

UCLA

UCLA Electronic Theses and Dissertations

Title

Bayesian Estimation of Finite Population Quantities from Spatially Correlated Data under Ignorable and Nonignorable Survey Designs

Permalink

<https://escholarship.org/uc/item/0jq8n7h3>

Author

Chan-Golston, Alec Michael

Publication Date

2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Bayesian Estimation of Finite Population Quantities
from Spatially Correlated Data under
Ignorable and Nonignorable Survey Designs

A dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Philosophy
in Biostatistics

by

Alec Michael Chan-Golston

2020

© Copyright by
Alec Michael Chan-Golston
2020

ABSTRACT OF THE DISSERTATION

Bayesian Estimation of Finite Population Quantities
from Spatially Correlated Data under
Ignorable and Nonignorable Survey Designs

by

Alec Michael Chan-Golston

Doctor of Philosophy in Biostatistics

University of California, Los Angeles, 2020

Professor Sudipto Banerjee, Chair

Data which is geographically referenced has become increasingly common in many fields of study, such as public health, education, forestry, medicine, and agriculture. When data is sampled from a population, there is often knowledge pertaining to the units not sampled, such as a total count and simple demographics. This knowledge can be leveraged to estimate finite population quantities such as the population total or mean, using design or model-based estimators. However, it is unknown how these estimators perform in the presence of spatial correlation, that is, when the outcome sampled is assumed to be a partial-realization of a spatial process. This dissertation first presents an analysis predicting store patronage and fruit and vegetable expenditures during a corner store intervention using Bayesian spatial techniques and then presents a brief example of finite population estimation in an ignorable sampling setting. Next a general Bayesian framework is presented that accounts for both study design and spatial association. Under this, posterior samples of finite population quantities can be retrieved. This framework is first given under the assumption of an ignorable sampling design and is used to construct four models to account for two-stage designs with

spatial dependence. These models are first applied to simulated data and then are used in an analysis of nitrate levels in California groundwater. We find that models accounting for both study design and spatial association perform best. This general framework is then extended to allow for a nonignorable sampling design, specifically to account for missing data patterns seen in reported annual household income in the corner store data. Through this, we are able to construct finite population estimates of the percent of income spent on fruits and vegetables. Such a framework provides a flexible way to account for spatial association and complex study designs in finite populations.

The dissertation of Alec Michael Chan-Golston is approved.

Michael L. Prelip

Catherine A. Sugar

Mark S. Handcock

Sudipto Banerjee, Committee Chair

University of California, Los Angeles

2020

DEDICATION

For my parents

Table of Contents

1	Introduction	1
1.1	Sampling	1
1.2	Design and Model-Based Finite Population Estimates	2
1.3	Design Ignorability	5
1.4	Bayesian Methods for Spatial Data	7
1.5	Finite Population Estimation for Spatial Data	8
1.6	Contributions and Dissertation Outline	8
2	Estimating the Impact of Competing Stores on Corner Store Patronage and Purchasing Behaviors During an Intervention	10
2.1	Introduction	10
2.2	Data	13
2.3	A Bayesian Spatial Interaction Model	17
2.3.1	Logistic and Linear Models	17
2.3.2	The Coregionalization and Conditional Approaches	19
2.3.3	Modeling Using the Data	20
2.3.4	A Conjugate Bayesian Model	22
2.3.5	A Finite Population Framework	23
2.3.6	Model Comparison and Assessment	25
2.4	Simulation	25
2.5	Data Modeling	27
2.5.1	Implementation	27
2.5.2	Results	28
2.6	Discussion	31

3	Multistage Bayesian FPS Models for Spatial Data with Ignorable Designs	33
3.1	Introduction	33
3.2	Bayesian modeling of multi-stage sampling	34
3.3	Bayesian spatial process modeling for multi-stage sampling	37
3.4	Model Implementation and Assessment	39
3.4.1	General framework	39
3.4.2	Exact Monte Carlo Estimation	41
3.4.3	Model Comparison and Assessment	43
3.5	Simulation	44
3.5.1	Data Generation	44
3.5.2	Exact Monte Carlo Simulation	45
3.5.3	Markov Chain Monte Carlo Simulation	47
3.6	Data Analysis: Nitrate in Central California Groundwater	50
3.7	Discussion	54
3.8	Appendix	56
4	Finite Population Estimation of Food Expenditure in the Presence of Spatially Correlated Data	61
4.1	Introduction	61
4.2	Data	63
4.3	A General Framework	66
4.3.1	Model Comparison and Assessment	70
4.4	Simulation	71
4.5	Data Analysis	76
4.5.1	Implementation	76
4.5.2	Results	77
4.6	Discussion	79

5 Conclusion	83
Bibliography	85

List of Figures

2.1	Locations of Stores	14
2.2	K-L Divergence Comparisons from 3 Simulations	27
2.3	Corner and Rival Store Random Effects from the Coregionalization Model	29
2.4	Corner and Rival Store Random Effects from the Independent Model	30
3.1	Centered Population Mean Estimates from 2 Exact Models with 95% CI	46
3.2	Centered Population Mean Estimates from 4 MCMC Models with 95% CI	48
3.3	Centered Population Mean Estimates from 2 MCMC Models with 95% CI.	49
3.4	Plot of California zip code tabulation areas.	52
3.5	Variograms from Population and Sampled Data.	52
3.6	Interpolated surfaces using the population (truth) and posterior predictive samples from the four models.	53
3.7	Spatial residual plots from the three spatial models.	54
4.1	Linear Interpolation Plots from Full Simulated Data and 3 Scenarios	73
4.2	Comparison of Standard and Orthogonalized η_y Estimates from Simulation 3, Model 4	76

List of Tables

2.1	Description of the Sample	16
2.2	Comparing KL-Divergence Results of 3 Simulations	26
2.3	Results of Logistic Model Predicting Store Patronage	29
2.4	Results of Linear Model Predicting Percentage of Food Expenditure Spent on FV	31
3.1	Comparison of Parameter Estimation and Model Fit in Two Exact Models . .	47
3.2	Comparison of Estimates and Model Fit from MCMC	49
3.3	Results of Data Analysis	54
4.1	Annual Income and FV Expenditures by Site and Time-point.	64
4.2	Simulation Results of Scenario 1: Spatial Outcome, Random Response . . .	74
4.3	Simulation Results of Scenario 2: Spatial Outcome, Preferential Sampling . .	75
4.4	Simulation Results of Scenario 3: Spatial Outcome, Preferential Sampling, Spatial Inclusion	81
4.5	Results of regression models predicting percentage of income spent on fruits and vegetables (log-scale)	81
4.6	Finite Population Estimates and 95% CI of Percentage of Income Spent on Fruits and Vegetables by Community and Timepoint.	82

ACKNOWLEDGEMENTS

Chapter 1 and 3 contain work previously printed in Chan-Golston et al. (2020). As per the permission guidelines provided by Environmetrics since I am the primary author of Chan-Golston et al. (2020) and this dissertation, no prior permission is required to reprint this material. This article has two co-authors. Sudipto Banerjee assisted in the writing of all sections of the manuscript. Mark S. Handcock provided important recommendations for Sections 3.2 and 3.4.

Chapter 2 has one co-author. Sudipto Banerjee assisted in the writing of all sections of the manuscript.

Chapter 4 has four co-authors. Sudipto Banerjee assisted in the writing of Sections 4.3 and 4.4 and provided feedback on all sections of the manuscript. Sarah E. Roth, Thomas R. Belin, and Michael L. Prelip provided feedback for all sections of the manuscript.

This dissertation work was supported by Grants from the National Heart, Lung and Blood Institute (NHLBI) at the National Institutes of Health P50 HL105188 and R25 HL108854, the Division of Information and Intelligent Systems of the National Science Foundation under Grant 1562303, the Division of Mathematical Sciences of the National Science Foundation under Grant 1916349, and the National Institute of Environmental Health Sciences of the National Institutes of Health under Grants 1R01ES027027 and R01ES030210-01.

BIOGRAPHICAL SKETCH

Alec Chan-Golston received his Masters of Science in Biostatistics in the Fielding School of Public Health at the University of California, Los Angeles (UCLA) after completing a Bachelors of Science in Applied Mathematics and Statistics at UCLA. His current research is focused on creating hierarchical models which can correctly account for both spatial correlation and study design. Alec is interested in Bayesian techniques, hierarchical data, and data visualization. During this time as a graduate student, he worked closely with the Department of Community Health Sciences as a statistician on multiple projects, including the evaluation of corner store makeovers in East Los Angeles and a physical activity intervention in middle school classes in Los Angeles Unified School District. In addition to receiving the 2019-2020 UCLA Graduate Division Dissertation Year Fellowship, he was the recipient of the 2015 Abdelmonem A. Afifi Fellowship, 2016 Judith Blake Memorial Fellowship, and the 2017 Celia G. and Joseph G. Blann Fellowship.

Chapter 1

Introduction

The aim of this dissertation is to introduce a flexible Bayesian framework that can account for both study design and spatial dependencies, first given the assumption of design ignorability, and then without it. A brief review of four topics are provided in this chapter: sampling, finite population estimation (from the design and model-based perspectives), design ignorability, Bayesian modeling of spatial data, and finite population estimation in the presence of spatial data. The chapter concludes with a summary of the main contributions of this work and a brief outline of the dissertation.

1.1 Sampling

The development of random sampling (Neyman 1934) has been incredibly influential in all fields of science. As inference is desired for large populations, simple random sampling is often replaced with more complicated sampling designs, such as stratified, probability-proportional-to-size, and cluster sampling. In particular, multistage sampling has become a common technique to collect data on a population that is geographically diverse, has naturally discrete groups, or can be split into more homogenous districts. It is the process of first randomly selecting a subset of predefined groups, referred to as primary sampling units, before taking a random sample of the elements contained in each chosen primary sampling unit, referred to as the secondary sampling units. This process is continued for the desired number of levels of sampling. This technique has been implemented in numerous large scale surveys in a variety of fields such as health (the National Health Interview Survey and the National Health and Nutrition Examination Survey) and education (the National Postsecondary Student Aid Study and the National Assessment of Educational Progress), as well as forestry, medicine, agriculture, and many other disciplines. While a simple random sampling is may be preferred to multistage sampling, it is often the case that time and monetary considerations

make multistage sampling a more viable alternative to collect a representative sample.

In order to collect such datasets, a population of interest must first be identified and all members must be recorded so that some form of sampling may take place. It is important to recognize, especially in surveys which sample a high percentage of the population, that in selecting one individual for the study, the probability of another individual being selected changes (this is not true in the case of a theoretically infinite population). Additionally, an indicator of inclusion in the sample can be defined by:

$$I_i = \begin{cases} 1, & \text{if the } i^{\text{th}} \text{ case is included in the sample} \\ 0, & \text{otherwise} \end{cases} .$$

Comparing this indicator to the commonly used indicator of missing, M_i discussed by Little and Rubin (2002) it is evident that $I_i = |1 - M_i|$. This notation allows quick reference to the wide literature of missing data. Given a random sample of size n from the population, $n < N$, denote "s" the "seen" index set $\{i : I_i = 1\}$ and "ns" the "not seen" index set $\{i : I_i = 0\}$. For notational convenience and without loss of generality, take the sampled units to be the first n elements of y , then $y = \{y_s, y_{ns}\}$, where where $y_s = (y_1, \dots, y_n)$ are the sampled elements and $y_{ns} = (y_{n+1}, \dots, y_N)$ are the unsampled elements.

1.2 Design and Model-Based Finite Population Estimates

Finite population survey sampling concerns statistical modeling and inference on finite populations from sampling designs; see, for example, Cochran (1977), Hartley and Sielken Jr. (1975), Royall (1970), and Horvitz and Thompson (1952). These estimates are derived from statistical techniques which take such probabilities into account. For instance, suppose there is interest in estimating the population total $T = \sum_{i=1}^N y_i$ for some observable characteristic y in a population of size N .

One of the most common approaches to estimating T is the Horvitz-Thompson estimator (Horvitz and Thompson 1952), $\hat{T} = \sum_{i=1}^N I_i \frac{y_i}{\pi_i} = \sum_{i=1}^n \frac{y_i}{\pi_i}$, where $P(I_i) = \pi_i$. This statistic is

unbiased if all units are independent and is easily interpreted as a weighted sum where each element contributes proportional to their inverse probability of selection. In the case of each element having an equal probability of being sampled, the estimate simplifies to $\hat{T} = \frac{N}{n} \sum_{i=1}^n y_i$, which is the sample total multiplied by the inverse of the sampling fraction, n/N . This concept can be extended to other sampling schemes where π_i is known, such as two-stage, stratified, and systematic sampling.

Horvitz-Thompson estimators are popular design-based statistics; estimates that are derived from knowledge of the sampling scheme and the observed data. This style of inference relies the assumption that the outcome y in the population is fixed and that randomness is only introduced by the random selection process. One main criticism of the design-based inference is that it can perform poorly in small sample settings. In addition, it has no techniques to account for non-response or measurement error (Kalton 2002), which are both common in large survey datasets.

The assumption of a fixed outcome is in stark contrast to the model-based approach, which assumes that the outcome is a random sample from a specified distribution. Inferences is performed on the joint distribution of y and I and values are imputed for the unobserved values in the population. The complete likelihood for y is

$$p(y, I | \theta, \psi) = p(y | \theta)p(I | y, \psi) , \tag{1.1}$$

where θ are the set of parameters associated with the distribution of the outcome, y , and ψ are the set of parameters associated with the probability of inclusion, I . We refer to the parameter(s) θ as superpopulation parameters. Little (2004) explains that these model-based approaches can be further divided into superpopulation and Bayesian methods.

Superpopulation techniques assume the parameters of the distribution are fixed and aim to maximize 1.1. For example, consider the simple case when $p(I | y, \psi) \propto \text{constant}$ and each y_i are independent, following an exponential distribution with rate parameter θ . Then the likelihood is maximized when the estimate of the rate parameter is taken to be $\hat{\theta} = \bar{y}_{obs}$ and in the absence of any other covariates, we take $\hat{y}_i = \hat{\theta}$, $y_i \in y_{ns}$. Then $\hat{T} = \sum_{i=1}^n y_i + \sum_{i=n+1}^N \hat{y}_i = \sum_{i=1}^n y_i + \sum_{i=n+1}^N \bar{y} = N\bar{y}$. It is immediately apparent that in this example, the superpopulation estimate

is identical to the Horvitz-Thompson estimator, but this need not always be the case. Royall (1970) extended this concept a linear regression framework by predicting non-sampled outcomes with known characteristics and made direct comparisons to the Horvitz-Thompson estimators. A discussion of the distribution of the inclusion mechanism, $p(I|y, \psi)$, and its implications for inference and interpretation will be discussed in the following section. A comprehensive review of superpopulation techniques has been compiled by Valliant et al. (2000). Little and Rubin (2002) present a similar missing data approach, making the distinction between likelihood-based estimation and bayesian estimation.

The Bayesian approach specifies prior distributions for these superpopulation parameters and makes inferences on the posterior distribution of the finite population estimate. Bayesian imputation is performed using the predictive distribution:

$$p(y_{ns} | y_s, I) = \int p(y_{ns} | y_s, I, \theta, \psi) p(\theta, \psi | y_s, I) \, d\theta d\psi. \quad (1.2)$$

If non-informative priors are taken, the results will often yield more classical conclusions. For instance, suppose that we make the additional assumption that $y_i | \mu, \pi_i \stackrel{iid}{\sim} N\left(\pi_i \mu, \frac{\pi_i^2}{1-\pi_i}\right)$ and take $p(\mu) \propto 1$. Integrating out μ , Ghosh and Sinha (1990) note that

$$y_{ns} | y_s \sim N\left(\frac{\sum_{i=1}^n (1-\pi_i) y_i / \pi_i}{\sum_{i=1}^n (1-\pi_i) / \pi_i} \pi_{ns}, \text{diag}\left(\frac{\pi_{n+1}^2}{1-p_{n+1}}, \dots, \frac{\pi_N^2}{1-p_N}\right) + \frac{1}{\sum_{i=1}^n (1-\pi_i) / \pi_i} \pi_{ns} \pi_{ns}^\top\right),$$

where $\pi_{ns} = (\pi_{n+1} \dots \pi_N)^\top$. As $\sum_{i=n+1}^N \pi_i = n - \sum_{i=1}^n \pi_i = \sum_{i=1}^n (1-\pi_i)$,

$$E[T | y_s] = \sum_{i=1}^n y_i + \frac{\sum_{i=1}^n (1-\pi_i) y_i / \pi_i}{\sum_{i=1}^n (1-\pi_i) / \pi_i} \times \sum_{i=n+1}^N \pi_i = \sum_{i=1}^n y_i + \sum_{i=1}^n y_i (1-\pi_i) / \pi_i = \sum_{i=1}^n y_i / \pi_i,$$

which is the Horvitz-Thompson estimator.

As another example, consider estimating the population mean \bar{y} using a fully specified conjugate model where $y_i | \mu, \sigma^2 \stackrel{iid}{\sim} N(\mu, \sigma^2)$, $\mu | \theta \sim N(\theta, \frac{\sigma^2}{n_0})$, and $\sigma^2 \sim \text{IG}(a, b)$. Integrating out μ , notice that $\bar{y}_{ns} \sim N\left(\frac{n_0 \theta + n \bar{y}_s}{n_0 + n}, \frac{\sigma^2}{n_0 + n} \frac{1+n_0/N}{1-n/N}\right)$ and therefore $\frac{\bar{y} - \{\frac{n}{N} \bar{y}_s + (1-n/N) \frac{n_0 \theta + n \bar{y}_s}{n_0 + n}\}}{\sqrt{\sigma^2 \frac{(1-n/N)(1+n_0/N)}{n+n_0}}} | y_s, \sigma^2 \sim N(0, 1)$. Letting $n_0 \rightarrow 0$, e.g. the prior for μ becomes uninformative, and $p(\sigma^2) \propto 1/\sigma^2$, we have that

$\frac{\bar{y}-\bar{y}_s}{\sqrt{\frac{s^2}{n}(1-n/N)}} \sim t_{n-1}$. Taking N to be large so that $\frac{n}{N} \approx 0$, we arrive at our standard statistic for the one-sample t-test.

Bayesian inference for finite population survey sampling is discussed in great detail in Gelman (2007), Little (2004), Ghosh and Meeden (1997), and Ericson (1969). In this domain, there is a substantial literature on small area estimation for regionally aggregated data (see, e.g., Rao, 2003; Ghosh et al., 1998; Ghosh and Rao, 1994; Clayton and Kaldor, 1987), where interest lies in modeling dependencies across regions.

1.3 Design Ignorability

In an example presented in Section 1.2, the conclusion was only valid under the assumption that $p(I|y, \psi) \propto \text{constant}$, meaning that the study design was ignorable. Such designs make the assumption that the probability of inclusion is independent from the outcome measurement, y . This section presents an introduction to design ignorability, summarizing remarks made by Gelman et al. (2014) in Chapter 8, "Modeling Accounting for Data Collection" and by Little and Rubin (2002) in Chapters 1, "Introduction, and 6, "Theory of Inference Based on the Likelihood Function".

When considering the complete data likelihood, we can incorporate covariates x observed for the entire population by rewriting (1.1) as

$$p(y, I | x, \theta, \psi) = p(y | x, \theta)p(I | x, y, \psi) . \tag{1.3}$$

We would like to perform our standard Bayesian inference on θ by ignoring the inclusion mechanism, $p(I | x, y, \psi)$, e.g.

$$p(\theta | x, y_s) \propto p(\theta | x)p(y_s | x, \theta) . \tag{1.4}$$

However, the posterior distribution of θ stemming from (1.3) is dependent on I :

$$p(\theta | x, y_s, I) \propto p(\theta | x) \iint p(\psi | x, \theta)p(y_s | x, \theta)p(I | x, y, \theta)dy_{ns}d\psi . \tag{1.5}$$

The inclusion mechanism can be ignored if it can be assumed that y_{ns} are missing at random (MAR)

and parameters ψ and θ are *a priori* independent, such that $p(\theta, \psi) = p(\theta)p(\psi)$. This independence condition is often referred to as "distinctness of parameters" and an equivalent definition is given by Gelman, which is $p(\psi | x, \theta) = p(\psi | x)$. Data is said to be MAR if $p(I | x, y_s, y_{ns}, \psi) = p(I | x, y_s, \phi)$ for all y_{ns} , which means that the inclusion mechanism does not depend on unobserved values of y . Under these two conditions, (1.5) immediately simplifies to (1.4), as the function no longer depends on y_{ns} and ψ is easily integrated out.

In the special case that $p(I | x, y_s, y_{ns}, \psi) = p(I | \psi) = \psi$, where ψ is a constant, we say that the data is missing completely at random (MCAR). This means that the probability of inclusion is independent of all characteristics of sampled and nonsampled observations. For instance, when performing simple random sampling on a population of size N , $\psi = \frac{1}{N}$. As it is uncommon to find true instances of MCAR, the hope is if a large enough set of covariates, x , is collected, the data can be assumed to be MAR. However, if this assumption cannot be made, then the data is said to be not missing at random (NMAR), meaning the data has a nonignorable design.

When considering these patterns of missing, it is important to understand if the inclusion mechanism is known, meaning that $p(I | x, y_s, y_{ns}, \psi)$ can be assigned a distribution, even if ψ is unknown. If data is MCAR, then the inclusion mechanism is known and ignorable. In the MAR case, if the probability of inclusion only depends on a set of covariates observed in the population, $p(I | x, y_s, \psi) = p(I | x, \psi)$, then the mechanism is known and said to be *strongly* ignorable. As an example, in stratified sampling the probability of inclusion is dependent on which strata a unit falls into, so conditioning on level of strata, the inclusion mechanism is understood. If $p(I | x, y_s, \psi)$ cannot be simplified further, then the mechanism is known and ignorable. One example of this given by Gelman et al. (2014) is a sequential sampling scheme in which the selection probability of the current member is dependent on a measurement of the previously selected member.

More complicated cases arise when the designs are nonignorable. Censored and rounded data are examples of nonignorable data with known mechanisms. The most complex problems stem for nonignorable designs which have unknown inclusion mechanisms. As discussed earlier, this can occur when the superpopulation parameters and the parameters that define the inclusion mechanism are not distinct. In an example given by Little and Rubin (2002), consider a stochastic censoring model in which a bivariate outcome $\{y_1, y_2\}$, where y_1 is observed only if y_2 is greater

than 0. Then the inclusion mechanism for y_1 is always dependent on the parameters defining the distribution of the outcome y_2 , and therefore ψ and θ cannot be assumed to be *a priori* independent. Alternatively, suppose that we cannot make the assumption of MAR, meaning the inclusion mechanism is dependent on all values of y . For instance, suppose we would like to estimate average income and perform a simple random sample to survey a community. If the probability of response differs by income then the data cannot be assumed to be MAR.

The assumption of ignorability can be further complicated if the covariates x are only observed for the sampled units. Sugden and Smith (1984) identify 6 combinations of design variables, x , and proxy design variables (e.g. a summary function of x , $w(x)$) observed for some part of the population: $\{x\}$, $\{w, x_s\}$, $\{w\}$, $\{w_s, x_s\}$, $\{x_s\}$, and $\{w_s\}$. Considering these different cases, they comment on validity of the ignorability assumption for various sampling schemes.

1.4 Bayesian Methods for Spatial Data

Data believed to be correlated as a function of geographic distance is typically described using a spatial process model. Unlike the literature of small area estimation (see, e.g., Rao (2003), Ghosh and Rao (1994), and Clayton and Kaldor (1987)), where the sampling units are regions such as counties, states or census-tracts, spatial process models consider quantities that, at least conceptually, exist in continuum over the entire domain. The process assigns a probability law to an uncountable subset within a d -dimensional Euclidean domain. In general, spatial process modeling (Banerjee et al. 2014; Cressie and Wikle 2011; and Ripley 2004) follows the generic paradigm

$$[\text{data} \mid \text{process}] \times [\text{process} \mid \text{parameters}] \times [\text{parameters}] , \quad (1.6)$$

which accommodates complex dependencies and multiple sources of variation.

The data is assumed to be a partial realization of a Gaussian process with dependencies between elements defined by an isotropic covariance function, $C(d)$, where d is the distance between any two points. Several choices for $C(d)$ are available (see, e.g. Banerjee et al., 2014), but a versatile family is the Matérn, defined as

$$C(d_{ab}) = \begin{cases} \sigma^2 + \tau^2 & \text{if } d_{ab} = 0 \\ \frac{\tau^2}{2^{\nu-1}\Gamma(\nu)} (\sqrt{2\nu}d_{ab}\phi)^\nu K_\nu(\sqrt{2\nu}d_{ab}\phi) & \text{if } d_{ab} > 0 \end{cases},$$

for a given distance, d_{ab} , between two locations ℓ_a and ℓ_b . Here $K_\nu(\cdot)$ is the modified Bessel function, ν is a smoothness parameter, σ^2 describes the variation due to measurement error, τ^2 measures the spatial variance, and ϕ is a decay parameter which determines the rate of decline in spatial association. The exponential covariance function is a special case of Matérn when $\eta = 1/2$ and is of the form:

$$C(d_{ab}) = \begin{cases} \sigma^2 + \tau^2 & \text{if } d_{ab} = 0 \\ \tau^2 \exp(-\phi d_{ab}) & \text{if } d_{ab} > 0 \end{cases}.$$

In this specific instance, the decay parameter is used to calculate the effective spatial range, $3/\phi$, which is the distance where spatial correlation between two points drops below 0.05.

1.5 Finite Population Estimation for Spatial Data

With regard to finite population sampling in spatial process settings, the literature appears to be considerably more scant than small area estimation. Here, Hoef (2002) discuss connections between geostatistical models and classical design-based sampling and develop methods for executing model-based block kriging. Cicchitelli and Montanari (2012) present a spline regression model-assisted, design-based estimator of the mean for use on a random sample from both finite and infinite spatial populations. A linear spatial interpolator is used by Bruno et al. (2013) to create a design-based predictor of values at unobserved locations which outperforms non-spatial predictors. While related to these developments, none of these techniques have presented a Bayesian approach.

1.6 Contributions and Dissertation Outline

In this article, we will concern ourselves with Bayesian inference for finite populations when the sampling units are spatially oriented. For instance, one may consider estimating the total biomass

in a forest given a sample of trees, the average income of a city given a sample of individuals and their addresses, or the total amount of air pollution attributable to cars on a freeway given a sample of pollution measurements.

This dissertation provides a Bayesian framework by which posterior estimates of finite population quantities can be collected while correctly accounting for spatial correlation in both sampled and nonsampled units. To our knowledge, this is the first work to examine the implications of finite population sampling performed on a partial realization of a spatial process. This flexible framework allows for a variety of study designs which may have complex inclusion mechanisms, which may be appealing to researchers working with geographically referenced data. This current chapter has provided a brief introduction to some of the literature discussed in this dissertation, including sampling, finite population estimation, ignorability assumptions, spatial data, and finite population estimation of spatial data. In Chapter 2 an Bayesian spatial data analysis is used to predict store patronage and fruit and vegetable expenditures during a corner store intervention, incorporating multiple spatial random effects. Using this, a brief example of finite population estimation under an ignorable sampling setting is provided. Chapters 3 and 4 work to create and describe a cohesive Bayesian framework that allows for the estimation of finite population quantities while accounting for study design and various spatial associations. In Chapter 3 an ignorable sampling design is assumed and a Bayesian framework is presented. In particular, a two-stage sampling design in the presence of spatial correlation is examined. Four models are considered to analyze such data, and these are first applied to simulated data and then for a data analysis of nitrate levels in California groundwater. This Bayesian framework is then extended to account for nonignorable sampling designs in Chapter 4. Specifically, this is used to account for complex missing data patterns seen in reported annual household income in the aforementioned corner store intervention data. Through this, we are able to construct finite population estimates of the percent of income spent on fruits and vegetables at the community level. The dissertation concludes with a brief discussion in Chapter 5.

Chapter 2

Estimating the Impact of Competing Stores on Corner Store Patronage and Purchasing Behaviors During an Intervention

2.1 Introduction

In a response to rising obesity and other health disparities that have been linked to diet, more Public Health research has targeted the food environment. One common approach theorized to change food purchasing behaviors at the community level are corner store interventions (Langellier et al., 2013). These interventions are particularly attractive for communities in “food swamps” (Rose et al., 2009), regions with a higher density of stores which primarily sell unhealthy products (such as fast-food or junk food) than stores with healthier food options (such as grocery stores with fresh produce). In a review of this literature, Langellier et al. (2013) notes that corner stores are often the main food supplier for low-income families but primarily sell staples (i.e. eggs and milk) and unhealthy items (i.e. chips and soda). However, fresh fruit and vegetables (FV) are sold infrequently at these stores, and if sold, are often of poor quality and expensive. The authors summarize that most commonly, researchers partner with existing stores to increase the amount of fresh FV, while reorganizing the store to make unhealthy less visible. These interventions may provide business consultations and refrigeration units, as well as provide support to increase awareness in the community by improving signage, holding cooking demonstrations, and remodeling the corner store.

Among these studies, findings have been mixed. Cavanaugh et al. (2014) reports large increases in the availability of FV were seen in conversion stores and Thorndike et al. 2017 concludes that stocking and displaying quality produce increased FV purchases in customers using WIC. Increased sales of healthy foods were also seen by Song et al. 2009. Paek et al. (2014) finds increases in FV,

bean, and nut purchasing, post-conversion, as well a higher awareness of the intervention stores. Similarly, more positive perceptions of the converted stores are reported by Albert et al. 2017. However, the authors find no difference in the reported consumption of FV or money spent on FV. Lawman et al. 2015 also reports that the conversion had no effect on the nutrient content of purchases made in the store.

Mixed findings are also seen in an intervention in eight neighborhoods in East Los Angeles and Boyle Heights, California (Ortega et al., 2016). While community perceptions of healthy food access and corner stores improved, there were no intervention effects corresponding to store patronage, as well as FV purchasing and consumption. One unique feature of this study is that respondents were asked to list all stores where they shopped for food. We are interested in determining if patronage at an individual's corner store and FV purchasing are influenced by that individual's other shopping locations. As all stores have been geocoded, we are challenged to incorporate this spatial information into a cohesive analysis that can answer such questions. To accomplish this, we first consider how the analysis of spatial data is approached in the food environment literature.

A majority food environment studies appear utilize areal data, where units are referenced by geographic regions, such a zip codes or counties. For instance, Frank et al. (2012) constructs measures of the nutritional and physical activity environments at the census block level to identify clusters of blocks which have poor measures of both environments. Smith et al. (2010) measures accessibility to food in data zones, which define geographic regions in Scotland, by measuring the travel time from the centroid of each data zone to the nearest food store. They find that deprived neighborhoods tended to have better access to grocery stores, especially those that sell fresh produce. A similar conclusion was found in rural Texas neighborhoods (Sharkey et al., 2010), where the authors conclude that access to fruits and vegetables was higher in regions with higher percentages of social and material deprivation or low rates of vehicle ownership.

To present evidence of spatial autocorrelation for area data, Moran's I is often presented. Sisiopiku and Barbour (2014) use this technique in identifying food deserts in Alabama counties and Dekker et al. (2017) demonstrates spatial clustering of neighborhood dietary scores. Koleilat et al. (2012) examine the association between rates of obesity in 3 and 4 year old WIC participants and the retail food environment index in Los Angeles County zip codes, but did not perform

spatial analyses as Moran’s I found no evidence of spatial association in obesity. Such analyses typically involve spatial regression models for areal data, which typically define correlation between a unit and its neighbors (see, for example, Banerjee et al. 2014 and Cressie and Wikle 2011). For instance, Smiley et al. (2010) and Hillier et al. (2009) perform spatial lag models to estimate the density of health resources and to predict the number of outdoor advertisements near child-serving institutions, respectively. Neelon et al. (2017) found associations between nursery manager’s perceptions in food insecurity by area-level deprivation using an adjusted geographically weighted logistic regression. Further, such models can be extended to permit spatiotemporal data. Using such models, Lamichhane et al. (2015) detects a positive association between poverty and larger number of food stores at the census tract level.

The literature that involves spatial locations that are point-referenced (e.g. a pair of coordinates can be identified for that location) is much more sparse. One approach is to assess evidence of spatial clustering using Ripley’s K function. Hillier et al. (2009) employs this to describe clustering of outdoor advertisements around child serving institutions, where the authors detect such clustering in Los Angeles and Philadelphia, but not Austin. Day and Pearce (2011) uses an extension of this K function that allows for multiple types of spatial patterns (in this instance, type of food outlet) and presents evidence of spatial clustering of food outlets around schools, with more outlets closer to primary and intermediate schools than secondary schools. Auchincloss et al. (2007) detected spatial association in the availability of healthy foods and accounted for this when interpolating to non-sampled locations in their survey data. The authors obtain these predictions in two ways, first by using residual kriging after performing a standard linear regression and second by fitting a linear model with a spatial error term and then drawing predictions from this.

To approach a dataset which has two spatially referenced locations (i.e. an individual’s corner store and a rival store that they may go to), we employ the linear coregionalization models described by Goulard and Voltz (1992). These analyses draw inspiration from Banerjee et al. (2000), who accounts for spatial association in observations with pairs of locations using a bivariate Gaussian spatial process. In their application, they sought to examine postal performance by considering both the location of where a package was sent and where it was received. In our scenario, we propose that attributes of an individual’s neighborhood corner store may dictate if an individual shops at this

location and in general, their FV expenditures. In the same vein, rival stores (such as supermarkets and other corner stores) may negatively affect patronage of an individual's neighborhood corner store and also may be associated with FV expenses. Additionally, by considering a coregionalization model, we are able to examine if these corner store and rival store effects are independent.

