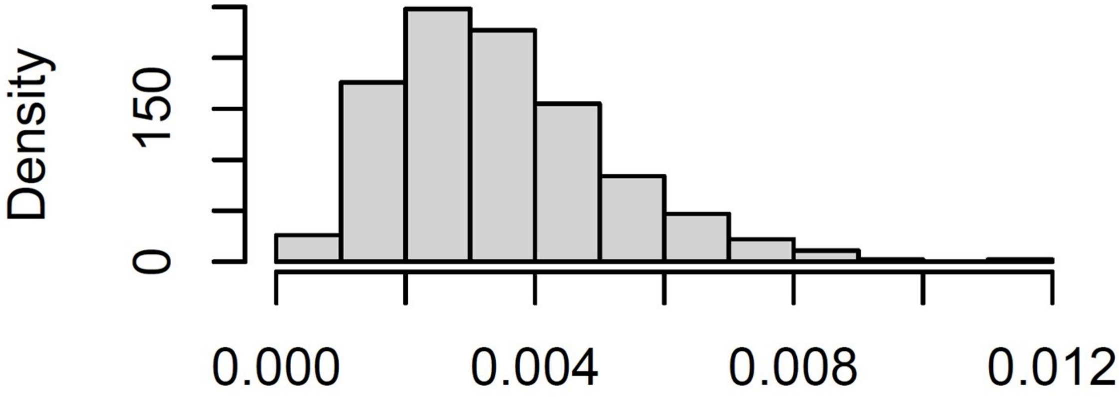

Figure 3: With ρ location between (0,0.1)

We can also see in the trace plot that ρ converges between $(0, 0.008)$, and the chain appears to be strongly mixing:

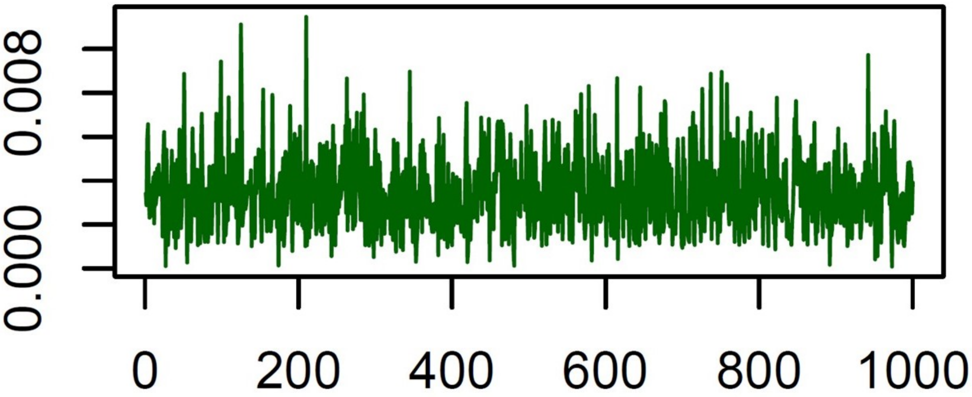

Figure 4: Trace plot of ρ values

Appendix C - Logistic Regression Model

In this section we describe the details of the logistic regression model used for comparison to the multinomial-Dirichlet models presented in Section 2.2. In the logistic regression model, our data contain binary information, and we make inference about $\boldsymbol{\beta}$, the regression coefficient(s) relating to some characteristic(s). In the BMI example presented in Section

2.3, the response variable is the obesity indicator where $y_i = 0$ represents "not obese" and $y_i = 1$ represents "obese". The covariates included in this example are age, race, and sex and their interactions: age×race, age×sex, race×sex, and age×race×sex. Therefore, there are 8 β's in the model, $\beta_0, \ldots, \beta_7$, where β_0 is the intercept coefficient. We include all interaction terms in the formula to be estimated, so the logistic model can be directly compared to the multinomial-Dirichlet model that contains all interactions between variables by default. The logistic model can be written:

$$\begin{gathered} y_i \mid \boldsymbol{\beta}, \boldsymbol{x_i} \overset{iid}{\sim} \text{Ber}\left(\frac{e^{\boldsymbol{x_i}'\beta}}{1+e^{\boldsymbol{x_i}'\beta}}\right), \qquad i = 1, \ldots, n, \\ \pi(\boldsymbol{\beta}) = 1, \end{gathered} \tag{47}$$

where $\boldsymbol{X} = (\boldsymbol{x_i}')$ is the design matrix and $\boldsymbol{x_i}'$ correspond to the rows of $\boldsymbol{X}$. We are using an improper prior for $\boldsymbol{\beta}$. The joint density is:

$$f(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{x_i}) = \prod_{i=1}^{n} \left\{ \frac{e^{\boldsymbol{x_i}'\beta y_i}}{1+e^{\boldsymbol{x_i}'\beta}} \right\}. \tag{48}$$

Therefore, we obtain the following conditional posterior density (CPD) for $\boldsymbol{\beta}$:

$$\pi(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{x_i}) = \prod_{i=1}^{n} \left\{ \frac{e^{\boldsymbol{x_i}'\beta y_i}}{1+e^{\boldsymbol{x_i}'\beta}} \right\}, \qquad y_i = 0, 1. \tag{49}$$

Since $\boldsymbol{\beta}$ is unbounded and does not have a standard posterior density, use the Metropolis-Hastings algorithm to obtain a sample of $\boldsymbol{\beta}$, with $\pi(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{x_i})$ as the target density. For the algorithm we also need a candidate-generating density (or proposal density) denoted $q(\boldsymbol{\beta_i})$. We use the mode-Hessian approximation to obtain the Student's T distribution as our proposal density with mode $\boldsymbol{m}$ and Hessian matrix H to get the covariance matrix Σ.

To obtain the mode, $\boldsymbol{m}$, we take the log of $\pi(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{x_i})$ denoted $l(\boldsymbol{\beta})$,

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\boldsymbol{x_i}'\boldsymbol{\beta} y_i - \log\left(1 + e^{\boldsymbol{x_i}'\beta}\right)\right]. \tag{50}$$

Then we use the Nelder-Mead algorithm to obtain the mode of $l(\boldsymbol{\beta})$. We obtain the Hessian matrix manually:

$$H = \frac{d^2}{d\boldsymbol{\beta}^2} l(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \left[\frac{\boldsymbol{x_i}\boldsymbol{x_i}'e^{\boldsymbol{x_i}'\beta}}{(1+e^{\boldsymbol{x_i}'\beta})^2} \right]. \tag{51}$$

Then we let $\Sigma = -H^{-1}$ evaluated at $\boldsymbol{m}$. Therefore, the proposal density has mode $\boldsymbol{m}$ and covariance matrix Σ. The steps for the Metropolis-Hastings algorithm are as follows:

1. Start with an initial value $\boldsymbol{\beta}^0$;

2. Generate $\boldsymbol{\beta}^1$ from $q\left(\boldsymbol{\beta^1}\right)$;
3. Compute $\alpha\left(\boldsymbol{\beta}^0, \boldsymbol{\beta}^1\right) = \min\left\{1, \frac{\pi\left(\beta^1\right)q\left(\beta^0\right)}{\pi\left(\beta^0\right)q\left(\beta^1\right)}\right\}$;
4. Draw $U \sim \text{Uniform}(0,1)$;
5. If $u \leq \alpha\left(\boldsymbol{\beta}^0, \boldsymbol{\beta}^1\right)$ accept value of $\boldsymbol{\beta^1}$, otherwise stay at $\boldsymbol{\beta^0}$;
6. Repeat steps 1-5 until convergence.

We also tune the degrees of freedom on the Student's T distribution to ensure the jumping rate is between $25\% - 75\%$, and for our BMI example the degrees of freedom is set equal to 8. Using this method, the Metropolis-Hastings sampler shows good performance as observable by the trace plots, auto-correlations, Geweke test of stationarity, the effective sample sizes, and the jumping rate of approximately 49%. The p-values for the Geweke test on $\boldsymbol{\beta}$ are $0.638, 0.999, 0.168, 0.189, 0.034, 0.438, 0.713,$ and 0.745 for $\beta_0, \ldots, \beta_7$, respectively. The p-value for β_4 is slightly too low, however the trace plot shows that stationarity does not seem to be an issue. The effective sample sizes for $\boldsymbol{\beta}$ are $1000, 1180, 1000, 1000, 1000, 1000, 1147,$ and 1147 for $\beta_0, \ldots, \beta_7$, respectively. Table 3 and Table 5 contain the results of the logistic model, and Figure 1 illustrates the comparison between the logistic model and the two multinomial-Dirichlet models. Section 2.2.2 describes how to make finite population predictions of the response variable in this logistic regression model.

Although we use an improper prior for $\boldsymbol{\beta}$ in (47), we prove the posterior distribution of $\pi\left(\boldsymbol{\beta} \mid \boldsymbol{y}\right)$ is proper. Let, $s_k = \sum\limits_{i=1}^{n} y_i x_{ik}$ and $t_k = \sum\limits_{i=1}^{n} x_{ik}$.

Theorem: Assume $t_k - s_k > 1$ and $s_k > 1$, $k = 1, \ldots, c$ (this is an extremely mild condition). Then, the posterior distribution of $\pi\left(\boldsymbol{\nu} \mid \boldsymbol{y}\right)$ is proper for all $0 \leq \nu_1, \ldots, \nu_c \leq 1$.

Proof:

We show that the posterior distribution of $\pi\left(\boldsymbol{\beta} \mid \boldsymbol{y}\right)$ is proper, even using an improper prior for $\boldsymbol{\beta}$. Given the model setup in (47), we can write the CPD in (49) as:

$$\pi(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{x_i}) = \frac{e^{\left(\sum\limits_{i=1}^{n} y_i \boldsymbol{x_i}'\right)\boldsymbol{\beta}}}{\prod\limits_{i=1}^{n}\left[1 + e^{\boldsymbol{x_i}'\boldsymbol{\beta}}\right]} \qquad = \frac{e^{\sum\limits_{k=1}^{c}\left(\sum\limits_{i=1}^{n} y_i x_{ik}\right)\beta_k}}{\prod\limits_{i=1}^{n}\left[1 + e^{\sum\limits_{k=1}^{c} x_{ik}\beta_k}\right]}. \tag{52}$$

Let $\nu_k = \frac{e^{\beta_k}}{1+e^{\beta_k}}$ therefore $0 < \nu_k < 1$, $k = 1, \ldots, c$. Now:

$$
\begin{aligned}
\nu_k + \nu_k e^{\beta_k} &= e^{\beta_k}, \\
\left(1-\nu_k\right) e^{\beta_k} &= \nu_k, \\
e^{\beta_k} &= \frac{\nu_k}{1-\nu_k}, \\
\text{So, } \beta_k = \log\left(\frac{\nu_k}{1-\nu_k}\right) & \qquad k = 1, \ldots, c, \\
d\beta_k = \frac{1-\nu_k}{\nu_k} \cdot \frac{1-\nu_k+\nu_k}{\left(1-\nu_k\right)^2} &= \frac{1}{\nu_k\left(1-\nu_k\right)} d\nu_k \qquad k = 1, \ldots, c.
\end{aligned}
\tag{53}
$$

Therefore,

$$
\pi\left(\boldsymbol{\nu} \mid \boldsymbol{y}\right) \propto \frac{\prod\limits_{k=1}^{c}\left[\frac{\nu_k}{1-\nu_k}\right]^{\sum\limits_{i=1}^{n} y_i x_{ik}}}{\prod\limits_{i=1}^{n}\left(1+\prod\limits_{k=1}^{c}\left[\frac{\nu_k}{1-\nu_k}\right]^{x_{ik}}\right)} \cdot \frac{1}{\nu_k\left(1-\nu_k\right)}. \tag{54}
$$

With, $s_k = \sum\limits_{i=1}^{n} y_i x_{ik}$ and $t_k = \sum\limits_{i=1}^{n} x_{ik}$,

$$
\pi\left(\boldsymbol{\nu} \mid \boldsymbol{y}\right) \propto \frac{\prod\limits_{k=1}^{c} {\nu_k}^{s_k-1}\left(1-\nu_k\right)^{t_k-s_k-1}}{\prod\limits_{i=1}^{n}\left[\prod\limits_{k=1}^{c} {\nu_k}^{x_{ik}} + \prod\limits_{k=1}^{c}\left(1-\nu_k\right)^{x_{ik}}\right]}, \qquad 0 < \nu_k < 1 \quad k = 1, \ldots, c. \tag{55}
$$

We want $s_k > 1$ and $t_k - s_k > 1$, $k = 1, \ldots, c$, and $x_{ik} \geq 0$. We add $\min\limits_{i}\left(x_{ik}\right)$ to x_{ik}, $k = 1, \ldots, c$, if $x_{ik} < 0$. Then,

$$
\pi\left(\boldsymbol{\nu} \mid \boldsymbol{y}\right) \propto \frac{\prod\limits_{k=1}^{c} {\nu_k}^{s_k-1}\left(1-\nu_k\right)^{t_k-s_k-1}}{\prod\limits_{i=1}^{n}\left[\prod\limits_{k=1}^{c} {\nu_k}^{x_{ik}} + \prod\limits_{k=1}^{c}\left(1-\nu_k\right)^{x_{ik}}\right]}, \qquad 0 \leq \nu_k \leq 1 \quad k = 1, \ldots, c. \tag{56}
$$

Now, ν_k is in the closed interval $[0, 1]$. $\pi\left(\boldsymbol{\nu} \mid \boldsymbol{y}\right)$ is well defined for all $0 \leq \nu_k \leq 1$, $k = 1, \ldots, c$. Therefore, $\pi\left(\boldsymbol{\nu} \mid \boldsymbol{y}\right)$ is proper with virtually no condition at all.

Other proofs for the propriety of the logistic regression model require stronger assumptions compared to the proof presented here. For instance, in Chen, Ibrahim, and Kim (2008), they use Jeffreys prior for $\boldsymbol{\beta}$ instead of a non-informative prior and they assume that the design matrix, $\boldsymbol{X}$, is full rank (Chen, Ibrahim, and Kim 2008).

Appendix D - Griddy Gibbs Sampler

Here we discuss the details of the griddy Gibbs sampler (Ritter and Tanner 1992). We are sampling values of $\boldsymbol{\mu}, \rho, \phi$, and γ within the griddy Gibbs sampler. It is called the griddy Gibbs sampler since we are using the grid method within a Gibbs sampler to draw each of the parameters. The grid method is the preferred method to sample $\boldsymbol{\mu}, \rho, \phi$, and γ since all of these parameters have values in the interval $[0,1]$.

The conditional posterior density for $\boldsymbol{\mu}$:

$$\pi\left(\boldsymbol{\mu} \mid \rho, \phi, \gamma, \boldsymbol{n}\right) \propto \frac{\prod_{i=1}^{c}\prod_{m=1}^{n_i}(\alpha_i + m - 1)}{\prod_{k=1}^{N}\left(\frac{1-\rho}{\rho} + k - 1\right)} \cdot \left(\prod_{j=1}^{c-1}\mu_j\right)^{\left(\phi\frac{1-\gamma}{\gamma}-1\right)} \left(\prod_{j=1}^{c-1}(1-\mu_j)\right)^{\left((1-\phi)\frac{1-\gamma}{\gamma}-1\right)}. \tag{57}$$

Recall that $\boldsymbol{\alpha}$ contains $\boldsymbol{\mu}$ as defined in (24).

The conditional posterior density for ρ:

$$\pi\left(\rho \mid \boldsymbol{\mu}, \phi, \gamma, \boldsymbol{n}\right) \propto \frac{\prod_{i=1}^{c}\prod_{m=1}^{n_i}(\alpha_i + m - 1)}{\prod_{k=1}^{N}\left(\frac{1-\rho}{\rho} + k - 1\right)}. \tag{58}$$

The conditional posterior density for ϕ:

$$\begin{aligned}\pi\left(\phi \mid \boldsymbol{\mu}, \rho, \gamma, \boldsymbol{n}\right) \propto &\left[\frac{\Gamma\left(\frac{1-\gamma}{\gamma}\right)}{\Gamma\left(\phi\frac{1-\gamma}{\gamma}\right)\Gamma\left((1-\phi)\frac{1-\gamma}{\gamma}\right)}\right]^{c-1} \\ &\cdot \left(\prod_{j=1}^{c-1}\mu_j\right)^{\left(\phi\frac{1-\gamma}{\gamma}-1\right)} \left(\prod_{j=1}^{c-1}(1-\mu_j)\right)^{\left((1-\phi)\frac{1-\gamma}{\gamma}-1\right)}.\end{aligned} \tag{59}$$

The conditional posterior density for γ:

$$\begin{aligned}\pi\left(\gamma \mid \boldsymbol{\mu}, \rho, \phi, \boldsymbol{n}\right) \propto &\left[\frac{\Gamma\left(\frac{1-\gamma}{\gamma}\right)}{\Gamma\left(\phi\frac{1-\gamma}{\gamma}\right)\Gamma\left((1-\phi)\frac{1-\gamma}{\gamma}\right)}\right]^{c-1} \\ &\cdot \left(\prod_{j=1}^{c-1}\mu_j\right)^{\left(\phi\frac{1-\gamma}{\gamma}-1\right)} \left(\prod_{j=1}^{c-1}(1-\mu_j)\right)^{\left((1-\phi)\frac{1-\gamma}{\gamma}-1\right)}.\end{aligned} \tag{60}$$

The initial values we used to start the Gibbs sampler are $\phi^0 = \frac{1}{c-1}\sum_{j=1}^{c-1}\mu_j^0$, $\rho^0 = 0.05$, $\gamma^0 = 0.2$, and $\boldsymbol{\mu}^0$ was found by recursively solving for μ_j^0, $j = 1, \ldots, c-1$, from the stick-breaking weight definition given in (24). Using the sample data counts we defined the initial weights of the cells to be $w_i^0 = \frac{n_i}{T}$, $i = 1, \ldots, c$. Therefore, using the initial values $\boldsymbol{w}^0$ we are able to solve for $\boldsymbol{\mu}^0$ by using (24).

Chapter 3

Spatial Modeling Techniques with a Continuous Response Variable

3.1 Introduction

The methods we present avoid defining a parametric relationship by considering each unique combination of the covariates in the population as an individual stratum. Then we use small area estimation techniques to make inference about each subset of the population based on its underlying covariates. Finally, we can estimate the overall finite population mean by pooling predictions of the strata together.

We present two versions of spatial models, a conditional autoregressive (CAR) model and a simple conditional autoregressive (SCAR) model (Chung and Datta 2022). We accommodate the covariates by using the spatial model instead of a regression model. For the spatial models, we include the spatial effects by creating the incidence matrix (or adjacency matrix) via the Mahalanobis distance between the covariates for each stratum. We use these spatial models to see if creating a neighborhood relationship between similar strata allows for less global pooling to the overall sample mean (Tang, Ghosh, Ha, and Sedransk 2018 and Jo, Nandram, and Kim 2021). By enabling strata to have neighbors, we expect neighborhoods to pool together without all strata pooling together.

We also present two non-spatial models for comparison, a form of the Scott-Smith model (Scott and Smith 1969) and a form of the Battese, Harter, and Fuller (BHF) model (Battese, Harter, and Fuller 1988). The BHF model is a more general version of the Scott-Smith model that includes the covariates where the Scott-Smith model does not. We use both non-spatial models as a baseline for comparison to see how including a spatial relationship in the models in this chapter impacts the results. Appendix E contains the full technical details of the Scott-Smith model, and Appendix F contains the technical details of the BHF model.

For the remainder of the chapter, we will discuss the methodology of the two spatial models in Section 3.2. We also discuss an extension of including survey weights into the previously mentioned spatial models in Section 3.2. Then in Section 3.3, an application using BMI data with each of the models is given, followed by a conclusion in Section 3.4. The appendix contains technical details for the Scott-Smith and BHF models.

3.2 Methodology

In this section, we show two spatial models and how we include survey weights in the spatial models. First, in Section 3.2.1 we present the two spatial models with the CAR model in Section 3.2.1.1 and the SCAR model in Section 3.2.1.2. Then, Section 3.2.2 illustrates how the survey weights can be included in the spatial models presented in Section 3.2.1. The methodologies for the Scott-Smith model and the BHF model can be found in Appendix E and Appendix F, respectively.

In all four models we observe a continuous response variable y_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, \ell$, and these responses are grouped together based on their covariate values. Each possible combination of covariates is taken into consideration, and each unique combination is considered to be an stratum. Therefore, each $\boldsymbol{y_i}$ has a unique corresponding covariate combination denoted $\boldsymbol{x_i}$, where $\boldsymbol{y_i}$ is the aggregated vector of responses of length n_i for each stratum. The covariate matrix $\boldsymbol{X} = (\boldsymbol{x_i}')$ has dimension $\ell \times p$ where $(p-1)$ is the number of covariates in the data. The first column of $\boldsymbol{X}$ represents the intercept, and $\boldsymbol{x_i}$ corresponds to the unique rows of $\boldsymbol{X}$. We make inference about the finite population mean, $\bar{Y}_i = \sum\limits_{j=1}^{N_i} y_{ij}/N_i$, based on the observed values of $\boldsymbol{y_i}$. Denote the sampling fraction as $f_i = n_i/N_i$, where n_i represents the sample size and N_i represents the population size for a given stratum. The N_i are unknown and we discuss later how to estimate them using inverse probability weighting.

3.2.1 Spatial Models

For the spatial models, we include the spatial effects by creating the symmetric incidence matrix, $\boldsymbol{W}$ of size $\ell \times \ell$, via the Mahalanobis distance between $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$, $i = 1, \ldots, \ell$, $j = 1, \ldots, \ell$, and $i \neq j$. The Mahalanobis distance is defined as:

$$d_{ij} = \sqrt{(\boldsymbol{x_i} - \boldsymbol{x_j})' \boldsymbol{S}^{-1} (\boldsymbol{x_i} - \boldsymbol{x_j})}, \tag{61}$$

where $\boldsymbol{S}$ is the covariance matrix of $\boldsymbol{X}$, and $d_{ii} = 0$. We define $\boldsymbol{W}$ by letting $w_{ij} = 1$ if $d_{ij} \leq d_0$ and $w_{ij} = 0$ if $d_{ij} > d_0$ with zeroes on the diagonal (i.e., $w_{ii} = 0$). The value d_0 yields the $\boldsymbol{W}$ matrix that maximizes Moran's I, which is defined as:

$$I = \frac{\ell}{w_{..}} \frac{\sum_i \sum_j w_{ij} (\bar{y}_i - \bar{y}) (\bar{y}_j - \bar{y})}{\sum_i (\bar{y}_i - \bar{y})^2}, \tag{62}$$

where $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ is the response variable; $\bar{y} = \sum_{i=1}^{\ell} \bar{y}_i/\ell$ is the overall sample mean response; w_{ij} corresponds to the elements of $\boldsymbol{W}$; and $w_{..} = \sum_i \sum_j w_{ij}$.

In Section 3.2.1.1 we describe the conditional autoregressive (CAR) model and in Section 3.2.1.2 we state the difference between this model and the simple conditional autoregressive

(SCAR) model.

