CHAPTER III

Accounting for Survey Design in Bayesian Disaggregation of Survey-based Estimates of Proportions

3.1 Introduction

Neighborhoods are dynamic entities, whose demographic and economic compositions can change over time. Despite resurgent research on the matter (see Zuk et al. (2018) for a thorough literature review), debates exist as for how neighborhoods' changes occur in time (Schelling, 1971; Card et al., 2008; Ellen and O'Regan, 2011; Reibel and Regelson, 2013; Wei and Knox, 2014; Zuk et al., 2018) and/or space (Lees et al., 2008; Heidkamp and Lucas, 2006). The inability to elucidate this issue can be partially attributed to the difficulty of procuring and disseminating data that characterize neighborhoods evolution, especially over large spatial domains.

Understanding neighborhood changes and their effects on health requires data on neighborhood characteristics, which would ideally be available at as fine spatial and temporal resolution as possible. Privacy concerns and the previously mentioned difficulty of procuring large amounts of neighborhood data often lead to the aggregation of estimates of neighborhood characteristics over space and/or

time. This aggregation can result in estimates whose spatial and/or temporal resolution is misaligned with the target spatial and/or temporal resolution of a research study. Multi-year estimates derived from the American Community Survey (ACS), a multi-year national survey administered by the United States Census Bureau, are a prime example of a massive, publicly available dataset whose levels of spatio-temporal aggregation may diminish its utility for researchers.

The ACS is administered annually to a sample of approximately 3.5 million Americans, including those residing in unincorporated territories (U.S. Census Bureau, 2008). Every year, estimates derived from the ACS are released to the public, providing up-to-date, timely, and accurate population and housing information to data-users. Currently, estimates for small municipal sub-divisions, such as census tracts, are aggregated over 5-year time periods, whereas estimates corresponding to 1-year time periods are only available for municipal sub-divisions with populations greater than 65,000. We provide additional details on the ACS and the estimates derived from it in Section 3.2.

Researchers conducting longitudinal studies of population health may wish to incorporate ACS estimates in order to account for social determinants of health in their epidemiological analysis, in which case they would be faced with a choice: (a) utilize the estimates with fine spatial resolution whose 5-year temporal resolution is unlikely to conform to other data sources or properly characterize yearly changes; or (b) utilize 1-year estimates, whose aggregation over large areal units diminishes their ability to characterize neighborhood surroundings in a meaningful way. To circumvent this issue, in this chapter we present a modeling framework that

allows one to disaggregate spatially and temporally estimates of proportions derived from sampling surveys. While other approaches to handle the spatial and/or spatio-temporal change of support problem for ACS estimates have already been proposed in the literature, see for example Bradley et al. (2015), Bradley et al. (2016b), Bradley et al. (2016a), Savitsky (2016), or Simpson et al. (2019), the distinguishing feature of our model is that we explicitly account for the survey design effect, thus merging survey methodology methods with spatial statistical modeling frameworks.

In application to the ACS multi-year estimates, our model yields estimates of neighborhood socioeconomic and demographic indicators that are more precise and have finer spatio-temporal support than estimates available through the ACS. We believe that this will render them of great utility to researchers who wish to incorporate socioeconomic indicators characterizing individuals' neighborhood surroundings into health studies.

Our modeling framework belongs to the class of Bayesian hierarchical models addressing the spatio-temporal change of support problem (COSP; see Banerjee et al. (2004)), that is, the problem of performing inference about a spatial or spatio-temporal process at a resolution (or support) that differs from that of the data. Additionally, our model incorporates aspects of survey methodology in order to account for the survey design, a relatively novel feature for spatio-temporal models (see Mercer et al. (2014) and Chen et al. (2014) for examples), especially those addressing the change of support problem. Readers interested in learning about COSP may refer, among others, to: (i) Gotway and Young (2002) and Banerjee

et al. (2004) for a review of the spatial COSP; (ii) Gelfand et al. (2001) for a spatio-temporal statistical model that addresses the spatio-temporal change of support; and (iii) Bradley et al. (2015) and Bradley et al. (2016a) for a model addressing the spatio-temporal change of support problem relative to ACS data.

At the basis of our model formulation is the assumption that the estimate at the areal level can be thought as an aggregation of an underlying, point-referenced spatio-temporal process. Such modeling specification allows us to derive estimates over spatio-temporal resolutions that are equal or larger than the smallest spatial and temporal resolution for which we have data. In our application, we focus on generating estimates at the 1-year time scale and census tract spatial resolution.

The remainder of this chapter is organized as follows. Section 3.2 provides more detailed background information on the ACS. Section 3.3 describes our modeling framework to disaggregate spatially and temporally ACS estimates of proportions while accounting for the survey design effect. Section 3.4 illustrates the capabilities of our model through simulation experiments, while Section 3.5 presents results from our model when applied, respectively, to: ACS estimates of the proportion of families living in poverty in Michigan from 2006 through 2016 (Section 3.5.1) and proportion of residents who identify as Black/African-American living in Michigan from 2006 to 2016 (Section 3.5.4). Finally, the chapter concludes with a discussion in Section 3.6.

3.2 The American Community Survey

In this section, we provide a brief overview of the American Community Survey (ACS), its design, and the methodology used to derive ACS estimates. Interested readers may refer to the ACS User Guide (U.S. Census Bureau, 2008) for background information on the ACS, and the ACS Design and Methodology Report (U.S. Census Bureau, 2014) for extensive details on the survey design and estimation procedures.

The American Community Survey (ACS) is an ongoing survey conducted by the U.S. Census Bureau (U.S. Census Bureau, 2008). It was first implemented after the 2000 Census with the intention of replacing the Census long form. The ACS samples approximately 3.5 million households annually, collecting data pertaining to social, housing, economic, and other community characteristics. In contrast to the Census long form, for which data were gathered every 10 years, the ACS surveys are administered continuously, allowing for the timely dissemination of up-to-date community information. Furthermore, the continuous collection of data allows the ACS estimates to be statistically representative of the time period during which the surveys were administered. This is yet another contrast to the Census long form, whose information is only reflective of the brief period every ten years during which it is conducted.

Surveys are administered according to a careful design, and the Census Bureau performs extensive follow-up over the phone, online, and in person. Samples are drawn independently for each county in the United States, with sampling rates

being inversely proportional to the population of the sampling unit and the anticipated probability of response. Estimates derived from the ACS are computed using surveys administered within a given geographic area and time period. The survey weights used for the derivation account for various forms of bias, including sampling and non-response, and are calibrated at various stages so that the ACS estimates conform to the Census Bureau's Population Estimates Program as well as administrative records. As with the decennial Census estimates, ACS estimates are provided with a margin of error, computed through successive differences replication U.S. Census Bureau (2014).

For various reasons, including statistical accuracy and precision, as well as privacy concerns, ACS estimates are released with varying spatial and temporal resolution. Specifically, estimates for small municipal subdivisions, such as census tracts, are aggregated and provided in the form of averages over a 5-year time period, whereas yearly estimates are provided for administrative regions that have over 65,000 inhabitants. Certain counties meet this criterion, however a number of counties in the US have populations less than 65,000 and would therefore be excluded from a dataset with a 1-year temporal resolution. An alternative to using county-level estimates would be to use estimates at the Public Use Microdata Areas (PUMA) level. PUMAs are collections of contiguous counties and/or census tracts whose total population exceeds 100,000 people. Because of their definition, all PUMAs are provided with 1-year estimates.

3.3 Modeling Approach

Our model to disaggregate ACS estimates of proportions while accounting for the survey design effect leverages the information contained in ACS estimates at two spatio-temporal resolutions: the 5-year estimates at census tract level, and the 1-year estimates at the PUMA level. To keep a general notation that is flexible enough to accommodate the different spatio-temporal resolutions, we follow Bradley et al. (2015) and we denote with $z_t^{(l)}(A)$ the estimate of a proportion corresponding to areal unit A for the l-year time period ending in time t, while we use $\tau_t^{2(l)}(A)$ to denote the design based variance of $z_t^{(l)}(A)$. Thus, $z_t^{(5)}(A_{ig})$ indicates the ACS estimate for the 5-year time period ending at year t for census tract g, $g = 1, \ldots, G_i$, within PUMA i, $i = 1, \ldots, N$, while $z_t^{(1)}(A_i)$ refers to the 1-year ACS estimate for PUMA i at year t. These estimates admit sampling-based variance $\tau_t^{2(5)}(A_{ig})$ and $\tau_t^{2(1)}(A_i)$, respectively, which are quantified through the margins of error provided in the ACS data.

3.3.1 Modeling survey-based estimates of areal proportions accounting for the design effect

Following Bradley et al. (2015), we assume that the survey-based estimate of the proportion corresponding to areal unit A over the l-unit time period ($l = 1$ or 5), ending at time t, $z_t^{(l)}(A)$, is related to the true proportion, $\pi_t^{(l)}(A)$, through some distribution function. Building upon the work of Korn and Graubard (1998), and later Mercer et al. (2014) and Chen et al. (2014), who proposed models for spatial smoothing of estimates of proportions from complex surveys, our working likelihood is based on a random variable $q_t^{*(l)}(A)$ that we construct from the ACS

estimate $z_t^{(l)}(A)$ and the effective sample size $m_t^{*(l)}(A)$. The latter represents the sample size of a simple random sample (SRS) that will yield an estimator for the proportion with a variance that matches the design-based variance of the ACS estimate. We derive the effective sample size using the notion of design effect introduced by Kish (1995), who calls a survey's design effect the ratio of the variance of an estimator under SRS to the sampling-based variance of survey-based estimator. By setting the survey's design effect equal to 1 and solving the equation for the SRS sample size $m_t^{(l)}(A)$, we obtain the sample size corresponding to the ACS design-based variance. We call this sample size the *effective sample size* (ESS). Including the ESS in the distribution function that relates a function of the estimate $z_t^{(l)}(A)$ to the true proportion $\pi_t^{(l)}(A)$ allows us to account for the survey's design effect in our modeling framework.

More specifically, in the case of $z_t^{(l)}(A)$ - the ACS estimate of a proportion corresponding to areal unit A for the l-unit time period ending at time t - the estimated variance of an estimator under simple random sampling would be of the form $\frac{z_t^{(l)}(A)(1-z_t^{(l)}(A))}{m_t^{(l)}(A)}$, thus yielding the following effective sample size, $m_t^{*(l)}(A)$:

$$m_t^{*(l)}(A) = \left[\frac{z_t^{(l)}(A)(1 - z_t^{(l)}(A))}{\tau_t^{2(l)}(A)} \right] \tag{3.1}$$

with $[\cdot]$ denoting rounding to the nearest integer and with rounding introduced to ensure that the effective sample size is an integer.

Having derived the effective sample size $m_t^{*(l)}(A)$ for areal unit A through the function in (3.1), which involves both the ACS estimate and its sampling-based variance, we use the ACS estimate of a proportion for areal unit A in the l-year

time period ending in year t to derive the *effective number of cases*, $q_t^{*(l)}(A)$, for the same areal unit and for the same time period, is:

$$q_t^{*(l)}(A) := \left[m_t^{*(l)}(A) \cdot z_t^{(l)}(A) \right] \tag{3.2}$$

again rounded to the nearest integer to ensure that also the effective number of cases is an integer.

Using now the effective number of cases in our likelihood, our Bayesian hierarchical model specifies at the first stage a Binomial likelihood for $q_t^{*(l)}(A)$ with number of trials equal to $m_t^{*(l)}(A)$ and success probability equal to the true proportion, $\pi_t^{(l)}(A)$, our parameter of interest, e.g.:

$$q_t^{*(l)}(A) | \, \pi_t^{(l)}(A) \sim \text{Binomial} \left(m_t^{*(l)}(A), \pi_t^{(l)}(A) \right) \tag{3.3}$$

As we fit our model to 5-year and 1-year ACS estimates of areal proportions, from (3.1) and (3.2) it follows that we work with the following effective sample sizes and effective number of cases: for $g = 1, \ldots, G_i$; $i = 1, \ldots, N$

$$\begin{aligned}
m_t^{*(5)}(A_{ig}) &= \left[\frac{z_t^{(5)}(A_{ig})(1 - z_t^{(5)}(A_{ig}))}{\tau_t^{2(5)}(A_{ig})} \right] \\
m_t^{*(1)}(A_i) &= \left[\frac{z_t^{(1)}(A_i)(1 - z_t^{(1)}(A_i))}{\tau_t^{2(1)}(A_i)} \right] \\
q_t^{(5)}(A_{ig}) &:= \left[m_t^{*(5)}(A_{ig}) \cdot z_t^{(5)}(A_{ig}) \right] \\
q_t^{(1)}(A_i) &:= \left[m_t^{*(1)}(A_{ig}) \cdot z_t^{(1)}(A_{ig}) \right]
\end{aligned}$$

According to (3.3), our data likelihood is then made of the following two components

$$\begin{aligned}
q_t^{*(5)}(A_{ig}) | \, \pi_t^{(5)}(A_{ig}) &\overset{ind}{\sim} \text{Binomial} \left(m_t^{*(5)}(A_{ig}), \pi_t^{(5)}(A_{ig}) \right) \\
q_t^{*(1)}(A_i) | \, \pi_t^{(1)}(A_i) &\overset{ind}{\sim} \text{Binomial} \left(m_t^{*(1)}(A_i), \pi_t^{(1)}(A_i) \right)
\end{aligned} \tag{3.4}$$

with $g = 1, \ldots, G_i$ and $i = 1, \ldots, I$.

The assumption of conditional independence on the effective number of cases in (3.4) follows in the tradition of spatial generalized linear models (see Diggle et al. (1998) for details). However, marginally, this modeling framework does impose dependence across the effective number of cases. Appendix B.2 presents derivations of the marginal covariance of the effective number of cases under our full modeling framework, the remainder of which is stated hereafter.

With the goal of disaggregating spatially and temporally the ACS estimates and deriving estimates of the true proportion at the census tract resolution for each year, we link $\pi_t^{(5)}(A_{ig})$ and $\pi_t^{(1)}(A_i)$ to $\pi_t^{(1)}(A_{ig})$, $g = 1, \ldots, G_i; i = 1, \ldots, N$, the true proportions at our desired spatial and temporal resolution:

$$\begin{aligned} \pi_t^{(5)}(A_{ig}) &= \frac{1}{5} \sum_{k=t-4}^{t} \pi_k^{(1)}(A_{ig}) \\ \pi_t^{(1)}(A_i) &= \frac{1}{N_t(A_i)} \sum_{h=1}^{G_i} N_t(A_{ih}) \pi_t^{(1)}(A_{ih}) \end{aligned} \tag{3.5}$$

where $N_t(A)$ typically denotes the number of households in areal unit A at time t. Alternatively, it may denote the population size, number of families, etc., in areal unit A at time t.

3.3.2 Handling the Change of Support Problem (COSP)

We proceed by describing how we handle the temporal and spatial disaggregation of the ACS estimates.

The underlying assumption of our change of support model is that a random variable relative to an areal unit and a time period can be expressed as the aggrega-

tion over the areal unit and the time period of a point-referenced spatio-temporal process, continuous in space and discrete in time.

More formally, we link $\pi_t^{(l)}(A)$, the true proportion over areal unit $A \subset \mathcal{S}$ (with $\mathcal{S}$ denoting the spatial domain) and over the l-unit time period ending at time t, to an underlying spatio-temporal process $\zeta_k(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$ through the probit link function, $\Phi^{-1}(\cdot)$, that is:

$$\begin{aligned} \Phi^{-1}\left(\pi_t^{(l)}(A)\right) &= \frac{1}{l}\sum_{k=t-l+1}^{t}\frac{1}{|A|}\int_{\mathbf{s}\in A}\zeta_k(\mathbf{s})ds + \xi(C_A) \\ \xi(C_A) &\overset{iid}{\sim} N(0\tau_C^2) \end{aligned} \tag{3.6}$$

In (3.6), $\xi(C_A)$ is a random effect defined at the same areal unit level as the clustering units of the sampling survey. Here, C_A denotes the cluster that contains areal unit A. We introduce the cluster-level random effect, $\xi(C_A)$, in order to construct estimates whose dependence reflects the survey sampling design.

In our first application, discussed in 3.5.1, we are interested in estimating the true proportion of families in poverty in any areal unit A. In this case, $\zeta_k(\mathbf{s})$, $\mathbf{s} \in A$, represents a function of the likelihood that a family living at location $\mathbf{s} \in A$ is in poverty in year $k, k = t-l+1, \ldots, t$. Decomposing the point-referenced spatio-temporal process $\zeta_k(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, into a large-scale spatio-temporal trend, $\mu_k(\mathbf{s})$, representing the mean of the process, and a spatio-temporal random effect, $w_k(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, (3.6) becomes:

$$\Phi^{-1}\left(\pi_t^{(l)}(A)\right) = \frac{1}{l}\sum_{k=t-l+1}^{t}\frac{1}{|A|}\int_{\mathbf{s}\in A}\left(\mu_k(\mathbf{s}) + w_k(\mathbf{s})\right)d\mathbf{s} + \xi(C_A) \qquad \forall k = t-l+1, \ldots, t \tag{3.7}$$

In turn, we model the spatio-temporal random effect, $w_k(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, as a Gaussian spatio-temporal process with a separable space-time covariance function with an

AR(1) structure in time and spatial dependence encoded through the covariance function $C(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$, $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$.

Anticipating a large number of areal units, for computational convenience, we approximate the spatio-temporal process $w_k(\mathbf{s}), \mathbf{s} \in \mathcal{S}$, with a linear combination of spatial basis functions with appropriate spatio-temporal basis function weights. For this purpose, we extend the basis function approximation method of Katzfuss (2017), the Multi-Resolution Approximation (MRA), to the spatio-temporal setting.

In the MRA, the spatial basis functions are defined through repeated implementations of the Predictive Process approximation (Banerjee et al., 2008) over recursive partitions (indexed by j) of the spatial domain at various resolutions (indexed by m). In our extension of the MRA to the spatio-temporal setting, which we call the Spatio-temporal Multi-Resolution Approximation (ST-MRA), we maintain the same idea, however we provide the basis function weights with a dynamic temporal structure so to obtain the AR(1) structure in time of the original spatio-temporal random effect $w_k(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$.

3.3.3 The Spatio-temporal Multi-resolution Approximation (ST-MRA)

For a detailed description of the MRA, interested readers may refer to Katzfuss (2017). Here, we provide details on the ST-MRA.

Let $w_t(\mathbf{s}), \mathbf{s} \in \mathcal{S}, t = 1, \ldots, T$, denote a mean-zero spatio-temporal Gaussian processes defined on a spatial domain $\mathcal{S}$, equipped with a separable space-time covariance function, with a first-order autoregressive structure in time and spatial

covariance function $C(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$, $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$. As in the MRA, suppose that we first introduce a set of r knots on the spatial domain $\mathcal{S}$ (level 0), and then at each level m ($m = 0, \ldots, M$), we recursively partition the spatial domain $\mathcal{S}$ in J^m non-overlapping subregions in which we introduce r knots. Let $S^*_{m,j}$ denote the set of r knots defined on partition j of level m. Then, from the MRA approach, we know that by defining the basis functions $\mathbf{b}_{m,j}(\mathbf{s})$, for $j = 1, \ldots, J^m; m = 0, \ldots, M$ recursively as:

$$
\begin{aligned}
v_0(\mathbf{s}_1, \mathbf{s}_2) &= C(\mathbf{s}_1, \mathbf{s}_2; \boldsymbol{\theta}) \\
v_{m+1}(\mathbf{s}_1, \mathbf{s}_2) &= 0 \qquad \text{if } \mathbf{s}_1 \text{ and } \mathbf{s}_2 \text{ are in different regions at resolution } m \\
v_{m+1}(\mathbf{s}_1, \mathbf{s}_2) &= v_m(\mathbf{s}_1, \mathbf{s}_2) - \mathbf{b}_{m,j}(\mathbf{s}_1)' \mathbf{K}_{m,j} \mathbf{b}_{m,j}(\mathbf{s}_2) \qquad \text{otherwise} \\
\mathbf{K}^{-1}_{m,j} &= v_m(S^*_{m,j}, S^*_{m,j}) \\
\mathbf{b}_{m,j}(\mathbf{s}) &= v_m(\mathbf{s}, S^*_{m,j})
\end{aligned}
$$

and by providing the basis function weights $\boldsymbol{\eta}_{m,j}$ with the following distribution $\boldsymbol{\eta}_{m,j} \sim N_r(\mathbf{0}, \mathbf{K}_{m,j})$, the linear combination $\sum_{m=0}^{M} \sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s}) \boldsymbol{\eta}_{m,j}$ yields an M-level approximation to a spatial process with covariance function $C(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$.

Since here we are working with a spatio-temporal process $w_t(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, $t = 1, \ldots, T$, for which $\text{Corr}\left(w_t(\mathbf{s}), w_{t'}(\mathbf{s})\right) = \alpha^{|t-t'|}$, $\forall \mathbf{s} \in \mathcal{S}$, to capture the temporal dependence structure among the $w_t(\mathbf{s})$'s, we let the basis function weights $\boldsymbol{\eta}_{t,m,j}$ vary in time and we provide them with a stationary first-order autoregressive structure (Gelfand et al., 2005). Hence, at time $t = 1$ we assume $\boldsymbol{\eta}_{1,m,j} \sim N_r(\mathbf{0}, \mathbf{K}_{m,j})$

while for $t = 2, \ldots, T$:

$$\begin{aligned} \boldsymbol{\eta}_{t,m,j} | \boldsymbol{\eta}_{t-1,m,j}, \boldsymbol{\eta}_{t-2,m,j}, \ldots, \boldsymbol{\eta}_{1,m,j} &\sim N_r(\alpha \boldsymbol{\eta}_{t-1,m,j}, \mathbf{U}_{m,j}) \\ \mathbf{U}_{m,j} &= (1-\alpha^2)\mathbf{K}_{m,j} \end{aligned} \tag{3.8}$$

Then, we call

$$w_{t,M}(\mathbf{s}) := \sum_{m=0}^{M} \sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s}) \boldsymbol{\eta}_{t,m,j} \tag{3.9}$$

the M-level ST-MRA approximation of the separable, spatio-temporal process $w_t(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, $t = 1, \ldots, T$, with AR(1) dependence in time and spatial covariance function $C(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$. Appendix B shows that the above expression does indeed provide an approximation of said spatio-temporal dependence structure.

3.3.4 The Bayesian spatio-temporal disaggregation model

Returning to our spatio-temporal change of support model in (3.7), we assume that the true proportion $\pi_t^{(l)}(A)$ corresponding to areal unit $A \subset \mathcal{S}$ during the l-unit time period ending at time t, can be represented as:

$$\begin{aligned} \Phi^{-1}\left(\pi_t^{(l)}(A)\right) &= \frac{1}{l} \sum_{k=t-l+1}^{t} \frac{1}{|A|} \int_{\mathbf{s} \in A} \left(\mu_k(\mathbf{s}) + w_k(\mathbf{s})\right) d\mathbf{s} + \xi(C_A) \\ &\approx \frac{1}{l} \sum_{k=t-l+1}^{t} \frac{1}{|A|} \int_{\mathbf{s} \in A} \left(\mu_k(\mathbf{s}) + w_{k,M}(\mathbf{s})\right) d\mathbf{s} + \xi(C_A) \\ &\approx \frac{1}{l} \sum_{k=t-l+1}^{t} \frac{1}{|A|} \int_{\mathbf{s} \in A} \left(\mu_k(\mathbf{s}) + \sum_{m=0}^{M} \sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s}) \boldsymbol{\eta}_{k,m,j}\right) d\mathbf{s} + \xi(C_A), \end{aligned} \tag{3.10}$$

with the last RHS obtained by replacing the spatio-temporal random effect $w_k(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, with its M-level ST-MRA approximation $w_{k,M}(\mathbf{s})$ defined as in (3.9).

Continuing with our model specification tailored specifically for the ACS data, and defining $\xi(C_A)$ at the same areal unit level of the clusters in the ACS, we will introduce notation for the county containing census tract A_{ig}. Let $C_{A_{ig}}$ denote the county containing census tract A_{ig}. Using (3.10) in (3.5), we obtain our complete Bayesian hierarchical model specification:

$$
\begin{aligned}
q_t^{*(5)}(A_{ig})|\, \pi_t^{(5)}(A_{ig}) &\overset{ind}{\sim} \text{Binomial}\left(m_t^{*(5)}(A_{ig}), \pi_t^{(5)}(A_{ig})\right) \\
q_t^{*(1)}(A_i)|\, \pi_t^{(1)}(A_i) &\overset{ind}{\sim} \text{Binomial}\left(m_t^{*(1)}(A_i), \pi_t^{(1)}(A_i)\right) \\
\pi_t^{(5)}(A_{ig}) &= \frac{1}{5}\sum_{k=t-4}^{t} \pi_k^{(1)}(A_{ig}) \qquad (3.11) \\
\pi_t^{(1)}(A_i) &= \frac{1}{N_t(A_i)}\sum_{h=1}^{G_i} N_t(A_{ih})\pi_t^{(1)}(A_{ih}) \\
\Phi^{-1}\left(\pi_t^{(1)}(A_{ig})\right) &\approx \frac{1}{|A_{ig}|}\int_{\mathbf{s}\in A_{ig}} \left(\mu_t(\mathbf{s}) + \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s})\boldsymbol{\eta}_{t,m,j}\right) d\mathbf{s} + \xi(C_{A_{ig}}) \\
\xi(C_{A_{ig}}) &\overset{iid}{\sim} N(0, \tau_C^2)
\end{aligned}
$$

The county-level random effect imposes that ACS estimates for census tracts within the same county exhibit greater dependence with one another than ACS estimates for census tracts in different counties, even when the distances between those tracts are the same. We speculate this phenomenon is realized in practice when factors such as the sampling procedure or response rate within a county-wide sampling frame systematically affect ACS estimates corresponding to most or all of the census tracts within that county.

Next, in order to take the integral over space of the underlying spatio-temporal process in (3.11), we define $\mathbf{b}_{m,j}(A_{ig})$ to be the integral of the basis functions

$\mathbf{b}_{m,j}(\mathbf{s})$ as $\mathbf{s}$ varies in areal unit A_{ig}. We will define $\mu_t(A_{ig})$ similarly.

(3.12)

$$\begin{aligned} \Phi^{-1}\left(\pi_t^{(1)}(A_{ig})\right) &\approx \mu_t(A_{ig}) + \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{t,m,j} + \xi(C_{A_{ig}}) + \epsilon_t(A_{ig}) \\ \epsilon_t(A_{ig}) &\overset{iid}{\sim} N(0,\tau^2), \end{aligned}$$

with $g = 1,\ldots,G_i$; $i = 1,\ldots,N$, $t = 0,\ldots,T$.

The term $\epsilon_t(A_{ig})$ in (3.12) is introduced to account for errors due to the aggregation from a point-referenced spatial support to an areal one (A_{ig}), as well as any error that occurs as a result of the multi-resolution space-time approximation.

3.3.5 Prior distributions

Our spatio-temporal model is provided in a Bayesian framework, hence a complete model specification includes priors for all the remaining model parameters. We use an Inverse-Gamma as prior distribution for the aggregation-error variance parameter τ^2 in (3.7) as well as for the variance of the county-level random effects, τ_C^2. On the other hand, the large-scale spatio-temporal mean functions, $\mu_t(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, $t = 1,2,\ldots,T$, in (3.7) are assumed to be constant in space and are thus simplified to μ_t. Such a modeling choice allows the spatio-temporal random effects $w_t(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, $t = 1,2,\ldots,T$, to account for all the spatio-temporal variation in the ACS estimates. We have investigated relaxing this assumption and allowing the μ_t's to vary in space, but results did not change substantially. Hence, we opt for a more parsimonious model, setting $\mu_t(\mathbf{s}) \equiv \mu_t$, for each $t = 1,2,\ldots,T$, and assuming they are independent in time a priori. We provide each μ_t, $t = 1,2,\ldots,T$ with an improper, flat prior, e.g. $p(\mu_t) \propto 1$, $\forall t$.

Moving onto priors for the ST-MRA terms: we place a Uniform ($[0, 1]$) prior on the autoregressive parameter α of the basis function weights (see (3.8)), while we define the ST-MRA basis functions by taking the spatial covariance function $C(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$, $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$, to be a stationary Matèrn covariance function, that is: for $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$

$$C(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \left(\frac{||\mathbf{s} - \mathbf{s}'||}{\phi} \right)^{\nu} \mathcal{K}_{\nu} \left(\frac{||\mathbf{s} - \mathbf{s}'||}{\phi} \right) \tag{3.13}$$

with $\mathcal{K}_{\nu}(\cdot)$ modified Bessel function of the second kind.

We specify a non-informative Gamma prior on the marginal variance parameter σ^2, while we place Gamma$(1, 1)$ and Uniform$((0, 2))$ priors on the range parameter ϕ and the smoothness parameter ν, respectively.

3.3.6 Computation

We fit our Bayesian hierarchical model using a Markov Chain Monte Carlo (MCMC) algorithm, with Gibbs sampling and Metropolis-Hastings steps. We assess convergence both visually by inspecting trace plots, and numerically using Geweke's diagnostic for Markov chains (Geweke, 1992). We run the MCMC algorithm for a number of iterations large enough that the effective sample size post burn-in for each model parameter exceeds 1,000. Tuning of the proposal distributions used in the Metropolis-Hastings steps is performed during burn-in so that desirable acceptance rates (between 15-40%) are achieved (see Roberts et al. (1997) for details on optimal acceptance rates for the Metropolis-Hastings algorithm).

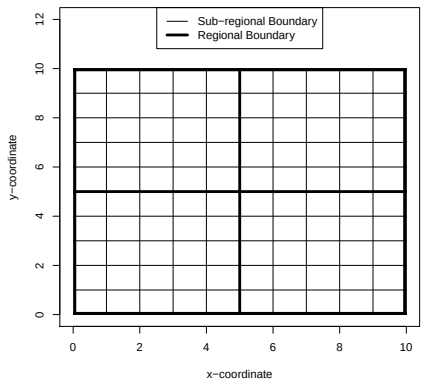


Figure 3.1: Areal units utilized for simulations.

3.4 Simulation Study

3.4.1 Simulation study data generation

As a preliminary evaluation of our modeling framework, we perform a simulation study on 30 datasets consisting of 100 randomly generated proportions each corresponding to a sub-region created on a 10×10 square grid. Sub-regions are nested within one of four regions (see Figure 3.1). This geographical configuration is analogous to that of census tracts and PUMAs respectively, and we will use the same notation adopted for the ACS data. Thus, $\pi_t(A_{ig})$ denotes the true proportion at time t for sub-region g, with $g = 1, \ldots, 25$, nested within region i, with $i = 1, \ldots, 4$. For each subregion, we generate the true proportions at 10 time points, indexed by $t = 1, \ldots, 10$.