The rest of this chapter is as follows: Section 2.2 briefly describes the study and presents the data, Section 2.3 presents the models used in the paper, and Section 2.4 provides a brief simulation comparing the spatial models employed in this paper. Section 2.5 presents the results of the data analysis and the paper concludes with a brief discussion of our findings in Section 2.6.

2.2 Data

Researchers carefully identified 4 corner stores in East Los Angeles for conversion and 4 corner stores in Boyle Heights to act as control sites. Corner store conversions included a reorganization of store items to promote healthy food purchasing, an external transformation of the store, a social marketing campaign and cooking demonstrations put on by local youth, connections to local wholesale markets, and refrigeration units. A full discussion of the study design and implementation is described by Ortega et al. (2015). In order to assess the effect of this intervention, a survey was given to residents within a five block radius of each of the eight corner stores. This community survey sought to extensively catalog the food purchasing of residents, including where they shopped, what types of food they bought, and who they were purchasing food for. As such, the survey was limited to only adults who were the main food purchaser of the family. Many other items, such as demographics, health problems, family history of residency, food program participation (such as food stamps or WIC), and others were also collected. This survey was conducted before the conversion and then again roughly one year after the conversion. There were 1,035 observations collected at baseline and 1,052 observations collected at follow-up. Roughly 60% of the individuals surveyed at baseline were surveyed again at follow-up.

There are two primary outcomes of interest: store patronage and the percent of food purchasing money spent on fruits and vegetables. More formal definitions of these quantities are provided later. In the survey, respondents are first asked where they usually shop, which allows up to four responses,

and then asked what corner stores they shop at, which allows up to three responses. In practice, no respondent listed seven stores in the survey, so there is confidence that the data adequately captures the shopping regions in the area. Stores mentioned were then catalogued, tabulated, and finally assigned geographic coordinates using Google Maps. While there were one hundred and ninety-one stores, besides the eight study stores, mentioned, it was decided to exclude stores with cumulative frequencies (combining baseline and total) less than ten, which reduced the number to forty-four. This was done so that there would be a large enough sample to draw from to make valid estimates regarding the stores effects on corner store patronage. These forty-four stores will be referred to as the "rival stores". Before exclusion, the mean number of stores reported at baseline was 2.67 while the mean at follow-up was 2.81, with standard deviations of 0.70 and 0.69 respectively. After exclusion the mean number of stores reported at baseline was 1.91 while the mean at follow-up was 2.07, each with standard deviations of 0.77. Figure 2.1 shows the relative locations of these stores, as well as the 8 corner stores included in the study.

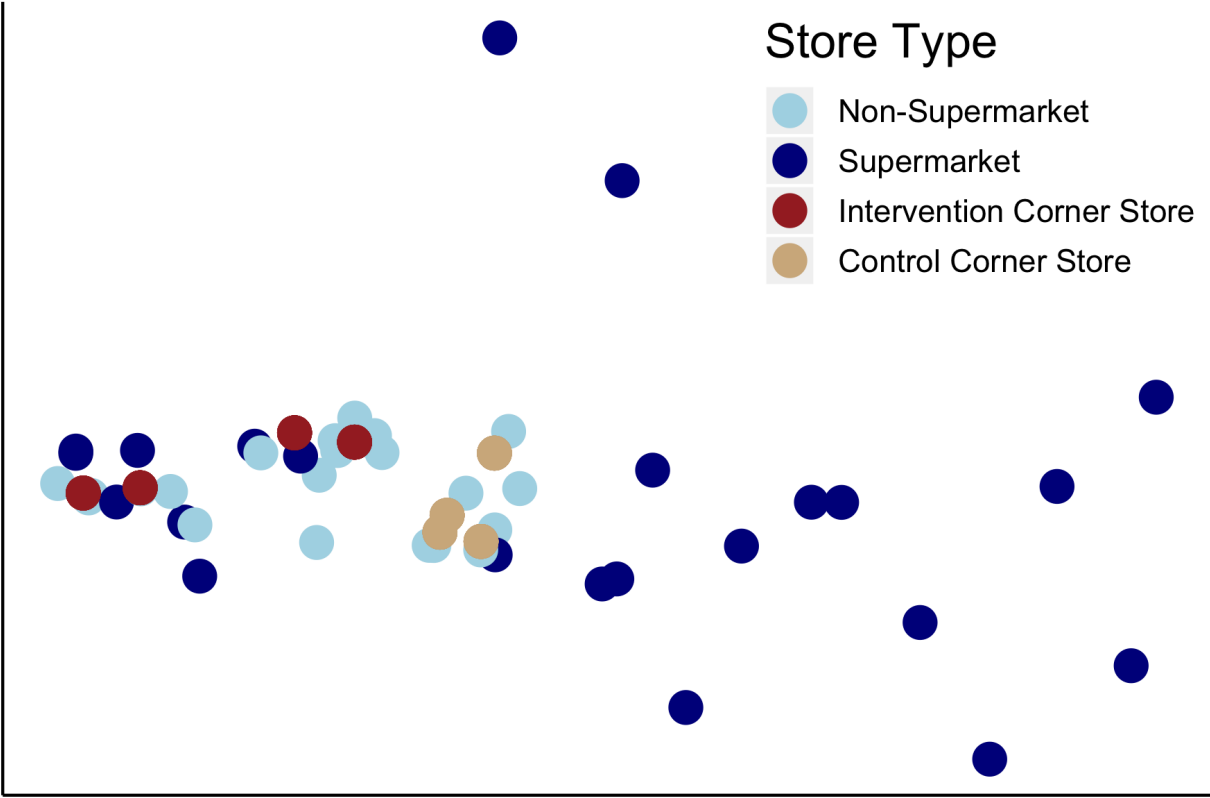


Figure 2.1: Locations of Stores

Other individual-level variables that were considered to affect patronage or fruits and vegetables purchasing were intervention status, age at time of interview, gender, household size, language of interview (English or Spanish), WIC participation, number of hours worked per week, marital status, and education level. Language spoken at home was considered (with categories English, Spanish, English/Spanish, and Other) but preliminary analyses revealed similar relationships with the outcomes when comparing Spanish only speakers and English/Spanish speakers, so Language of Interview was used as there was no missingness in this variable. While marital status initially consisted of six categories (single, married, separated, divorced, widowed, and living with a partner in a marriage-like relationship) this was consolidated whether or not the respondent was in a marriage or marriage-like relationship, as one might suspect an individual in a long term relationship might have different shopping trends than a single individual. Education reduced from twenty-seven distinct categories to three: less than a high school education, a high school education, and greater than a high school education. Income was excluded due to high levels of non-response and a concern that this might bias the analyses. A brief summary of these variables is presented in Table 1. Except for education level and age, there were differences in demographics between the baseline and follow-up data. A majority of respondents were female, Hispanic, and spoke Spanish. Roughly one-fifth utilized WIC and more than half were married. About half of the respondents had at least a high school education. The average respondent at baseline was 45.6, lived in a household with 3 other individuals, and worked 36.9 hours per week. Due to the homogeneity of ethnicity in the sample, identifying as Hispanic was not considered in the analyses.

Also included was how (drove, public transportation/ride with someone, taxi/public transportation, or other) and how frequently (high, medium, or low) a patron went to a given store. High frequency was defined as going to a store at least three times a week, medium frequency as at least once a month and at most twice a week, and low frequency as less than once a month. A single store-level covariate was included, whether or not the store was a supermarket, with the assumption that an individual's food purchasing and patronage behaviors may be very different for supermarkets compared to other types of stores (such as meat markets, corner stores, and convenience stores).

This primary goal of this paper is to assess the effect of rival stores on the outcomes of the eight

Table 2.1: Description of the Sample

	Baseline N = 1,035	Follow-up N = 1,052
Demographic	% or M (SD)	% or M (SD)
Intervention Status		
Intervention	50.1	50.6
Control	49.9	49.4
Sex		
Male	21.9	20.0
Female	78.1	80.0
Language of Interview		
English	39.5	42.7
Spanish	60.5	57.3
Ethnicity		
Hispanic	97.2	97.3
Non-Hispanic	2.8	2.7
WIC Participation		
Participant	19.5	18.5
Non-Participant	80.5	81.5
Marital Status		
Married	57.1	58.2
Not Married	42.9	41.8
Education Level		
> High School	24.9	29.7
High School	27.5	23.1
< High School	47.7	47.2
Age	45.6 (16.6)	47.1 (15.6)
Household Size	4.0 (1.9)	4.0 (2.0)
Hours Worked per Week	36.9 (15.3)	35.2 (12.1)

corner stores at baseline and follow-up. A secondary aim is to determine if controlling for these effects helps better understand the effect of the intervention of these outcomes. It is also important to note that one of the stores assigned to intervention gained a liquor license during the course of the study and did not convert their store. This paper presents an intent to treat analysis, but it is possible that this fifth “control” store diminishes the treatment effect.

2.3 A Bayesian Spatial Interaction Model

2.3.1 Logistic and Linear Models

As there is interest in determining what is associated with corner store patronage, the outcome variable is defined as follows:

$$Y_{ijk} = \begin{cases} 1 & \text{if individual } k \text{ went to corner store } i \text{ and rival store } j, \\ 0 & \text{if not,} \end{cases}$$

where $i = 1, \dots, N_{cs}$, $j = 1, \dots, N_{rs}$, and $k = 1, \dots, n$. Here N_{cs} is the number of corner stores, N_{rs} is the number of rival stores, and n is the total number of individuals. To account for covariates and neighborhood effects, one can formulate a logistic regression model,

$$\text{logit}(P(Y_{ijk} = 1)) = q_{ijk}^\top \alpha + u(s_i) + v(s_j), \quad (2.1)$$

where q_{ijk} is a vector of covariates as well as an intercept term, $u(s_i)$ and $v(s_j)$ are spatial effects corresponding to corner store at location s_i and rival store at location s_j , respectively. The covariates will include individual-level information such as gender, store-level information such as store type, and individual-store level such as an individual’s frequency of store patronage.

Our other primary outcome is percent of food purchasing money spent on fruits and vegetables, defined as

$$Y_{ijk} = \frac{\text{Dollars individual } k \text{ reported spending on FV}}{\text{Dollars individual } k \text{ reported spending on food}}$$

where $i = 1, \dots, N_{cs}$, $j = 1, \dots, N_{rs}$, and $k = 1, \dots, n_{ij}$. Here N_{cs} is the number of corner stores,

N_{rs} is the number of rival stores, and n_{ij} is the number of individuals who reside in the region with corner store i and report shopping at rival store j . In this way, individuals who reported shopping a multiple rival stores are repeated in a dataset. However, as variability in the number of stores reported was low, we are less concerned that individuals who reported more stores are over represented in the dataset. A percentage was preferred to a raw amount of money spent on produce because it controls for amount of money an individual spent on food generally, and is therefore less confounded by income than the raw amount. Similar to (2.1), one can formulate a linear regression model,

$$Y_{ijk} = q_{ijk}^\top \alpha + u(s_i) + v(s_j) + e_{ijk}, \quad e_{ijk} \sim N(0, \tau^2). \quad (2.2)$$

In both (2.1) and (2.2), interest lies in modeling the spatial effects $u(s_i)$'s and $v(s_j)$'s appropriately. We opt for process-based models because (i) we have the geographic coordinates for each store (i.e., point-referenced data), and (ii) we are interested in producing predictive probability surfaces for the spatial effects over the entire region of interest (including at arbitrary points). As there are two sets of point referenced data, those of the corner stores and those of the rival stores, we are interested in modeling a bivariate Gaussian spatial process

$$z(s) = \begin{pmatrix} u(s) \\ v(s) \end{pmatrix} \sim GP \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, K_\theta(\cdot, \cdot) \right],$$

where for any corner store location s_i and rival store location s_j , we define $K_\theta(\cdot, \cdot)$ as the cross-covariance matrix

$$K_\theta(s, t) = \text{cov}(z(s), z(t)) = \begin{pmatrix} \text{cov}(u(s), u(t)) & \text{cov}(u(s), v(t)) \\ \text{cov}(v(s), u(t)) & \text{cov}(v(s), v(t)) \end{pmatrix}.$$

The cross-covariance function maps a pair of spatial locations to a 2×2 matrix such that (i) $K_\theta(s, t) = K_\theta^\top(t, s)$ and (ii) $\sum_{i,j=1}^n a_i^\top K_\theta(s_i, s_j) a_j > 0$ for all $a_i \in \mathbb{R}^2 \setminus \{0\}$ and for any finite collection of spatial locations $\mathcal{S} = \{s_1, s_2, \dots, s_n\}$. The latter ensures that $\text{var}(z)$ is positive definite, where $z = (z(s_1), z(s_2), \dots, z(s_n))^\top$ and $z \sim N(0, K_\theta)$, where K_θ is the $2n \times 2n$ block matrix with $K_\theta(s_i, s_j)$ as the (i, j) -th block. We consider two different approaches for modeling the cross-

covariance function that we describe in the following section.

2.3.2 The Coregionalization and Conditional Approaches

For the coregionalization approach, suppose that $z(s)$ can be described as a linear combination of two underlying independent Gaussian processes. Then we write

$$z(s) = \begin{pmatrix} u(s) \\ v(s) \end{pmatrix} = \begin{pmatrix} a_{11} & 0 \\ a_{12} & a_{22} \end{pmatrix} \begin{pmatrix} w_1(s) \\ w_2(s) \end{pmatrix} = Aw(s),$$

where $w_1(s)$ and $w_2(s)$ are Gaussian processes that have unit variances and are independent of each other. Therefore, if $K_z(s, t)$ is the cross-covariance matrices corresponding to $z(s)$ then

$$K_z(s, t) = \begin{pmatrix} a_{11} & 0 \\ a_{12} & a_{22} \end{pmatrix} \begin{pmatrix} C_{w_1}(s, t) & 0 \\ 0 & C_{w_2}(s, t) \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} \\ 0 & a_{22} \end{pmatrix}. \quad (2.3)$$

If $s = t$, then $K_z(s, s) = AA^\top$ and A is the lower-triangular Cholesky factor of $K_z(s, s)$.

Alternatively, we can build bivariate process models by regressing one process on the complete realizations of the other (Cressie and Zammit-Mangion, 2016). Let $u(s)$ and $e(s)$ be zero-mean Gaussian processes with covariance functions $C_{\theta_u}(s, t)$ and $C_{\theta_e}(s, t)$, respectively, and suppose we construct

$$v(s) = \beta_0 + \int \beta(s, x)u(x)dx + e(s). \quad (2.4)$$

Then, $z(s) = (u(s), v(s))^\top$ is a legitimate bivariate Gaussian process with mean $(0, \beta_0)^\top$ and cross-covariance matrix function

$$K_z(s, t) = \begin{pmatrix} C_{\theta_u}(s, t) & \int \beta(t, x)C_{\theta_u}(s, x)dx \\ \int \beta(s, x)C_{\theta_u}(t, x)dx & \int \int \beta(s, x)\beta(t, x')C_{\theta_u}(x, x')dxdx' + C_{\theta_e}(s, t) \end{pmatrix}. \quad (2.5)$$

Since our consideration here is primary one of practical efficiency, rather than full generality, we

focus upon a useful simplification. To be precise, we set

$$\beta(s, x) = \begin{cases} \beta_1 & \text{if } s = x \\ 0 & \text{otherwise,} \end{cases}$$

which simplifies (2.4) to $v(s) = \beta_0 + \beta_1 u(s) + e(s)$ and the cross-covariance function in (2.5) reduces to the more tractable form (without the integrals),

$$K_z(s, t) = \begin{pmatrix} C_{\theta_u}(s, t) & \beta_1 C_{\theta_u}(s, t) \\ \beta_1 C_{\theta_u}(s, t) & \beta_1^2 C_{\theta_u}(s, t) + C_{\theta_e}(s, t) \end{pmatrix}. \quad (2.6)$$

Since the spatial process is introduced after adjusting for the global mean, we further set $\beta_0 = 0$, so $z(s)$ is a zero-mean process. Then, for any finite collection of n locations in \mathcal{S} , the joint distribution of u and v , i.e., the $n \times 1$ vectors with elements $u(s_i)$ and $v(s_i)$, respectively, is multivariate normal

$$\begin{pmatrix} u \\ v \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} C_{u, \theta_u} & \beta_1 C_{u, \theta_u} \\ \beta_1 C_{u, \theta_u} & \beta_1^2 C_{u, \theta_u} + C_{e, \theta_e} \end{pmatrix} \right],$$

where C_{u, θ_u} and C_{e, θ_e} are the covariance matrices corresponding to $u(s_i)$'s and $e(s_i)$'s respectively.

In fact, it is easy to establish a one-one correspondence between the process parameters in (2.6) and those in (2.3). Letting $C_{\theta_u}(s, t) = \sigma_{\theta_u}^2 C_{w_1}(s, t)$ and $C_{\theta_e}(s, t) = \sigma_{\theta_e}^2 C_{w_2}(s, t)$, we easily deduce the following correspondence:

$$\sigma_{\theta_u}^2 = a_{11}^2, \quad \beta_1 = \frac{a_{12}}{a_{11}} \quad \text{and} \quad \sigma_{\theta_e}^2 = a_{22}^2. \quad (2.7)$$

2.3.3 Modeling Using the Data

We will extend (2.1) and (2.2) to a general Bayesian spatial hierarchical model,

$$p(\theta) \times p(\sigma) \times N(z | 0, K_\theta) \times N(\alpha | 0, V_{\alpha, \sigma}) \times \prod_{i=1}^{N_{cs}} \prod_{j=1}^{N_{rs}} \prod_{k=1}^{n_i} f(Y_{ijk}), \quad (2.8)$$

where θ is the set of parameters in the spatial variance-covariance matrix for the $2n \times 1$ vector of process realizations $z = (z^\top(s_1), \dots, z^\top(s_n))^\top$ with $n = N_{cs} + N_{rs}$, and σ is a set of parameters in the prior variance-covariance matrix for α . In the logistic case, $f(Y_{ijk}) = \text{Ber}(Y_{ijk} | p_{ijk} = \text{logit}^{-1}(q_{ijk}^\top \alpha + u(s_i) + v(s_j)))$, and in the linear case, $N(\mu_{ijk} | \mu_{ijk} = q_{ijk}^\top \alpha + u(s_i) + v(s_j), \tau^2)$.

Four models were considered for each outcome. In general, $q_{ijk}^\top \alpha = x_{1,k}^\top \zeta + x_{2,j}^\top \gamma + x_{3,jk}^\top \delta$. Individual-level covariates were selected using simple linear regressions, as additive missingness from covariates was a concern. Models of both outcomes had the individual-level covariate vector, x_{1k} include an intercept, gender, the number of persons that individual lives with, intervention status, time-point (baseline or follow-up) and the interaction between intervention status and time point (to detect an intervention effect). The linear regression also included WIC participation, weekly food budget, and language of interview. The rival store covariate for store j , $x_{2,j}$, is an indicator variable denoting whether the store being supermarket or not. Finally, $x_{3,jk}$, comprise covariates specific to rival store j and individual k , include the method of transportation used and frequency of shopping. The prior distributions for ζ , γ and δ are zero-mean normal with variance-covariance matrices $\sigma_\zeta^2 I_{p_1}$, $\sigma_\gamma^2 I_{p_2}$, and $\sigma_\delta^2 I_{p_3}$, respectively, where p_1 , p_2 , and p_3 are the corresponding dimensions of ζ , γ and δ . For simplicity, we set $\sigma^2 = \sigma_\zeta^2 = \sigma_\gamma^2 = \sigma_\delta^2$ and a half-cauchy prior distribution was assigned to this parameter. $V_{\alpha,\sigma}$ is block-diagonal with the preceding variance-covariance matrices as the diagonal blocks.

The first model, referred to as the ‘‘simple model’’, does not incorporate any spatial random effects and is obtained by setting $z(s) = (u(s), v(s))^\top = 0$ in (2.8), which excludes $p(\theta)$ and $N(z | 0, K_\theta)$ from (2.8). The most complex model incorporates the proposed spatial random effects using the coregionalization technique presented in 2.3.2. The matrix K_θ in (2.8) is a $2n$ by $2n$ block variance-covariance matrix with $K_z(s_i, s_j)$ as the (i, j) -th block, where $K_z(s_i, s_j)$ is as described in (2.3) with $C_{w_i}(t) = \exp(-\phi_i t)$ for $i = 1, 2$, reflecting the exponential covariance function. The parameter $\theta = \{\phi_1, \phi_2, A\}$ in (2.8). These are taken to be independent apriori, with ϕ_1 and ϕ_2 having uniform prior distributions and AA^\top an Inverse-Wishart prior distribution. We refer to this as the ‘‘coregionalization model’’ and it will be the fourth model presented. The second model, referred to as the ‘‘independent model’’, modifies the ‘‘coregionalization model’’ as well, restricting A to a diagonal matrix by setting $a_{12} = 0$ in (2.3), simplifying $K_z(s_i, s_j)$ to a diagonal matrix. The

third model simplifies the “coregionalization” model by setting $\phi_1 = \phi_2$, thus $C_{w_1}(t) = C_{w_2}(t)$, and θ only comprises ϕ_1 and A . This is called the “separable model”.

Given the correspondence between (2.3) and (2.6) and that we have already obtained posterior samples for the parameters in (2.3), there is no benefit to estimating the conditional model again. We can obtain the posterior samples for the process parameters for (2.6) from those of (2.3) using the correspondence in (2.7).

2.3.4 A Conjugate Bayesian Model

Here we present a conjugate bayesian model for the linear case. Let y_{ijk} be a continuous measurement or score associated with individual $k = 1, 2, \dots, n_{ij}$ who visited corner store $i = 1, 2, \dots, n_{cs}$ and rival store $j = 1, 2, \dots, n_{rs}$ and q_{ijk} be a vector of covariates associated with the outcome through a regression model

$$y_{ijk} = q_{ijk}^\top \alpha + u(s_i) + v(s_j) + e_{ijk}, \quad e_{ijk} \sim N(0, \tau^2). \quad (2.9)$$

Here α is the vector of unknown regression coefficients, $u(s_i)$ and $v(s_j)$ are spatial random effects corresponding to corner store i and rival store j , and e_{ijk} represents Gaussian nonspatial error or noise with zero mean and variance τ^2 . Stacking measurements over individuals, we obtain

$$y_{ij} = Q_{ij} \alpha + 1 \otimes u(s_i) + 1 \otimes v(s_j) + e_{ij}, \quad e_{ij} \stackrel{ind}{\sim} N(0, \tau^2 D_{ij}),$$

where y_{ij} is the $n_{ij} \times 1$ vector obtained by stacking y_{ijk} 's over individual indices, Q_{ij} is the matrix with rows q_{ijk}^\top , 1 is the $n_{ij} \times 1$ vector of ones, and e_{ij} is an $n_{ij} \times 1$ vector of nonspatial error with $n_{ij} \times n_{ij}$ covariance matrix $\tau^2 D_{ij}$ and D_{ij} is diagonal with variance components. Next, stacking over the corner stores and then the rival stores produces

$$y = Q \alpha + R_1 u + R_2 v + e, \quad e \sim N(0, \tau^2 D), \quad (2.10)$$

where y , u , v and e are vectors of y_{ij} 's, $u(s_i)$'s, $v(s_j)$'s and e_{ij} 's, respectively, Q is a matrix formed by stacking the Q_{ij} 's, R_1 and R_2 are design matrices corresponding to u and v , and $\tau^2 D$ is the

covariance matrix for e . We rewrite (2.10) as

$$y = X\Omega + e, \quad e \sim N(0, \tau^2 D),$$

where $X = [Q : R_1 : R_2]$ is an $n \times r$ matrix and $\Omega = [\alpha^\top : u^\top : v^\top]^\top$ is an $r \times 1$ column vector. Assuming $\Omega \sim N(\mu_\Omega, \tau^2 V_\Omega)$ and $\tau^2 \sim \text{IG}(a, b)$, the joint posterior distribution is

$$\begin{aligned} & \text{IG}(\tau^2 | a, b) \times N(\Omega | \mu_\Omega, \tau^2 V_\Omega) \times N(y | X\Omega, \tau^2 D) \\ & \propto \text{IG}(\tau^2 | a_*, b_*) \times N\left(\Omega | \hat{\Omega}, \tau^2 (\tilde{X}^\top \tilde{V}^{-1} \tilde{X})^{-1}\right), \end{aligned} \quad (2.11)$$

where $\hat{\Omega} = (\tilde{X}^\top \tilde{V}^{-1} \tilde{X})^{-1} \tilde{X}^\top \tilde{V}^{-1} \tilde{y}$, $\tilde{y} = \begin{bmatrix} y \\ \mu_\Omega \end{bmatrix}$, $\tilde{X} = \begin{bmatrix} X \\ I \end{bmatrix}$, and $\tilde{V} = \begin{bmatrix} D & 0 \\ 0 & V_\Omega \end{bmatrix}$, $a_* = a + \frac{n}{2}$ and $b_* = b + \frac{1}{2} (\tilde{y} - \tilde{X} \hat{\Omega})^\top \tilde{V}^{-1} (\tilde{y} - \tilde{X} \hat{\Omega})$.

V_Ω has been chosen and assuming D is fixed, sampling from this posterior is achieved by first drawing τ^2 from its marginal posterior $\text{IG}(a_*, b_*)$ and then drawing Ω from its conditional posterior distribution $N\left(\hat{\Omega}, \tau^2 (\tilde{X}^\top \tilde{V}^{-1} \tilde{X})^{-1}\right)$ given the sampled value of τ^2 . Repeating this M times results in M exact posterior samples from (2.11).

2.3.5 A Finite Population Framework

As only 60% of individuals participated at both baseline and follow-up, suppose we are interested in estimating the average percent of food purchasing money spent on fruits and vegetables at baseline and follow-up, as modeled in (2.2). To achieve this, we predict responses at follow-up for individuals who only responded at baseline and predict responses at baseline for individuals who only responded at follow-up. All covariates associated with these individuals remain unchanged, except the indicator of time point. We therefore modify our notation slightly to make the distinction that $Y_{ijk}^{(1)}$ is the k^{th} individual in the region served by corner store i who went to the j^{th} rival store at baseline, $Y_{ijk}^{(2)}$ is such an individual at follow-up. These individual's corresponding covariates would then be denoted $q_{ijk}^{(1)}$ and $q_{ijk}^{(2)}$, respectively. Decomposing $n_{ij} = n_{ij}^{(both)} + n_{ij}^{(1)} + n_{ij}^{(2)}$, where $n_{ij}^{(both)}$, $n_{ij}^{(1)}$, and $n_{ij}^{(2)}$ as the observed number of individuals in the region served by corner store i who shop at rival store j

at both timepoints, only at baseline, and only at follow-up, respectively. Then assuming that unobserved individuals at baseline and follow-up did not change the rival stores they shop at, the number of individuals in the baseline and follow-up finite populations is $n_{ij}^{(max)} = n_{ij}^{(both)} + 2n_{ij}^{(1)} + 2n_{ij}^{(2)}$. Using this fact, we define the set of individuals at baseline who are observed to be $Y_s^{(1)} = [Y_{111}^{(1)}, \dots, Y_{11(n_{11}^{(both)} + n_{11}^{(1)})}^{(1)}, \dots, Y_{N_{cs}N_{rs}1}^{(1)}, \dots, Y_{N_{cs}N_{rs}(n_{N_{cs}N_{rs}}^{(both)} + n_{N_{cs}N_{rs}}^{(1)})}^{(1)}]^\top$ and unobserved to be $Y_{ns}^{(1)} = [Y_{11(n_{11}^{(both)} + n_{11}^{(1)} + 1)}^{(1)}, \dots, Y_{11n_{11}^{(max)}}^{(1)}, \dots, Y_{N_{cs}N_{rs}(n_{N_{cs}N_{rs}}^{(both)} + n_{N_{cs}N_{rs}}^{(1)} + 1)}^{(1)}, \dots, Y_{N_{cs}N_{rs}n_{N_{cs}N_{rs}}^{(max)}}^{(1)}]^\top$. Similarly, for follow-up, we have: $Y_s^{(2)} = [Y_{111}^{(2)}, \dots, Y_{11(n_{11}^{(both)} + n_{11}^{(2)})}^{(2)}, \dots, Y_{N_{cs}N_{rs}1}^{(2)}, \dots, Y_{N_{cs}N_{rs}(n_{N_{cs}N_{rs}}^{(both)} + n_{N_{cs}N_{rs}}^{(2)})}^{(2)}]^\top$ and $Y_{ns}^{(2)} = [Y_{11(n_{11}^{(both)} + n_{11}^{(2)} + 1)}^{(2)}, \dots, Y_{11n_{11}^{(max)}}^{(2)}, \dots, Y_{N_{cs}N_{rs}(n_{N_{cs}N_{rs}}^{(both)} + n_{N_{cs}N_{rs}}^{(2)} + 1)}^{(2)}, \dots, Y_{N_{cs}N_{rs}n_{N_{cs}N_{rs}}^{(max)}}^{(2)}]^\top$. We now stack the observed measurements and unobserved measurements into a single vector, $Y = [Y_s^\top, Y_{ns}^\top]^\top = [Y_s^{(1)\top}, Y_s^{(2)\top}, Y_{ns}^{(1)\top}, Y_{ns}^{(2)\top}]^\top$. Defining the total population size at baseline to be T , e.g. the dimension of $[Y_s^{(1)\top}, Y_{ns}^{(1)\top}]^\top$ is $T \times 1$, we also have that the total population size at follow-up is T and therefore the dimension of Y is $2T \times 1$. We construct the following linear regression model:

$$\begin{bmatrix} Y_s \\ Y_{ns} \end{bmatrix} = \begin{bmatrix} Q_s \\ Q_{ns} \end{bmatrix} \alpha + \begin{bmatrix} R_{1s} \\ R_{1ns} \end{bmatrix} v + \begin{bmatrix} R_{2s} \\ R_{2ns} \end{bmatrix} \nu + \begin{bmatrix} \epsilon_s \\ \epsilon_{ns} \end{bmatrix}; \quad \begin{bmatrix} \epsilon_s \\ \epsilon_{ns} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tau^2 I\right). \quad (2.12)$$

Here Q_s , R_{1s} , and R_{2s} are constructed such that the r^{th} element of Y_s , corresponds to the r^{th} row of each of these matrices, similar to those presented in 2.3.4. For example, if the r^{th} element of Y_s is $Y_{121}^{(1)}$, the r^{th} row of Q_s will be the corresponding set of covariates, $q_{121}^{(1)}$. Similarly, the r^{th} row of R_{1s} will a vector of length N_{cs} with a 1 in the first position and 0 elsewhere, while the r^{th} row of R_{2s} will a vector of length N_{rs} with a 1 in the second position and 0 elsewhere. The matrices Q_{ns} , R_{1ns} , and R_{2ns} are constructed similarly, corresponding to Y_{ns} . Additionally, ϵ_s and ϵ_{ns} are constructed in a manner identical to Y_s and Y_{ns} .

We have that $p(Y_{ns} | Y_s, \alpha, v, \nu, \tau^2) = N(Y_{ns} | \mu_{ns|s}, \tau^2 I)$, where $\mu_{ns|s} = Q_{ns}\alpha + R_{1ns}v + R_{2ns}\nu$. Therefore for each posterior sample of $\{\alpha, v, \nu, \tau^2\}$ from 2.8, we draw $Y_{ns} \sim N(Y_{ns} | \mu_{ns|s}, \tau^2 I)$. These samples from the posterior predictive distribution can be then used to obtain posterior finite population estimates. Specifically, at each iteration g , estimates of the unobserved units are drawn

and estimates for the baseline population mean,

$$\bar{Y}^{(1,g)} = \frac{1}{T} \left(\sum_{i=1}^{N_{cs}} \sum_{j=1}^{N_{rs}} \sum_{k=1}^{n_{ij}^{(both)} + n_{ij}^{(1)}} Y_{ijk}^{(1)} + \sum_{i=1}^{N_{cs}} \sum_{j=1}^{N_{rs}} \sum_{k=n_{ij}^{(both)} + n_{ij}^{(1)} + 1}^{n_{ij}^{(max)}} Y_{ijk}^{(1,g)} \right),$$

and the follow-up population mean,

$$\bar{Y}^{(2,g)} = \frac{1}{T} \left(\sum_{i=1}^{N_{cs}} \sum_{j=1}^{N_{rs}} \sum_{k=1}^{n_{ij}^{(both)} + n_{ij}^{(2)}} Y_{ijk}^{(2)} + \sum_{i=1}^{N_{cs}} \sum_{j=1}^{N_{rs}} \sum_{k=n_{ij}^{(both)} + n_{ij}^{(2)} + 1}^{n_{ij}^{(max)}} Y_{ijk}^{(2,g)} \right),$$

are calculated.

2.3.6 Model Comparison and Assessment

Model fit is evaluated using the Watanabe-Akaike Information Criteria (WAIC), expressed as $WAIC = -2\widehat{elpd} = -2(\widehat{lpd} + \hat{p}_{WAIC})$ in Vehtari et al. (2017), where \widehat{elpd} is the estimated expected log pointwise predictive density. To calculate this, at each iteration, $g = 1, \dots, G$, $p(y_h | \Theta^{(g)})$ is computed; the likelihood of each observed value conditional on that iteration's parameters. For a sample of size k , the estimated log pointwise predictive density is the sum of the log average likelihood for each observation, $\widehat{lpd} = \sum_{h=1}^k \log \left[\frac{1}{G} \sum_{g=1}^G p(y_h | \Theta^{(g)}) \right]$. The sample variance of the log-likelihood for each observation is $s_{lp(y_h)}^2 = \frac{1}{G-1} \sum_{g=1}^G \left[\log(p(y_h | \Theta^{(g)})) - \frac{1}{G} \sum_{g=1}^G \log(p(y_h | \Theta^{(g)})) \right]^2$ and the estimated effective number of parameters is the sum of these variances: $\hat{p}_{WAIC} = \sum_{h=1}^k s_{lp(y_h)}^2$.