3.2.1.1 CAR Model The Bayesian hierarchical CAR model is:

$$
\begin{aligned}
y_{ij} \mid \boldsymbol{\mu}, \sigma^2 &\sim \text{Normal}\left(\mu_i, \sigma^2\right), \\
\boldsymbol{\mu} \mid \theta, \rho, \sigma^2, \gamma &\sim \text{Normal}\left(\theta \mathbf{1}, \frac{\rho}{1-\rho}\sigma^2 \left(\boldsymbol{R} - \gamma \boldsymbol{W}\right)^{-1}\right), \\
\pi\left(\theta, \sigma^2, \rho\right) &\propto \frac{1}{\sigma^2}, \\
\gamma &\sim \text{Uniform}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_\ell}\right), \\
\frac{1}{\lambda_1} \leq \gamma \leq \frac{1}{\lambda_\ell}, \quad -\infty < \theta < \infty, &\quad 0 < \rho < 1, \quad \sigma^2 > 0, \\
j = 1, \ldots, n_i, &\quad i = 1, \ldots, \ell,
\end{aligned}
\tag{63}
$$

where ℓ in this case represents the total number of possible covariate combinations, which are considered to be the individual strata. We discretize any continuous variables so there is a finite number of possible covariate combinations. Then, we store the continuous responses, $\boldsymbol{y_i}$, such that there are n_i responses for each stratum $i = 1, \ldots, \ell$. Here $\boldsymbol{R}$ is a diagonal $\ell \times \ell$ matrix defined as $\boldsymbol{R} = \text{diag}\left\{w_{i\cdot}\right\}_{i=1}^{\ell}$ where $w_{i\cdot} = \sum_{j=1}^{n_i} w_{ij}$ is the sum of the ith row of $\boldsymbol{W}$. Also, λ_1 is the minimum eigenvalue of $\boldsymbol{R}^{-1}\boldsymbol{W}$ and λ_ℓ is the maximum eigenvalue of $\boldsymbol{R}^{-1}\boldsymbol{W}$, and since $\sum_{i=1}^{\ell} w_{ii} = 0$ this results in $\lambda_1 < 0 < \lambda_\ell$ (Chung and Datta 2022). Here, $(\boldsymbol{R} - \gamma\boldsymbol{W})$ is guaranteed to be positive definite as long as γ is in the range $\frac{1}{\lambda_1} \leq \gamma \leq \frac{1}{\lambda_\ell}$. To obtain samples from the joint posterior density of this model, we can integrate out $\boldsymbol{\mu}$, θ, and σ^2, and then we only need to draw γ and ρ using a griddy Gibbs sampler (Ritter and Tanner 1992).

We can vectorize the continuous response variable, y_{ij}, to be $\boldsymbol{y}$ with dimension $n \times 1$ where $n = \sum\limits_{i=1}^{\ell} n_i$ such that:

$$
\boldsymbol{y}_{(n\times 1)} \mid \boldsymbol{\mu}, \sigma^2 \sim \text{Normal}\left(A\boldsymbol{\mu}, \sigma^2 \mathbb{I}_{n\times n}\right), \tag{64}
$$

where $\mathbb{I}_{n\times n}$ is the identity matrix and A has dimension $n \times \ell$ and can be defined as

$$
A = \begin{pmatrix} \mathbf{1_1} & 0 & \cdots & 0 \\ 0 & \mathbf{1_2} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \mathbf{1_\ell} \end{pmatrix}, \tag{65}
$$

and $\mathbf{1_1}$ through $\mathbf{1_\ell}$ are vectors of ones with lengths corresponding to the number of observations in that stratum. Therefore, $\mathbf{1_1}$ is a vector of ones with length n_1, $\mathbf{1_2}$ is a vector of ones with length n_2, and so on through $\mathbf{1_\ell}$. The purpose of writing the model in this way is so we can use the lemma from Section 2 in Lindley and Smith (1972) to obtain the posterior distribution of $\boldsymbol{\mu}$ that we draw samples of $\boldsymbol{\mu}$ from:

$$\begin{aligned} \boldsymbol{\mu} \mid \Omega, \boldsymbol{y} \sim \text{Normal}\Bigg[&\left(\boldsymbol{D} + \frac{1-\rho}{\rho}\left(\boldsymbol{R} - \gamma\boldsymbol{W}\right)\right)^{-1}\left(A'\boldsymbol{y} + \left(\frac{1-\rho}{\rho}\left(\boldsymbol{R} - \gamma\boldsymbol{W}\right)\right)\theta\mathbf{1}\right), \\ &\sigma^2\left(\boldsymbol{D} + \frac{1-\rho}{\rho}\left(\boldsymbol{R} - \gamma\boldsymbol{W}\right)\right)^{-1}\Bigg], \end{aligned} \tag{66}$$

where $\boldsymbol{D} = \text{diag}\left(n_1, \ldots, n_\ell\right)$. Let $\Omega = (\theta, \rho, \sigma^2, \gamma)$ for simplicity of notation.

Another way of writing this spatial model to make it simpler to integrate out $\boldsymbol{\mu}$ would be:

$$\begin{aligned} \bar{\boldsymbol{y}} \mid \boldsymbol{\mu}, \sigma^2 &\sim \text{Normal}\left(\boldsymbol{\mu}, \sigma^2\text{diag}\left(\frac{1}{n_1}, \ldots, \frac{1}{n_\ell}\right)\right), \\ \boldsymbol{\mu} \mid \theta, \rho, \sigma^2, \gamma &\sim \text{Normal}\left(\theta\mathbf{1}, \frac{\rho}{1-\rho}\sigma^2\left(\boldsymbol{R} - \gamma\boldsymbol{W}\right)^{-1}\right). \end{aligned} \tag{67}$$

Now, if we integrate out $\boldsymbol{\mu}$ from this model we are left with the posterior density:

$$\begin{aligned} \pi\left(\theta, \rho, \sigma^2, \gamma \mid \boldsymbol{y}\right) \propto &\det\left[\sigma^2\left(\text{diag}\left(\frac{1}{n_1}, \ldots, \frac{1}{n_\ell}\right) + \frac{\rho}{1-\rho}\left(\boldsymbol{R} - \gamma\boldsymbol{W}\right)^{-1}\right)\right]^{-1/2} \\ &\times \exp\left\{\frac{-1}{2\sigma^2}\left(\bar{\boldsymbol{y}} - \theta\mathbf{1}\right)'\left(\text{diag}\left(\frac{1}{n_1}, \ldots, \frac{1}{n_\ell}\right) + \frac{\rho}{1-\rho}\left(\boldsymbol{R} - \gamma\boldsymbol{W}\right)^{-1}\right)^{-1}\left(\bar{\boldsymbol{y}} - \theta\mathbf{1}\right)\right\} \\ &\times \left(\frac{1}{\sigma^2}\right)^{(n-\ell)/2}\exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{\ell}(n_i - 1)s_i^2\right\} \times \frac{1}{\sigma^2}. \end{aligned} \tag{68}$$

From this density, we can see that θ follows a normal distribution:

$$\theta \mid \sigma^2, \rho, \gamma, \bar{\boldsymbol{y}} \sim \text{Normal}\left(\hat{\theta}, \frac{\sigma^2}{\mathbf{1}'\Sigma\mathbf{1}}\right), \tag{69}$$

where $\hat{\theta} = \mathbf{1}'\Sigma\bar{\boldsymbol{y}}/\mathbf{1}'\Sigma\mathbf{1}$ and $\Sigma = \left[\text{diag}\left(\frac{1}{n_1}, \ldots, \frac{1}{n_\ell}\right) + \frac{\rho}{1-\rho}\left(\boldsymbol{R} - \gamma\boldsymbol{W}\right)^{-1}\right]^{-1}$. We can use this fact to integrate out θ, so we have,

$$\begin{aligned}\pi\left(\rho,\sigma^2,\gamma\mid\boldsymbol{y}\right)&\propto\det\left[\sigma^2\left(\operatorname{diag}\left(\frac{1}{n_1},\ldots,\frac{1}{n_\ell}\right)+\frac{\rho}{1-\rho}\left(\boldsymbol{R}-\gamma\boldsymbol{W}\right)^{-1}\right)\right]^{-1/2}\\&\times\exp\left\{\frac{-1}{2\sigma^2}\left(\hat{\theta}\mathbf{1}-\bar{\boldsymbol{y}}\right)'\Sigma\left(\hat{\theta}\mathbf{1}-\bar{\boldsymbol{y}}\right)\right\}\times\sqrt{2\pi\sigma^2/\mathbf{1}'\Sigma\mathbf{1}}\\&\times\left(\frac{1}{\sigma^2}\right)^{(n-\ell)/2}\exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{\ell}(n_i-1)s_i^2\right\}\times\frac{1}{\sigma^2}.\end{aligned}\tag{70}$$

From this density, we can find the inverse-gamma distribution,

$$\sigma^2\mid\rho,\gamma,\bar{\boldsymbol{y}}\sim\operatorname{InvGam}\left(\frac{n-1}{2},\left[\left(\hat{\theta}\mathbf{1}-\bar{\boldsymbol{y}}\right)'\Sigma\left(\hat{\theta}\mathbf{1}-\bar{\boldsymbol{y}}\right)+\sum_{i=1}^{\ell}(n_i-1)s_i^2\right]/2\right).\tag{71}$$

Finally, after integrating out σ^2 we have the nonstandard joint posterior density,

$$\begin{aligned}\pi\left(\rho,\gamma\mid\boldsymbol{y}\right)\propto\det&\left[\left(\operatorname{diag}\left(\frac{1}{n_1},\ldots,\frac{1}{n_\ell}\right)+\frac{\rho}{1-\rho}\left(\boldsymbol{R}-\gamma\boldsymbol{W}\right)^{-1}\right)\right]^{-1/2}\\\times&\left[\left(\hat{\theta}\mathbf{1}-\bar{\boldsymbol{y}}\right)'\Sigma\left(\hat{\theta}\mathbf{1}-\bar{\boldsymbol{y}}\right)+\sum_{i=1}^{\ell}(n_i-1)s_i^2\right]^{\frac{-n+1}{2}}\times\left(\mathbf{1}'\Sigma\mathbf{1}\right)^{-1/2}.\end{aligned}\tag{72}$$

Using the griddy Gibbs sampler, we can draw samples of ρ and γ from (72) (Ritter and Tanner 1992). We use the same conditional posterior density to draw both parameters using a grid method, however the grids of ρ and γ differ since their ranges of support are not equivalent. This method mixes well and converges to the target distribution quickly. Then, continuing in reverse order we can input our samples of ρ and γ to directly obtain samples of σ^2 from (71), next θ from (69), and finally $\boldsymbol{\mu}$ from (66). Obtaining samples of σ^2, θ, and $\boldsymbol{\mu}$ is straight-forward since they all have known distributions. Based on our samples of $\boldsymbol{\mu}$, σ^2 and the observed values of $\boldsymbol{y_i}$, we make inference for the finite population mean $\bar{Y}_i$, using the model:

$$\bar{Y}_i\mid\mu_i,\sigma^2,\boldsymbol{y_i}\overset{\text{ind}}{\sim}\operatorname{Normal}\left(f_i\bar{y}_i+(1-f_i)\mu_i,(1-f_i)\frac{\sigma^2}{N_i}\right).\tag{73}$$

We examine the performance of this model in Section 3.3 with an application using BMI data.

3.2.1.2 SCAR Model Below we describe the simple conditional autoregressive (SCAR) model and state the differences between this model and the previously discussed CAR model. The main computational difference between the two models is that in the CAR model the matrix $\boldsymbol{R}$ is used in the variance of the prior on $\boldsymbol{\mu}$, and the SCAR model replaces $\boldsymbol{R}$ with the identity matrix, $\boldsymbol{I}$, hence simplifying the model. The Bayesian hierarchical SCAR model is given as:

$$\begin{gathered}
y_{ij} \mid \boldsymbol{\mu}, \sigma^2 \sim \text{Normal}\left(\mu_i, \sigma^2\right), \\
\boldsymbol{\mu} \mid \theta, \rho, \sigma^2, \gamma \sim \text{Normal}\left(\theta \mathbf{1}, \frac{\rho}{1-\rho}\sigma^2 \left(\boldsymbol{I} - \gamma \boldsymbol{W}\right)^{-1}\right), \\
\pi\left(\theta, \sigma^2, \rho\right) \propto \frac{1}{\sigma^2}, \\
\gamma \sim \text{Uniform}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_\ell}\right), \\
\frac{1}{\lambda_1} \leq \gamma \leq \frac{1}{\lambda_\ell}, \quad -\infty < \theta < \infty, \quad 0 < \rho < 1, \quad \sigma^2 > 0, \\
j = 1, \ldots, n_i, \quad i = 1, \ldots, \ell.
\end{gathered} \tag{74}$$

Here, λ_1 is the minimum eigenvalue of $\boldsymbol{W}$ and λ_ℓ is the maximum eigenvalue of $\boldsymbol{W}$. Also $(\boldsymbol{I} - \gamma \boldsymbol{W})$ is guaranteed to be positive definite as long as γ is in the range $\frac{1}{\lambda_1} \leq \gamma \leq \frac{1}{\lambda_\ell}$. Aside from the few small computational changes mentioned, the distributions and methods we use to obtain a sample from the posterior density and the method we use to make inference about the finite population mean, $\bar{Y}_i$, is the same as described in Section 3.2.1.1. We simply substitute $\boldsymbol{R}$ for the matrix $\boldsymbol{I}$ and use the updated values of λ_1 and λ_ℓ accordingly. We illustrate the performance of this model in Section 3.3 with an application using BMI data.

In the SCAR model, the precision matrix is set to be the identity matrix, $\boldsymbol{I}$. While the diagonal elements of the precision matrix are all equal, the diagonal elements of the inverse may not be all equal thus allowing for heteroscedasticity of random effects. In the CAR model, diagonal entries of the precision matrix, $\boldsymbol{R}$, are the number of neighbors corresponding to each stratum. Therefore, the matrix $\boldsymbol{R}$ weights each row by the number of neighbors it has, and acts as a normalizing matrix. Both the SCAR and CAR models assume that μ_i depends only on neighboring strata means. That is, μ_i is correlated with μ_j, $j \neq i$, only through the means of surrounding strata, and μ_i is not correlated with remote strata (Chung and Datta 2022). Note that in the CAR and SCAR models, it is important that ρ and γ are not too small, because we want to emphasize the spatial structure in order to accommodate the covariates.

3.2.2 Including Survey Weights

In this section, we show how to include survey weights in the two spatial models we are advocating for. Here we use the original survey weights, denoted v_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots \ell$, to calculate the effective sample size and the adjusted weights a_{ij}. First, we calculate the effective sample size, $\hat{n}$:

$$\hat{n} = \frac{\left(\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v_{ij}\right)^2}{\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v_{ij}^2}. \tag{75}$$

The effective sample size, $\hat{n}$, illustrates how severely the variance is increased by the unequal weighting (Nandram and Rao 2021). Then, we calculate the adjusted weights, a_{ij}, which are used to eliminate the bias present in the original survey weights:

$$a_{ij} = \hat{n}\frac{v_{ij}}{\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v_{ij}}, \tag{76}$$

where $\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v_{ij} = \hat{N}$ is the Horvitz-Thompson estimator of population size, and $\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} a_{ij} = \hat{n}$, the effective sample size.

These adjusted weights a_{ij} are able to be used in a model when the data do not have outliers present. However, in the BMI example in Section 3.3 the original data do have outliers. Therefore, we use Winsorization, which is an effective method to deal with outliers by trimming the survey weights (Yang, Nandram, and Choi 2023). Outliers here are defined as observed survey weights greater than $v_0 = Q_3 + 1.5\,(Q_3 - Q_1)$, where Q_1 is the first quartile and Q_3 is the third quartile. Let v^* denote weights after trimming,

$$v_{ij}^* = \begin{cases} v_0, & v_{ij} \geq v_0 \\ r v_{ij}, & v_{ij} < v_0 \end{cases}, \tag{77}$$

where r is a rescaling parameter such that $\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v_{ij}^* = \sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v_{ij} = \hat{N}$. Then we obtain the adjusted and trimmed weights a_{ij}^*,

$$\begin{aligned} \hat{n}^* &= \frac{\left(\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v_{ij}^*\right)^2}{\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v_{ij}^{*2}}, \\ a_{ij}^* &= \hat{n}^*\frac{v_{ij}^*}{\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v_{ij}^*}. \end{aligned} \tag{78}$$

These adjusted and trimmed weights a_{ij}^* are used in both the CAR and SCAR model.

3.2.2.1 Inlcuding Survey Weights in CAR Model The CAR model with the adjusted weights included can be written:

$$
\begin{gathered}
y_{ij} \mid \boldsymbol{\mu}, \sigma^2 \sim \text{Normal}\left(\mu_i, \frac{\sigma^2}{a_{ij}^*}\right), \\
\boldsymbol{\mu} \mid \theta, \rho, \sigma^2, \gamma \sim \text{Normal}\left(\theta \mathbf{1}, \frac{\rho}{1-\rho} \sigma^2 \left(\boldsymbol{R} - \gamma \boldsymbol{W}\right)^{-1}\right), \\
\pi\left(\theta, \sigma^2, \rho\right) \propto \frac{1}{\sigma^2}, \\
\gamma \sim \text{Uniform}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_\ell}\right), \\
\frac{1}{\lambda_1} \leq \gamma \leq \frac{1}{\lambda_\ell}, \quad -\infty < \theta < \infty, \quad 0 < \rho < 1, \quad \sigma^2 > 0, \\
j = 1, \ldots, n_i, \quad i = 1, \ldots, \ell.
\end{gathered} \tag{79}
$$

We use the same logic for obtaining a sample from this model with the adjusted weights as we used in Section 3.2.1.1. The difference is in how we make population predictions, by including the survey weights we now need to use surrogate sampling techniques. We obtain population predictions by:

$$
\overline{Y_i} \mid \boldsymbol{\mu}, \sigma^2 \sim \text{Normal}\left(\mu_i, \frac{\sigma^2}{N_i}\right), \qquad i = 1, \ldots, \ell, \tag{80}
$$

where $N_i = \sum_{j=1}^{n_i} v_{ij}$ represents the Horvitz-Thompson estimator of population size for each strata $i = 1, \ldots, \ell$. We no longer need to include the sample means, $\bar{y}_i$, or the sampling fraction, f_i, in this population prediction.

Previously when making population predictions, we combined both the sampled part and non-sampled part of the population to obtain a sample of $\overline{Y_i}$. However, now that we are utilizing the survey weights from the probability sample we are able to use surrogate sampling techniques and no longer need to include the sampled part. The adjusted and trimmed survey weights, $\boldsymbol{a}^*$, are used to simulate an unbiased sample from the population. With the adjusted and trimmed weights included in the model we must sample the entire population using surrogate sampling techniques, because the survey weights of both the sample and the nonsample are biased (Nandram 2007, Nandram and Rao 2021). Then, when we make population predictions in (80) we use the unadjusted survey weights for prediction since these weights accurately represent the entire population. We examine the performance of this CAR model with the inclusion of survey weights in Section 3.3 with an application using BMI data.

3.2.2.2 Inlcuding Survey Weights in SCAR Model Similarly, the SCAR model with the adjusted weights included can be written:

$$
\begin{gathered}
y_{ij} \mid \boldsymbol{\mu}, \sigma^2 \sim \text{Normal}\left(\mu_i, \frac{\sigma^2}{a^*_{ij}}\right), \\
\boldsymbol{\mu} \mid \theta, \rho, \sigma^2, \gamma \sim \text{Normal}\left(\theta \mathbf{1}, \frac{\rho}{1-\rho}\sigma^2 \left(\boldsymbol{I} - \gamma \boldsymbol{W}\right)^{-1}\right), \\
\pi\left(\theta, \sigma^2, \rho\right) \propto \frac{1}{\sigma^2}, \\
\gamma \sim \text{Uniform}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_\ell}\right), \\
\frac{1}{\lambda_1} \leq \gamma \leq \frac{1}{\lambda_\ell}, \quad -\infty < \theta < \infty, \quad 0 < \rho < 1, \quad \sigma^2 > 0, \\
j = 1, \ldots, n_i, \quad i = 1, \ldots, \ell.
\end{gathered} \tag{81}
$$

We use the same logic for obtaining a sample from this model with the adjusted weights as we used in Section 3.2.1.2. Similar to the CAR model with survey weights included we now need to use surrogate sampling techniques in the SCAR model with survey weights to make population predictions. We use (80) to make our population predictions in the SCAR model with survey weights included. We examine the performance of this SCAR model with the inclusion of survey weights in Section 3.3 with an application using BMI data.

3.3 Application using BMI Data

When interested in the health of a population, the the BMI levels of individuals may be an important indicator. We illustrate our various non-spatial and spatial models using a probability sample of BMI data with 1,867 individuals from eight counties in California recorded in the Third National Health and Nutrition Examination Survey (NHANES III). The survey weights sum to 12,232,099, which means our sample accounts for 0.015% of the population. These data have four variables we are interested in including age, race, sex, and a continuous measure of BMI in kg/m^2. The values for race are "white" or "non-white", and the values for sex are "male" or "female". The age variable is a continuous variable included in the data, and age ranges from 20 years old to 90 years old. We bin the age variable into groups of two years, therefore the bins are 20-21 years old, 22-23 years old, and so on, in order to have a finite number of possible covariate combinations. This idea of binning variables is commonly used in practice, as it lessens the need to rely on exact accuracy of the data and allows for inference to be made about a broader age group. The value for the continuous BMI response variable ranges from 15.8 kg/m^2 to 58.4 kg/m^2.

After aggregating over all possible age, race, sex, and obesity combinations there are 144 strata each with its own unique set of covariate values. However, in our sample of BMI data we have 12 strata with no observations and we assume these are structural zeroes in the data. This means that we assume these 12 groups of individuals do not exist in our population. Therefore, the total number of strata, ℓ, in this case represents the total number of possible covariate combinations available in our sample and $\ell = 132$ after removing the 12 structural zeroes from the data. If necessary, we can mitigate the structural zeroes by using coarser groups of the covariates.

In Section 3.3.1 we illustrate and compare the results between the two spatial and two non-spatial models, with results both including and excluding survey weights. Then in Section 3.3.2, we show how including the spatial component in our models reduced the amount of global pooling, compared to the non-spatial models.

3.3.1 BMI Application Model Comparison

Before sampling from either of the spatial models, we first create the symmetric incidence matrix, $\boldsymbol{W}$ of size 132×132, using the Mahalanobis distance described in (61). Recall that we define $\boldsymbol{W}$ by letting $w_{ij} = 1$ if $d_{ij} \leq d_0$ and $w_{ij} = 0$ if $d_{ij} > d_0$ with zeroes on the diagonal (i.e., $w_{ii} = 0$), where d_0 is the value yielding the $\boldsymbol{W}$ matrix that maximizes Moran's I from (62). After performing a grid search for the optimal value of d_0, we achieved the maximum value of Moran's I at $I = 0.212$ when we let $d_0 = \text{mean}(d_{ij})/38 \approx 0.157$. In general, we found decreasing d_0 increases Moran's I up to a certain point. In our case, if we continue to decrease d_0 to be less than 0.157, we will not see any increase in Moran's I. However, if we increase d_0 to be greater than 0.157 then Moran's I will decrease.