We use 4 different data generation mechanisms to assess the performance of our model in settings where the data are not generated according to our model. In each case we assume that in each subregion A_{ig}, there is a latent covariate

$x(A_{ig})$ distributed according to a standard normal distribution, which drives the true proportion $\pi_t(A_{ig})$, $g = 1, 2, \ldots, 25$, $i = 1, \ldots, 4$. In a first simulation setting, for each subregion A_{ig} and at each time point t, the true proportion $\pi_t(A_{ig})$ is obtained by applying the inverse logistic function, simply referred to as *expit*, to the latent covariate $x(A_{ig})$ and adding to it white noise to allow for temporal and spatial variability in the true proportions. Although the true proportions will not be the same across space and time, this data generation mechanism will yield true proportions that are independent in space and correlated in time.

To induce spatial correlation in the true proportions $\pi_t(A_{ig})$, $g = 1, 2, \ldots, 25$; $i = 1, \ldots, 4$, in the second simulation setting we introduce a point-referenced spatial process, $w(\mathbf{s})$, equipped with a Matérn covariance function, $C(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}), \boldsymbol{\theta} = \{\sigma^2, \phi, \nu\}$ with unit marginal variance (i.e. $\sigma^2 = 1$), and range and smoothness parameters (ϕ and ν, respectively) equal to 0.5 and 1, respectively. This implies that the effective range of the spatial process $w(\mathbf{s})$ is in the interval $[1, 2]$ resulting in true proportions for neighboring sub-regions that are spatially dependent. The true proportion, $\pi_t(A_{ig})$, for areal unit A_{ig} at time t is obtained by applying the expit function to the sum of the latent covariate $x(A_{ig})$ and the spatial process $w(\mathbf{s}_{ig})$ evaluated at the centroid $\mathbf{s}_{ig}$ of areal unit A_{ig}. To this quantity, we again add white noise to allow for temporal variability in the true proportions $\pi_t(A_{ig})$. This second data generation mechanism yields true proportions that are correlated in space and time.

The third data generation mechanism allows for a temporal trend in the true proportions by introducing a linear time trend $\alpha_0 + \alpha_1 t$. Thus, the true proportion

for areal unit A_{ig} at time t is now obtained by applying the expit function to the sum of the latent covariate $x(A_{ig})$ and the linear trend, $\alpha_0 + \alpha_1 t$, and adding to it white noise. We use $\alpha_0 = -1.0$ and $\alpha_1 = 0.2$ in the linear temporal trend, resulting in a noticeable increase over time of the true proportions. Adding white noise to the expit applied to $x(A_{ig}) + \alpha_0 + \alpha_1 t$, with $\alpha_0 = -1.0$ and $\alpha_1 = 0.2$, allows the true proportions to not all be clustered around 1 in the later part of the 10-unit time period.

Finally, despite the temporal structure in the true proportions generated under the third simulation setting, the $\pi_t(A_{ig})$'s are not spatially correlated. To address this shortcoming, the fourth data generation mechanism combines the second and third data generation mechanism together yielding true proportions $\pi_t(A_{ig})$ that display a temporal trend, and are correlated in space and time. The data generation procedures are summarized as follows:

$$
\begin{aligned}
\mathbf{X} &= \{x(A_{ig})\}_{i=1,\ldots,4, g=1,\ldots,25} \\
x(A_{ig}) &\overset{iid}{\sim} N(0,1) \\
\mathbf{w} &= \{w(\mathbf{s}_{ig})\}_{i=1,\ldots,4, g=1,\ldots,25} \sim \text{MVN}(0, C(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})) \\
C(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) &= \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \left(\frac{||\mathbf{s} - \mathbf{s}'||}{\phi} \right)^{\nu} \mathcal{K}_{\nu} \left(\frac{||\mathbf{s} - \mathbf{s}'||}{\phi} \right) \\
\boldsymbol{\theta} &= \{\sigma^2, \phi, \nu\} = \{1.0, 0.5, 1.0\}
\end{aligned}
\tag{3.14}
$$

The "observed estimates" $z_t^{(5)}(A_{ig})$ and $z_t^{(1)}(A_i)$, which play the equivalent role to the ACS estimates and to which we fit our model, are obtained by aggregating

Setting	Equation
1	$\pi_t(A_{ig}) = \frac{\exp\{x(A_{ig})\}}{1+\exp\{x(A_{ig})\}} + \epsilon, \epsilon \overset{iid}{\sim} N(0, 0.01)$
2	$\pi_t(A_{ig}) = \frac{\exp\{x(A_{ig}) + w(\mathbf{s}_{ig})\}}{1+\exp\{x(A_{ig}) + w(\mathbf{s}_{ig})\}} + \epsilon, \epsilon \overset{iid}{\sim} N(0, 0.01)$
3	$\pi_t(A_{ig}) = \frac{\exp\{x(A_{ig}) + \alpha_0 + \alpha_1 t\}}{1+\exp\{x(A_{ig})) + \alpha_0 + \alpha_1 t\}} + \epsilon, \epsilon \overset{iid}{\sim} N(0, 0.01)$
4	$\pi_t(A_{ig}) = \frac{\exp\{x(A_{ig}) + w(\mathbf{s}_{ig}) + \alpha_0 + \alpha_1 t\}}{1+\exp\{x(A_{ig}) + w(\mathbf{s}_{ig}) + \alpha_0 + \alpha_1 t\}} + \epsilon, \epsilon \overset{iid}{\sim} N(0, 0.01)$

Table 3.1: Data generation mechanism used in each of the four simulation settings.

the true proportions, as follows:

$$z_t^{(5)}(A_{ig}) = \frac{1}{5}\sum_{k=t-4}^{t} \pi_k(A_{ig}) + \epsilon_t^{(5)}(A_{ig}) \qquad \epsilon_t^{(5)}(A_{ig}) \overset{iid}{\sim} N(0, 0.04)$$

$$z_t^{(1)}(A_i) = \frac{1}{25}\sum_{g=1}^{25} \pi_t(A_{ig}) + \epsilon_t^{(1)}(A_i) \qquad \epsilon_t^{(1)}(A_i) \overset{iid}{\sim} N(0, 0.01) \tag{3.15}$$

Specifically, $z_t^{(5)}(A_{ig})$ is aggregated over 5-unit time intervals while $z_t^{(1)}(A_i)$ is aggregated at each time point over the regions. Both sets of observed data have random normal error added onto them, with variance chosen so that the variation in the observed data at adjacent time periods resembles the year-to-year variation observed in the ACS estimates of the proportion of families in poverty, our first case study discussed in Section 3.5.1. Without loss of generality, we assume equal population sizes in each region and sub-region. The effective sample sizes for the $z_t^{(1)}(A_i)$'s are set to 400 for each region A_i, while the effective sample sizes for the $z_t^{(5)}(A_{ig})$ are set to 100 for each subregion A_{ig}. These values were chosen based on the computed ESS's for the ACS data on proportions of families in poverty. Finally, to ensure that the simulated estimates represent proportions, we truncate them to lie in the [0, 1] interval.

We infer upon the true proportions by fitting our Bayesian hierarchical model

in (3.11) and (3.12) to the data simulated under the four simulation settings. To each of the 30 simulations in the four setting, we run the MCMC algorithm for 10,000 iterations, discarding the first 2,000 iterations for burn-in.

3.4.2 Simulation results

Taking the posterior means of the $\pi_t(A_{ig})$'s as estimates $\hat{\pi}_t(A_{ig})$ of the true proportions, we summarize the performance of our model by reporting the mean squared and the mean absolute (relative) error as well as the empirical coverage of the 95% credible interval for each $\pi_t(A_{ig})$.

Tables 3.2 and 3.3 presents results for our simulation studies, including the average empirical probability that a 95% credible interval for $\pi_t(A_{ig})$ covers the true value at each time t, the mean squared error, and the mean absolute error. The latter two are defined as the mean squared and the mean absolute difference between the $\hat{\pi}_t(A_{ig})$'s and the true values, the $\pi_t(A_{ig})$'s. In addition, Tables 3.2 and 3.3 shows the average mean squared relative error (MSRE) and the mean absolute relative error (MARE), defined respectively as the average over 30 simulations of:

$$MSRE = \frac{1}{100}\sum_{g=1}^{4}\sum_{i=1}^{25}\frac{(\hat{\pi}_t(A_{ig}) - \pi_t(A_{ig}))^2}{\pi_t(A_{ig})}$$

$$MARE = \frac{1}{100}\sum_{g=1}^{4}\sum_{i=1}^{25}\frac{|\hat{\pi}_t(A_{ig}) - \pi_t(A_{ig})|}{\pi_t(A_{ig})} \tag{3.16}$$

The 95% credible intervals, computed by taking the 2.5^{th} and 97.5^{th} percentile of the posterior samples for each $\pi_t(A_{ig})$, contain the true values between 87% and 97% of the time, with the credible intervals corresponding to the middle of the time-series ($t = 3, 4, 5, 6, 7$) having the highest coverage probabilities. The low values

for the squared and absolute errors indicate successful recovery of the true $\pi_t(A_{ig})$. Figure 3.2 presents scatterplots of the true proportions against the $\hat{\pi}_t(A_{ig})$: all plots illustrate our model's ability to disaggregate survey-based estimates of areal proportions regardless of the data generation mechanism.

3.5 ACS Data Analysis Results

3.5.1 Families in poverty

We apply the modeling framework presented in Section 3.3 to the ACS estimates of the proportion of families in Michigan living in poverty from 2006 to 2016, with the goal of obtaining annual estimates for each year at the census tract level. Of the 2,813 census tracts in Michigan, 84 (3%) did not have estimates due to insufficient surveys being administered within those tracts. A number of the excluded census tracts were primarily water-based, while others contain few to no residential buildings.

We present results from our model in a variety of ways, including a comparison of the mean and variance of our model-based estimates to those provided in the ACS dataset. However, our primary focus is on the disaggregated estimates for selected neighborhoods of two Michigan cities: Detroit and Flint. For Detroit, we chose a set of census tracts in Midtown, a mixed-use area in Detroit located north of downtown and comprising several business districts, Wayne State University, and some residential neighborhoods. Some of the census tracts in Midtown have very high poverty, while others host various sporting arenas and other downtown attractions, and thus exhibit considerably lower poverty rate. Some of the high-

(a) Simulation setting 1

t	Average Coverage	MSE	MAE	MSRE	MARE
1	0.892	0.012	0.075	0.030	0.251
2	0.930	0.007	0.056	0.024	0.199
3	0.949	0.005	0.044	0.015	0.153
4	0.958	0.004	0.042	0.011	0.128
5	0.966	0.003	0.040	0.009	0.121
6	0.965	0.003	0.041	0.009	0.115
7	0.962	0.004	0.043	0.011	0.128
8	0.955	0.006	0.045	0.016	0.159
9	0.932	0.008	0.059	0.021	0.120
10	0.910	0.013	0.078	0.029	0.263

(b) Simulation setting 2

t	Average Coverage	MSE	MAE	MSRE	MARE
1	0.875	0.011	0.088	0.030	0.234
2	0.898	0.009	0.073	0.024	0.201
3	0.921	0.007	0.062	0.021	0.153
4	0.937	0.006	0.056	0.019	0.143
5	0.944	0.004	0.044	0.015	0.104
6	0.945	0.004	0.050	0.011	0.109
7	0.926	0.006	0.054	0.012	0.136
8	0.922	0.007	0.061	0.017	0.161
9	0.908	0.010	0.075	0.019	0.210
10	0.893	0.010	0.082	0.027	0.255

Table 3.2: Results corresponding to 30 simulated datasets generated under the first two of the four settings described in Table 3.1. For each time t, $t = 1, \ldots, 10$, the table reports: (i) the average empirical coverage of the 95% credible interval of $\pi_t(A_{ig})$, $i = 1, \ldots, 4$, $g = 1, \ldots, 25$ averaged across the 100 subregions A_{ig} and 30 simulations; (ii) the mean squared error (MSE); (iii) the mean absolute error (MAE); (iv) the mean squared relative error (MSRE); and (v) the mean absolute relative error (MARE), defined in (3.16) and averaged across the 30 simulations.

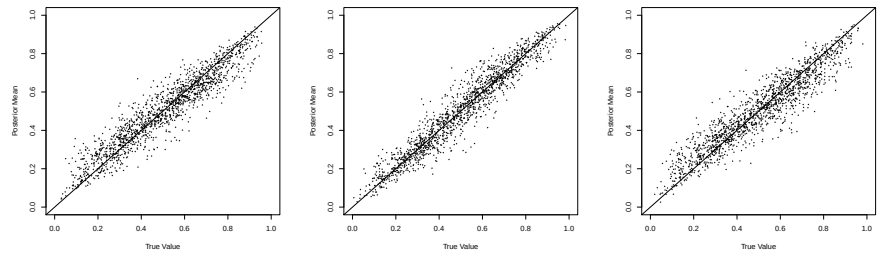
(a) Simulation setting 3

t	Average Coverage	MSE	MAE	MSRE	MARE
1	0.899	0.012	0.072	0.023	0.200
2	0.936	0.007	0.054	0.021	0.159
3	0.953	0.005	0.045	0.012	0.134
4	0.965	0.004	0.041	0.008	0.129
5	0.963	0.003	0.040	0.006	0.109
6	0.970	0.004	0.042	0.006	0.105
7	0.969	0.004	0.041	0.007	0.097
8	0.956	0.005	0.048	0.008	0.105
9	0.930	0.007	0.056	0.011	0.116
10	0.901	0.011	0.073	0.016	0.135

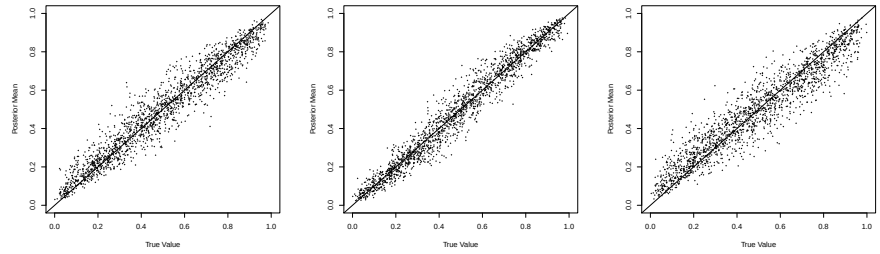
(b) Simulation setting 4

t	Average Coverage	MSE	MAE	MSRE	MARE
1	0.871	0.013	0.088	0.032	0.251
2	0.881	0.008	0.066	0.027	0.167
3	0.907	0.005	0.055	0.022	0.155
4	0.928	0.004	0.046	0.018	0.124
5	0.942	0.003	0.043	0.014	0.122
6	0.935	0.004	0.044	0.016	0.136
7	0.930	0.004	0.047	0.014	0.152
8	0.901	0.005	0.055	0.015	0.152
9	0.876	0.008	0.067	0.019	0.165
10	0.873	0.014	0.091	0.034	0.216

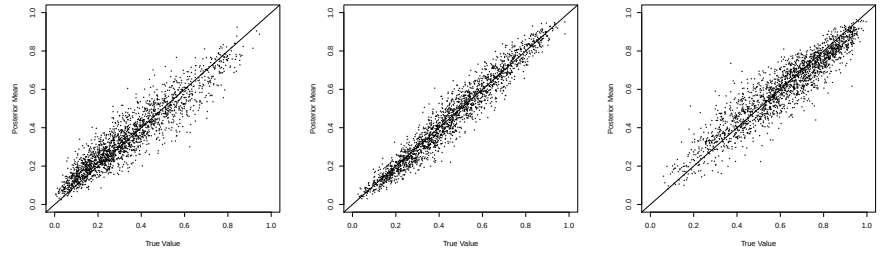
Table 3.3: Results corresponding to 30 simulated datasets generated under the last two of the four settings described in Table 3.1. For each time t, $t = 1, \ldots, 10$, the table reports: (i) the average empirical coverage of the 95% credible interval of $\pi_t(A_{ig})$, $i = 1, \ldots, 4$, $g = 1, \ldots, 25$ averaged across the 100 subregions A_{ig} and 30 simulations; (ii) the mean squared error (MSE); (iii) the mean absolute error (MAE); (iv) the mean squared relative error (MSRE); and (v) the mean absolute relative error (MARE), defined in (3.16) and averaged across the 30 simulations.



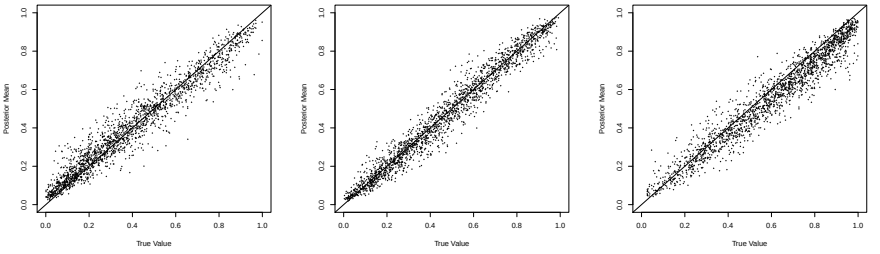
(a) Setting 1; $t = 2$ (b) Setting 1; $t = 5$ (c) Setting 1; $t = 9$



(d) Setting 2; $t = 2$ (e) Setting 2; $t = 5$ (f) Setting 2; $t = 9$



(g) Setting 3; $t = 2$ (h) Setting 3; $t = 5$ (i) Setting 3; $t = 9$



(j) Setting 4; $t = 2$ (k) Setting 4; $t = 5$ (l) Setting 4; $t = 9$

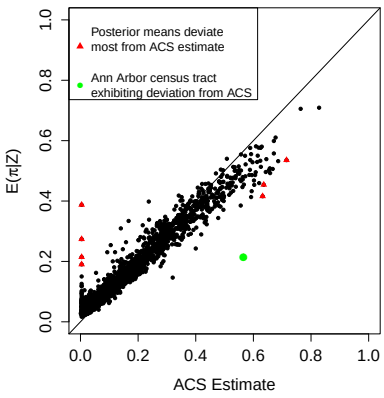
Figure 3.2: Scatterplots of the true $\pi_t(A_{ig})$'s against their corresponding estimates, $\hat{\pi}_t(A_{ig})$, over 30 simulated datasets generated under the four different simulation settings described in Table 3.1.

poverty tracts have been subject to gentrification and development in recent years (Moehlman and Robins-Somerville, 2016; Aguilar, 2015). Due to these spatial inhomogeneities and temporal changes, we believe that these tracts will effectively illustrate the need for fine scale spatio-temporal estimates in order to properly characterize neighborhood surroundings.

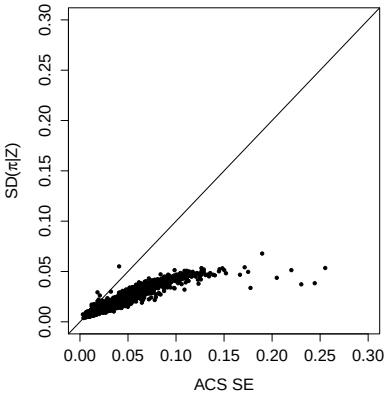
In Flint, a city in mid-Michigan that received attention in 2014 and beyond for the water crisis, we selected an urban area of the city. This part of the city has been exposed to lead contamination in drinking water and has been the center of various studies assessing the effect of poor water quality on health outcomes (Goovaerts, 2017b,a; Kennedy et al., 2016). As these studies often synthesize and combine census tract poverty data with temporally-resolved data from other sources, this example illustrates the issue of temporal misalignment often faced by health researchers when working with the 5-year estimates from the ACS.

Comparison of 5-year model-based estimates to ACS 5-year estimates

Even though we have presented our modeling framework as one whose primary utility is spatial and temporal disaggregation, using the model presented in Section 3.3 we can also generate model-based multi-year estimates. It is expected that our model-based estimates will approximately resemble the ACS estimates. Figure 4.7(a) presents a scatter plot of the 5-year census tract ACS estimates for years 2009–2013 against our multi-year model-based estimates for the same time period. On the other hand, Figure 4.7(b) presents a comparison of the ACS standard error for the 5-year census tract estimates with the posterior standard deviation of our



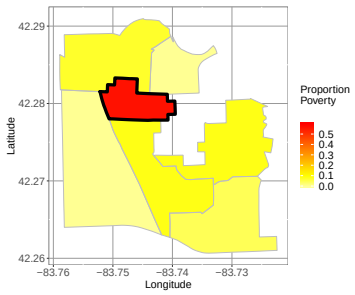
(a) Model-based vs. ACS estimates



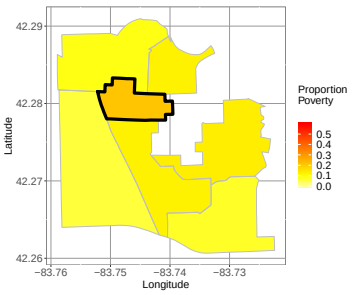
(b) Posterior SD vs. ACS SE's

Symbol	Tract ID	ACS Estimate	ACS SE
▲	26163517200	0.00	0.19
●	**26161400100**	**0.57**	**0.20**
▲	26163550800	0.00	0 .04
▲	26073000700	0.63	0.11
▲	26057000400	0.00	0.61
▲	26037011200	0.00	0.17
▲	26163500400	0.64	0.10
▲	26163512900	0.72	0.11

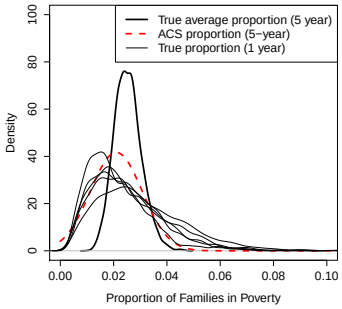
(c) ACS estimates and SE's for tracts highlighted in (a)



(d) ACS estimates, Ann Arbor



(e) Model-based estimates, Ann Arbor



(f) Posterior density

Figure 3.3: (a) Average model-based estimates at census tract level for years 2009-2013 against ACS estimates for the same time period. Census tracts deviating greatly from the identity line are denoted in red and blue. (b) Posterior standard deviation of our model-based estimates at census tract level against ACS standard error of the estimates at census tract level for years 2009-2013. (c) Tabulation of Tract ID's, ACS estimates, and ACS standard errors for census tracts in which the posterior mean deviates most greatly from the ACS estimate. Census tract 26161400100, located in Ann Arbor, is indicated in bold. (d) ACS estimate for Ann Arbor census tract 26161400100. (e) Model-based estimate for census tract 26161400100. (f) Probability density functions for census tract 26163563500: posterior densities of the (i) 5-year average proportion and (ii) 1-year proportion of families living in poverty as provided by our model, and (iii) truncated normal density function with mean and variance based on the 5-year ACS estimate.

estimates at the same spatial and temporal resolution. As Figure 4.7(a) shows, the points in the scatterplot tend to fall around the identity line, indicating a rather good agreement between our model-based estimates and the ACS ones. In addition, Figure 4.7 (b) demonstrates that our estimates are characterized by smaller uncertainty. These results may be viewed as a preliminary validation of our model.

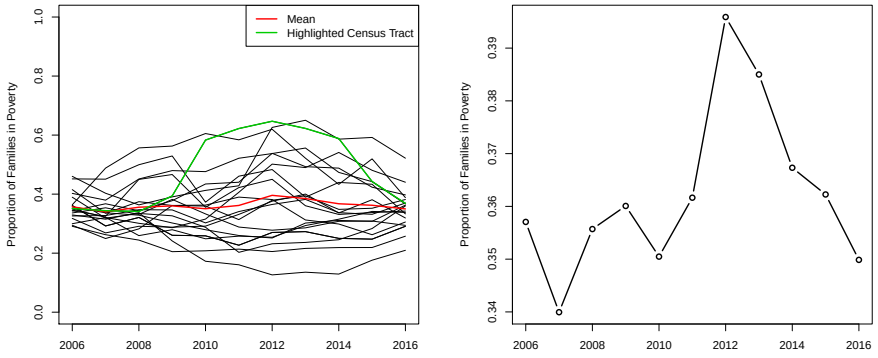
Returning to Figure 4.7(a), we can observe that a few points deviate greatly from the identity line. Many of these points correspond to census tracts with zero-valued ACS estimates. In addition, upon further inspection, we notice that these points are also the ones for which the ACS standard errors are quite large (see Figure 4.7(c)) when compared to the average ACS standard error of 0.041, averaged across all census tracts.

An example of such a census tract and its neighboring tracts are displayed in panels (d) and (e) of Figure 4.7. Census tract 26161400100 is located in downtown Ann Arbor and, according to the ACS estimates, has a poverty rate of 0.59 for the 5-year time period from 2009 to 2013. This estimate deviates greatly from neighboring counties, and has a design-based standard error around 5 times the average standard error for ACS estimates of poverty in Michigan. As our model borrows information from the neighboring census tracts in yielding an estimate, the model-based estimate for this census tract is closer to those of the neighboring tracts than the raw ACS estimates. We note that we only observe regression of the model-based estimates towards the average estimates of their neighbors in situations where the design-based standard errors are quite large.

Posterior density of true population proportions

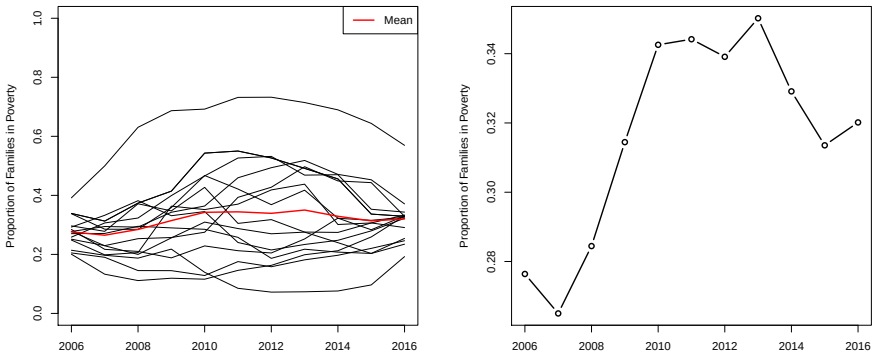
ACS estimates are provided with margins of error based on standard errors derived from an iterative estimation procedure (see U.S. Census Bureau (2014) for details) multiplied by standard normal quantiles. These may in turn be utilized to construct confidence intervals for the estimates by adding and subtracting their margins of error, with users being cautioned to use "logical boundaries when creating confidence bounds from the margins of error" U.S. Census Bureau (2008) (e.g. zero and one for proportions). This implies a truncated normal distribution centered at the ACS estimate, with variance depending on the standard error of the ACS estimate.

Through our Bayesian modeling framework, we can infer upon the posterior distribution of the true proportions at the 1-year census tract scale without imposing assumptions of symmetry or truncation. For example, Figure 4.7(e) plots the posterior density of the average proportion of families living in poverty in census tract 26163563500 located in Midtown Detroit for the 5-year period 2009-2013. For this census tract, the confidence interval for the 5-year ACS estimate for 2009-2013 would be subject to truncation at zero. Figure 4.7(e) shows the truncated normal density with mean and standard deviation equal, respectively, to the ACS estimate and its standard error. To facilitate direct comparison to the ACS estimates, Figure 4.7(e) also presents the posterior density of the 5-year average proportion of families living in poverty as provided by our model, as well as the posterior density for the 1-year proportions for years 2009, 2010, 2011, 2012 and



(a) Spaghetti plot of poverty for Midtown census tracts

(b) Mean poverty for Midtown census tracts



(c) Spaghetti plot of poverty for Flint census tracts

(d) Mean poverty for Midtown census tracts

Figure 3.4: Spaghetti plots displaying the proportion of families in poverty over time and the average poverty rate across census tracts in: (a)-(b) Midtown Detroit; (c)-(d) Flint.

2013. As the figure shows, thanks to the borrowing of information over time and from neighboring census tracts, the posterior density of the 5-year average proportion is characterized by smaller uncertainty than the normal density centered at the ACS estimate.

Disaggregated estimates of poverty for Detroit and Flint

Disaggregating the ACS estimates allows us to examine yearly trends in poverty for individual census tracts, as well as combinations of census tracts which do not form a PUMA or are not part of a highly populated county, for both of which

we could not assess temporal trends using the ACS estimates alone. Figure 3.5 presents various maps of the disaggregated estimates of the proportion of families in poverty in Michigan yielded by our model. Panel (a) displays census tract estimates for all of Michigan for year 2010. Panel (b) presents the same results for Wayne County, which contains areas of extreme poverty, as well as some of the wealthiest neighborhoods in Michigan, while panel (l) displays the estimated percentage of families in poverty in Genesee county, the county where the city of Flint lies.

Panels (c)-(k) of Figure 3.5 present yearly results over time for a subset of census tracts in Wayne County located in Midtown Detroit. This set of census tracts was selected because the poverty rates exhibit spatial heterogeneity, with certain pairs of neighboring census tracts differing by over 20%. This means that a single poverty estimate for this area (e.g. a 1-year ACS estimate at the PUMA level) would not properly characterize the neighborhood conditions of its residents. Furthermore, recent changes in these census tracts have been well-documented (Moehlman and Robins-Somerville, 2016), particularly in those tracts that comprise the Cass Corridor, an area of downtown Detroit which has faced high crime and poverty, but has recently experienced sudden gentrification (Aguilar, 2015). Therefore, the 5-year ACS estimates at the census tract level may not properly characterize yearly changes in these census tracts.

Figures 3.4(a)-(b) show the change over time in poverty rates in a set of census tracts in the Midtown area of Detroit. Consistent with national trends, Figure 3.4(b) indicates that on average the area experienced an increase in poverty

following the 2008 financial crisis in the US, which eventually decreases in later years. Figure 3.5, panels (c)-(k), highlight a census tract of particular interest, indicated in green in Figure 3.4(a), which did not experience a decrease in poverty until 2014. Year to year, it has had among the highest poverty rates in Detroit. However, recent developments such as the groundbreaking of Little Caesars Arena in 2014 and an influx of newly built restaurants and bars, might have contributed to the drop in poverty in from 2014 to 2015 (Moehlman and Robins-Somerville, 2016).

Figure 3.5, panels (m)-(u), present the disaggregated estimates for a set of census tracts in downtown Flint, MI, for years 2007-2015. Prior to its recent exposure in national media due to the water crisis, Flint had faced several financial emergencies. These are closely tied to the collapse of the auto industry in Michigan which, along with other manufacturing jobs, comprised a large portion of Flint's economy. Flint has experienced heavy outward migration, with the resident population decreasing by 18% from 2000 to 2014, and nearly 50% since 1960. This reduction ranks among the largest among cities in the United States (Murembya and Guthrie, 2016).