2.4 Simulation

In order to assess the ability of our models to detect the true underlying process, three simulations were performed using the conjugate model presented in 2.3.4. The first simulation assumed that true underlying process is described by the coregionalization method in Section 2.3.2 by setting the $u(s_i)$ and $v(s_j)$ as the true random effects coming from this process. At each iteration, 20 locations (10 locations for each process) had coordinates randomly sampled from two Uniform(0,50)

distributions. 10 replicates were sampled at each location pair such that

$$Y_{ij} \sim N(\mu_{ij}, \sigma^2); \quad \mu_{ij} = \beta_0 + u(s_i) + v(s_j); \quad i = 1, \dots, 10; \quad j = 1, \dots, 10.$$

β_0 was assigned -5 and σ^2 9. To define $u(s)$ and $v(s)$, ϕ_1 was set to 3, ϕ_2 to 1 to reflect similar conditions to the true data, and A to $\begin{pmatrix} 2 & -1 \\ 0 & 1 \end{pmatrix}$. One quarter of the data was set aside as a test set and an exact sampler was employed to fit the coregionalization model, separable model, and independent model to the remaining data. This process was repeated 50 times, so that the KL-Divergence (Kullback and Leibler 1951) could be calculated to assess deviations from the true parameter values. For this statistic, higher values indicate distributions that are farther from the true distribution. A second simulation repeated this process, but assumed that the true underlying process is the separable coregionalization model described in Section 2.3. Thus A was altered to $\begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$. The final simulation repeated the first process but made the assumption that the true underlying process is the independent coregionalization model described in Section 2.3. Thus $\phi_1 = \phi_2 = 10$.

Table 2.2: Comparing KL-Divergence Results of 3 Simulations

Model	Simulation (Truth)		
	Coregionalization Median (95% CI)	Separable Median (95% CI)	Independent Median (95% CI)
Coregionalization	154.5 (114.2, 193.4)	164.9 (124.6, 215.0)	276.3 (230.2, 323.2)
Separable	153.2 (113.3, 196.4)	166.5 (121.8, 221.3)	276.7 (228.2, 319.7)
Independent	236.6 (189.6, 307.2)	224.5 (178.1, 285.5)	275.6 (224.6, 325.8)

Table 2.2 suggests that with respect to the KL-Divergence measure, while coregionalization models are adept at detecting truly correlated processes, there are no benefits from using this model instead of a separable model. While different values of ϕ_1 were considered, it's possible that larger differences in effect spatial range between the two processes may result in the coregionalization model fitting better. Density plots corresponding to Table 2.2 are provided in Figure 2.2.

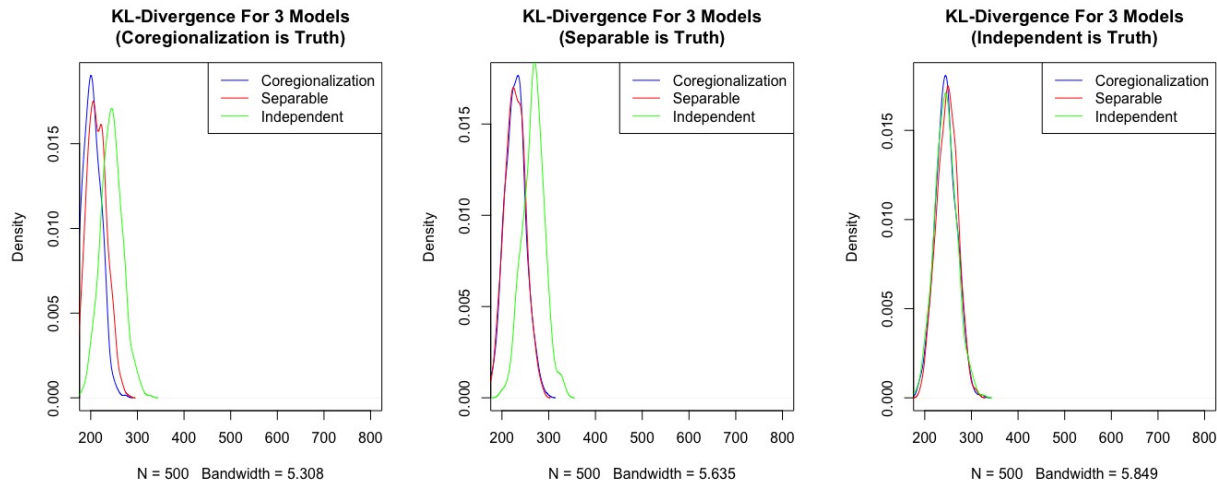


Figure 2.2: K-L Divergence Comparisons from 3 Simulations

2.5 Data Modeling

2.5.1 Implementation

Each logistic model was run for 3 chains of 12,000 iterations with 500 burn-in. Initial values for regression coefficients were drawn from a Normal(0, 1) distribution, while the initial values for ϕ parameters were set to 1, the σ^2 (and τ^2 , in the linear case) parameters set to 10, and the spatial random effects set to 0. The regression coefficients were given Normal(0, σ^2) priors, with σ having a half-cauchy prior (Gelman et al., 2008) centered at 0 with 5 degrees of freedom. In the coregionalization and independent models, the priors for ϕ_1 and ϕ_2 were established by allowing the spatial range to vary between 10% to 40% of the maximum distance between stores, which was 5.46 for the corner stores and 14.23 for the rival stores. As $3/\phi$ is the effective spatial range, ϕ_1 was given a Uniform(1.37, 5.49) prior and ϕ_2 a Uniform(0.53, 2.11) prior. The maximum distance between all stores was 15.86, so using the same spatial range guidelines ϕ_1 was given a Uniform(0.47, 1.89) in the Separable Model. An Inverse-Wishart prior was set on AA^\top with a starting value of the identity matrix. In the independent case, rather than employing an Inverse-Wishart distribution, A_{11} and A_{22} were given Uniform(0,10) priors to save on computation costs. Mixing was assessed using the Gelman-Rubin Diagnostic. After combining the three chains, fit was compared using the WAIC. All modeling was performed using the STAN software (Stan Development Team and others,

2018) in R 3.3.3 (R Core Team, 2018).

2.5.2 Results

The results of the logistic model predicting store patronage is presented in Table 2.3. While it is evident that the spatial models fit the data better than the non-spatial model, there does not seem to be a large difference in WAIC between the three spatial models. As the independent model provides negligibly better fit and the 95% Credible Interval for the A_{21} parameters in Models 3 and 4 include 0, we conclude that the corner store and rival store processes may in fact be independent. Importantly, we do not find any evidence of an intervention effect improving patronage. Additionally, we do find in both Model 1 and 4 that the odds are lower in the intervention sites at baseline.

Few covariates were associated with patronage, but the supermarket covariate is significant in Models 1 and 4, and sees a similar trend in Models 2 and 3 (although their credible intervals contain 0). This suggests that individuals are more likely to be patrons of their local corner store and a supermarket than their local corner store and another non-supermarket. When considering this result and the locations of the study stores (presented in Figure 2.1), we note that the the intervention stores have twelve non-supermarkets within close proximity, compared to six for the comparison stores. This higher density of rival, non-supermarket stores may have contributed to a non-significant intervention effect, as there is more competition among these stores. Additionally, transportation type and frequency of visit to the rival stores were not found to be associated with the outcome.

Finally, the posterior estimates of the spatial random effects of the coregionalization and independent models are provided in Figures 2.3 and 2.4. In both cases, the rival store random effects identify five stores with positive random effects. Further work is needed understand what qualities of these stores can explain such results. We note that while the intercept term in the independent model is closer to 0, the negative random effects in the independent case are negative, suggesting that the mean has simply been absorbed into the random effect term.

Evidence of process independence can be seen in the 95% Credible Interval for the A_{21} param-

eter, which in both models contains 0.

Table 2.3: Results of Logistic Model Predicting Store Patronage

	Model 1	Model 2	Model 3	Model 4
	M (95% CI)	M (95% CI)	M (95% CI)	M (95% CI)
Intercept	-5.92 (-6.2, -5.64)	-0.07 (-0.41, 0.19)	-4.63 (-7.05, 0.05)	-6.08 (-7.15, -4.88)
Male	-0.04 (-0.19, 0.11)	-0.05 (-0.20, 0.04)	-0.11 (-0.28, 0.03)	-0.13 (-0.29, 0.03)
Household Size	-0.02 (-0.05, 0.01)	0.01 (-0.02, 0.04)	0.01 (-0.02, 0.04)	0.01 (-0.02, 0.04)
Intervention	-1.12 (-1.32, -0.92)	-0.03 (-0.37, 0.20)	-1.11 (-2.49, 0.07)	-1.29 (-2.32, -0.33)
Followup	0.01 (-0.13, 0.15)	0.01 (-0.09, 0.10)	0.00 (-0.15, 0.14)	-0.01 (-0.16, 0.14)
Intervention×Followup	0.13 (-0.14, 0.41)	0.03 (-0.08, 0.23)	0.12 (-0.13, 0.41)	0.14 (-0.14, 0.42)
Supermarket	2.32 (2.09, 2.56)	0.09 (-0.09, 0.86)	1.42 (-0.03, 2.54)	1.71 (0.82, 2.57)
Transportation				
Don't go (ref)				
Drive	0.52 (-1.05, 2.10)	-0.01 (-0.18, 0.13)	-0.03 (-1.46, 1.37)	-0.03 (-1.60, 1.54)
Walk/Public Trans.	-0.04 (-1.62, 1.56)	-0.05 (-0.36, 0.08)	-0.27 (-1.77, 1.11)	-0.32 (-1.90, 1.28)
Taxi/Ride	0.52 (-1.07, 2.12)	-0.02 (-0.24, 0.13)	-0.10 (-1.56, 1.29)	-0.11 (-1.71, 1.47)
Other	0.80 (-0.94, 2.52)	0.02 (-0.19, 0.32)	0.24 (-1.31, 1.84)	0.30 (-1.43, 2.03)
Freq. of Visit				
Don't go (ref)				
Low	0.52 (-1.09, 2.12)	-0.06 (-0.40, 0.07)	-0.30 (-1.79, 1.08)	-0.37 (-1.96, 1.23)
Medium	0.64 (-0.94, 2.21)	-0.02 (-0.19, 0.11)	-0.03 (-1.43, 1.39)	-0.03 (-1.60, 1.54)
High	0.62 (-1.12, 2.34)	0.01 (-0.20, 0.26)	0.10 (-1.44, 1.69)	0.12 (-1.61, 1.84)
σ	2.05 (1.38, 3.16)	0.10 (0.01, 0.40)	1.55 (0.02, 2.96)	1.99 (1.29, 3.11)
ϕ_1		3.51 (1.5, 5.39)	1.35 (0.49, 1.87)	3.65 (1.66, 5.39)
ϕ_2		1.32 (0.57, 2.07)		1.32 (0.57, 2.07)
A_{11}		6.11 (3.54, 9.43)	1.02 (0.46, 3.45)	0.72 (0.42, 1.28)
A_{21}		0.00 (0.00, 0.00)	-0.17 (-1.52, 2.47)	-0.39 (-1.29, 0.69)
A_{22}		1.86 (1.39, 2.49)	1.77 (1.02, 3.74)	1.29 (0.87, 1.77)
WAIC	10281.6	8420.9	8428.2	8426.5

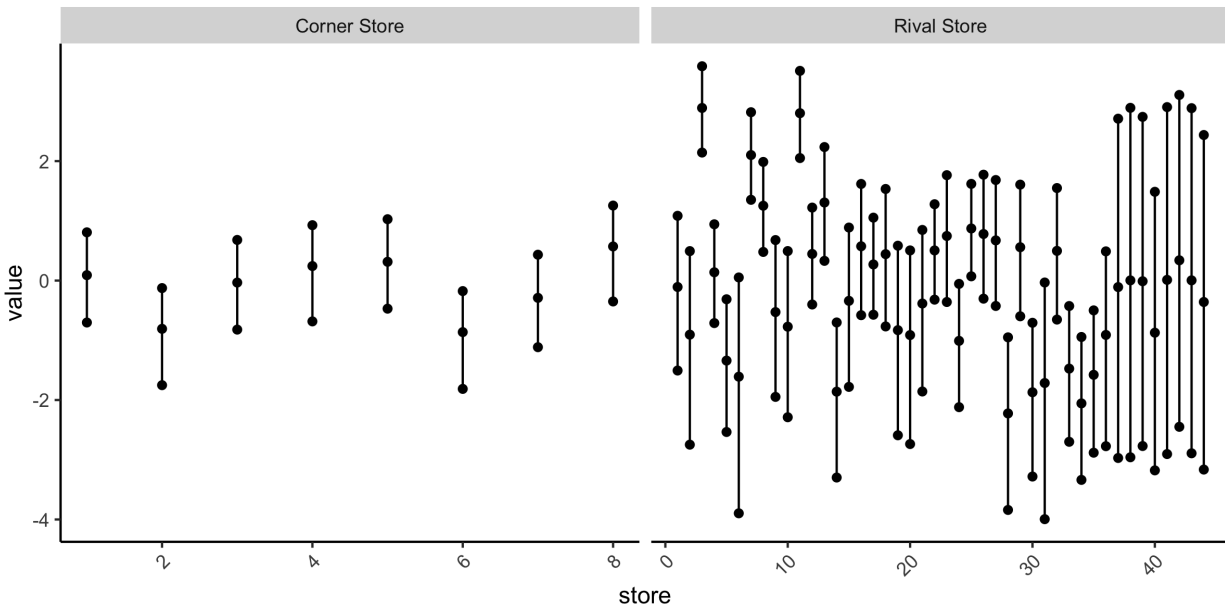


Figure 2.3: Corner and Rival Store Random Effects from the Coregionalization Model

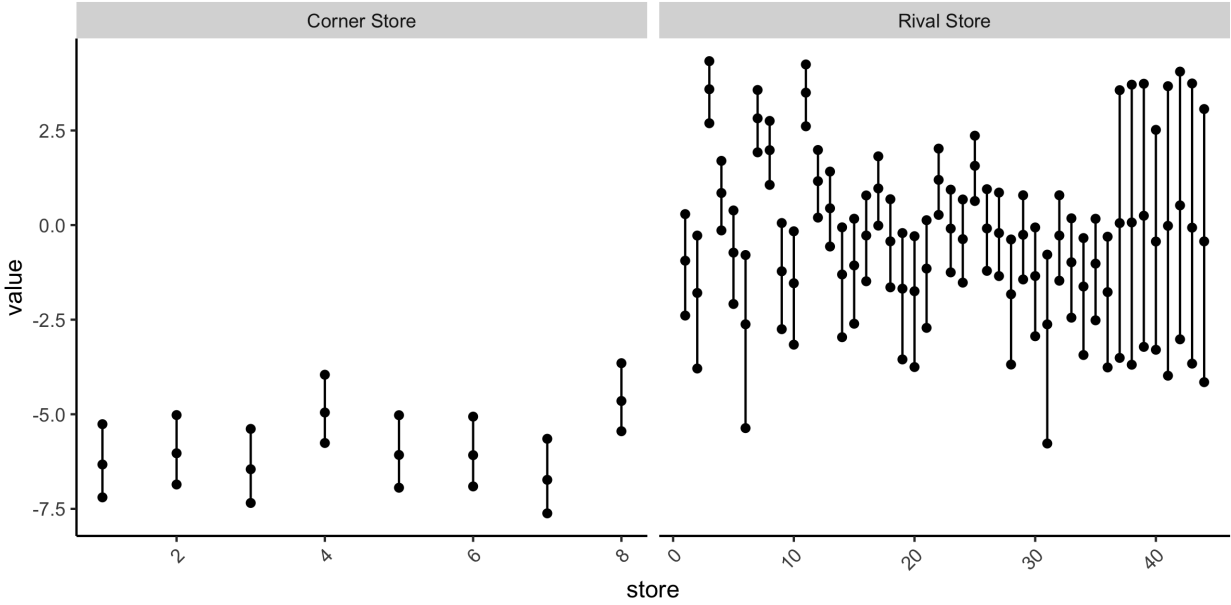


Figure 2.4: Corner and Rival Store Random Effects from the Independent Model

In the linear models predicting percent of food budget spent on FV presented in Table 2.4, there is little evidence of spatial association in the outcome, as the non-spatial model fits better than the separable and coregionalization model and is very similar to the independent model. While the independent model shows a slight improvement to model fit, this may be due to simply adding a random effect term. The poorer model fit of 3 and 4 suggests that permitting association between the two spatial random effects is not appropriate in this instance. Interpreting the coefficients from the independent spatial model, we find that men spent on average about 5% less on FV than women and Spanish speakers spent about 5% more on FV than non-Spanish speakers. No differences by WIC participation, food budget, store type, or transportation were detected. While there was no difference between intervention and control sites at either time point, while the interaction term has a 95% credible interval that contains 0, as a large majority of this interval excludes 0, we note this may be evidence of a slight, positive intervention effect. Similarly, while the credible interval contains 0, there may be some evidence that individuals who report shopping at specific stores more frequently spend 2% less of their food budget on FV.

Comparing the finite population estimates of baseline and follow-up, we see that the addition of random effects resulted in more variability and an overall decrease in the estimates. However, as Models 3 and 4 may not be appropriate for this data, we focus on the non-spatial and independent

spatial models. Comparing these estimates to the averages found from the raw data, 34.99 at baseline and 34.46 at follow-up, both models agree that the finite population mean is lower than the sampled data at baseline. At follow-up, the non-spatial model presents a 2% increase, while the independent spatial model presents a 1.6% decrease. While the apparent disagreement, we note that the estimate from non-spatial case is contained in the credible interval of the spatial case (as is the mean from the raw data), and cannot conclude that there is a significant difference.

Table 2.4: Results of Linear Model Predicting Percentage of Food Expenditure Spent on FV

	Model 1 M (95% CI)	Model 2 M (95% CI)	Model 3 M (95% CI)	Model 4 M (95% CI)
Intercept	0.35 (0.32, 0.37)	0.12 (-0.03, 0.37)	0.00 (-0.05, 0.06)	0.01 (-0.04, 0.07)
WIC Participant	-0.009 (-0.022, 0.004)	-0.009 (-0.021, 0.005)	-0.008 (-0.021, 0.004)	-0.008 (-0.021, 0.004)
Spanish Speaker	0.05 (0.04, 0.06)	0.05 (0.04, 0.06)	0.05 (0.04, 0.06)	0.05 (0.04, 0.06)
Food Budget	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)
Male	-0.05 (-0.06, -0.04)	-0.05 (-0.06, -0.04)	-0.05 (-0.06, -0.03)	-0.05 (-0.06, -0.03)
Household Size	0.001 (-0.002, 0.004)	0.001 (-0.002, 0.004)	0.001 (-0.002, 0.004)	0.001 (-0.002, 0.004)
Intervention	0.00 (-0.02, 0.01)	0.01 (-0.04, 0.06)	0.00 (-0.05, 0.06)	0.00 (-0.05, 0.06)
Followup	-0.009 (-0.023, 0.006)	-0.009 (-0.023, 0.005)	-0.01 (-0.024, 0.003)	-0.011 (-0.024, 0.003)
Intervention×Followup	0.013 (-0.007, 0.033)	0.014 (-0.004, 0.033)	0.017 (-0.002, 0.036)	0.018 (-0.001, 0.037)
Supermarket	0.00 (-0.02, 0.02)	0.00 (-0.02, 0.02)	0.00 (-0.04, 0.04)	0.00 (-0.04, 0.05)
Transportation				
Drive (ref)				
Walk/Public Trans.	0.01 (-0.01, 0.02)	0.01 (-0.01, 0.02)	0.01 (-0.01, 0.02)	0.01 (-0.01, 0.02)
Taxi/Ride	0.01 (-0.01, 0.02)	0.01 (-0.01, 0.02)	0.01 (0.00, 0.02)	0.01 (0.00, 0.02)
Other	-0.01 (-0.05, 0.03)	0.00 (-0.04, 0.03)	0.00 (-0.03, 0.03)	0.00 (-0.04, 0.03)
Freq. of Visit				
Low	0.00 (-0.02, 0.01)	0.00 (-0.02, 0.01)	0.00 (-0.02, 0.01)	0.00 (-0.02, 0.01)
Medium (ref)				
High	-0.03 (-0.057, -0.003)	-0.024 (-0.051, 0.001)	-0.02 (-0.045, 0.004)	-0.02 (-0.045, 0.004)
σ	0.10 (0.07, 0.15)	0.05 (0.02, 0.13)	0.03 (0.02, 0.04)	0.03 (0.02, 0.04)
σ_e	0.15 (0.15, 0.16)	0.15 (0.15, 0.16)	0.15 (0.15, 0.16)	0.15 (0.15, 0.16)
ϕ_1		3.12 (1.42, 5.36)	0.51 (0.47, 0.65)	1.64 (1.37, 2.7)
ϕ_2		1.32 (0.57, 2.07)		1.32 (0.57, 2.07)
A_{11}		0.24 (0, 0.61)	0.37 (0.24, 0.58)	0.4 (0.26, 0.64)
A_{21}		0.00 (0.00, 0.00)	0.00 (-0.06, 0.07)	0 (-0.05, 0.06)
A_{22}		0.01 (0, 0.02)	0.18 (0.14, 0.22)	0.17 (0.13, 0.21)
$\bar{Y}^{(1)} * 100\%$	32.84 (32.47, 33.20)	26.69 (22.06, 33.16)	23.31 (21.46, 25.23)	23.54 (21.65, 25.64)
$\bar{Y}^{(2)} * 100\%$	36.54 (36.33, 36.74)	32.86 (30.08, 36.70)	30.78 (29.61, 31.96)	30.92 (29.72, 32.22)
WAIC	-3293.5	-3296.5	-3276.7	-3272.1

2.6 Discussion

This paper has demonstrated a method for flexible modeling where the researchers suspect that multiple spatial processes are influencing the outcome. In our case, we hypothesized that there

might be two spatial processes, one related to the location of an individual's closest corner store and another related to the location of other rival stores that an individual shopped at. Both binary and continuous data were modeled using this technique and in the logistic case, the inclusion of spatial random effects resulted in significant improvements in model fit compared to the standard logistic models. By employing coregionalization models, we allowed for, but did not force, spatial processes to be correlated with one another. In both outcomes, there was no evidence that the local corner store spatial process and rival store spatial process were correlated.

While much work has been done to describe the food environment and health, regression models which account for spatial models are not commonly used, particularly if locations are point-referenced. Specifically, we believe that the use of coregionalization models as a technique to account for pairs of geographic points (rather than a single point) attributed with one observation may be beneficial to this field of work. For instance, pairs such as home-work, work-preferred supermarket, school-advertisements, could all be described using this technique.

Chapter 3

Multistage Bayesian FPS Models for Spatial Data with Ignorable Designs

3.1 Introduction

Bayesian finite population survey sampling is essentially model-based (see, e.g., Little 2004). In a Gaussian setting, Scott and Smith (1969) devised Bayesian hierarchical models for inferring with two-stage designs, while Malec and Sedransk (1985) extended this framework to general multi-stage (more than two-stages) models and also discussed handling unknown variances. Our current contribution focuses on incorporating survey sampling designs within (1.6). We extend this framework to spatial process settings under the context of ignorable sampling designs (Rubin 1976; Sugden and Smith 1984), where the probability of element selection is assumed independent of the measured outcome given the design variables. We specifically develop the distribution theory and algorithms for implementing (1.6) in the context of two-stage designs that encompass simple random, cluster and stratified sampling (as defined in Cochran, 1977) as special cases. Extension of this work to multi-stage present no new methodological difficulties, building upon Malec and Sedransk (1985).

The remainder of this chapter evolves as follows. In Section 3.2, we review a general framework for Bayesian modeling for multi-stage sampling and how simple, two-stage, and stratified random sampling designs arise as special cases. Section 3.3 presents modeling strategies for spatially correlated data sampled using a two-stage design, the implementation of which, along with the model proposed by Scott and Smith (1969), is discussed in Section 3.4 using Bayesian exact and Markov chain Monte Carlo (MCMC) sampling. These models are then applied to simulated data in Section 3.5 and then used in an analysis of nitrate levels in California groundwater in Section 3.6. The chapter concludes with a brief discussion of the results in Section 3.7.

3.2 Bayesian modeling of multi-stage sampling

Suppose that n samples are randomly drawn from a finite population of size N , $n \leq N$ and for the i -th sampled unit, the outcome y_i is measured. Without loss of generality, let the set of outcomes from the finite population be stacked into a vector $y = [y_s^\top : y_{ns}^\top]^\top$, where $y_s = [y_1, \dots, y_n]^\top$ and $y_{ns} = [y_{n+1}, \dots, y_N]^\top$ are vectors of outcome values from the sampled and nonsampled units, respectively. This vector has corresponding design matrix $X = [X_s^\top : X_{ns}^\top]^\top$, which observed for the entire finite population and denotes group membership.

From a superpopulation perspective, we consider the finite population to be a random sample of size N from an infinitely large population. This superpopulation is assumed to follow a Gaussian distribution with mean ν and a covariance function defined by parameters θ . In general, we can construct the following linear regression model:

$$\begin{bmatrix} y_s \\ y_{ns} \end{bmatrix} = \begin{bmatrix} X_s \\ X_{ns} \end{bmatrix} \beta + \begin{bmatrix} \epsilon_s \\ \epsilon_{ns} \end{bmatrix}; \quad \begin{bmatrix} \epsilon_s \\ \epsilon_{ns} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} V_s(\theta) & V_{s,ns}(\theta) \\ V_{ns,s}(\theta) & V_{ns}(\theta) \end{bmatrix}\right). \quad (3.1)$$

Bayesian specifications further model $\beta \sim N(A\nu, V_\beta)$, where A and ν are a vector with length equal to the number of groups and a scalar, respectively, and V_β is the variance of β . The hierarchy continues with probabilistic specifications on ν and θ .

Our goal is to estimate linear finite population quantities of the form $\alpha^\top y$, where α is a given, fixed vector of weights defined for the entire population. Suppose $\nu \sim N(0, \gamma^2)$. Define $V_{\beta|y_s} = [(\gamma^2 AA^\top + V_\beta)^{-1} + X_s^\top V_s^{-1} X_s]^{-1}$ and $Q = X_{ns} - V_{ns,s} V_s^{-1} X_s$. Fixing the variance parameters, the posterior expectation of the finite population quantity is:

$$E[\alpha^\top y | y_s] = \{\alpha_s^\top + \alpha_{ns}^\top [V_{ns,s} + \alpha_{ns}^\top Q V_{\beta|y_s} X_s^\top] V_s^{-1}\} y_s.$$

Defining $B_V = V_{ns,s} + Q V_{\beta|y_s} X_s^\top$, the variance of this expectation is:

$$\text{Var}[E[\alpha^\top y | y_s]] = \alpha_s^\top V_s \alpha_s + 2\alpha_s^\top B_V^\top \alpha_{ns} + \alpha_{ns}^\top B_V V_s^{-1} B_V^\top \alpha_{ns}.$$

Additionally, the posterior variance of $\alpha^\top y$ is:

$$\text{Var}[\alpha^\top y | y_s] = \alpha_{ns}^\top (QV_{\beta|y_s}Q^\top + V_{ns} - V_{ns,s}V_s^{-1}V_{s,ns})\alpha_{ns}.$$

For the special case of a census, in which all members of the population are sampled, e.g. $y_s = y$, the conditional expectation of the finite population quantity is finite population consistent, in the sense that $E[\alpha^\top y | y_s] = E[\alpha^\top y | y] = \alpha^\top y$. Different sampling designs can be incorporated by appropriately structuring the sampled and nonsampled elements. We provide a few examples below. All derivations can be located in the supplementary materials.

Example 1 *Simple Random Sampling*

In simple random sampling, n units are randomly drawn from a population of size N , where each unit in the population is independent and identically distributed with mean μ and variance σ^2 . To express this as in (3.1), define $y_s = [y_1, \dots, y_n]^\top$ and $y_{ns} = [y_{n+1}, \dots, y_N]^\top$, with corresponding design matrices $X_s = 1_n$ and $X_{ns} = 1_{N-n}$, respectively, where 1_n represents the $n \times 1$ vector of ones. Let $A = 1$ and take β to be a scalar μ with mean $\nu = 0$ and variance $V_\beta = \xi^2$. Additionally, let $V_s = \sigma^2 I_n$, $V_{ns} = \sigma^2 I_{N-n}$, and $V_{s,ns} = V_{ns,s}^\top = O$, where O is a matrix of zeroes of appropriate order. Define finite population weights $\alpha = [\alpha_1, \dots, \alpha_N]^\top$. Fixing the variance parameters, the posterior expectation of $\alpha^\top y$ is:

$$E[\alpha^\top y | y_s] = \sum_{i=1}^n \left(\alpha_i + \frac{\sum_{i=n+1}^N \alpha_i}{\frac{\sigma^2}{\xi^2} + n} \right) y_i \quad (3.2)$$

□

Example 2 *Two-Stage Sampling*

In a more complex case, suppose that the population is divided into N distinct groups defined by geography or other characteristics, with the i -th group of size M_i . Assume that within the i -th group, each unit is independent and identically distributed with mean μ_i and variance σ_i^2 . The group means μ_1, \dots, μ_N are independent and follow a normal distribution centered at ν with a variance of δ^2 , hence $\beta = \mu = [\mu_1, \dots, \mu_N]^\top$, $A = 1_N$ and $V_\beta = \delta^2 I_N$. Suppose that only n of the N groups are randomly sampled, where $n \leq N$. Without loss of generality, take the first n

groups to be sampled, and then within the chosen i -th group, m_i units are randomly selected; $m_i \leq M_i$, $i = 1, \dots, n$. As $N - n$ groups are not sampled, the number of observed units in these groups are zero, e.g. $m_i = 0$, $i = n + 1, \dots, N$. Hence, we can define the number of sampled units as $k = \sum_{i=1}^N m_i = \sum_{i=1}^n m_i$, the number of unsampled units as $K = \sum_{i=1}^N (M_i - m_i)$, and the population total to be $T = K + k = \sum_{i=1}^N M_i$.

To examine (3.1) in the context of a two-stage design, define outcome vectors $y_s = [y_1^\top, \dots, y_n^\top]^\top$ and $y_{ns} = [y_1^{*\top}, \dots, y_n^{*\top}]^\top$, with i^{th} components $y_i = [y_{i1}, \dots, y_{im_i}]^\top$ and $y_i^* = [y_{im_i+1}, \dots, y_{iM_i}]^\top$, respectively. The design matrix for the sampled units can be modified by fixing the $k \times N$ matrix $X_s = [\oplus_{i=1}^n \mathbf{1}_{m_i} : O]$, reflecting that $N - n$ of the N sites are unobserved. Similarly, for the unobserved units, define X_{ns} as a block diagonal, $K \times N$ matrix with upper block $[\oplus_{i=1}^n \mathbf{1}_{M_i - m_i}]$ and lower block $[\oplus_{i=n+1}^N \mathbf{1}_{M_i - m_i}]$. For notational convenience, we also define $X_{s1} = [\oplus_{i=1}^n \mathbf{1}_{m_i}]$ and divide the group mean vector μ into sampled, $\mu_s = [\mu_1, \dots, \mu_n]^\top$, and nonsampled, $\mu_{ns} = [\mu_{n+1}, \dots, \mu_N]^\top$, components such that $\mu = [\mu_s^\top, \mu_{ns}^\top]^\top$. Note that distributional mean of y_s , $X_s \mathbf{1}_N \mu$, can be simplified to $X_{s1} \mathbf{1}_n \mu_s$, as the mean of the sampled units does not depend on μ_{ns} . Define the sampled and nonsampled covariance matrices to be $V_s = V_s^{(\sigma)} = [\oplus_{i=1}^n \sigma_i^2 I_{m_i}]$ and $V_{ns} = V_{ns}^{(\sigma)} = [\oplus_{i=1}^N \sigma_i^2 I_{M_i - m_i}]$, respectively, and set $V_{s,ns} = V_{ns,s}^\top = O$. Additionally, define $V^{(\sigma)} = \begin{bmatrix} V_s^{(\sigma)} & O \\ O & V_{ns}^{(\sigma)} \end{bmatrix}$.

To make this model fully Bayesian, let $\nu \sim N(0, \gamma^2)$ and $\delta^2 \sim IG(a, b)$. As our interest lies in estimating $\alpha^T y$, we can derive the posterior distributions of $p(\delta^2 | y_s)$ and $p(\nu | y_s)$ for exact sampling of the superpopulation parameters, the details of which are provided in Section 3.4.2. This approach yields results similar to those derived by Scott and Smith (1969), but has the added strength of including a prior distribution on ν , and Ghosh and Meeden (1997), who replaced distributional assumptions with the assumption of posterior linearity and fixed the variance parameters.