Now that we have $\boldsymbol{W}$, we can proceed drawing samples from the CAR and SCAR models described in Section 3.2.1. Sampling from these models is extremely similar as they both begin using the griddy Gibbs sampler to obtain samples of ρ and γ simultaneously (Ritter and Tanner 1992). For the CAR model the grid interval for γ is: $(-1.94, 1)$ and for the SCAR model the grid interval for γ is: $(-0.390, 0.169)$. The grid intervals for γ in these models differ since the range of γ is based on the eigenvalues of $\boldsymbol{R}^{-1}\boldsymbol{W}$ for the CAR model and the eigenvalues of $\boldsymbol{W}$ for the SCAR model. The grid interval from the CAR model is better in the sense that it brings γ closer to unity. We also present the results with the survey weights included in the CAR and SCAR models.

We ran 10,000 iterations of the sampler, then dropped the first 1,000 sampled values and chose every 9th sampled value to end with a final sample size of 1,000 for both parameters. In both the CAR and SCAR models, the griddy Gibbs sampler shows good performance as evident by the trace plots, auto-correlations, Geweke test of stationarity, and the effective sample sizes. For ρ and γ the P-values for the Geweke test are 0.237 and 0.286, respectively, in the CAR model and 0.938 and 0.833, respectively, in the SCAR model, meaning both parameters pass stationarity requirements in each model. As seen in Table 7 in the CAR model excluding survey weights, the posterior means for ρ and γ are 0.188 and 0.937, respectively, and in the SCAR model excluding survey weights, the posterior means for ρ and γ are 0.044 and 0.165, respectively. Since it is important that ρ and γ are not too small considering we want to emphasize the spatial structure that accommodates the covariates, then we prefer the CAR model that has larger values of both ρ and γ. In the CAR model, γ is close to unity, which is a good sign that our spatial component will have more of an impact in the model compared to the much lower γ value in the SCAR model. As ρ and γ decrease, our posterior standard error of the population predictions will also decrease.

In Table 8, we observe that the SCAR model has a slightly lower posterior standard error compared to the CAR model, and this is due to the lower values of ρ and γ in the SCAR model. After successfully running the griddy Gibbs sampler to obtain values for ρ and γ, we continue to sample the remaining parameters σ^2, θ, and $\boldsymbol{\mu}$ directly from their known posterior densities for both the CAR and SCAR models. Each model contains 136 total parameters that we then use to predict the BMI of the population using (73). The method for obtaining samples from the CAR and SCAR models with weights included is the same as described in the models with weights excluded, and the griddy Gibbs sampler has a very similar good performance (Ritter and Tanner 1992). Making population predictions is different in the CAR and SCAR models when we include the survey weights, as shown in (80).

Table 7: Posterior Estimates of ρ and γ

	PM	PSE	CV	95% HPDI
Models excluding Survey Weights				
ρ (CAR)	0.188	0.045	0.244	(0.108, 0.289)
γ (CAR)	0.937	0.050	0.054	(0.848, 1.000)
ρ (SCAR)	0.044	0.013	0.292	(0.021, 0.069)
γ (SCAR)	0.165	0.004	0.026	(0.159, 0.169)
Models including Survey Weights				
ρ (CAR)	0.185	0.048	0.262	(0.098, 0.282)
γ (CAR)	0.940	0.052	0.055	(0.851, 1.000)
ρ (SCAR)	0.042	0.013	0.316	(0.018, 0.068)
γ (SCAR)	0.166	0.004	0.023	(0.159, 0.169)

A sample from the Scott-Smith model (presented in Appendix E) can be obtained directly without the need for any MCMC algorithm. Once we sample ρ using the grid method, then we can draw samples of σ^2, θ, and $\boldsymbol{\mu}$ in order from their known conditional posterior densities. Since there are 132 strata our model contains $\mu_1, \ldots, \mu_{132}$, resulting in 135 total parameters sampled in this model. Once these parameters are obtained we are able to make predictions for the BMI of the population using (90).

Similar to the Scott-Smith model, a sample from the BHF model (presented in Appendix F) can also be obtained without the need for an MCMC sampler. For this model, we sample ρ using the grid method, then we can draw samples of σ^2, $\boldsymbol{\beta}$, and $\boldsymbol{\nu}$ in order from their known conditional posterior densities. This is the only model containing covariates directly in the model, so we have $\beta_0, \ldots, \beta_3$ where β_0 represents the intercept's coefficient. Therefore, there are 138 total parameters that once sampled we are able to predict the BMI of the population using (104). The Scott-Smith and BHF models with weights included are fit similar to the case without weights, except again population predictions are different now. The population predictions can be made using (92) and (106) for the Scott-Smith and BHF models with weights included, respectively. The Scott-Smith and BHF models are used as a baseline to compare to the performance of the spatial model.

Table 8: BMI Population Prediction Model Comparison

	Predicted $\bar{Y}$	SE of $\bar{Y}$	CV of $\bar{Y}$	95% HPDI of $\bar{Y}$
Models excluding Survey Weights				
CAR	27.402	0.091	0.003	(27.233, 27.584)
SCAR	27.418	0.088	0.003	(27.237, 27.579)
Scott-Smith	27.375	0.129	0.005	(27.117, 27.634)
BHF	27.347	0.132	0.005	(27.109, 27.608)
Models including Survey Weights				
CAR	27.070	0.100	0.004	(26.879, 27.263)
SCAR	27.090	0.100	0.004	(26.898, 27.268)
Scott-Smith	27.380	0.147	0.005	(27.070, 27.656)
BHF	27.294	0.161	0.006	(27.007, 27.614)

Table 8 contains the results of the population prediction of BMI for all four models, both excluding and including survey weights. In our application the response variable is BMI, therefore $\overline{Y}$ represents the overall mean of BMI for the population of the eight counties in California. The results from the two non-spatial models, the Scott-Smith and the BHF models, are very similar in all four measures of the posterior mean, posterior standard error (SE), coefficient of variation (CV), and highest posterior density interval (HPDI). The two spatial models, CAR and SCAR, also perform similar to each other. The CAR and SCAR models resulted in a slightly higher prediction of the posterior finite population mean BMI of the population. The spatial models also have a smaller posterior standard error and CV, which yields a tighter HPDI compared to the non-spatial models. Since strata are gaining strength from neighboring strata in the spatial models, we see the posterior standard errors decrease while the posterior means are more tailored to each neighborhood. Strata with a very small sample size are no longer relying on solely their limited number of observations in the spatial models, since they are now included in neighborhoods that collectively have a larger number of observations. In the models without the spatial component, predictions for strata are more general leading to more vague predictions centered around the overall sample mean with larger posterior standard error.

Table 8 also contains the results of the population prediction of BMI when the adjusted and trimmed survey weights are included in the models. From the table we can see that the overall population prediction for the finite population mean BMI is similar to that of the models without weights. However, the posterior standard error and the CV increase in the models with the weights compared to the models without weights. Having larger standard error in turn also leads to the models with survey weights having wider HPDIs. By including the adjusted and trimmed survey weights in the model, we do expect the posterior standard error to increase since including the weights decreases the sample size to the effective sample size. Naturally, with a smaller sample size the posterior standard error will be larger.

Table 9 contains the results of the population prediction of BMI for all four models split by the eight counties included in our BMI survey data. The models were not fit to each county, rather the results were simply separated by county. The results for each county are similar to the overall results described from Table 8, however since the sample sizes are reduced when we group by county then the posterior standard errors will increase. All of the counties have roughly the same sample size, except for County 3 that has an exceptionally large sample size of 795 observations. County 3 accounts for about 43% of the total BMI data sample size. The remainder of the counties each have sample sizes ranging from 125 to 176 observations. The comparatively large sample size of County 3 yields smaller posterior standard errors compared to the other counties.

Table 10 contains the results of the population prediction of BMI for all four models with survey weights included split by the eight counties included in our BMI survey data. Again, the models were not fit to each county, and instead the results were simply separated by county. The results for each county are similar to the overall results described from Table

Table 9: County Level BMI Population Prediction Model Comparison

	Predicted $\bar{Y}$	SE of $\bar{Y}$	CV of $\bar{Y}$	95% HPDI of $\bar{Y}$
County 1				
CAR	27.242	0.101	0.004	(27.061, 27.457)
SCAR	27.232	0.097	0.004	(27.044, 27.414)
Scott-Smith	27.194	0.149	0.005	(26.913, 27.481)
BHF	27.198	0.151	0.006	(26.891, 27.480)
County 2				
CAR	27.549	0.112	0.004	(27.308, 27.759)
SCAR	27.554	0.112	0.004	(27.336, 27.768)
Scott-Smith	27.491	0.178	0.006	(27.154, 27.852)
BHF	27.487	0.174	0.006	(27.138, 27.801)
County 3				
CAR	27.460	0.094	0.003	(27.268, 27.635)
SCAR	27.470	0.093	0.003	(27.292, 27.656)
Scott-Smith	27.424	0.140	0.005	(27.142, 27.701)
BHF	27.389	0.140	0.005	(27.100, 27.646)
County 4				
CAR	27.446	0.139	0.005	(27.178, 27.700)
SCAR	27.467	0.144	0.005	(27.164, 27.729)
Scott-Smith	27.462	0.212	0.008	(27.029, 27.842)
BHF	27.428	0.219	0.008	(27.008, 27.829)
County 5				
CAR	27.557	0.123	0.004	(27.310, 27.794)
SCAR	27.563	0.124	0.004	(27.329, 27.801)
Scott-Smith	27.551	0.188	0.007	(27.188, 27.939)
BHF	27.497	0.194	0.007	(27.125, 27.899)
County 6				
CAR	27.481	0.129	0.005	(27.236, 27.725)
SCAR	27.502	0.131	0.005	(27.260, 27.781)
Scott-Smith	27.408	0.205	0.007	(26.984, 27.766)
BHF	27.366	0.206	0.008	(26.967, 27.753)
County 7				
CAR	27.083	0.114	0.004	(26.856, 27.287)
SCAR	27.109	0.110	0.004	(26.892, 27.322)
Scott-Smith	27.113	0.163	0.006	(26.827, 27.450)
BHF	27.087	0.166	0.006	(26.769, 27.405)
County 8				
CAR	27.218	0.120	0.004	(26.983, 27.445)
SCAR	27.273	0.113	0.004	(27.050, 27.490)
Scott-Smith	27.205	0.163	0.006	(26.895, 27.503)
BHF	27.184	0.170	0.006	(26.885, 27.581)

Table 10: Including Survey Weights in County Level BMI Population Prediction

	Predicted $\bar{Y}$	SE of $\bar{Y}$	CV of $\bar{Y}$	95% HPDI of $\bar{Y}$
County 1				
CAR	27.158	0.114	0.004	(26.949, 27.411)
SCAR	27.185	0.111	0.004	(26.970, 27.401)
Scott-Smith	27.347	0.169	0.006	(27.011, 27.666)
BHF	27.202	0.175	0.006	(26.888, 27.575)
County 2				
CAR	26.945	0.125	0.005	(26.709, 27.202)
SCAR	27.021	0.117	0.004	(26.806, 27.259)
Scott-Smith	27.275	0.164	0.006	(26.950, 27.575)
BHF	27.416	0.200	0.007	(27.048, 27.814)
County 3				
CAR	27.196	0.102	0.004	(26.999, 27.383)
SCAR	27.199	0.104	0.004	(27.014, 27.417)
Scott-Smith	27.362	0.145	0.005	(27.070, 27.631)
BHF	27.329	0.171	0.006	(26.987, 27.634)
County 4				
CAR	26.983	0.122	0.005	(26.749, 27.229)
SCAR	26.990	0.123	0.005	(26.734, 27.206)
Scott-Smith	27.268	0.165	0.006	(26.979, 27.607)
BHF	27.352	0.253	0.009	(26.852, 27.816)
County 5				
CAR	27.162	0.118	0.004	(26.932, 27.391)
SCAR	27.171	0.117	0.004	(26.943, 27.400)
Scott-Smith	27.284	0.174	0.006	(26.953, 27.633)
BHF	27.393	0.224	0.008	(26.967, 27.859)
County 6				
CAR	27.142	0.115	0.004	(26.924, 27.379)
SCAR	27.135	0.118	0.004	(26.887, 27.344)
Scott-Smith	27.320	0.179	0.007	(26.972, 27.659)
BHF	27.298	0.233	0.009	(26.871, 27.761)
County 7				
CAR	26.935	0.121	0.004	(26.706, 27.180)
SCAR	26.923	0.120	0.004	(26.694, 27.158)
Scott-Smith	27.107	0.174	0.006	(26.794, 27.477)
BHF	27.085	0.194	0.007	(26.725, 27.459)
County 8				
CAR	26.840	0.127	0.005	(26.590, 27.090)
SCAR	26.903	0.126	0.005	(26.669, 27.150)
Scott-Smith	27.054	0.171	0.006	(26.742, 27.403)
BHF	27.156	0.198	0.007	(26.774, 27.585)

8. In Table 10, the sample sizes are being reduced by the inclusion of survey weights, in addition to the sample sizes being split by county. The posterior standard errors continue to increase due to both factors reducing the sample sizes. Recall that all of the counties have roughly the same sample size, except for County 3, which has an exceptionally large sample size. This large sample size of County 3 yields the smallest posterior standard errors compared to the other counties.

3.3.2 Reduction in Global Pooling via Spatial Modeling

Our main goal is to make inference about the finite population mean without directly including the covariates in our models, which we have shown. A reason we chose to introduce a spatial component in our models is to have the posterior means of the individual strata result in less global pooling (Tang, Ghosh, Ha, and Sedransk 2018 and Jo, Nandram, and Kim 2021). We do not want the posterior mean of each individual stratum to simply approach the overall population posterior mean. Instead we would rather have strata with similar covariate attributes (i.e. neighbors in the $\boldsymbol{W}$ matrix) borrow strength from each other to have stratum posterior means gravitate towards the neighborhood posterior mean. Therefore, we study the $\boldsymbol{\mu}$ from the Scott-Smith model and the CAR model to show that by including the spatial component, as seen in the CAR model, we are able to increase the variability of the posterior predictions for $\boldsymbol{\mu}$. Since the results from the Scott-Smith model and the BHF model are similar we only use the Scott-Smith model in this comparison. In general, we avoid using covariates in our models (as seen in the BHF model) when possible and the similar results from the Scott-Smith and BHF models are evidence that including the covariates in the model did not improve the prediction results. Similarly, since the CAR and SCAR models yield similar results and the SCAR model is a simpler version of the CAR model, then we proceed making this comparison using the CAR model.

We are analyzing the $\boldsymbol{\mu}$ from each model, which means we have a sample from the posterior density of μ_i for each $i = 1, \dots, \ell$. There are three main indications that illustrate how including the spatial component in the CAR model reduces global pooling when compared to the Scott-Smith model. First, the estimates of $\mu_1, \dots, \mu_{132}$ from the CAR model have a standard deviation of 1.237, compared to the estimates from the Scott-Smith model, which have a standard deviation of 0.967. From this we can already see increased variation in the $\boldsymbol{\mu}$ estimates in the CAR model, and this increased variation is a sign that less global pooling occurs in the CAR model as $\boldsymbol{\mu}$ contains more disperse values. Second, by looking at Figure 5 that contains the two kernel density curves for $\boldsymbol{\mu}$ from each model, we are able to see that $\boldsymbol{\mu}$ from the CAR model has a lower peak and heavier tails compared to the Scott-Smith model. The values of $\boldsymbol{\mu}$ in the CAR model are not converging to the overall population mean as aggressively as the Scott-Smith model.

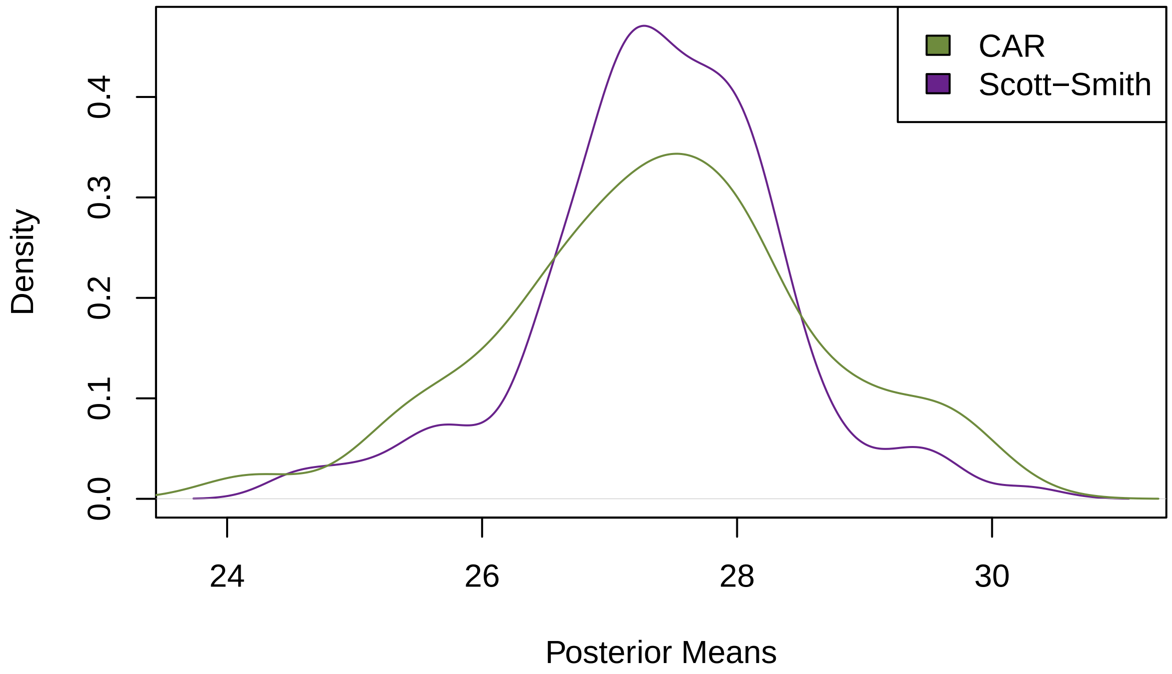

Figure 5: Comparing Posterior Distribution of μ

Thirdly, we look at the shrinkage parameters in the posterior mean of $\boldsymbol{\mu}$ from each model. Recall that the posterior mean of μ_i for the Scott-Smith model in (84) is: $(\lambda_i \bar{y}_i + (1-\lambda_i)\theta)$ where $\lambda_i = n_i\rho/\left((n_i-1)\,\rho+1\right)$, $i = 1\ldots,\ell$. We can rewrite the posterior mean of $\boldsymbol{\mu}$ in the CAR model in (66) as: $(\Lambda\bar{\boldsymbol{y}} + (\boldsymbol{I}-\Lambda)\theta\boldsymbol{1})$ where $\Lambda = \left(\text{diag}\left(\frac{\sigma^2}{n_1},\ldots,\frac{\sigma^2}{n_\ell}\right)^{-1} + \left(\frac{\rho}{1-\rho}\sigma^2\left(\boldsymbol{R}-\gamma\boldsymbol{W}\right)^{-1}\right)^{-1}\right)^{-1}\text{diag}\left(\frac{\sigma^2}{n_1},\ldots,\frac{\sigma^2}{n_\ell}\right)^{-1}$. When the shrinkage parameters from the Scott-Smith model, $(1-\lambda_i)$, are greater than the sum of the row values of the shrinkage parameters from the CAR model, $(\boldsymbol{I}-\Lambda)$, then the non-spatial Scott-Smith model tends more towards the global pooling parameter, θ, instead of maintaining the characteristics of the individual strata. This is exactly what we see in Figure 6.

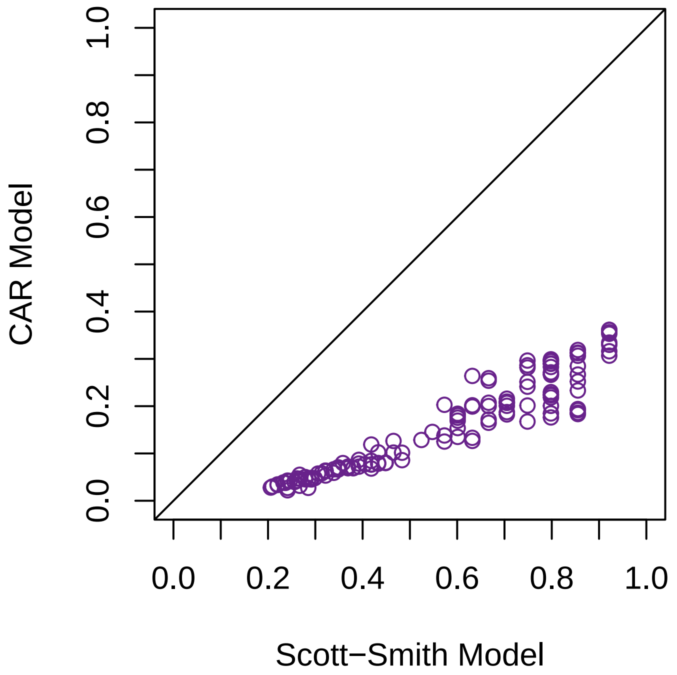

Figure 6: Comparing Shrinkage Parameters

Figure 6 shows that the Scott-Smith model puts significantly more weight on the global pooling parameter, θ, compared to the CAR model. This maintains our objective that the CAR model would result in less global pooling overall by including the neighborhood relationships. It is also important to point out that all of the shrinkage parameter values for both the Scott-Smith and the CAR models are in the range $[0, 1]$.

3.4 Conclusion

Our main goal in introducing the spatial component in these models is to accommodate the covariates without using a regression model. In doing so, we also reduce the severity of global pooling and instead allow for neighbors with similar attributes to have predictions closer together. We have shown in our comparison of the $\boldsymbol{\mu}$ in the CAR and Scott-Smith models that including this spatial relationship of the strata does indeed limit the global pooling. This point was made by looking at the shrinkage parameters, the posterior densities, and the posterior variation of $\boldsymbol{\mu}$ from both models. The CAR and SCAR models both work well as small area estimation models that reduce global pooling effects without defining the relationship between the response and the covariates. However, we prefer the CAR model, which has larger values of both ρ and γ. The CAR model is favored, since it is important that ρ and γ are not too small in order to emphasize the spatial structure that accommodates the covariates.

In the CAR model γ is close to unity, which is a good sign that our spatial component will have more of an impact in the model compared to the much lower γ value in the SCAR model. As ρ and γ decrease, our posterior standard error of the population predictions will also decrease. We also presented how to use these spatial models we are advocating for with and without survey weights, and how to make population predictions in both cases. Ultimately, we are not interested in defining a relationship between the response variable $\boldsymbol{y}$ and $\boldsymbol{X}$ by having $\boldsymbol{\beta}$ in the model, as is the case in the BHF model. We avoid making strong assumptions about this relationship, and we maximize the number of potential applications our models can be applied to.