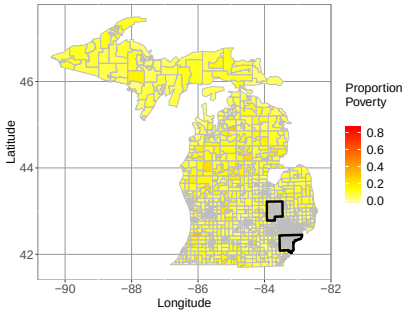
Due to its recent prominence in illustrating massive disparities in population level health-promoting services (i.e. clean drinking water), as well as the need for monitoring population health, we envision that our disaggregated estimates could be a useful resource for investigators that aim to conduct health studies in Flint. We believe that these investigators, in addition to using health outcomes and environmental exposure data, might also want to characterize neighborhoods'

conditions at a fine spatio-temporal resolution. Indeed, several papers that have investigated the effect of the Flint water crisis on health in the local population have included neighborhood poverty data in their statistical analyses (Goovaerts, 2017b,a; Kennedy et al., 2016). In particular, in a study that looked into potential sampling bias when lead levels were measured in the Flint water supply, (Goovaerts, 2017a) characterizes poverty at the census-tract level using 5-year ACS estimates that were derived from surveys administered from 2011 to 2015. However, the water samples used to measure lead levels were all taken in 2016. By using our 1-year model-based estimates, estimates of poverty could be temporally aligned with the other data sources used in the study.

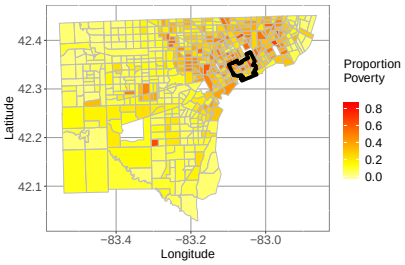
Much like the census tracts in Detroit, the temporal trends in the poverty rates in Flint, shown in Figure 3.4(c)-(d), illustrate an urban area with high overall poverty, a feature that is both heterogeneous spatially and dynamic temporally. As Figure 3.4(c)-(d) show, the average poverty rate exceeds 30% with certain census tracts having rates that exceed 60% in certain years.

3.5.2 Out-of-sample prediction

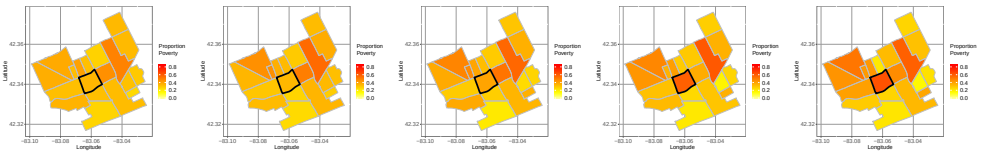
To assess our model's out-of-sample predictive performance, we consider the county-wide proportion of families in poverty during the 3-year time period from 2011-2013. We generate 3-year county-level predictions as a weighted average of the disaggregated estimates within each county census tracts for the three years from 2011 to 2013, with weights proportional to the number of families living in each tract. Then, those yearly county estimates are averaged over the 3-year time



(a) Model-based estimates: Michigan 2010

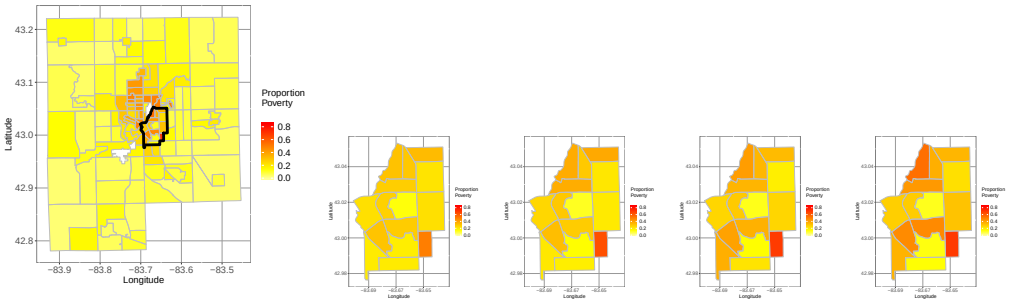


(b) Model-based estimates: Wayne County 2010

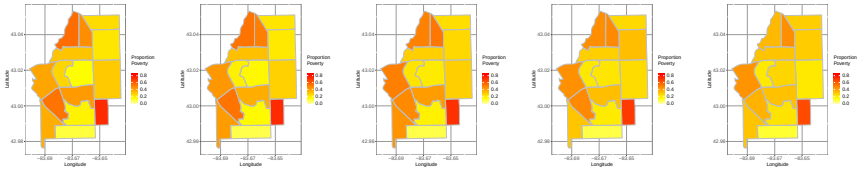


(c) 2007 (d) 2008 (e) 2009 (f) 2010 (g) 2011

(h) 2012 (i) 2013 (j) 2014 (k) 2015



(l) Model-based estimates: Genesee County 2010 (m) 2007 (n) 2008 (o) 2009 (p) 2010



(q) 2011 (r) 2012 (s) 2013 (t) 2014 (u) 2015

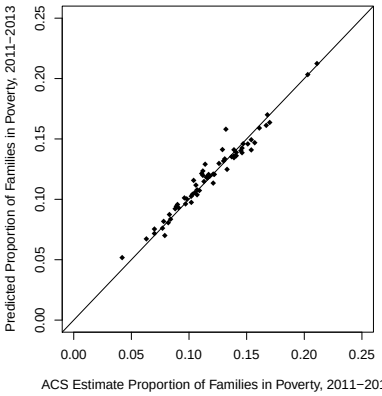
Figure 3.5: Maps showing the proportion of families in poverty in (a) Michigan, (b) Wayne County for 2010; as it changes over time in selected census tracts in: (c)-(k) Midtown Detroit; (l) Genesee County for 2010; and over time in selected census tracts in: (m)-(u) Flint.

period from 2011 to 2013. Our "true values" are the 3-year county-level ACS estimates for 64 Michigan counties, which we did not use for model fitting. This allows us to assess our model's ability to predict over time periods and areal units that are not utilized in model fitting.

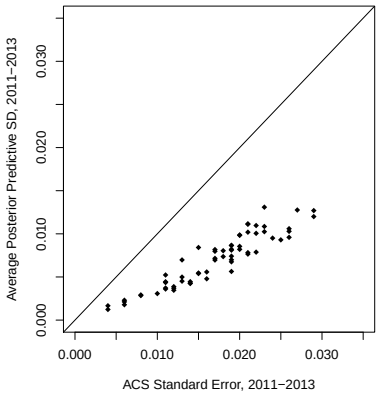
The mean squared and absolute prediction errors are 0.00004 and 0.005 respectively. The mean squared relative prediction error is 0.0003 and the mean absolute relative prediction error is 0.042. These demonstrate minimal predictive error in our modeling framework. Figure 3.6(a) compares the estimated 3-year proportion yielded by our model with the corresponding ACS estimates. As the figure shows, the two sets of estimates tend to be very similar, further indicating the strong predictive performance of our model. Finally, Figure 3.6(b) compares the posterior predictive standard deviations of the 3-year average proportions of families in poverty at the county level to the standard errors of the 3-year ACS estimates. As the figure shows, the posterior predictive standard deviations are consistently smaller than the ACS standard errors, because of the borrowing of information over time and across neighboring spatial units. Additionally, the empirical probability that a 95% prediction interval for a 3-year average proportion covers the corresponding 3-year ACS estimate is near nominal, at 92%.

3.5.3 Posterior distribution of county-level random effects

It is of interest to examine further the posterior distribution of the county-level random effects. Figure 3.7 presents in panel (a) the posterior means of the county-level random effects, $\xi(C_{A_{ig}}), \forall C_{A_{ig}}$, and in panel (b) the posterior standard



(a) ACS estimates vs. posterior predictive means; county-level random effect.



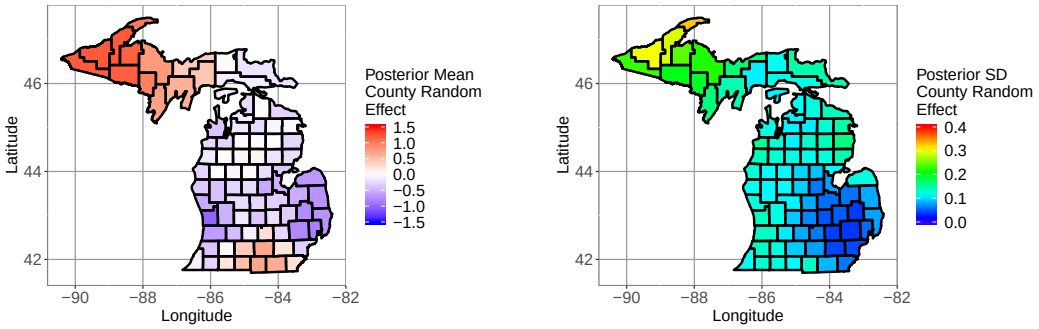
(b) ACS SE's vs. posterior predictive standard deviations; county-level random effect.

Figure 3.6: (a) Model-based estimates of the 3-year average proportion of families in poverty at counties across Michigan for the period 2011-2013 vs. the 3-year ACS estimates for the same time period. (b) Posterior predictive standard deviations of the 3-year average proportion of families in poverty at counties across Michigan for the period 2011-2013 vs ACS standard errors for the 3-year estimates at the county level

deviations of the county-level random effects.

We observe a set of positive county-level random effects in Washtenaw, Hillsdale, Lenawee, and Ingham counties (in the south of Michigan), as well as in the Upper Peninsula. The posterior standard deviations of the county-level random effects are smallest in the highly populated areas of southeast Michigan, including Wayne, Oakland, Genesee, and other counties, and greatest in the Upper Peninsula, an area of Michigan that is sparsely populated.

Sixty-six point two percent of Michigan's counties (55/83) have random effects whose posterior credible intervals did not contain zero. Among Michigan's 83 counties, 25% had positive random effects whose credible intervals did not contain zero, whereas 41% had negative random effects whose credible intervals did not contain



(a) Posterior mean of $\xi(C_{A_{ig}})$.

(b) Posterior standard deviation of $\xi(C_{A_{ig}})$.

Figure 3.7: Maps of (a) the posterior mean and (b) the posterior standard deviation of the county-level random effects $\xi(C_{A_{ig}}), \forall C_{A_{ig}}$.

zero. We posit that the posterior distribution of the county-level random effects may be informative to the stewards of the ACS who wish to better understand the survey's individual sampling frames.

3.5.4 Secondary Analysis: Gerrymandering in southeast Michigan

The term *Gerrymander* is a portmanteau that originated in an 1812 political cartoon, which drew attention to the highly irregular, salamander-like shape of one of Massachusetts' state senate districts commissioned by then Governor Elbridge Gerry (Schuck, 1987; Martis, 2008). The districts were drawn with the intention of providing a political advantage to the Democratic Republican Party, of which Gerry was a member. Gerrymandering today is a more sophisticated, data-driven process. However, the goal remains to draw districts in a fashion that provides political advantage to the party that is tasked with redistricting. Gerrymandering

often involves drawing districts that maximize the efficiency of voters of a certain political party. This may be achieved be either splitting groups of like-minded voters in order to ensure they do not represent a majority in any district, or by packing like-minded voters into a single district, such that their voting power is consolidated into one election rather than several (Schuck, 1987).

Racial gerrymandering was practiced overtly prior to the Civil Rights era, and was most often achieved by splitting communities of African American voters into several districts (Durst, 2018). The practice of racial gerrymandering has faced numerous legal challenges in the last 50 years. Through a series of Supreme Court decisions, congressional districts in which race was the predominant factor in drawing boundaries are now considered to be in violation to the Voting Rights Act and The Equal Protection Clause (National Conference of State Legislatures, 2018). Although these cases create strong precedent against racial gerrymandering, race is often a reliable proxy for political affiliation. In particular, because African American (and other racial minorities) tend to vote for Democratic candidates at a higher rate than white voters (Tyson, 2018), determining whether a case of gerrymandering was truly politically motivated or had race as a predominant factor is a somewhat subjective, and often contentious topic. In this secondary analysis, we focus on the ACS estimates of proportion of individuals who identified themselves as being Black/African-American, which will simply refer to as "proportion of Black/African-American" for the remainder of this chapter. As in the analysis of poverty, we will leverage ACS 1-year estimates for the proportion of Black/African-American in PUMAs in Michigan over the years 2006-2016, along

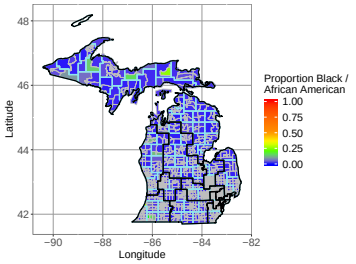
with 5-year census-tract estimates over the same time period. Our goal is to obtain 1-year census-tract estimates of proportion of Black/African-American over census tracts in Michigan. Here we present results for year 2010, the year preceding the redrawing of congressional electoral district boundaries. Specifically, our goal here is to examine the degree to which the 2011 redistricting in Michigan was carried out with the intention of concentrating (or packing) Black/African-American voters into a small number of electoral districts.

To this end, Figure 3.8 (a) and (b) present the estimated proportion of Black/African-American in census tracts across Michigan in year 2010, as estimated by our Bayesian hierarchical model, with overlaid the old (pre-2011) and the new (post-2011) congressional boundaries. Note that the number of congressional electoral districts in Michigan decreased from 15 to 14 in 2011. As Black/African-Americans in Michigan are concentrated mostly in specific areas, Figures 3.8 (f) through (h) focuses on the South West subregion in Michigan around Detroit, that is, the region covered by the pre-2011 congressional districts 9, 11, 12, 13, 14 and post-2011 congressional districts 9, 11, 13, 14.

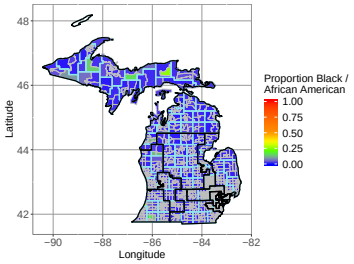
As Figure 3.8 (f) and (g) show, the new congressional districts 13 and 14 comprise all census tracts that, according to our estimates, in 2010 were predominantly Black/African-American. On the other hand, prior to redistricting in 2011, census tracts that in 2010 had large proportion of Black/African-Americans were included in congressional districts 9, 12, 13 and 14. Particularly indicative is the extremely irregular shape of congressional district 14, which has been showcased as an illustrative example of gerrymandering in Michigan (Winowiecki, 2018) and whose

(a) Map of old congressional boundaries.

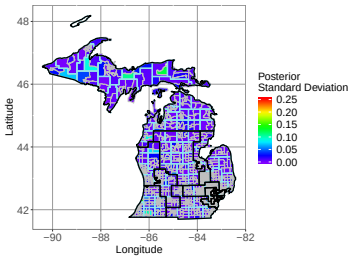
(b) Map of new congressional boundaries.



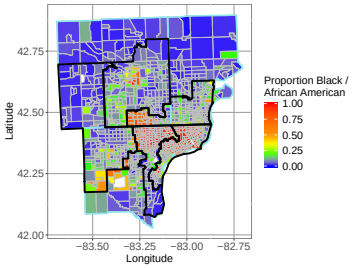
(c) Proportion Black/African American: old boundaries.



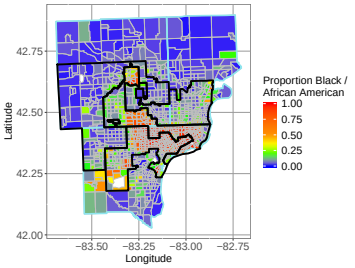
(d) Proportion Black/African American: new boundaries.



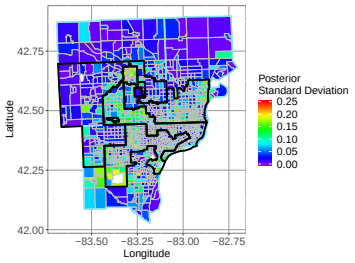
(e) Posterior SD.



(f) Proportion Black/African American: old boundaries.



(g) Proportion Black/African American: new boundaries.



(h) Posterior SD.

Figure 3.8: (a)-(b) Numbered map of the congressional district boundaries in Michigan pre-2011 (a) and post-2011 (b). (c)-(d) Estimated proportion of Black/African-American in census tracts in Michigan in year 2010 with overlaid the congressional boundaries pre-2011 (c) and post-2011 (d). (e) Posterior standard deviation of the proportion of Black/African American in Michigan in year 2010. (f)-(g) Estimated proportion of Black/African-American in census tracts in southwest Michigan in year 2010 with overlaid congressional boundaries pre-2011 (f) and post-2011 (g). (h) Posterior standard deviation of the proportion of Black/African American in southwest Michigan in year 2010.

boundaries seems to follow the boundaries of census tracts that were predominantly Black/African-American in 2010.

3.6 Discussion

This chapter has presented and discussed results obtained from applying a spatio-temporal Bayesian hierarchical model to disaggregate multi-year estimates of proportions over areal units derived from sampling surveys. Our modeling framework breaks with previous approaches with a similar goal in that our model explicitly accounts for the survey design in modeling the survey-based estimates. A notable outcome of our model is that it allows for the generation of data on socioeconomic indicators at fine spatial and temporal resolution, which could be employed by social and health science researchers in their statistical analyses, resolving the typical problem of spatial and temporal misalignment often encountered when multiple data sources are queried for research investigations and analyses. We demonstrate the utility of our Bayesian hierarchical modeling framework by applying it to ACS estimates, highlighting not only the fact that our model yields annual estimates at census tract spatial resolution, but showcasing also the smaller uncertainty of our estimates compared to that of the ACS estimates, a result that we attribute to the borrowing of information over time and from neighboring units. We use our disaggregated estimates to examine trends over time of family poverty in Michigan's census tracts as well as assess the proportion of Black/African-Americans in Michigan's census tracts in year 2010, the year prior to the re-drawing of congressional districts in Michigan.

In addition to the modeling framework presented here, Chapter IV presents a modeling framework to disaggregate ACS estimates of count-valued neighborhood characteristics while explicitly accounting for the sampling survey design. In addition, as other community characteristics beyond poverty might be of interest, we are exploring the possibility of using our modeling framework on other community indicators, leveraging the breadth of indicators provided by the ACS.

We conclude by noting that when fitting our Bayesian hierarchical model to a large set of municipal subdivisions (say every census tract in the U.S.), users may wish to incorporate dimension-reduction techniques to reduce the computational burden associated with estimating the point-reference spatio-temporal process $w_t(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$. Examples of dimension reduction techniques include fixed-rank Kriging (Cressie and Johannesson, 2008), predictive process modeling (Banerjee et al., 2008), nearest neighbor Gaussian process models (Datta et al., 2016), and the M-RA modeling approach of Katzfuss (2017), which we use within our modeling framework and for which we have proposed here an extension to the spatio-temporal setting.

CHAPTER IV

A Spatio-temporal Change of Support Model for Survey-based Estimates of Births in Michigan

4.1 Introduction

In recent years, large spatial and spatio-temporal datasets have become increasingly available to the public. The efforts of data stewards at entities such as the U.S. Census Bureau and the Department of Health and Human Services in collecting and releasing these data has created great opportunity for public health researchers to better understand environmental, demographic, and economic trends of administrative areal units and their residents. To gain these insights, it's customary for researchers to leverage data from multiple sources, often available at different spatial or spatio-temporal resolutions. Combining incongruous spatial data is one of the classical goals of spatial statistical analyses, and a vast literature in spatial statistics is devoted to addressing what is called the change of support problem (COSP) (Banerjee et al., 2004; Gotway and Young, 2002), that is the problem of inferring upon a spatial process at a resolution that differs from that of the data.

Although several papers have discussed methods to address the COSP in dif-

ferent settings – see Gelfand et al. (2001, 2010); Gotway and Young (2002) – efforts to address the COSP in the context of survey-based estimates are still sparse. The most prominent works in this context include Bradley et al. (2015) and Bradley et al. (2016b), whose modeling frameworks are motivated by multi-year estimates derived from the American Community Survey (ACS), an ongoing survey conducted by the U.S. Census Bureau. Because the ACS is also the focus of this chapter, we will provide a more comprehensive introduction in the sequel. The work of Bradley et al. (2015) and Bradley et al. (2016b) can be considered pioneering in this field, and introduces models to generate estimates of characteristics such as median household income or number of households living below the poverty line at new spatial resolutions starting from the ACS estimates. In Bradley et al. (2015), the authors focused on survey estimates of variables that can be reasonably thought as having Gaussian sampling errors. There, the change of support problem was handled through the introduction of a point-referenced Gaussian spatio-temporal process, to which the ACS estimates were linked through an integration over space. Moving onto survey-based estimates of counts, Bradley et al. (2016b) proposed a Bayesian hierarchical model that specified a Poisson distribution at the first level for the ACS estimates. Even though, as in the case of normally-distributed random variables, Bradley et al. (2016b) related the parameters of the likelihood for the survey-based estimates to a latent, point referenced spatial process, ultimately, during model fitting, they worked with an aggregation of the process at the finest plausible areal-unit level, thus working with a latent random field, further expressed as a linear combination of basis functions derived

from the Moran's I propagator operator. To improve their estimation of the latent random field, starting from the remark that the ACS estimates are also provided with survey-based variances that utilize the same source of information as the ACS estimates, they added to their Bayesian hierarchical model an additional level that linked the logarithm of the design-based ACS variance to the latent, underlying random field. This COSP model for survey-based estimates of counts was presented only in a spatial version, and was not used to disaggregate temporally the multi-year ACS estimates. As the model by Bradley et al. (2016b) can be considered as the gold standard for disaggregation of ACS estimates, later in the Chapter we offer an extension to the space-time setting and we use this extension in an out-of-sample predictive performance comparison with our model.

Other modeling efforts that address the COSP for ACS estimates include the multiresolution Bayesian model of Savitsky (2016) and the Bayesian hierarchical model of Simpson et al. (2019). Although both models share the same goal of generating estimates at a finer resolution than what is publicly available, their modeling frameworks take very different approaches. Savitsky (2016) proposes a Bayesian non-parametric multi-resolution modeling framework specified following the nested structure of the ACS estimate and whose goal is to borrow information across nested levels to shrink the variance in the 1-year survey estimates at the county level. On the other hand, Simpson et al. (2019) present a Bayesian hierarchical model that combines the information in the ACS areal-level estimates with the information in the individual-level Public-use Microdata Sample (PUMS) data.

This Chapter follows this research line and presents a Bayesian hierarchical model to disaggregate estimates of areal-level, count-valued indicators obtained from multi-year surveys. The contribution of our modeling framework, compared to the previous ones cited above, is the fact that our model accounts for the survey design through the design effect (Kish, 1995), that is the ratio of an estimator's design-based variance over its variance under simple random sampling. Incorporation of the design effect, a metric well known in the survey statistical literature, enables us to propagate the often inflated survey variances of the estimates during inference. We have already illustrated in Chapter III how the design effect can be accounted for in a Bayesian hierarchical model to disaggregate areal-level estimates of proportions provided by the ACS. Here, we work again with the ACS but we focus on areal-level estimates of counts.

The ACS is an ongoing survey that was initiated in 2005 and is conducted by the U.S. Census Bureau with the intent of replacing the Census Long Form. By administering approximately 3.5 million surveys per year, the Census Bureau can release publicly up to date estimates of demographic, economic, and housing statistics at various spatial and temporal resolutions. Due to both statistical precision and subject privacy, estimates for small municipal sub-divisions, such as census tracts or block groups, refer to 5-year time periods, whereas municipal subdivisions with populations greater than 65,000 receive annual estimates. While not all counties meet this population threshold, the U.S. Census Bureau partitions states into Public Use Microdata Areas (PUMAs), which are comprised of contiguous census tracts and counties whose population exceed 100,000, meaning

that researchers have access to yearly estimates at the PUMA level. The spatio-temporal resolution of these estimates represent a trade-off for researchers. While the spatial support of the 5-year estimates is desirable for researchers aiming to understand the neighborhood dynamics, the 5-year temporal support may not properly characterize temporal trends. On the other hand, the 1-year estimates have a desirable temporal support, but their aggregation over large spatial regions diminishes their ability to characterize more local regions in a meaningful way. A solution to this problem is to model the 5- and 1-year estimates jointly through a spatio-temporal change of support model, which can yield estimates at the 1-year temporal scale at as fine of a spatial resolution as is available through the ACS. Because ultimately, our goal is to generate estimates at the finest spatial and temporal resolution available in the ACS, we call our model a disaggregation model.

As case study for our model, we will focus on an ACS count-valued estimate that might be of interest to public health researchers who want to understand spatial and temporal changes in maternity across administrative areal units: the number of births in Michigan. In the last thirty years, the live birth rate in the state of Michigan has decreased from 16.4 live births per 1,000 population to 11.2 (MDHHS, 2017), with the number of live births in 2017 reaching its lowest number since 1941 (Mack, 2019). However, the decrease in live births rate is not distributed equally among counties in Michigan nor it is homogeneous across race: the decrease in live birth rate is not as steep for black mothers, and the trend is the opposite when considering other racial subgroups (Mack, 2019). Although here we will work

with estimates of the overall number of births in areal units in Michigan, our model could be applied to number of births within specific demographic subgroups and could be used to highlight estimated changes at the yearly temporal scale and at the census tract resolution within each demographic subgroup. This in turn would enable us to identify potential disparities as well as areas where such disparities are more pronounced. Working with the overall number of births in areal units across Michigan, in this Chapter we speculate that our disaggregated estimates of the number of births could be used to identify locations where additional hospitals or resources are needed. We highlight this by developing a rather rudimentary metric for hospital needs: the number of births per hospital bed, and we flag a particular hospital whose number of births per bed has a high posterior mean and standard deviation.

The remainder of this Chapter is organized as follows: in Section 4.2 we present our modeling framework to disaggregate ACS count-valued estimates, as well as a spatio-temporal extension of the model of Bradley et al. (2016b), who first proposed a model to address the COSP for such ACS estimates. We will use this spatio-temporal extension as a benchmark model to assess the out-of-sample predictive performance of our model in real data analysis, and to quantify the gains offered by a model that incorporates the concept of survey design effect in its formulation. Section 4.3 is devoted to results from: (1) simulation studies, where we test the capabilities of our model in disaggregating multi-year estimates of counts when we know the true counts at the finest spatial and temporal resolution, and (2) results relative to the ACS case study. For the latter, we present both results

related to in-sample predictions and out-of-sample predictions. Finally, Section 4.4 concludes the Chapter with a discussion on the proposed model and avenues for further research.

4.2 Methods

This section presents two modeling frameworks for the spatio-temporal disaggregation of estimates of count-valued community characteristics: our own model, and a spatio-temporal extension of a model previously offered only in a spatial context by Bradley et al. (2016b) to address the COSP for ACS estimates of counts. The model we propose is a Bayesian hierarchical model that, at the first stage, models the ACS estimates of counts using a Poisson distribution and incorporates the survey's design effect in order to properly propagate the inflated survey variance of the ACS estimates. Because of this, hereafter we will refer to our model as the "Poisson with Design Effect" modeling framework, shortened in the rest of the Chapter as either "Poisson DEFF" or "PD" in some instances, while we will refer to the spatio-temporal extension of the model by Bradley et al. (2016b) as the BWH model.

4.2.1 A Poisson with Design Effect modeling framework for disaggregating ACS estimates: a general formulation

We now present the model we propose to disaggregate ACS multi-year estimates of counts while accounting for the survey's design effect. For the sake of clarity, we will formulate it using as an example the problem of estimating the number of births in census tracts at the 1-year temporal resolution when data available are

the 5-year census-tract and the 1-year PUMA estimates provided by the ACS.

According to our model, the process underlying the ACS estimates, that is the true number of births in areal unit A over the l-year time period ending in time t, is distributed according to a Poisson distribution, that is:

$$W_t^{(l)}(A) = \text{Pois}(N_t(A)\lambda_t^{(l)}(A)) \tag{4.1}$$

where $N_t(A)$ is some measure of the population size at time t and $\lambda_t^{(l)}(A)$ is a rate parameter. Ideally, $N_t(A)$ is known and therefore not a random quantity, however in practice it might not be, and it has to be estimated through some proxies.

For the sake of stating our modeling framework, consider briefly the hypothetical case in which we could estimate $W_t^{(l)}(A)$ via a simple random sample of size $m_t^{(l)}(A)$. Then, this SRS-based estimator, denoted by $\hat{W}_{SRS,t}^{(l)}(A)$, would be the product of the number of cases observed in the SRS, denoted by $y_t^{(l)}(A)$, times the inverse probability of selection, that is, $\hat{W}_{SRS,t}^{(l)}(A)$ is:

$$\hat{W}_{SRS,t}^{(l)}(A) \;=\; y_t^{(l)}(A)\frac{N_t(A)}{m_t^{(l)}(A)} \tag{4.2}$$

where, assuming that (4.1) holds for any "population", $y_t^{(l)} \sim \text{Pois}(m_t^{(l)}(A)\lambda_t^{(l)}(A))$. Under this distributional assumption, the estimator $\hat{W}_{SRS,t}^{(l)}(A)$ admits variance:

$$\text{Var}\left(\hat{W}_{SRS,t}^{(l)}(A)\right) = \lambda_t^{(l)}(A)\frac{N_t(A)^2}{m_t^{(l)}(A)}$$

Now, let $\hat{W}^{(l)}_{ACS,t}(A)$ denote the ACS estimate of $W^{(l)}_t(A)$. ACS estimates are provided with design-based variances which we denote by $\tau^{2(l)}_t(A)$, for varying t, l and A. To derive the *effective sample size*, $m^{*(l)}_t(A)$, that is, the sample size needed in a SRS to yield an estimator of $W^{(l)}_t(A)$ with the same variance as the design-based variance of the ACS estimate, we set the design effect equal to 1, thus equating the design-based variance $\tau^{2(l)}_t(A)$ to the variance of the estimator under SRS, and we solve it for the sample size. Setting:

$$\tau^{2(l)}_t(A) = \frac{\lambda^{(l)}_t(A)N_t(A)^2}{m^{(l)}_t(A)}$$

yields:

$$m^{(l)}_t(A) = \frac{\lambda^{(l)}_t(A)N_t(A)^2}{\tau^{2(l)}_t(A)} \tag{4.3}$$

Finally, substituting the ACS estimate $\hat{W}^{(l)}_{ACS,t}(A)$ for the mean of the random variable $W^{(l)}_t(A)$, that is $\lambda^{(l)}_t(A)N_t(A)$, in (4.3) provides the *effective sample size* $m^{*(l)}_t(A)$ or ESS:

$$m^{*(l)}_t(A) = \left[\frac{\hat{W}^{(l)}_{ACS,t}(A)N_t(A)}{\tau^{2(l)}_t(A)}\right] \tag{4.4}$$

with $[\cdot]$ denoting the nearest integer operator, introduced to ensure that the effective sample size $m^{*(l)}_t(A)$ is an integer number. Having derived the effective sample size for the SRS that would yield an estimator with the same variance as the ACS design-based variance, we now need to derive the number of births we expect

to observe in this SRS of size $m_t^{*(l)}(A)$, or the *effective number of cases*, $y_t^{*(l)}(A)$, which we will subsequently model as a Poisson random variable. To compute the effective number of cases $y_t^{*(l)}(A)$, we use (4.2), solving it for $y_t^{(l)}(A)$, the number of births in the SRS of size $m_t^{(l)}(A)$ and adjust it to reflect the fact that now we have a SRS of size $m_t^{*(l)}(A)$. Finally, substituting the ACS estimate, $\hat{W}_{ACS,t}^{(l)}(A)$ for the SRS-based estimate in (4.2), we obtain the following expression for the *effective number of cases* (ENC):

$$y_t^{*(l)}(A) = \left[\frac{\hat{W}_{ACS,t}^{(l)}(A) m_t^{*(l)}(A)}{N_t(A)} \right] \tag{4.5}$$

where $[\cdot]$ denotes rounding to the nearest integer.