In the two-stage case, for a set of weights $\alpha = [\alpha_{11}, \dots, \alpha_{NM_N}]^\top$, define the group mean of sampled units as $\bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$, $i = 1, \dots, n$, and the group weight of nonsampled units as $\alpha_i = \sum_{j=m_i+1}^{M_i} \alpha_{ij}$, $i = 1, \dots, N$. Also, let $\tilde{\gamma}^2 = \gamma^2 / \delta^2$ and define $\lambda_i = \delta^2 / (\delta^2 + \sigma_i^2 / m_i)$ if $i \in \{1, \dots, n\}$ and $\lambda_i = 0$ if $i \in \{n + 1, \dots, N\}$. Fixing all variance parameters, the expected value of the finite

population estimate is

$$\mathbb{E}[\alpha^\top y | y_s] = \sum_{i=1}^n \sum_{j=1}^{m_i} \left(\alpha_{ij} + \left[\alpha_i + \frac{\sum_{i=1}^N \alpha_i (1 - \lambda_i)}{1/\tilde{\gamma}^2 + \sum_{i=1}^n \lambda_i} \right] \frac{\lambda_i}{m_i} \right) y_{ij} \quad (3.3)$$

Additionally, the two-stage case can be extended to a three-stage case by assuming that the j^{th} element of the i^{th} group has m_{ij}^* subelements. Malec and Sedransk (1985) derive posterior distributions for the means for a three-stage sampling scheme and provide a framework to extend this to data with t stages of sampling. \square

Example 3 Stratified Random Sampling

Stratified sampling is a special case of two-stage sampling where all groups are sampled (e.g. $n = N$ and $m_i > 0$, $i = 1, \dots, N$), and, therefore, considering the same population described in Example 2, the number of sampled units is $k = \sum_{i=1}^N m_i$, the number of nonsampled units is $K = \sum_{i=1}^N (M_i - m_i)$, and the population total is again $T = K + k = \sum_{i=1}^N M_i$. Thus, to express this design as (3.1), let $y_s = [y_1^\top, \dots, y_N^\top]^\top$ and $y_{ns} = [y_1^{*\top}, \dots, y_N^{*\top}]^\top$, with i^{th} components, $y_i = [y_{i1}, \dots, y_{im_i}]^\top$ and $y_i^* = [y_{im_i+1}, \dots, y_{iM_i}]^\top$, respectively. To reflect a membership to one of N groups, we take $X_s = [\oplus_{i=1}^N \mathbf{1}_{m_i}]$, $X_{ns} = [\oplus_{i=1}^N \mathbf{1}_{M_i - m_i}]$, $\beta = \mu = [\mu_1, \dots, \mu_N]^\top$, and $A = \mathbf{1}_N$. The variance components also reflect this and are defined as $V_\beta = \delta^2 I_N$, $V_s = [\oplus_{i=1}^N \sigma_i^2 I_{m_i}]$, $V_{ns} = [\oplus_{i=1}^N \sigma_i^2 I_{M_i - m_i}]$, and $V_{s,ns} = V_{ns,s}^\top = O$.

The posterior expectation of $\alpha^\top y$ is given by (3.3), noting that $n = N$ and $\lambda_i = \frac{\delta^2}{\delta^2 + \sigma_i^2/m_i}$, $i = 1, \dots, N$, is well-defined as $m_i > 0$ for all i . In fact, if non-informative priors are taken for the means, e.g. $\gamma^2 \rightarrow \infty$ and $\delta^2 \rightarrow \infty$, then $\lambda_i \rightarrow 1$, $i = 1, \dots, n$ and the stratified finite population mean is $\mathbb{E} \left[\frac{1}{T} \mathbf{1}_T^\top y | y_s \right] = \sum_{i=1}^N \frac{M_i}{T} \bar{y}_i$, (see, e.g. Little, 2004). \square

3.3 Bayesian spatial process modeling for multi-stage sampling

Extending Example 2 to a geographic context, our spatial domain comprises N regions. Let l_{ij} denote the j -th location in region i . The finite population is described by values $y(l_{ij})$, $i = 1, \dots, N$

and $j = 1, \dots, M_i$. Let y_s be the $k \times 1$ vector corresponding to measurements from the sampled locations and y_{ns} be the $K \times 1$ vector of unsampled measurements. Consider the following spatial regression model for the two-stage finite population,

$$y(\ell_{ij}) = \mu(\ell_{ij}) + \omega(\ell_{ij}) + \epsilon(\ell_{ij}) ; \omega \sim N(0, \Omega) ; \epsilon \sim N(0, V^{(\sigma)}) , \quad (3.4)$$

where $\mu(\ell_{ij})$ is the mean of the outcome at ℓ_{ij} , $\omega = [\omega_s^\top : \omega_{ns}^\top]^\top$ and $\epsilon = [\epsilon_s^\top : \epsilon_{ns}^\top]^\top$ are $T \times 1$ vectors formed by stacking up $\omega(\ell_{ij})$'s and $\epsilon(\ell_{ij})$'s, respectively (analogous to y in Example 2). Here, ω accounts for spatial effects and Ω is the $T \times T$ spatial covariance matrix constructed with $C(d_{ab})$ and is partitioned as $\Omega = \begin{bmatrix} \Omega_s & \Omega_{s,ns} \\ \Omega_{ns,s} & \Omega_{ns} \end{bmatrix}$. Introducing spatial effects in Example 2 yields $\mu(\ell_{ij}) = \mu_i$, $V_s = \Omega_s + V_s^{(\sigma)}$, $V_{ns} = \Omega_{ns} + V_{ns}^{(\sigma)}$, and $V_{s,ns} = V_{ns,s}^\top = \Omega_{s,ns}$ in (3.1). This also accommodates spatial versions of Examples 1 and 3 by setting $N = 1$ and $N = n$, respectively.

Analogous to (3.3), the posterior estimate of a linear function of the population values is

$$\begin{aligned} E[\alpha^\top y | y_s] &= \sum_{i=1}^n \sum_{j=1}^{m_i} \left(\alpha_{ij} + \alpha_{ns}^\top \left\{ \Omega_{ns,s} Q_s + [X_{ns} - \Omega_{ns,s} Q_s^{-1} X_s] \times \left(\frac{1}{\delta^2} I_N + X_s^\top Q_s^{-1} X_s \right)^{-1} \right. \right. \\ &\quad \left. \left. \times \left[X_s^\top Q_s^{-1} + \frac{\frac{1}{\delta^2} 1_N 1_N^\top X_s^\top (\delta^2 X_s X_s^\top + Q_s)^{-1}}{\frac{1}{\gamma^2} + \sum_{i=1}^n \lambda_i^*} \right] \right\} q_{ij} \right) y(\ell_{ij}) , \end{aligned} \quad (3.5)$$

where $\lambda^{*\top} = [\lambda_1^* \dots, \lambda_N^*] = 1_N^\top X_s^\top (\delta^2 X_s X_s^\top + \Omega_s + V_s^{(\sigma)})^{-1} X_s$, $Q_s = \Omega_s + V_s^{(\sigma)}$, and q_{ij} is a set of k indicator vectors of length k , $i = 1, \dots, n$, $j = 1, \dots, m_i$. For $i = 1$, $q_{1j} = j$ and 0 elsewhere, and if $i > 1$, q_{ij} is 1 at element $\sum_{i=1}^{i-1} m_i + j$ and 0 elsewhere. This two-stage spatial model, (3.4), can be written as an intercept-only spatial model by setting $\mu(\ell_{ij}) = \mu$ and $\sigma_i^2 = \sigma^2$, $i = 1, \dots, N$, i.e., simplifying $V^{(\sigma)}$ to $\sigma^2 I_T$. As region is not accounted for, the design matrices X_s and X_{ns} are replaced with 1_k and 1_K , respectively.

However, as the size of the finite population, T , grows, the scalability of (3.4) diminishes due to an increased computational burden stemming from the inversion of the $T \times T$ matrix Ω . To address this, we also consider a more computationally efficient model which also allows for region specific means, but specifies that each region is defined by its own process parameters and is independent

from all other regions. To reflect this regional independence, we specify the covariance function specifying the spatial process $\omega(\ell)$ in (3.4) to be 0 for any two points in different regions, and equal to the value of the Matérn covariance function for any two points within the same region.

Comparing the finite populations estimates given in (3.3) and (3.5), it is evident that accounting for spatial variation results in a more complex equation, as all observed and unobserved outcome values in a population can no longer be assumed to be independent conditional on the group means. This can also be seen in the calculation of the λ parameters, which in the two-stage model, are a simple ratio of variances. In the spatial case, however, the complexity of the parameters is increased by the addition of the spatial covariance matrix.

3.4 Model Implementation and Assessment

3.4.1 General framework

A Bayesian linear model corresponding to the likelihood of the sampled data in (3.1) is

$$p(\theta, \nu, \beta | y_s) \propto p(\theta) \times N(\nu | 0, V_\nu) \times N(\beta | A\nu, V_\beta) \times N(y_s | X_s\beta, V_s(\theta)). \quad (3.6)$$

We use Markov chain Monte Carlo algorithms (see, e.g. Robert and Casella, 2004) for sampling from (3.6). Subsequent Bayesian inference for y_{ns} is available in posterior predictive fashion by drawing samples from

$$p(y_{ns} | y_s) = \int p(y_{ns} | y_s, \theta, \nu, \beta) \times p(\theta, \nu, \beta | y_s) d\theta d\nu d\beta. \quad (3.7)$$

Using the conditional independence of parameters in (3.1), we obtain $p(y_{ns} | y_s, \theta, \nu, \beta) = N(y_{ns} | \mu_{ns|s}, V_{ns|s})$, where $\mu_{ns|s} = X_{ns}\beta + V_{ns,s}(\theta)V_s(\theta)^{-1}(y_s - X_s\beta)$ and $V_{ns|s} = V_{ns}(\theta) - V_{ns,s}(\theta)V_s(\theta)^{-1}V_{s,ns}(\theta)$. Therefore, sampling from (3.7) is achieved by drawing one $\{\beta, \theta\}$ from (3.6) followed by one $y_{ns} \sim N(X_{ns}\beta, V_{ns}(\theta))$, for each posterior sample of $\{\beta, \theta\}$. The resulting samples provide inference on the nonsampled group means μ_{ns} and Bayesian imputation for the nonsampled population units, y_{ns} .

These samples from the posterior predictive distribution can be used to obtain posterior finite population estimates of the form $\alpha^\top y$. We consider four models using (3.6).

Model 1. Two-Stage

For the model provided in Example 2, we take $\theta = [\gamma^2, \delta^2, \sigma_1^2, \dots, \sigma_N^2]^\top$, $p(\theta) = IG(\gamma^2 | a_\gamma, b_\gamma) \times IG(\delta^2 | a_\delta, b_\delta) \times \prod_{i=1}^N IG(\sigma_i^2 | a_{\sigma_i}, b_{\sigma_i})$ in (3.6), and follow the other specifications as in the two-stage setting in Example 2.

As the priors have been chosen to be fully conjugate, one can derive the full posterior conditional distributions for each of the parameters. Specifically, the variance parameters will have posterior distributions of the form $IG(a^*, b^*)$, while the rest of the parameters will have posterior distributions of the form $N(Mm, M)$. However, as only n of the N groups are observed, the variance terms of the unsampled groups, $\sigma_{n+1}^2, \dots, \sigma_N^2$, must either be fixed or given informative priors. If not, draws from the posterior predictive distribution corresponding to units in the nonsampled groups will have arbitrary variability and could spuriously dominate the finite population estimates.

Model 2. Spatial

Under (3.6), the intercept-only spatial model defines $\theta = [\phi, \delta^2, \sigma^2, \tau^2]^\top$ with corresponding prior distribution $p(\theta) = p(\phi) \times IG(\delta^2 | a_\delta, b_\delta) \times IG(\sigma^2 | a_\sigma, b_\sigma) \times IG(\tau^2 | a_\tau, b_\tau)$ and $V_\beta = \delta^2$, $V_s = \Omega_s + \sigma^2 I_k$. As there are no group terms, replace X_s with 1_k and take $\nu = 0$ with probability 1, e.g. $V_\nu^{-1} = 0$.

Unlike model 1, regardless of the prior distribution placed on $p(\phi)$, a closed-form posterior distribution cannot be found for ϕ . In practice, ϕ is often fixed using an estimate found from a variogram and then full posterior conditional distributions can be found using the same techniques described for the non-spatial case. However, MCMC can still be implemented by specifying a prior distribution for ϕ (Banerjee et al. 2014), which is often taken to be a uniform distribution.

To recover the spatial effects ω absorbed into the variance parameter of y , note that $\beta = \mu_s$ and $p(\omega | y, \theta, \mu_s) \propto N(\omega | 0, \Omega) \times N(y | 1_T \mu_s + \omega, \sigma^2 I_T) \propto N(M_\omega m_\omega, M_\omega)$, where $m_\omega = \frac{1}{\sigma^2}(y - 1_T \mu_s)$ and $M_\omega = (\Omega^{-1} + \frac{1}{\sigma^2} I_T)^{-1}$. Thus, drawing one $\omega \sim N(M_\omega m_\omega, M_\omega)$ for each posterior sample of $\{\theta, \mu_s, y_{ns}\}$ will result in a set of posterior samples of ω .

Model 3. Two-Stage + Spatial

The spatial model in (3.4) can be rewritten using (3.6) by letting $\theta = [\phi, \gamma^2, \delta^2, \tau^2, \sigma_1^2, \dots, \sigma_N^2]^\top$,

with $p(\theta) = p(\phi) \times IG(\gamma^2 | a_\gamma, b_\gamma) \times IG(\delta^2 | a_\delta, b_\delta) \times IG(\tau^2 | a_\tau, b_\tau) \times \prod_{i=1}^N IG(\sigma_i^2 | a_{\sigma_i}, b_{\sigma_i})$, $V_\nu = \gamma^2$, $V_\beta = \delta^2 I_N$, $A = 1_N$, $V_s = \Omega_s + V_s^{(\sigma)}$, and $\beta = \mu$. After posterior samples of $\{\theta, \mu, y_{ns}\}$ are drawn as described in (3.7), posterior samples of the spatial effects can be recovered by sampling one $\omega \sim N(M_{\omega 2} m_{\omega 2}, M_{\omega 2})$ for each posterior sample of $\{\theta, \mu, y_{ns}\}$, where $m_{\omega 2} = V^{(\sigma)-1}(y - X\mu)$ and $M_{\omega 2} = (\Omega^{-1} + V^{(\sigma)-1})^{-1}$.

Model 4. Regional Spatial

To rewrite the region-specific spatial model given using (3.6), let $V_\nu = \gamma^2$, $V_\beta = \delta^2 I_N$, $A = 1_N$, $V_s = \Omega_{s*} + V_s^{(\sigma)}$, and $\beta = \mu$. Also take $\theta = [\phi_1, \dots, \phi_N, \gamma^2, \delta^2, \tau_1^2, \dots, \tau_N^2, \sigma_1^2, \dots, \sigma_N^2]^\top$ with $p(\theta) = \prod_{i=1}^N p(\phi_i) \times IG(\gamma^2 | a_\gamma, b_\gamma) \times IG(\delta^2 | a_\delta, b_\delta) \times \prod_{i=1}^N IG(\tau_i^2 | a_{\tau_i}, b_{\tau_i}) \times \prod_{i=1}^N IG(\sigma_i^2 | a_{\sigma_i}, b_{\sigma_i})$. Similar to model 1, as not all locations are sampled, informative priors must be placed on the $\phi_{n+1}, \dots, \phi_N$ spatial decay parameters. Additionally, to recover posterior samples of ω , sample one $\omega \sim N(M_{\omega 3} m_{\omega 3}, M_{\omega 3})$ for each posterior sample of $\{\theta, \mu, y_{ns}\}$ drawn using (3.7), where $m_{\omega 3} = V^{(\sigma)-1}(y - X\mu)$ and $M_{\omega 3} = (\Omega_*^{-1} + V^{(\sigma)-1})^{-1}$.

To achieve computation efficiency, redefine $y = [y_1^\top, y_1^{*\top}, \dots, y_n^\top, y_n^{*\top}, y_{n+1}^{*\top}, \dots, y_N^{*\top}]^\top$ so that the outcome is organized by region and then Ω_* becomes a $T \times T$ block diagonal matrix composed of N blocks. This allows us to instead invert N covariance matrices of size $M_1 \times M_1, \dots, M_N \times M_N$, rather than one $T \times T$ matrix, in the estimation of ω .

3.4.2 Exact Monte Carlo Estimation

If we are able to provide reasonable fixed values of the parameters, (3.1) can be simplified into a conjugate Bayesian linear model resembling:

$$IG(\delta^2 | a, b) \times N(\nu | 0, \delta^2 \tilde{V}_\nu) \times N(\beta | A\nu, \delta^2 \tilde{V}_\beta) \times N(y_s | X\beta, \delta^2 \tilde{V}_s). \quad (3.8)$$

For a model such as this, the components a , b , \tilde{V}_ν , \tilde{V}_β , and \tilde{V}_s are fixed, reducing the model to three unknown parameters, δ^2 , ν , and β . Thus, we can avoid MCMC and sample from the joint posterior $p(\delta^2, \nu, \beta | y_s)$ using the following steps. First sample δ^2 from $p(\delta^2 | y - S) = IG(a^*, b^*)$ and then for each δ^2 drawn, draw a corresponding ν from $N(M_\nu m_\nu, \delta^2 M_\nu)$. Next, for each pair of $\{\delta^2, \nu\}$, draw β from $N(M_\beta m_\beta, \delta^2 M_\beta)$ (see the Supporting Information for details). As an example, we recast

each model presented in Section 3.4.1 in the form of (3.8) and derive the posterior conditional distributions for model 1 and model 2, details of which are provided in the Appendix at the end of this chapter.

Model 1. Two-Stage

To create a conjugate Bayesian model such as (3.8) from the non-spatial model, define $A = 1_N$, $\tilde{V}_\nu = \tilde{\gamma}^2 = \frac{\gamma^2}{\delta^2}$, $\tilde{V}_\beta = I_N$, and $\tilde{V}_s = \tilde{V}_s^{(\sigma)} = [\oplus_{i=1}^n \frac{\sigma_i^2}{\delta^2} I_{m_i}]$. Noting that $p(\nu | y_s) \propto N(\nu | 0, \delta^2 \tilde{\gamma}^2) \times N(y_s | X_{s1} 1_n \nu, \delta^2 [X_{s1} X_{s1}^\top + \tilde{V}_s^{(\sigma)}])$ a little algebra reveals

$$\nu | y_s, \delta^2 \sim N(\nu | c, \delta^2 d), \quad (3.9)$$

where $c = \frac{\sum_{i=1}^n \lambda_i \bar{y}_i}{\frac{1}{\tilde{\gamma}^2} + \sum_{i=1}^n \lambda_i}$ and $d = \left[\frac{1}{\tilde{\gamma}^2} + \sum_{i=1}^n \lambda_i \right]^{-1}$. The mean of the posterior distribution, c , is the weighted average of the sampled group means, where each mean is weighted by a function of each group's element-wise variance. Integrating out ν and μ from $p(\delta^2, \nu, \mu | y_s)$ yields $p(\delta^2 | y_s)$, which is:

$$\delta^2 | y_s \sim IG \left(a + \frac{k}{2}, b + \frac{1}{2} \left[y_s^\top (X_{s1} X_{s1}^\top + \tilde{V}_s^{(\sigma)})^{-1} y_s + \frac{c^2}{d} \right] \right). \quad (3.10)$$

Taking the limits of c and d as $\tilde{\gamma}^2 \rightarrow \infty$ (e.g. $\gamma^2 \rightarrow \infty$) we recover the findings of Scott and Smith (1969), who assigned $p(\nu) \propto 1$:

$$\lim_{\tilde{\gamma}^2 \rightarrow \infty} c = \frac{\sum_{i=1}^n \lambda_i \bar{y}_i}{\sum_{i=1}^n \lambda_i} \quad \text{and} \quad \lim_{\tilde{\gamma}^2 \rightarrow \infty} \delta^2 d = \frac{\delta^2}{\sum_{i=1}^n \lambda_i}.$$

As $p(\mu_s | y_s, \delta^2, \nu) \propto N(\mu_s | \nu 1_n, \delta^2 I_n) \times N(y_s | X_{s1} \mu_s, \delta^2 \tilde{V}_s^{(\sigma)})$ we have that:

$$\mu_s | y_s, \nu, \delta^2 \sim N(\mu_s | c_*, \delta^2 d_*) , \quad (3.11)$$

where $c_* = \begin{bmatrix} (1 - \lambda_1)\nu + \lambda_1 \bar{y}_1 \\ \vdots \\ (1 - \lambda_n)\nu + \lambda_n \bar{y}_n \end{bmatrix}$ and $d_* = \left[\oplus_{i=1}^n (1 - \lambda_i) \right]$. The posterior mean is appealing for

interpretation, as its i -th element is the weighted average of the i -th group's sample mean and the superpopulation mean estimate. Finally, note that $\mu_{ns} | y_s, \nu, \delta^2 \sim N(\mu_{ns} | \nu, \delta^2 I_{N-n})$, as y_s provides no information pertaining to the nonsampled groups.

Model 2. Spatial

The spatial model can be recast as (3.8) by defining $\tilde{V}_\nu^{-1} = 0$ and $\tilde{V}_s = \tilde{\Omega}_s = \frac{1}{\delta^2}\Omega_s + I_k$, where \tilde{V}_β is fixed to 1. Defining $V_{\tilde{\Omega}_s} = (1 + \mathbf{1}_k^\top \tilde{\Omega}_s^{-1} \mathbf{1}_k)^{-1}$, the posterior conditionals are:

$$\delta^2 | y_s \sim IG \left[a + \frac{k}{2}, b + \frac{1}{2} y_s^\top \left(\tilde{\Omega}_s^{-1} - \tilde{\Omega}_s^{-1} \mathbf{1}_k V_{\tilde{\Omega}_s} \mathbf{1}_k^\top \tilde{\Omega}_s^{-1} \right) y_s \right] \text{ and} \quad (3.12)$$

$$\mu_s | y_s, \delta^2 \sim N \left[V_{\tilde{\Omega}_s} \mathbf{1}_k^\top \tilde{\Omega}_s^{-1} y_s, \delta^2 V_{\tilde{\Omega}_s} \right]. \quad (3.13)$$

Model 3. Two-Stage + Spatial

The form of (3.8) is achieved by defining $\tilde{V}_\nu = \tilde{\gamma}^2$, $\tilde{V}_\beta = I_N$, $A = \mathbf{1}_N$, and $\tilde{V}_s = \frac{1}{\delta^2}\Omega_s + \tilde{V}_s^{(\sigma)}$.

Model 4. Regional Spatial

The form of (3.8) is achieved by defining $\tilde{V}_\nu = \tilde{\gamma}^2$, $\tilde{V}_\beta = I_N$, $A = \mathbf{1}_N$, and $\tilde{V}_s = \frac{1}{\delta^2}\Omega_{s^*} + \tilde{V}_s^{(\sigma)}$.

3.4.3 Model Comparison and Assessment

Model fit was evaluated in two ways. In general, consider a sample of size k drawn from a population of size T with outcome $y = [y_s^\top : y_{ns}^\top]^\top$. Without loss of generality, say $y_h \in y_s$ if $h = 1, \dots, k$ and $y_h \in y_{ns}$ if $h = k + 1, \dots, T$. First we evaluate the predictive accuracy of the models using the Watanabe-Akaike Information Criteria (WAIC), which is expressed as $WAIC = -2\widehat{elpd} = -2(\widehat{lpd} + \hat{p}_{WAIC})$ in Vehtari et al. (2017), where \widehat{elpd} is the estimated expected log pointwise predictive density and is multiplied by -2 to be on the deviance scale. To calculate this, at each iteration, $l = 1, \dots, L$, $p(y_h | \Theta^{(l)})$ is computed; the likelihood of each observed value conditional on that iteration's parameters. The estimated log pointwise predictive density is the sum of the log average likelihood for each observation, $\widehat{lpd} = \sum_{h=1}^k \log \left[\frac{1}{L} \sum_{l=1}^L p(y_h | \Theta^{(l)}) \right]$. The sample variance of the log-likelihood for each observation is $s_{lp(y_h)}^2 = \frac{1}{L-1} \sum_{l=1}^L \left[\log(p(y_h | \Theta^{(l)})) - \frac{1}{L} \sum_{l=1}^L \log(p(y_h | \Theta^{(l)})) \right]^2$ and the estimated effective number of parameters is the sum of these variances: $\hat{p}_{WAIC} = \sum_{h=1}^k s_{lp(y_h)}^2$. To calculate the standard error of the WAIC, rewrite $-2(\widehat{elpd} + \hat{p}_{WAIC}) = -2 \sum_{h=1}^k \widehat{elpd}_h = \sum_{h=1}^k \left\{ \log \left[\frac{1}{L} \sum_{l=1}^L p(y_h | \Theta^{(l)}) \right] + s_{lp(y_h)}^2 \right\}$. Under the assumption that each \widehat{elpd}_h is independent, the sample variance of each individual \widehat{elpd}_h is $s_{elpd,ind}^2 = \frac{1}{N-1} \sum_{h=1}^k \left[\widehat{elpd}_h - \frac{1}{k} \sum_{h=1}^k \widehat{elpd}_h \right]^2$. Then $SE(WAIC) = \sqrt{Var(-2 \sum_{h=1}^k \widehat{elpd}_h)} = 2\sqrt{n Var(\widehat{elpd}_h)} = 2s_{elpd,ind}\sqrt{n}$.

Second, for simulated data the true values $y = [y_s : y_{ns}]^\top$ are known and so we compare these with replicated datasets, $y_{rep}^{(l)} = [y_{rep,1}^{(l)} \cdots y_{rep,k}^{(l)}]^\top$, generated from the pointwise posterior predictive distribution at each iteration l . These are used to formulate the goodness of fit measurement $D = G + P$ described in Gelfand and Ghosh (1998), composed of an error sum of squares term and a penalty term for large predictive variances. For L iterations, $G = \sum_{h=1}^k (y_h - E[y_{rep,h} | y_s])^2$ and $P = \sum_{h=1}^k var(y_{rep,h} | y_s)$. We approximate $E[y_{rep,h} | y_s] \approx \frac{1}{L} \sum_{l=1}^L y_{rep,h}^{(l)}$ and $var(y_{rep,h} | y_s) \approx \frac{1}{L-1} \sum_{l=1}^L (y_{rep,h}^{(l)} - \frac{1}{L} \sum_{l=1}^L y_{rep,h}^{(l)})^2$. For non-simulated datasets, where y_{ns} is unknown, D can still be calculated by restricting the replicate datasets to the observed units, y_s , e.g. $y_{rep}^{(l)} = [y_{rep,1}^{(l)} \cdots y_{rep,k}^{(l)}]^\top$, at the l -th iteration.

3.5 Simulation

3.5.1 Data Generation

To simulate spatial correlation and allow for two-stage random sampling, a unit square was divided into 100 equally sized square regions and 2,500 locations were randomly drawn from the unit square. Data was simulated from the intercept-only spatial model described in Model 2 with $\mu = 2$. A distance matrix for all points was constructed and used to create a covariance matrix that reflects an exponential covariance function described in Section 3.3, where ϕ was assigned a value of 10, reflecting an effective spatial range of 3/10. The spatial variance, τ^2 , was fixed at 9, while the non-spatial variance, σ^2 , was set to 4. After a dataset was generated, a cluster random sampling scheme was implemented. 25 regions were randomly selected and then in each cluster, a random number of individuals were selected (the minimum and maximum percent of those selected from a region was set to be 20% and 90%, respectively). 20 datasets containing information of both the sampled and nonsampled units were generated in this way. To examine Models 1 and 4 for larger dataset, this process was then repeated with the same parameters to generate 20 datasets with 8,100 locations from 324 regions, where 81 regions were randomly sampled. All data generation and analyses were performed using R version 3.5.1 (R Core Team, 2018).

3.5.2 Exact Monte Carlo Simulation

To perform the two-stage procedure using the conditional distributions and methods described in Section 3.4.2, sample means, $\hat{\mu}_i$, and sample variances, $\hat{\sigma}_i^2$, were calculated from each observed region, $i = 1, \dots, 25$. The variance matrix of the sampled units was fixed to be $\tilde{V}_s^{(\sigma)} = \left[\oplus_{i=1}^n \frac{\hat{\sigma}_i^2}{\text{Var}(\hat{\mu})} I_{m_i} \right]$ where $\text{Var}(\hat{\mu})$ represents the sample variance of the observed sample means. Similar to fixing $V_s^{(\sigma)}$, the variance matrix of the nonsampled units was fixed at $V_{ns} = \left[\oplus_{i=1}^N \frac{\tilde{\sigma}_i^2}{\text{Var}(\hat{\mu})} I_{M_i - m_i} \right]$, where $\tilde{\sigma}_i^2 = \hat{\sigma}_i^2$ if $i \in \{1, \dots, n\}$ and $\tilde{\sigma}_i^2 = \frac{1}{n} \sum_{i=1}^n \hat{\sigma}_i^2$ if $i \in \{n+1, \dots, N\}$. The value of γ^2 was fixed to be half of the value of δ^2 , reflecting the belief that there was less variability in the population mean than between group means. The prior distribution for δ^2 was assigned to be $IG(3, 5)$. Sampling from the posterior was performed using the conditional distributions and methods described in Section 3.4.2. At each iteration g , the population mean estimate for that iteration was then calculated as $\bar{y}^{(g)} = \frac{1}{T} \left(\sum_{i=1}^n \sum_{j=1}^{m_i} y_{ij} + \sum_{i=1}^N \sum_{j=m_i+1}^{M_i} y_{ij}^{(g)} \right)$.

As we have fixed the ratios of all variance components, we have also fixed $\tilde{\lambda}_i = \text{Var}(\hat{\mu}) / (\text{Var}(\hat{\mu}) + \hat{\sigma}_i^2 / m_i)$ if $i \in (1, \dots, n)$ and $\tilde{\lambda}_i = 0$ if $i \in (n+1, \dots, N)$. Define $c = \frac{\sum_{i=1}^n \tilde{\lambda}_i \bar{y}_i}{\frac{1}{2} + \sum_{i=1}^n \tilde{\lambda}_i}$, $d = \left[\frac{1}{2} + \sum_{i=1}^n \tilde{\lambda}_i \right]^{-1}$, $c^{*(g)} = \begin{bmatrix} (1 - \tilde{\lambda}_1) \nu^{(g)} + \tilde{\lambda}_1 \bar{y}_1 \\ \vdots \\ (1 - \tilde{\lambda}_n) \nu^{(g)} + \tilde{\lambda}_n \bar{y}_n \end{bmatrix}$, and $d^* = \left[\oplus_{i=1}^n (1 - \tilde{\lambda}_i) \right]$. The following procedure was

implemented to produce posterior estimates of the population mean, $\bar{y}^{(g)}$, for G iterations.

for(g in 1:G){

$$\delta^{2(g)} \sim IG\left(3 + \frac{1}{2} \sum_{i=1}^n m_i, 5 + \frac{1}{2} \left[y_s^\top (\tilde{V}_s^{(\sigma)} + X_s X_s^\top)^{-1} y_s + \frac{c^2}{d} \right]\right)$$

$$\nu^{(g)} \sim N(c, \delta^{2(g)} d)$$

$$\mu_s^{(g)} \sim N(c^{*(g)}, \delta^{2(g)} d^*)$$

$$\mu_{ns}^{(g)} \sim N(\nu^{(g)} \mathbf{1}_n, \delta^{2(g)} I_{N-n})$$

$$y_{ns}^{(g)} \sim N(X_{ns} \mu^{(g)}, \delta^{2(g)} \tilde{V}_{ns})$$

$$\bar{y}^{(g)} = \frac{1}{T} \left(\sum_{i=1}^n \sum_{j=1}^{m_i} y_{ij} + \sum_{i=1}^N \sum_{j=m_i+1}^{M_i} y_{ij}^{(g)} \right)$$

}

To perform the spatial random effect procedure, ϕ was set to its true value of 10 and the ratio of δ^2/τ^2 to its true value of 4/9. The posterior conditionals of $\delta^2 | y_s$ and $\mu_s | y_s, \delta^2$, (3.12) and (3.13) respectively, were sampled as outlined in Section 3.3. This sampling and the prediction of y_{ns} was performed using commands from the *spBayes* R package (Finley et al., 2015, 2007). The population mean estimate was calculated using the technique described in the non-spatial sampling case above.

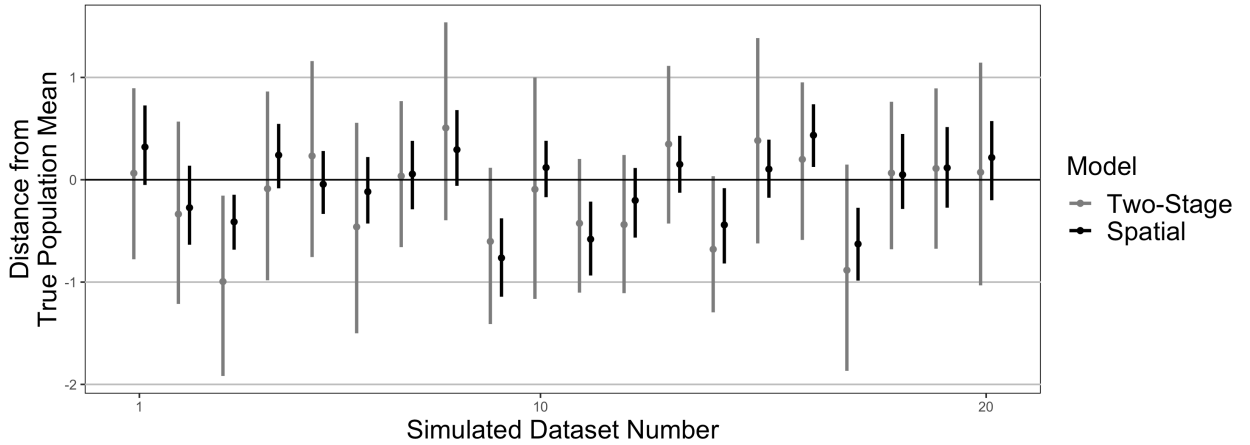


Figure 3.1: Centered Population Mean Estimates from 2 Exact Models with 95% CI

Figure 3.1 plots population average-centered mean estimates and 95% credible intervals from

both methods applied to the twenty simulated datasets. While the spatial cases consistently have a smaller credible interval, their point estimates are similar to the two-stage case. However, as the ratio of spatial and non-spatial variance is fixed for this method, this may result in a reduction in the overall variance of the population mean. Posterior mean estimates and their associated 95% credible intervals of the superpopulation parameters and finite population mean, \bar{y} , from the first generated dataset are given in Table 3.1, along with the WAIC, its standard error, and D values. While both models have similar estimates for ν , the two-stage model overestimates the non-spatial variance. This is expected, as we know that there is additional variance due to spatial correlation that is not being accounted for otherwise in the model. Similarly, both measures of goodness of fit prefer the spatial model.