Future work includes continuing to work on this type of problem by adapting the models to cover the situation with a binary response variable instead of a continuous response variable. While the binary case is more computationally intense, it has a lot of useful applications. For an example related to the BMI application, we may be more interested in the proportion of individuals in the population who are obese (i.e. BMI ≥ 30), instead of predicting the overall BMI of the finite population. There are cases where knowing the proportion of a characteristic possessed by a population is more informative than knowing the average value of that characteristic in the population.

Appendix E - Scott-Smith Model

In Appendix E we discuss the technical details of the Scott-Smith model (Scott and Smith 1969). The original version of the Scott-Smith model was used to model continuous data from two-stage cluster sampling, and there have since been many extensions to allow for a variety of sampling designs. See the references in Nandram, Toto, and Choi (2011) for an extension that accommodates binary data and shows other generalizations of this model.

The adapted Scott-Smith model we use can be written as:

$$
\begin{gathered}
y_{ij} \mid \mu_i, \sigma^2 \sim \text{Normal}\left(\mu_i, \sigma^2\right), \\
\mu_i \mid \theta, \rho, \sigma^2 \sim \text{Normal}\left(\theta, \frac{\rho}{1-\rho}\sigma^2\right), \\
\pi\left(\theta, \sigma^2, \rho\right) \propto \frac{1}{\sigma^2}, \\
-\infty < \theta < \infty, \quad 0 < \rho < 1, \quad \sigma^2 > 0, \\
j = 1, \ldots, n_i, \quad i = 1, \ldots, \ell.
\end{gathered}
\tag{82}
$$

Although the covariates are not present in the model, the responses are still grouped together using the covariate values, so each $\boldsymbol{y_i}$ has the same unique covariate combination for each stratum $i = 1, \ldots \ell$. Nandram, Toto, and Choi (2011) has shown that ρ is a common intra-class correlation. The joint posterior density is:

$$
\begin{aligned}
\pi\left(\boldsymbol{\mu}, \theta, \sigma^2, \rho \mid \boldsymbol{y}\right) \propto& \left(\frac{1}{\sigma^2}\right)^{(n+\ell)/2+1} \left(\frac{1-\rho}{\rho}\right)^{\ell/2} \\
&\times \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{\ell}\left\{\frac{n_i}{\lambda_i}\left(\mu_i - \left(\lambda_i \bar{y}_i + \left(1-\lambda_i\right)\theta\right)\right)^2 + \lambda_i\left(\frac{1-\rho}{\rho}\right)\left(\bar{y}_i - \theta\right)^2 + \left(n_i - 1\right)s_i^2\right\}\right]\right\},
\end{aligned}
\tag{83}
$$

where $\lambda_i = n_i\rho / \left(\left(n_i - 1\right)\rho + 1\right)$, $i = 1 \ldots, \ell$. From this joint posterior density we can see that μ_i follows a normal distribution:

$$
\mu_i \mid \theta, \sigma^2, \rho, \boldsymbol{y} \sim \text{Normal}\left(\lambda_i \bar{y}_i + (1-\lambda_i)\theta, (1-\lambda_i)\rho\sigma^2/(1-\rho)\right). \tag{84}
$$

We then integrate out $\boldsymbol{\mu}$ to obtain the posterior density of θ, σ^2, and ρ:

$$
\begin{aligned}
\pi_2\left(\theta, \sigma^2, \rho \mid \boldsymbol{y}\right) \propto \quad & \frac{\left(\prod_i \lambda_i\right)^{1/2}}{\left(\sigma^2\right)^{n/2+1}} \left(\frac{1-\rho}{\rho}\right)^{\ell/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{\ell}(n_i - 1)s_i^2\right\} \\
& \times \exp\left\{-\frac{1-\rho}{2\sigma^2\rho}\left\{\sum_{i=1}^{\ell}\lambda_i\left(\bar{y}_i - \widetilde{y}\right)^2 + \left(\sum_{i=1}^{\ell}\lambda_i\right)\left(\widetilde{y} - \theta\right)^2\right\}\right\}.
\end{aligned}
\tag{85}
$$

Note that $\widetilde{y} = \left(\sum_{i=1}^{\ell}\left(n_i / \left(\left(n_i - 1\right)\rho + 1\right)\right)\bar{y}_i\right) / \left(\sum_{i=1}^{\ell}\left(n_i / \left(\left(n_i - 1\right)\rho + 1\right)\right)\right)$ is well defined for all $0 \leq \rho \leq 1$ and $l \geq 2$. From (85) we can see the conditional posterior density of θ is:

$$\theta \mid \sigma^2, \rho, \boldsymbol{y} \sim \text{Normal}\left(\widetilde{y}, \frac{\sigma^2\rho}{(1-\rho)\sum\limits_{i=1}^{\ell}\lambda_i}\right). \tag{86}$$

Then we integrate out θ to obtain the joint posterior density of σ^2 and ρ:

$$\begin{aligned} \pi_3\left(\sigma^2, \rho \mid \boldsymbol{y}\right) \propto & \sqrt{\frac{\prod_i \lambda_i}{\sum_i \lambda_i}} \times \left(\frac{1-\rho}{\rho}\right)^{(l-1)/2} \left(\frac{1}{\sigma^2}\right)^{(n+1)/2} \\ & \times \exp\left\{-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{\ell}(n_i-1)s_i^2 + \frac{1-\rho}{\rho}\left(\sum_{i=1}^{\ell}\lambda_i\left(\bar{y}_i - \widetilde{y}\right)^2\right)\right\}\right\}. \end{aligned} \tag{87}$$

From (87) we can see the conditional posterior density of σ^2 is:

$$\sigma^2 \mid \rho, \boldsymbol{y} \sim \text{InvGamma}\left(\frac{n-1}{2}, \left\{\sum_{i=1}^{\ell}(n_i-1)s_i^2 + \frac{1-\rho}{\rho}\left(\sum_{i=1}^{\ell}\lambda_i\left(\bar{y}_i - \widetilde{y}\right)^2\right)\right\}/2\right). \tag{88}$$

Finally, once we integrate out σ^2 we are left with the nonstandard posterior density,

$$\begin{aligned} \pi_4(\rho \mid \boldsymbol{y}) \propto \quad & (1-\rho)^{(l-2)/2}\sqrt{\frac{\prod_{i=1}^{\ell} n_i/\left((n_i-1)\rho+1\right)}{\sum_{i=1}^{\ell} n_i/\left((n_i-1)\rho+1\right)}} \\ \times & \frac{1}{\left\{1 + (1-\rho)\left(\sum_{i=1}^{\ell}\left(n_i/\left((n_i-1)\rho+1\right)\right)\left(\bar{y}_i - \widetilde{y}\right)^2\right)/\left(\sum_{i=1}^{\ell}(n_i-1)s_i^2\right)\right\}^{(n-1)/2}}. \end{aligned} \tag{89}$$

This proves that the joint posterior density is proper, and this also shows how we can obtain a sample from the joint posterior density by sampling from $\pi_4(\rho \mid \boldsymbol{y})$ first and then continuing to draw samples from their known distributions in reverse order (Nandram, Toto, and Choi 2011). Therefore, we begin by drawing samples of ρ from (89) using the grid method. Next, we use the sample of ρ we obtained to draw a sample of σ^2 directly from (88). Then we use the samples of ρ and σ^2 to draw a sample of θ from its standard distribution (86). Finally, we use the samples of ρ, σ^2, and θ to draw a sample of μ_i, $i = 1, \dots, \ell$ from (84). Based on our samples from the posterior density and the observed values of $\boldsymbol{y_i}$, we make inference for the finite population mean $\bar{Y}_i$, using the model:

$$\bar{Y}_i \mid \boldsymbol{y_i} \overset{\text{ind}}{\sim} \text{Normal}\left(f_i\bar{y}_i + (1-f_i)\mu_i, (1-f_i)\frac{\sigma^2}{N_i}\right). \tag{90}$$

The results of this model are presented in Section 3.3 with an application using BMI data.

Inlcuding Survey Weights in Scott-Smith Model

We also can include survey weights in the Scott-Smith model, and we use the same adjusted and trimmed survey weights described in Section 3.2.2. We write the model with weights included as:

$$\begin{gathered}
y_{ij} \mid \mu_i, \sigma^2 \sim \text{Normal}\left(\mu_i, \frac{\sigma^2}{a^*_{ij}}\right), \\
\mu_i \mid \theta, \rho, \sigma^2 \sim \text{Normal}\left(\theta, \frac{\rho}{1-\rho}\sigma^2\right), \\
\pi\left(\theta, \sigma^2, \rho\right) \propto \frac{1}{\sigma^2}, \\
-\infty < \theta < \infty, \quad 0 < \rho < 1, \quad \sigma^2 > 0, \\
j = 1, \dots, n_i, \quad i = 1, \dots, \ell.
\end{gathered} \tag{91}$$

We use the same logic for obtaining a sample from this model with the adjusted weights as we used above in the model without weights. Making population predictions differs, because we obtain population predictions by:

$$\overline{Y_i} \mid \boldsymbol{\mu}, \sigma^2 \sim \text{Normal}\left(\mu_i, \frac{\sigma^2}{N_i}\right), \qquad i = 1, \dots, \ell, \tag{92}$$

where $N_i = \sum\limits_{j=1}^{n_i} v_{ij}$ represents the Horvitz-Thompson estimator of population size for each strata $i = 1, \dots, \ell$.

Appendix F - BHF Model

In Appendix F we discuss the technical details of the Battese, Harter, and Fuller (BHF) model (Battese, Harter, and Fuller 1988). We also fit the BHF model to compare with the spatial models. This model is a more general version of the Scott-Smith model, as the Scott-Smith model does not include the covariates. The BHF model is the only model used in this chapter that specifies the relationship between the response and the covariates. In general, we avoid defining this relationship between $\boldsymbol{y_i}$ and $\boldsymbol{x_i}$. The BHF model is:

$$\begin{aligned} y_{ij} \mid \boldsymbol{\nu}, \boldsymbol{\beta}, \sigma^2 &\overset{\text{ind}}{\sim} \text{Normal}\left(\boldsymbol{x_i}'\boldsymbol{\beta} + \nu_i, \sigma^2\right), \\ \boldsymbol{\nu} \mid \rho, \sigma^2 &\overset{\text{ind}}{\sim} \text{Normal}\left(0, \frac{\rho}{1-\rho}\sigma^2\right), \\ \pi\left(\boldsymbol{\beta}, \sigma^2, \rho\right) &\propto \frac{1}{\sigma^2}, \quad \sigma^2 > 0, \quad 0 < \rho < 1, \quad \boldsymbol{\beta} \in \mathbb{R}^p, \\ j &= 1, \ldots, n_i, \quad i = 1, \ldots \ell. \end{aligned} \tag{93}$$

This non-spatial model introduces covariates and includes the random effects for each stratum. With $n = \sum\limits_{i=1}^{\ell} n_i$, the joint posterior density of this model is:

$$\begin{aligned} \pi\left(\boldsymbol{\nu}, \boldsymbol{\beta}, \sigma^2, \rho \mid \boldsymbol{y}\right) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n+\ell}{2}+1} \left(\frac{1-\rho}{\rho}\right)^{\ell/2} &\times \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{\ell}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y_i}\right)^2\right\} \\ &\times \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{\ell}\left\{n_i\left(\bar{y_i} - \boldsymbol{x_i}'\boldsymbol{\beta} - \nu_i\right)^2 + \left(\frac{1-\rho}{\rho}\right)\nu_i^2\right\}\right\}. \end{aligned} \tag{94}$$

Letting $\lambda_i = \rho n_i/((1-\rho) + \rho n_i)$ we can write:

$$\begin{aligned} \pi\left(\boldsymbol{\nu}, \boldsymbol{\beta}, \sigma^2, \rho \mid \boldsymbol{y}\right) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n+\ell}{2}+1} \left(\frac{1-\rho}{\rho}\right)^{\ell/2} &\times \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{\ell}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y_i}\right)^2\right\} \\ &\times \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{\ell}\left\{n_i(1-\lambda_i)\left(\bar{y_i} - \boldsymbol{x_i}'\boldsymbol{\beta}\right)^2\right\}\right\} \\ &\times \exp\left\{-\frac{1-\rho}{2\rho\sigma^2}\sum_{i=1}^{\ell}\left\{\frac{1}{(1-\lambda_i)}\left[\nu_i - \lambda_i\left(\bar{y_i} - \boldsymbol{x_i}'\boldsymbol{\beta}\right)\right]^2\right\}\right\}. \end{aligned} \tag{95}$$

From (95) we can see that:

$$\nu_i \mid \boldsymbol{\beta}, \sigma^2, \rho, \boldsymbol{y} \overset{\text{ind}}{\sim} \text{Normal}\left(\lambda_i\left(\bar{y_i} - \boldsymbol{x_i}'\boldsymbol{\beta}\right), \frac{(1-\lambda_i)\rho\sigma^2}{(1-\rho)}\right). \tag{96}$$

Now, we can integrate out $\boldsymbol{\nu}$ to get:

$$\begin{aligned} \pi\left(\boldsymbol{\beta}, \sigma^2, \rho \mid \boldsymbol{y}\right) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+1} \prod_{i=1}^{\ell}(1-\lambda_i)^{\frac{1}{2}} &\times \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{\ell}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y_i}\right)^2\right\} \\ &\times \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{\ell}\left\{n_i(1-\lambda_i)\left(\bar{y_i} - \boldsymbol{x_i}'\boldsymbol{\beta}\right)^2\right\}\right\}. \end{aligned} \tag{97}$$

We can rewrite (97) as:

$$\begin{aligned}\pi\left(\boldsymbol{\beta},\sigma^2,\rho \mid \boldsymbol{y}\right) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+1} \prod_{i=1}^{\ell}\left(1-\lambda_i\right)^{\frac{1}{2}} \times \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{\ell}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y_i}\right)^2\right\} \\ \times \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{\ell} n_i\left(1-\lambda_i\right)\left(\bar{y_i}-\boldsymbol{x_i}'\hat{\boldsymbol{\beta}}\right)^2\right\} \\ \times \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\hat{\Sigma}^{-1}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right\},\end{aligned} \tag{98}$$

where

$$\hat{\boldsymbol{\beta}} = \hat{\Sigma}\sum_{i=1}^{\ell} n_i\left(1-\lambda_i\right)\bar{y}_i\boldsymbol{x_i}' \qquad \text{and} \qquad \hat{\Sigma} = \left(\sum_{i=1}^{\ell} n_i\left(1-\lambda_i\right)\boldsymbol{x_i}\boldsymbol{x_i}'\right)^{-1}. \tag{99}$$

Note that $\hat{\boldsymbol{\beta}}$ and $\hat{\Sigma}^{-1}$ are well defined for all ρ provided the design matrix $\boldsymbol{X} = (\boldsymbol{x_i}')$ is full rank, where $\boldsymbol{x_i}'$ correspond to the rows of $\boldsymbol{X}$, therefore,

$$\boldsymbol{\beta} \mid \sigma^2,\rho,\boldsymbol{y} \sim \text{Normal}\left(\hat{\boldsymbol{\beta}},\sigma^2\hat{\Sigma}\right). \tag{100}$$

We can use this to integrate out $\boldsymbol{\beta}$ to obtain the joint posterior density of σ^2 and ρ:

$$\begin{aligned}\pi\left(\sigma^2,\rho \mid \boldsymbol{y}\right) \propto \det\left[\left(\sum_{i=1}^{\ell} n_i\left(1-\lambda_i\right)\boldsymbol{x_i}\boldsymbol{x_i}'\right)^{-1}\right]^{1/2} \\ \times\left(\frac{1}{\sigma^2}\right)^{\frac{n-p}{2}+1}\prod_{i=1}^{\ell}\left(1-\lambda_i\right)^{\frac{1}{2}} \times \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{\ell}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y_i}\right)^2\right\} \\ \times \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{\ell} n_i\left(1-\lambda_i\right)\left(\bar{y_i}-\boldsymbol{x_i}'\hat{\boldsymbol{\beta}}\right)^2\right\}.\end{aligned} \tag{101}$$

From here we can see that:

$$\sigma^2 \mid \rho,\boldsymbol{y} \sim \text{InvGamma}\left(\frac{n-p}{2},\frac{\sum_{i=1}^{\ell}\left[n_i\left(1-\lambda_i\right)\left(\bar{y_i}-\boldsymbol{x_i}'\hat{\boldsymbol{\beta}}\right)^2+\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y_i}\right)^2\right]}{2}\right). \tag{102}$$

Therefore, after integrating out σ^2 we are finally left with the nonstandard posterior density of ρ:

$$
\begin{aligned}
\pi\left(\rho \mid \boldsymbol{y}\right) \propto \det\left[\left(\sum_{i=1}^{\ell} n_i\left(1-\lambda_i\right)\boldsymbol{x_i}\boldsymbol{x_i}'\right)^{-1}\right]^{1/2} & \prod_{i=1}^{\ell}\left(1-\lambda_i\right)^{\frac{1}{2}} \\
\times\left[\sum_{i=1}^{\ell} n_i\left(1-\lambda_i\right)\left(\bar{y_i}-\boldsymbol{x_i}'\hat{\boldsymbol{\beta}}\right)^2+\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y_i}\right)^2\right]^{\frac{-n+p}{2}} & .
\end{aligned}
\tag{103}
$$

We are now able to directly obtain a sample from the joint posterior density by beginning with using the grid method to sample ρ. Then sampling σ^2, $\boldsymbol{\beta}$, and $\boldsymbol{\nu}$ is straight forward since each of these parameters has a standard form. Based on our samples from the posterior density and the observed values of $\boldsymbol{y_i}$, we make inference for the finite population mean $\bar{Y_i}$, using the model:

$$
\bar{Y_i} \mid \boldsymbol{y_i} \overset{\text{ind}}{\sim} \text{Normal}\left(f_i\bar{y_i}+\left(1-f_i\right)\left[\boldsymbol{x_i}'\boldsymbol{\beta}+\nu_i\right], \left(1-f_i\right)\frac{\sigma^2}{N_i}\right). \tag{104}
$$

We explore the performance of this model in Section 3.3 with an application using BMI data.

Inlcuding Survey Weights in BHF Model

We also can include survey weights in the BHF model using the same adjusted and trimmed survey weights described in Section 3.2.2. The model with weights included is,

$$
\begin{gathered}
y_{ij} \mid \boldsymbol{\nu}, \boldsymbol{\beta}, \sigma^2 \overset{\text{ind}}{\sim} \text{Normal}\left(\boldsymbol{x_i}'\boldsymbol{\beta}+\nu_i, \frac{\sigma^2}{a_{ij}^*}\right), \\
\boldsymbol{\nu} \mid \rho, \sigma^2 \overset{\text{ind}}{\sim} \text{Normal}\left(0, \frac{\rho}{1-\rho}\sigma^2\right), \\
\pi\left(\boldsymbol{\beta}, \sigma^2, \rho\right) \propto \frac{1}{\sigma^2}, \quad \sigma^2>0, \quad 0<\rho<1, \quad \boldsymbol{\beta}\in\mathbb{R}^p, \\
j=1,\ldots,n_i, \quad i=1,\ldots\ell.
\end{gathered}
\tag{105}
$$

We use the same logic for obtaining a sample from this model with the adjusted weights as we used above in the model without weights. Making population predictions differs, because we obtain population predictions by:

$$
\overline{Y_i} \mid \boldsymbol{x_i}, \boldsymbol{\beta}, \nu_i, \sigma^2 \sim \text{Normal}\left(\boldsymbol{x_i}'\boldsymbol{\beta}+\nu_i, \frac{\sigma^2}{N_i}\right), \qquad i=1,\ldots,\ell, \tag{106}
$$

where $N_i = \sum\limits_{j=1}^{n_i} v_{ij}$ represents the Horvitz-Thompson estimator of population size for each strata $i=1,\ldots,\ell$.

Chapter 4

Spatial CAR Model with the Stick-Breaking Prior and a Binary Response Variable

4.1 Introduction

In this chapter, we continue to make inference about the study variable without explicitly assuming the relationship between the response and covariates. We also want to limit the extent of global pooling in our model predictions, since we want the unique characteristics of the strata in our data to gain strength from similar strata via spatial, heterogeneous, and cluster components (Tang, Ghosh, Ha, and Sedransk 2018 and Jo, Nandram, and Kim 2021). We build a model that includes these three components to make Bayesian predictive inference for a finite population proportion. We begin by presenting a spatial model with a binary response variable. Next, we include a heterogeneity parameter in the binary spatial model to account for the differences between strata. Ultimately, we end with the model we are most interested in that contains spatial, heterogeneous, and cluster components by incorporating the stick-breaking prior. By considering each unique combination of the covariates in the population as an individual stratum, we avoid directly including the covariates in the model. Then we use small area estimation techniques to make inference about each subset of the population based on its underlying covariates. Finally, we can estimate the overall finite population proportion by pooling predictions of the strata together.

We use spatial modeling techniques via the conditional autoregressive (CAR) model (Chung and Datta 2022). We include the spatial effects by creating the incidence matrix (or adjacency matrix) via the Mahalanobis distance between the covariates for each stratum. By enabling strata to have neighbors, we have seen neighborhoods pool together without all strata pooling together.

We incorporate a clustering component in our model using the stick-breaking prior to cluster the strata (Ishwaran and James 2001). This stick-breaking prior is a finite approximation of the Dirichlet-process prior, which allows the data to determine the number of appropriate clusters. This stick-breaking prior is ideal for when we do not know the distinct number of clusters present in the data before sampling. The spatial relationships are defined by the covariates before sampling, and the clustering component is determined by the data in the algorithm. While strata in the same spatial neighborhood will gain strength from each other, the strata in the same cluster will share cluster-specific parameters.

For the remainder of the chapter, we describe the structure of the we input data used in our models in Section 4.2. In Section 4.3 we discuss the methodology of the spatial and cluster models and show how to perform Bayesian predictive inference. Specifically, Section 4.3.1 presents the spatial model with a binary response variable. In Section 4.3.2, we include a heterogeneity parameter in the binary spatial model to account for the differences between strata. In Section 4.3.3, we incorporate the stick-breaking prior to obtain our model of interest that contains spatial, heterogeneous, and cluster components. Then in Section 4.4, an application using BMI data with each of the models is given, followed by a conclusion in Section 4.5. Appendix G provides the technical details for the Fay-Herriot model that we use as a baseline for comparison to our models.