Having derived the effective number of cases as a function of the ACS estimate, $\hat{W}_{ACS,t}^{(l)}(A)$, and of the effective sample size, $m_t^{*(l)}(A)$, which in turn depends on the ACS design-based variance $\tau_t^{2(l)}(A)$, we take this new variable as our datum and model it as arising from a Poisson distribution with a parameterization that is similar to that of (4.1), that is: for a time period l ending in year t and an areal unit A,

$$y_t^{*(l)}(A) \sim \text{Pois}(m_t^{*(l)}(A)\lambda_t^{(l)}(A))$$

To enable disaggregation of the ACS estimates to the finest spatio-temporal resolution possible in the ACS, that is, 1-year and census tract level, we model $\lambda_t^{(l)}(A)$ as an aggregation of an underlying Gaussian spatio-temporal process $\zeta_t(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, which we link to the rate $\lambda_t^{(l)}(A)$ through a log link. Hence we state that:

$$\log\left(\lambda_t^{(l)}(A)\right) = \frac{1}{l}\sum_{k=t-l+1}^{t}\frac{1}{|A|}\int_{\mathbf{s}\in A}\zeta_k(\mathbf{s})d\mathbf{s} \tag{4.6}$$

In turn, we decompose the point-referenced Gaussian spatio-temporal process $\zeta_k(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, into a large-scale spatio-temporal trend, $\mu_k(\mathbf{s})$, representing the mean of the process, and a spatio-temporal random effect $w_k(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, $\forall k = t-l+1, \ldots, t$, such that (4.6) becomes:

$$\log\left(\lambda_t^{(l)}(A)\right) = \frac{1}{l}\sum_{k=t-l+1}^{t}\frac{1}{|A|}\int_{\mathbf{s}\in A}\left(\mu_k(\mathbf{s}) + w_k(\mathbf{s})\right)d\mathbf{s} \tag{4.7}$$

Finally, to share information across space and time, we model the spatio-temporal random effect, $w_k(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, as a Gaussian spatio-temporal process with a separable space-time covariance function with an AR(1) structure in time and spatial dependence encoded via the covariance function $C(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$, $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$. Given that in our data analysis, we will work with a large number of areal units, to improve computational efficiency, we approximate the spatio-temporal process $w_t(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, via the basis function expansion described below.

4.2.2 The spatio-temporal multi-resolution approximation (ST-MRA)

In recent years, a number of methods have been proposed to alleviate the computational burden associated with fitting a spatial or spatio-temporal statistical model to large dimensional datasets. Many of these methods fall under the class of basis function expansions, which consist of expressing the spatial process in consideration as a linear combination of basis functions with appropriate basis function

weights, chosen so to ensure that the spatial process possesses an admittable covariance function. Due to the large number of areal units we anticipate to be working with, estimating the spatio-temporal process $w_t(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, in (4.7) without any approximation is computationally burdensome. Therefore, we approximate it via a basis function expansion or some other dimension reduction technique. Here, we elect to utilize the multi-resolution approximation (MRA) idea of Katzfuss (2017) which decomposes a spatial process through repeated implementation of a predictive process approximation (Banerjee et al., 2008) over recursive domain partitions. Given that we are working in a spatio-temporal setting we extend it to a space-time context as we have also done in Chapter III.

Our spatio-temporal extension of the MRA, hereafter referred to as the ST-MRA, decomposes a spatio-temporal Gaussian process, here $w_t(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, $t = 1, \ldots, T$, provided with a separable space-time covariance function with an AR(1) structure in time and spatial dependence encoded by a covariance function $C(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$, $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$, into a linear combination of spatial basis functions and dynamic spatio-temporal basis function weights.

Specifically, the M-level approximation of $w_t(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, $t = 1, \ldots, T$, denoted by $w_{t,M}(\mathbf{s})$ is given by a sum over multi-resolution levels (indexed by m) and recursive partitions of the domain (indexed by j):

$$w_{t,M}(\mathbf{s}) := \sum_{m=0}^{M} \sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s}) \boldsymbol{\eta}_{t,m,j} \tag{4.8}$$

The basis functions $\mathbf{b}_{m,j}(\mathbf{s})$ are defined as in the MRA, for which more details

are offered either in Katzfuss (2017) or in Chapter II within this dissertation. Here we offer a short description of how they are derived. To define the basis functions, we first introduce a set of r knots on the spatial domain $\mathcal{S}$ (level 0), and then at each level m ($m = 0, \ldots, M$), we recursively partition the spatial domain $\mathcal{S}$ into J^m non-overlapping subregions in which we introduce r knots. If $S^*_{m,j}$ denotes the set of r knots defined on partition j of level m, the MRA basis functions $\mathbf{b}_{m,j}(\mathbf{s})$, for $j = 1, \ldots, J^m; m = 0, \ldots, M$ are defined recursively as:

$$
\begin{aligned}
v_0(\mathbf{s}_1, \mathbf{s}_2) &= C(\mathbf{s}_1, \mathbf{s}_2; \boldsymbol{\theta}) \\
v_{m+1}(\mathbf{s}_1, \mathbf{s}_2) &= 0 \qquad \text{if } \mathbf{s}_1 \text{ and } \mathbf{s}_2 \text{ are in different regions at resolution } m \\
v_{m+1}(\mathbf{s}_1, \mathbf{s}_2) &= v_m(\mathbf{s}_1, \mathbf{s}_2) - \mathbf{b}_{m,j}(\mathbf{s}_1)' \mathbf{K}_{m,j} \mathbf{b}_{m,j}(\mathbf{s}_2) \qquad \text{otherwise} \\
\mathbf{K}^{-1}_{m,j} &= v_m(S^*_{m,j}, S^*_{m,j}) \\
\mathbf{b}_{m,j}(\mathbf{s}) &= v_m(\mathbf{s}, S^*_{m,j})
\end{aligned}
$$

To complete the description of the ST-MRA, we introduce the prior distributions specified for the basis function weights: at time $t = 0$ we assume $\boldsymbol{\eta}_{0,m,j} \sim N_r(\mathbf{0}, \mathbf{K}_{m,j})$ while for $t = 1, \ldots, T$:

$$
\begin{aligned}
\boldsymbol{\eta}_{t,m,j} | \boldsymbol{\eta}_{t-1,m,j}, \boldsymbol{\eta}_{t-2,m,j}, \ldots, \boldsymbol{\eta}_{0,m,j} &\sim N_r(\alpha \boldsymbol{\eta}_{t-1,m,j}, \mathbf{U}_{m,j}) \\
\mathbf{U}_{m,j} &= (1 - \alpha^2) \mathbf{K}_{m,j}
\end{aligned} \tag{4.9}
$$

Here, α is provided with a uniform distribution on the interval from 0 to 1. With the above construction for the basis functions $\mathbf{b}_{m,j}(\mathbf{s})$ and the prior distribution specification for the basis function weights $\boldsymbol{\eta}_{t,m,j}$, the ST-MRA in (4.8) provides an

approximation of a separable, spatio-temporal process $w_t(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, $t = 1, \ldots, T$, equipped with AR(1) dependence structure in time and spatial covariance function $C(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$. Readers may refer to Appendix B.1 for proofs showing that the above expression approximates a spatio-temporal Gaussian process with the stated spatio-temporal dependence structure.

4.2.3 The complete model

Adapting the modeling framework we introduced in Section 4.2.1 to our case study and combining it with the ST-MRA introduced in Section 4.2.2, we now provide the formulation of our model to disaggregate multi-year ACS estimates of count-valued community characteristics – in our case, the number of births in an areal unit. Let's denote by $\hat{W}^{(5)}_{ACS,t}(A_{ig})$ the 5-year ACS estimate for the number of births for census tract g within PUMA i, $i = 1, ..., I$, $g = 1, ..., G_i$, and by $\hat{W}^{(1)}_{ACS,t}(A_i)$ the 1-year PUMA estimate, and let's use the notation $\tau_t^{2(5)}(A_{ig})$ and $\tau_t^{2(1)}(A_i)$ to indicate their design-based variances, respectively. Casting these estimates and their variances into (4.4) and (4.5), using the ACS estimated number of women per census tract and PUMA respectively as proxies for $N_t(A_{ig})$ and $N_t(A_i)$, results in the following ESS and ENC for the 5-year census tract estimates and the 1-year PUMA estimates:

$$
\begin{aligned}
m_t^{*(5)}(A_{ig}) &= \left[\frac{\hat{W}_{ACS,t}^{(5)}(A_{ig})N_t^{(5)}(A_{ig})}{\tau_t^{2(5)}(A_{ig})}\right] \\
y_t^{*(5)}(A_{ig}) &= \left[\frac{\hat{W}_{ACS,t}^{(5)}(A_{ig})m_t^{*(5)}(A_{ig})}{N_t^{(5)}(A_{ig})}\right]
\end{aligned}
$$

$$
y_t^{*(5)}(A_{ig})|\lambda_t^{(5)}(A_{ig}) \overset{ind}{\sim} \text{Pois}(m_t^{*(5)}(A_{ig})\lambda_t^{(5)}(A_{ig})) \tag{4.10}
$$

$$
\begin{aligned}
m_t^{*(1)}(A_i) &= \left[\frac{\hat{W}_{ACS,t}^{(1)}(A_i)N_t^{(1)}(A_i)}{\tau_t^{2(1)}(A_i)}\right] \\
y_t^{*(1)}(A_i) &= \left[\frac{\hat{W}_{ACS,t}^{(1)}(A_i)m_t^{*(1)}(A_i)}{N_t^{(1)}(A_i)}\right]
\end{aligned}
$$

$$
y_t^{*(1)}(A_i)|\lambda_t^{(1)}(A_i) \overset{ind}{\sim} \text{Pois}(m_t^{*(1)}(A_i)\lambda_t^{(1)}(A_i))
$$

In this modeling framework, we have assumed that, conditional on their respective rate parameters $\lambda_t^{(5)}(A_{ig})$ and $\lambda_t^{(1)}(A_i)$, the effective number of cases $y_t^{*(5)}(A_{ig})$ and $y_t^{*(1)}(A_i)$ are independent. However, given that we model the rate parameters to be aggregations of the same underlying point-referenced spatio-temporal process, the effective number of cases are marginally dependent in space and time. Appendix C provides derivations of the marginal covariance of the effective number of cases under our proposed modeling framework.

The rate parameters, $\lambda_t^{(5)}(A_{ig})$ and $\lambda_t^{(1)}(A_i)$, for the 5-year and 1-year number of births, respectively, can be stated in terms of the finest desired spatio-temporal

resolution, that is the 1-year census tract resolution, through the following equations:

$$\begin{aligned}\lambda_t^{(5)}(A_{ig}) &= \frac{1}{5}\sum_{k=t-4}^{t}\lambda_k^{(1)}(A_{ig}) \\ \lambda_t^{(1)}(A_i) &= \frac{1}{N_t(A_i)}\sum_{h=1}^{G_i}N_t(A_{ih})\lambda_t^{(1)}(A_{ih})\end{aligned}$$

Being interested in the disaggregation of the ACS estimates, the next step consists into expressing $\lambda_t^{(1)}(A_{ig})$, the rate at our desired spatio-temporal support (the 1-year census tract level), as an aggregation of a spatio-temporal process to which we apply the ST-MRA to alleviate computations. Hence, from (4.7) and (4.9), we obtain:

$$\begin{aligned}\log\left(\lambda_t^{(1)}(A_{ig})\right) &= \frac{1}{|A_{ig}|}\int_{\mathbf{s}\in A_{ig}}\left(\mu_t(\mathbf{s})+\sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(\mathbf{s})\boldsymbol{\eta}_{t,m,j}\right)d\mathbf{s} \\ &\approx \mu_t(A_{ig})+\sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{t,m,j}+\epsilon_t(A_{ig}) \qquad (4.11)\end{aligned}$$

At the finest desired spatio-temporal support, we compute the integral in (4.7) via numerical approximation. The aggregated basis functions $\mathbf{b}_{m,j}(A_{ig})$ in (4.12) are taken to be the integral of the basis functions $\mathbf{b}_{m,j}(\mathbf{s})$ as $\mathbf{s}$ varies in areal unit A_{ig}, which we approximate by sampling several points within each A_{ig} and taking the mean of the basis functions at those selected points. A similar approximation can be performed for $\mu_t(A_{ig})$ in the case of spatially-varying large scale spatio-temporal trend, $\mu_t(\mathbf{s})$.

We note that the term $\epsilon_t(A_{ig})$ in (4.11) serves two purposes. First, it accounts for any residual error due to the ST-MRA approximation or the spatio-temporal

aggregation as well as any departure of the data from (4.7). Second, we wish to use this term to capture zero-valued counts that we would struggle to capture under a typical modeling framework that utilizes the Poisson distribution. To this end, we specify a bimodal prior distribution on $\epsilon_t(A_{ig})$, specifically a mixture of normal distributions for $i = 1, ..., I$ and $g = 1, ..., G_i$:

$$\epsilon_t(A_{ig}) \overset{ind}{\sim} \gamma_t(A_{ig})N(0, \tau_1^2) + (1 - \gamma_t(A_{ig}))N(c, \tau_2^2)$$

where c is some negative number, and $1 - \gamma_t(A_{ig})$ can be viewed as the probability that census tract A_{ig} has no births at time t. By specifying this prior distribution on the $\epsilon_t(A_{ig})$'s, we have accounted for zero-inflation at the spatio-temporal level in a fashion that adds little complexity to the modeling framework (as opposed to, say, specifying a data model involving a zero-inflated Poisson distribution).

Finally, we assume that $\mu_t(A_{ig}), i = 1, ..., I, g = 1, ..., G_i$ is constant over space and time, and can therefore just be expressed as μ. With such simplification, our final model to disaggregate ACS estimates of count-valued community characteristics is stated as:

$$
\begin{aligned}
y_t^{*(5)}(A_{ig}) \mid \lambda_t^{(5)}(A_{ig}) &\overset{ind}{\sim} \text{Pois}\left(m_t^{*(5)}(A_{ig})\lambda_t^{(5)}(A_{ig})\right) \\
y_t^{*(1)}(A_i) \mid \lambda_t^{(1)}(A_i) &\overset{ind}{\sim} \text{Pois}\left(m_t^{*(1)}(A_i)\lambda_t^{(1)}(A_i)\right) \\
\lambda_t^{(5)}(A_{ig}) &= \frac{1}{5}\sum_{k=t-4}^{t}\lambda_k^{(1)}(A_{ig}) \qquad (4.12) \\
\lambda_t^{(1)}(A_i) &= \frac{1}{N_t(A_i)}\sum_{h=1}^{G_i} N_t(A_{ih})\lambda_t^{(1)}(A_{ih}) \\
\log\left(\lambda_t^{(1)}(A_{ig})\right) &= \frac{1}{|A_{ig}|}\int_{\mathbf{s}\in A_{ig}}\left(\mu + \sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(\mathbf{s})\boldsymbol{\eta}_{t,m,j}\right)d\mathbf{s} \\
&\approx \mu + \sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{t,m,j} + \epsilon_t(A_{ig}) \\
p(\mu) &\propto 1 \\
\epsilon_t(A_{ig}) &\overset{ind}{\sim} \gamma_t(A_{ig})N(0,\tau_1^2) + (1-\gamma_t(A_{ig}))N(c,\tau_2^2) \\
\gamma_t(A_{ig}) &\overset{iid}{\sim} \text{Beta}(1,1) \\
\tau_1^2 &\sim \text{IG}(1,1) \\
\tau_2^2 &\sim \text{IG}(1,1)
\end{aligned}
$$

with $m_t^{*(5)}(A_{ig}), m_t^{*(1)}(A_i), y_t^{*(5)}(A_{ig}), y_t^{*(1)}(A_i)$, defined in (4.10) for $i = 1, ..., I$ and $g = 1, ..., G_i$.

The prior distributions for the $\boldsymbol{\eta}_{t,m,j}$ are provided in (4.9). Using a Matérn covariance function as spatial covariance function for the separable, spatio-temporal Gaussian process $w_t(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, $t = 1, \ldots, T$, we specify prior distributions for the covariance function parameters $\boldsymbol{\theta} = \{\sigma^2, \phi, \nu\}$. For the parameter σ^2, commonly referred to in the spatial statistical literature as the marginal or spatial variance, we provide a non-informative Inverse Gamma prior distribution with both hyper-

parameters equal to 1. The remaining covariance function parameters, the range ϕ and the smoothness ν, are provided with discrete uniform prior distributions over the values $\{0.1, 0.5, 1, 5, 10, 25, 100, 200\}$ and $\{0.1, 0.5, 1.0, 1.5\}$ respectively. These prior specifications facilitate more efficient computation for these parameters; for further details see Section 4.2.5.

4.2.4 A spatio-temporal extension of the COSP model by Bradley, Wikle, and Holan

In this section, we adapt to the spatio-temporal setting the model proposed by Bradley et al. (2016b) to address the COSP when dealing with ACS estimates of count-valued characteristics. Hereafter, we will refer to this model as the "BWH model" after the authors Bradley, Wikle, and Holan. First, we note that in its initial formulation the BWH model postulates that count-valued ACS estimates follow a Poisson distribution with mean equal to the integral of a point-referenced spatial process over the areal unit of interest. In practice however, since the underlying, latent point-referenced spatial process is in turn assumed to be constant over some very fine areal units, typically the smallest areal units for which data are available, in the BWH model the mean of the Poisson distributions of the ACS estimates are expressed as linear combination of spatially-correlated areal-based random variables defined over appropriate spatial supports. Taking the census tracts A_{ig} as the fine areal units over which the latent point-referenced spatial process is constant, in a purely spatial setting, the BWH model states the following distribution for the ACS estimates, $W_{ACS}(A_{ig})$:

$$\hat{W}_{ACS}(A_{ig})|Y(A_{ig}) \overset{ind}{\sim} \text{Pois}(\exp(Y(A_{ig})))$$

To accommodate the BWH model to multi-year ACS estimates, consisting of 5-year census tract estimates $\hat{W}^{(5)}_{ACS,t}(A_{ig})$, and 1-year PUMA-level estimates, $\hat{W}^{(1)}_{ACS,t}(A_i)$, we assume that the 5-year estimates, in our case, the estimated average number of births in census tract g within PUMA i over the 5-year time period ending in time t, is a Poisson random variable whose mean is equal to the average number of births in census tract g within PUMA i during the 5-year time period. In a similar fashion, we postulate that the 1-year PUMA estimates $\hat{W}^{(1)}_{ACS,t}(A_i)$, that is the estimated number of births in PUMA i, is a Poisson random variable whose expected value is equal to the sum of the expected number of births in that year over all the census tracts contained in the PUMA. In other words, the spatio-temporal version of the BWH model states:

$$\begin{aligned}
\hat{W}^{(5)}_{ACS,t}(A_{ig}) &\sim \text{Pois}\left(\frac{1}{5}\sum_{k=t-4}^{t}\exp(Y_k(A_{ig}))\right) \\
\hat{W}^{(1)}_{ACS,t}(A_i) &\sim \text{Pois}\left(\sum_{h=i}^{G_i}\exp(Y_t(A_{ih}))\right)
\end{aligned}$$

Following the BWH model specification, the $Y_t(A_{ig})$ is in turn decomposed as:

$$Y_t(A_{ig}) = \beta_t + \boldsymbol{\psi}\boldsymbol{\eta} + \xi_t(A_{ig}) \tag{4.13}$$

with the term β_t introduced to account for the additional temporal dimension of the data. Other terms in (4.13) are specified as in Bradley et al. (2016b): thus.

the terms $\xi_t(A_{ig})$, $t = 1, \ldots, T$, are assumed to be independent and identically distributed replicates in time of Gaussian random variables with mean zero and variance equal to σ^2_γ while the variance σ^2_γ is assumed to follow an Inverse Gamma distribution with shape and scale parameters α_γ and β_γ. These two parameters are altered from their specification in Bradley et al. (2016b) to accommodate spatio-temporal data.

For the rest of the model, we preserve the same specification as in Bradley et al. (2016b). Specifically, we maintain the same basis function construction for the spatial random effects as in Bradley et al. (2016b) since that allows for dimension reduction, important when working with a large number of areal units, while retaining complexity in the spatial dependence structure of the random field. Letting n be the number of areal units at the finest spatial scale (e.g. the number of census tracts) and $r (r << n)$ the number of basis functions, we indicate with $\boldsymbol{\psi}$ the $n \times r$ set of Moran's I basis functions constructed from the Moran's propagator operator, whereas $\boldsymbol{\eta}$ indicate the $r \times 1$ vector of basis function weights. These weights are assumed to follow a multivariate Gaussian distribution with mean zero and covariance matrix $\mathbf{K}$, with $\mathbf{K}$ an $r \times r$ matrix equal to $\boldsymbol{\Phi}(\phi \times \boldsymbol{\Lambda}_Q)\boldsymbol{\Phi}'$. The matrix $\boldsymbol{\Phi}$ is constructed through a Givens rotator product (Kang and Cressie, 2011) of matrices $\mathbf{O}_{i,j}$:.

$$\boldsymbol{\Phi} \equiv (\mathbf{O}_{1,2} \times \mathbf{O}_{1,3} \times ... \times \mathbf{O}_{1,r}) \times (\mathbf{O}_{2,3} \times ... \times \mathbf{O}_{3,r}) \times ... \times \mathbf{O}_{r-1,r}$$

with $\mathbf{O}_{i,j}$ $r \times r$ identity matrix with the entries (i, i) and (j, j) replaced by $\cos(\theta_{i,j})$, entry (i, j) replaced by $\sin(\theta_{i,j})$, and entry (j, i) replaced by $-\sin(\theta_{i,j})$.

In turn, the angles $\theta_{i,j}$ are postulated to belong to the interval $[-\pi/2; \pi/2]$ and depend on the Givens angles $g_{i,j}(\mathbf{\Phi}_Q)$ through the following model:

$$\begin{aligned} \zeta_{i,j} &\equiv 1/2 + \theta_{i,j}/\pi \\ \text{logit}(\zeta_{i,j}) &= a + b \times g_{i,j}(\mathbf{\Phi}_Q) \end{aligned} \tag{4.14}$$

Finally, the Givens angles $g_{i,j}(\mathbf{\Phi}_Q)$ are the eigenvector of $\mathbf{\Psi'Q\Psi}$, with $\mathbf{Q}$ equal to $\text{diag}(\mathbf{A1}) - \mathbf{A}$, $\mathbf{1}$ an n-dimensional vector of all 1's and $\mathbf{A}$ adjacency matrix, that is, a sparse, symmetric matrix with entry (i, j) equal to 1 if census tracts i and j share a border and 0 otherwise.

Since the ACS estimates are provided with design-based variances and these variances are derived using the same information that yields the ACS estimates, in an effort to improve inference on the latent random field $Y_t(A_{ig})$, Bradley et al. (2016b) provide also a model for the design based variances $\tau^{2(5)}(A_{ig})$ and $\tau^{2(1)}(A_i)$ for the 5-year and 1-year ACS estimates respectively, which they link to $Y_t(A_{ig})$ via:

$$\begin{aligned} \log(\tau_t^{2(5)}(A_{ig})) &\sim N\left(\log\left(\frac{1}{5}\sum_{k=t-4}^{t}\exp(Y_k(A_{ig}))\right), \sigma_t^{2(5)}(A_{ig})\right) \\ \log(\tau_t^{2(1)}(A_i)) &\sim N\left(\log\left(\sum_{h=i}^{G_i}\exp(Y_t(A_{ih}))\right), \sigma_t^{2(1)}(A_i)\right) \\ \sigma_t^{2(5)}(A_{ig}) &\sim \text{IG}(1,1) \\ \sigma_t^{2(1)}(A_i) &\sim \text{IG}(1,1) \end{aligned}$$

Using the same prior distributions and hyperparameters as recommended by

Bradley et al. (2016b), the spatio-temporal extension of the BWH model that we will use in our case study is for $i = 1, ..., I$ and $g = 1, ..., G_i$:

$$
\begin{aligned}
\hat{W}^{(5)}_{ACS,t}(A_{ig}) &\sim \text{Pois}\left(\frac{1}{5}\sum_{k=t-4}^{t}\exp(Y_k(A_{ig}))\right) \\
\hat{W}^{(1)}_{ACS,t}(A_i) &\sim \text{Pois}\left(\sum_{h=i}^{G_i}\exp(Y_t(A_{ih}))\right) \\
\log(\tau_t^{2(5)}(A_{ig})) &\sim N\left(\log\left(\frac{1}{5}\sum_{k=t-4}^{t}\exp(Y_k(A_{ig}))\right), \sigma_t^{2(5)}(A_{ig})\right) \\
\log(\tau_t^{2(1)}(A_i)) &\sim N\left(\log\left(\sum_{h=i}^{G_i}\exp(Y_t(A_{ih}))\right), \sigma_t^{2(1)}(A_i)\right) \\
Y_t(A_{ig}) &= \beta_t + \boldsymbol{\psi}\boldsymbol{\eta} + \xi_t(A_{ig}) \\
\boldsymbol{\eta} &\sim N(0, \mathbf{K}) \\
\xi_t(A_{ig}) &\overset{iid}{\sim} N(0, \sigma^2_\gamma) \qquad (4.15) \\
p(\beta_t) &\propto 1 \\
p(a) &\propto 1 \\
p(b) &\propto 1 \\
\sigma_t^{2(5)}(A_{ig}) &\sim \text{IG}(1,1) \\
\sigma_t^{2(1)}(A_i) &\sim \text{IG}(1,1) \\
\sigma^2_\gamma &\sim \text{IG}(1,1)
\end{aligned}
$$

Note that in Bradley et al. (2016b), the parameters $\beta_t, t = 1, ..., T$, as well as the intercept and slope parameters a and b in (4.14), are all provided with normal prior distributions with mean zero and variance 10^{15}, rather than the flat priors we specify in our spatio-temporal extension provided in (4.15).

4.2.5 Computation

Both modeling frameworks are fit using a Markov Chain Monte Carlo algorithm. In the Poisson DEFF model, the efficiency of estimating ϕ and ν, the range and smoothness parameters, can be increased by defining a priori a grid of possible values and pre-computing the basis functions and prior covariance matrix of the basis function weights. During MCMC, we sample ϕ from the values (in kilometers) (0.1, 0.5 ,1 ,5 , 10, 25, 100, 200) and ν from the values (0.1, 0.5, 1.0, and 1.5) via a random-walk Metropolis-Hastings algorithm. Additional Metropolis-Hastings steps are required in order to generate a sample from the posterior distribution of all model parameters, with the exception of the parameters τ_1^2 and τ_2^2 in the Poisson DEFF model (see (4.12)), and $\sigma_t^{2(5)}(A_{ig})$, $\sigma_t^{2(1)}(A_i)$, and $\sigma_\gamma^2(A_{ig})$ in the BWH model (see (4.15)). Proposal variances are tuned at every 100^{th} iteration during burn-in in order to achieve acceptance rate approximately equal to 23.4% for parameter vectors such as, for example, $\boldsymbol{\eta} := \{\boldsymbol{\eta}_{t,m,j}\}_{t=1,\ldots,T,m=0,\ldots,M,j=1,\ldots,J^M}$, and 35% for single parameters such as μ in (4.12). The former acceptance rate of 23.4% was selected based on the asymptotic optimal acceptance rate proposed by Roberts et al. (1997). We used a burn-in period of 10,000 iterations, and ran the algorithm for an additional 10,000 iterations after burn-in. Convergence was assessed through visual inspection of trace plots and through the Geweke statistic (Geweke, 1992). In addition, we confirmed that each model's individual parameters have effective sample sizes greater than 1,000.

4.3 Results

4.3.1 Simulation results

We now present results from the application of our modeling framework to 30 simulated data sets which we generated by simulating estimates of counts for areal units contained in a spatial domain made of 100 subregions (analogous to census tracts, indexed by g) nested within 4 larger regions (analogous to counties or PUMAs, indexed by i). Estimates were generated over a period of 10 time points, $t = 1, ..., 10$, and were obtained by following this procedure: we first simulated the population sizes $N_t(A_{ig})$ for the smallest areal units. These population sizes were randomly generated using a two-step approach: first we generated the population sizes at time $t = 1$ from a discrete uniform distribution on the set $\{2000; 2001; \ldots; 8000\}$. Then, for each subsequent time point they were generated according to the following scheme: $N_t(A_{ig}) = [N_{t-1}(A_i g) + \delta_t(A_{ig})], \delta_t(A_{ig}) \overset{iid}{\sim} N(0, 200), i = 1, ..., 4, g = 1, ..., 25$. Having obtained the population sizes, we moved onto generating realizations for the rate parameter $\lambda(A_{ig})$ at the smallest areal units, according to the following model:

$$\begin{aligned} \lambda(A_{ig}) &= \exp\left(-4 + X(A_{ig}) + w(A_{ig})\right) \qquad i = 1, ..., 4; g = 1, ..., 25 \\ X(A_{ig}) &\overset{iid}{\sim} N(0, 0.25) \end{aligned}$$

with $w(A_{ig})$ realizations of a mean-zero Gaussian process with Matèrn covariance function with parameters $\sigma^2 = 1$, $\phi = 0.5$, and $\nu = 1$, generated at the centroids of each A_{ig}. The spatial covariance function parameters, σ^2, and ϕ are chosen so that the spatial process has a correlation that is almost null at at a distance larger

than $\sqrt{2}$, the largest distance between two-adjacent areal units.