Table 3.1: Comparison of Parameter Estimation and Model Fit in Two Exact Models

	Two-Stage	Spatial
ν (2)	2.60 (1.55, 3.58)	2.79 (1.36, 4.23)
δ^2 (4)	6.36 (5.47, 7.45)	3.84 (3.35, 4.43)
τ^2 (9)	—	8.65 (7.53, 9.96)
\bar{y} (2.60)	2.66 (7.53, 9.96)	2.92 (7.53, 9.96)
WAIC	1803.83 (25.52)	1686.7 (27.6)
$D = G + P$	66100.83=34203.26+31897.56	45335.54=22875.45+22460.08

3.5.3 Markov Chain Monte Carlo Simulation

To explore these findings further, we implemented the four models described in Section 3.4 using the JAGS software in R on the same generated datasets. Models were run for 650 iterations with 50 burn-in, as examination of individual trace plots suggested sufficient mixing and convergence of the non-spatial parameters. At each iteration g , estimates of the nonsampled units were drawn and estimates for the population mean, $\bar{y}^{(g)} = \frac{1}{T} \left(\sum_{i=1}^n \sum_{j=1}^{m_i} y_{ij} + \sum_{i=1}^N \sum_{j=m_i+1}^{M_i} y_{ij}^{(g)} \right)$ were calculated. All variance parameters (σ^2 , τ^2 , and δ^2 , as well as site-specific variances such as σ_i^2 and τ_i^2) were given an inverse-gamma prior with shape 2 and scale 10, reflecting a weakly-informative prior distribution with mean 10. There is a substantial literature in theoretical spatial statistics regarding the identifiability, or lack there of, of the spatial process parameters, hence, non-informative or

completely flat priors are excluded from consideration. The prior families and specifics we use are fairly customary in spatial modeling. They exploit some information about the spatial domain and the extent of spatial association that can be detected from finite samples using variograms. For example, in practice, given a real data set, we would pass the data through an exploratory analysis tool, such as a variogram, glean some information about the spatial variance component and the measurement error component, and use the weakly informative centered inverse-gamma priors to reflect these values. In addition, ν was given a flat prior to not inform the estimation of the mean and all ϕ parameters were given Uniform(5,15) priors to allow the spatial range to vary from 0.2 (3/15) to 0.6 (3/5). While our priors were chosen to be weakly-informative to be conservative in estimation, more informative priors could easily be added in a data analysis if additional information regarding the parameters was known. MCMC sampling was performed using the computer program JAGS (Plummer, 2017) in R.

When assessing model fit in the first realization of the data with WAIC, the spatial model performed slightly worse than the rest of the models with a value of 1,912.70 (SE = 26.36). This may be due to the additional variation which comes from varying the spatial range parameter. This was followed closely by the two-stage model with 1,870.26 (35.00) , which was outperformed by both the regional spatial model with 1,202.08 (38.99) and the two-stage + spatial model with 455.67 (17.03). It is interesting that while the data was generated by the spatial model and sampled by a two-stage framework, neither of these models perform better than the two models which take both the spatial correlation and study design into account.

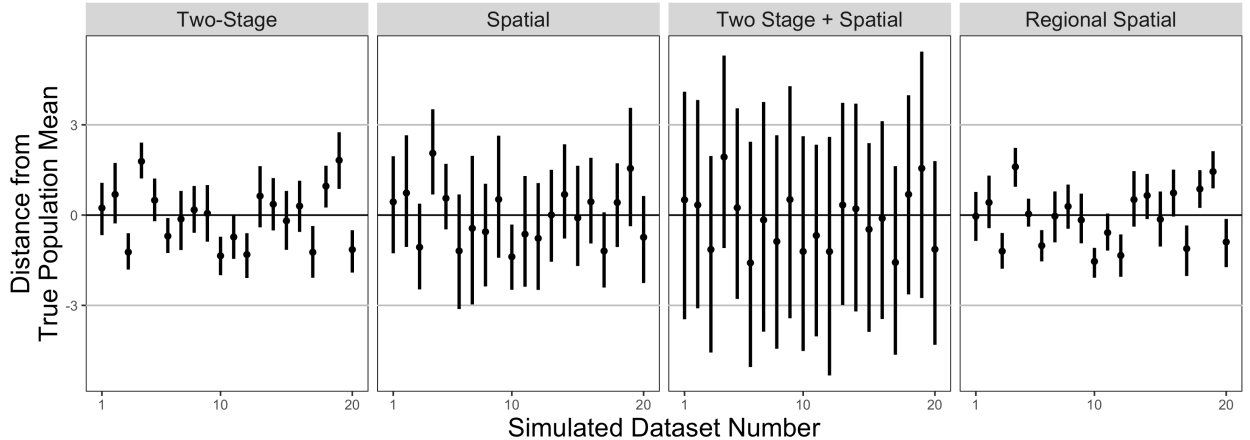


Figure 3.2: Centered Population Mean Estimates from 4 MCMC Models with 95% CI

Figure 3.2 shows the models' posterior mean estimates of the finite population mean, which are centered at the true population mean and presented with 95% credible intervals for the 20 simulated datasets. While point estimates remain similar across models, the best fitting model, Model 3, has the widest credible intervals for the population mean. Accounting for only regional effects results in tight credible intervals in Models 1 and 4, which are narrow compared to Model 2, which fails to take into account region specific variability. Table 3.2 compares the finite population mean estimate, ν estimate, and model fit for the first simulated dataset. Recall that $\nu = 2$ and the finite population mean was 2.60. Notice that while the true values are included in the credible intervals for all models, the credible intervals for ν are wider than those of the FP mean.

Table 3.2: Comparison of Estimates and Model Fit from MCMC

Model	FP Mean (95% CI)	ν (95% CI)	WAIC (SE)
1. Two-Stage	2.84 (1.93, 3.67)	2.84 (1.82, 3.82)	1870.26 (35.00)
2. Spatial	3.04 (1.33, 4.56)	3.24 (1.20, 5.01)	1912.70 (26.36)
3. Two-Stage + Spatial	3.11 (-0.86, 6.70)	3.22 (-1.23, 7.33)	455.67 (17.03)
4. Regional Spatial	2.56 (1.74, 3.37)	2.44 (1.49, 3.47)	1202.08 (38.99)

Similar results were found when applying Models 1 and 4 to the larger simulated datasets. Figure 3.3 recreates the centered mean plots presented in Figure 3.2 for the larger data case, in which the number of regions is 324. As in the $N = 100$ case, the point estimates and 95% credible intervals are similar for the two models. Additionally, the regional spatial model still outperforms the two-stage model with a WAIC of 1,052 ($SE = 38.89$) compared to 1,869 ($SE = 34.77$).

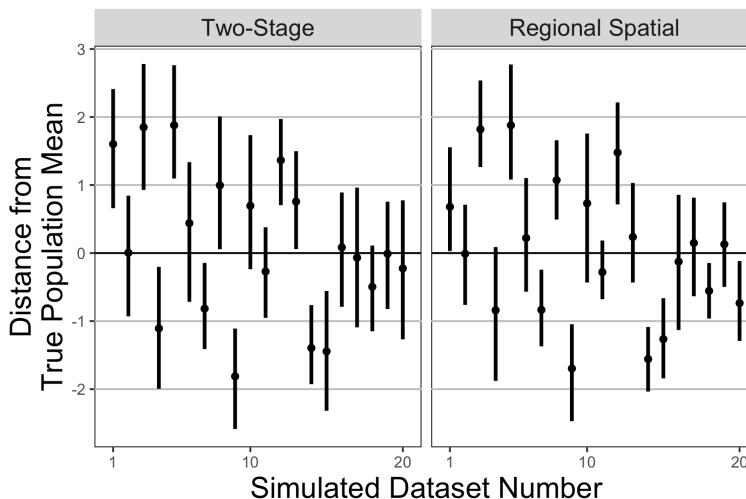


Figure 3.3: Centered Population Mean Estimates from 2 MCMC Models with 95% CI.

3.6 Data Analysis: Nitrate in Central California Groundwater

In this section, we provide an analysis of groundwater nitrate content of the Tulare Lake Basin (TLB) in Central California from the California Ambient Spatio-Temporal Information on Nitrate in Groundwater (CASTING) Database, which is described in Harter et al. (2017) and Boyle et al. (2012) and is available as the University of California, Davis, nitrate data in the data repository of the Groundwater Ambient Monitoring and Assessment Program (2019). Interest lies in identifying regions in which ground water nitrate levels exceed 45 mg/L, which is the maximum contaminant level established by the EPA Boyle et al. (2012). At high levels, infants and pregnant women are more susceptible to nitrate poisoning, which makes it more difficult for oxygen to be distributed to body and can be fatal to infants less than six months old. Besides human sources such as sewage disposal, many sources of nitrate are agricultural, such as fertilizer for crops and animal waste (Harter and Lund, 2012). Because of this, regions with high agricultural activity, such as the TLB, have experienced rising levels of nitrate over the past few decades. As groundwater, and therefore nitrate levels in groundwater, can be assumed to be present at all areas of the Central Valley, we can assume that water samples taken from wells come from a spatial field. Therefore, given a sample of readings from various wells, our primary goal is to estimate of the finite population average of all known wells, which represents an overall measure of water-health. Additionally, plots of posterior predictive distribution may be useful in identifying high-risk regions which exceed the maximum contaminant level.

The CASTING Database is an extensive collection of nitrate readings from the TLB and Salinas Valley collected by multiple agencies, over 70% of which were collected between 2000 and 2011. Of these, most wells had repeated measurements taken over the time. As the Salinas Valley and the TLB are geographically separate regions of California, only the TLB was included. In order to avoid associations over time, the data was restricted to a single year. The year 2009 was selected as variogram plots suggested nitrate levels followed a roughly exponential distribution. While directional variograms suggested that the measurements may be anisotropic, we continued with the

methods presented above, recognizing that a model accounting for directional spatial dependence may provide a better fit to the data.

While the data roughly covers the TLB, we construct a likely sampling scenario in which the TLB is separated into distinct geographic regions and due constraints (perhaps time or financial), a random subset of these regions are sampled. In our scenario, zip codes were used as it is common to collect such geographic information in large scale health surveys, but many alternatives such as cities or grid-based approach could have also been used. A map of California zip codes tabulation areas obtained from the *tigris* R package (Walker, 2018) was overlaid on the approximate geographic locations of each of the sampled wells, effectively assigning each well to one specific region, defined by a zip code. 1) Only the most recent observation was taken from each well so that each well was only represented once. 2) If unique wells had the same geographic coordinates, one was chosen at random to be removed. 3) Sparse zip codes with less than 10 wells were excluded to ensure that each selected zip code would have a large sample size and to avoid overfitting when modeling. These restrictions resulted in a dataset with 6,117 unique wells among 63 zip codes. Nitrate level had a mean of 37.9 mg/L, standard deviation of 52.3 mg/L, and ranged from 0.0 to 903.1 mg/L.

In order to recreate a cluster sampling scenario, 21 of the zip codes were randomly chosen and 50-90% of the population in that zip code was randomly sampled. This resulted in an observed sample size of 489 with a mean nitrate level of 34.2 mg/L and a standard deviation of 40.0 mg/L. The nitrate level ranged from 0.0 to 269.6. Figure 3.4 presents a map of these and surrounding zip codes, denoting them as either sampled or non-sampled, while all other zip codes are denoted as excluded.

The variograms provided in Figure 3.5 were fitted to the entire population of wells (left) and sampled wells (right). For the population variogram, the estimated values of the nugget, partial sill, and range were 2527.4, 319.9, and 2.29, respectively. For the sample variogram, the estimated values of the nugget, partial sill, and range were 1808.6, 1014.4, and 335.9, respectively.

Using this sampled data, all four models in Section 3.5.3 were implemented and the results are shown in Table 3.3. In all models, ν was given a flat prior. For model 1, the regional variance parameters were given an inverse-gamma prior with shape 2 and scale 1600 to reflect the sample variance. For the spatial models, a variogram was fit and spatial variances were given an inverse-

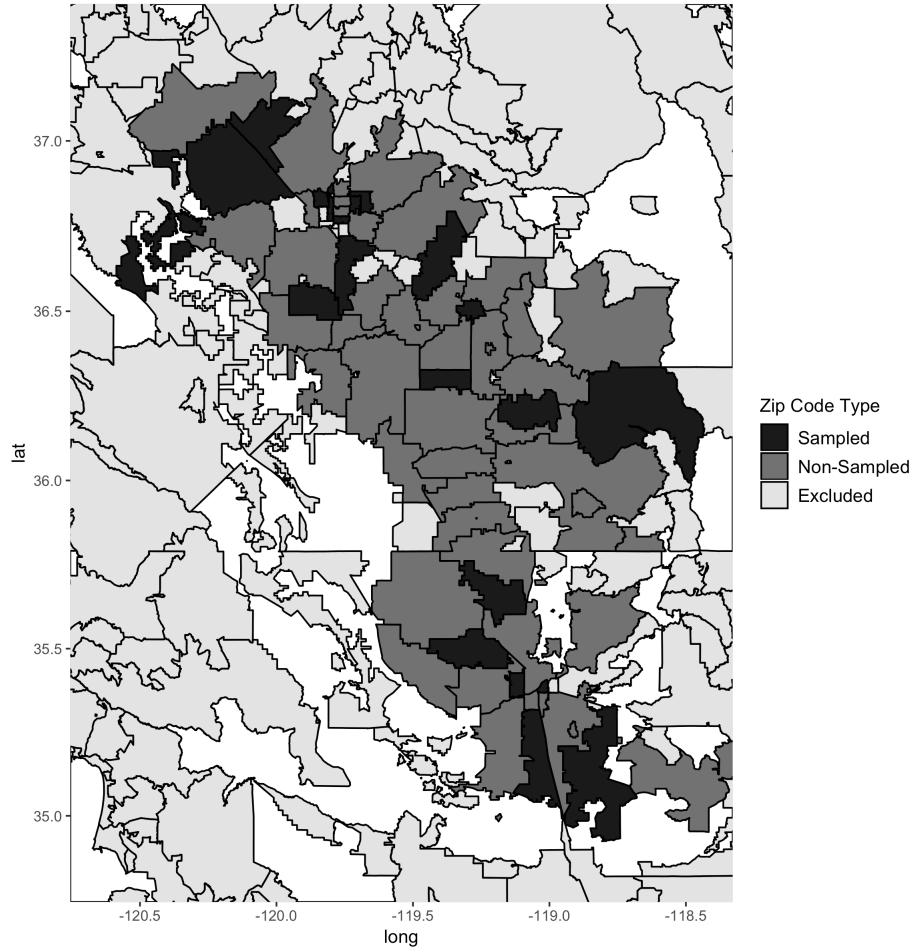


Figure 3.4: Plot of California zip code tabulation areas.

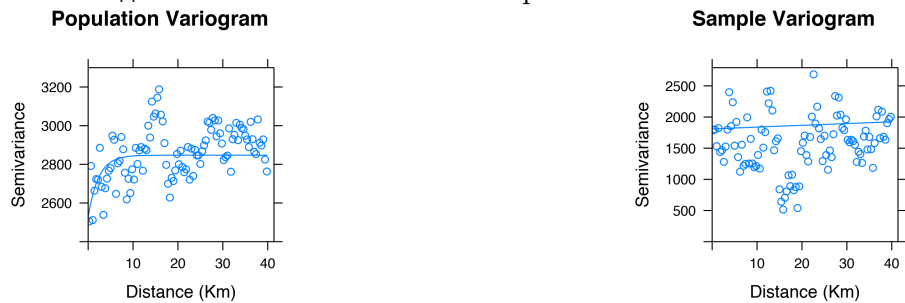


Figure 3.5: Variograms from Population and Sampled Data.

gamma prior with shape 2 and scale 1800. Similarly, non-spatial variance terms were assigned an inverse-gamma prior with shape 2 and scale 1000. The variance of regional means was assigned an inverse-gamma prior with shape 2 and scale 10 to allow for small, localized deviations. All ϕ parameters were given $\text{Uniform}(0.01,5)$ priors to reflect a spatial range varying between 0.6 km (3/5) and 300 km (3/0.01). MCMC sampling was performed using JAGS (Plummer, 2017) in R

(R Core Team, 2018).

With respect to the estimate of the true mean nitrate level, only the intercept-only spatial model contained the true mean value within its 95% credible intervals. However, as evidenced by the larger mean, standard deviation, and range in the complete dataset, it appears that the sampled units did not capture some of the larger outliers, so it is unsurprising that the estimates of the population mean are lower than the truth. Comparing WAIC, we see results similar to those found in Section 3.5.3. The spatial models which accounted for regional means had lower WAIC values than the two-stage model, which is evidence that this data is spatially correlated. However, the intercept-only spatial model did not fit the data as well as the two-stage model, which may be due to ignoring the study design. Additionally, the two-stage + spatial model again fits the model the best.

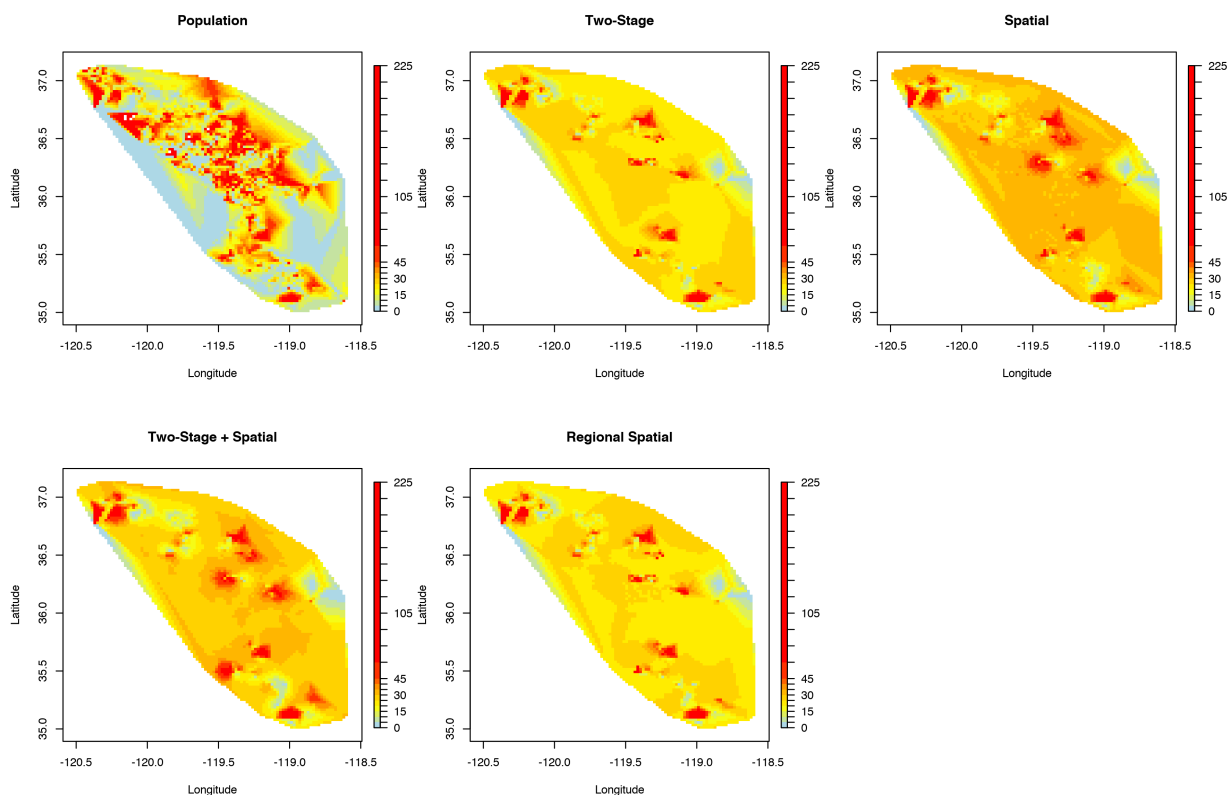


Figure 3.6: Interpolated surfaces using the population (truth) and posterior predictive samples from the four models.

Figure 3.6 shows the interpolated population surface from the complete sample and the interpolated surface from posterior predictive samples. While there are common regions at high risk (nitrate level greater than 45 mg/L) in all the posterior predictive maps, the spatial and two-stage + spatial maps predict larger regions. Also seen in Table 3.3, it is clear that Model 3 estimates a population mean that is larger than Models 1 and 4, but smaller than Model 3.

Table 3.3: Results of Data Analysis

Model	FP Mean (95% CI)	WAIC (SE)
1. Two-Stage	26.6 (19.2, 35.1)	4701.6 (78.6)
2. Spatial	32.6 (27.2, 38.7)	4858.4 (78.0)
3. Two-Stage + Spatial	31.2 (24.2, 37.3)	2697.0 (150.5)
4. Regional Spatial	26.2 (18.7, 34.4)	4536.6 (93.1)

Figure 3.7 provides spatial residual plots arising from the three spatial models. The spatial model which does not account for regional effects sees the most dispersed spatial effects, while the two-stage + spatial and regional spatial models show more localized spatial variability.

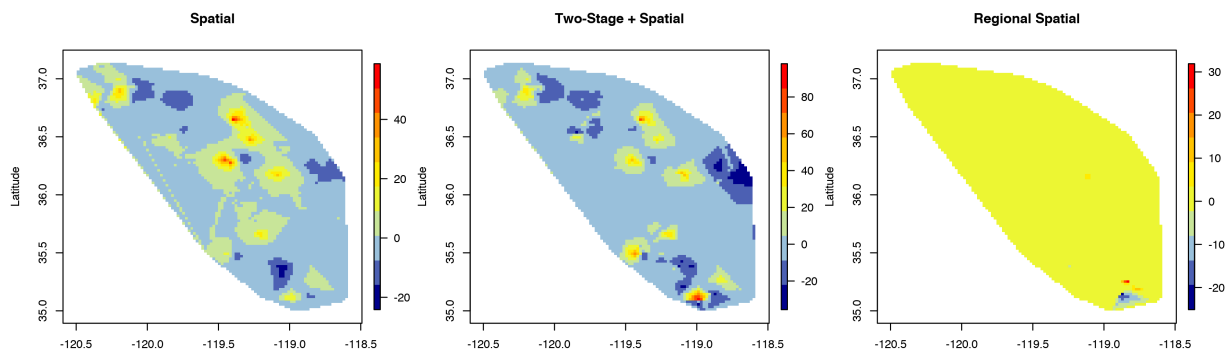


Figure 3.7: Spatial residual plots from the three spatial models.

3.7 Discussion

This paper examines the implications of performing two-stage random sampling on point-referenced data which exists in a spatial field. While Scott and Smith (1969) and Malec and Sedransk (1985) provided a Bayesian model-based framework to account for such a study design, we have demon-

strated that an analysis ignoring the underlying spatial correlation between locations or sampling design may lead to spurious inference and poorer model fit.

This work is a first step in developing an overarching framework for Bayesian finite population sampling from spatial process based populations. In our two-stage case, additional work may be done to further improve this model. For instance, CAR priors could be placed on regional parameters such as the μ_i 's, the regional means, to induce additional spatial correlation in the model. Also, many of the models presented account for regional differences in variance but if other sources of heteroscedasticity are suspected, new approaches (such as Zangeneh and Little 2015) may be needed to account for this. While an exponential covariance function was employed in the analyses in this paper, other spatial covariance functions could be used to create similar simulations and data analyses, as well as account for anisotropy.

Future work is needed to establish a more general framework that can account for more sophisticated sampling designs in a spatial context. The sampling designs presented in this paper are said to be ignorable (Rubin 1976; Sugden and Smith 1984), which allows us to perform inference on the superpopulation parameters while ignoring the inclusion probability distribution. However, designs in which the data cannot be assumed to be missing at random or where parameters define both the outcome and inclusion distributions are referred to as nonignorable and must account for the inclusion probability distribution. One example of this in the spatial context is preferential sampling (Diggle et al. 2010, Gelfand et al. 2012), in which the measurement values and sampling strategy are assumed to stem from the same spatial process. While Pati et al. (2011) have analyzed such data using Bayesian hierarchical models, an overall framework is needed to account for this and other non-ignorable design types.

Additionally, the implications of study design on finite population estimates when sampling from a spatially correlation population over time are unknown. In order to better understand this, these Bayesian models must first be extended to account for both study design and spatio-temporal associations.

Finally, while this paper provided a scaleable model which can account for study design and spatial correlation in massive survey data by assuming regional independence, further work should be done to incorporate recent strategies in modeling large spatial data (Heaton et al., 2018) when

analyzing survey data with spatial correlations, such as nearest neighbor processes (Datta et al., 2016), covariance tapering (Furrer et al., 2006), and metakriging (Guhaniyogi and Banerjee, 2018). Finite population models would particularly benefit from such techniques, as computation increases as a function of the population total, T , rather than the sample size, k .

3.8 Appendix

This section first provides the derivation of the empirical Bayesian estimators presented in Section 3.4.2 and then the finite population estimates presented in Sections 3.2 and 3.3.

The values a^* , b^* , M_ν , m_ν , M_β , and m_β for the general case (3.8) in Section 3.4.2 are presented below.

$$\begin{aligned}
a^* &= a + \frac{n}{2} \\
b^* &= b + \frac{1}{2} (y_s^\top y_s - y_s^\top \tilde{V}_s^{-1} X_s^\top [(\tilde{V}_\beta + A \tilde{V}_\nu A^\top)^{-1} + X_s^\top \tilde{V}_s^{-1} X_s]^{-1} X_s^\top \tilde{V}_s^{-1} y_s) \\
M_\nu &= [\tilde{V}_\nu^{-1} - A^\top \tilde{V}_\beta^{-1} (\tilde{V}_\beta^{-1} + A^\top X_s^\top \tilde{V}_s^{-1} X_s A)^{-1} \tilde{V}_\beta^{-1} A^\top]^{-1} \\
m_\nu &= A^\top \tilde{V}_\beta^{-1} (\tilde{V}_\beta^{-1} + A^\top X_s^\top \tilde{V}_s^{-1} X_s A)^{-1} A^\top X_s^\top \tilde{V}_s^{-1} y_s \\
M_\beta &= [\tilde{V}_\beta^{-1} + A^\top X_s^\top \tilde{V}_s^{-1} X_s A]^{-1} \\
m_\beta &= \tilde{V}_\beta^{-1} A \nu + A^\top X_s^\top \tilde{V}_s^{-1} y_s
\end{aligned}$$

The derivation of these values arise the conjugacy of the Normal and Inverse-Gamma distributions. We first derive (3.10) and (3.11). Take $\epsilon_s \sim N(0, \delta^2 \tilde{V}_s^{(\sigma)})$, $\epsilon_{ns} \sim N(0, \delta^2 \tilde{V}_{ns}^{(\sigma)})$, and $\nu \sim N(0, \delta^2 \tilde{\gamma}^2)$, where $\tilde{\gamma}^2 = \frac{1}{\delta^2} \gamma^2$, $\tilde{V}_s^{(\sigma)} = \frac{1}{\delta^2} V_s^{(\sigma)} = \left[\bigoplus_{i=1}^n \frac{\sigma_i^2}{\delta^2} I_{m_i} \right]$, and $\tilde{V}_{ns}^{(\sigma)} = \frac{1}{\delta^2} V_{ns}^{(\sigma)} = \left[\bigoplus_{i=1}^n \frac{\sigma_i^2}{\delta^2} I_{(M_i - m_i)} \right]$. Since the elements of y are independent conditional on μ , $V_{s,ns} = 0$ and $V_{ns,s} = 0$. Define observed group means as $\bar{y}_i = \frac{1}{m_i} \sum_{j=1}^n y_{ij}$ and the ratio of variances as $\lambda_i = \delta^2 / (\delta^2 + \sigma_i^2 / m_i)$ if $i \in (1, \dots, n)$ and $\lambda_i = 0$ if $i \in (n+1, \dots, N)$. Also define the vector of observed group variances to be $\tilde{\sigma}^2 = [\sigma_1^2, \dots, \sigma_n^2]^\top$. Recall $\mu_s = \nu 1_n + \epsilon_{\mu_s}; \epsilon_{\mu_s} \sim N(0, \delta^2 I_n)$, then $y_s = X_{s1} \mu_s + \epsilon_s = X_{s1} 1_n \nu + e_s^*$, where $e_s^* \sim N(0, \delta^2 [X_{s1} X_{s1}^\top + \tilde{V}_s^{(\sigma)}])$. Then we have that $p(\nu | y_s) \propto N(\nu | 0, \delta^2 \tilde{\gamma}^2) \times$

$N(y_s | X_{s1}1_n\nu, \delta^2[X_{s1}X_{s1}^\top + \tilde{V}_s^{(\sigma)}]) \propto N(\nu | Bb, \delta^2B)$, where

$$b = 1_n^\top X_{s1}^\top (X_{s1}X_{s1}^\top + \tilde{V}_s^{(\sigma)})^{-1} y_s = \left[1_{m_1}^\top, \dots, 1_{m_n}^\top \right] \left[\bigoplus_{i=1}^n \left(\frac{\delta^2}{\sigma_i^2} \right) \left(I_{m_i} - \frac{\frac{\delta^2}{\sigma_i^2} 1_{m_i} 1_{m_i}^\top}{1 + \frac{\delta^2}{\sigma_i^2} m_i} \right) \right] y_s \text{ and}$$

$$B^{-1} = \frac{1}{\tilde{\gamma}^2} + 1_n^\top X_{s1}^\top (X_{s1}X_{s1}^\top + \tilde{V}_s^{(\sigma)})^{-1} X_{s1}1_n = \frac{1}{\tilde{\gamma}^2} + \sum_{i=1}^n \left(\frac{\delta^2}{\sigma_i^2} \right) 1_{m_i}^\top \left(I_{m_i} - \frac{\frac{\delta^2}{\sigma_i^2} 1_{m_i} 1_{m_i}^\top}{1 + \frac{\delta^2}{\sigma_i^2} m_i} \right) 1_{m_i}.$$

Therefore, $B = \left[\frac{1}{\tilde{\gamma}^2} + \sum_{i=1}^n \lambda_i \right]^{-1}$ and $Bb = \frac{\sum_{i=1}^n \lambda_i \bar{y}_i}{\frac{1}{\tilde{\gamma}^2} + \sum_{i=1}^n \lambda_i}$. To solve $p(\delta^2 | y_s)$, split the posterior conditional distribution of the superpopulation parameters; $p(\delta^2, \nu | y_s) = IG(\delta^2 | a_\delta^*, b_\delta^*) \times N(\nu | Bb, \delta^2B)$, where $a_\delta^* = a_\delta + \frac{k}{2}$ and $b_\delta^* = b_\delta + \frac{1}{2} [y_s^\top (X_{s1}X_{s1}^\top + \tilde{V}_s^{(\sigma)})^{-1} y_s + b^\top Bb]$. As $p(\mu_{ns} | y_s, \nu, \delta^2) = p(\mu_{ns} | \nu, \delta^2)$, $\mu_{ns} | \nu, \delta^2 \sim N(1_{N-n}\nu, \delta^2 I_{N-n})$. To solve $p(\mu_s | y_s, \nu, \delta^2)$, note that $p(\mu_s | y_s, \nu, \delta^2) \propto N(\mu_s | 1_n\nu, \delta^2 I_n) \times N(y_s | X_{s1}\mu_s, \delta^2 \tilde{V}_s^{(\sigma)}) \propto N(\mu_s | B_* b_*, \delta^2 B_*)$, where

$$b_* = \nu 1_n + X_{s1}^\top (\tilde{V}_s^{(\sigma)})^{-1} y_s = \left[\nu + \frac{\delta^2}{\sigma_1^2} m_1 \bar{y}_1, \dots, \nu + \frac{\delta^2}{\sigma_n^2} m_n \bar{y}_n \right]^\top,$$

$$B_*^{-1} = I_n + X_{s1}^\top (\tilde{V}_s^{(\sigma)})^{-1} X_{s1} = \left[\bigoplus_{i=1}^n \frac{\sigma_i^2 + \delta^2 m_i}{\sigma_i^2} \right]; B_* = \left[\bigoplus_{i=1}^n \frac{\sigma_i^2}{\sigma_i^2 + \delta^2 m_i} \right] = \left[\bigoplus_{i=1}^n (1 - \lambda_i) \right], \text{ and}$$

$$B_* b_* = \left[\bigoplus_{i=1}^n (1 - \lambda_i) \right] \left[\nu + \frac{\delta^2}{\sigma_1^2} m_1 \bar{y}_1, \dots, \nu + \frac{\delta^2}{\sigma_n^2} m_n \bar{y}_n \right]^\top$$

$$= \left[(1 - \lambda_1)\nu + \lambda_1 \bar{y}_1, \dots, (1 - \lambda_n)\nu + \lambda_n \bar{y}_n \right]^\top.$$

To derive (3.12) and (3.13) define $\tilde{V}_\nu^{-1} = 0$ and $\tilde{V}_s = \tilde{\Omega}_s = \frac{1}{\delta^2} \Omega_s + I_k$, and $\tilde{V}_{\mu_s} = 1$. Note that $p(\mu_s | y_s, \delta^2) \propto N(\mu_s | 0, \delta^2) \times N(y_s | 1_k \mu_s, \delta^2 \tilde{\Omega}_s) \propto N(\mu_s | B_{\mu_s} b_{\mu_s}, \delta^2 B_{\mu_s})$, where $B_{\mu_s} = (1 + 1_k^\top \tilde{\Omega}_s^{-1} 1_k)^{-1}$ and $b_{\mu_s} = 1_k^\top \tilde{\Omega}_s^{-1} y_s$. Splitting the posterior conditional distribution of the superpopulation parameters, $p(\delta^2, \mu_s | y_s) = IG(\delta^2 | a_\delta^{**}, b_\delta^{**}) \times N(\mu_s | B_{\mu_s} b_{\mu_s}, \delta^2 B_{\mu_s})$, where $a_\delta^{**} = a + \frac{k}{2}$ and $b_\delta^{**} = b + \frac{1}{2} y_s^\top \left\{ \tilde{\Omega}_s^{-1} - \tilde{\Omega}_s^{-1} 1_k \left(1 + 1_k^\top \tilde{\Omega}_s^{-1} 1_k \right)^{-1} 1_k^\top \tilde{\Omega}_s^{-1} \right\} y_s$.