4.2 Data Structure

In our models we perform Bayesian predictive inference for a finite population proportion, therefore our variable of interest, y_{ij}, $j = 1, \dots, n_i$, $i = 1, \dots, \ell$, is binary. The responses are grouped together based on their covariate values, where each unique combination of covariates is considered to be a stratum. Therefore, each $\boldsymbol{y_i}$ has a unique corresponding covariate combination denoted $\boldsymbol{x_i}$, where $\boldsymbol{y_i}$ is the aggregated vector of responses of length n_i for each stratum. The design matrix $\boldsymbol{X} = (\boldsymbol{x_i}')$ has dimension $\ell \times p$ where $(p-1)$ is the number of covariates in the data. The first column of $\boldsymbol{X}$ represents the intercept, and $\boldsymbol{x_i}'$ corresponds to the unique rows of $\boldsymbol{X}$. To mitigate computational difficulties, we do not model the data with the Bernoulli distribution and instead transform the binary response variable into a continuous proportion. We denote this proportion for each stratum as $\hat{\theta}_i$ that can be calculated as

$$\hat{\theta}_i = \frac{\sum\limits_{j=1}^{n_i} v_{ij}^* y_{ij}}{\sum\limits_{j=1}^{n_i} v_{ij}^*}, \qquad i = 1, \dots, \ell, \tag{107}$$

where y_{ij} are the binary responses and v_{ij}^* are the adjusted trimmed survey weights we describe below.

In addition we calculate,

$$\hat{\sigma}_i^2 = \frac{\hat{\theta}_i(1-\hat{\theta}_i)}{n_i^*}, \tag{108}$$

and $n_i^* = \sum_{j=1}^{\ell} v_{ij}^*$ are the corresponding effective sample sizes for each stratum $i = 1, \dots, \ell$. One possible issue in defining $\hat{\theta}_i$ in this way is that for any particular stratum, the y_{ij} could be all zeroes or all ones for some i. If y_{ij} are all zeroes, we set the corresponding $\hat{\theta}_i$ as smallest nonzero $\hat{\theta}_i$ value $i = 1, \dots, \ell$. Similarly, if y_{ij} are all ones, we set the corresponding $\hat{\theta}_i$ as largest nonzero $\hat{\theta}_i$ value $i = 1, \dots, \ell$.

Here we use the original survey weights, denoted v_{ij}, $j = 1, \dots, n_i$, $i = 1, \dots \ell$, to calculate the effective sample size and the adjusted trimmed weights v^*_{ij}. In the BMI example in Section 4.4 our original data have outliers, so we use Winsorization to deal with outliers by trimming the survey weights (Yang, Nandram, and Choi 2023). Outliers here are defined as observed survey weights greater than $v_0 = Q_3 + 1.5\,(Q_3 - Q_1)$, where Q_1 is the first quartile and Q_3 is the third quartile. Let v^* denote weights after trimming,

$$v^*_{ij} = \begin{cases} v_0, & v_{ij} \geq v_0 \\ rv_{ij}, & v_{ij} < v_0 \end{cases}, \tag{109}$$

where r is a rescaling parameter such that $\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v^*_{ij} = \sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v_{ij} = \hat{N}$ and $\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v_{ij} = \hat{N}$ is the Horvitz-Thompson estimator of population size.

We calculate the effective sample size, $\hat{n}^*$:

$$\hat{n}^* = \frac{\left(\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v^*_{ij}\right)^2}{\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v^{*2}_{ij}}. \tag{110}$$

Then we obtain the adjusted and trimmed weights v^*_{ij},

$$v^*_{ij} = \hat{n}^* \frac{v^*_{ij}}{\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{n_i} v^*_{ij}}. \tag{111}$$

These adjusted and trimmed weights v^*_{ij} are used in calculating $\hat{\theta}_i$, $i = 1, \dots, \ell$.

4.3 Methodology

All of our models begin with an adapted Fay-Herriot model (Rao and Molina 2015; Nandram, Erciulescu, and Cruze 2019). The Bayesian Fay-Herriot model without covariates is,

$$\begin{gathered} \hat{\theta}_i \mid \boldsymbol{\mu} \overset{\text{ind}}{\sim} \text{Normal}\left(\mu_i, \hat{\sigma}^2_i\right), \quad i = 1, \dots, \ell, \\ \mu_i \mid \theta, \delta^2 \overset{\text{ind}}{\sim} \text{Normal}\left(\theta, \delta^2\right), \quad i = 1, \dots, \ell, \\ \pi\left(\theta, \delta^2\right) \propto \frac{1}{(1+\delta^2)^2}, \end{gathered} \tag{112}$$

where $\pi\left(\delta^2\right)$ is a proper shrinkage prior. We do not include covariates directly in our models to improve the robustness and allow for more applications. Typically, in this model it is assumed that $\hat{\sigma}_i^2$, $i = 1, \ldots, \ell$ are fixed. We construct our model by letting $G = \left\{\prod_{i=1}^{\ell} \left(\hat{\sigma}_i^2\right)^{n_i}\right\}^{1/n}$, the weighted geometric mean where n_i is the sample size of the i^{th} stratum and $n = \sum_{i=1}^{\ell} n_i$ is the total sample size. Also let $\kappa_i = \frac{G}{\hat{\sigma}_i^2}$, $i = 1, \ldots, \ell$, which are also fixed.

We adjust the standard Fay-Herriot model by,

$$\begin{gathered}
\hat{\theta}_i \mid \boldsymbol{\mu}, \sigma^2 \overset{\text{ind}}{\sim} \text{Normal}\left(\mu_i, \frac{\sigma^2}{\kappa_i}\right), \quad i = 1, \ldots, \ell, \\
\mu_i \mid \theta, \sigma^2 \overset{\text{ind}}{\sim} \text{Normal}\left(\theta, \sigma^2\right), \quad i = 1, \ldots, \ell, \\
\pi\left(\theta, \sigma^2\right) \propto \frac{1}{\sigma^2}.
\end{gathered} \tag{113}$$

By writing the Fay-Herriot model in this way, we no longer require a proper prior for σ^2. Now that σ^2 is in the likelihood and therefore directly connected to the data, it is unlikely that σ^2 would suffer from impropriety. Having σ^2 connected to the data improves identifiability and allows for σ^2 to have a known conditional posterior density following an inverse-gamma distribution.

In the remainder of this section, we show various spatial and clustering models. First, in Section 4.3.1 we present the binary spatial CAR model. Section 4.3.2 illustrates how the heterogeneity parameter can be included in the spatial model to account for the differences between strata. Section 4.3.3 incorporates the stick-breaking prior to perform clustering of the strata in the spatial model. Lastly, Section 4.3.4 explains Bayesian predictive inference for a finite population proportion. Appendix G contains the technical details of the Fay-Herriot model with covariates included, which we use as a baseline for comparison to our models.

4.3.1 Binary Spatial Model

First, we begin with a simple binary response model with the spatial component only. We include the spatial effects using the CAR model described in Section 3.2.1.1. We create the symmetric incidence matrix, $\boldsymbol{W}$ of size $\ell \times \ell$, via the Mahalanobis distance between $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$, $i = 1, \ldots, \ell$, $j = 1, \ldots, \ell$, and $i \neq j$. The Mahalanobis distance is defined as:

$$d_{ij} = \sqrt{\left(\boldsymbol{x_i} - \boldsymbol{x_j}\right)' \boldsymbol{S}^{-1} \left(\boldsymbol{x_i} - \boldsymbol{x_j}\right)}, \tag{114}$$

where $\boldsymbol{S}$ is the covariance matrix of $\boldsymbol{X}$, and $d_{ii} = 0$. We define $\boldsymbol{W}$ by letting $w_{ij} = 1$ if $d_{ij} \leq d_0$ and $w_{ij} = 0$ if $d_{ij} > d_0$ with zeroes on the diagonal (i.e., $w_{ii} = 0$). The value d_0 yields the $\boldsymbol{W}$ matrix that maximizes Moran's I, which is defined as:

$$I = \frac{\ell}{w_{\cdot\cdot}} \frac{\sum_i \sum_j w_{ij} \left(\bar{y_i} - \bar{y}\right) \left(\bar{y_j} - \bar{y}\right)}{\sum_i \left(\bar{y_i} - \bar{y}\right)^2}, \tag{115}$$

where $\bar{y_i} = \sum_{j=1}^{n_i} y_{ij}/n_i$ is the response variable; $\bar{y} = \sum_{i=1}^{\ell} \bar{y}_i/\ell$ is the overall sample mean response; w_{ij} corresponds to the elements of $\boldsymbol{W}$; and $w_{\cdot\cdot} = \sum_i \sum_j w_{ij}$.

This model serves as our base model, which we will expand upon in the later sections to include the heterogeneous and clustering components. The binary response spatial model is,

$$\begin{gathered}
\hat{\theta}_i \mid \boldsymbol{\mu}, \sigma^2 \sim \text{Normal}\left(\mu_i, \frac{\sigma^2}{\kappa_i}\right), \\
\boldsymbol{\mu} \mid \theta, \sigma^2, \psi \sim \text{Normal}\left(\theta \mathbf{1}, \sigma^2 \left(\boldsymbol{R} - \psi \boldsymbol{W}\right)^{-1}\right), \\
\pi\left(\theta, \sigma^2\right) \propto \frac{1}{\sigma^2}, \\
\psi \sim \text{Uniform}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_\ell}\right), \\
\frac{1}{\lambda_1} \leq \psi \leq \frac{1}{\lambda_\ell}, \quad -\infty < \theta < \infty, \quad \sigma^2 > 0, \\
i = 1, \ldots, \ell,
\end{gathered} \tag{116}$$

where ℓ in this case represents the total number of possible covariate combinations, which are considered to be the strata we use to perform small area estimation. Also, λ_1 is the minimum eigenvalue of $\boldsymbol{R}^{-1}\boldsymbol{W}$ and λ_ℓ is the maximum eigenvalue of $\boldsymbol{R}^{-1}\boldsymbol{W}$, and since $\sum_{i=1}^{\ell} w_{ii} = 0$ this results in $\lambda_1 < 0 < \lambda_\ell$ (Chung and Datta 2022). Here, $(\boldsymbol{R} - \psi\boldsymbol{W})$ is guaranteed to be positive definite as long as ψ is in the range $\frac{1}{\lambda_1} \leq \psi \leq \frac{1}{\lambda_\ell}$. Now, we describe how to obtain samples from this model.

The conditional posterior distribution of $\boldsymbol{\mu}$ that we draw samples from is

$$\boldsymbol{\mu} \mid \theta, \sigma^2, \psi, \hat{\boldsymbol{\theta}} \sim \text{Normal}\left[\boldsymbol{\Delta_0}\hat{\boldsymbol{\theta}} + (\boldsymbol{I} - \boldsymbol{\Delta_0})\theta\mathbf{1}, \sigma^2(\boldsymbol{I} - \boldsymbol{\Delta_0})\left(\boldsymbol{R} - \psi\boldsymbol{W}\right)^{-1}\right], \tag{117}$$

where $\boldsymbol{\Delta_0} = \left(\text{diag}(\kappa_1, \ldots, \kappa_\ell) + (\boldsymbol{R} - \psi\boldsymbol{W})\right)^{-1} \text{diag}(\kappa_1, \ldots, \kappa_\ell)$.

The conditional posterior density for θ also follows a normal distribution that we obtain samples from,

$$\theta \mid \mu, \sigma^2, \psi, \hat{\boldsymbol{\theta}} \sim \text{Normal}\left(\frac{\sum_{i=1}^{\ell} r_{ii}\mu_i}{\sum_{i=1}^{\ell} r_{ii}}, \frac{\sigma^2}{(1-\psi)\sum_{i=1}^{\ell} r_{ii}}\right). \tag{118}$$

Next we obtain samples of σ^2 from the inverse-gamma distribution,

$$\sigma^2 \mid \mu, \theta, \psi, \hat{\boldsymbol{\theta}} \sim \text{InvGam}\left(\ell, Q/2\right), \tag{119}$$

where $Q = \left[\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\mu}\right)' \operatorname{diag}(\kappa_1, \ldots, \kappa_\ell) \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\mu}\right) + (\boldsymbol{\mu} - \theta\mathbf{1})'(\boldsymbol{R} - \psi\boldsymbol{W})(\boldsymbol{\mu} - \theta\mathbf{1})\right]$.

Finally, the nonstandard conditional posterior distribution for ψ is,

$$\pi\left(\psi \mid \mu, \theta, \sigma^2, \hat{\boldsymbol{\theta}}\right) \propto \det\left[(\boldsymbol{R} - \psi\boldsymbol{W})^{-1}\right]^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\mu} - \theta\mathbf{1})'(\boldsymbol{R} - \psi\boldsymbol{W})(\boldsymbol{\mu} - \theta\mathbf{1})\right), \tag{120}$$

which we sample by using the grid method.

In Section 4.3.4 we discuss how to perform Bayesian predictive inference for the finite population proportion after obtaining samples from this model.

4.3.2 Including Heterogeneity Parameter in Binary Spatial Model

In this section we present the first extension of the simple spatial model, as we are now including an additional heterogeneity parameter, $\boldsymbol{\eta}$. The spatial component is used to define relationships between similar strata, while the heterogeneity parameters account for the strata's differences. The binary response spatial model with $\boldsymbol{\eta}$ included is,

$$\begin{gathered}
\hat{\theta}_i \mid \boldsymbol{\mu}, \sigma^2 \sim \text{Normal}\left(\mu_i, \frac{\sigma^2}{\kappa_i}\right), \\
\boldsymbol{\mu} \mid \boldsymbol{\eta}, \sigma^2, \psi \sim \text{Normal}\left(\boldsymbol{\eta}, \sigma^2\left(\boldsymbol{R} - \psi\boldsymbol{W}\right)^{-1}\right), \\
\boldsymbol{\eta} \mid \theta, \sigma^2 \sim \text{Normal}\left(\theta\mathbf{1}, \sigma^2\boldsymbol{I}\right), \\
\pi\left(\theta, \sigma^2\right) \propto \frac{1}{\sigma^2}, \\
\psi \sim \text{Uniform}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_\ell}\right), \\
\frac{1}{\lambda_1} \leq \psi \leq \frac{1}{\lambda_\ell}, \quad -\infty < \theta < \infty, \quad \sigma^2 > 0, \\
i = 1, \ldots, \ell,
\end{gathered} \tag{121}$$

where ℓ in represents the total number of possible covariate combinations, which are considered to be the individual strata. In this section we describe how to obtain samples from this model.

The conditional posterior distribution of $\boldsymbol{\mu}$ that we draw samples from is,

$$\boldsymbol{\mu} \mid \boldsymbol{\eta}, \sigma^2, \psi, \hat{\boldsymbol{\theta}} \sim \text{Normal}\left[\boldsymbol{\Delta_1}\hat{\boldsymbol{\theta}} + (\boldsymbol{I} - \boldsymbol{\Delta_1})\boldsymbol{\eta}, \sigma^2\left(\boldsymbol{I} - \boldsymbol{\Delta_1}\right)\left(\boldsymbol{R} - \psi\boldsymbol{W}\right)^{-1}\right], \tag{122}$$

where $\boldsymbol{\Delta_1} = (\operatorname{diag}(\kappa_1, \ldots, \kappa_\ell) + (\boldsymbol{R} - \psi\boldsymbol{W}))^{-1} \operatorname{diag}(\kappa_1, \ldots, \kappa_\ell)$.

Then, the conditional posterior distribution for $\boldsymbol{\eta}$ also follows a normal distribution,

$$\boldsymbol{\eta} \mid \boldsymbol{\mu}, \theta, \sigma^2, \psi, \hat{\boldsymbol{\theta}} \sim \text{Normal}\Bigg[\boldsymbol{\Delta_2}\boldsymbol{\mu} + (\boldsymbol{I} - \boldsymbol{\Delta_2})\theta\mathbf{1}, \sigma^2\left(\boldsymbol{I} - \boldsymbol{\Delta_2}\right)\Bigg], \tag{123}$$

where $\boldsymbol{\Delta_2} = ((\boldsymbol{R} - \psi\boldsymbol{W}) + \boldsymbol{I})^{-1}(\boldsymbol{R} - \psi\boldsymbol{W})$.

Next, we sample θ from the normal distribution,

$$\theta \mid \boldsymbol{\eta}, \sigma^2, \hat{\boldsymbol{\theta}} \sim \text{Normal}\left(\frac{\sum_{i=1}^{\ell}\eta_i}{\ell}, \frac{\sigma^2}{\ell}\right). \tag{124}$$

The conditional posterior distribution for σ^2 follows the inverse-gamma distribution,

$$\sigma^2 \mid \boldsymbol{\mu}, \boldsymbol{\eta}, \theta, \psi, \hat{\boldsymbol{\theta}} \sim \text{InvGam}\left(\frac{3\ell}{2}, Q/2\right). \tag{125}$$

where

$$Q = \left[\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\mu}\right)' \text{diag}(\kappa_1, \ldots, \kappa_\ell)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\mu}\right) + (\boldsymbol{\mu} - \boldsymbol{\eta})'(\boldsymbol{R} - \psi\boldsymbol{W})(\boldsymbol{\mu} - \boldsymbol{\eta}) + (\boldsymbol{\eta} - \theta\mathbf{1})'(\boldsymbol{\eta} - \theta\mathbf{1})\right].$$

Lastly, the nonstandard conditional posterior density for ψ is,

$$\pi\left(\psi \mid \boldsymbol{\mu}, \boldsymbol{\eta}, \sigma^2, \hat{\boldsymbol{\theta}}\right) \propto \det\left[(\boldsymbol{R} - \psi\boldsymbol{W})^{-1}\right]^{-1/2} \exp\left(\frac{-1}{2\sigma^2}(\boldsymbol{\mu} - \boldsymbol{\eta})'(\boldsymbol{R} - \psi\boldsymbol{W})(\boldsymbol{\mu} - \boldsymbol{\eta})\right), \tag{126}$$

which we obtain samples from using the grid method.

In Section 4.3.4 we discuss how to perform Bayesian predictive inference for the finite population proportion after obtaining samples from this model.

4.3.3 Including the Stick-Breaking Algorithm for Clustering and Heterogeneity in Binary Spatial Model

In this final form of this model, we use the stick-breaking prior to sample $\boldsymbol{\eta}$. This is the model we are the most interested in as it includes the spatial, cluster, and heterogeneous components. By including the stick-breaking prior we are clustering strata together when we do not know the distinct number of clusters present in the data before sampling. The clusters are determined by the data in the algorithm, unlike the spatial relationships that are defined by the covariates before sampling. Strata in the same cluster will share cluster-specific parameters. The binary response spatial model with the stick-breaking prior included is,

$$
\begin{aligned}
\hat{\theta}_i \mid \boldsymbol{\mu}, \sigma^2 &\sim \text{Normal}\left(\mu_i, \frac{\sigma^2}{\kappa_i}\right), \\
\boldsymbol{\mu} \mid \boldsymbol{\eta}, \sigma^2, \psi &\sim \text{Normal}\left(\boldsymbol{\eta}, \sigma^2\left(\boldsymbol{R} - \psi \boldsymbol{W}\right)^{-1}\right), \\
\eta_i \mid \boldsymbol{z}, \sigma^2 &\sim \sum_{s=1}^{\ell} p_s \cdot \text{Normal}\left(z_s, \sigma^2\right), \\
z_s \mid \theta, \rho, \sigma^2 &\sim \text{Normal}\left(\theta, \frac{\rho}{1-\rho}\sigma^2\right), \\
\pi\left(\theta, \sigma^2\right) &\propto \frac{1}{\sigma^2}, \\
\psi &\sim \text{Uniform}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_\ell}\right), \\
\frac{1}{\lambda_1} \leq \psi \leq \frac{1}{\lambda_\ell}, &\quad -\infty < \theta < \infty, \quad \sigma^2 > 0, \\
i &= 1, \ldots, \ell,
\end{aligned} \tag{127}
$$

where ℓ in this case represents the total number of possible covariate combinations, which are considered to be the individual strata.

Also p_s are the stick-breaking weights defined as:

$$
\begin{aligned}
p_1 &= \nu_1, \\
p_s &= \nu_s \prod_{k<s}\left(1 - \nu_k\right), \quad s = 2, \ldots, \ell - 1, \\
p_\ell &= \prod_{s=1}^{\ell-1}\left(1 - \nu_s\right),
\end{aligned} \tag{128}
$$

with priors $\nu_s \overset{\text{ind}}{\sim} \text{Beta}\left(1, \frac{1-\gamma}{\gamma}\right)$ and $\pi(\gamma) \propto 1$, $0 < \gamma < 1$.

We first tried to use the Pitman-Yor prior described in Ishwaran and James (2001), namely $\nu_s \overset{\text{ind}}{\sim} \text{Beta}\left(1 - \delta, \frac{1-\gamma}{\gamma} + s \cdot \delta\right)$ and $\pi(\delta, \gamma) = 1$, $0 < \delta, \gamma < 1$. However, there are some computational difficulties with this prior if δ approaches 1, then the prior becomes unstable and not well-defined. Therefore, to improve the behavior of sampling ν_s, $s = 1, \ldots, \ell$, we use the beta prior typically used in the Dirichlet-process prior $\nu_s \overset{\text{ind}}{\sim} \text{Beta}\left(1, \alpha\right)$, $\alpha > 0$ where $\alpha = \frac{1-\gamma}{\gamma}$, $0 < \gamma < 1$ (Sethuraman 1994; Kalli, Griffin, and Walker 2011). Here α is a concentration parameter such that larger values of α will lead to more clusters, and smaller values of α will lead to fewer unique clusters. By letting $\alpha = \frac{1-\gamma}{\gamma}$, $0 < \gamma < 1$, we are able to sample γ using a grid method, which we could not do for $\alpha > 0$. To allow for more clusters, we make an adjustment by setting $\alpha > 1$ and therefore $0 < \gamma < 1/2$.

The stick-breaking weights are used to determine the cluster each stratum belongs in. We use d_i, $i = 1 \ldots, \ell$, as classification variables to identify the z_s, $s = 1, \ldots, m$, associated with each η_i, $i = 1 \ldots, \ell$, where m is the number of unique clusters and $m \leq \ell$. Therefore, the d_i identify the cluster each η_i belongs in, and the η_i in each cluster will share the same z_s. The stick-breaking weights are a component of the probability used to sample each d_i, $i = 1, \ldots, \ell$, as seen in the second block of our algorithm. Once we have the d_i, $1 = 1, \ldots, \ell$, we can fit the model with the blocked Gibbs sampler (Ishwaran and James 2001). We must initialize the blocked Gibbs sampler with values for d_i in order to begin the algorithm.