To generate simulations that mimic what we observe in the ACS estimates, that is, an excess number of zero for some areal units, we introduce into our data generating mechanism a structural zero probability $\pi(A_{ig}), i = 1, ..., 4, g = 1, ..., 25$, which is distributed according to a Beta distribution with shape and scale parameters equal to 0.001 and 0.01, respectively. These parameter values for the Beta distribution were chosen so that the structural zero probabilities were either close to zero or close to one, with the large majority being close to zero. To determine subregions with zero-valued counts, at each time point $t = 1, ..., 10$, a Bernoulli random variable, $z_t(A_{ig})$ is generated corresponding to each location A_{ig}, $i = 1, ..., 4$, $g = 1, ..., 25$, with probability $\pi(A_{ig})$.

Finally, having generated the population sizes $N_t(A_{ig})$ and the rate parameters $\lambda(A_{ig})$, for $t = 1, \ldots, 10$, $g = 1, \ldots, 25$ and $i = 1, \ldots, 4$, we simulate the true counts, $W_t(A_{ig})$ according to:

$$p(W_t(A_{ig})) = \begin{cases} \text{Pois}(N_t(A_{ig})\lambda(A_{ig})) & \text{if } z_t(A_{ig}) = 0 \\ 0 & \text{if } z_t(A_{ig}) = 1 \end{cases} \tag{4.16}$$

The decision to keep $\lambda(A_{ig})$ constant over time was made with the intention of simulating data based on a process that does not directly imitate our modeling framework. Although $\lambda(A_{ig})$ is constant in space, the $W_t(A_{ig})$ still vary temporally due to: (i) the temporally varying $N_t(A_{ig})$, and (ii) the fact that they are randomly generated according to the Poisson distribution described in (4.16).

To mimic the 5- and 1-year ACS estimates, the "pseudo-estimates" $\hat{W}^{(5)}_{ACS,t}(A_{ig})$ and $\hat{W}^{(5)}_{ACS,t}(A_{ig})$, $i = 1, \ldots, 4; g = 1, \ldots, 25$, are then obtained by aggregating the true counts $W_t(A_{ig})$ over time and space, adding to them random errors. Specifically, we add $\xi^{(5)}_t(A_{ig})$ to the true counts at the smallest areal unit level and $\xi^{(1)}_t(A_i)$ to the true counts at the coarser resolution. Standard deviations of these random errors are chosen to reflect the ACS standard errors for the number of births, our case study. Hence:

$$\begin{aligned}
\hat{W}^{(5)}_{ACS,t}(A_{ig}) &= \left[\frac{1}{5}\sum_{k=t-4}^{t} W_k(A_{ig}) + \xi^{(5)}_t(A_{ig})\right] \\
\xi^{(5)}_t(A_i) &\overset{iid}{\sim} N(0, 100) \qquad t = 5, ..., 10 \\
\hat{W}^{(1)}_{ACS,t}(A_i) &= \left[\sum_{h-1}^{25} W_k(A_{ih}) + \xi^{(1)}_t(A_i)\right] \\
\xi^{(1)}_t(A_i) &\overset{iid}{\sim} N(0, 10000) \qquad t = 1, ..., 10
\end{aligned}$$

After truncating these pseudo-estimates at zero, we use them as data in our modeling framework. We then generate disaggregated estimates of the counts using our model and we compare such estimates to the true values at the 1-year subregion resolution (corresponding in practice to the 1-year census tract resolution). Specifically, we report the percentage of times a 95% credible interval for the disaggregated value covers the true value, the mean absolute error, and the average length of a 95% credible interval.

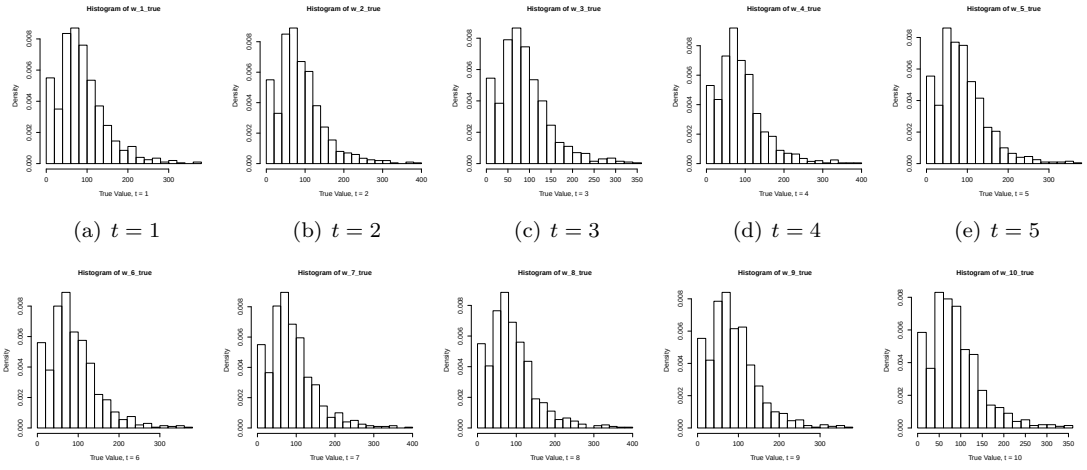
Table 4.1 presents summary statistics for the true counts over the 30 simulated datasets at times $t = 1, ..., 10$, with the columns "Minimum" and "Maximum" presenting respectively the minimum and maximum values that were generated

t	Mean	SD	Minimum	Maximum
1	86.4	58.9	0	372
2	87.1	59.6	0	390
3	87.0	59.3	0	348
4	87.0	60.0	0	383
5	87.2	59.7	0	370
6	87.0	59.7	0	375
7	87.2	60.5	0	383
8	87.1	60.7	0	394
9	87.1	60.3	0	378
10	87.4	61.0	0	358

Table 4.1: Summary statistics for the true counts across 30 simulated data sets: mean and standard deviation averaged over 30 simulated data sets; minimum and maximum values generated for all 30 data sets.

over all 30 simulated data sets. The true counts range from 0 to just under 400, with means around 87 and standard deviations around 60. Figure 4.1 presents histograms of the simulated values for each time point, using data from all 30 simulated datasets.

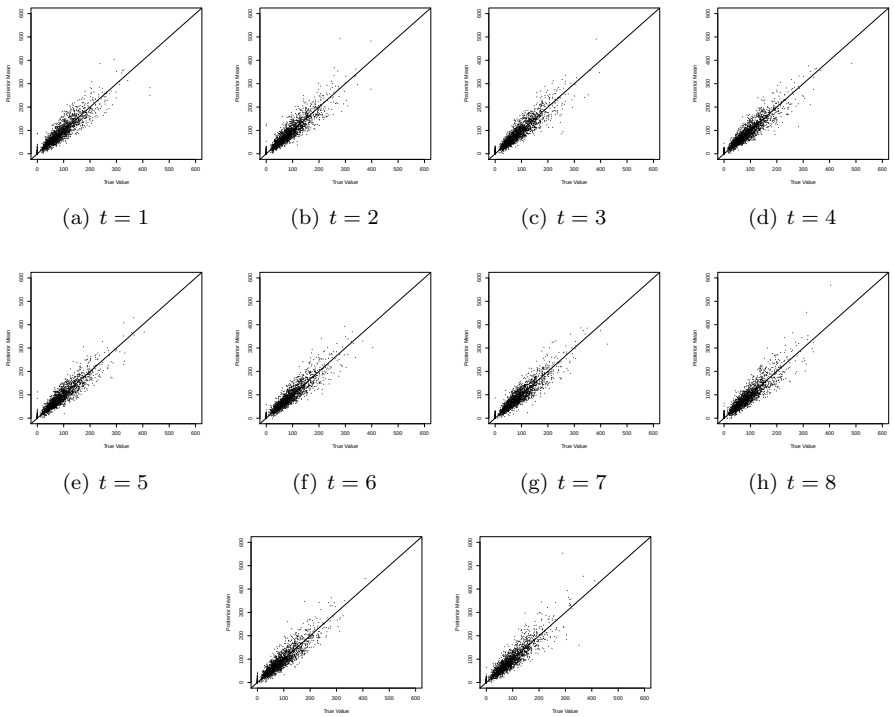
Figure 4.2 presents a scatterplot of our disaggregated estimates, denoted as $\hat{W}_t(A_{ig})$, which we take as the posterior mean of the $W_t(A_{ig})$, against the true counts for all times $t = 1, ..., 10$. As the figure shows, points tend to fall on the identity line, indicating successful recovery of the true counts. On the other hand, Table 4.2 presents, for each time point, the average coverage probability of the 95% credible intervals, averaged across the 30 simulated datasets, the Mean Absolute Error (MAE), and the average length of 95% credible intervals, averaged across the 30 simulated datasets. Table 4.3 presents similar statistics only for non-zero data. In addition, we present the Average Symmetric Mean Absolute Relative Error (SMARE), defined as:



(a) $t = 1$ (b) $t = 2$ (c) $t = 3$ (d) $t = 4$ (e) $t = 5$

(f) $t = 6$ (g) $t = 7$ (h) $t = 8$ (i) $t = 9$ (j) $t = 10$

Figure 4.1: Histograms of the true counts for 30 simulated data sets at each of 10 time points.



(a) $t = 1$ (b) $t = 2$ (c) $t = 3$ (d) $t = 4$

(e) $t = 5$ (f) $t = 6$ (g) $t = 7$ (h) $t = 8$

(i) $t = 9$ (j) $t = 10$

Figure 4.2: Scatter plots of true counts vs. posterior means of the disaggregated counts for 30 simulations at each time point.

$$SMARE = \frac{1}{1000}\sum_{t=1}^{10}\sum_{i=1}^{4}\sum_{g=1}^{25}\frac{|\hat{W}_t(A_{ig}) - W_t(A_{ig})|}{(|\hat{W}_t(A_{ig})| + |W_t(A_{ig})|)/2} \tag{4.17}$$

where again $\hat{W}_t(A_{ig})$ is the estimated value of $W_t(A_{ig})$, $t = 1, ..., 10$, $i = 1, ..., 4$, $g = 1, ..., 25$. In (4.17), the denominator of 1,000 is the product of the number of time points times the number of subregions for a single simulation.

We note that we do not include the average SMARE in Table 4.2 due to the denominators of the statistic for certain census tracts and time points being zero-valued.

In Table 4.2, coverage probabilities are somewhat low (all around 85% compared to the 95% nominal level) but reasonably close to nominal coverage. The MAE seems to indicate that our model is able to disaggregate the estimates over space and time and recover the true counts. When we focus only on the performance of the model relative to the non-zero-counts, we can see, as indicated in Table 4.3, that our model achieves near nominal coverage for the 95% credible interval of the true counts at the smallest spatial and temporal resolution.

4.3.2 Data analysis results: number of births in Michigan, 2006 – 2016

The results that follow refer to the application of the Poisson DEFF model detailed in Section 4.2.3 to the ACS datasets on the number of births in Michigan from 2006-2016. We leverage 1-year and 5-year data in order to generate generate estimates of the number of births at the 1-year census tract level. For the analysis of the number of births, we use as a proxy for $N_t(A_{ig})$ the 5-year ACS estimates of the number of women aged 15-50. We acknowledge that one may reasonably

t	Coverage Probability	Average MAE	Average CI Length
1	0.824	17.1	60.5
2	0.832	16.8	60.0
3	0.840	16.6	59.6
4	0.844	16.2	59.4
5	0.852	16.1	59.5
6	0.850	16.1	59.7
7	0.843	16.6	59.9
8	0.819	17.7	60.9
9	0.822	17.3	59.8
10	0.824	17.2	60.3

Table 4.2: Statistics pertaining to the recovery of true values for all randomly generated data (zero- and non-zero-valued): average coverage probability for the 95% credible intervals of the true counts, average Mean Absolute Error, average length of the 95% credible interval. All statistics pertain to the indicated time point $t = 1, ..., 10$ and is the average across the 30 simulated datasets.

t	Coverage Probability	Average MAE	Average $SMARE$	Average CI Length
1	0.913	18.3	0.211	66.3
2	0.923	17.9	0.207	65.7
3	0.929	17.9	0.205	65.1
4	0.934	17.3	0.203	64.9
5	0.945	17.2	0.200	65.1
6	0.942	17.3	0.200	65.5
7	0.934	17.8	0.208	65.6
8	0.908	18.9	0.221	66.6
9	0.911	18.7	0.223	65.5
10	0.913	18.4	0.214	66.0

Table 4.3: Statistics pertaining to the recovery of true values for randomly generated data that are non-zero: average coverage probability for the 95% credible intervals of the true counts, average Mean Absolute Error, average Symmetric Mean Absolute Relative Error, and average length of the 95% credible interval. All statistics pertain to the indicated time point $t = 1, ..., 10$ and is the average across the 30 simulated datasets.

criticize these proxies, as they refer to a 5-year temporal resolution and are survey-based estimates whose variance we do not incorporate. An alternative would be to use census tract population sizes from the 2010 decennial census, which may more reasonably be treated as fixed.

Exploratory Analysis

In this section, we provide an exploratory analysis of the seven (7) 5-year census tract data sets that we utilize for the data analysis. Table 4.4 presents various summary statistics for each dataset, including the mean and standard deviation of the estimated number of births, the average ACS standard error, the minimum and maximum number of births, and the percentage of census tracts for which the estimated number of births is zero. The average number of births per census tract has been decreasing over the 11-year duration of our study, despite Michigan's population having an overall positive trend over the last decade. In addition, there is a decrease in the percentage of census tracts with 0 estimated births, from 7.3% in the time period 2006–2010, to 4.7% and 4.8% in the time periods 2011–2015 and 2012–2016 respectively. It is possible that this is indicative of the improved quality of the ACS sampling and estimation over time rather than population trends.

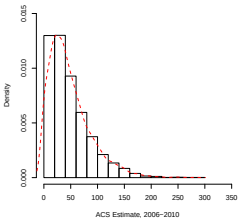
Figure 4.3 presents histograms and a boxplot of the ACS estimates of the number of births for census tracts in Michigan. Each one indicates right skewness with maxima between 240 and 360. Figure 4.4 presents histograms of the ACS standard errors. The standard errors for all sets of 5-year estimates have a modal standard error value that is observed in approximately 200 census tracts. This is

Time Period	Mean number of births	SD number of births	Mean ACS standard error	Minimum number of births	Maximum number of births	Percentage zero-valued estimates
2006–2010	47.8	38.8	27.9	0	282	7.3%
2007–2011	46.6	38.0	26.0	0	256	7.1%
2008–2012	45.3	36.5	22.4	0	245	6.4%
2009–2013	45.2	37.7	22.2	0	333	6.1%
2010–2014	45.0	37.2	21.9	0	325	5.5%
2011–2015	44.0	36.5	21.4	0	323	4.7%
2012–2016	43.8	36.5	21.2	0	352	4.8%

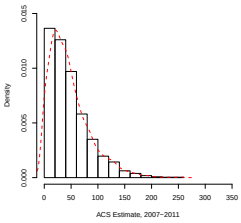
Table 4.4: Summary statistics for the 5-year ACS estimates and standard errors: Mean number of births; standard deviation of the number of births; mean ACS standard error; minimum number of births; maximum number of births; percentage of estimates that are zero-valued.

most noticeable in Figure 4.4 for the 2006–2010 and the 2007-2011 estimates. For the 2006–2010 estimates, 208 census tracts have a standard error of 66.3, and for the 2007–2011 estimates, 206 census tracts have a standard error of 47.4. Among those 206 census tracts that have modal standard error in the 2007–2011 dataset, 129 (63%) have model standard error in the 2006–2010 data set as well. Figure 4.5 plots the census tracts with modal standard errors in both 2006-2010 and 2007-2011. These tracts often have zero-valued estimates for the number of births. We speculate that in these cases an insufficient number of surveys were administered within those census tracts and time periods to provide valid estimates.

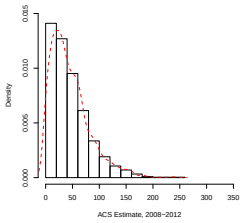
Figure 4.6 presents a scatter plot matrix of the ACS estimates for all seven 5-year time periods that are available, with Pearson correlation coefficients in the lower triangle. Zero-valued estimates are highlighted in red. The plots demonstrate a high degree of discordance in ACS estimates that differ by more than two years, a phenomenon that is especially true for zero-valued estimates. That is, many ACS estimates that are zero-valued for one 5-year year time period are not zero-valued



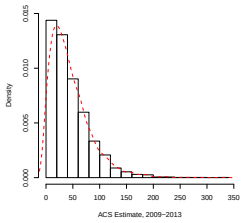
(a) 2006–2010



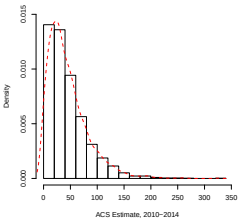
(b) 2009–2011



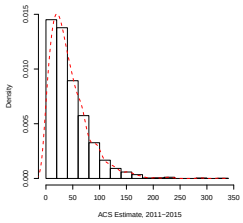
(c) 2008–2012



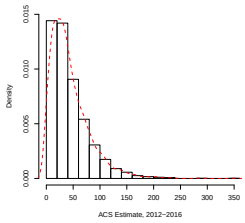
(d) 2009–2013



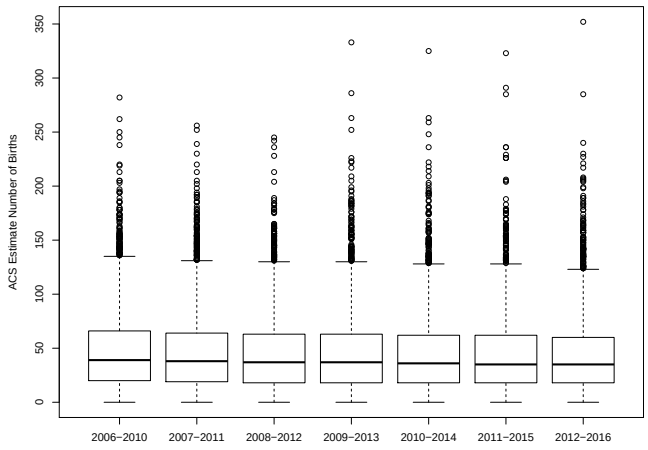
(e) 2010–2014



(f) 2011–2015

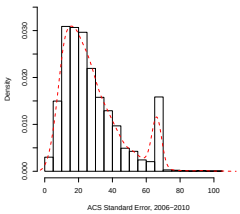


(g) 2012–2016

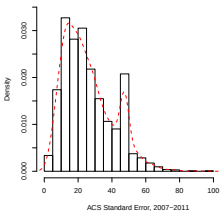


(h) Boxplots of ACS Estimates

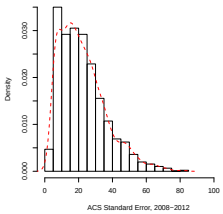
Figure 4.3: Histograms and boxplots of the 5-year ACS estimates for all seven 5-year time periods. In panels (a)–(g), the density of the ACS estimates is overlaid in red.



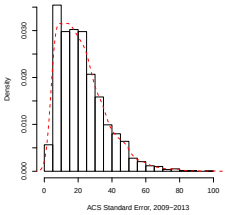
(a) 2006–2010



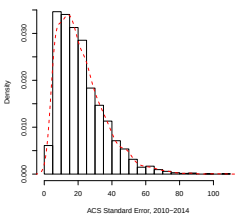
(b) 2007–2011



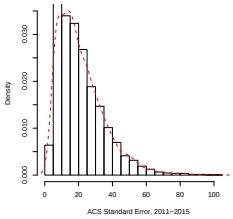
(c) 2008–2012



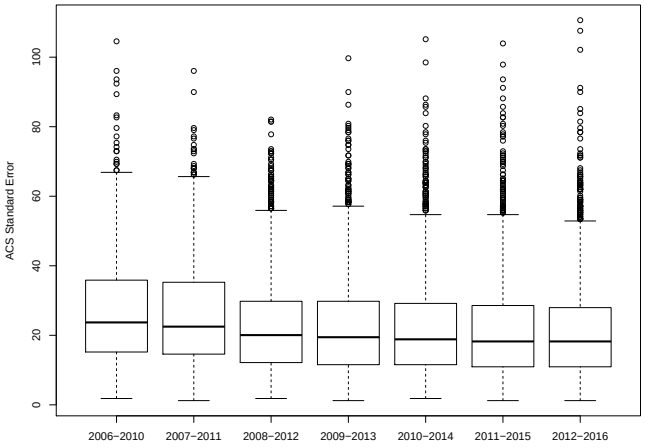
(d) 2009–2013



(e) 2010–2014



(f) 2011–2015



(g) Boxplots of ACS Estimates

Figure 4.4: Histograms and boxplots of the standard errors of the 5-year ACS estimates. In panels (a)–(g), the density of the ACS standard errors is overlaid in red.

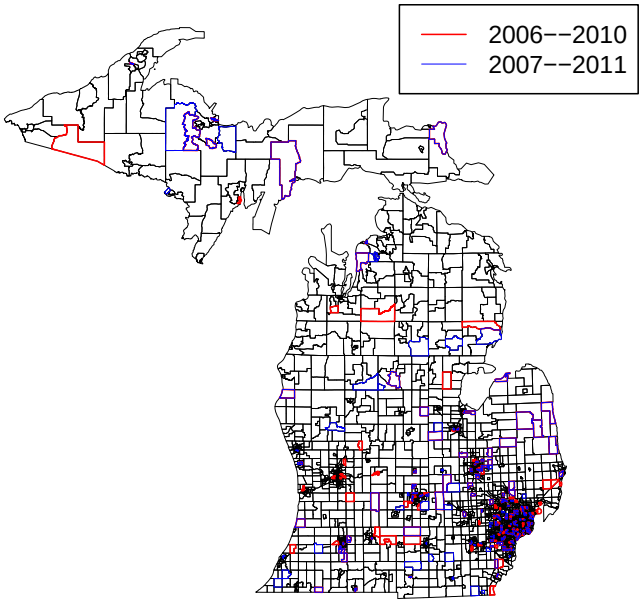


Figure 4.5: Census tracts with modal standard errors in the 2006–2010 and 2007-2011 ACS datasets. Note that census tracts with modal standard errors in both time periods appear as purple due to the overlay of blue and red lines.

in that 5-year data set that immediately follows, for example from 2006–2010 to 2007–2011.

Validation and Comparison

We begin by comparing our 5-year estimates at the census tract level, i.e. the posterior means of the disaggregated, model-based estimates to the ACS estimates at the census tract level for the 5-year period from 2009-2013. To do so, we average the disaggregated estimates over the years 2009-2013, ensuring that we are comparing model based and ACS estimates for the same time period. These results are presented in Figure 4.7. Starting with panel (a), we can observe that our estimates approximately resemble the ACS estimates. In Figure 4.7 (a), discordant zero-valued tracts are highlighted in red. Discordance is defined as any census tract whose ACS estimate was zero-valued in 2009-2013 and non-zero in either 2008-2012 or 2010-2014. The converse is included as well, that is, any census tract whose ACS estimate was non-zero in 2009-2013 and zero-valued in either 2008-2012 or 2010-2014. When a census tract has ACS estimates that are zero-valued for only a small number of time periods, i.e. those classified as "discordant zeros" in Figure 4.7 (a), the posterior means of the disaggregated estimates from our model tend to be dispersed away from zero. Conversely, census tracts whose ACS estimates are consistently zero-valued over time, i.e. those that are classified as "concordant zeros" and denoted in green in Figure 4.7, have posterior means that are consistently 0, indicating success of the mixture prior on $\epsilon_t(A_{ig})$, $\forall i, g$ in shrinking those tracts to zero. In Figure 4.7 (b), we once again present the poste-

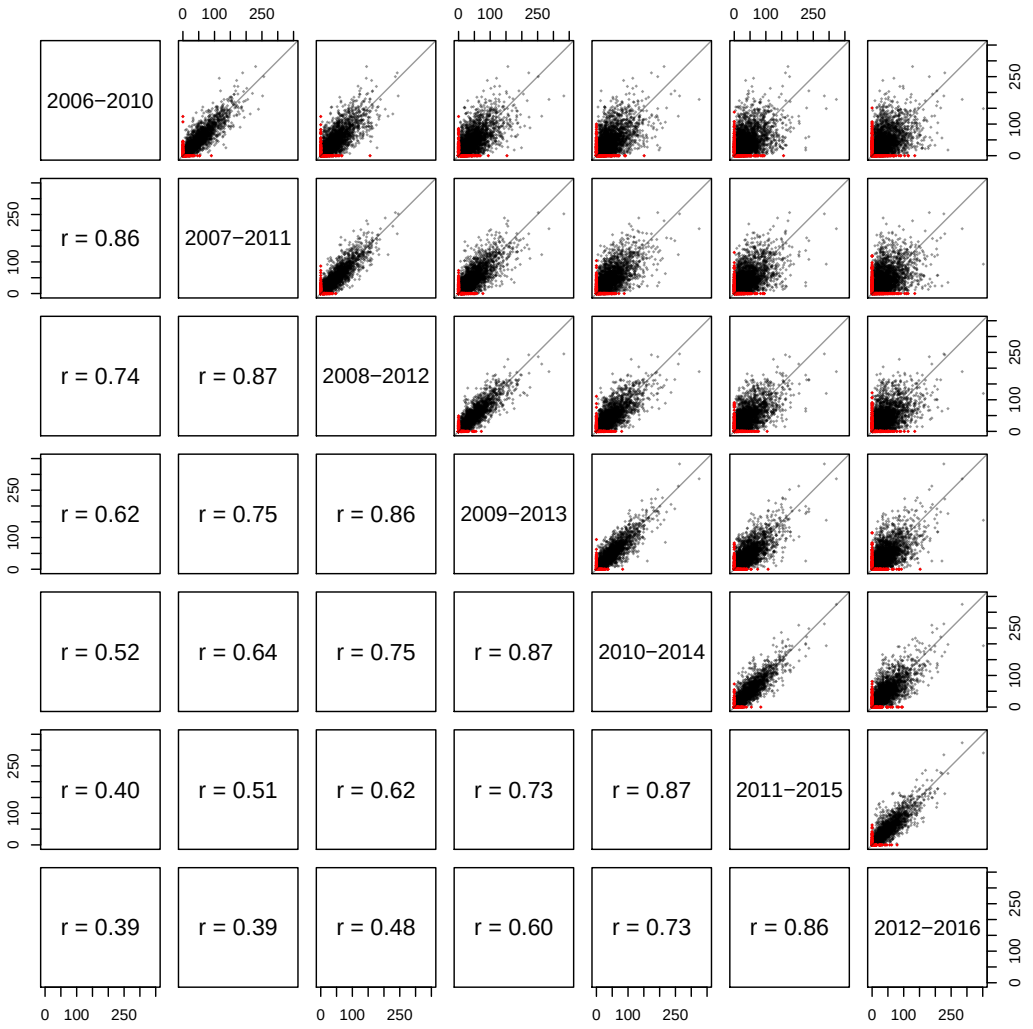
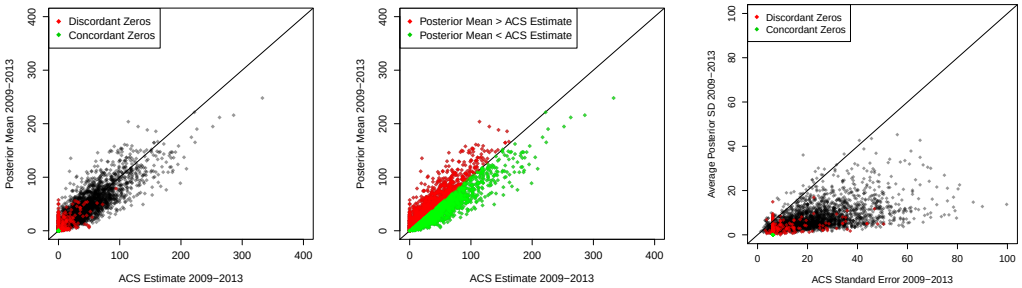


Figure 4.6: Scatter plot matrix of ACS 5-year estimates for each of seven 5-year time periods available. Zero-valued estimates are highlighted in red.



(a) ACS Estimate vs Average Posterior Mean.

(b) ACS Estimate vs Average Posterior Mean

(c) ACS SE vs Average Posterior SD

Figure 4.7: (a) ACS estimates of the number of births for 2009-2013 vs. the average posterior means of the disaggregated estimates for the same time period with zero-valued ACS estimates highlighted; (b) ACS estimates of the number of births for 2009-2013 vs. the average posterior means of the disaggregated estimates for the same time period with indicator of whether the ACS Estimate is greater than the posterior mean; (c) ACS standard errors vs. the posterior standard deviations.

rior means of the disaggregated estimates from our model, this time highlighting whether or not the model-based estimate is greater or less than the corresponding ACS estimate. Forty-nine point one percent of the model-based estimates are greater than the corresponding ACS estimate, while 50.9% are less, indicating no systematic over- or under-estimation. Figure 4.7 (c) plots the ACS standard error of the number of births for the time period 2009-2013 vs. the posterior standard deviations for the same time period. It indicates that the model-based estimates tend to have lower posterior variances than the ACS variances.