We now continue by deriving the general cases presented in Section 3.2. As $\beta | \nu \sim N(A\nu, V_\beta)$ and $\nu \sim N(0, \gamma^2)$, we have that $\beta \sim N(0, \gamma^2 AA^\top + V_\beta)$. Note that $p(\beta | y_s) \propto N(0, \gamma^2 AA^\top + V_\beta) \times N(X_s \beta, V_s) \propto N(V_\beta | y_s X_s^\top V_s^{-1} y_s, V_\beta | y_s)$, where $V_\beta | y_s = [(\gamma^2 AA^\top + V_\beta)^{-1} + X_s^\top V_s^{-1} X_s]^{-1}$. Defining

$B_V = V_{ns,s} + (X_{ns} - V_{ns,s}V_s^{-1}X_s)V_{\beta|y_s}X_s^\top$ and $Q = X_{ns} - V_{ns,s}V_s^{-1}X_s$, we have that:

$$\begin{aligned}
\mathbb{E}[\alpha^\top y | y_s] &= \alpha_s^\top y_s + \alpha_{ns}^\top \mathbb{E}[\mathbb{E}[y_{ns} | \beta, y_s] | y_s] \\
&= \alpha_s^\top y_s + \alpha_{ns}^\top \mathbb{E}[X_{ns}\beta + V_{ns,s}V_s^{-1}(y_s - X_s\beta) | y_s] \\
&= \alpha_s^\top y_s + \alpha_{ns}^\top V_{ns,s}V_s^{-1}y_s + QV_{\beta|y_s}X_s^\top V_s^{-1}y_s \\
&= \{\alpha_s^\top + \alpha_{ns}^\top [V_{ns,s} + \alpha_{ns}^\top QV_{\beta|y_s}X_s^\top]V_s^{-1}\}y_s, \\
\text{Var}[\mathbb{E}[\alpha^\top y | y_s]] &= \text{Var}[(\alpha_s^\top + \alpha_{ns}^\top B_V V_s^{-1})y_s] \\
&= \alpha_s^\top V_s \alpha_s + 2\alpha_{ns}^\top B_V \alpha_s + \alpha_{ns}^\top B_V V_s^{-1} B_V^\top \alpha_{ns}, \text{ and} \\
\text{Var}[\mathbb{E}[\alpha^\top y | y_s]] &= \text{Var}[\alpha_{ns}^\top y_{ns} | y_s] \\
&= \text{Var}[\mathbb{E}[\alpha_{ns}^\top y_{ns} | \beta, y_s] | y_s] + \mathbb{E}[\text{Var}[\alpha_{ns}^\top y_{ns} | \beta, y_s] | y_s] \\
&= \text{Var}[\alpha_{ns}^\top Q\beta | y_s] + \mathbb{E}[\alpha_{ns}^\top (V_{ns} - V_{ns,s}V_s^{-1}V_{s,ns})\alpha_{ns} | y_s] \\
&= \alpha_{ns}^\top (QV_{\beta|y_s}Q^\top + V_{ns} - V_{ns,s}V_s^{-1}V_{s,ns})\alpha_{ns}.
\end{aligned}$$

To derive the estimate given in Example 1, note that $p(\mu | y_s) \propto N(\mu | 0, \xi^2) \times N(y_s | 1_n \mu, \sigma^2 I_n) \propto N(\mu | B_{srs} b_{srs}, B_{srs})$, where $B_{srs} = (\frac{1}{\xi^2} + \frac{n}{\sigma^2})^{-1}$, $b_{srs} = \frac{1}{\sigma^2} 1_n^\top y_s$, and $B_{srs} b_{srs} = \frac{\frac{1}{\sigma^2} 1_n^\top y_s}{\frac{1}{\xi^2} + \frac{n}{\sigma^2}} = \frac{\sum_{i=1}^n y_i}{\frac{\sigma^2}{\xi^2} + n}$. Fixing the variance components, the finite population estimate is

$$\begin{aligned}
\mathbb{E}[\alpha^\top y | y_s] &= \alpha_s^\top y_s + \alpha_{ns}^\top \mathbb{E}[\mathbb{E}[y_{ns} | \mu, y_s] | y_s] = \alpha_s^\top y_s + \alpha_{ns}^\top 1_{(N-n)} \mathbb{E}[\mu | y_s] \\
&= \sum_{i=1}^n \alpha_i y_i + \frac{\sum_{i=n+1}^N \alpha_i}{\frac{\sigma^2}{\xi^2} + n} \sum_{i=1}^n y_i = \sum_{i=1}^n \left(\alpha_i + \frac{\sum_{i=n+1}^N \alpha_i}{\frac{\sigma^2}{\xi^2} + n} \right) y_i.
\end{aligned}$$

To derive (3.3), it is helpful to first make a note regarding X_{s1} vs X_s in the calculation of $p(\nu | y_s)$ and $p(\mu | \nu, y_s)$. Specifically, $p(\nu | y_s)$ does not change, since $X_s = [X_{s1} : 0]$, $b = 1_n^\top X_{s1}^\top (X_{s1} X_{s1}^\top + \tilde{V}_s^{(\sigma)})^{-1} y_s = 1_N^\top X_s^\top (X_s X_s^\top + \tilde{V}_s^{(\sigma)})^{-1} y_s$. Similarly, $B^{-1} = \frac{1}{\tilde{\gamma}^2} + 1_n^\top X_{s1}^\top (X_{s1} X_{s1}^\top + \tilde{V}_s^{(\sigma)})^{-1} X_{s1} 1_n = \frac{1}{\tilde{\gamma}^2} + 1_N^\top X_s^\top (X_s X_s^\top + \tilde{V}_s^{(\sigma)})^{-1} X_s 1_N$. However, while computing $p(\mu_s | \nu, y_s)$ using X_{s1} is computationally convenient for interpretation, employing X_s provides us the posterior distribution $p(\mu | \nu, y_s)$. We have that $p(\mu | y_s, \nu) \propto N(\mu | \nu 1_N, \delta^2 I_N) \times N(y_s | X_s \mu, \delta^2 \tilde{V}_s^{(\sigma)}) \propto N(\mu | B_{**} b_{**}, \delta^2 B_{**})$. Some algebra simplifies the expressions for b_{**} and $B_{**} b_{**}$ and matches the conclusions found by deriving

$p(\mu_s | \nu, y_s)$ and $p(\mu_{ns} | \nu, y_s)$ separately:

$$\begin{aligned}
b^{**} &= \nu 1_N + X_s^\top (\tilde{V}_s^{(\sigma)})^{-1} y_s = \left[\nu + \frac{\delta^2}{\sigma_1^2} m_1 \bar{y}_1, \dots, \nu + \frac{\delta^2}{\sigma_n^2} m_n \bar{y}_n, \nu 1_{(N-n)}^\top \right]^\top ; \\
B_{**}^{-1} &= I_N + X_s^\top (\tilde{V}_s^{(\sigma)})^{-1} X_s = \begin{bmatrix} \bigoplus_{i=1}^n \frac{\sigma_i^2 + \delta^2 m_i}{\sigma_i^2} & 0 \\ 0 & I_{(N-n)} \end{bmatrix} ; B_{**} = \left[\bigoplus_{i=1}^N (1 - \lambda_i) \right] ; \text{ and} \\
B_{**} b_{**} &= \left[(1 - \lambda_1) \nu + \lambda_1 \bar{y}_1, \dots, (1 - \lambda_n) \nu + \lambda_n \bar{y}_n, \nu 1_{(N-n)}^\top \right]^\top .
\end{aligned}$$

Using these derivations, define $\lambda = [\lambda_1, \dots, \lambda_N]^\top$ and $\bar{y} = [\bar{y}_1, \dots, \bar{y}_n, 0_{(N-n)}]^\top$. Then fixing the variance components, we have:

$$\begin{aligned}
\mathbb{E}[\alpha^\top y | y_s] &= \alpha_s^\top y_s + \alpha_{ns}^\top \mathbb{E}[\mathbb{E}[y_{ns} | \mu, \nu, y_s] | \nu, y_s | y_s] = \alpha_s^\top y_s + \alpha_{ns}^\top \mathbb{E}[\mathbb{E}[X_{ns} \mu | \nu, y_s] | y_s] \\
&= \alpha_s^\top y_s + \alpha_{ns}^\top X_{ns} \mathbb{E}[\left[(1 - \lambda_1) \nu + \lambda_1 \bar{y}_1, \dots, (1 - \lambda_n) \nu + \lambda_n \bar{y}_n, \nu 1_{(N-n)}^\top \right]^\top | y_s] \\
&= \alpha_s^\top y_s + \alpha_{ns}^\top X_{ns} \left[\bigoplus_{i=1}^N (1 - \lambda_i) \right] 1_N \frac{\sum_{i=1}^n \lambda_i \bar{y}_i}{1/\tilde{\gamma}^2 + \sum_{i=1}^n \lambda_i} + \alpha_{ns}^\top X_{ns} \left[\bigoplus_{i=1}^N \lambda_i \right] \bar{y} \\
&= \alpha_s^\top y_s + [\alpha_1(1 - \lambda_1), \dots, \alpha_N(1 - \lambda_N)]^\top 1_N \frac{\sum_{i=1}^n \lambda_i \bar{y}_i}{1/\tilde{\gamma}^2 + \sum_{i=1}^n \lambda_i} + [\alpha_1 \lambda_1, \dots, \alpha_N \lambda_N]^\top \bar{y} \\
&= \sum_{i=1}^n \sum_{j=1}^{m_i} \alpha_{ij} y_{ij} + \left\{ \sum_{i=1}^N \alpha_i (1 - \lambda_i) \right\} \frac{\sum_{i=1}^n \frac{\lambda_i}{m_i} \sum_{j=1}^{m_i} y_{ij}}{1/\tilde{\gamma}^2 + \sum_{i=1}^n \lambda_i} + \sum_{i=1}^n \alpha_i \frac{\lambda_i}{m_i} \sum_{j=1}^{m_i} y_{ij} \\
&= \sum_{i=1}^n \sum_{j=1}^{m_i} \left(\alpha_{ij} + \left[\alpha_i + \frac{\sum_{i=1}^N \alpha_i (1 - \lambda_i)}{1/\tilde{\gamma}^2 + \sum_{i=1}^n \lambda_i} \right] \frac{\lambda_i}{m_i} \right) y_{ij}
\end{aligned}$$

Now consider the stratified case for estimating the population mean. Taking non-informative priors for the group means, μ , is equivalent to letting $\delta^2 \rightarrow \infty$ and $\gamma^2 \rightarrow \infty$. Therefore $\lambda_i = \frac{\delta^2}{\delta^2 + \sigma_i^2/m_i} \rightarrow 1$, for all $i = 1, \dots, N$. Note $\alpha_i = \sum_{j=m_i+1}^{M_i} \frac{1}{T} = \frac{M_i - m_i}{T}$, $i = 1, \dots, n$. We have that:

$$\lim_{\delta^2, \gamma^2 \rightarrow \infty} \mathbb{E} \left[\frac{1}{T} 1_T^\top y | y_s \right] = \sum_{i=1}^n \sum_{j=1}^{m_i} \left(\frac{1}{T} + \left[\frac{M_i - m_i}{T} + 0 \right] \frac{1}{m_i} \right) y_{ij} = \sum_{i=1}^n \frac{M_i}{T} \bar{y}_i$$

To derive (3.5), note $p(\nu | y_s, \tau^2, \Omega_s, V_s^{(\sigma)}) \propto N(y_s | X_s 1_N \nu, \delta^2 X_s X_s^\top + \Omega_s + V_s^{(\sigma)}) \times N(\nu | 0, \gamma^2) \propto N(\nu | B_{sp2} b_{sp2}, B_{sp2})$, where $B_{sp2} = (\frac{1}{\gamma^2} + 1_N^\top X_s^\top (\delta^2 X_s X_s^\top + \Omega_s + V_s^{(\sigma)})^{-1} X_s 1_N)^{-1}$ and $b_{sp2} =$

$1_N^\top X_s^\top (\delta^2 X_s X_s^\top + \Omega_s + V_s^{(\sigma)})^{-1} y_s$. Consider the non-spatial case and define $\lambda^\top = [\lambda_1, \dots, \lambda_N] = 1_N^\top X_s^\top (X_s X_s^\top + \tilde{V}_s^{(\sigma)})^{-1} X_s$, then $B = \left[\frac{1}{\tilde{\gamma}^2} + 1_N^\top X_s^\top (X_s X_s^\top + \tilde{V}_s^{(\sigma)})^{-1} X_s 1_N \right]^{-1} = \left[\frac{1}{\tilde{\gamma}^2} + \lambda^\top 1_N \right]^{-1} = \left[\frac{1}{\tilde{\gamma}^2} + \sum_{i=1}^n \lambda_i \right]^{-1}$, which agrees with our previous findings.

Similarly, define $\lambda^{*\top} = [\lambda_1^*, \dots, \lambda_N^*] = 1_N^\top X_s^\top (\delta^2 X_s X_s^\top + \Omega_s + V_s^{(\sigma)})^{-1} X_s$.

Then $B_{sp2} = \left[\frac{1}{\tilde{\gamma}^2} + \lambda^{*\top} 1_N \right]^{-1} = \left[\frac{1}{\tilde{\gamma}^2} + \sum_{i=1}^n \lambda_i^* \right]^{-1}$.

Additionally, $p(\mu | y_s, \nu) \propto N(\mu | \nu 1_N, \delta^2 I_N) \times N(y_s | X_s \mu, \Omega_s + V_s^{(\sigma)}) \propto N(\mu | B_{sp2*} b_{sp2*}, B_{sp2*})$.

Here $B_{sp2*} = \left(\frac{1}{\delta^2} I_N + X_s^\top (\Omega_s + V_s^{(\sigma)})^{-1} X_s \right)^{-1}$ and $b_{sp2*} = \frac{1}{\delta^2} 1_N \nu + X_s^\top (\Omega_s + V_s^{(\sigma)})^{-1} y_s$.

Fixing the variance parameters, we have that:

$$\begin{aligned}
\mathbb{E}[\alpha^\top y | y_s] &= \alpha_s^\top y_s + \alpha_{ns}^\top \mathbb{E}[\mathbb{E}[y_{ns} | \mu, \nu, y_s] | \nu, y_s] \\
&= \alpha_s^\top y_s + \alpha_{ns}^\top \mathbb{E}[\mathbb{E}[X_{ns} \mu + \Omega_{ns,s} (\Omega_s + V_s^{(\sigma)})^{-1} (y_s - X_s \mu) | \nu, y_s] | y_s] \\
&= \alpha_s^\top y_s + \alpha_{ns}^\top \Omega_{ns,s} (\Omega_s + \sigma^2 V_s^{(\sigma)})^{-1} y_s + \alpha_{ns}^\top [X_{ns} - \Omega_{ns,s} (\Omega_s + V_s^{(\sigma)})^{-1} X_s] \times \\
&\quad \mathbb{E} \left[\left(\frac{1}{\delta^2} I_N + X_s^\top (\Omega_s + V_s^{(\sigma)})^{-1} X_s \right)^{-1} \left(\frac{1}{\delta^2} 1_N \nu + X_s^\top (\Omega_s + V_s^{(\sigma)})^{-1} y_s \right) | y_s \right] \\
&= \alpha_s^\top y_s + \alpha_{ns}^\top \Omega_{ns,s} (\Omega_s + \sigma^2 V_s^{(\sigma)})^{-1} y_s + \alpha_{ns}^\top [X_{ns} - \Omega_{ns,s} (\Omega_s + V_s^{(\sigma)})^{-1} X_s] \times \\
&\quad \left(\frac{1}{\delta^2} I_N + X_s^\top (\Omega_s + V_s^{(\sigma)})^{-1} X_s \right)^{-1} X_s^\top (\Omega_s + V_s^{(\sigma)})^{-1} y_s + \\
&\quad \alpha_{ns}^\top [X_{ns} - \Omega_{ns,s} (\Omega_s + V_s^{(\sigma)})^{-1} X_s] \times \\
&\quad \left(\frac{1}{\delta^2} I_N + X_s^\top (\Omega_s + V_s^{(\sigma)})^{-1} X_s \right)^{-1} \frac{1}{\delta^2} 1_N \frac{1_N^\top X_s^\top (\delta^2 X_s X_s^\top + \Omega_s + V_s^{(\sigma)})^{-1} y_s}{\frac{1}{\tilde{\gamma}^2} + \sum_{i=1}^n \lambda_i^*} \\
&= \sum_{i=1}^n \sum_{j=1}^{m_i} \left(\alpha_{ij} + \alpha_{ns}^\top \left\{ \Omega_{ns,s} (\Omega_s + V_s^{(\sigma)}) + \right. \right. \\
&\quad [X_{ns} - \Omega_{ns,s} (\Omega_s + V_s^{(\sigma)})^{-1} X_s] \left(\frac{1}{\delta^2} I_N + X_s^\top (\Omega_s + V_s^{(\sigma)})^{-1} X_s \right)^{-1} \times \\
&\quad \left. \left. \left[X_s^\top (\Omega_s + V_s^{(\sigma)})^{-1} + \frac{\frac{1}{\delta^2} 1_N 1_N^\top X_s^\top (\delta^2 X_s X_s^\top + \Omega_s + V_s^{(\sigma)})^{-1}}{\frac{1}{\tilde{\gamma}^2} + \sum_{i=1}^n \lambda_i^*} \right] \right\} q_{ij} \right) y_{ij}.
\end{aligned}$$

Chapter 4

Finite Population Estimation of Food Expenditure in the Presence of Spatially Correlated Data

4.1 Introduction

Income is associated with numerous health outcomes, including obesity (Pickett et al., 2005), mental health (Wildman, 2003), self-rated health (Kennedy et al., 1998), and mortality (Lynch et al., 1998). Specifically, in the realm of food purchasing, disparities by income exist in fruit and vegetable (FV) consumption, (Grimm et al., 2012), nutrition (Casey et al., 2001), and overall food insecurity (Ribar and Hamrick 2003 and Rose 1999). These problems are observed in “food swamps”, communities with higher number of unhealthy establishment which serve fast-food or sell junk food (Rose et al., 2009) than stores with healthy food options. Corner store interventions are one public health strategy to change the food environment in the hope of improving eating behaviors at the individual and community level (Langellier et al., 2013). Such interventions commonly increase the amount of fresh FVs sold in a store (Langellier et al., 2013), and may provide refrigeration units (Paek et al., 2014), a store remodeling, cooking demonstrations (Ortega et al., 2015), increased signage (Lawman et al., 2015), and business consulting (Ortega et al., 2015). Among these studies, findings regarding availability and sales of fruits, vegetables, and other healthy foods has been mixed (Thorndike et al. 2017; Albert et al. 2017; Paek et al. 2014; Lawman et al. 2015; and Song et al. 2009).

However, the focus of analysis in these interventions have been the patrons of these corner stores while few studies have examined the effect of these interventions at the community level. Notably, in such an intervention in two low-income, predominantly Hispanic communities in California, East Los Angeles and Boyle Heights, Ortega et al. (2016) reported no significant improvements to FV purchasing or consumption. However, one variable of interest, the percent of annual reported income spent on fruits and vegetables, was unable to be examined due to a high rate of missing data in reported income. As many of the intervention components targeted the community, it is

important to assess if the intervention impacted the percent of annual reported income spent on FV.

Non-response of household income is a common occurrence in survey research (Watson and Starick 2011; Yan et al. 2010; and Schenker et al. 2006), but the imputation of income in these communities provides two challenges. First, there is evidence that reported income is spatially associated in neighborhoods (Chakravorty 1996 and Breau et al. 2018). Secondly, individuals with lower incomes may be less likely to report their income and this should be accounted for. Similar descriptions of this income response have been given in Greenlees et al. (1982) and Riphahn and Serfling (2005), although both of these studies suggested individuals with higher incomes are less likely to respond. As we suspect our outcome is spatially associated, however, we turn to the recent literature regarding preferential sampling to better understand this problem. First described in Diggle et al. (2010), preferential sampling is a technique in which the probability of selection on a spatial domain increases as a function of intensity of the measurement. Diggle presents a joint model in which the selection sites and the measured values arise from the same spatial process. Pati et al. (2011) presents a model for preferential sampling in a fully Bayesian framework by including a function of intensity as a predictor of the outcome to account for informative sampling. Additionally, preferential sampling are shown to give biased predictions (Gelfand et al. 2012 and Lee et al. 2015) and parameter estimation (Antonelli et al. 2016). In our corner store scenario, we consider “preferential” response, in which the probability of a spatially associated variable being reported is dependent on the value of that variable. In the non-spatial case, such scenarios of missing can be accounted for using “selection” models described in Little and Rubin (2002). Finally, as we are interested in estimating average percentage of income spent on fruits and vegetables for all individuals in a community, we examine the problem from a finite population perspective, considering those who reported income to be the sampled or observed cases.

The rest of the chapter is as follows: Section 4.2 elaborates on data collected during the corner store intervention alluded to above in Ortega et al. (2016) and provides an in depth explanation of the income non-response by community, Section 4.3 presents a Bayesian framework that allows us to account for preferential nonresponse, and Section 4.4 examines a simulation study of our framework. Section 4.5 presents a data analysis to determine if there is a significant intervention effect on the

percentage of income spent on fruits and vegetables and describes the finite population estimates pre- and post-intervention. The chapter concludes with a brief discussion of our framework and the data analysis in Sec 4.6.

4.2 Data

Researchers carefully identified 4 corner stores in East Los Angeles for conversion and 4 corner stores in Boyle Heights to act as control sites. Corner store conversions included a reorganization of store items to promote healthy food purchasing, an external transformation of the store, a social marketing campaign and cooking demonstrations led by local youth, connections to local wholesale markets, and refrigeration units. A full discussion of the study design and implementation is described by Ortega et al. (2015). In order to assess the effect of this intervention, a survey was given to residents within a five block radius of each of the eight corner stores. This community survey sought to extensively catalog the food purchasing of residents, including where they shopped, what types of food they bought, and who they were purchasing food for. As such, the survey was limited to only adults who were the main food purchaser of the family. Many other items, such as demographics, health problems, family history of residency, and food program participation (such as food stamps or WIC), were also collected. This survey was conducted in each of the eight communities surrounding the store before the conversion and then again roughly one year after the conversion. There were 1,035 observations collected at baseline and 1,052 observations collected at follow-up. Roughly 60% of the individuals surveyed at baseline were surveyed again at follow-up.

While there is a strong interest in describing the percentage of income spent on fruits and vegetables in each community, the sample had high levels of missingness in income (one-third) at both baseline and follow-up, which are presented in Table 4.1. Noticeably, this outcome is highest at baseline in communities 1 and 7, 26.0% and 46.5%, respectively, which also observed lower levels of response and income compared to the averages of the total. With high levels of non-response, it is important to know if this value is being inflated due to the missing values of income. Additionally, while the number of sampled units ranged from 114 to 143, the percentage of missingness ranged widely from 5.90% to 66.62%. For this paper, we consider the sampled data to be the finite

population of eight communities in East Los Angeles and Boyle Heights. This is a reasonable assumption, as the response rate of 80% and 71% at baseline and follow-up suggest that a majority of individuals in these communities are represented in this dataset. Amount spent on fruits and vegetables was reported on weekly, bi-weekly, or monthly scale. These values were multiplied by 52, 26, and 12, respectively, to reflect the annual amount spent on fruits and vegetables in a household. Reported yearly income is continuous and ranged from \$0 to \$300,000. Both of these values were log-transformed to produce a more normal distribution of the outcome. Twenty-four individuals reported a higher amount spent on fruits and vegetables than their income, so their income was imputed to the value spent on food, so that the percent of income spent on fruits and vegetables was bounded by 0 and 100. Data was restricted to observations which had no missing covariates and reported the amount spent on fruits and vegetables. This resulted in a final dataset with 982 observations at baseline and 1033 at followup.

Table 4.1: Annual Income and FV Expenditures by Site and Time-point.

Site	Time	N	Income		FV Expenditure	
			% Response	M (SD)	M (SD)	% of Income M(SD)
1	B	117	54.70	24.79 (25.23)	2.10 (1.39)	26.0 (33.3)
	F	124	49.19	37.57 (27.98)	2.62 (1.81)	11.2 (14.9)
2	B	137	72.99	33.04 (25.20)	2.25 (1.41)	11.4 (13.2)
	F	134	78.36	32.64 (33.63)	2.58 (1.65)	13.6 (16.0)
3	B	122	59.84	25.77 (16.28)	2.37 (1.52)	16.8 (19.5)
	F	130	88.46	26.17 (22.33)	2.40 (1.70)	13.6 (11.4)
4	B	131	58.02	24.68 (17.32)	2.19 (1.66)	13.6 (14.8)
	F	128	34.38	29.90 (20.41)	2.29 (1.87)	11.4 (10.9)
5	B	117	70.94	39.82 (43.28)	2.26 (1.68)	10.6 (11.3)
	F	143	95.10	35.95 (36.19)	2.36 (1.59)	10.9 (10.7)
6	B	114	64.04	28.67 (22.72)	1.95 (1.50)	9.2 (7.1)
	F	129	72.09	29.77 (28.58)	1.80 (1.30)	10.5 (14.0)
7	B	125	61.60	17.52 (20.02)	2.18 (1.28)	46.5 (43.7)
	F	122	52.46	31.68 (29.38)	2.28 (1.50)	15.6 (16.7)
8	B	119	54.62	39.39 (46.85)	2.23 (1.55)	15.2 (21.9)
	F	123	52.85	34.12 (33.52)	2.31 (1.76)	11.3 (10.4)
Total	B	982	62.22	29.39 (29.70)	2.20 (1.50)	18.3 (25.7)
	F	1033	66.12	32.13 (30.31)	2.33 (1.67)	12.3 (13.3)

Note: B corresponds to Baseline and F corresponds to Follow-up. Income and FV expenditure are presented in units of \$1000.

Other individual-level variables that were hypothesized to affect the percent of annual income spent on fruits and vegetables were age at time of interview, gender, household size, marital status, and education level. Marital status was reported as one of six categories (single, married, separated, divorced, widowed, and living with a partner in a marriage-like relationship), but this was consolidated as to whether the respondent was in a marriage or marriage-like relationship or not. A small number of don't know/refused were placed in the non-marriage category. Education reduced from twenty-seven distinct categories to two: less than a high school education and at least a high school education. Similarly, a small number of non-responses were placed in the less than high school category. Due to the homogeneity of ethnicity in the sample, identifying as Hispanic was not considered in the analyses.

Individual locations (addresses) were provided and geographic coordinates were assigned to each address. As there were multiple apartment complexes in these communities, individuals living in different units of the same complex were assigned the same geographic coordinates. Thus among the 8 communities, there were 635 identified locations. At baseline, 518 of these locations were observed, 366 of these locations had a least one individual who reported their income, and on average 1.90 individuals shared the same location. At follow-up, 562 of these locations were observed, 472 of these locations had a least one individual who reported their income, and on average 2.38 individuals shared the same location. 555 locations had at least one reported income at either time-point.

Variograms of the outcome and log-income were constructed and both suggested evidence of spatial association. To explore our primary outcome and determine if there is any evidence of preferential response in income, a linear model was first fit using the previously described covariates, as well as a indicators for time-point, intervention status, and the interaction of these two indicators to detect an interaction effect, predicting the log-percent of income spent on fruits and vegetables. For individuals who did not report income, predictions of this log-percent were made using the results of the linear model. By dividing the reported amount spent on fruits and vegetables by this percent, we have constructed a prediction of income for the non-respondents. Then, a logistic regression model predicting the response of income was fit with an intercept term and income (either the reported value for those that responded or the predicted value from the linear model for those that did not respond). This model found income was a significantly associated with the probability

of response. An observed coefficient estimate of 0.12 (SE = 0.05) suggests that individuals with higher values of income are more likely to report income, and conversely, lower income in these communities are more likely to be underreported. Further, a logistic regression model with random intercepts for location was fit and the standard deviation corresponding to the random intercept was 0.67.

4.3 A General Framework

Formally, define a spatial domain $\mathcal{L} \subseteq \mathfrak{R}^2$, where a finite population of size T is located in N locations, $\mathcal{L}_{FP} = \{\ell_1, \dots, \ell_N\}$, $T \geq N$. Suppose there are the M_i units at the i^{th} location, then $T = \sum_{i=1}^N M_i$. Further, suppose that t , $t \leq T$, units are sampled from the finite population and thus n , $n \leq N$, locations are represented in this sample. Taking the first n locations to be sampled, define the sampled and nonsampled location sets as $\mathcal{L}_s = \{\ell_1, \dots, \ell_n\}$ and $\mathcal{L}_{ns} = \{\ell_{n+1}, \dots, \ell_N\}$, respectively. Additionally, denoting m_i the number of sampled units at the i^{th} location, $i = 0, \dots, M_i$, we have that $t = \sum_{i=1}^N m_i = \sum_{i=1}^n m_i$, as $m_i = 0$ for $i = n + 1, \dots, N$. In the context the data, we have that $T = 2015$, $t = 1294$, $N = 635$, and $n = 555$. We are interested in measuring annual reported income on the natural log scale, \mathbf{y} , which is a vector of sampled and nonsampled measurements, e.g. $\mathbf{y} = [\mathbf{y}_s^\top, \mathbf{y}_{ns}^\top]^\top$. Denoting $y_j(\ell_i)$ as the annual income on the natural log scale of the j^{th} individual at the i^{th} location, let $\mathbf{y}_s = [y_1(\ell_1), \dots, y_{m_1}(\ell_1), \dots, y_1(\ell_n), \dots, y_{m_n}(\ell_n)]^\top$ and $\mathbf{y}_{ns} = [y_{m_1+1}(\ell_1), \dots, y_{M_1}(\ell_1), \dots, y_{m_N+1}(\ell_N), \dots, y_{M_N}(\ell_N)]^\top$. Additionally, let $z_j(\ell_i)$ be the reported amount spent on fruits and vegetables on the natural log scale corresponding to $y_j(\ell_i)$. This is measured for all members of the finite population and therefore vectors \mathbf{z}_s and \mathbf{z}_{ns} , defined in the same manner as \mathbf{y}_s and \mathbf{y}_{ns} , denote reported values of FV expenditures corresponding to individuals who reported and did not report income, respectively. We examine the log percent of income spent on fruits and vegetables, which can be written as $\mathbf{z} - \mathbf{y}$, by modeling \mathbf{y} with an offset term of \mathbf{z} . Assume that there is a Gaussian spatial process, $\omega(\cdot)$, defined on \mathcal{L} with covariance function $K_\omega(d)$, and that \mathbf{y} is a partial realization of this process. Finally, define the inclusion mechanism as a spatial process on \mathcal{L} , which is dependent on \mathbf{y} and another Gaussian spatial process, $v(\cdot)$, defined on the same domain with covariance function $K_v(d)$. A joint model defined

in the form of our generic spatial paradigm (1.6) is:

$$[y(\cdot) | \omega(\cdot)] \times [I(\cdot) | y(\cdot), v(\cdot)] \times [\omega(\cdot)] \times [v(\cdot)] \quad (4.1)$$

The first component of (4.1) is the conditional distribution of \mathbf{y} , $[y(\cdot) | \omega(\cdot)]$. Assuming \mathbf{y} is an $T \times 1$ vector, this conditional distribution can be written as:

$$y_j(\ell_i) = z_j(\ell_i) - \mathbf{x}_j(\ell_i)^\top \boldsymbol{\beta} + \omega(\ell_i) + \epsilon_j(\ell_i) \quad ; \quad \epsilon_j(\ell_i) \stackrel{iid}{\sim} N(0, \sigma^2) \quad . \quad (4.2)$$

Here $i = 1, \dots, 635$, $j = 1, \dots, M_i$, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$, where $\boldsymbol{\Sigma}_\epsilon = \sigma^2 \mathbf{I}$. Each $\epsilon_j(\ell_i)$ corresponds to $y_j(\ell_i)$ and $\boldsymbol{\epsilon}$ is defined in the same manner as \mathbf{y} . Similarly, define the covariates corresponding to the j^{th} unit at the i^{th} location as $\mathbf{x}_j(\ell_i)$. Here each 10×1 vector $\mathbf{x}_j(\ell_i)$ is corresponds to the outcome $y_j(\ell_i)$. This vector of coefficients corresponds to the 10×1 vector $\boldsymbol{\beta}$, $\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\beta)$, and includes an intercept term, gender, household size, relationship status, age, age², timepoint, intervention status, and an interaction between intervention status and timepoint. Following the notational convention of \mathbf{y}_s and \mathbf{y}_{ns} , define the 2015×10 matrix $\mathbf{X} = [\mathbf{X}_s^\top, \mathbf{X}_{ns}^\top]^\top$ as the collection of covariates from sampled and nonsampled individuals, where $\mathbf{X}_s = [\mathbf{x}_1(\ell_1), \dots, \mathbf{x}_{m_1}(\ell_1), \dots, \mathbf{x}_1(\ell_n), \dots, \mathbf{x}_{m_n}(\ell_n)]^\top$ and $\mathbf{X}_{ns} = [\mathbf{x}_{m_1+1}(\ell_1), \dots, \mathbf{x}_{M_1}(\ell_1), \dots, \mathbf{x}_{m_N+1}(\ell_N), \dots, \mathbf{x}_{M_N}(\ell_N)]^\top$. Additionally, note that as the offset $z_j(\ell_i)$ is placed on the right-hand side of this equation, we subtract the $\mathbf{x}_j(\ell_i)^\top \boldsymbol{\beta}$ term to improve interpretation. In this way, a positive component in $\boldsymbol{\beta}$ corresponds to a positive increase in $\mathbf{z} - \mathbf{y}$, our outcome of interest.