By utilizing the blocked Gibbs sampler it is straightforward to obtain the conditional posterior distributions used to obtain samples of this model's parameters. For the first block of our algorithm we sample $\boldsymbol{\mu}$, $\boldsymbol{\eta}$, and ψ. Let $\boldsymbol{d} = \{d_1, \ldots, d_\ell\}$ denote the current set of cluster labels taking values $\{1, \ldots, m\}$, given $\boldsymbol{d}$ we have:

$$\begin{gathered}
\hat{\theta}_i \mid \boldsymbol{\mu}, \sigma^2 \sim \text{Normal}\left(\mu_i, \frac{\sigma^2}{\kappa_i}\right), \\
\boldsymbol{\mu} \mid \boldsymbol{\eta}, \sigma^2, \psi \sim \text{Normal}\left(\boldsymbol{\eta}, \sigma^2 \left(\boldsymbol{R} - \psi \boldsymbol{W}\right)^{-1}\right), \\
\eta_i \mid d_i = s, z_s, \sigma^2 \sim \text{Normal}\left(z_s, \sigma^2\right), \\
\psi \sim \text{Uniform}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_\ell}\right), \\
i = 1, \ldots, \ell, \quad s = 1, \ldots, m.
\end{gathered} \tag{129}$$

From this first block of our model we can obtain the conditional posterior densities of $\boldsymbol{\mu}$, $\boldsymbol{\eta}$, and ψ. We sample $\boldsymbol{\mu}$ from the conditional posterior density,

$$\boldsymbol{\mu} \mid \boldsymbol{\eta}, \sigma^2, \psi, \hat{\boldsymbol{\theta}} \sim \text{Normal}\left[\Delta_1 \hat{\boldsymbol{\theta}} + (\boldsymbol{I} - \Delta_1)\boldsymbol{\eta}, \sigma^2 (\boldsymbol{I} - \Delta_1)(\boldsymbol{R} - \psi \boldsymbol{W})^{-1}\right], \tag{130}$$

where $\boldsymbol{\Delta_1} = (\text{diag}(\kappa_1, \ldots, \kappa_\ell) + (\boldsymbol{R} - \psi \boldsymbol{W}))^{-1} \text{diag}(\kappa_1, \ldots, \kappa_\ell)$.

Then, we sample $\boldsymbol{\eta}$ from the conditional posterior density,

$$\boldsymbol{\eta} \mid \boldsymbol{d}, \boldsymbol{z}, \sigma^2, \hat{\boldsymbol{\theta}} \sim \text{Normal}\left[\Delta_2 \boldsymbol{\mu} + (\boldsymbol{I} - \Delta_2)(\boldsymbol{A}\boldsymbol{z}), \sigma^2 (\boldsymbol{I} - \Delta_2)\right], \tag{131}$$

where $\boldsymbol{\Delta_2} = ((\boldsymbol{R} - \psi \boldsymbol{W}) + \boldsymbol{I})^{-1} (\boldsymbol{R} - \psi \boldsymbol{W})$ and $\boldsymbol{A}$ is an $\ell \times m$ mapping matrix that assigns the corresponding z_s to each η_i. Therefore, $\boldsymbol{A}$ is a sparse matrix of mostly zeroes, except $\boldsymbol{A}[i, d_i] = 1$, $i = 1, \ldots, \ell$.

For example, assume we have five η_i, $i = 1, \ldots, 5$, and the d_i assign the η_i to two clusters as given in the table below.

i	d_i	z_s
1	2	z_2
2	1	z_1
3	2	z_2
4	2	z_2
5	1	z_1

Therefore, there are two values of z_s, $s = 1, 2$ and $m = 2$. Then, the 5×2 matrix $\boldsymbol{A}$ would be,

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix},$$

such that,

$$\boldsymbol{A}\boldsymbol{z} = \boldsymbol{A} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} z_2 \\ z_1 \\ z_2 \\ z_2 \\ z_1 \end{bmatrix}.$$

The nonstandard conditional posterior density for ψ is,

$$\pi\left(\psi \mid \mu, \boldsymbol{\eta}, \sigma^2, \hat{\boldsymbol{\theta}}\right) \propto \det\left[(\boldsymbol{R} - \psi \boldsymbol{W})^{-1}\right]^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\mu} - \boldsymbol{\eta})'(\boldsymbol{R} - \psi\boldsymbol{W})(\boldsymbol{\mu} - \boldsymbol{\eta})\right), \tag{132}$$

recall that $1/\lambda_1 \leq \psi \leq 1/\lambda_\ell$, so we sample ψ using the grid method.

In the second block of the Gibbs sampler we will obtain the d_i, $i = 1, \ldots, \ell$. We draw the d_i using probability sampling with probability equal to

$$P(d_i = s \mid \boldsymbol{\eta}, \boldsymbol{z}, \sigma^2, \hat{\boldsymbol{\theta}}) = \frac{p_s \text{Normal}_{\eta_i}(z_s, \sigma^2)}{\sum\limits_{s=1}^{\ell} p_s \text{Normal}_{\eta_i}(z_s, \sigma^2)}, \quad s = 1, \ldots, \ell, \tag{133}$$

therefore, d_i are indicators that track the cluster each η_i belongs to. There are m unique values of d_i, $i = 1, \ldots, \ell$, which correspond to m unique clusters. Here the p_s, $s = 1, \ldots, \ell$, are the stick-breaking weights defined in (124).

In the third block of the Gibbs algorithm we obtain samples of $\boldsymbol{z}$, θ, and ρ. Using the

d_i that indicate the cluster each η_i belongs to, we can arrange the η_i as: ϕ_{sj}, $s = 1, \ldots, m$, $j = 1, \ldots, n_s$. Here, m is the number of clusters (i.e. the number of unique d_i values) and n_s is the number of η_i in each cluster. Note that m is subject to change at each iteration of the sampler. Writing in this way, we have,

$$\begin{aligned} \phi_{sj} \mid \boldsymbol{z}, \theta, \sigma^2 &\sim \text{Normal}(z_s, \sigma^2), \\ z_s \mid \theta, \rho, \sigma^2 &\sim \text{Normal}\left(\theta, \frac{\rho}{1-\rho}\sigma^2\right), \\ \pi(\theta, \rho) &= 1, \end{aligned} \tag{134}$$

which is simply the Scott-Smith model (Scott and Smith 1969).

We obtain samples of $\boldsymbol{z}$ from the conditional posterior distribution,

$$z_s \mid \rho, \boldsymbol{\phi}, \theta, \sigma^2, \hat{\boldsymbol{\theta}} \sim \text{Normal}\left(\delta_s \bar{\phi}_s + (1-\delta_s)\theta, \frac{\delta_s \sigma^2}{n_s}\right), \tag{135}$$

$s = 1, \ldots, m$ where m is the number of clusters, $\bar{\phi}_s = 1/n_s \sum_{j=1}^{n_s} \phi_{sj}$, and $\delta_s = \frac{n_s\rho}{(n_s-1)\rho+1}$. For the remaining z_s from $s = m+1, \ldots, \ell$ we sample from the prior,

$$z_s \mid \theta, \rho, \sigma^2 \sim \text{Normal}\left(\theta, \frac{\rho}{1-\rho}\sigma^2\right). \tag{136}$$

Note that we only use the samples of z_s from $s = m+1, \ldots, \ell$ in block two of the algorithm when we sample the d_i, $i = 1, \ldots, \ell$. For all other blocks of the Gibbs sampler, $\boldsymbol{z}$ is a vector of length m that corresponds to the z_s, $s = 1 \ldots, m$ sampled from the conditional posterior distribution (131) described above.

The conditional posterior density of θ is,

$$\theta \mid \rho, \boldsymbol{z}, \sigma^2, \hat{\boldsymbol{\theta}} \sim \text{Normal}\left(\frac{\sum\limits_{s=1}^{m} z_s}{m}, \frac{\rho}{1-\rho}\frac{\sigma^2}{m}\right). \tag{137}$$

The last conditional posterior density in block three is the nonstandard distribution for ρ,

$$\pi(\rho \mid \sigma^2, \boldsymbol{z}, \hat{\boldsymbol{\theta}}) \propto \left(\frac{1-\rho}{\rho}\right)^{\frac{m}{2}} \exp\left[-\frac{1}{2}\left(\frac{1-\rho}{\rho\sigma^2}\right)\sum_{s=1}^{m}(z_s-\theta)^2\right], \tag{138}$$

and we use the grid method to obtain samples of ρ since $0 < \rho < 1$.

In the fourth block, we sample the parameters that give us the stick-breaking weights, including $\boldsymbol{\nu}$ and γ. The conditional posterior density of $\boldsymbol{\nu}$ is,

$$\nu_s \mid \gamma, \boldsymbol{d}, \hat{\boldsymbol{\theta}} \sim \text{Beta}\left(1 + n_s, \frac{1-\gamma}{\gamma} + \sum_s I(d_i > s)\right), \quad s = 1, \ldots, m, \tag{139}$$

where $I(d_i > s)$ represents the indicator function, $I(d_i > s) = 1$ if $d_i > s$, otherwise $I(d_i > s) = 0$. For the remaining ν_s from $s = m+1, \ldots, \ell$, we sample from the prior,

$$\nu_s \mid \gamma \sim \text{Beta}\left(1, \frac{1-\gamma}{\gamma}\right), \quad s = m+1, \ldots, \ell. \tag{140}$$

We must sample the ν_s, $s = 1, \ldots, \ell$, to use in construction of the stick-breaking weights in the second block of this algorithm, as this allows for the number of clusters to vary each iteration.

Next, γ has the nonstandard conditional posterior density,

$$\pi(\gamma \mid \boldsymbol{\nu}, \hat{\boldsymbol{\theta}}) \propto \prod_{s=1}^{m} \frac{\Gamma\left(1 + \frac{1-\gamma}{\gamma}\right)}{\Gamma\left(\frac{1-\gamma}{\gamma}\right)} (1 - \nu_s)^{\frac{1-\gamma}{\gamma} - 1}, \tag{141}$$

that we sample using the grid method, since $0 < \gamma < 1$.

Finally, in the fifth block of the Gibbs algorithm, we sample σ^2 from the inverse-gamma distribution,

$$\sigma^2 \mid \mu, \theta, \psi, \boldsymbol{\eta}, \boldsymbol{z}, \rho, \hat{\boldsymbol{\theta}} \sim \text{InvGam}\left(\frac{3\ell + m}{2}, \frac{Q}{2}\right), \tag{142}$$

where

$$\begin{aligned} Q = {} & \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\mu}\right)' \boldsymbol{D}^{-1} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\mu}\right) + (\boldsymbol{\mu} - \boldsymbol{\eta})'(\boldsymbol{R} - \psi \boldsymbol{W})(\boldsymbol{\mu} - \boldsymbol{\eta}) + (\boldsymbol{\eta} - \boldsymbol{A}\boldsymbol{z})'(\boldsymbol{\eta} - \boldsymbol{A}\boldsymbol{z}) \\ & + (\boldsymbol{z} - \theta \mathbf{1})' \left(\frac{1-\rho}{\rho}\right) (\boldsymbol{z} - \theta \mathbf{1}). \end{aligned}$$

We intentionally sample σ^2 in the last block of the Gibbs sampler, since σ^2 is present in every level of the hierarchical model we want all of the involved parameters to be updated prior to sampling σ^2.

In Section 4.3.4 we discuss how to perform Bayesian predictive inference for the finite population proportion after obtaining samples from this model.

4.3.4 Bayesian Predictive Inference of a Finite Population Proportion

Once we have fit the three aforementioned models and obtained samples of all the parameters, we make Bayesian predictive inference for the finite population proportion. We begin by sampling $\boldsymbol{\mu}$ from each model's conditional posterior density now truncated in $[0, 1]$. We must have $\boldsymbol{\mu}$ contained in $[0, 1]$ since we need to predict the population values for the original binary response variable Y_{ij}, $i = 1, \ldots, \ell$, $j = 1, \ldots, N_i$, from the Bernoulli distribution with probability μ_i. We sample $\boldsymbol{\mu}$ from the unconstrained normal distribution within the Gibbs sampler, then in the output analysis we sample the $\boldsymbol{\mu}$ again from the normal distribution truncated in $[0, 1]$.

After fitting the base model, we obtain $\boldsymbol{\mu}$ from,

$$\boldsymbol{\mu} \mid \theta, \sigma^2, \psi, \hat{\boldsymbol{\theta}} \sim \text{Truncated Normal}\left[\boldsymbol{\Delta_0}\hat{\boldsymbol{\theta}} + (\boldsymbol{I} - \boldsymbol{\Delta_0})\theta\mathbf{1}, \sigma^2(\boldsymbol{I} - \boldsymbol{\Delta_0})\left(\boldsymbol{R} - \psi\boldsymbol{W}\right)^{-1}\right], \quad (143)$$

using the samples of θ, σ^2, and ψ obtained in the Gibbs sampler with the normal distribution truncated in $[0, 1]$, and $\boldsymbol{\Delta_0} = (\text{diag}(\kappa_1, \ldots, \kappa_\ell) + (\boldsymbol{R} - \psi\boldsymbol{W}))^{-1}\,\text{diag}(\kappa_1, \ldots, \kappa_\ell)$.

Next, in the output analysis of the first and second extensions of the base model, the conditional posterior distribution of $\boldsymbol{\mu}$ is

$$\boldsymbol{\mu} \mid \boldsymbol{\eta}, \sigma^2, \psi, \hat{\boldsymbol{\theta}} \sim \text{Truncated Normal}\left[\Delta_1\hat{\boldsymbol{\theta}} + (\boldsymbol{I} - \Delta_1)\boldsymbol{\eta}, \sigma^2(\boldsymbol{I} - \Delta_1)(\boldsymbol{R} - \psi\boldsymbol{W})^{-1}\right], \quad (144)$$

using the samples of η, σ^2, and ψ obtained from the corresponding Gibbs sampler with $\boldsymbol{\Delta_1} = (\text{diag}(\kappa_1, \ldots, \kappa_\ell) + (\boldsymbol{R} - \psi\boldsymbol{W}))^{-1}\,\text{diag}(\kappa_1, \ldots, \kappa_\ell)$, and the normal distribution truncated in $[0, 1]$.

Then, using the respective $\boldsymbol{\mu}$ for each model described above, we can predict the binary responses for each stratum's population by,

$$Y_i \mid \boldsymbol{\mu}, \hat{\boldsymbol{\theta}} \sim \text{Binomial}\left(N_i, \mu_i\right), \quad (145)$$

where $N_i = \sum_{j=1}^{n_i} v_{ij}$ is the Horvitz-Thompson estimator of population size for each stratum $i = 1, \ldots, \ell$, and v_{ij} are the original survey weights. Finally, the overall population proportion is,

$$P = \frac{\sum\limits_{i=1}^{\ell} Y_i}{N}, \quad (146)$$

where $N = \sum_{i=1}^{\ell} N_i$ is the Horvitz-Thompson estimator of total population size.

We also fit the Fay-Herriot model with covariates as a baseline for comparison to our models. The full technical details of the model are in Appendix G, and we describe how to make Bayesian predictive inference for the model here. The Fay-Herriot model does not require the Gibbs sampler, so instead we only have to sample θ_i, $i = 1, \ldots, \ell$, once. The conditional posterior distribution is,

$$\theta_i \mid \boldsymbol{\beta}, \delta^2, \hat{\boldsymbol{\theta}} \sim \text{Truncated Normal}\left[\lambda_i \hat{\theta}_i + (1-\lambda_i)\boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{\prime}}\boldsymbol{\beta}, (1-\lambda_i)\delta^2\right], \tag{147}$$

where $\lambda_i = \frac{\delta^2}{\hat{\sigma}_i^2+\delta^2}$, $i = 1, \ldots, \ell$ and the normal distribution truncated in $[0, 1]$. Then, using $\boldsymbol{\theta}$, we can predict the binary responses for each stratum's population by,

$$Y_i \mid \boldsymbol{\theta}, \hat{\boldsymbol{\theta}} \sim \text{Binomial}\left(N_i, \theta_i\right). \tag{148}$$

Now, we calculate the overall population proportion as seen in (143).

4.4 Application using BMI Data

The proportion of the population suffering from obesity is an important indicator when interested in studying the health of a finite population. We illustrate our spatial and cluster models using a probability sample of BMI data with 1,867 individuals from eight counties in California recorded in the Third National Health and Nutrition Examination Survey (NHANES III). These data have four variables including age, race, sex, and an obesity indicator. We are interested in predicting the proportion of the finite population that are obese. The values for sex are “male” or “female”, and the values for race are “white” or “non-white”. The obesity indicator is defined such that if an individual's BMI is greater than or equal to 30 kg/m^2 then the indicator's value is “obese”, otherwise the indicator's value is “not obese”. The age values range from 20 years old to 90 years old, and this variable is the only continuous variable included in the data. We bin the age variable into groups of two years, therefore the bins are 20-21 years old, 22-23 years old, and so on, in order to have a finite number of possible covariate combinations.

After aggregating over all possible age, race, sex, and obesity combinations there are 144 strata each with its own unique set of covariate values. However, in our sample of BMI data we have 12 strata with no observations and we assume these are structural zeroes in the data. This means that we assume these 12 groups of individuals do not exist in our population. Therefore, the total number of strata, ℓ, in this case represents the total number of possible covariate combinations available in our sample and $\ell = 132$ after removing the 12 structural zeroes from the data. If necessary, we can mitigate the structural zeroes by using coarser groups of the covariates.

In Section 4.4.1 we describe and compare the results between our three models and the Fay-Herriot model with covariates included.

4.4.1 BMI Application Model Comparison

Before sampling from any of the models, we first create the symmetric incidence matrix, $\boldsymbol{W}$ of size 132×132, using the Mahalanobis distance described in (61). Recall that we define $\boldsymbol{W}$ by letting $w_{ij} = 1$ if $d_{ij} \leq d_0$ and $w_{ij} = 0$ if $d_{ij} > d_0$ with zeroes on the diagonal (i.e., $w_{ii} = 0$), where d_0 is the value yielding the $\boldsymbol{W}$ matrix that maximizes Moran's I from (62). After performing a grid search for the optimal value of d_0, we achieved the maximum value of Moran's I at $I = 0.212$ when we let $d_0 = \text{mean}(d_{ij})/38 \approx 0.157$. Now that we have $\boldsymbol{W}$, we find the grid interval for ψ to be $(-1.94, 1)$. In all three of our models the matrix $\boldsymbol{W}$ and the grid interval for ψ are the same.

Our model including the stick-breaking algorithm needs to be initialized with clusters in place, since we assume the first block of the Gibbs sampler has clusters already. Therefore, we use the the posterior mean results from the first model extension to define the initial values in our stick-breaking model. We also order the posterior means of η_i, $i = 1, \ldots, \ell$, from least to greatest such that $\eta^{(1)} < \eta^{(2)} < \cdots < \eta^{(\ell)}$ and group these ordered values into ten roughly equal sized clusters, so clusters have similar values of η_i, $i = 1, \ldots, \ell$. These ten clusters are the initial clusters used in the stick-breaking model, then for the remainder of the algorithm the number of clusters is determined by the model.

We ran 50,000 iterations of the Gibbs sampler, then dropped the first 5,000 sampled values and chose every 45th sampled value to end with a final sample size of 1,000 for all parameters in the base model and the first extension of the model. In the second extension of the model with the stick-breaking algorithm we ran 200,000 iterations of the blocked Gibbs sampler, then dropped the first 20,000 sampled values and chose every 180th sampled value to end with a final sample size of 1,000 for all parameters. The model with the stick-breaking algorithm is slower mixing, so it requires more iterations. However, all three models eventually have Gibbs samplers with good performance as evident by the trace plots, auto-correlations, Geweke test of stationarity, and the effective sample sizes.

In the base model, θ, σ^2, and ψ have P-values for stationarity of 0.346, 0.896, and 0.085, respectively, meaning all three parameters pass the Geweke test - results for $\boldsymbol{\mu}$ are similar. Additionally, for θ, σ^2, and ψ the effective samples sizes are 883, 904, and 1151, respectively, which are close to our true sample size - results for $\boldsymbol{\mu}$ are similar.

In the first extension of the model, θ, σ^2, and ψ have P-values for stationarity of 0.910, 0.773, and 0.432, respectively, meaning all three parameters pass the Geweke test - results for $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$ are similar. Additionally, for θ, σ^2, and ψ the effective samples sizes are 1086, 1243, and 1000, respectively, which are an accurate representation of our true sample size - results for $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$ are similar.

In the second extension of the model, θ, σ^2, ψ, ρ, and γ have P-values for stationarity of 0.855, 0.347, 0.873, 0.790, and 0.085, respectively, meaning all five parameters pass the Geweke test - results for $\boldsymbol{\mu}$, $\boldsymbol{\eta}$, $\boldsymbol{z}$, and $\boldsymbol{\nu}$ are similar. Additionally, for θ, σ^2, ψ, ρ, and γ the effective samples sizes are 1000, 753, 889, 863, and 1000, respectively, which are close to our true sample size - results for $\boldsymbol{\mu}$, $\boldsymbol{\eta}$, $\boldsymbol{z}$, and $\boldsymbol{\nu}$ are similar.

A sample from the Fay-Herriot model (presented in Appendix G) can be obtained directly without the need for any MCMC algorithm. Once we sample δ^2 using the grid method, then we can draw samples of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ in order from their known conditional posterior densities. After these parameters are obtained we are able to make predictions for the proportion of obesity in the population using the methods described in Section 4.3.4. The Fay-Herriot model is used as a baseline to compare to the performance of our spatial and clustering models. The Fay-Herriot model is the only model that contains the covariates directly in the model, which we avoid intentionally in our models.

Table 12: Model Comparison of Obesity Population Proportion Predictions

	Predicted P	SE of P	CV of P	95% HPDI of P
Base	0.204	0.028	0.139	(0.150, 0.258)
First Ext.	0.240	0.021	0.089	(0.200, 0.285)
Second Ext.	0.231	0.028	0.119	(0.181, 0.288)
F-H	0.193	0.039	0.202	(0.118, 0.267)
Bootstrap	0.248	0.012	0.047	(0.228, 0.272)

Table 12 contains the results of the population prediction of the obesity proportion for all four models. In our application the response variable is the obesity indicator, therefore P represents the overall proportion of obesity for the population of the eight counties in California. The Fay-Herriot model has the lowest predicted proportion of obesity in the population of 0.193 with the highest posterior standard error (SE) of 0.039, thus leading to also having the largest coefficient of variation (CV) of 0.202. The base model without the clustering component has the next smallest predicted proportion of obesity in the population of 0.204 with a SE of 0.028.