Table 4.5 presents results pertaining to the parameter $\gamma_t(A_{ig})$, the probability that census tract A_{ig} has a non-zero number of births at time t. Rows of the table are determined based on the number times in the study period (2006–2016) a census tract exhibits a zero-valued 5-year census tract estimate. For example, census tracts grouped into the first row of Table 4.5 have non-zero values for all seven

Number of zero-valued ACS Estimates	Number of Census Tracts	Average posterior mean $\gamma_t(A_{ig})$	SD posterior mean $\gamma_t(A_{ig})$	Minimum posterior mean $\gamma_t(A_{ig})$	Maximum posterior mean $\gamma_t(A_{ig})$
0	2,233	1.00	0.00	1.00	1.00
1	217	0.80	0.40	0.00	1.00
2	108	0.64	0.48	0.00	1.00
3	74	0.57	0.49	0.00	1.00
4	38	0.51	0.50	0.00	1.00
5	26	0.36	0.48	0.00	1.00
6	8	0.27	0.45	0.00	1.00
7	22	0.00	0.00	0.00	0.00

Table 4.5: Summary statistics pertaining to the of the posterior means of $\gamma_t(A_{ig})$ over the 11 years of the study. $\gamma_t(A_{ig})$ denotes the probability that census tract A_{ig} has non-zero-valued count at time t. Statistics are grouped by the number of zero-valued 5-year census tract ACS estimates and contain the average posterior mean, standard deviation of the posterior means, minimum posterior mean, and maximum posterior mean.

of the 5-year ACS estimates, and so on. We present the average posterior means over years 2006–2016 of $\gamma_t(A_{ig})$, along with the standard deviation of the posterior means of $\gamma_t(A_{ig})$, and the minimum and maximum posterior mean $\gamma_t(A_{ig}), \forall i, g$. Focusing on the average posterior mean, Table 4.5 shows a clear gradient relationship between the number of zero-valued ACS estimates that a census tract has over the 11-year study period and the probability that our disaggregated estimates are zero-valued. Furthermore, when a census tract has no zero-valued ACS estimates, $\epsilon_t(A_{ig})$ is never drawn from the second mixture component, meaning $\lambda_t(A_{ig})$ is not shrunk to zero. Alternatively, when a census tract has all zero-valued estimates, its corresponding rate parameter $\lambda_t(A_{ig})$ is always shrunk towards zero.

Comparison of Out-of-Sample Prediction to the BWH Model

To assess the out-of-sample predictive performance of our model, we utilize 3-year ACS estimates of 64 Michigan counties with populations greater than 20,000

for the time periods of 2006-2008, 2007-2009, and so on through 2011-2013. This results in a total of 384 out-of-sample points for which we perform prediction. Furthermore, we compare our model to that of the BWH model presented in Section 4.2.4. Table 4.6 presents the average over the 64 counties and 6 time periods of the Symmetric Relative Bias, Mean Absolute Error, Symmetric Mean Absolute Relative Error, empirical probability that a 95% credible interval covers the 3-year ACS estimate, and average posterior predictive standard deviations. With the exception of the latter two, these metrics are obtained by comparing the posterior predictive mean of the 3-year county-level number of births to the corresponding 3-year ACS estimates. Letting $W_t^{(3)}(C_i)$ denote the ACS estimate for the number of births in county C_i for the 3-year time period ending in time t, and $\hat{W}_t^{(3)}(C_i)$ denote the posterior predictive mean of county C_i for the 3-year time period ending in time t, the Symmetric Mean Relative Bias and Symmetric Mean Absolute Relative Error are defined as follows:

$$\begin{aligned} SMRB &= \sum_{i=1}^{64} \frac{\hat{W}_t^{(3)}(C_i) - W_t^{(3)}(C_i)}{(|\hat{W}_t^{(3)}(C_i)| + |W_t^{(3)}(C_i)|)/2} \\ SMARE &= \sum_{i=1}^{64} \frac{|\hat{W}_t^{(3)}(C_i) - W_t^{(3)}(C_i)|}{(|\hat{W}_t^{(3)}(C_i)| + |W_t^{(3)}(C_i)|)/2} \end{aligned}$$

Table 4.6 indicate relatively similar performance in prediction accuracy, with the BWH model slightly having better performance for Wayne and Oakland, the Michigan counties with ACS estimates greater than 13,000 births, and the Poisson DEFF model having slightly better performance for the remaining 62 other counties (averaged over all six 3-year time periods). For Oakland and Wayne counties, the average relative error of the Poisson DEFF model is 0.009, whereas for the

BWH model it is 0.008. The fact that these relative errors are both less than 1% indicate that both models perform very well for these counties. More generally, Figure 4.8 indicates that both models are quite accurate, with the plot of predicted vs. ACS estimate falling on the identity line for both models. This notion of good predictive accuracy is further supported by the absolute relative errors being around 3% for both methods. Both models have posterior predictive standard deviations that fall well below the ACS standard errors, with the average posterior predictive standard deviation of the Poisson DEFF model being greater than that of the BWH model. Finally, the empirical probability that a 95% prediction covers the true value is close to nominal for the Poisson DEFF model, whereas for the BWH model the empirical coverage is extremely low. This implies that our modeling framework properly propagates the uncertainty in the ACS estimates, quantified in the data by the ACS design-based variance, by incorporating the design effect.

We wish to emphasize at this time that the modeling framework we are designating as the BWH model, specifically our spatio-temporal extension to the model presented in Bradley et al. (2016b) might not be the one the authors would have proposed had they been presented with the same problem. Therefore, we are not willing to state that the low coverage probability we observe in this model would necessarily occur if the authors were to analyze these data. Rather, we intend for the main findings of this section to be as follows:

1. The predictive accuracy of our model is similar to, and for most counties slightly better than that of the BWH model, which is at this time the gold

Subset	Model	SMRB	SMARE	Coverage	PPSD
Full Data	PD*	−0.019	**0.032**	**0.881**	68.1
	BWH	**−0.013**	0.032	0.395	39.5
$W_t^{(3)}(C_i) < 5000$	PD	−0.020	**0.032**	**0.898**	48.6
	BWH	**−0.012**	0.033	0.422	31.4
$5000 \leq W_t^{(3)}(C_i) < 13000$	PD	**−0.008**	**0.009**	**0.902**	212.9
	BWH	−0.038	0.013	0.239	104.3
$W_t^{(3)}(C_i) \geq 13000$	PD	−0.026	0.009	**0.877**	453.7
	BWH	**−0.021**	**0.008**	0.267	185.0

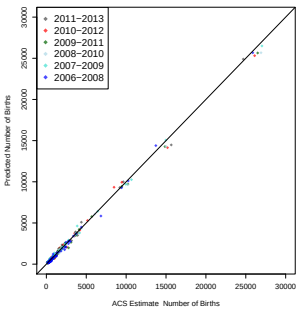
Table 4.6: Prediction results comparing the Poisson DEFF model to the BWH Model: Symmetric Mean Relative Bias, Mean Absolute Error, Symmetric Mean Absolute Relative Error, and empirical probability that a 95% credible interval covers the 3-year county-level ACS estimate, average posterior predictive standard deviation. *The abbreviation "PD" denotes results pertaining to our model: Poisson model with design effect.

standard for addressing the change of support problem for ACS data.

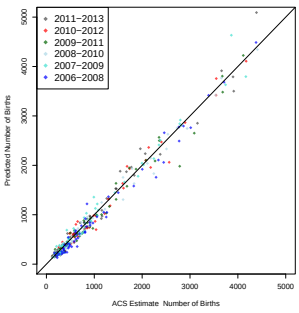
2. By incorporating the design effect into our modeling framework, our 95% prediction intervals achieve near nominal coverage when predicting out-of-sample 3-year county-level estimates.

Hospital demand in Michigan

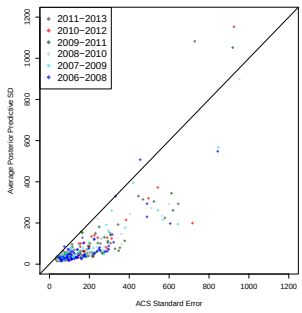
We have interest in using these results to demonstrate where resources may need to be allocated in order to address unmet needs pertaining to maternity in Michigan. Specifically, we want to identify which hospitals experience excess demand relative to their available resources. To this end, we start in Figure 4.9 by plotting the locations of the 145 general practice hospitals in Michigan. The color code for a hospital is based on the number of births in 2012 expected at each hospital based on census tract proximity to each hospital. Specifically, the expected number of births at a hospital is derived at each MCMC iteration after burn-in as the sum of the disaggregated number of births in census tracts whose centroids are closest to that hospital. Figure 4.9 (a) shows the posterior mean number of births



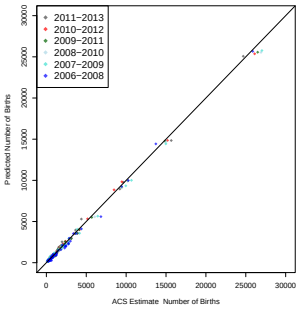
(a) $\hat{W}^{(3)}_{ACS,t}(A_c)$ vs $W^{(3)}_{ACS,t}(A_c)$: Pois. DEFF



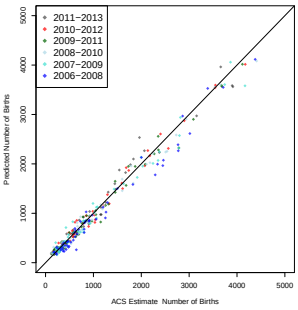
(b) $\hat{W}^{(3)}_{ACS,t}(A_c)$ vs $W^{(3)}_{ACS,t}(A_c)$: Pois. DEFF, $\hat{W}^{(3)}_{ACS,t}(A_c) < 5$k



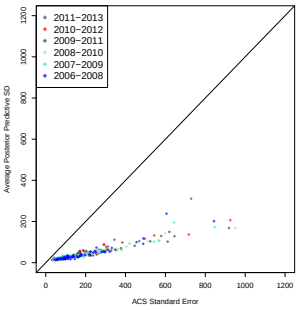
(c) Posterior predictive SD vs. ACS SE: Pois. DEFF



(d) $\hat{W}^{(3)}_{ACS,t}(A_c)$ vs $W^{(3)}_{ACS,t}(A_c)$: BWH model.



(e) $\hat{W}^{(3)}_{ACS,t}(A_c)$ vs $W^{(3)}_{ACS,t}(A_c)$: Pois. DEFF, $\hat{W}^{(3)}_{ACS,t}(A_c) < 5$k



(f) Posterior predictive SD vs. ACS SE: BWH model.

Figure 4.8: Posterior predictive means and standard deviations for the 3-year county level number of births against the corresponding ACS estimates and standard errors for our model and the BWH model. Panels (a) and (d) display the posterior predictive means vs. ACS estimates for (a) the Poisson DEFF model and (d) the BWH model. Panels (b) and (e) present the same results as (a) and (d) respectively, but zoomed in on 3-year county-level ACS estimates less than 5,000. Panels (c) and (f) display the posterior predictive standard deviation against the ACS standard errors for (c) the Poisson DEFF model and (f) the BWH model.

for each hospital and (b) the posterior standard deviation. Whereas the number of births is used as a proxy for demand, we utilize publicly available data on the number of beds per hospital as a proxy for each hospital's available resources. Figure 4.9 panels (c) and (d) plot the posterior mean and standard deviation respectively of the number of births per bed in the hospital.

We note that not everyone elects to use their nearest general practice hospital to give birth, be it based on simply haversine distance or rather driving distance. Furthermore, not every bed in a hospital is occupied by a person giving birth. Rather, both of these metrics represent rudimentary proxies for demand (the number of births) and hospital's capacity to meet that demand (number of beds).

Based on these metrics, one can see that urban areas see the greatest demand in terms of number of births occurring in close proximity to the hospital (Figure 4.9 (a) and (b)). However, relative to the amount of resources available to the hospital, quantified by the number of beds, it is clear that hospitals in more rural areas of Michigan, such as the northern lower peninsula, have very high demand as well (Figure 4.9 (c) and (d)).

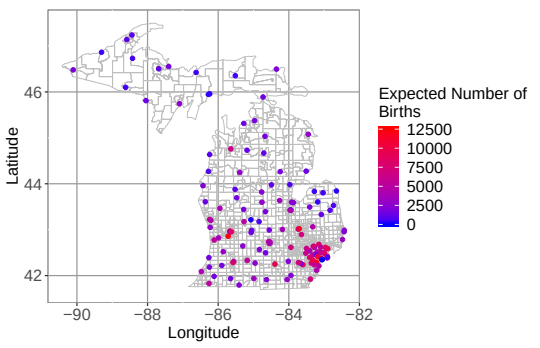
In order to better understand the distribution of the number of births per bed, as well as to illustrate the utility of casting the ACS estimates into a Bayesian hierarchical model, Figure 4.10 presents results pertaining to the posterior distribution, not just the mean and standard deviation, of the number births per bed for hospitals in Michigan. Panels (a) and (b) present maps of the lower and upper limits of a 95% credible interval obtained by taking the 2.5^{th} and 97.5^{th} percentile of the posterior samples of the number of births per bed. Panel (c) presents pos-

terior density of the number of births per bed for the Kalkaska County Health Center, which exhibits both high posterior mean and standard deviation in terms of the number of births per bed.

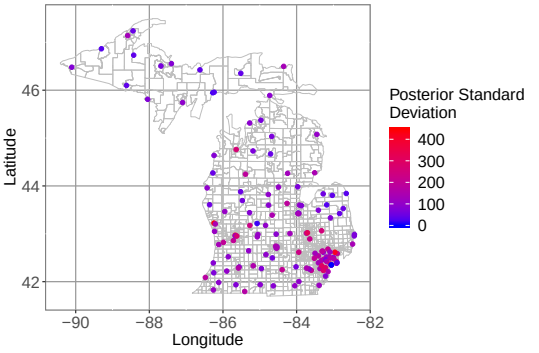
4.4 Discussion

This chapter has presented a modeling framework for disaggregation of spatio-temporal estimates of count-valued community characteristics that are derived from sampling surveys. Our model differs from that of Bradley et al. (2016b) in two ways. First, we accommodate spatio-temporal estimates of count-valued estimates, whereas the model presented in Bradley et al. (2016b) is purely spatial. Second, we incorporate into our modeling framework the survey's design effect in order to propagate the inflated design-based variance of the ACS estimates. Our simulation and out-of-sample prediction results demonstrate that this modeling decision results in near nominal coverage of, respectively, credible and prediction intervals. Furthermore, the predictive accuracy of our model is on par with that of the model of Bradley et al. (2016b). Finally, we illustrate the practical utility of our method by modeling the expected number of births per bed at 145 general practice hospitals in Michigan. We contend that, although this metric is rather rudimentary, our model could help identify hospitals with unmet needs, an application that is well-suited to a model for count-valued characteristics.

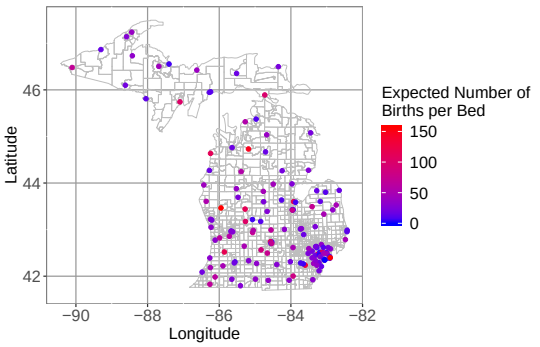
In order to account for an excess number of zero-valued estimates, we specified a straightforward mixture prior on the term $\epsilon_t(A_{ig})$, the residual term that was initially included to account for errors due to spatio-temporal aggregation.



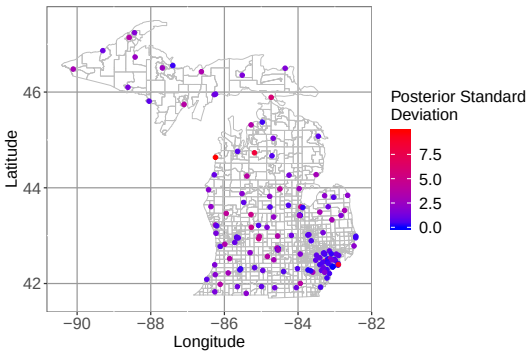
(a) Posterior mean number of births.



(b) Posterior SD number of births.

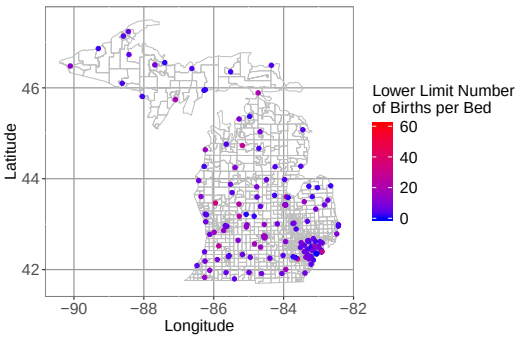


(c) Posterior mean number of births per bed.

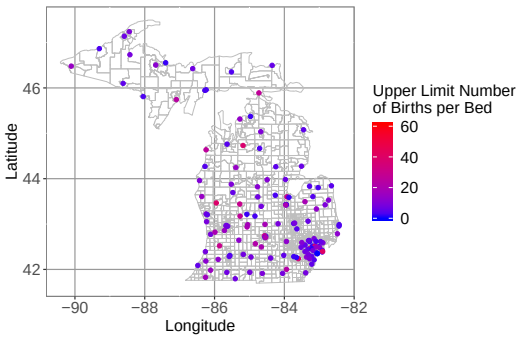


(d) Posterior SD number of births per bed.

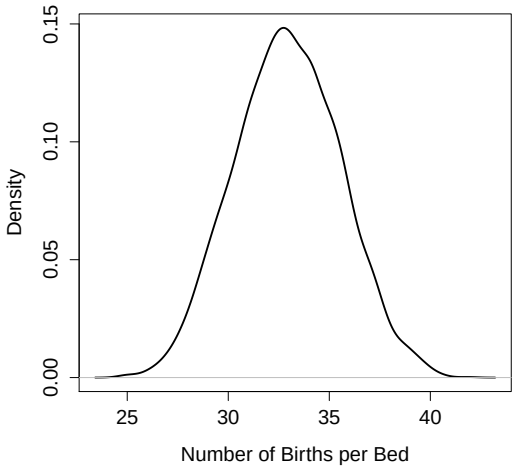
Figure 4.9: Plots pertaining to the expected number of births in 2010 for 145 general practice hospitals in Michigan: (a) Expected number of births; (b) posterior standard deviation of the number of births; (c) expected number of births per bed; (d) posterior standard deviation of the number of births per bed. The expected number of births is derived at each MCMC iteration after burn-in by identifying for each hospital, which census tracts have population centroids that are closest to that hospital, and then taking the sum of the number of births occurring in those census tracts and attributing them to the hospital.



(a) 2.5^{th} percentile, number of births per bed.



(b) 97.5^{th} percentile, number of births per bed.



(c) Posterior distribution, number of births per bed.

Figure 4.10: Plots pertaining to the posterior distribution of the number of births per bed in 2010 for 145 general practice hospitals in Michigan: (a) Lower bounds of 95% credible intervals for the number of births per bed; (b) upper bounds of 95% credible intervals for the number of births per bed; (c) posterior density of the number of births per bed corresponding to a hospital in Michigan with high posterior mean and standard deviation.

In doing so, we can account for zero-inflation without departing from the Poisson distribution in our modeling framework. Future work may explore other ways of accounting for zero-inflation, including the traditional zero-inflated Poisson model (Lambert, 1992) or the hurdle model (Mullahy, 1986). In addition, while we are encouraged by the strong predictive performance of our model, we note that our prediction intervals cover the true value approximately 90% of the time, rather than the desired 95%. Future work should address this issue, perhaps by specifying a negative binomial distribution on the effective number of cases, commonly used to account for over-dispersion in count-valued data, which can't be easily accommodated in a Poisson modeling framework.

A reasonable criticism of the model we have presented here is that it does not fully account for the ACS sampling design because it does not account for the fact that ACS samples are drawn independently at the county level. One could be more accommodating to the ACS sampling design by including in the model for $\lambda_t^{(1)}(A_{ig})$ a county-level random effect. Doing so would increase the marginal covariance between the effective number of cases corresponding to census tracts within the same county. Such a model is provided in Chapter III, and readers may refer to Appendix B.2 for a derivation of the marginal covariance of the effective number of cases under a modeling framework for proportions when a county-level random effect is included.

An inherent limitation of our data analysis is the quality of the data. While a number of metrics of the ACS are measured with high statistical precision, for example, the proportion of families in poverty (see Chapter III), we observed that

a large number of ACS estimates of the number of births exhibit extremely high standard errors. A number of these were zero-valued, indicating perhaps that an insufficient number of surveys were administered in order to capture these and other relatively rare events over small areal units. As our model is equipped to jointly model survey-based estimates of differing spatial supports, regardless or their source, future work could incorporate data from sources other than the ACS. For instance, The Michigan Department of Health and Human Services provides county-level data on the number of births in Michigan based on the Geocoded Michigan Birth Certificate Registry. These data are likely more accurate than the ACS, and would improve yearly estimates of birth at the sub-county level. Continuing with this line of thought, the disaggregation model could be utilized alongside other data fusion models, perhaps leveraging both individual and aggregated data.

CHAPTER V

A Point Pattern Modeling Framework to Address Sampling Bias in Electronic Health Records

5.1 Introduction

In recent years, the use of Electronic Health Records (EHRs) in United States hospitals has neared complete coverage. As of 2017, approximately 86% of hospitals in the U.S. use some form of EHR (U.S. Centers for Disease Control & Prevention, 2019). With the advent of EHRs, physicians are given faster access to patient information such as diagnostic tests, demographics, and various disease risk factors. Furthermore, patients may consent to their information being used for research purposes, creating exciting new opportunities for public health researchers who may use EHR data in place of a costly and time-consuming cross-sectional or cohort study. However, using EHR data for epidemiological or clinical research presents various challenges, with one of the most troublesome being that EHR data are not collected with research purposes in mind. Rather, they are generated through patient interactions with hospital systems, and are potentially restricted further by only including individuals who consent to their data being used for research purposes. This non-probabilistic sampling mechanism can result

in samples that are not representative of the target population and that could yield biased inference on, say, the association between disease and exposure. In this chapter, we propose to correct for sampling bias in EHR data by modeling the locations of EHR subjects as a spatial point process, transforming the intensity of the process to a sampling probability, and using those estimated probabilities to compute sampling weights for the EHR data.

5.1.1 Sampling bias in EHR data

Sampling bias is widely recognized as a potential pitfall of statistical science. In the context of hospital data, the term *Berkson's bias* or *Berkson's paradox* is often used to describe sampling bias that occurs due to a study being performed using only hospital patients, when the target population of the study is more broadly defined (say, for example, an entire county or state) (Berkson, 1946; Westreich, 2012). While Berkson's bias is typically attributed to the fact that hospital patients are, on average, sicker than the general population (Weiskopf et al., 2013), individuals whose EHR data are used in research studies can also have differing demographic characteristics and risk behaviors, higher rates of private insurance usage (Bower et al., 2017), and other departures from the general population.

In studies of the association between exposure and disease, the presence of sampling bias depends on the relationship between exposure and selection and the relationship between disease and selection. Westreich (2012) provides a concise summary of how these relationships affect inference from both a sampling bias and a missing data perspective. When both exposure and disease are not related

to selection, there is no need to correct inference, as sampling is equivalent to simple random sampling. When exposure is related to selection and disease is not, contrasts of disease risk are unbiased, however it is worth noting that this scenario is rarely encountered in practice. When disease is related to selection and exposure is not, odds ratios are unbiased, a phenomenon that is often noted in the context of case-control studies. Finally, when both exposure and disease are related to selection, all inference is subject to bias. Beesley et al. (2018) present results that are consistent with the findings of Westreich (2012), while also providing a modeling framework for characterizing both sampling and misclassification bias in studies with EHR data. These results suggest that users performing analyses of EHR data should expect little bias as long as the sampling mechanism is independent of exposure status, which is perhaps a tenable assumption in, say, studies of gene-related associations (Avery et al., 2009). However, we envision a number of scenarios in which this does not hold. Consider, for example, the association between risk factors such as smoking or exposure to particulate matter and bronchial disease such as lung cancer or chronic obstructive pulmonary disease (COPD). In this case, we expect diseased individuals to be more likely to use the hospital system, meaning that there is an association between disease and the sampling mechanism. Furthermore, the rates of aforementioned exposures may very well be different for those in the hospital system compared to those in the target population, perhaps due to their proximity to the hospital in the case of particulate exposure, especially if the hospital is in an urban area.

In situations where the sampling mechanism of the EHR data could poten-

tially result in biased inference, it is necessary to attempt some sort of correction. For example, Goldstein et al. (2016) suggest controlling for the number of health encounters and demonstrate via simulation that such an adjustment can alleviate bias due to individuals with EHRs being, on average, sicker than the general population. Adjusting for factors that impact selection is in general viewed as a viable method of controlling for sampling bias, provided that data pertaining to those factors are available. Another frequently utilized method for controlling for sampling bias, and one which will be the focus of this chapter, is weighting based on the inverse probability of selection (Gelman, 2007).

5.1.2 Preferential sampling in geostatistics

Closely related to the issue of sampling bias is the issue of preferential sampling in point-referenced spatial data, also known as *geostatistical* data. In this section, we will provide an overview of preferential sampling in geostatistics and two models that were among the first to address this issue, specifically the models of Diggle et al. (2010) and Pati et al. (2011).

In a typical geostatistical model, the outcome of interest corresponds to a point-referenced location that is considered to be fixed. In addition, the data locations are assumed to be independent of the underlying spatial process. However, in practice this assumption can be violated, and the observation locations might be stochastically dependent on the spatial process that underlies the response field. A typical instance of this is encountered when collecting and modeling air pollution data: monitors are typically located in more highly polluted and densely

populated areas. In their seminal paper, Diggle et al. (2010) demonstrate that such preferential sampling leads to biased inference and prediction. They curtail this problem through a framework that models the locations as a spatial point process, and the outcomes at those locations through a traditional geostatistical model, when conditioned on the sampled locations. Crucial to their methodology is the notion that the geostatistical process underlying the outcome, and the spatial point process underlying the locations, depend on the same spatial random effect. By jointly modeling the observations and their locations, the authors achieve a substantial reduction in bias in simulations in which data are generated via a preferential sampling scheme. A similar idea was at the basis of the model by Pati et al. (2011), in which locations are once again modeled through a spatial point process, whose intensity is then used as a linear covariate for the model of the outcome of interest.

While our model does not explicitly follow those of Diggle et al. (2010) or Pati et al. (2011), we do take inspiration from them in our proposed modeling framework, in that we too model the locations of the EHR data using a spatial point process. In his comments on Diggle et al. (2010), Rathbun (2010) draws an insightful connection between the topic of preferential sampling in geostatistics and the effect of sampling design on inference in the field of survey statistics. He posits that by modeling the locations as a spatial point process, and then weighting the observed responses by the inverse intensity of that spatial point process, one may also reduce bias in inference and prediction of geostatistical data. He points out that this procedure would be analogous to weighting by the inverse probability

of selection. We take this observation to be the starting point of our modeling framework. Specifically, while also modeling the association between disease and exposure, we aim to recover the sampling probabilities of the EHR subjects by modeling their locations as an inhomogeneous Poisson process. Our model also allows for additional calibration of the sampling probabilities by incorporating publicly available health data that are aggregated at the areal (e.g. county) level.

5.1.3 Data sources

Because we envision a model that synthesizes publicly available aggregate data with EHRs, we proceed by describing several sources that provide health-related data and statistics. The aggregate data that we focus on in this chapter will hereafter be referred to, generally, as areal data. In this section, we will describe several data sources that we believe will be useful to users in fitting our model to EHR data, however we note that the sources we provide here are far from representing a comprehensive list.

The Behavioral Risk Factor Surveillance System (BRFSS) is a health survey coordinated by the United States Centers for Disease Control and Prevention (CDC). BRFSS conducts more than 400,000 interviews per year, and provides data on factors that include overall health, health care coverage, certain chronic health conditions including some cancers, and risk behavior such as smoking and lack of physical activity (U. S. Centers for Disease Control & Prevention, 2019a). Also conducted by the CDC, specifically the National Center for Health Statistics, the National Health Interview Survey (NHIS) consists of questions pertaining to ba-

sic health and demographics, as well as specific information on mental health, injuries, disabilities, and chronic conditions including cancer (U. S. Centers for Disease Control & Prevention, 2019b).

Both BRFSS and NHIS provide highly de-identified individual-level survey data to the public. In addition, they are used by the CDC and National Institutes of Health (NIH) to create the county-level estimates on the State Cancer Profiles (SCP) website. Users of SCP can obtain county-level estimates of cancer incidence and mortality, risk factors, and various demographics. Although these data lack individual-level details, their geographic information allows users to characterize the cancer burden of sub-state areas while exploring associations between disease and exposure at an ecological level (U. S. Centers for Disease Control & Prevention, 2019c).

Finally, the American Community Survey (ACS), an ongoing national survey conducted by the U.S. Census Bureau, is a source of up-to-date information on demographic, housing, and economic factors. ACS estimates are provided at various levels of geographic resolution, including sub-county resolutions such as census tracts and block groups (U.S. Census Bureau, 2008), making the ACS highly useful in understanding neighborhood dynamics in the U.S.

5.1.4 Chapter organization

Having summarized the data that we intend to leverage, we now provide a road map for the remainder of this chapter. In Section 5.2, we present our model to recover the EHR sampling probabilities by using in tandem both areal and point-

referenced data, while also modeling the probability of disease via a straightforward logistic regression model. In Section 5.3, we present results from three simulation studies that demonstrate that our modeling framework is successful in recovering the sampling probabilities and, through weighting by the inverse selection probabilities, effective in reducing bias of the parameter that quantifies the association between disease and exposure in simulated data. Section 5.3 also includes a case-control study of the association between smoking and incident lung cancer using EHR data from the Michigan Genomics Initiative. We conclude with a discussion of limitations and future work in Section 5.4

5.2 Methods

In this section, we state our modeling framework for jointly modeling the association between disease and exposure as well as the probability of selection for EHR data. We begin by stating our disease model, which models the probability of disease via a logistic regression model. Next, we model the probability of selection by transforming the intensity function of the inhomogeneous Poisson process model that we use to model the EHR locations. Finally, we provide a framework to further calibrate the selection probabilities by incorporating external data that are aggregated over space.

5.2.1 Disease modeling framework

Let us denote the outcome $Y(\mathbf{s})$ as the disease status of an individual at location $\mathbf{s}$ in the spatial domain $\mathcal{S}$. Specifically, $Y(\mathbf{s}) = 1$ if a subject at location $\mathbf{s}$ has the disease of interest, and $Y(\mathbf{s}) = 0$ otherwise. Our main exposure of interest is

denoted $X(\mathbf{s})$, which may be binary, ordinal, or numeric depending on the data that are available. Our goal is to estimate the association between $Y(\mathbf{s})$ and $X(\mathbf{s})$, quantified by a parameter β_1. Disease status may be associated with q additional risk factors (which can be considered confounders or comorbidities in this setting), which we will denote $Z_1(\mathbf{s}), ..., Z_q(\mathbf{s})$. These factors should be controlled for in our model, and their associations are quantified through parameters $\gamma_1, ..., \gamma_q$. We model the association between disease and exposure using a logistic regression model, with the outcome being the log odds of the probability of disease $\pi(\mathbf{s})$.

$$\log\left(\frac{\pi(\mathbf{s})}{1-\pi(\mathbf{s})}\right) = \beta_0 + \beta_1 X(\mathbf{s}) + \gamma_1 Z_1(\mathbf{s}) + ... + \gamma_q Z_q(\mathbf{s}) \tag{5.1}$$

As discussed in Section 5.1, the modeling framework in (5.1) would yield an unbiased estimate of β_1 and other parameters under circumstances that may be untenable in the EHR setting (e.g. that the data are a simple random sample from the population of interest). Alternatively, if the probability of selection was known, as it would be for a survey created with research purposes in mind, one could produce unbiased estimates of β_1 and other parameters through, for example, inverse probability weighting. However, given that these data are sampled through patient interactions with the health care system rather than a pre-specified sampling design, we must propose another approach to estimate the selection probabilities.