Spatial variation is accounted for with the 635×1 vector $\boldsymbol{\omega} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\omega)$, where $\boldsymbol{\Sigma}_\omega$ is a 635×635 matrix defined by the covariance function $K_\omega(d)$. Finally, construct a 2015×635 site indicator matrix $\mathbf{A} = [\mathbf{A}_s^\top, \mathbf{A}_{ns}^\top]^\top$, where $\mathbf{A}_s = [\oplus_{i=1}^{555} \mathbf{1}_{m_i} : \mathbf{0}]$ and $\mathbf{A}_{ns} = [\oplus_{i=1}^{635} \mathbf{1}_{M_i - m_i}]$. Thus the row in \mathbf{A} corresponding to measurement $y_j(\ell_i)$ has value 1 in i^{th} column and 0 elsewhere. We then have that $\mathbf{y} \sim N(\mathbf{z} - \mathbf{X}\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\omega}, \boldsymbol{\Sigma}_\epsilon)$.

The second component of (4.1), $[I(\cdot) | y(\cdot), v(\cdot)]$ describes the response mechanism. Here the $T \times 1$ vector \mathbf{I} has element $I_j(\ell_i) = 1$ if the corresponding j^{th} individual in the i^{th} location reported their income, e.g. $y_j(\ell_i)$ is observed, and $I_j(\ell_i) = 0$ if they did not report their income. This can

be expressed as:

$$I_j(\boldsymbol{\ell}_i) \sim \text{Ber}(\pi_j(\boldsymbol{\ell}_i)) \quad ; \quad \text{logit}(\pi_j(\boldsymbol{\ell}_i)) = y_j(\boldsymbol{\ell}_i)\eta_y + \mathbf{q}_j(\boldsymbol{\ell}_i)^\top \boldsymbol{\eta} + v(\boldsymbol{\ell}_i) \quad . \quad (4.3)$$

The probability of response for each individual in the finite population is permitted to vary by its corresponding value of y , which is captured in the regression coefficient η_y , $\eta_y \sim N(0, \sigma_{\eta_y}^2)$. Similar to our modeling of the outcome, $\mathbf{q}_j(\boldsymbol{\ell}_i)$ is a 2×1 vector composed of an intercept term and age, which corresponds to a 2×1 vector of coefficients $\boldsymbol{\eta}$, $\boldsymbol{\eta} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$. Additional spatial variability in the probability of inclusion is accounted for with \mathbf{v} , $\mathbf{v} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_v)$, where $\boldsymbol{\Sigma}_v$ is a 635×635 matrix and is defined by covariance function $K_v(d)$. If we wish to orthogonalize v with respect to y , define \mathbf{Z} to be the 635×635 lower diagonal matrix that arises from the Cholesky decomposition of $\boldsymbol{\Sigma}_v$, such that $\boldsymbol{\Sigma}_v = \mathbf{Z}\mathbf{Z}^\top$, and let $\mathbf{u} \sim N(\mathbf{0}, I_{635})$. Let $\boldsymbol{\pi}$ be defined analogously to \mathbf{y} and \mathbf{q} to \mathbf{x} , then Equation 4.3 can be rewritten as $\text{logit}(\boldsymbol{\pi}) = \mathbf{y}\eta_y + \mathbf{q}^\top \boldsymbol{\eta} + \mathbf{AZu}$. Defining the projection matrix $\mathbf{p}_y = \mathbf{y}(\mathbf{y}^\top \mathbf{y})^{-1} \mathbf{y}^\top$, we can decompose $\mathbf{AZu} = \mathbf{p}_y \mathbf{AZu} + (\mathbf{I} - \mathbf{p}_y) \mathbf{AZu}$. Thus, $\text{logit}(\boldsymbol{\pi}) = \mathbf{y}\eta_y^* + \mathbf{q}^\top \boldsymbol{\eta} + (\mathbf{I} - \mathbf{p}_y) \mathbf{AZu}$, where $\eta_y^* = (\mathbf{y}^\top \mathbf{y})^{-1} \mathbf{y}^\top \mathbf{AZu}$.

Additionally, we take the two processes, ω and v , to be independent. Collecting additional variance parameters in $\boldsymbol{\theta}$, the joint posterior distribution of (4.1) is proportional to:

$$\begin{aligned} & p(\boldsymbol{\omega}, \mathbf{v}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\eta}, \eta_y, \mathbf{y}_{ns} | \mathbf{y}_s, \mathbf{I}) \\ & \propto p(\boldsymbol{\theta}) \times N(\boldsymbol{\omega} | \mathbf{0}, \boldsymbol{\Sigma}_\omega) \times N(\mathbf{v} | \mathbf{0}, \boldsymbol{\Sigma}_v) \times N(\boldsymbol{\beta} | \mathbf{0}, \boldsymbol{\Sigma}_\beta) \times N(\boldsymbol{\eta} | \mathbf{0}, \boldsymbol{\Sigma}_\eta) \times N(\eta_y | 0, \sigma_{\eta_y}^2) \quad (4.4) \\ & \times \prod_{i=1}^N \prod_{j=1}^{M_i} \text{Ber}(I_j(\boldsymbol{\ell}_i) | \pi_j(\boldsymbol{\ell}_i)) \times \prod_{i=1}^N \prod_{j=1}^{M_i} N(y_j(\boldsymbol{\ell}_i) | z_j(\boldsymbol{\ell}_i) - \mathbf{x}_j(\boldsymbol{\ell}_i)^\top \boldsymbol{\beta} + \omega(\boldsymbol{\ell}_i), \sigma^2), \end{aligned}$$

where

$$\begin{aligned} \text{Ber}(I_j(\boldsymbol{\ell}_i) | \pi_j(\boldsymbol{\ell}_i)) &= \left(\frac{\exp[y_j(\boldsymbol{\ell}_i)\eta_y + \mathbf{q}_j(\boldsymbol{\ell}_i)^\top \boldsymbol{\eta} + v(\boldsymbol{\ell}_i)]}{1 + \exp[y_j(\boldsymbol{\ell}_i)\eta_y + \mathbf{q}_j(\boldsymbol{\ell}_i)^\top \boldsymbol{\eta} + v(\boldsymbol{\ell}_i)]} \right)^{I_j(\boldsymbol{\ell}_i)} \\ & \times \left(\frac{1}{1 + \exp[y_j(\boldsymbol{\ell}_i)\eta_y + \mathbf{q}_j(\boldsymbol{\ell}_i)^\top \boldsymbol{\eta} + v(\boldsymbol{\ell}_i)]} \right)^{1 - I_j(\boldsymbol{\ell}_i)}. \end{aligned}$$

Markov chain Monte Carlo must be used to sample from (4.4). A Gibbs update can be employed

to sample the posterior distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$, which are

$$\boldsymbol{\beta}|\cdot \sim N\left(\left(\boldsymbol{\Sigma}_\beta^{-1} + \mathbf{X}_s^\top \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{X}_s\right)^{-1} \mathbf{X}_s^\top \boldsymbol{\Sigma}_\epsilon^{-1} (\mathbf{y}_s - \mathbf{z}_s - \mathbf{A}_s \boldsymbol{\omega}), \left(\boldsymbol{\Sigma}_\beta^{-1} + \mathbf{X}_s^\top \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{X}_s\right)^{-1}\right) \quad \text{and}$$

$$\boldsymbol{\omega}|\cdot \sim N\left(\left(\boldsymbol{\Sigma}_\omega^{-1} + \mathbf{A}_s^\top \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{A}_s\right)^{-1} \mathbf{A}_s^\top \boldsymbol{\Sigma}_\epsilon^{-1} (\mathbf{y}_s - \mathbf{z}_s + \mathbf{X}_s \boldsymbol{\beta}), \left(\boldsymbol{\Sigma}_\omega^{-1} + \mathbf{A}_s^\top \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{A}_s\right)^{-1}\right),$$

respectively. The conditional distributions for the remaining parameters are not available in closed form and must be sampled using a Metropolis-Hastings step. Specifically, we have:

$$\begin{aligned} \mathbf{y}_{ns}|\mathbf{y}_s, \boldsymbol{\omega}, \boldsymbol{\theta}, \boldsymbol{\beta} &\propto p(\boldsymbol{\theta}) \times N(\boldsymbol{\omega}|\mathbf{0}, \boldsymbol{\Sigma}_\omega) \times N(\boldsymbol{\beta}|\mathbf{0}, \boldsymbol{\Sigma}_\beta) \times \prod_{i=1}^N \prod_{j=1}^{M_i} \text{Ber}(I_j(\boldsymbol{\ell}_i)|\pi_j(\boldsymbol{\ell}_i)) \\ &\times \prod_{i=1}^N \prod_{j=1}^{M_i} N(y_j(\boldsymbol{\ell}_i)|z_j(\boldsymbol{\ell}_i) - \mathbf{x}_j(\boldsymbol{\ell}_i)^\top \boldsymbol{\beta} + \boldsymbol{\omega}(\boldsymbol{\ell}_i), \sigma^2), \end{aligned}$$

$$\boldsymbol{\eta}|\mathbf{y}, \boldsymbol{\theta} \propto p(\boldsymbol{\theta}) \times N(\boldsymbol{\eta}|\mathbf{0}, \boldsymbol{\Sigma}_\eta) \times \prod_{i=1}^N \prod_{j=1}^{M_i} \text{Ber}(I_j(\boldsymbol{\ell}_i)|\pi_j(\boldsymbol{\ell}_i)),$$

$$\eta_y|\mathbf{y}_s, \boldsymbol{\theta} \propto p(\boldsymbol{\theta}) \times N(\eta_y|0, \sigma_{\eta_y}^2) \times \prod_{i=1}^N \prod_{j=1}^{M_i} \text{Ber}(I_j(\boldsymbol{\ell}_i)|\pi_j(\boldsymbol{\ell}_i)), \text{ and}$$

$$\mathbf{v}|\mathbf{y}_s, \boldsymbol{\theta} \propto p(\boldsymbol{\theta}) \times N(\mathbf{v}|\mathbf{0}, \boldsymbol{\Sigma}_v) \times \prod_{i=1}^N \prod_{j=1}^{M_i} \text{Ber}(I_j(\boldsymbol{\ell}_i)|\pi_j(\boldsymbol{\ell}_i)).$$

The posterior samples of \mathbf{y}_{ns} are then used to obtain posterior finite population estimates. Specifically, we are interested in the mean income of finite population, $\exp[\frac{1}{T} \sum_{i=1}^N \sum_{j=1}^{M_i} y_j(\boldsymbol{\ell}_i)]$, and the mean of the percent of income spent on fruits and vegetables, $\frac{1}{T} \sum_{i=1}^N \sum_{j=1}^{M_i} \exp[z_j(\boldsymbol{\ell}_i) - y_j(\boldsymbol{\ell}_i)]$. These values are calculated overall, by site and by timepoint.

Four models are considered in the form of (4.1) and are described below. For these models, regression parameters are considered independent, e.g. $\boldsymbol{\Sigma}_\beta = \sigma_\beta^2 \mathbf{I}_{10}$ and $\boldsymbol{\Sigma}_\eta = \sigma_\eta^2 \mathbf{I}_2$, and their associated variance parameters, σ_β^2 and σ_η^2 , are fixed in both the simulation and data analysis. Similarly, σ_{η_y} is fixed. The spatial covariance functions are taken to be exponential, as described in Sec 1.4.

Model 1. Non-Spatial in Outcome with Ignorable Response This model is a standard linear regression model and therefore spatial effects ($\boldsymbol{\omega}$ and \mathbf{v}) are fixed at 0. Inclusion parameters

are also fixed at 0 so that the probability of inclusion is a fixed number. We take $\boldsymbol{\theta} = \sigma^2$ and $p(\boldsymbol{\theta}) = IG(\sigma^2|a, b)$.

Model 2. Non-Spatial Association in Outcome with Preferential Response Preferential response is now accounted for through η_y but spatial effects are again fixed at 0. Similar to Model 1, $\boldsymbol{\theta} = \sigma^2$ and $p(\boldsymbol{\theta}) = IG(\sigma^2|a, b)$.

Model 3. Spatial Association in Outcome with Preferential Response This model accounts for spatial association in the outcome but fixes $v = 0$. Therefore $\boldsymbol{\theta} = [\sigma^2, \delta_\omega^2, \phi_\omega]^\top$ and $p(\boldsymbol{\theta}) = IG(\sigma^2|a, b) \times IG(\delta_\omega^2|a_\omega, b_\omega) \times Unif(\phi_\omega|c_\omega, d_\omega)$.

Model 4. Spatial Association in Outcome and Probability of Inclusion with Preferential Response This model expands upon Model 3 by permitting spatial association in the probability of response. We take $\boldsymbol{\theta} = [\sigma^2, \delta_\omega^2, \phi_\omega, \delta_v^2, \phi_v]^\top$ and $p(\boldsymbol{\theta}) = IG(\sigma^2|a, b) \times IG(\delta_\omega^2|a_\omega, b_\omega) \times Unif(\phi_\omega|c_\omega, d_\omega) \times IG(\delta_v^2|a_v, b_v) \times Unif(\phi_v|c_v, d_v)$.

4.3.1 Model Comparison and Assessment

Model fit was evaluated in two ways. In general, consider a sample of size t drawn from a population of size T with outcome $\mathbf{y} = [\mathbf{y}_s^\top, \mathbf{y}_{ns}^\top]^\top$. Without loss of generality, say $y_h \in \mathbf{y}_s$ if $h = 1, \dots, t$ and $y_h \in \mathbf{y}_{ns}$ if $h = t + 1, \dots, T$. Replicated datasets, $\mathbf{y}_{rep}^{(l)} = [y_{rep,1}^{(l)} \dots y_{rep,t}^{(l)}]^\top$, can be generated from the pointwise posterior predictive distribution at each iteration l . These are used to formulate the predictive model choice criteria,

$$D = \sum_{h=1}^t (y_h - \mathbb{E}[y_{rep,h} | \mathbf{y}_s])^2 + \sum_{h=1}^t \text{var}(y_{rep,h} | \mathbf{y}_s)$$

described in Gelfand and Ghosh (1998), and the Gneiting-Raftery Score (Gneiting and Raftery, 2007),

$$GRS = - \sum_{h=1}^t \frac{(y_h - \mathbb{E}[y_{rep,h} | \mathbf{y}_s])^2}{\text{var}(y_{rep,h} | \mathbf{y}_s)} - \sum_{h=1}^t \log \text{var}(y_{rep,h} | \mathbf{y}_s) .$$

In this formulation, lower values of D and higher values of GRS are indicative of better model fit. For L iterations, we approximate $\mathbb{E}[y_{rep,h} | \mathbf{y}_s] \approx \frac{1}{L} \sum_{l=1}^L y_{rep,h}^{(l)}$ and $\text{var}(y_{rep,h} | \mathbf{y}_s) \approx \frac{1}{L-1} \sum_{l=1}^L (y_{rep,h}^{(l)} - \frac{1}{L} \sum_{l=1}^L y_{rep,h}^{(l)})^2$. For simulated datasets, where \mathbf{y}_{ns} is known, these measures can be extended to

all observations, e.g. summing to T instead of t in each score.

4.4 Simulation

To examine the ability of the proposed models to capture various sampling schemes, a simplified dataset was simulated and three response scenarios were implemented. For simplicity, in this simulation study we predict income (on the log-scale) with only the covariates gender and household size for a finite population of size 2,000, e.g. $z_j(\ell_i)$ is fixed at 0 and $-\mathbf{x}_j(\ell_i)$ is replaced by $\mathbf{x}_j(\ell_i)$ for all i and j in (4.2). For each unit of the population, gender was drawn from a bernoulli distribution with the probability of female set to 0.8 and household size was drawn from a poisson distribution with a mean of 4. To induce spatial correlation, a 5 x 5 square was created and 500 locations were randomly assigned within the square and distance matrix was constructed from these locations. The spatial process parameters were fixed at $\sigma^2 = 1$, $\delta_\omega^2 = 1$, and $\phi_\omega = 0.5$. Each unit of the population was randomly assigned to a location, with the requirement that at least one unit was located at each location. Regression parameters were fixed at $\boldsymbol{\beta} = [\beta_0, \beta_{fem}, \beta_{hhs}]^\top = [10, -0.2, 0.1]^\top$, to reflect an average income of $\exp(10) = \$22,000$ in the reference group, a small average reduction in income for females, and a small average increase in income for larger household sizes. Log-income values were generated from (4.2).

Three scenarios were considered to reflect possible response scenarios in which there is spatial association in the outcome. In scenario 1, income is from a spatial process but there is no preferential response. This arises from Model 3, fixing $\eta_y = 0$ and $\mathbf{q} = [1, \dots, 1]^\top$. The probability of inclusion was set at 0.5, which is equivalent to fixing $\eta = 0$. This resulted in a selection of 54% of the simulated data. In the second scenario, income is from a spatial process which is reported preferentially, as described in Model 3. Here, η_y was set to 0.5 and $\boldsymbol{\eta} = [\eta_0, \eta_{fem}] = [-4, -1]^\top$, to reflect higher odds of response for larger values of income and lower odds of response for women. The choice of these coefficients resulted in 54.15 % of the simulated data having income responses. The third scenario considers income as coming from a spatial process whose response is preferential and whose inclusion probability is dependent on another spatial process, which is described in Model 4. To reflect this, we set $\phi_v = 1.5$ and $\delta_v^2 = 1$; this resulted in responses in 48.1% of the simulated data.

All data generation and analyses were performed using R version 3.6.1 (R Core Team, 2018).

Linear interpolation plots from the full simulated data and the subset data from the three scenarios are shown in Fig 4.1. As expected, scenario 1 (a simple random sample) is the most similar to the full dataset. In the cases of preferential response (scenarios 2 and 3), the interpolated plots have larger regions of high income than the true dataset. This is most apparent in the western region of the graph, where values below 8 are rare in this instance. Comparing scenario 2 and 3, there appears to be some smoothing, with fewer pockets of low income in the west and northeast of the graph, which is due to the spatial association induced on the probability of response in scenario 3.

Models were run for 10,000 iterations with 1,000 burn-in, as examination of individual trace plots suggested sufficient mixing and convergence of the non-spatial parameters. At each iteration g , estimates of the nonsampled units were drawn and estimates for the population mean, $\bar{y}^{(g)} = \exp \left[\frac{1}{T} \left(\sum_{i=1}^n \sum_{j=1}^{m_i} y_j(\ell_i) + \sum_{i=1}^N \sum_{j=m_i+1}^{M_i} y_j(\ell_i)^{(g)} \right) \right]$ were calculated. The variance parameter σ_β^2 was fixed at 1,000 to reflect an uninformative prior, while the σ_η^2 and σ_{η_y} terms were fixed at 10 as a weakly informative prior restricting the range of the logistic regression coefficients. The non-spatial, σ^2 , and spatial, δ_ω^2 and δ_v , variance components were assigned prior distributions of IG(2,10), to reflect a small point mass centered at 10. The spatial range parameters, ϕ_ω and ϕ_v , were assigned prior distributions of Unif(0.1, 2), to reflect a spatial range of 1.5 (3/2) to 30 (3/0.1). MCMC sampling was performed using the computer program JAGS (Plummer, 2017) in R.

The results of Scenario 1 are presented in Table 4.2. While the credible intervals for each model contain the true value of regression coefficients for female and household size, as well as the true finite population mean, the non-spatial models fail to contain the true intercept and the non-spatial variance values in their credible intervals. As expected, both spatial models were able to correctly capture the spatial parameters, ϕ_ω , and δ_ω^2 , for the outcome. Additionally, the coefficients η_0 and η_y are small and have credible intervals containing 0 for Models 2 - 4, which suggests that these models correctly demonstrate no evidence of preferential response. The response-level spatial parameters in Model 4 also suggest no evidence of spatial variability, as the credible interval of ϕ_v is nearly the same range as the prior distribution given and the spatial variance, δ_v^2 , is very close to 0. Additionally, the fit of Model 4 is negligibly poorer than Model 3, as there is no spatial association

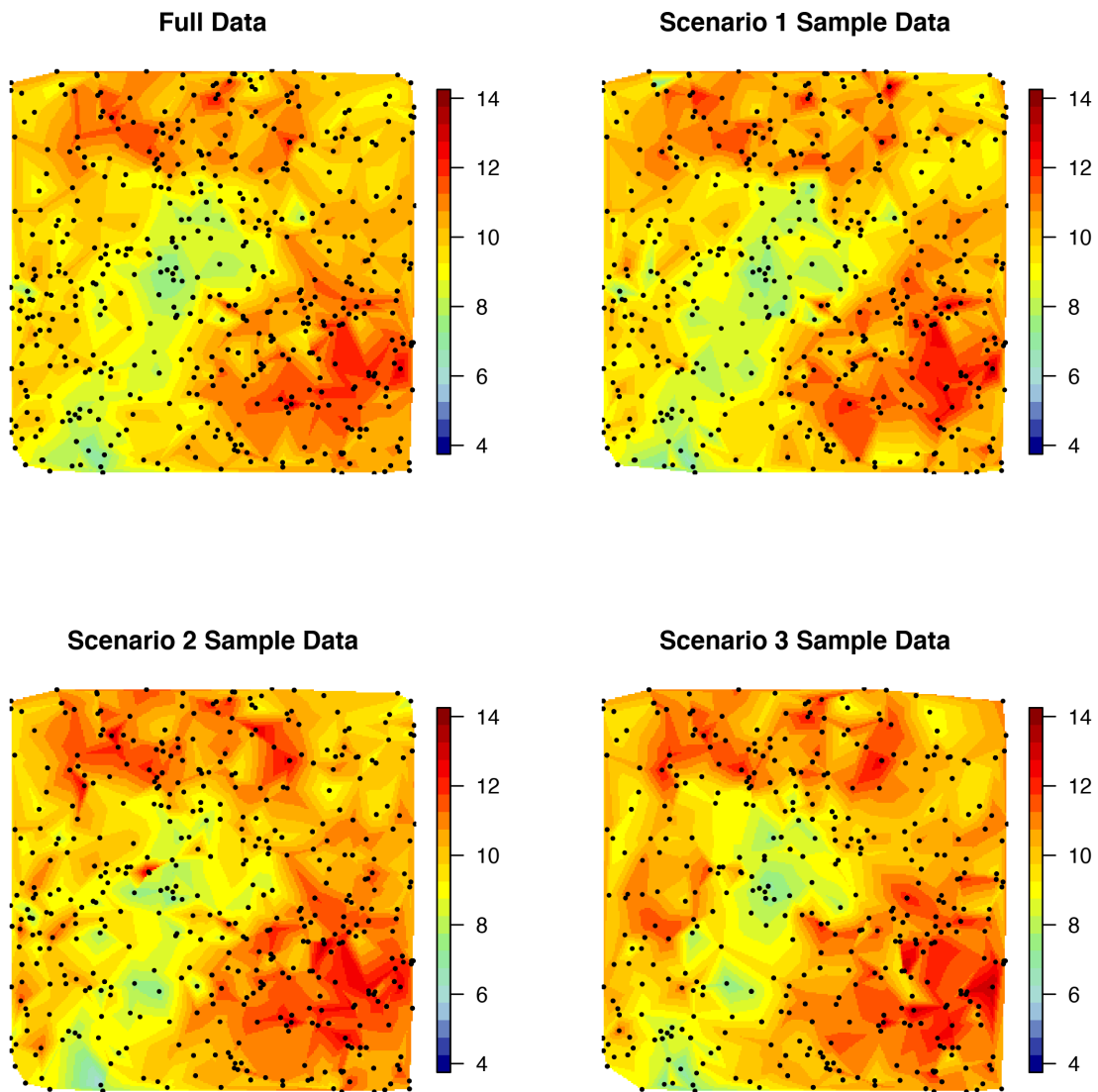


Figure 4.1: Linear Interpolation Plots from Full Simulated Data and 3 Scenarios in the probability of response.

The results of Scenario 2 are given in Table 4.3 and examines a preferential response of a spatially associated outcome. Importantly, unlike Scenario 1, the two non-spatial models fail to capture the true finite population mean of 9.94 within their 95% credible intervals. This is also true of the intercept term, β_0 , and non-spatial variance, σ^2 , although we expect σ^2 to be larger, as it

Table 4.2: Simulation Results of Scenario 1: Spatial Outcome, Random Response

	Model 1	Model 2	Model 3	Model 4
	Mean (95% CI)	Mean (95% CI)	Mean (95% CI)	Mean (95% CI)
\bar{y} (9.94)	9.94 (9.88, 10.00)	9.90 (9.41, 10.14)	9.91 (9.86, 9.98)	9.90 (9.83, 9.97)
β_0 (10)	9.55 (9.29, 9.82)	9.52 (9.06, 9.91)	9.49 (8.64, 10.32)	9.47 (8.54, 10.31)
β_{fem} (-0.2)	-0.21 (-0.42, 0.00)	-0.2 (-0.42, 0.01)	-0.17 (-0.32, -0.02)	-0.17 (-0.33, -0.01)
β_{hhs} (0.10)	0.14 (0.10, 0.19)	0.14 (0.10, 0.18)	0.12 (0.09, 0.15)	0.12 (0.09, 0.15)
σ^2 (1)	2.04 (1.87, 2.22)	2.08 (1.89, 2.38)	1.01 (0.92, 1.11)	1.01 (0.92, 1.11)
η_0 (0)		-0.24 (-5.24, 2.11)	-0.17 (-0.91, 0.72)	-0.45 (-1.49, 0.70)
η_y (0)		0.04 (-0.19, 0.58)	0.03 (-0.06, 0.11)	0.06 (-0.05, 0.17)
ϕ_ω (0.5)			0.79 (0.36, 1.29)	0.77 (0.30, 1.29)
δ_ω^2 (1)			0.87 (0.5, 1.59)	0.91 (0.52, 1.73)
ϕ_ν (0)				1.31 (0.33, 1.97)
δ_ν^2 (0)				0.05 (0.02, 0.13)
D	6799.8	6857.1	4163.0	4166.1
GRS	-3474.1	-3460.6	-2041.2	-2042.3

absorbing the variability in the outcome attributed to spatial association. Model 1 also incorrectly provides a positive estimative of β_{fem} whose credible interval does not contain the true value of -2 . Moreover, while Model 2 - 4 provide similar estimates of η_{fem} , Model 2 fails to capture the true values of η_0 and η_y in its credible intervals, unlike the two spatial models. Possibly due to the poor modeling of income, Model 2 spuriously concludes that there is no evidence of preferential sampling. Finally, similar to Scenario 1, the spatial models correctly capture the spatial parameters ϕ_ω and δ_ω^2 and Model 4 suggests little evidence of spatial association in the probability of response. The model fit statistics both slightly favor Model 3 to Model 4, due to the lack of response level spatial association, and prefer the spatial to non-spatial models.

When incorporating spatial association into the probability of income response, seen in Table 4.4, Model 4 outperforms the other three models in terms of model fit by correctly accounting for this additional association in the logistic regression component of the model. As before, non-spatial models have poorer model fit and larger estimates of the non-spatial variance term. Unlike Models 2-4, Model 1 fails to include the true finite population mean in its credible interval, which may be attributable to a disregard for the preferential response. However, the credible interval provided by Model 3 does not contain the true finite population mean as well. We investigate this

in the following paragraph. As in Scenario 2, Model 1 incorrectly provides a positive estimate of β_{fem} , and all models except Model 2 contain the true intercept in their credible intervals. Models 2 - 4 each correctly capture the logistic regression coefficients, η_0 , η_{fem} , and η_y . Additionally, the spatial models provide reasonable estimates of ϕ_ω and δ_ω^2 , and in the case of Model 4, ϕ_v and δ_v^2 .

Interestingly, in this third simulation, the credible intervals for the finite population mean given by Models 2 and 4 are similar and contain the truth. One concern arises from the inclusion of two spatial components in Equation 4.3, specifically y and v . It is possible that the two components result in spatial confounding (Hodges and Reich, 2010), which may influence inference on the finite population mean. To address this, we orthogonalize these two components, so that spatial random effect v is restricted to the residual space of y , as described in Section 4.3. Figure 4.2 presents the estimates of η_y (denoted standard), η_y^* (denoted orthogonal), and the difference, $\eta_y - \eta_y^*$. As the difference is centered at zero and the posterior distribution of η_y and η_y^* are similar, we conclude there is no evidence that spatial confounding contributes to this result.

Table 4.3: Simulation Results of Scenario 2: Spatial Outcome, Preferential Sampling

	Model 1	Model 2	Model 3	Model 4
	Mean (95% CI)	Mean (95% CI)	Mean (95% CI)	Mean (95% CI)
\bar{y} (9.94)	10.36 (10.31, 10.42)	10.17 (9.97, 10.41)	9.99 (9.92, 10.05)	10.00 (9.91, 10.08)
β_0 (10)	9.87 (9.66, 10.09)	9.74 (9.47, 10.01)	9.56 (8.63, 10.4)	9.54 (8.64, 10.37)
β_{fem} (-0.2)	0.08 (-0.10, 0.25)	-0.01 (-0.20, 0.19)	-0.14 (-0.28, 0.01)	-0.13 (-0.28, 0.01)
β_{hhs} (0.10)	0.11 (0.07, 0.15)	0.11 (0.07, 0.15)	0.11 (0.08, 0.14)	0.11 (0.07, 0.14)
σ^2 (1)	1.71 (1.57, 1.86)	1.77 (1.61, 1.96)	0.99 (0.89, 1.09)	0.98 (0.89, 1.09)
η_0 (-4)		-1.56 (-3.93, 1.42)	-3.50 (-4.52, -2.63)	-3.34 (-4.86, -1.94)
η_{fem} (-1)		-0.93 (-1.18, -0.69)	-0.92 (-1.18, -0.67)	-0.92 (-1.17, -0.68)
η_y (0.5)		0.25 (-0.05, 0.49)	0.44 (0.36, 0.55)	0.43 (0.29, 0.58)
ϕ_ω (0.5)			0.75 (0.29, 1.32)	0.73 (0.27, 1.28)
δ_ω^2 (1)			0.84 (0.46, 1.66)	0.84 (0.46, 1.68)
ϕ_v (0)				0.96 (0.11, 1.95)
δ_v^2 (0)				0.04 (0.01, 0.11)
D	6928.4	6721.4	4141.9	4144.2
GRS	-3829.3	-3609.0	-2048.3	-2055.0

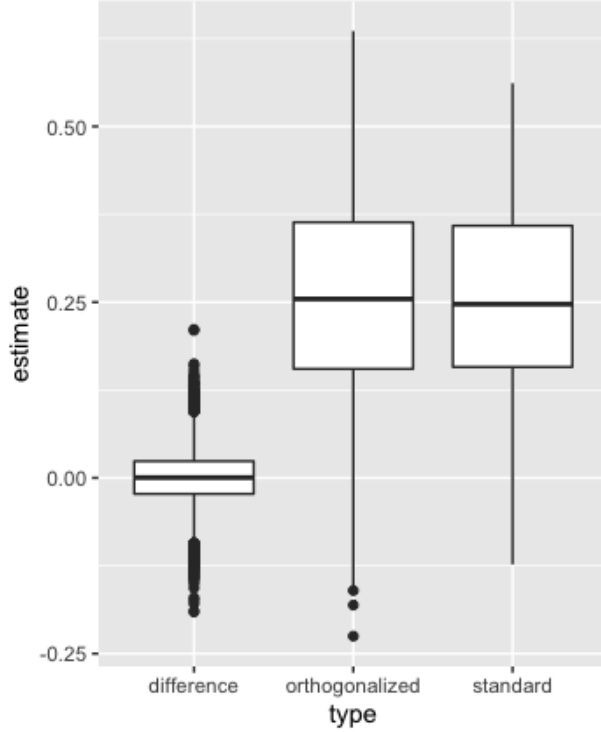


Figure 4.2: Comparison of Standard and Orthogonalized η_y Estimates from Simulation 3, Model 4

4.5 Data Analysis

4.5.1 Implementation

Similar to our simulations, Models 1-4 were implemented using JAGS (Plummer, 2017) in R and run for 10,000 iterations with 1,000 burn-in, as examination of individual trace plots suggested sufficient mixing and convergence of the non-spatial parameters. At each iteration g , the finite population mean income, $\bar{y}^{(g)} = \exp \left[\frac{1}{T} \left(\sum_{i=1}^n \sum_{j=1}^{m_i} y_j(\mathbf{l}_i) + \sum_{i=1}^N \sum_{j=m_i+1}^{M_i} y_j(\mathbf{l}_i)^{(g)} \right) \right]$, and the finite population mean percentage of income spent on fruits and vegetables, $\bar{y}^{(g)} = \frac{1}{T} \left(\sum_{i=1}^n \sum_{j=1}^{m_i} \exp [z_j(\mathbf{l}_i) - y_j(\mathbf{l}_i)] + \sum_{i=1}^N \sum_{j=m_i+1}^{M_i} \exp [z_j(\mathbf{l}_i) - y_j(\mathbf{l}_i)^{(g)}] \right)$, were calculated using estimates of the nonsampled units drawn at that iteration. The variance parameter σ_β^2 was fixed at 1,000 to reflect an uninformative prior, while the σ_η^2 and σ_{η_y} terms were fixed at 0.68 as a weakly informative prior restricting the range of the exponentiated logistic regression coefficients to $\frac{1}{5}$ and 5. The non-spatial σ^2 , and spatial, δ_ω^2 , variance components were assigned prior distribu-

tions of $IG(2,10)$ and $IG(2,2)$, respectively, to reflect small point masses centered at 10 and 2. The prior for δ_v was assigned to be uniform distribution ranging from 0 to 0.75, so that the standard deviation reported in Section 4.2 is included in this range. This tight prior was found to improve convergence in the other logistic regression parameters. The spatial range parameters, ϕ_ω and ϕ_v , were assigned prior distributions of $Unif(0.1, 2)$, to reflect a spatial range of 1.5 (3/2) to 30 (3/0.1).