The first and second extensions of our model have the most similar prediction proportions of obesity in the population of 0.240 and 0.231, respectively. Both of these obesity proportion predictions more accurately represent the presence of obesity in the data. The first extension of our model has the shortest 95% highest posterior density interval (HPDI) of P. Since the second extension of our model has a slightly higher SE, this also leads to a wider HPDI of P. This shows how including the stick-breaking algorithm for clustering adds more variation to

Table 13: County Level Obesity Population Prediction Model Comparison

	Predicted P	SE of P	CV of P	95% HPDI of P
County 1				
Base	0.221	0.032	0.144	(0.160, 0.282)
First Ext.	0.265	0.025	0.093	(0.217, 0.313)
Second Ext.	0.264	0.029	0.112	(0.208, 0.322)
F-H	0.195	0.040	0.206	(0.120, 0.277)
County 2				
Base	0.236	0.038	0.161	(0.168, 0.308)
First Ext.	0.285	0.032	0.113	(0.219, 0.349)
Second Ext.	0.279	0.037	0.131	(0.200, 0.342)
F-H	0.188	0.040	0.214	(0.120, 0.275)
County 3				
Base	0.206	0.029	0.141	(0.151, 0.263)
First Ext.	0.240	0.023	0.094	(0.194, 0.284)
Second Ext.	0.232	0.026	0.113	(0.179, 0.277)
F-H	0.195	0.040	0.207	(0.116, 0.274)
County 4				
Base	0.183	0.032	0.177	(0.119, 0.246)
First Ext.	0.213	0.037	0.176	(0.143, 0.284)
Second Ext.	0.203	0.039	0.191	(0.141, 0.290)
F-H	0.197	0.045	0.228	(0.112, 0.286)
County 5				
Base	0.196	0.030	0.152	(0.144, 0.256)
First Ext.	0.231	0.029	0.126	(0.178, 0.291)
Second Ext.	0.228	0.034	0.148	(0.165, 0.296)
F-H	0.188	0.042	0.223	(0.111, 0.276)
County 6				
Base	0.221	0.037	0.166	(0.153, 0.295)
First Ext.	0.267	0.032	0.120	(0.211, 0.338)
Second Ext.	0.267	0.035	0.131	(0.196, 0.327)
F-H	0.193	0.043	0.222	(0.111, 0.281)
County 7				
Base	0.222	0.034	0.154	(0.162, 0.291)
First Ext.	0.263	0.029	0.110	(0.212, 0.327)
Second Ext.	0.258	0.031	0.120	(0.199, 0.314)
F-H	0.197	0.043	0.216	(0.114, 0.278)
County 8				
Base	0.206	0.030	0.146	(0.148, 0.262)
First Ext.	0.233	0.025	0.107	(0.189, 0.282)
Second Ext.	0.222	0.026	0.116	(0.172, 0.273)
F-H	0.199	0.042	0.209	(0.125, 0.284)

the predictions of the strata. By increasing the variation among the strata, we are reducing the amount of global pooling present in the first model extension.

Table 13 contains the results of the population prediction of obesity for all four models split by the eight counties included in our BMI survey data. The model parameters in the Gibbs sampler were not fit to each county, rather the predictions in the output analysis were simply separated by county. The values $\hat{\boldsymbol{\theta}}_k$, $k = 1, \ldots, 8$, are separated into the corresponding eight counties.

That is, for the base model we obtain $\boldsymbol{\mu}_k$, $k = 1, \ldots, 8$, from,

$$\boldsymbol{\mu}_k \mid \theta, \sigma^2, \psi, \hat{\boldsymbol{\theta}}_k \sim \text{Truncated Normal}\left[\boldsymbol{\Delta_0}\hat{\boldsymbol{\theta}}_k + (\boldsymbol{I} - \boldsymbol{\Delta_0})\theta\mathbf{1}, \sigma^2(\boldsymbol{I} - \boldsymbol{\Delta_0})\left(\boldsymbol{R} - \psi\boldsymbol{W}\right)^{-1}\right], \tag{149}$$

using the samples of θ, σ^2, and ψ obtained in the Gibbs sampler with the normal distribution truncated in $[0, 1]$, and $\boldsymbol{\Delta_0} = (\text{diag}(\kappa_1, \ldots, \kappa_\ell) + (\boldsymbol{R} - \psi\boldsymbol{W}))^{-1}\,\text{diag}(\kappa_1, \ldots, \kappa_\ell)$.

Similarly, in the output analysis of the first and second extensions of the base model, we sample $\boldsymbol{\mu}_k$, $k = 1, \ldots, 8$, for

$$\boldsymbol{\mu}_k \mid \boldsymbol{\eta}, \sigma^2, \psi, \hat{\boldsymbol{\theta}}_k \sim \text{Truncated Normal}\left[\Delta_1\hat{\boldsymbol{\theta}}_k + (\boldsymbol{I} - \Delta_1)\boldsymbol{\eta}, \sigma^2(\boldsymbol{I} - \Delta_1)(\boldsymbol{R} - \psi\boldsymbol{W})^{-1}\right], \tag{150}$$

using the samples of η, σ^2, and ψ obtained from the corresponding Gibbs sampler with $\boldsymbol{\Delta_1} = (\text{diag}(\kappa_1, \ldots, \kappa_\ell) + (\boldsymbol{R} - \psi\boldsymbol{W}))^{-1}\,\text{diag}(\kappa_1, \ldots, \kappa_\ell)$, and the normal distribution truncated in $[0, 1]$.

Then, using the respective $\boldsymbol{\mu}_k$, $k = 1, \ldots, 8$, for each county and model described above, we can predict the binary responses for each county's stratum by,

$$Y_{ik} \mid \boldsymbol{\mu}_k, \hat{\boldsymbol{\theta}}_k \sim \text{Binomial}\left(N_{ik}, \mu_{ik}\right), \tag{151}$$

where $N_{ik} = \sum_{j=1}^{n_i} v_{ijk}$ is the Horvitz-Thompson estimator of population size for each county's stratum $i = 1, \ldots, \ell$, $k = 1, \ldots, 8$, and v_{ijk} are the original survey weights. Finally, the overall county proportions are,

$$P_k = \frac{\sum\limits_{i=1}^{\ell} Y_{ik}}{N_k}, \tag{152}$$

where $N_k = \sum_{i=1}^{\ell} N_{ik}$ is the Horvitz-Thompson estimator of total county size.

We also predict the county proportions for Fay-Herriot model with covariates. The full technical details of the model are in Appendix G. We sample θ_{ik}, $i = 1, \ldots, \ell$, $k = 1, \ldots, 8$,

$$\theta_{ik} \mid \boldsymbol{\beta}, \delta^2, \hat{\boldsymbol{\theta}}_k \sim \text{Truncated Normal}\left[\lambda_i\hat{\theta}_{ik} + (1 - \lambda_i)\boldsymbol{x_i'}\boldsymbol{\beta}, (1 - \lambda_i)\delta^2\right], \tag{153}$$

where $\lambda_i = \frac{\delta^2}{\hat{\sigma}_i^2+\delta^2}$, $i = 1, \ldots, \ell$ and the normal distribution truncated in $[0, 1]$. Then, using $\boldsymbol{\theta}_k$, we can predict the binary responses for each county's stratum by,

$$Y_{ik} \mid \boldsymbol{\theta}_k, \hat{\boldsymbol{\theta}}_k \sim \text{Binomial}\left(N_{ik}, \theta_{ik}\right). \tag{154}$$

Now, we calculate the overall county proportions P_k as seen in (152).

The results for each county are similar to the overall results described from Table 12, however since the sample sizes are reduced when we group by county then the posterior standard errors increase. All of the counties have roughly the same sample size, except for County 3 that has an exceptionally large sample size of 795 observations. County 3 accounts for about 43% of the total BMI data sample size. The remainder of the counties each have sample sizes ranging from 125 to 176 observations. The comparatively large sample size of County 3 yields smaller posterior standard errors compared to the other counties. The base model and the Fay-Herriot model have obesity proportion predictions lower than the first and second extension across all counties.

4.5 Conclusion

Our main goal in introducing the clustering component in these models is to further accommodate the covariates without explicitly assuming the relationship between the response and covariates. By combining the cluster model with the spatial model, we continue to reduce the severity of global pooling and instead allow for neighbors with similar attributes to have predictions closer together. Specifically, the spatial model with the stick-breaking component gains additional information from the covariates without directly including them in the model. We see the reduction in global pooling in the spatial model with the stick-breaking component since this model has a wider 95% HPDI of P compared to the model without the stick-breaking algorithm.

Future work includes adapting the model with the stick-breaking component to apply to the situation with a polychotomous response variable instead of a binary response variable. The polychotomous response model could provide a more detailed understanding of the study variable. For an example related to the BMI application, we may be more interested in the proportions of individuals in the population who are underweight, healthy, overweight, obese, and extremely obese - instead of simply predicting the proportion of obesity. By understanding all levels of BMI present in the population we can see the more extreme values of BMI on both ends of the spectrum.

Appendix G - Fay-Herriot Model

In Appendix G we discuss the technical details of the Fay-Herriot model (Nandram, Erciulescu, and Cruze 2019). We fit the Fay-Herriot model as a baseline for comparison to our spatial and cluster models with a binary response. Recall $\hat{\theta}_i$ and $\hat{\sigma}_i^2$ are observed data and let $\boldsymbol{x_i}$ be the covariates. Consider the Fay-Herriot model with covariates included:

$$\begin{aligned}\hat{\theta}_i \mid \theta_i &\overset{ind}{\sim} \text{Normal}\left(\theta_i, \hat{\sigma}_i^2\right), \\ \theta_i \mid \boldsymbol{\beta}, \delta^2 &\overset{ind}{\sim} \text{Normal}\left(\boldsymbol{x_i}'\boldsymbol{\beta}, \delta^2\right), \\ \pi\left(\boldsymbol{\beta}, \delta^2\right) &= \frac{1}{(1+\delta^2)^2}, \quad i = 1, \dots, \ell.\end{aligned} \tag{155}$$

Beginning with the joint posterior density,

$$\pi\left(\boldsymbol{\theta}, \boldsymbol{\beta}, \delta^2 \mid \hat{\boldsymbol{\theta}}\right) \propto \frac{1}{(1+\delta^2)^2}\left(\frac{1}{\delta^2}\right)^{\frac{\ell}{2}} \prod_{i=1}^{\ell}\left\{\exp\left[-\frac{1}{2}\left(\frac{1}{\hat{\sigma}_i^2}(\hat{\theta}_i - \theta_i)^2 + \frac{1}{\delta^2}(\theta_i - \boldsymbol{x_i}'\boldsymbol{\beta})^2\right)\right]\right\}, \tag{156}$$

we can first integrate out θ_i. Letting $\lambda_i = \frac{\delta^2}{\hat{\sigma}_i^2 + \delta^2}$ the conditional posterior density for θ_i is,

$$\theta_i \mid \boldsymbol{\beta}, \delta^2, \hat{\boldsymbol{\theta}} \sim \text{Normal}\left[\lambda_i\hat{\theta}_i + (1-\lambda_i)\boldsymbol{x_i}'\boldsymbol{\beta}, (1-\lambda_i)\delta^2\right], i = 1, \dots, \ell. \tag{157}$$

Then, we integrate out θ_i and obtain the density,

$$\pi\left(\boldsymbol{\beta}, \delta^2 \mid \hat{\boldsymbol{\theta}}\right) \propto \prod_{i=1}^{\ell}\left[(1-\lambda_i)\right]^{\frac{1}{2}} \frac{1}{(1+\delta^2)^2} \exp\left[-\frac{1}{2}\sum_{i=1}^{\ell}\frac{\lambda_i}{\delta^2}\left(\hat{\theta}_i - \boldsymbol{x_i}'\boldsymbol{\beta}\right)^2\right]. \tag{158}$$

We rewrite the above equation to get a normal kernel in terms of $\boldsymbol{\beta}$,

$$\begin{aligned}\pi\left(\boldsymbol{\beta}, \delta^2 \mid \hat{\boldsymbol{\theta}}\right) \propto &\prod_{i=1}^{\ell}\left[(1-\lambda_i)\delta^2\right]^{\frac{1}{2}} \frac{1}{(1+\delta^2)^2}\left(\frac{1}{\delta^2}\right)^{\frac{\ell}{2}} \\ &\times \exp\left[-\frac{1}{2}\sum_{i=1}^{\ell}\frac{\lambda_i}{\delta^2}\left(\hat{\theta}_i - \boldsymbol{x_i}'\hat{\boldsymbol{\beta}}\right)^2 - \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\hat{\Sigma}^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right],\end{aligned} \tag{159}$$

where

$$\hat{\boldsymbol{\beta}} = \hat{\Sigma}\sum_{i=1}^{\ell}\frac{\hat{\theta}_i\boldsymbol{x_i}}{\hat{\sigma}_i^2 + \delta^2} \quad \text{and} \quad \hat{\Sigma}^{-1} = \sum_{i=1}^{\ell}\frac{\boldsymbol{x_i}\boldsymbol{x_i}'}{\hat{\sigma}_i^2 + \delta^2}. \tag{160}$$

Here, $\hat{\boldsymbol{\beta}}$ and $\hat{\Sigma}$ are well defined so long as the design matrix $\boldsymbol{X}$ is full rank where $\boldsymbol{X'} = (\boldsymbol{x_1}, \dots, \boldsymbol{x_\ell})$.

The conditional posterior density of $\boldsymbol{\beta}$ is,

$$\boldsymbol{\beta} \mid \delta^2, \hat{\boldsymbol{\theta}} \sim \text{Normal}\left(\hat{\boldsymbol{\beta}}, \hat{\Sigma}\right). \tag{161}$$

Once we can integrate out $\boldsymbol{\beta}$ we are left with the nonstandard conditional posterior density of δ^2,

$$\pi\left(\delta^2 \mid \hat{\boldsymbol{\theta}}\right) \propto \frac{1}{(1+\delta^2)^2} \cdot |\hat{\Sigma}|^{\frac{1}{2}} \prod_{i=1}^{\ell} \left[\frac{1}{\hat{\sigma}_i^2 + \delta^2}\right]^{\frac{1}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{\ell} \frac{1}{\hat{\sigma}_i^2 + \delta^2}\left(\hat{\theta}_i - \boldsymbol{x_i}'\hat{\boldsymbol{\beta}}\right)^2\right]. \tag{162}$$

While $\pi\left(\delta^2 \mid \hat{\boldsymbol{\theta}}\right)$ is not in a standard form, we know this distribution is proper because $Q(\delta^2) = |\hat{\Sigma}|^{\frac{1}{2}} \prod\limits_{i=1}^{l} \left[\frac{1}{\hat{\sigma}_i^2+\delta^2}\right]^{\frac{1}{2}} \exp\left[-\frac{1}{2}\sum\limits_{i=1}^{l} \frac{1}{\hat{\sigma}_i^2+\delta^2}\left(\hat{\theta}_i - \boldsymbol{x_i}'\hat{\boldsymbol{\beta}}\right)^2\right]$ is bounded in δ^2 as long as the design matrix, $\boldsymbol{X}$, is full rank. The only piece of $\pi\left(\delta^2 \mid \hat{\boldsymbol{\theta}}\right)$ that remains is the prior $\pi(\delta^2) = \frac{1}{(1+\delta^2)^2}$, which is a proper prior (Nandram, Erciulescu, and Cruze 2019). Using an improper prior for δ^2 such as $\pi(\delta^2) = \frac{1}{\delta^2}$ would lead to an improper joint posterior.

Note that $\delta^2 > 0$, so we cannot directly use the grid method to sample δ^2. We must first make the transformation $\delta^2 = \frac{1-\phi}{\phi}$ with $0 < \phi < 1$. The transformed conditional posterior density of ϕ is,

$$\pi\left(\phi \mid \hat{\boldsymbol{\theta}}\right) \propto |\hat{\Sigma}|^{\frac{1}{2}} \prod_{i=1}^{\ell} \left[\frac{1}{\hat{\sigma}_i^2 + \frac{1-\phi}{\phi}}\right]^{\frac{1}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{\ell} \frac{1}{\hat{\sigma}_i^2 + \frac{1-\phi}{\phi}}\left(\hat{\theta}_i - \boldsymbol{x_i}'\hat{\boldsymbol{\beta}}\right)^2\right], \tag{163}$$

that we are now able to use the grid method to sample. Then we transform the values of ϕ back into values of δ^2. Once we have samples of δ^2, we can sample $\boldsymbol{\beta}$ then $\boldsymbol{\theta}$ directly from their known conditional posterior distributions. Finally, we can obtain the predicted population proportion as described in Section 4.3.4.

Chapter 5

Summary and Future Work

We conclude this dissertation with a summary of the previous chapters, a description of future work, and questions that arose at the dissertation defense. The work presented in Chapters 2 and 3 have been submitted as individual papers, see Lockwood and Nandram (2023a) and (2023b), respectively. Chapter 4 will also be submitted after graduation, see Lockwood (2023).

In Section 5.1, we present a summary of the methods explored in Chapters 2, 3, and 4, as well as the difficulties we addressed with each model. In Section 5.2, we discuss future work, which includes a polychotomous extension of our model with spatial, cluster, and heterogeneous components. In Section 5.3, we address questions received from the audience and the dissertation committee at the defense.

5.1 Summary

We presented various models with the theme of avoiding defining the relationship between the response variable and the covariates directly in our models. By performing Bayesian predictive inference for a study variable without having to estimate regression coefficients, we consequently overcame some limitations of traditional regression analysis. Not specifying the relationship between the response variable and the covariates adds flexibility and robustness to our models and allows for more applications. For real applications, we took care of three effects (spatial, heterogeneity, and clustering) simultaneously. We have explored several multinomial-Dirichlet models with stick-breaking representation on the mean vector to address a polychotomous regression problem. We also presented a continuous regression problem addressed by including a spatial component in our Bayesian hierarchical model. Finally, we demonstrated a solution to the binary predictive inference problem while also incorporating a clustering stick-breaking prior. We illustrated all these models using an application with BMI data.

In Chapter 2, we first explored several different versions of the multinomial-Dirichlet model. Each version of the model addresses a different concern or situation, and therefore gives us flexibility to choose what model best fits our problem. The variations of the multinomial-Dirichlet model include an unordered and ungrouped model, an ordered and grouped model, a pooled area estimation model, and we show how to include survey weights in the unordered and ungrouped model. We illustrated that this model can be applied to include, making inference about any characteristic (or multiple characteristics) for a finite population, for a number of small areas, or for a finite population using survey weight adjusted data.

The unordered and ungrouped multinomial-Dirichlet model shows a good performance compared to the logistic regression model by providing comparable results and requiring less assumptions. The main issue with the unordered and ungrouped multinomial-Dirichlet model is that the number of parameters drawn in the Gibbs sampler is greater than the number of data points we have. We attempted to solve the issue of having too many parameters in the unordered and ungrouped model by introducing the ordered and grouped multinomial-Dirichlet model. However, the population predictions from this model move closer to simply predicting the sample proportion and relying less on the model. For this reason, we began to explore methods that would reduce the amount of global pooling and better accommodate the covariates.

In Chapter 3, we illustrated the advantage of including a spatial component to better account for the covariates in our models to make Bayesian predictive inference about the finite population mean. We treated each unique covariate combination as an individual stratum. Then, we used small area estimation techniques to make inference about the finite population mean of the continuous response variable. The two spatial models used are the conditional autoregressive (CAR) and simple conditional autoregressive models (SCAR). We showed that neighboring strata yield similar predictions and that the variation between strata that are not neighbors increases. Ultimately, the spatial models have less global pooling compared to the non-spatial models, which was the desired outcome.

The CAR and SCAR models both work well as small area estimation models that reduce global pooling effects without defining the relationship between the response and the covariates. However, in order to emphasize the spatial structure that accommodates the covariates it is important that ρ and γ are not too small. Hence, we prefer the CAR model that has larger values of both ρ and γ. In the CAR model γ is close to unity, which is a positive sign that our spatial component will have more of an impact in the model compared to the much lower γ value in the SCAR model. As ρ and γ decrease, our posterior standard error of the population predictions will also decrease. Therefore, we continued to expand upon the CAR model in the next phase of our research.

In Chapter 4, we continued to exclude covariates from being used directly in our models by combining spatial, heterogeneous, and clustering components. After finding success with the CAR model, we used this spatial model joined with a clustering stick-breaking prior to gain more information from the covariates. The main advantage of including a stick-breaking component in our model is that the number of clusters is determined by the algorithm and is subject to change at every iteration of the blocked Gibbs sampler. This is unlike the spatial component that manually defines neighborhood relationships before the Gibbs sampler. Allowing the number of clusters to fluctuate gives the model and the data the opportunity to select more relevant clusters.

The spatial model with the stick-breaking prior proved to be the most computationally involved. We first tried to use the Pitman-Yor prior described in Ishwaran and James (2001), namely $\nu_s \overset{\text{ind}}{\sim} \text{Beta}\left(1-\delta, \frac{1-\gamma}{\gamma} + s \cdot \delta\right)$. However, there are some computational difficulties with this prior if δ approaches unity, then the prior becomes unstable and is not well-defined. We experience this instability of the prior in our application. Therefore, to improve the behavior of sampling $\boldsymbol{\nu}$, we use the beta prior typically used in the Dirichlet-process prior, $\nu_s \overset{\text{ind}}{\sim} \text{Beta}(1, \alpha)$, $\alpha > 0$ (Sethuraman 1994; Kalli, Griffin, and Walker 2011). This change of prior drastically improved our computation of the stick-breaking parameters and yielded a strongly mixing Gibbs sampler.

We addressed concerns with traditional regression analysis by providing multiple Bayesian hierarchical models that allow for inference to be made about a study variable including covariates and without the need for estimating regression coefficients. All projects completed thus far have successful results that pave the way for further extensions of alternate regression models. The progression from the multinomial-Dirichlet model to the spatial CAR and SCAR models reduces the amount of global pooling seen in our predictions. Then by including clustering components via the stick-breaking prior in our binary spatial model, we are able to extract even more information from the covariates without directly including them in our models. Our models expand the scope of applications we can explore with minimal assumptions, when compared to traditional regression models.

5.2 Future Work

One possible extension of this research is to use a multivariate Dirichlet process prior on both the spatial component containing the covariates and the clustering component. This multivariate prior would be used instead of the conditional autoregressive model to accommodate the spatial relationship between the covariates. Janicki, Raim, Holan, and Maples (2022) described this extension using Dirichlet process mixing on the latent Gaussian process and regression coefficients. The underlying base distribution in the Dirichlet process prior would be a product of Gaussian distributions. This extension would be useful if a dataset contained a large amount of covariate combinations. The Dirichlet process would determine the number of clusters then we would not have to define the incidence matrix via the Mahalanobis distance.

Another possible extension of this research is to expand the spatial cluster model to incorporate polychotomous data. The data used as input in this model are similar to the input used in the proposed binary model, however now our response variable, y, can have more than two responses. We aggregate the data into a finite number of possible covariate combinations to use in our design matrix $\boldsymbol{X}$. Therefore, each row of $\boldsymbol{X}$, denoted $\boldsymbol{x}_i', i = 1, \ldots, \ell$ is a unique covariate combination that has polychotomous responses recorded y_{ijk}, $j = 1, \ldots, n_i$ and $k = 1, \ldots, c$. Here n_i is the number of observations in the stratum corresponding to the unique covariate combinations of $\boldsymbol{x_i}$, and each observation has a corresponding survey weight, v_{ij}. These survey weights can be summed for each stratum to obtain the population size of that stratum, denoted N_i. Now that we are using polychotomous data, the value c represents the number of categories the response variable, y_{ijk}, can take. For example, in our BMI example using a binary response variable we are interested in if an individual is obese or not, therefore only having two potential response values. However, now using polychotomous data in our BMI example we will classify each individual as either (1) underweight, (2) healthy, (3) overweight, (4) obese, or (5) extremely obese. Therefore, now the response variable, y_{ijk}, can take 5 different values, meaning $c = 5$ in this example. The format of the data can be seen below in Table 14.