5.2.2 Modeling EHR locations with an inhomogeneous Poisson process

In an attempt to model the sampling probabilities of the EHR data, we will employ a spatial point process model, specifically an inhomogeneous Poisson process (hereafter referred to simply as IPP). That is, the IPP generates a countable

set of points on the spatial domain $\mathcal{S}$ and adheres to the following conditions:

1. For any B belonging to the borel σ-algebra $\mathcal{B}$ of subsets of $\mathcal{S}$, the number of points in B, denoted $N(B)$, is a Poisson random variable.
2. For any integer n and for any $B \in \mathcal{B}$, with $0 < E(N(B)) < \infty$, conditional on $N(B) = n$, the events are located independently an uniformly over B.

The IPP is characterized by the intensity function $\lambda(\mathbf{s})$, defined for each $\mathbf{s} \in \mathcal{S}$ as:

$$\lambda(\mathbf{s}) = \lim_{|\Delta\mathbf{s}| \to 0} \frac{E(N(\Delta\mathbf{s}))}{|\Delta\mathbf{s}|}$$

It follows that $E(N(B)) = \int_{\mathbf{s} \in B} \lambda(\mathbf{s}) d\mathbf{s}$. $\lambda(\mathbf{s})$ is modeled as in (5.2), in which $D(\mathbf{s})$ denotes the distance between location $\mathbf{s}$ and the hospital, and $V(\mathbf{s})$ denotes the population density at location $\mathbf{s}$.

$$\begin{aligned} \log(\lambda(\mathbf{s})) &= \varphi_0 + \varphi_1 Z_1(\mathbf{s}) + ... + \varphi_q Z_q(\mathbf{s}) + \varphi_{q+1} Y(\mathbf{s}) + \\ &\quad \varphi_{q+2} X(\mathbf{s}) + \gamma_{q+3} D(\mathbf{s}) + \gamma_{q+4} V(\mathbf{s}) \end{aligned} \tag{5.2}$$

We now derive rudimentary estimates for the sampling probabilities of the subjects in the EHR dataset. Here for convenience of notation we introduce the subscript i, and proceed by denoting the probability of selection of an EHR subject at location $\mathbf{s}_i$ as $p(\mathbf{s}_i)$, $i = 1, ..., m$, with m being the number of EHR subjects. Specifically, for a subject residing at location $\mathbf{s}_i$, we define the probability of selection as:

$$p(\mathbf{s}_i) := \lim_{|\Delta(\mathbf{s}_i)|\to 0} \frac{E(N(\Delta(\mathbf{s}_i)))}{M(\Delta(\mathbf{s}_i))} = \lim_{|\Delta(\mathbf{s}_i)|\to 0} \frac{\int_{\mathbf{s}\in\Delta(\mathbf{s}_i)} \lambda(\mathbf{s}_i)d\mathbf{s}}{M(\Delta(\mathbf{s}_i))}, \qquad i = 1, ..., m$$

In other words, we assume *a priori* that the sampling probability of an individual at location $\mathbf{s}_i$ is equal to the limit as $\Delta(\mathbf{s}_i)$ becomes infinitesimally small of the expected number of EHR subjects in $\Delta(\mathbf{s}_i)$ divided by the population of $\Delta(\mathbf{s}_i)$. In practice, it may be necessary to define $\Delta(\mathbf{s}_i)$ as a relatively small unit, such as a census tract or block group, for which the population $M(\Delta(\mathbf{s}_i))$ can be estimated with a reasonable degree of statistical precision. In this case, $p(\mathbf{s}_i)$ is truly an areal quantity, meaning that sujbects residing in $\Delta(\mathbf{s}_i)$ would have the same estimated sampling probability, provided they exhibit the same values for the factors on the right hand side of (5.2).

A more exact approach, specifically one that defines probabilities at the point-referenced spatial resolution rather than an areal one, would be to also model the population locations that underlie $M(\Delta(\mathbf{s}_i))$ as a spatial point pattern with intensity $\lambda_M(\mathbf{s}_i)$. We would then then define the sampling probability of the subject residing at location $\mathbf{s}_i$ as:

$$p(\mathbf{s}_i) := \frac{\lambda(\mathbf{s}_i)}{\lambda(_M\mathbf{s}_i)}, \qquad i = 1, ..., m$$

5.2.3 Calibrating sampling probabilities using areal data

Having defined an estimate for the sampling probabilities, we will now propose a method to incorporate available areal data by invoking the Horvitz-Thompson estimators of various quantities used in our model. The Horvitz-Thompson estimator of a stratum-wide population total is equal to the sum of the sample

data times the inverse probability of selection. Thus, we can define a number of Horvitz-Thompson estimators based on the EHR data.

Let $M(A_k)$ and $m(A_k)$ denote respectively the number of people and then number of EHR subjects residing in areal unit A_k, $k = 1, ..., K$, with $M = \sum_{k=1}^{K} M(A_k)$ being the total population of the domain $\mathcal{S}$ and $m = \sum_{k=1}^{K} m(A_k)$ being the total number of EHR subjects in the spatial domain $\mathcal{S}$. Let $\tilde{Y}(A_k)$ be the true number of people with the disease indicated by $Y(\mathbf{s})$ in areal unit A_k. Similarly let $\tilde{X}(A_k)$ and $\tilde{Z}_j(A_k), j = 1..., q$ denote the total number of people in areal unit A_k exhibiting the characteristic denoted by the random variables $X(\mathbf{s})$ and $Z_j(\mathbf{s}), j = 1, ..., q$ respectively. Alternatively, for continuous variables, the aggregate quantity for A_k could also be defined as the mean value of the random variables for areal unit A_k. The areal-level statistics $\tilde{Y}(A_k)$, $\tilde{X}(A_k)$, $\tilde{Z}_j(A_k), j = 1, ..., q$, and $\tilde{M}(A_k)$, can be obtained through public sources such as State Cancer Profiles or the American Community Survey. Here, $\tilde{M}(A_k)$ denotes the population of A_k as provided by the aforementioned sources. For simplicity, we will consider them fixed and known, however we acknowledge that $\tilde{Y}(A_k)$, $\tilde{X}(A_k)$, $\tilde{Z}_j(A_k), j = 1, ..., q$, and $\tilde{M}(A_k)$ may truly be estimates derived from sampling surveys, and are therefore subject to sampling error.

Given the locations of the EHR subjects, $\mathbf{s}_i, i = 1, ..., m$, we can estimate $\tilde{Y}(A_k)$, $\tilde{X}(A_k)$, $\tilde{Z}_j(A_k), j = 1, ..., q$, and $\tilde{M}(A_k)$ using only the EHR data. Using $\tilde{Y}(A_k)$ as an example, the EHR-based estimator of $\tilde{Y}(A_k)$, denoted $\tilde{Y}_{EHR}(A_k)$, is:

$$\tilde{Y}_{EHR}(A_k) = \sum_{i=1}^{m} \frac{Y(\mathbf{s}_i)}{p(\mathbf{s}_i)} I(\mathbf{s}_i \in A_k)$$

Where $I(\cdot)$ denotes an indicator function. $\tilde{X}_{EHR}(A_k)$, $\tilde{Z}_{j,EHR}(A_k), j = 1, ..., q$, and $\tilde{M}_{EHR}(A_k)$ are defined similarly and respectively as the EHR-based estimates of $\tilde{X}(A_k)$, $\tilde{Z}_j(A_k), j = 1, ..., q$, and $\tilde{M}(A_k)$.

For the sake of casting these areal data into our modeling framework, we will assume *a priori* that the $\tilde{Y}_{EHR}(A_k)$ are normally distributed with means equal to their true values $\tilde{Y}(A_k), k = 1, ..., K$, and variance τ_1^2. Applying similar prior distributions to the remaining areal estimates, we get:

$$\begin{aligned}
\tilde{Y}_{EHR}(A_k) &\overset{ind}{\sim} N(\tilde{Y}(A_k), \tau_1^2), \qquad k = 1, ..., K \\
\tilde{X}_{EHR}(A_k) &\overset{ind}{\sim} N(\tilde{X}(A_k), \tau_2^2), \qquad k = 1, ..., K \\
\tilde{Z}_{j,EHR}(A_k) &\overset{ind}{\sim} N(\tilde{Z}_j(A_k), \tau_{j+2}^2), \qquad j = 1, ..., q; k = 1, ..., K \\
\tilde{M}_{EHR}(A_k) &\overset{ind}{\sim} N(\tilde{M}(A_k), \tau_{q+3}^2), \qquad k = 1, ..., K
\end{aligned}$$

5.2.4 The complete modeling framework

Combining the above three modeling components, our full modeling framework can be stated as follows:

$$
\begin{aligned}
L(\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_m|\lambda(s)) &\propto \exp\left(-\int_{\mathbf{s}\in\mathcal{S}} \lambda(\mathbf{s})d\mathbf{s}\right)\prod_{i=1}^{m}\lambda(\mathbf{s}_i) \\
\log\left(\lambda(\mathbf{s})\right) &= \varphi_0 + \varphi_1 Z_1(\mathbf{s}) + ... + \varphi_q Z_q(\mathbf{s}) + \varphi_{q+1}Y(\mathbf{s}) + \\
&\quad \varphi_{q+2}X(\mathbf{s}) + \gamma_{q+3}D(\mathbf{s}) + \gamma_{q+4}V(\mathbf{s}) \\
p(\mathbf{s}_i) &= \lim_{|\Delta(\mathbf{s}_i)|\to 0}\frac{\int_{\mathbf{s}\in\Delta(\mathbf{s}_i)}\lambda(\mathbf{s}_i)d\mathbf{s}}{M(\Delta(\mathbf{s}_i))}, \qquad i = 1, ..., m \qquad (5.3)
\end{aligned}
$$

$$
\begin{aligned}
Y(\mathbf{s}_i)|\pi(\mathbf{s}_i) &\overset{ind}{\sim} \text{Bernoulli}(\pi(\mathbf{s}_i)), \qquad i = 1, ..., m \\
\log\left(\frac{\pi(\mathbf{s}_i)}{1-\pi(\mathbf{s}_i)}\right) &= \beta_0 + \beta_1 X(\mathbf{s}_i) + \gamma_1 Z_1(\mathbf{s}_i) + ... + \gamma_q Z_q(\mathbf{s}_i) \qquad (5.4)
\end{aligned}
$$

$$
\begin{aligned}
\tilde{Y}_{EHR}(A_k) &\overset{ind}{\sim} N(\tilde{Y}(A_k), \tau_1^2), \qquad k = 1, ..., K \\
\tilde{X}_{EHR}(A_k) &\overset{ind}{\sim} N(\tilde{X}(A_k), \tau_2^2), \qquad k = 1, ..., K \\
\tilde{Z}_{j,EHR}(A_k) &\overset{ind}{\sim} N(\tilde{Z}_j(A_k), \tau_{j+2}^2), \qquad j = 1, ..., q; k = 1, ..., K \\
\tilde{M}_{EHR}(A_k) &\overset{ind}{\sim} N(\tilde{M}(A_k), \tau_{q+3}^2), \qquad k = 1, ..., K \qquad (5.5)
\end{aligned}
$$

$$
\begin{aligned}
\tau_j^2 &\sim IG(1,1), \qquad j = 1, \ldots, q+3 \\
p(\beta_j) &\propto 1, \qquad j = 1, 2 \\
p(\gamma_j) &\propto 1, \qquad j = 1, \ldots, q \\
p(\phi_j) &\propto 1, \qquad j = 1, \ldots, q+4
\end{aligned}
$$

The data likelihood of $Y(\mathbf{s}_i), i = 1, ..., m$, is defined based on the pseudo-

posterior method proposed by Savitsky and Toth (2016), in which the data likelihood is weighted by a set of weights that are proportional to the inverse sampling probability and normalized so that they sum to the total sample size m. Specifically, let $\mathbf{Y} = \{Y(\mathbf{s}_1), ..., Y(\mathbf{s}_m)\}$ denote the disease status for the EHR subjects and $\boldsymbol{\Theta}$ be the set of all parameters in the model, then the pseudo-likelihood of the data defined by Savitsky and Toth (2016), denoted $L_P(\mathbf{Y}|\boldsymbol{\Theta})$, is given by:

$$L_P(\mathbf{Y}|\boldsymbol{\Theta}) = \prod_{i=1}^{m} L(Y(\mathbf{s}_i)|\boldsymbol{\Theta})^{\tilde{w}(\mathbf{s}_i)}$$

where $L(Y(\mathbf{s}_i)|\boldsymbol{\Theta})$ is the data likelihood of $Y(\mathbf{s}_i)$ given in (5.4) and $\tilde{w}(\mathbf{s}_i)$ is defined as follows:

$$\begin{aligned} w(\mathbf{s}_i) &= \frac{1}{p(\mathbf{s}_i)}, \qquad i = 1, ..., m \\ \tilde{w}(\mathbf{s}_i) &= m\frac{w(\mathbf{s}_i)}{\sum_{j=1}^{m} w(\mathbf{s}_j)}, \qquad i = 1, ..., m \end{aligned}$$

In Section 5.3, we will fit various forms of the model described in (5.3) through (5.5), at times excluding certain portions in order to more effectively demonstrate the utility of the full model. For added clarity, we will refer to the model components described up to equation (5.3) are referred to as the *IPP* component of our model; the model components described in the equations through (5.4) as the *regression* component of our full model; finally the equations through (5.5) are referred to as the *calibration* component of our model.

5.2.5 Computation

The model is fit via Markov Chain Monte Carlo (MCMC), with Gibbs sampling steps being used to generate posterior samples of the variance parameters $\tau_j^2, j =$

$1, ..., q + 3$ in the calibration component of the model, and Metropolis-Hastings steps being used to generate posterior samples of the parameters for the IPP and regression components. Proposal values for the regression parameters are generated at each MCMC iteration via a uniform distribution centered at the current value of each parameter, with proposal variance adjusted automatically every 100 iterations during burn-in to achieve an acceptance rate of 30%, based on the optimal acceptance rates of Roberts et al. (1997). In the three simulation studies presented in Sections 5.3.1 through 5.3.4, the MCMC algorithm is run for 10,000 iterations, with the first 5,000 iterations discarded as a burn-in period. In the data analysis, we increase the number of iterations to 15,000, with the first 10,000 being discarded as a burn-in period. Convergence is assessed via visual inspection of trace plots for all model parameters and through computation of Geweke statistics (Geweke, 1992). The 5,000 iterations that occur post-burn-in are sufficient to ensure that effective sample sizes for all parameters exceed 1,000.

5.3 Results

In this section, we begin by presenting results from two simulation studies with the aim of testing our proposed model for adjusting for sampling bias in EHR data. In both simulations, we first generate the underlying population at a set of locations, each with disease, exposure, and comorbidity variables. From the population, we then select a sample representing the EHR subjects. In both simulation studies, we define a priori a location to represent the site of the hospital from which the EHR sample are taken. In the first simulation study, the hospital

location is randomly generated, whereas in the second one we placed the hospital in a densely populated area.

In both of these studies, the sampled observations representing the EHR subjects differ from the overall population in that they live close to the hospital, are more likely to have the disease (i.e. $Y(\mathbf{s}) = 1$), and exhibit differing levels of exposure and certain other factors. Furthermore, for each variable generated during these simulations ($Y(\mathbf{s})$, $X(\mathbf{s})$, etc.), we also compute a set of areal data corresponding to 100 square subregions of the domain. For example, $\tilde{Y}(A_k)$ is the number of simulated data points in areal unit A_k, $k = 1, ..., 100$, for which $Y(\mathbf{s}) = 1$. These aggregated quantities are considered fixed and known in these simulations, and are analogous to areal data such as those obtained from the ACS or SCP.

After this, we present a third simulation study whose data generation procedure is very similar to that of the first two simulation studies. However, in lieu of the sampling probabilities being a function of the distance to the hospital, we generate another, unmeasured factor, that is strongly associated with selection, and that we do not utilize in our model fitting. In doing so, we evaluate the performance of our model in a less convenient simulation setting; that is one in which we do not account for all factors that are associated with selection.

In addition to the simulation studies, we present results from a case-control analysis that assesses the association between smoking and incident lung cancer using EHRs from the Michigan Genomics Initiative. This analysis incorporates county-level data from the State Cancer Profiles website (U. S. Centers for Disease

Control & Prevention, 2019c) in order to further calibrate sampling probabilities as described in Section 5.2.3.

5.3.1 Simulation Study 1

In our first simulation study we generate a population of size 10,000 at random locations $\mathbf{s}_l^*, l = 1, .., 10000$ on the unit square, that is the spatial domain $\mathcal{S} \in (0,1) \times (0,1)$. These population data are used in each of the 30 simulations created under this setting, with each simulation drawing a new sample from the population. In addition, we generate a hospital at a single random location shared between all simulations. Figure 5.1 presents various plots pertaining to the data generation procedure. Figure 5.1 (a) presents the randomly generated population locations along with the location of the hospital.

Various attributes for each datum are then randomly generated, starting with exposure $X(\mathbf{s}_l^*), l = 1, ..., 10000$, which are Bernoulli random variables whose success probabilities are the expit of independent and identically distributed normal random variables with mean 0 and variance 2. Figure 5.1 (b) presents a histogram of the probabilities of exposure, and Figure 5.1 (e) plots the population locations with exposure status for a single simulated data set denoted in red.

Next, comorbidities $Z_1(\mathbf{s}_l^*)$ and $Z_2(\mathbf{s}_l^*), l = 1, ..., 10000$ are both generated as standard normal random variables. Disease status, $Y(\mathbf{s}_l^*), l = 1, ..., 10000$, is generated as a Bernoulli random variable with disease probability $\pi(\mathbf{s}_l^*)$ generated as follows:

$$\log\left(\frac{\pi(\mathbf{s}_l^*)}{1-\pi(\mathbf{s}_l^*)}\right) = -3 + 2X(\mathbf{s}_l^*) - 0.5Z_1(\mathbf{s}_l^*), \qquad l = 1, ..., 10000$$

Figure 5.1 (c) shows a histogram of the disease probabilities colored by exposure status, while in panel (f) it presents plots of the population locations with disease status for a single simulated data set denoted in red.

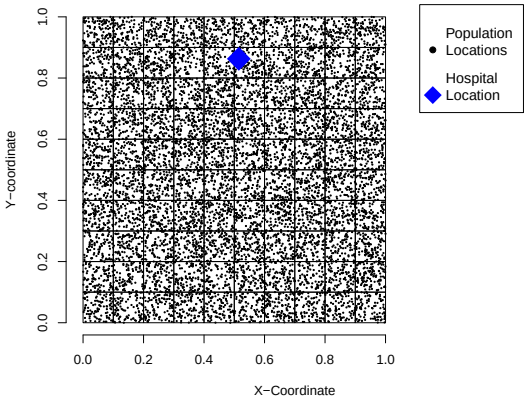
Finally, each population datum has a sampling probability probability $p(\mathbf{s}_i^*)$ generated as follows:

$$\log\left(\frac{p(\mathbf{s}_l^*)}{1-p(\mathbf{s}_l^*)}\right) \equiv -1.5 + 2Y(\mathbf{s}_l^*) - 0.5X(\mathbf{s}_l^*) - 2D(\mathbf{s}_l^*), \qquad l = 1, ..., 10000$$

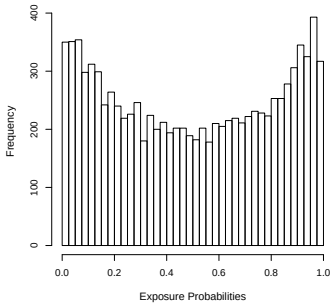
In other words, those with the disease are more likely to be in the EHR sample, those who live close to the hospital are more likely to be sampled, and those who are exposed have lower probability to be in the sampled. Simulated datasets are then created by generating Bernoulli random variables $S(\mathbf{s}_l^*)$ with success probabilities $p(\mathbf{s}_l^*), l = 1, ..., 10000$. All locations for whom $S(\mathbf{s}_l^*) = 1$ are considered part of the EHR data set, whose locations we have previously denoted as $\{\mathbf{s}_1, ..., \mathbf{s}_m\}$, and to which we fit our model. This procedure is repeated 30 times to create the 30 datasets we utilize in the full simulation study. Figure 5.1 (d) shows a histogram of the sampling probabilities while panel (g) plots the population locations with sampled locations relative to a single simulated dataset indicated in red.

5.3.2 Simulation Study 2

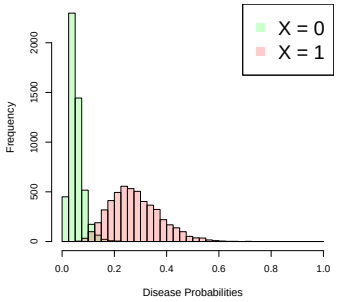
In the second simulation study, points are generated instead using a non-uniform pattern in order to more closely reflect true population distributions around a city. We begin by generating a city center on the unit square at arbitrary location



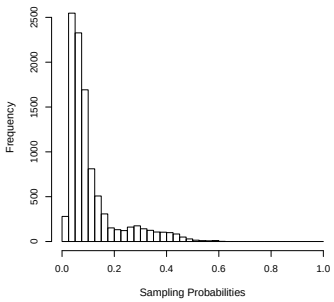
(a) Population and hospital locations.



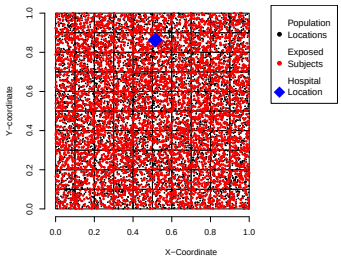
(b) Histogram of exposure probabilities.



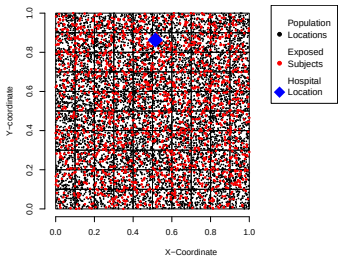
(c) Histogram of disease probabilities grouped by exposure status.



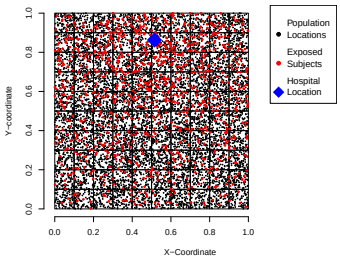
(d) Histogram of sampling probabilities.



(e) Plot of exposed subjects.



(f) Plot of diseased subjects.



(g) Plot of sampled subjects.

Figure 5.1: Plots pertaining to data generation in Simulation Study 1. (a) Population locations with hospital location indicated in green; (b) histogram of the probabilities of being exposed ($\Pr(X(\mathbf{s}) = 1)$); (b) histogram of probabilities of being diseased ($\Pr(Y(\mathbf{s}) = 1)$) broken up by exposure status; (d) histogram of sampling probabilities ($\Pr(S(\mathbf{s}) = 1)$); (e) plot of population locations with exposure status for a single simulated dataset indicated in red; (f) plot of population locations with disease status for a single simulated dataset indicated in red; (g) plot of population locations with sampling status for a single simulated dataset indicated in red.

$(0.8, 0.3)$, and then generating population locations according to an inhomogeneous Poisson process with the intensity function depending on the distance to the city center. This procedure results in 9,924 population locations. To the southwest of the city center, we generate a hospital location. We note that the city center, hospital location, and population locations are kept constant for all of the 30 simulations, with new EHR samples being generated for each simulated data set. Figure 5.2 (a) presents a plot of the population locations along with the hospital location and city center.

Next, we generate our exposure variable, $X(\mathbf{s}_l^*), l = 9924$, which is a Bernoulli random variable with success probability being higher when $\mathbf{s}_l^*$ is closer to the city center. Specifically, let $D_c(\mathbf{s}_l^*)$ be the distance between location $\mathbf{s}_l^*$ and the city center, $l = 1, ..., 9924$. Then, let:

$$(5.6) \log\left(\frac{\Pr(X(\mathbf{s}_l^*)=1)}{1-\Pr(X(\mathbf{s}_l^*)=1)}\right) = -1 - D_c(\mathbf{s}_l^*) + \epsilon_X(\mathbf{s}_l^*), \qquad l = 1, ..., 9924$$
$$\epsilon_X(\mathbf{s}_l^*) \overset{iid}{\sim} N(0, 0.1)$$

The random noise in (5.6) allows for $X(\mathbf{s}_l^*)$ to be a linear function of $D_c(\mathbf{s}_l^*)$ without the association being too extreme (see Figure 5.2 (b)). The probability of exposure, that is $Pr(X(\mathbf{s}_l^*) = 1), l = 1, ..., 9924$, ranges from 0.06 to 0.50 (see Figure 5.2 (c)). The resulting data are shown in Figure 5.2 (f), with exposed subjects corresponding to a single simulated data set denoted in red.

Comorbidities $Z_1(\mathbf{s}_l^*)$ and $Z_2(\mathbf{s}_l^*), l = 1, ..., 9924$ are again sampled from standard normal distributions, and disease status $Y(\mathbf{s}_l^*)$ is generated as a Bernoulli

random variable with success probability $\pi(\mathbf{s}_l^*)$ as follows:

$$\log\left(\frac{\pi(\mathbf{s}_l^*)}{1-\pi(\mathbf{s}_l^*)}\right) = -3 + 2X(\mathbf{s}_l^*) - 0.5Z_1(\mathbf{s}_l^*), \qquad l = 1, ..., 9924$$

Figure 5.2 panel (d) shows a histogram of the probability of disease broken up by exposure status, while panel (g) presents a plot of the population locations with disease status for a single simulated data set denoted in red.

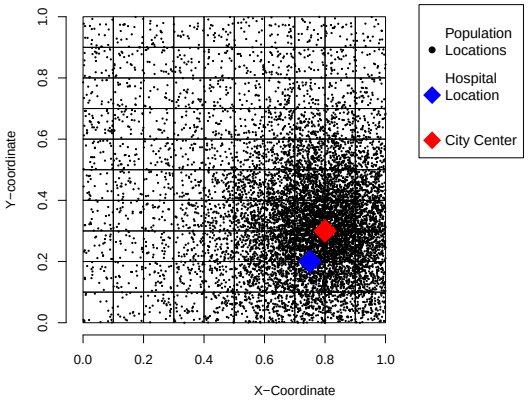
As in Simulation Study 1, each population datum has sampling probability $p(\mathbf{s}_l^*)$ generated as follows:

$$\log\left(\frac{p(\mathbf{s}_l^*)}{1-p(\mathbf{s}_l^*)}\right) = -2 + 2Y(\mathbf{s}_l^*) - X(\mathbf{s}_l^*) - 3D(\mathbf{s}_l^*) + 0.5Z_1(\mathbf{s}_l^*), \qquad l = 1, ..., 9924 \tag{5.7}$$

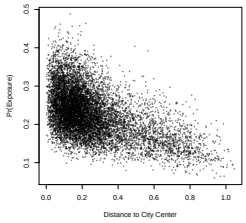
Figure 5.2 panel (e) presents a histogram of the probability of being sampled, while panel (h) plots the population locations with the locations of the sampled data for a single simulation indicated in red. Simulated datasets are created using the same procedure as in Simulation Study 1, with a Bernoulli random variable $S(\mathbf{s}_l^*)$ with probability $p(\mathbf{s}_l^*), l = 1, ..., 9924$, determining sampling status in each of 30 data sets, and with the resulting EHR data locations being denoted as $\{\mathbf{s}_1, ..., \mathbf{s}_m\}$.

5.3.3 Simulation Study 3

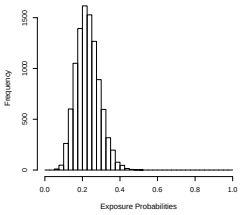
For our third simulation study, we proceed in an identical fashion to Simulation Study 2, up until the generation of sampling probabilities defined in (5.7). This means that the population data for this study are identical to those of Simulation Study 2, and differ only in their selection probabilities that determine the 30



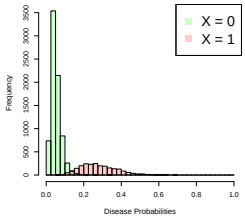
(a) Population, hospital, and city center locations.



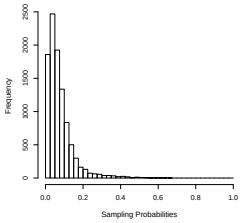
(b) Probabilities of exposure vs. $D_c(\mathbf{s}_i)$.



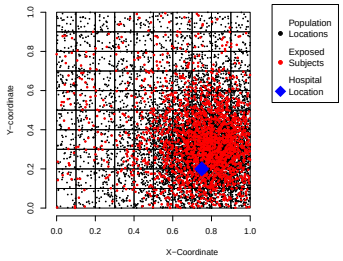
(c) Histogram of exposure probabilities.



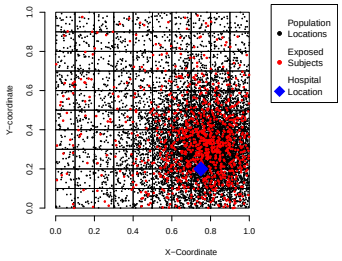
(d) Histogram of disease probabilities grouped by exposure status.



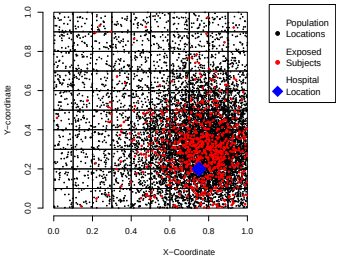
(e) Histogram of sampling probabilities.



(f) Plot of exposed subjects.



(g) Plot of diseased subjects.



(h) Plot of sampled subjects.