4.5.2 Results

The results of this analysis are presented in Table 4.5. Notably, there is no evidence of an intervention effect on the percent of income spent on fruits and vegetable in any of the models, denoted by the coefficient $\beta_{treat*follow}$ being small and all credible intervals containing 0. Evidence of a significant intervention effect would have seen a larger positive coefficient. This finding supports previous findings of no community-level changes as reported in Ortega et al. (2016). The four models have similar agreement on all β regression coefficients, so the following interpretations are based on Model 4. All else equal, males spent 58% ($\exp(-0.54)$) of the outcome spent by women. Larger reported households were associated with higher amounts of household income spent on fruits and vegetables, with the outcome multiplicatively increasing by 15% for every additional household member. Food purchasers who reported having less than a high school education spent 1.8 times of the outcome of those with a high school diploma or more. There was a small negative linear effect of age on the outcome, as well as a small positive quadratic term. The percent of income spent on fruits and vegetables was also lower at follow-up, which is consistent with the raw percentages presented in Table 4.1. There were no differences were detected for partner status.

Confirming the preliminary analyses discussed in Section 4.2, all three models that account for preferential response conclude that larger incomes are more likely to provide their income. Models 2-4 agree that age is not associated with the probability of response. Accounting for association in the probability of response appears to also best fits the data, as evidenced by the lowest value of D and highest GRS value. Interestingly, the model fit for Model 2 is poorest (on the GRS scale), suggesting that accounting for preferential sampling while not accounting for spatial association (either at the outcome or response levels) leads to poorer fit. Also, Model 3 fits poorer than Model

1 (and Model 2 on the D scale), which suggests that spatial association at the outcome level may have been accounted for with the inclusion of additional covariates.

However, our estimation of the finite population mean of the percent of income spent on fruits and vegetables is very model specific. Most importantly, it is evident that in ignoring the presence of preferential sampling, Model 1 spuriously underestimates this percentage. The reason for this is clearly explained by examining each corresponding model's finite population estimate of income. As Model 1 does not account for the fact that individuals with lower incomes are less likely to report their income, there is much less variability in the average income of the community. This leads to a spurious estimate almost \$10,000 and 30% larger than the next closest estimate of \$29,364.66, given by Model 2. It is important to note that Model 1's estimates are also much larger than the averages presented in Table 4.1, while Models 2-4 present credible intervals that contain these values. While it is true that the additional variability from accounting for preferential sampling leads to larger posterior credible intervals, we note that no part of Model 1's credible interval is contained in any of the other models. Despite this apparent disagreement, Model 4's incorporation of spatial association in the response mechanism results in a compromise between Model 1 and 3. This trend is also observed in the finite population mean fraction, where higher estimated incomes in Model 1 correspond to much lower estimated fractions than the other models. Based on model fit statistics, we conclude that Model 4 provides the best estimate of the finite population fraction mean, which is 26%.

Additionally, as posterior samples are drawn for all individuals with non-response, finite population estimates can be constructed for each community at both timepoints, which are presented for each model in Table 4.6. Bolded estimates represent instances where the 95% credible intervals do not include the raw average reported in Table 4.1. Interpolated maps corresponding to these finite population estimates are presented in the supplemental material. Importantly, these results emphasize the importance of imputation. Models 2-4 show remarkable similarity in these estimates and conclude that raw data of 6 of the 8 sites underestimates the percentage of income spent on fruits and vegetables at baseline and all but 1 of the 8 underestimate at follow-up. Model 3 additionally identifies site 1 at baseline, but this is not supported by the rest of the models. Even in the case of Model 1, at baseline 3 were found to underestimate the percentage and 1 suggested overes-

timation, and at follow-up, 2 communities were found to underestimate as well. Encouragingly, in all but one case (site 7 at baseline) of the disagreements with the raw data that Model 1 identified, Models 2-4 also identified these cases. Additionally, Models 2-4 suggest that the baseline total is underestimating the true average and all models agree that the follow-up total is underestimated.

4.6 Discussion

This paper presents a new framework to account for data whose outcome is spatially associated and whose probability of response is assumed to be associated with the value of the outcome. We examine the implications of this data on finite population quantities and demonstrate how to perform bayesian estimation on these values. This work builds on an existing literature in spatial statistics, bayesian finite population estimation, and missing data and has a wide range of applications in health, economics, and environmental work.

Specifically, in our presented data analysis, we find that accounting for spatial association at both the outcome and probability levels provides the best model fit. By accounting for such associations and preferential responses in income, we are more confident in concluding that there was no effect on the percent of income spent on fruits and vegetables at the community level attributable to the corner store intervention. We were however, able to more accurately describe the individual communities by estimating finite population means at each site level. In fact, the finite population estimates of income that stem from the modeling ignoring both spatial association and preferential response are substantially larger than the other models and are less believable, given the community. This directly contributed to lower estimates of the percent of income spent on fruits and vegetables in these communities, compared to the other models. The public health importance of such estimation work is two-fold. In future projects in these regions, interventions that focus on FV access and knowledge could target areas with high estimated percentages. Additionally, future work can examine ways in which income information can be solicited from lower income neighborhoods and what factors may be driving this non-response (besides the level of income).

The literature of bayesian finite population estimation in the presence of spatial association is

small and future extensions to the work presented in this paper are numerous. While this model draws inspiration from the preferential sampling described by Diggle et al. (2010), we examined a missing data case that had similar evidence of preferential response. However, a data analysis implementing this technique on a dataset with preferential sampling from a finite population would be a strong addition to the literature. The authors view the framework discussed in Section 4.3 to be flexible enough to allow for other, more complicated sampling schemes as well, although more simulation work would be needed to fully understand the implications of these on finite population quantities, especially if spatial association is assumed.

Additionally, while the sample size presented in the data analysis of this paper was small, this framework can be extended to account for massive sample sizes. The problem of spatial modeling for big data stems from the inversion of dense covariance matrices, but modern work in covariance approximation has made this feasible. Such techniques include low-rank models, sparsity-inducing processes, and map reducing approaches. An excellent review of these techniques is given by Heaton et al. (2018).

Table 4.4: Simulation Results of Scenario 3: Spatial Outcome, Preferential Sampling, Spatial Inclusion

	Model 1	Model 2	Model 3	Model 4
	Mean (95% CI)	Mean (95% CI)	Mean (95% CI)	Mean (95% CI)
\bar{y} (9.94)	10.38 (10.32, 10.44)	9.91 (9.73, 10.07)	9.84 (9.74, 9.93)	9.99 (9.84, 10.13)
β_0 (10)	9.97 (9.73, 10.20)	9.58 (9.33, 9.85)	9.50 (8.54, 10.39)	9.60 (8.65, 10.54)
β_{fem} (-0.2)	0.06 (-0.13, 0.26)	-0.12 (-0.33, 0.09)	-0.14 (-0.3, 0.01)	-0.09 (-0.25, 0.07)
β_{hhs} (0.1)	0.09 (0.05, 0.13)	0.11 (0.07, 0.15)	0.10 (0.07, 0.14)	0.10 (0.07, 0.13)
σ^2 (1)	1.81 (1.66, 1.98)	2.04 (1.81, 2.29)	1.14 (1.02, 1.27)	1.07 (0.96, 1.19)
η_0 (-4)		-4.48 (-6.25, -2.87)	-5.04 (-6.15, -3.79)	-3.52 (-6.09, -1.07)
η_{fem} (-1)		-0.85 (-1.11, -0.61)	-0.84 (-1.1, -0.60)	-0.97 (-1.25, -0.70)
η_y (0.5)		0.52 (0.35, 0.72)	0.58 (0.46, 0.70)	0.44 (0.21, 0.70)
ϕ_ω (0.5)			0.67 (0.26, 1.20)	0.64 (0.22, 1.17)
δ_ω^2 (1)			0.84 (0.45, 1.73)	0.82 (0.40, 1.83)
ϕ_ν (1.5)				1.53 (0.58, 1.98)
δ_ν^2 (1)				0.91 (0.48, 1.82)
D	7035.3	6798.4	4397.1	4295.4
GRS	-3819.5	-3466.9	-2153.9	-2119.1

Table 4.5: Results of regression models predicting percentage of income spent on fruits and vegetables (log-scale)

	Model 1	Model 2	Model 3	Model 4
FP Avg. %	0.17 (0.16, 0.19)	0.26 (0.22, 0.31)	0.35 (0.28, 0.43)	0.26 (0.22, 0.32)
FP Avg. Inc.	40.85 (37.19, 45.86)	29.36 (26.66, 32.58)	25.58 (24.14, 27.71)	28.72 (26.18, 31.83)
β_0	-2.81 (-3.41, -2.21)	-2.22 (-2.86, -1.58)	-2.25 (-3.24, -1.35)	-2.42 (-3.31, -1.65)
β_{male}	-0.5 (-0.67, -0.34)	-0.53 (-0.7, -0.36)	-0.55 (-0.72, -0.37)	-0.54 (-0.71, -0.37)
$\beta_{partner}$	-0.11 (-0.25, 0.03)	-0.07 (-0.22, 0.07)	-0.02 (-0.17, 0.13)	-0.04 (-0.19, 0.1)
β_{hhs}	0.13 (0.09, 0.17)	0.14 (0.1, 0.18)	0.14 (0.1, 0.18)	0.14 (0.1, 0.18)
$\beta_{treatment}$	-0.01 (-0.21, 0.18)	-0.07 (-0.27, 0.12)	-0.02 (-0.9, 0.8)	0.06 (-0.74, 1.3)
$\beta_{followup}$	-0.25 (-0.45, -0.05)	-0.25 (-0.45, -0.06)	-0.2 (-0.4, 0)	-0.24 (-0.44, -0.04)
$\beta_{treat*follow}$	0.1 (-0.17, 0.37)	0.05 (-0.22, 0.32)	0.06 (-0.22, 0.33)	0.1 (-0.17, 0.37)
$\beta_{<HS}$	0.59 (0.44, 0.73)	0.63 (0.48, 0.78)	0.6 (0.45, 0.76)	0.6 (0.45, 0.74)
β_{age}	-0.012 (-0.036, 0.012)	-0.028 (-0.055, -0.002)	-0.0267 (-0.050, 0.003)	-0.025 (-0.048, -0.003)
β_{age^2}	1e-04 (-1e-04, 4e-04)	3e-04 (1e-04, 6e-04)	3e-04 (0, 5e-04)	3e-04 (1e-04, 5e-04)
σ^2	1.5 (1.39, 1.63)	1.64 (1.49, 1.81)	1.7 (1.52, 1.89)	1.53 (1.39, 1.7)
η_0		-4.19 (-5.63, -2.91)	-6.6 (-7.8, -5.04)	-3.57 (-5.01, -2.18)
η_{age}		0 (-0.01, 0.01)	0 (-0.01, 0.01)	0 (-0.01, 0.01)
η_y		0.5 (0.37, 0.65)	0.76 (0.6, 0.89)	0.54 (0.38, 0.71)
ϕ_ω			1.31 (0.21, 1.98)	1.38 (0.25, 1.98)
δ_ω^2			0.36 (0.15, 1.09)	0.31 (0.12, 0.77)
ϕ_ν				0.79 (0.24, 1.87)
δ_ν^2				0.67 (0.52, 0.75)
D	3527.8	3692.2	3701.2	3482
GRS	-1841.2	-1903.6	-1863	-1767.8

Note: The finite population mean income is presented in units of \$ 10,000.

Table 4.6: Finite Population Estimates and 95% CI of Percentage of Income Spent on Fruits and Vegetables by Community and Timepoint.

Site	Time	Data	Model 1	Model 2	Model 3	Model 4
1	B	26.0	23.2 (19.6, 29.5)	34.2 (25.5, 49.4)	45.9 (31.9, 70.0)	33.3 (24.8, 48.3)
	F	11.2	13.6 (10.4, 18.8)	24.4 (16.0, 38.8)	38.2 (23.7, 63.5)	23.9 (15.9, 37.4)
2	B	11.4	14.8 (11.8, 20.5)	21.9 (15.2, 34.5)	23.5 (16.0, 37.4)	20.4 (14.4, 31.5)
	F	13.6	14.7 (12.8, 18.4)	19.0 (14.7, 26.9)	21.0 (15.8, 30.6)	18.9 (14.7, 26.8)
3	B	16.8	19.6 (15.6, 26.6)	30.9 (21.7, 46.5)	35.5 (24.3, 54.4)	30.5 (21.5, 45.8)
	F	13.6	14.3 (13.0, 17.2)	17.1 (14.0, 23.9)	18.1 (14.5, 25.6)	16.7 (13.9, 22.3)
4	B	13.6	17.4 (13.6, 23.7)	29.2 (19.9, 45.5)	42.1 (26.8, 68.4)	27.4 (18.6, 42.9)
	F	11.4	17.0 (12.3, 24.4)	33.4 (21.6, 53.2)	55.5 (33.5, 90.8)	31.8 (20.1, 50.7)
5	B	10.6	15.3 (11.6, 22.2)	23.8 (15.7, 38.8)	22.6 (15.2, 36.0)	20.5 (14.1, 32.3)
	F	10.9	11.1 (10.6, 12.7)	12.0 (10.8, 15.3)	12.1 (10.8, 15.5)	11.8 (10.7, 14.6)
6	B	9.2	12.9 (9.8, 18.5)	21.5 (14.3, 34.6)	24.2 (15.6, 39.1)	19.4 (13.2, 30.3)
	F	10.5	11.7 (9.8, 15.2)	17.2 (12.4, 26)	20.0 (14.0, 31.3)	16.3 (12.0, 24.1)
7	B	46.5	38.0 (34.0, 45.2)	48.1 (39.2, 64.2)	69.8 (50.7, 100.00)	57.5 (44.0, 81.8)
	F	15.6	17.5 (13.8, 23.9)	26.7 (18.8, 40.4)	54.0 (33.8, 90.6)	41.1 (26.3, 66.6)
8	B	15.2	19.1 (14.7, 26.6)	31.0 (20.9, 48.3)	43.1 (27.5, 69.0)	30.8 (20.6, 47.9)
	F	11.3	15.4 (11.6, 21.6)	24.7 (16.8, 38.5)	37.8 (23.9, 61.8)	26.5 (17.3, 42.4)
Total	B	18.3	20.1 (18.2, 22.6)	30.1 (24.9, 37.6)	38.4 (30.7, 48.5)	30.0 (24.7, 37.4)
	F	12.3	14.4 (13.1, 16.1)	21.6 (17.7, 27.1)	31.5 (24.7, 40.9)	23.0 (18.8, 29.2)

Note: Models whose 95% credible intervals do not contain the raw mean average are bolded.
 One percentage has been capped at 100.0.

Chapter 5

Conclusion

This dissertation is the first work to examine finite population sampling in the presence of spatial association using Bayesian modeling techniques. The general Bayesian framework presented in Chapter 3 considers an ignorable sampling design and allows for posterior estimates of finite population quantities to be collected while correctly accounting for spatial correlation in both sampled and nonsampled units. Two-stage sampling was explored first with a simulation study and then a data analysis of the nitrate content California groundwater obtained through well samples. Both works concluded that a model which both correctly accounted for study design and spatial association had the best model fit and improved finite population estimates.

Survey data obtained from a corner store intervention was first analyzed in Chapter 2 using a more standard spatial analysis approach and then in Chapter 4, where we fully develop the general framework of Chapter 3 to allow for preferentially missing data. Importantly, Chapter 2 presents the coregionalization model as a technique to allow for two, possibly correlated, spatial random effects. While point-referenced data is common in food environment studies, analyses investigate and test for spatial association are rare. The dissemination of the analyses presented here can offer modeling advice to data where each observation is not only point-referenced, but is identified by a pair of locations. Further, this analysis provided another example of finite population estimation for ignorable missing data. The Bayesian framework presented in Chapter 4 is more flexible by accounting for nonignorability designs (or missingness) through a logistic regression. Additionally, we permit spatial association in both the outcome and sampling levels, which improved the model fit and lead to more reasonable finite population estimates in the real data analysis. “Preferential” sampling is permitted by incorporating the spatially associated outcome variable in the modeling of the sampling mechanism.

This dissertation combines themes found in the Bayesian statistics, spatial statistics, bayesian finite population estimation, and missing data literatures and many extensions to this work should

be researched. For instance, a spatiotemporal model would be able to incorporate the entire groundwater data discussed in Chapter 3, rather than be restricted to a single value at each well as presented in our analysis, and would provide population estimates at multiple time points. Further, this work has only considered a Gaussian Process to describe the outcome variable but our process based approach could be modified to allow various other processes, such as mixtures of Gaussian Processes, a generalized Gaussian Process, or a spatial Dirichlet Process. Such work would greatly improve the flexibility of such models to allow for more sophisticated associations. In continuing to develop this framework, our goal is to improve the quality and accuracy of the estimation of these finite population quantities, particularly in the context of Public Health.

Bibliography

- Albert, S. L., Langellier, B. A., Sharif, M. Z., Chan-Golston, A. M., Prelip, M. L., Garcia, R. E., Glik, D. C., Belin, T. R., Brookmeyer, R., and Ortega, A. N. (2017). A corner store intervention to improve access to fruits and vegetables in two latino communities. *Public health nutrition*, 20(12):2249–2259.
- Antonelli, J., Cefalu, M., and Bornn, L. (2016). The positive effects of population-based preferential sampling in environmental epidemiology. *Biostatistics*, 17(4):764–778.
- Auchincloss, A. H., Roux, A. V. D., Brown, D. G., Raghunathan, T. E., and Erdmann, C. A. (2007). Filling the gaps: spatial interpolation of residential survey data in the estimation of neighborhood characteristics. *Epidemiology (Cambridge, Mass.)*, 18(4):469.
- Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, Boca Raton, FL, 2nd edition.
- Banerjee, S., Gelfand, A. E., and Polasek, W. (2000). Geostatistical modelling for spatial interaction data with application to postal service performance. *Journal of statistical planning and inference*, 90(1):87–105.
- Boyle, D., King, A., Kourakos, G., Lockhart, K., Mayzelle, M., Fogg, G. E., and Harter, T. (2012). Groundwater Nitrate Occurrence, Technical Report 4. Technical report, Center for Watershed Sciences, University of California, Davis, Davis, CA.
- Breau, S., Shin, M., and Burkhart, N. (2018). Pulling apart: new perspectives on the spatial dimensions of neighbourhood income disparities in canadian cities. *Journal of Geographical Systems*, 20(1):1–25.
- Bruno, F., Cocchi, D., and Vaghegini, A. (2013). Finite population properties of individual predictors based on spatial pattern. *Environmental and Ecological Statistics*, 20(3):467–494.

- Casey, P. H., Szeto, K., Lensing, S., Bogle, M., and Weber, J. (2001). Children in Food-Insufficient, Low-Income Families: Prevalence, Health, and Nutrition Status. *Archives of Pediatrics and Adolescent Medicine*, 155(4):508–514.
- Cavanaugh, E., Green, S., Mallya, G., Tierney, A., Brensinger, C., and Glanz, K. (2014). Changes in food and beverage environments after an urban corner store intervention. *Preventive Medicine*, 65:7–12.
- Chakravorty, S. (1996). A measurement of spatial disparity: The case of income inequality. *Urban Studies*, 33(9):1671–1686.
- Chan-Golston, A. M., Banerjee, S., and Handcock, M. S. (2020). Bayesian inference for finite populations under spatial process settings. *Environmetrics*, 31(3):e2606. e2606 env.2606.
- Cicchitelli, G. and Montanari, G. E. (2012). Model-assisted estimation of a spatial population mean. *International Statistical Review*, 80(1):111–126.
- Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43(3):671–681.
- Cochran, W. G. (1977). *Sampling Techniques*. John Wiley & Sons, Hoboken, NJ, 3rd edition.
- Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. John Wiley & Sons, Hoboken, NJ.
- Cressie, N. and Zammit-Mangion, A. (2016). Multivariate spatial covariance models: a conditional approach. *Biometrika*, 103(4):915–935.
- Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812.
- Day, P. L. and Pearce, J. (2011). Obesity-promoting food environments and the spatial clustering of food outlets around schools. *American journal of preventive medicine*, 40(2):113–121.

- Dekker, L. H., Rijnks, R. H., Strijker, D., and Navis, G. J. (2017). A spatial analysis of dietary patterns in a large representative population in the north of the netherlands—the lifelines cohort study. *international journal of behavioral nutrition and physical activity*, 14(1):166.
- Diggle, P. J., Menezes, R., and Su, T.-L. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C*, 59(2):191–232.
- Ericson, W. A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society, Series B*, 31(2):195–233.
- Finley, A. O., Banerjee, S., and Carlin, B. P. (2007). spBayes: An R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, 19(4):1–24.
- Finley, A. O., Banerjee, S., and E.Gelfand, A. (2015). spBayes for large univariate and multivariate point-referenced spatio-temporal data models. *Journal of Statistical Software*, 63(13):1–28.
- Frank, L. D., Saelens, B. E., Chapman, J., Sallis, J. F., Kerr, J., Glanz, K., Couch, S. C., Learnihan, V., Zhou, C., Colburn, T., et al. (2012). Objective assessment of obesogenic environments in youth: geographic information system methods and spatial findings from the neighborhood impact on kids study. *American journal of preventive medicine*, 42(5):e47–e55.
- Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523.
- Gelfand, A. E. and Ghosh, S. K. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika*, 85(1):1–11.
- Gelfand, A. E., Sahu, S. K., and Holland, D. M. (2012). On the effect of preferential sampling in spatial prediction. *Environmetrics*, 23(7):565–578.
- Gelman, A. (2007). Struggles with Survey Weighting and Regression Modeling. *Statistical Science*, 22(2):153–164.

- Gelman, A., Carlin, J. B., Stern, H. S., stern, Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL, 3rd edition.
- Gelman, A., Jakulin, A., Pittau, M. G., Su, Y.-S., et al. (2008). A weakly informative default prior distribution for logistic and other regression models. *The annals of applied statistics*, 2(4):1360–1383.
- Ghosh, M. and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman & Hall, London.
- 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):273–282.
- Ghosh, M. and Rao, J. N. K. (1994). Small Area Estimation: An Appraisal. *Statistical Science*, 9(1):55–93.
- Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(447):359–378.
- Goulard, M. and Voltz, M. (1992). Linear coregionalization model: tools for estimation and choice of cross-variogram matrix. *Mathematical Geology*, 24(3):269–286.
- Greenlees, J. S., Reece, W. S., and Zieschang, K. D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77(378):251–261.
- Grimm, K. A., Foltz, J. L., Blanck, H. M., and Scanlon, K. S. (2012). Household income disparities in fruit and vegetable consumption by state and territory: Results of the 2009 behavioral risk factor surveillance system. *Journal of the Academy of Nutrition and Dietetics*, 112(12):2014–2021.
- Groundwater Ambient Monitoring and Assessment Program (2019). GAMA Data Download. Available from <https://gamagroundwater.waterboards.ca.gov/gama/datadownload> [last accessed June 1, 2019].

- Guhaniyogi, R. and Banerjee, S. (2018). Meta-kriging: Scalable bayesian modeling and inference for massive spatial datasets. *Technometrics*, 60(4):430–444.
- Harter, T., Dzurella, K., Kourakos, G., Hollander, A., Bell, A., Santos, N., Hart, Q., King, A., Quinn, J., Lampinen, G., Liptzin, D., Rosenstock, T., Zhang, M., Pettygrove, G., and Tomich, T. (2017). Nitrogen Fertilizer Loading to Groundwater in the Central Valley, Final Report to the Fertilizer Research Education Program Projects 11-0301 and 15-0454. Technical report, California Department of Food and Agriculture and University of California Davis, Davis, CA.
- Harter, T. and Lund, J. R. (2012). Addressing Nitrate in California’s Drinking Water with a Focus on Tulare Lake Basin and Salinas Valley Groundwater, Report for the State Water Resources Control Board report to the Legislature ? executive summary. Technical report, Center for Watershed Sciences, University of California, Davis, Davis, CA.
- Hartley, H. O. and Sielken Jr., R. L. (1975). A ”Super-Population Viewpoint” for Finite Population Sampling. *Biometrics*, 31(2):411–422.
- Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., Lindgren, F., Nychka, D. W., Sun, F., and Zammit-Mangion, A. (2018). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*.
- Hillier, A., Cole, B. L., Smith, T. E., Yancey, A. K., Williams, J. D., Grier, S. A., and McCarthy, W. J. (2009). Clustering of unhealthy outdoor advertisements around child-serving institutions: a comparison of three cities. *Health & Place*, 15(4):935–945.
- Hodges, J. S. and Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, 64(4):325–334.
- Hoef, J. V. (2002). Sampling and geostatistics for spatial data. *Écoscience*, 9(2):152–161.
- Horvitz, D. G. and Thompson, D. J. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, 47(260):663–685.

- Kalton, G. (2002). Models in the Practice of Survey Sampling (Revisited). *Journal of Official Statistics*, 18(2):129–154.
- Kennedy, B. P., Kawachi, I., Glass, R., and Prothrow-Stith, D. (1998). Income distribution, socioeconomic status, and self rated health in the united states: multilevel analysis. *BMJ*, 317(7163):917–921.
- Koleilat, M., Whaley, S. E., Afifi, A. A., Estrada, L., and Harrison, G. G. (2012). Understanding the relationship between the retail food environment index and early childhood obesity among wic participants in los angeles county using geoda. *Online journal of public health informatics*, 4(1).
- Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.
- Lamichhane, A. P., Warren, J. L., Peterson, M., Rummo, P., and Gordon-Larsen, P. (2015). Spatial-temporal modeling of neighborhood sociodemographic characteristics and food stores. *American journal of epidemiology*, 181(2):137–150.
- Langellier, B. A., Garza, J. R., Prelip, M. L., Glik, D., Brookmeyer, R., and Ortega, A. N. (2013). Corner Store Inventories, Purchases, and Strategies for Intervention: A Review of the Literature. *Californian Journal of Health Promotion*, 11(3):1–13.
- Lawman, H. G., Veur, S. V., Mallya, G., McCoy, T. A., Wojtanowski, A., Colby, L., Sanders, T. A., Lent, M. R., Sandoval, B. A., Sherman, S., Wylie-Rosett, J., and Foster, G. D. (2015). Changes in quantity, spending, and nutritional characteristics of adult, adolescent and child urban corner store purchases after an environmental intervention. *Preventive Medicine*, 74:81–85.
- Lee, A., Szpiro, A., Kim, S. Y., and Sheppard, L. (2015). Impact of preferential sampling on exposure prediction and health effect inference in the context of air pollution epidemiology. *Environmetrics*, 26(4):255–267.
- Little, R. J. (2004). To Model or Not To Model? Competing Modes of Inference for Finite Population Sampling. *Journal of the American Statistical Association*, 99(466):546–556.

- Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Little/Statistical Analysis with Missing Data. John Wiley & Sons, Hoboken, NJ, 2nd edition.
- Lynch, J. W., Kaplan, G. A., Pamuk, E. R., Cohen, R. D., Heck, K. E., Balfour, J. L., and Yen, I. H. (1998). Income inequality and mortality in metropolitan areas of the united states. *American Journal of Public Health*, 88(7):1074–1080. PMID: 9663157.
- Malec, D. and Sedransk, J. (1985). Bayesian Inference for Finite Population Parameters in Multi-stage Cluster Sampling. *Journal of the American Statistical Association*, 80(392):897–902.
- Neelon, S. E. B., Burgoine, T., Gallis, J. A., and Monsivais, P. (2017). Spatial analysis of food insecurity and obesity by area-level deprivation in children in early years settings in england. *Spatial and spatio-temporal epidemiology*, 23:1–9.
- Neyman, J. (1934). On the Two Different Aspects of the Representative Method. *Journal of the Royal Statistical Society*, 97(4):558–625.
- Ortega, A. N., Albert, S. L., Chan-Golston, A. M., Langellier, B. A., Glik, D. C., Belin, T. R., Rosa Elena Garcia, R. B., Sharif, M. Z., and Prelip, M. L. (2016). Substantial improvements not seen in health behaviors following corner store conversions in two Latino food swamps. *BMC Public Health*, 16(389).
- Ortega, A. N., Albert, S. L., Sharif, M. Z., Langellier, B. A., Garcia, R. E., Glik, D. C., Brookmeyer, R., Chan-Golston, A. M., Friedlander, S., and Prelip, M. L. (2015). A Multi-level, Community-Engaged Corner Store Intervention in East Los Angeles and Boyle Heights. *Journal of Community Health*, 40(347-356).
- Paek, H.-J., Oh, H. J., Jung, Y., Thompson, T., Alaimo, K., Risley, J., and Mayfield, K. (2014). Assessment of a healthy corner store program (fit store) in low-income, urban, and ethnically diverse neighborhoods in michigan. *Family & community health*, 37(1):86–99.
- Pati, D., Reich, B. J., and Dunson, D. B. (2011). Bayesian geostatistical modelling with informative sampling locations. *Biometrika*, 98(1):35–48.

- Pickett, K. E., Kelly, S., Brunner, E., Lobstein, T., and Wilkinson, R. G. (2005). Wider income gaps, wider waistbands? an ecological study of obesity and income inequality. *Journal of Epidemiology & Community Health*, 59(8):670–674.
- Plummer, M. (2017). *JAGS Version 4.3.0 user manual*. International Agency for Research on Cancer, Lyon, France.
- R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rao, J. N. K. (2003). *Small Area Estimation*. John Wiley & Sons, Hoboken, NJ.
- Ribar, D. C. and Hamrick, K. S. (2003). Dynamics of poverty and food sufficiency. page 30.
- Riphahn, R. T. and Serfling, O. (2005). Item non-response on income and wealth questions. *Empirical Economics*, 30(2):521–538.
- Ripley, B. D. (2004). *Spatial Statistics*. John Wiley & Sons, Hoboken, NJ.
- Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, New York, NY, 2nd edition.
- Rose, D. (1999). Economic Determinants and Dietary Consequences of Food Insecurity in the United States. *The Journal of Nutrition*, 129(2):517–520S.
- Rose, D., Bodor, J. N., Swalm, C. M., Rice, J. C., Farley, T. A., and Hutchinson, P. L. (2009). Deserts in new orleans? illustrations of urban food access and implications for policy. *Ann Arbor, MI: University of Michigan National Poverty Center/USDA Economic Research Service Research*.
- Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57(2):377–387.
- Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63(3):581–592.

- Schenker, N., Raghunathan, T. E., Chiu, P.-L., Makuc, D. M., Zhang, G., and Cohen, A. J. (2006). Multiple imputation of missing income data in the national health interview survey. *Journal of the American Statistical Association*, 101(475):924–933.
- Scott, A. and Smith, T. M. F. (1969). Estimation in Multi-Stage Surveys. *Journal of the American Statistical Association*, 64(327):830–840.
- Sharkey, J. R., Horel, S., and Dean, W. R. (2010). Neighborhood deprivation, vehicle ownership, and potential spatial access to a variety of fruits and vegetables in a large rural area in texas. *International Journal of Health Geographics*, 9(1):26.
- Sisiopiku, V. P. and Barbour, N. (2014). Use of gis spatial analysis to identify food deserts in the state of alabama. *Athens Journal of Health*, 1(2):91–103.
- Smiley, M. J., Roux, A. V. D., Brines, S. J., Brown, D. G., Evenson, K. R., and Rodriguez, D. A. (2010). A spatial analysis of health-related resources in three diverse metropolitan areas. *Health & place*, 16(5):885–892.
- Smith, D. M., Cummins, S., Taylor, M., Dawson, J., Marshall, D., Sparks, L., and Anderson, A. S. (2010). Neighbourhood food environment and area deprivation: spatial accessibility to grocery stores selling fresh fruit and vegetables in urban and rural settings. *International journal of epidemiology*, 39(1):277–284.
- Song, H.-J., Gittelsohn, J., Kim, M., Suratkar, S., Sharma, S., and Anliker, J. (2009). A corner store intervention in a low-income urban community is associated with increased availability and sales of some healthy foods. *Public Health Nutrition*, 12(11):2060–2067.
- Stan Development Team and others (2018). Rstan: the r interface to stan. r package version 2.17.3.
- Sugden, R. A. and Smith, T. M. F. (1984). Ignorable and Informative Designs in Survey Sampling Inference. *Biometrika*, 71(3):495–506.
- Thorndike, A. N., Bright, O.-J. M., Dimond, M. A., Fishman, R., and Levy, D. E. (2017). Choice architecture to promote fruit and vegetable purchases by families participating in the special

- supplemental program for women, infants, and children (wic): randomized corner store pilot study. *Public Health Nutrition*, 20(7):1297-1305.
- Valliant, R., Dorfman, A. H., and Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. John Wiley & Sons, Hoboken, NJ.
- Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432.
- Walker, K. (2018). *tigris: Load Census TIGER/Line Shapefiles*. R package version 0.7.
- Watson, N. and Starick, R. (2011). Evaluation of alternative income imputation methods for a longitudinal survey. *Journal of Official Statistics*, 27(4):693.
- Wildman, J. (2003). Income related inequalities in mental health in great britain: analysing the causes of health inequality over time. *Journal of Health Economics*, 22(2):295 – 312.
- Yan, T., Curtin, R., and Jans, M. (2010). Trends in income nonresponse over two decades. *Journal of Official Statistics*, 26(1):145.
- Zangeneh, S. Z. and Little, R. J. A. (2015). Bayesian inference for the finite population total from heteroscedastic probability proportional to size sample. *Journal of Survey Statistics and Methodology*, 3:162–192.