Table 14: Data Format used in Polychotomus Spatial Model

$\boldsymbol{x}$	$\boldsymbol{n}$	$(\boldsymbol{v}, \boldsymbol{y})$	$\boldsymbol{N}$
$\boldsymbol{x_1}$	n_1	$(v_{1j}, y_{1jk}), j = 1, \ldots, n_1, k = 1, \ldots, c$	$\sum_{j=1}^{n_1} v_{1j} = N_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{x_\ell}$	n_ℓ	$(v_{\ell j}, y_{\ell jk}), j = 1, \ldots, n_\ell, k = 1, \ldots, c$	$\sum_{j=1}^{n_\ell} v_{\ell j} = N_\ell$

The setup of our proposed polychotomous spatial model with stick-breaking prior is:

$$
\begin{aligned}
\boldsymbol{m_i} \mid \boldsymbol{\pi_i} &\sim \text{Multinomial}\left(n_i, \boldsymbol{\pi_i}\right), \\
\text{with } \pi_{ik} &= \frac{e^{\theta_k+\mu_i+\eta_i}}{1+e^{\mu_i+\eta_i}\sum\limits_{k=1}^{c-1} e^{\theta_k}}, \\
\boldsymbol{\mu} \mid \sigma^2, \psi &\sim \text{Normal}\left(\boldsymbol{0}, \sigma^2(\boldsymbol{R}-\psi\boldsymbol{W})^{-1}\right), \\
\eta_i \mid \boldsymbol{z}, \sigma^2 &\sim \sum_{s=1}^{\ell} p_s \cdot \text{Normal}\left(z_s, \sigma^2\right), \\
z_s \mid \rho, \sigma^2 &\sim \text{Normal}\left(0, \frac{\rho}{1-\rho}\sigma^2\right), \\
\pi\left(\boldsymbol{\theta}, \sigma^2\right) &\propto \frac{1}{(1+\sigma^2)^2}, \\
\psi &\sim \text{Uniform}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_\ell}\right), \\
\boldsymbol{\theta} \in \mathbb{R}^{c-1}, &\quad \sigma^2 > 0, \quad 0 < \rho < 1, \\
i = 1, \ldots, \ell, \qquad j &= 1, \ldots, n_i, \qquad k = 1, \ldots, c-1,
\end{aligned}
\tag{164}
$$

where ℓ in this case represents the total number of possible covariate combinations, which are considered to be the individual stratum. Also n_i are the number of observations in each stratum, and $m_{ik} = \sum_{j=1}^{n_i} y_{ijk}$, $i = 1, \ldots, \ell$, $j = 1, \ldots, n_i$, $k = 1, \ldots, c-1$. The values of k range from $k = 1, \ldots, c-1$ since $\pi_{ic} = 1 - \sum\limits_{k=1}^{c-1} p_{ik}$ that is the last value of $\boldsymbol{\pi_i}$ is calculated using the previous values to ensure that $\sum\limits_{k=1}^{c} \pi_{ik} = 1$. Also λ_1 is the minimum eigenvalue of $\boldsymbol{R}^{-1}\boldsymbol{W}$ and λ_ℓ is the maximum eigenvalue of $\boldsymbol{R}^{-1}\boldsymbol{W}$, and since $\sum_{i=1}^{\ell} w_{ii} = 0$ this results in $\lambda_1 < 0 < \lambda_\ell$ (Chung and Datta 2022). Also $(\boldsymbol{R} - \psi\boldsymbol{W})$ is guaranteed to be positive definite as long as ψ is in the range $\frac{1}{\lambda_1} \leq \psi \leq \frac{1}{\lambda_\ell}$. In this model, θ_k are considered the fixed effects, μ_i are the spatial effects, and then η_i account for the clustering. Also p_s are the stick-breaking weights defined as:

$$
\begin{aligned}
p_1 &= \nu_1, \\
p_s &= \nu_s \prod_{k<s} \left(1-\nu_k\right), \quad s = 2, \ldots, \ell-1, \\
p_\ell &= \prod_{s=1}^{\ell-1} \left(1-\nu_s\right),
\end{aligned}
\tag{165}
$$

with priors $\nu_s \overset{\text{ind}}{\sim} \text{Beta}\left(1, \frac{1-\gamma}{\gamma}\right)$ and $\pi(\gamma) \propto 1$, $0 < \gamma < 1$. Using this model we will be able to make predictive inference for the finite population proportion of each category.

We can also include survey weights in our polychotomous spatial model with stick-breaking prior by letting $m^*_{ik} = \sum_{j=1}^{n_i} v^*_{ij} y_{ijk}$, $i = 1, \ldots, \ell$, $j = 1, \ldots, n_i$, $k = 1, \ldots, c-1$. The v^*_{ij} are the adjusted and trimmed survey weights. By including the adjusted and trimmed survey weights, we eliminate the bias present in the original survey weights. Our model with survey weights included is,

$$
\begin{gathered}
\boldsymbol{m_i^*} \mid \boldsymbol{\pi_i} \sim \text{Multinomial}\left(n_i, \boldsymbol{\pi_i}\right), \\
\text{with } \pi_{ik} = \frac{e^{\theta_k + \mu_i + \eta_i}}{1 + e^{\mu_i + \eta_i} \sum\limits_{k=1}^{c-1} e^{\theta_k}}, \\
\boldsymbol{\mu} \mid \sigma^2, \psi \sim \text{Normal}\left(\boldsymbol{0}, \sigma^2 (\boldsymbol{R} - \psi \boldsymbol{W})^{-1}\right), \\
\eta_i \mid \boldsymbol{z}, \sigma^2 \sim \sum_{s=1}^{\ell} p_s \cdot \text{Normal}\left(z_s, \sigma^2\right), \\
z_s \mid \rho, \sigma^2 \sim \text{Normal}\left(0, \frac{\rho}{1-\rho} \sigma^2\right), \\
\pi\left(\boldsymbol{\theta}, \sigma^2\right) \propto \frac{1}{(1+\sigma^2)^2}, \\
\psi \sim \text{Uniform}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_\ell}\right), \\
\boldsymbol{\theta} \in \mathbb{R}^{c-1}, \quad \sigma^2 > 0, \quad 0 < \rho < 1, \\
i = 1, \ldots, \ell, \qquad j = 1, \ldots, n_i, \qquad k = 1, \ldots, c-1,
\end{gathered}
\tag{166}
$$

where ℓ represents the total number of possible covariate combinations, which are considered to be the individual stratum. Also n_i are the number of observations in each stratum. The p_s are the stick-breaking weights defined in (159).

5.3 Questions from the Audience and Dissertation Committee

At the defense for this dissertation, many additional directions for future work were mentioned including:

a. Alternate methods for estimating the sub-population sizes (strata);
b. Inclusion of survey weights as covariates;
c. Incorporate a simulation study to assess the predictive power of the models;
d. Demonstrate the handling of two or more response variables;
e. Calculating Bayesian diagnostics for model assessment;
f. Explore optimal methods to discretize the continuous variables.

For (a), we currently use the Horvitz-Thompson estimator of population size for each strata, and the same estimator for the entire finite population size. Using the Horvitz-

Thompson estimator is not optimal, although it is widely used. Instead, we can let $T_i = \log\left(\sum_{j=1}^{n_i} v_{ij}\right)$, $\boldsymbol{t_i} = (1, \log(n_i))'$, where v_{ij} are the original survey weights $j = 1, \ldots, n_i$, $i = 1, \ldots, \ell$. We can then make the assumption,

$$\begin{gathered} T_i \mid \boldsymbol{\beta}, \sigma^2 \sim \text{Normal}\left(\boldsymbol{t_i}'\boldsymbol{\beta}, \sigma^2\right), \quad i = 1, \ldots, \ell, \\ \pi(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}. \end{gathered} \tag{167}$$

Therefore, the joint posterior density is $\pi(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{T}, \boldsymbol{t_i}) = \pi(\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{T}, \boldsymbol{t_i})\pi(\sigma^2 \mid \boldsymbol{T}, \boldsymbol{t_i})$,

$$\begin{gathered} \boldsymbol{\beta} \mid \sigma^2, \boldsymbol{T}, \boldsymbol{t_i} \sim \text{Normal}\left[\left(\sum_{i=1}^{\ell} \boldsymbol{t_i}\boldsymbol{t_i}'\right)^{-1}\left(\sum_{i=1}^{\ell} T_i\boldsymbol{t_i}\right), \left(\sum_{i=1}^{\ell} \boldsymbol{t_i}\boldsymbol{t_i}'\right)^{-1}\sigma^2\right], \\ \sigma^2 \mid \boldsymbol{T}, \boldsymbol{t_i} \sim \text{InverseGamma}\left(\frac{\ell-2}{2}, \left(\sum_{i=1}^{\ell} T_i - \boldsymbol{t_i}'\hat{\boldsymbol{\beta}}\right)^2 /2\right), \end{gathered} \tag{168}$$

where, $\hat{\boldsymbol{\beta}} = \left(\sum\limits_{i=1}^{\ell} \boldsymbol{t_i}\boldsymbol{t_i}'\right)^{-1}\left(\sum\limits_{i=1}^{\ell} T_i\boldsymbol{t_i}\right)$. Now we can draw σ^2, followed by $\boldsymbol{\beta}$. Finally, we can calculate the estimator of the population strata size by,

$$N_i^{(h)} = \exp\left\{\boldsymbol{t_i}'\boldsymbol{\beta}^{(h)}\right\}, \quad i = 1, \ldots, \ell, \tag{169}$$

for each drawn $\boldsymbol{\beta}^{(h)}$. This calculation can be run in parallel to the main part of the original model. It is well-known that the log transformation will lead to an improper predictive distribution. It is better to use a power transformation within the Box-Cox family, say $T_i = \left(\sum_{j=1}^{n_i} v_{ij}\right)^{1/k}$ and $\boldsymbol{t_i} = \left(1, n_i^{1/k}\right)'$ (e.g. $k = 10$ or so). Makela, Si, and Gelman (2018) address the challenges that arise when the sizes of the nonsampled strata are unknown. They propose nonparametric and parametric Bayesian approaches for predicting the unknown strata sizes.

For (b), there are many different ways to include survey weights in the models presented. One suggestion was to include the survey weights as a covariate. However, since we intentionally avoid linking the study variable to the covariates in our models, then the survey weights would not be directly included in our models. Additionally, we only have the survey weights for the sampled observations and do not have the survey weights for the non-sampled values. Therefore, when we want to perform Bayesian predictive inference, a key component of the dissertation, we would be missing these survey weights in the covariate structure.

For (c), including a simulation study would be useful in assessing the predictive power of the models. For one thing, when we are currently making predictions for the BMI data applications, we do not know the true values of the non-sampled observations we are predicting.

We can use the design-based method presented in Chen, Li and Wu (2020) for generating the finite population and the samples. This simulated data would have the same structure as the BMI data in our applications. We can modify the sampling process of Chen, Li and Wu (2020) to draw x_{1i}, x_{2i}, and x_{3i}. Then, we can construct the population values of y_i by,

$$\begin{gathered} y_i = 23.8449 + 0.0559x_{1i} + 2.2656x_{2i} + 0.2525x_{3i} + e_i, \quad i = 1, \dots, N, \\ e_i \overset{\text{iid}}{\sim} \text{Normal}(0, \sigma^2), \end{gathered} \tag{170}$$

where N is the population size, see Nandram and Rao (2021). Emulating Chen, Li and Wu (2020), we select σ^2 such that the correlation, $\text{Cor}(23.8449 + 0.0559x_{1i} + 2.2656x_{2i} + 0.2525x_{3i}, y_i) = \rho$, and we selected ρ by trial and error. We would now know the true value of the population mean, $\bar{Y} = \frac{1}{N}\sum_{i=1}^{\ell} y_i$. We select π_i such that,

$$\pi_i = \frac{nz_i}{\sum\limits_{i=1}^{N} z_i}, z_i = \theta + x_{1i} + 0.2x_{2i} + 0.1x_{3i} \quad (\text{arbitrary } z_i), \tag{171}$$

where n is the probability sample size and θ is selected, again by trial and error, to ensure $\max(z_i)/\min(z_i) \approx 50$. Also the survey weights in the simulated data are given by,

$$W_i = N\frac{\frac{1}{\pi_i}}{\sum_{i=1}^{\ell}\frac{1}{\pi_i}}, \tag{172}$$

where N is the population size and π_i are the selection probabilities. The probability sample with target sample size n is taken using randomized systematic probability-proportional-to-size sampling. Then, we would use the simulated sample to fit our models and compare how accurately our models predict the true values of the simulated population mean.

For (d), one recommendation during the defense was to utilize a copula regression approach. However, we can handle the inclusion of two or more response variables straightforwardly in Chapter 2, by simply splitting the contingency table for the additional variable(s). We also do not need to use copula regression to extend the models in Chapter 4 to include two binary response variables. We can write the two binary response variables as a multinomial response instead, and then we can proceed with the model described in Section 5.2. For example, if we have y_1 and y_2 as our two binary response variables, we can write

$$\begin{gathered} y_2 \mid y_1 = 0, \pi_0 \sim \text{Bernoulli}(\pi_0), \\ y_2 \mid y_1 = 1, \pi_1 \sim \text{Bernoulli}(\pi_1), \\ y_1 \mid \gamma \sim \text{Bernoulli}(\gamma), \end{gathered} \tag{173}$$

to get the joint probability mass function of (y_1, y_2). Therefore, we have the 2×2 categorical table,

	$y_2 = 0$	$y_2 = 1$
$y_1 = 0$	$(1-\gamma)(1-\pi_0)$	$(1-\gamma)\pi_0$
$y_1 = 1$	$\gamma(1-\pi_1)$	$\gamma\pi_1$

Now we can write,

$$\boldsymbol{z} \mid \boldsymbol{p} \sim \text{Multinomial}(1, \boldsymbol{p}), \tag{174}$$

where $p_1 = (1-\gamma)(1-\pi_0)$, $p_2 = (1-\gamma)\pi_0$, $p_3 = \gamma(1-\pi_1)$, and $p_4 = \gamma\pi_1$ (Yu, Bhadra, and Nandram 2017). Then we can use (174) as the study variable in the model described in (166).

For (e), the Bayesian diagnostics are a useful tool for comparing model performance. However, Bayesian diagnostics are not sensible to use when survey weights are included in the data to account for selection bias. The Bayesian diagnostics check how well our models fit the observed data, but this not sensible if there is bias in the data. Model comparison can be useful still.

For (f), with any application if there are doubts about the best granularity for discretizing the continuous variables, a sensitivity analysis can be used. The models can be fit using different bin sizes of the continuous variables, and the results can be compared to see how sensitive the model is to the bin sizes. For our models, we discretized the continuous variables based on the number of strata and structural zeroes created. We want a reasonable number of strata without too many structural zeroes. The number of structural zeroes can be reduced by using coarser groups of the covariates. For the age variable, the National Center for Health Statistics uses bins of four years, 20-24, 25-29, 30-34, and so on.

References

- Agresti, A. (2012). *Categorical Data Analysis, 3rd Edition*, Hoboken, NJ, USA: Wiley. ISBN: 978-0-470-46363-5. p. 339-368.

- Albert, J. H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, 88(422), p. 669–679. [Link]

- Antoniak, C. E. (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2(6), p. 1152–1174. [Link]

- Battese, G.E.; Harter, R.M.; and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data, *Journal of the American Statistical Association.* 83, p. 28–36.

- Blackwell, D., and MacQueen, J. B. (1973). Ferguson Distributions Via Polya Urn Schemes. *The Annals of Statistics*, 1(2), p. 353–355. [Link]

- Box, G. E. P., and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Longman Higher Education. ISBN 10: 0201006227 / ISBN 13: 9780201006223. p. 421-477.

- Chen, M.H.; Ibrahim, J.; and Kim, S. (2008). Properties and Implementation of Jeffreys's Prior in Binomial Regression Models, *Journal of the American Statistical Association*, 103:484, p. 1659-1664. [Link]

- Chen, Y., Li, P. and Wu, C. (2020). Doubly Robust Inference With Nonprobability Survey Samples, *Journal of the American Statistical Association*, 115 (532), p. 2011-2021.

- Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART Model Search. *Journal of the American Statistical Association*, 93 (443), p. 935-948. [Link]

- Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Statistics.* 4 (1) p. 266-298. [Link]

- Chung, H.C., and Datta, G.S. (2022). Bayesian spatial models for estimating means of sampled and non-sampled small areas. *Survey Methodology*, Statistics Canada, Catalogue No. 12-001-X, Vol. 48, No. 2. p. 463-489. [Link]

- Department of Health and Human Services. (2018). Body mass index: Considerations for Practitioners. Centers for Disease Control and Prevention. p. 1-4. [Link]

- Ding, N. and Vishwanathan, S.V. N. (2010). T -logistic Regression, *Journal of Machine Learning Research.* p. 1-9. [Link]

- Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2), p. 209–230. [Link]

- Ghosh, M., Natarajan, K., Stroud, T. W. F., and Carlin, B. P. (1998). Generalized Linear Models for Small-Area Estimation. *Journal of the American Statistical Association*, 93(441), p. 273–282. [Link]

- Hill, J., Linero, A., and Murray, J. (2019). Bayesian Additive Regression Trees: A Review and Look Forward. *Annual Review of Statistics and Its Application*, Vol. 7. p. 251-278. [Link]

- Ishwaran, H and James, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors, *Journal of the American Statistical Association*, 96:453, p. 161-173. [Link]

- Jackson A.S., Stanforth P.R., Gagnon J., Rankinen T., Leon A.S., Rao D.C., Skinner J.S., Bouchard C., Wilmore J.H. (2002). The effect of sex, age and race on estimating percentage body fat from body mass index: The Heritage Family Study. *International Journal of Obesity and Related Metabolic Disorders.* 26(6). p. 789-796. [Link]

- Janicki, R., Raim, A. M., Holan, S. H., and Maples, J. J. (2022). Bayesian nonparametric multivariate spatial mixture mixed effects models with application to American Community Survey special tabulations. *The Annals of Applied Statistics*, 16(1), p. 144-168. [Link]

- Jo, A., Nandram, B. and Kim, D.H. (2021). Bayesian pooling for analyzing categorical data from small areas. *Survey Methodology*, Statistics Canada, Catalogue No. 12-001-X, Vol. 47, No. 1. p. 191-213. [Link]

- Kalli, M., Griffin, J.E. and Walker, S.G. (2011). Slice sampling mixture models. *Statistics and Computing*, 21, p. 93–105. [Link]

- Lindley, D. V., and Smith, A. F. M. (1972). Bayes Estimates for the Linear Model. *Journal of the Royal Statistical Society.* Series B (Methodological), 34(1), p. 1–41. [Link]

- Lockwood, A. and Nandram, B. (2023a). Bayesian Predictive Inference Using a Multinomial- Dirichlet Model Without Specifying the Relation Between a Binary Variable and the Covariates (submitted to the *Journal of Survey Statistics and Methodology*).

- Lockwood, A. and Nandram, B. (2023b). Bayesian Predictive Inference of a Finite Population Mean Without Specifying the Relation Between the Study Variable and the Covariates (submitted to *Survey Methodology*).

- Lockwood, A. (2023). Bayesian Predictive Inference Using the Stick-Breaking Algorithm Without Specifying the Relation Between the Study Variable and the Covariates. *Technical Report*, Department of Mathematical Sciences, Worcester Polytechnic Institute, p. 1-20.

- Lukman, P. A., Abdullah, S., and Rachman, A. (2021). Bayesian logistic regression and its application for hypothyroid prediction in post-radiation nasopharyngeal cancer patients. *Journal of Physics: Conference Series.* Ser. 1725. 012010. p. 1-9. [Link]

- Makela, S., Si, Y., and Gelman, A. (2018) Bayesian inference under cluster sampling with probability proportional to size. *Statistics in Medicine.* Vol. 37(26). p. 3849-3868.

- Nandram, B. (1998). A Bayesian analysis of the three-stage hierarchical multinomial model, *Journal of Statistical Computation and Simulation*, 61:1-2, p. 97-126. [Link]

- Nandram, B. (2007). Bayesian predictive inference under informative sampling via surrogate samples. *Bayesian Statistics and its Applications*, (Eds. S.K. Upadhyay, U. Singh and D. Dey), Anamaya, New Delhi, Chapter 25, p. 356-374.

- Nandram, B. (2021). A Bayesian Approach to Linking a Survey and a Census via Small Areas. *Stats*. 4, p. 509–528. [Link]

- Nandram, B. and Choi, J.W. (2010). A Bayesian Analysis of Body Mass Index Data From Small Domains Under Nonignorable Nonresponse and Selection. *Journal of the American Statistical Association*, 105:489, p. 120-135. [Link]

- Nandram, B., Erciulescu, A.L. and Cruze, N.B. (2019). Bayesian benchmarking of the Fay-Herriot model using random deletion. *Survey Methodology*, Statistics Canada, Catalogue No. 12-001-X, Vol. 45, No. 2. p. 365-390. [Link].

- Nandram, B., Toto, M.C., and Choi, J.W. (2011). A Bayesian benchmarking of the Scott–Smith model for small areas. *Journal of Statistical Computation and Simulation*, 81, p. 1593 - 1608. [Link]

- Nandram, B., Kim, D., and Zhou, J. (2019). A pooled Bayes test of independence for sparse contingency tables from small areas, *Journal of Statistical Computation and Simulation*, 89:5, p. 899-926. [Link]

- Nandram, B. and Rao, J. N. K. (2021). A Bayesian Approach for Integrating a Small Probability Sample with a Non-probability Sample. Survey Research Methods Section JSM. p. 1568 - 1603. [Link]

- NHANES III. The Third National Health and Nutrition Examination Survey. (1988-1994). National Center for Health Statistics (NCHS). Centers for Disease Control and Prevention. p. 1-67. [Link]

- Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*, Wiley Series in Survey Methodology. p. 75-96, 333-404.

- Ritter, C., and Tanner, M. A. (1992). Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler. *Journal of the American Statistical Association*, 87(419), p. 861–868. [Link]

- Scott, A. and Smith, T.M.F. (1969). Estimation in multi-stage surveys, *Journal of the American Statistical Association.* 101, p. 1387–1397

- Sethuraman, J.(1994). A constructive definition of Dirichlet priors, *Statistica Sinica* 4. p. 639–650. [Link]

- Tang, X., Ghosh, M., Ha, N.S., and Sedransk, J. (2018). Modeling Random Effects Using Global–Local Shrinkage Priors in Small Area Estimation, *Journal of the American Statistical Association*, 113:524, p. 1476-1489, [Link]

- Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476), p. 1566–1581. [Link]

- Tibshirani, J. and Manning, C (2013). Robust Logistic Regression using Shift Parameters. 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference. 2. p. 124–129. [Link]

- Yang, L., Nandram, B., and Choi, J.W. (2023). Bayesian Predictive Inference Under Nine Methods for Incorporating Survey Weights. *International Journal of Statistics and Probability*; Vol. 12, No. 1. p. 33-53. [Link]

- Yin, J., and Nandram, B. (2020). A Bayesian Small Area Model with Dirichlet Processes on the Responses. *Statistics in Transition New Series*, 21(3), p. 1-19. [Link]

- Yu, Y., Bhadra, D., and Nandram, B. (2017). Tests of Independence for a 2 × 2 Contingency Table with Random Margins. *International Journal of Statistics and Probability*. 6. p. 106-121.