Figure 5.2: Plots pertaining to data generation in Simulation Study 2. (a) Population locations with hospital location indicated in green and city center indicated in red; (b) scatterplot of distance to city center ($D_c(\mathbf{s})$) vs. probabilities of being exposed ($\Pr(X(\mathbf{s}) = 1)$); (c) histogram of the probabilities of being exposed ($\Pr(X(\mathbf{s}) = 1)$); (d) histogram of probabilities of being diseased ($\Pr(Y(\mathbf{s}) = 1)$) broken up by exposure status; (e) histogram of sampling probabilities ($\Pr(S(\mathbf{s}) = 1)$); (f) plot of population locations with exposure status for a single simulated dataset indicated in red; (g) plot of population locations with disease status for a single simulated dataset indicated in red; (h) plot of population locations with sampling status for a single simulated dataset indicated in red.

sampled datasets. Instead of generating sampling probabilities according to (5.7), for this third simulation study, we define another variable, $Q(\mathbf{s}_l^*)$, that is related to selection as follows:

$$\begin{aligned} \Pr(Q(\mathbf{s}_l^*) = 1) &= \frac{\exp(\xi_Q(\mathbf{s}_l^*))}{1 + \exp(\xi_Q(\mathbf{s}_l^*))}, \qquad l = 1, ..., 9924 \\ \xi_Q(\mathbf{s}_l^*) &\overset{iid}{\sim} N(0, 1) \\ Q(\mathbf{s}_l^*) &\overset{ind}{\sim} \text{Bernoulli}(\Pr(Q(\mathbf{s}_l^*) = 1)) \end{aligned}$$

Then, the probability of selection is:

$$\log\left(\frac{p(\mathbf{s}_l^*)}{1 - p(\mathbf{s}_l^*)}\right) = -3.5 + Y(\mathbf{s}_l^*) - 1.5X(\mathbf{s}_l^*) + 2Q(\mathbf{s}_l^*) - Z_1(\mathbf{s}_l^*), \qquad l = 1, ..., 9924$$

$Q(\mathbf{s}_l^*)$ may be viewed as a variable that we know to be associated with inclusion in the EHR sample, but that is inaccessible to researchers due to privacy concerns. For example, $Q(\mathbf{s}_l^*)$ could denote whether or not a subject has private insurance coverage.

We proceed as in Simulation Studies 1 and 2, generating Bernoulli random variables $S(\mathbf{s}_l^*), l = 1, ..., 9924$ in order to define which members of the population are sampled into the dataset to which our model is fit. The resulting EHR data set is denoted as $\{\mathbf{s}_1, ..., \mathbf{s}_m\}$. This procedure is repeated 30 times.

Because the population data are identical to Simulation Study 2, here we do not include figures describing the data generating procedures, with the exception of Figure 5.3, in which we present a histogram of the sampling probabilities under Simulation Study 3.

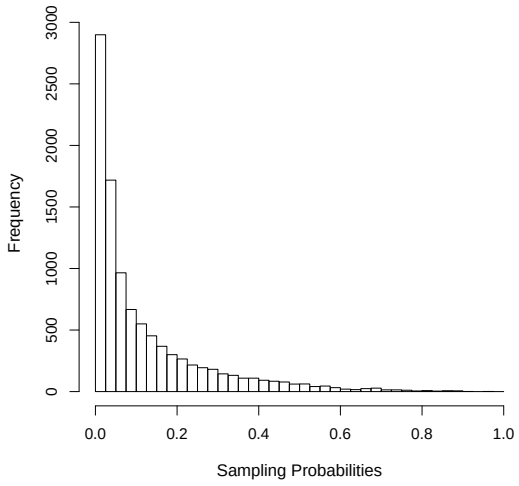


Figure 5.3: Histogram of sampling probabilities for Simulation Study 3.

5.3.4 Simulation results

We fit our model to each of the 30 simulated datasets generated under the three simulation procedures. For comparison, we also fit to the same data several other models so to be able to assess the performance of our full modeling framework. The competing models considered are:

1. Naive: A logistic regression model that regresses the log odds of disease on exposure and comorbidities with no adjustment for sampling bias. For this model, we simply use off-the-shelf software.
2. True inverse probability weight (True IPW): A logistic regression model that regresses the log odds of disease on exposure and comorbidities that is weighted based on the true inverse sampling probabilities. This is our "gold standard" that we could not achieve in practice because we don't know the true sampling probabilities for EHR data.

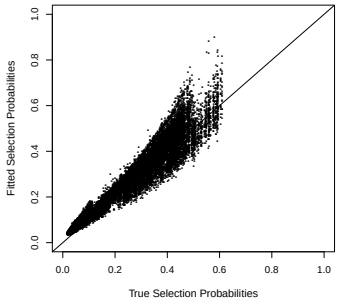
3. Fitted IPW: A logistic regression model that regresses the log odds of disease on exposure and comorbidities with likelihood weighted based on the inverse of the posterior sampling probabilities $p(\mathbf{s}_i), i = 1, ..., 1000$. This model captures our entire modeling framework, with the IPP, calibration, and logistic regression modeling components of a hierarchical Bayesian model. The survey weights are applied to the likelihood as in Savitsky and Toth (2016).
4. Fitted IPW (no calibration): A logistic regression model that regresses the log odds of disease on exposure and comorbidities that is weighted based on the inverse of the posterior sampling probabilities $p(\mathbf{s}_i), i = 1, ..., 1000$. This model excludes the calibration component of our full Bayesian hierarchical model, meaning we do not incorporate areal data into this model. We do not fit this model in Simulation Study 3.
5. Fitted IPW (off-the-shelf): A logistic regression model that regresses the log odds of disease on exposure and comorbidities and for which the likelihood is weighted based on the inverse of the estimated sampling probabilities $p(\mathbf{s}_i), i = 1, ...1000$. Note that in this case we still use the publicly available data in the calibration component of our model. However, only the calibration and IPP components of our model are fit in a Bayesian hierarchical model. The logistic regression stage is fit using off-the-shelf software, with the posterior means of the sampling probabilities being used to derive sampling weights.

We will start by examining results that pertain to the recovery of the true sampling probabilities. Figure 5.4 (a) presents the posterior mean selection prob-

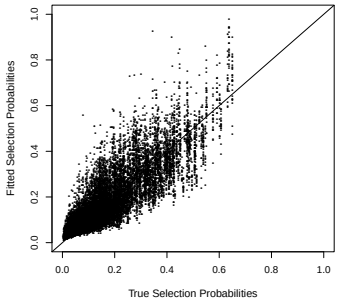
abilities against the true selection probabilities using data from all 30 simulations in Simulation Study 1. Figure 5.4 (b) presents a plot of analogous quantities, but referring to Simulation Study 2. These results are obtained when fitting Model 3 in the above enumeration, which is our full Bayesian hierarchical model. While we do not recover as accurately the true sampling probabilities in Simulation Study 2, both settings are characterized by successful recovery of the true selection probabilities when the full model is specified. In Figure 5.4 panels (c) and (d), we present the same scatterplots comparing the estimated selection probabilities with the true selection probabilities, the former now estimated using Model 4, the fitted IPW with no calibration. In this case, we fail to recover the selection probabilities, highlighting the importance of the calibration stage of our modeling framework. The key takeaway pertaining to the sampling probabilities is that by incorporating both point-level and areal-level data in the IPP and calibration components of our modeling framework, we are able to recover with some success the true selection probabilities.

In Figure 5.4 (e), we present for Simulation Study 3 a similar scatterplot comparing the estimated selection probabilities to the true ones, under our full model which is enumerated above as Model 3, "Fitted IPW." Compared to Simulation Study 2, in which population data are generated in an identical fashion, we struggle in this setting to recover the true sampling probabilities. Specifically, we underestimate the true sampling probabilities, which is unsurprising given that we fail to account for a factor that is positively associated with selection.

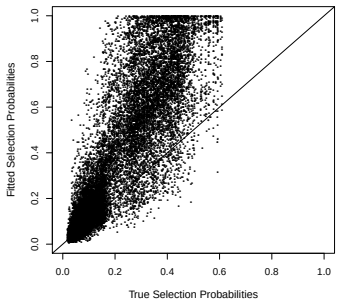
Results pertaining to the recovery of the parameter quantifying the association



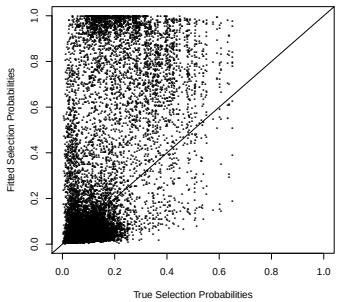
(a) Posterior mean vs. true sampling probabilities: Simulation Study 1.



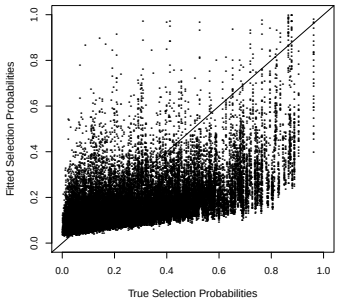
(b) Posterior mean vs. true sampling probabilities: Simulation Study 2.



(c) Posterior mean vs. true sampling probabilities without calibration component: Simulation Study 1 (no calibration).



(d) Posterior mean vs. true sampling probabilities without calibration component: Simulation Study 2 (no calibration).



(e) Posterior mean vs. true sampling probabilities: Simulation Study 3.

Figure 5.4: Posterior mean probabilities of selection vs. true probabilities of selection for 30 simulated data sets generated under the various simulation studies. Panels (a) and (b) refer to Simulation Studies 1 and 2 respectively, and plot the true vs. fitted selection probabilities for our full modeling framework. Panels (c) and (d) refer to Simulation Studies 3 and 4 respectively, and plot true vs. fitted selection probabilities under model 4, Fitted IPW (no calibration), in which we exclude the portion of the model that calibrates the sampling probabilities using areal data. Panel (e) presents the posterior mean probabilities of selection vs. true probabilities of selection for Simulation Study 3 and for our full modeling framework.

between exposure and disease (β_1) are displayed in Table 5.1. Specifically, we present: the bias of estimates of β_1; the percent reduction in bias compared to the naive analysis; the average empirical probability that a 95% credible/confidence interval for β_1 covers the true value; and the average length of a 95% credible/confidence interval for β_1, averaged over the 30 simulated data sets in each simulation study. Weighting by the fitted inverse sampling probabilities yields a substantial reduction in bias in estimating β_1, with this reduction being comparable to the gold standard, true IPW model. Furthermore, when we use the posterior mean sampling probabilities to construct sampling weights to be used in off-the-shelf software, there is little to no decrease in the model performance, both in terms of bias and coverage probability. Failing to adjust in any way for sampling bias of the EHR data, as well as failing to calibrate the sampling probabilities using areal data, yields highly biased results, highlighting the utility of our full modeling framework in settings where sampling bias is present in EHR data.

In Simulation Study 3, the reduction in the bias of the estimates of β_1 is less pronounced than in Simulation Study 2. This is unsurprising given that our model struggled to recover the data generating sampling probabilities when we fail to account for all of the factors associated with selection (see Figure 5.4 (e)). However, it highlights the importance in practice of considering a comprehensive list of factors that may be associated with selection in order to achieve unbiased inference on the association between disease and exposure.

Setting	Bias Correction	Bias β_1 (% reduction)	Coverage β_1	CI Length
Simulation Study 1	Naive	0.15 (0.0%)	63.3%	0.58
	True IPW	−0.05 (65.3%)	**93.3%**	0.62
	Fitted IPW	**0.05 (66.8%)**	**93.3%**	0.59
	Fitted IPW (No calibration)	−0.16 (−7.1%)	60.0%	0.78
	Fitted IPW (off-the-shelf)	0.05 (65.9%)	**93.3%**	0.61
Simulation Study 2	Naive	0.15 (0.0%)	80.0%	0.82
	True IPW	−0.05 (66.4%)	**96.7%**	1.09
	Fitted IPW	0.05 (66.1%)	**93.3%**	0.89
	Fitted IPW (No calibration)	0.35 (−132.7%)	50.0%	1.51
	Fitted IPW (off-the-shelf)	**0.04 (70.1%)**	**93.3%**	0.95
Simulation Study 3	Naive	0.12 (0.0%)	86.7%	0.82
	True IPW	**−0.04 (66.6%)**	**96.7%**	1.12
	Fitted IPW	0.08 (35.1%)	**96.7%**	0.91
	Fitted IPW (off-the-shelf)	0.08 (32.9%)	**93.3%**	0.87

Table 5.1: Results averaged over the 30 simulations generated under the three simulation studies: average bias for the exposure parameter, defined as the posterior mean of β_1 minus the true value of β_1 employed to generate the data; percent reduction in bias and the percent reduction in bias compared to the naive analysis; average empirical probability that a 95% credible/confidence interval for β_1 covers the true value; average length of a 95% credible/confidence interval for β_1. Numbers given in bold indicate the models that achieve the greatest reduction in bias relative to the naive model (column 3) and the coverage probability that is closest to nominal (column 4).

5.3.5 An analysis of smoking and lung cancer using subjects from Michigan Genomics Initiative

We will proceed with an analysis of the association between incident lung cancer in 2016-2018 and smoking status, adjusting for race and age. Our definition of lung cancer is provided in detail in Appendix D. EHR data refer to patients that are part of the Michigan Genomic Initiative (MGI) cohort, a set of subjects who have had an encounter with the University of Michigan Hospital System (UMHS) and consented to having their data user for research purposes. The data analysis includes all cases, defined as those who were diagnosed with lung cancer between 2016 and 2018, and a subset of 1,000 controls, defined as those who have not been diagnosed with lung cancer according to the EHR generated at their most recent interaction with UMHS. Additional inclusion criteria include subjects who identified their race as either White or Black. Due to extremely low representation, for our analysis we excluded all patients that belong to any race other than White or Black. In addition, we only included those those for whom smoking status was identified.

Table 5.2 presents descriptive statistics for the cases and controls from our data analysis, as well as for a more comprehensive subset of the MGI cohort. The latter is comprised of all MGI subjects who reside in the state of Michigan and for whom smoking status, race, ethnicity, and gender were all available. As out target population for this study is the entire state of Michigan, Table 5.2 also presents similar statistics for the entire state of Michigan.

Recall that from Westreich (2012) and Beesley et al. (2018), we should expect

biased analyses of the association between disease and exposure when sampling presence is associated with both disease and exposure. From Table 5.2, we see that, MGI subjects are less likely to be current smokers and more likely to be former smokers compared to the state of Michigan. Seemingly coincidentally, when combined into the categories of ever having smoked vs. never having smoked, MGI subjects are quite similar to the state of Michigan (46.6% for MGI vs. 46.2% for the State of Michigan). Despite this, it is clear that the smoking behavior of MGI subjects differs from that of the target population. Moreover, the 3-year incidence of lung cancer among MGI subjects is 0.5%, 2.5 times greater than the state-wide 3-year incidence of 0.2%, meaning that sampling presence is also associated with disease status. Therefore, a naive analysis of these data, i.e. one that does not adjust for the sampling mechanism, could yield biased estimates of the association between smoking and lung cancer.

Table 5.2 demonstrates that MGI subjects are more likely to be white, more likely to be non-Hispanic, and slightly more likely to be female than the target population. Table 5.3 presents descriptive statistics on the age of subjects in the MGI cohort. Compared to the state of Michigan as a whole, MGI subjects and, to a greater extent, MGI subjects with incident lung cancer, are older on average.

Figure 5.5 presents synthetic locations for subjects in the MGI database. To protect subject privacy, we were provided with each subject's county and the distance between their residential zip code centroid and the University of Michigan Medical Center. Although this information could potentially allow us to recover deterministically most if not all subject zip codes, we forego a procedure that does

Variable	Category	MGI Cohort	MGI Cases	MGI Controls	State of Michigan
Gender	Male	46.8%	56.7%	46.6%	49.2%
	Female	53.2%	43.3%	53.4%	50.8%
Smoking Status	Current Smoker	11.6%	10.2%	10.6%	20.4%
	Former Smoker	35.0%	67.0%	37.5%	25.8%
	Ever Smoker	46.6%	77.2%	48.1%	46.2%
	Never Smoker	53.4%	22.7%	51.8%	53.8%
Race	White	91.0%	95.5%	95.3%	78.7%
	Black	5.3%	4.5%	4.7%	13.9%
	Asian	1.6%	NA	NA	2.9%
	American Indian/ Alaska Native	0.4%	NA	NA	0.5%
	Native Hawaiian/ Pacific Islander	0.1%	NA	NA	0.3%
	Other	1.6%	NA	NA	1.2%
Ethnicity	Hispanic	2.0%	1.1%	2.1%	5.2%
	Non-Hispanic	98.0%	98.9%	97.9%	94.8%

Table 5.2: Gender, smoking status, race, and ethnicity for: (i) the full MGI cohort; (ii) cases in our data analysis who had incident lung cancer between 2016 and 2018; (iii) controls in our data analysis who had never had lung cancer according the EHR generated for their most recent visit; and (iv) the entire State of Michigan.

Source	Mean Age	Minimum Age	Maximum Age
MGI Full Data	56.0	18.0	103.0
MGI Cases	68.5	40.0	97.0
MGI Controls	56.0	19.0	97.0
State of Michigan	48.1	NA	NA

Table 5.3: Age of MGI subjects, broken up by lung cancer status, and age for Michigan residents. Note that for the state of Michigan, the metric presented is the mean age of adults (18+) in Michigan.

so in order to maintain the same level of deidentification as the original data set. Rather, we use the following procedure:

1. A list of possible zip codes is determined based on the subjects' county.
2. Using publicly available shapefiles, we compute on our own the distance between the zip-code centroid and the Michigan Medical Center.
3. For each possible zip code, we compute the absolute difference between the subjects' distance to the hospital and the zip code distance to the hospital.
4. A subject is randomly allocated to a zip code with probability inversely proportional to the difference in distances.
5. The subject is then assigned a random location within that zip code, which is used as its synthetic location.

In Figure 5.5, locations appear concentrated near the hospital, around Metropolitan Detroit, and to a lesser degree in highly populated areas of the state of Michigan.

To the 311 cases, who were all diagnosed with lung cancer between 2016 and 2018, and to the 1,000 controls, we fit the modeling framework presented in Section 5.2. Our outcome of interest is $Y(\mathbf{s}_i), i = 1, ..., 1311$, the binary indicator of incident lung cancer for a subject at location $\mathbf{s}_i$. Our exposure of interest, $X(\mathbf{s}_i)$ is a binary indicator of whether the subject a location $\mathbf{s}_i$ ever smoked cigarettes, whereas $Z_1(\mathbf{s}_i)$ and $Z_2(\mathbf{s}_i)$ denote respectively the race (either Black or White) and age of the subject at $\mathbf{s}_i, 1 = 1, ..., 1311$.

In addition to our full Bayesian hierarchical modeling framework, enumerated

Figure 5.5: Synthetic locations of EHR subjects.

as Model 3 in Section 5.3.4, we fit to these data the models 1 and 5 based on the enumeration and descriptions in Section 5.3.4.

Results relative to the application of models 1, 3, and 5 to the MGI data are presented in Table 5.4, including the parameter estimates, defined as the posterior means of each parameter, as well as the lower and upper limits of a 95% credible interval, which are obtained by taking the 2.5^{th} and 97.5^{th} percentile of the posterior samples of each parameter. For the naive model, we find that, after controlling for race and age, ever having smoked is associated with 4.7 (95% confidence interval: 3.4, 6.6) times higher odds of being diagnosed with lung cancer during the 3-year duration of the study. In the models that adjust for the sampling mechanism, that is, our full model, and the "Fitted IPW (off-the-shelf)" model, ever having smoked is associated respectively with 2.8 (95% credible interval: 1.6,4.5) and 2.9 (95% confidence interval: 1.6, 5.1) times higher odds of being diagnosed with lung cancer.

Hence, we observe that adjusting for the sampling probabilities yields odds ratios quantifying the association between smoking and lung cancer that are closer to the null value, although it is worth acknowledging that the confidence/credible intervals for the exposure-disease association parameter and other model parameters overlap for all three modeling frameworks. We may attribute attenuation of the exposure-disease association to the fact that sickest members of the target population are included in the EHR sample, thus corresponding to Berkson's bias in the most traditional sense. However, we want to point out that such a finding does not imply that previous analyses of the association between smoking and lung

Model	Parameter	Estimate	95% Confidence/Credible Interval	
			Lower Limit	Upper Limit
Naive	β_0	-5.90	-6.74	-5.06
	β_1	1.55	1.21	1.89
	γ_1	0.51	-0.18	1.21
	γ_2	0.06	0.05	0.07
Fitted IPW	β_0	-7.71	-8.99	-5.84
	β_1	1.04	0.50	1.50
	γ_1	-0.46	-1.89	0.62
	γ_2	0.08	0.05	0.10
Fitted IPW (off-the-shelf)	β_0	-7.51	-8.85	-6.18
	β_1	1.07	0.49	1.63
	γ_1	-0.38	-1.42	0.65
	γ_2	0.08	0.06	0.10

Table 5.4: Parameter estimates from three models estimating the association between smoking and the log odds incident lung cancer, adjusting for race and age. Columns denote the modeling framework; the parameter of interest; their estimates, defined as the posterior means of the parameters; the lower and upper limits of a 95% credible interval, obtained by taking the 2.5^{th} and 97.5^{th} percentiles of the posterior samples of each parameter. β_1 quantifies the association between ever smoking and incident lung cancer; γ_1 quantifies the association between race (Black vs. White) and incident lung cancer; γ_2 quantifies the association between age and incident lung cancer.

cancer have over-estimated their association.

5.4 Discussion

We have a presented a modeling framework that can be used to correct for sampling bias in EHR data through recovery of sampling probabilities and inverse probability weighting. Our procedure allows users to incorporate areal-level data to further calibrate sampling probabilities, a step which, in simulation studies, was shown to greatly improve the performance of our modeling framework. We demonstrated via simulation studies that the posterior means of the sampling probabilities match relatively well with the data generating probabilities, with the best performance occurring in simulations where all of the factors associated with the sampling mechanism are accounted for in the model. By then deriving

sampling weights using the sampling probabilities, we are able to reduce the bias of the parameter quantifying the association between disease and exposure. Our data analysis demonstrates how our modeling framework can be used for clinical case-control data.

This work is still in the early stages and a number of improvements could be made. First, the disease model, referred to earlier as the regression component of our full model, assumes that the disease status of the EHR subjects is conditionally independent in space. This assumption may be untenable in practice, possibly due to unmeasured spatial covariates that are associated with the disease. In this case, it would be appropriate to specify a spatial random effect in this modeling component in order to capture residual spatial variance and dependence that is unaccounted for in our original modeling framework. In the same vein, we may include a spatial random effect in the intensity function of the spatial point process that specifies the EHR subjects' location. In this case, the spatial point process would be a log Gaussian Cox process and not an IPP. We could also investigate further the model of Diggle et al. (2010) for our application, which would specify a shared spatial random effect between the point-process and regression components of the model.

Our specification of normal prior distributions for the Horwitz-Thompson estimators in the calibration component of our model is based on the asymptotic distribution of the estimator, which may not be appropriate in every instance. Greater care can be given to defining the prior distributions, although we are encouraged at this time by the demonstrated utility of this model component.

In a follow-up paper to Savitsky and Toth (2016), Leon-Novelo and Savitsky (2017) specify a joint Bayesian model for the data and their inclusion probabilities, which one might consider preferable to the pseudo-posterior approach of Savitsky and Toth (2016). We are open to exploring these and other means of fitting the model, including ones that will more readily allow users who are unfamiliar with Bayesian model fitting to apply our method to their data.

The scientific validity of our data analysis relies on the assumption that we have not overlooked in our model for the sampling mechanism any factors that are strongly associated with selection. Given the finding in Simulation Study 3 that failing to account for all factors associated with the sampling mechanism yields poor model performance, future analyses should consider a more comprehensive set of factors to model the intensity of the locations of EHR subjects. Future work will also explore how to modify appropriately the modeling framework when this assumption is violated.

It would be preferable to include the entire set of available MGI data, rather than only a subset of controls. We were also limited in terms of what covariates we could use in the model based on what factors were publicly available. For instance, it would have been beneficial to group smokers into "current," "former," and "never," however only rates of "ever having smoked" were available from State Cancer Profiles. Further exploration of publicly available data may yield an analysis that is of greater practical utility.

CHAPTER VI

Discussion

The availability of spatial data has increased dramatically in recent decades. These data provide opportunities for public health researchers to better understand environmental and social determinants of health. However, challenges are posed when classic spatial statistical models are unable to capture complex spatial dependencies in the data, when the spatial or spatio-temporal scale of various data sources are incompatible with one another, or when the data available to researchers are not representative of the target population. The overarching goal of this dissertation is to present a set of spatial statistical modeling frameworks that allow users to confront the practical issues that arise in newly emerging, often publicly available spatial data sets, including issues of dependence structure (Chapter II) and issues of scale/support (Chapters III - V). A secondary goal is to present modeling frameworks that accommodate the source of the data, including estimates from complex sampling surveys (Chapters III and IV), and data pulled from Electronic Health Records (Chapter V). In this concluding section, we will summarize the findings and scientific contributions of the four projects that

comprise this dissertation and present ideas for future work.

In Chapter II, we build off of the Multi-resolution Approximation (M-RA) of Katzfuss (2017), and present a Bayesian hierarchical model that allows users to both explore and accommodate spatial dependence structures that depart from the typical assumption of stationarity in geostatistical models. By specifying a mixture of normal prior distributions on the basis function weights, with one of the mixture components being centered at zero with small variance, users can better understand the dependence structure of the spatial process that underlies their data. Specifically, we demonstrate via multiple simulation studies and an analysis of Soil Organic Carbon (SOC) that regions of the domain whose spatial dependence is characterized by a slow rate of decay require a lower-resolution approximation than regions in which spatial dependence decays quickly. Thus, in regions where spatial dependence persists at large distances, basis function weights at higher levels are shrunk towards zero in our modeling framework. Furthermore, in our analysis of SOC, we demonstrate that the out-of-sample predictive performance of our model exceeds that of classical stationary modeling frameworks and matches that of an existing non-stationary modeling framework. Because of this, we not only accomplish our primary goal of introducing an exploratory tool that allows users to determine analytically specific regions of spatially varying range, we have also provided a spatial statistical modeling framework whose out-of-sample predictive performance for non-stationary spatial data exceeds several established models.

In Chapter II, we discussed the possibility of treating L as a random variable,

rather than treating it as fixed and tuning it as part of the burn-in period. In that same vein, the number and locations of the knots could also be treated as random. Although analogous techniques have been used on other areas of statistics (see, for example, Denis and Molinari (2010), who model a hazard function via a B-spline basis, with the number and locations of knots being sampled through reversible jump MCMC), to our knowledge, this would be relatively novel in the spatial statistical setting. Given the findings in Chapter II about the relationship between the number and spacing of knots and the scale of a spatial process, treating the number and locations of knots as random would be a natural next step to further our understanding of this relationship. Finally, while we have already proposed a spatio-temporal M-RA, a spatio-temporal *mixture M-RA* could be explored in future work.

In Chapter III, we introduce a Bayesian hierarchical modeling framework that addresses the spatio-temporal Change of Support Problem (COSP) for estimates derived from sampling surveys of community characteristics that are characterized as proportions (e.g. the proportion of people below the poverty level in a census tract). Our primary methodological contribution in Chapter III is the incorporation of each survey estimate's design effect into our modeling framework. In doing so, our model accounts for the design-based variance of the estimates while also permitting convenient distributional assumptions about the data. The former feature is rarely seen in spatial statistical modeling frameworks despite the frequency with which survey-based estimates are encountered in application. Furthermore, to our knowledge this is one of the first spatio-temporal change of support models

that accounts for survey design. An additional methodological contribution of this chapter is a spatio-temporal extension of the M-RA of Katzfuss (2017), which we have called the spatio-temporal multi-resolution approximation (ST-MRA).

In Chapter IV, we shift our focus to estimates of count-valued community characteristics, such as the number of births in a census tracts. Once again, we incorporate the survey's design effect in order to properly propagate the design based variance of the estimates. However, we found that simply adapting our modeling framework from Chapter IV for count-valued data was insufficient to handle the data we chose to illustrate our model: the number of births in census tracts in Michigan. In particular, to address the excess zeros that arise in the ACS data, we added to our model a simple yet effective spatio-temporal zero-inflation term. After doing so, we found that the predictive accuracy of our model was on par with that of Bradley et al. (2016b). Due to our modeling of the survey variance by incorporating the design effect, our prediction intervals achieved near nominal coverage.

Chapters III and III both utilize the ST-MRA to approximate the spatio-temporal random effect of the spatio-temporal process that we assume underlies our data. Additional novel representations of the spatio-temporal random effect, especially ones that are well-suited to modeling areal data in an efficient way without loss of statistical precision, are an avenue for future research. Furthermore, a more thorough study of the ST-MRA, perhaps for point-referenced spatio-temporal data, could be the motivation for a future paper.

Our modeling framework allows for covariate effects to be built into the large-

scale mean term $\mu_t(A_{ig})$, although we never exploited this option in our data analyses. It may be of interest to see if inference and prediction improve through the incorporation of additional covariates into the mean term. Likewise, a multivariate extension of our modeling framework, analogous to the model of Bradley et al. (2016a), could be developed moving forward.

In Chapter V, we present a modeling framework that allows clinical and population health researchers to use Electronic Health Record (EHR) data to greater utility by correcting the sampling bias that exists in them. Through stages of both estimation (via an inhomogeneous Poisson process) and calibration (using, in practice, publicly available aggregated data), we are able to recover sampling probabilities of simulated EHR data that are selected from a larger simulated population. By then constructing sampling weights that are inversely proportional to the sampling probabilities, we are able to reduce bias in an analysis of the association between disease and exposure. Our model is relatively unique in that it synthesizes data of both point-level and areal support. The issue of sampling bias in EHR data is a popular topic in statistics due to the recent increase of EHRs in clinical and epidemiological research. We hope that our contribution can move researchers toward the goal of unbiased inference in population health studies, therefore increasing the already great utility of EHR data.

This chapter provides numerous avenues for future work. Our simulation studies only illustrated the utility of our model under rather simple data generating mechanisms. For instance, the data generating probabilities in the first two simulation studies were simply logistic-linear functions of the outcome, exposure, and

distance, whereas the model assumed that the intensity of EHR locations had a log-linear relationship with the same factors. More rigorous studies will be required in order to evaluate our model under more complex (and perhaps realistic) sampling mechanisms. Our third simulation study, in which we introduced an unobserved factor that was associated with sampling probability, is a first step in that direction.

In our data analysis, we treated both the outcome and exposure data as fixed temporally, whereas in reality the MGI data provide us with a cohort that is monitored over time. Incorporation into our modeling framework of temporal components, either pertaining to exposure or disease progression, could provide us with a better understanding of the disease etiology.

As spatial and spatio-temporal data become larger and more widely available, major methodological developments in the field of spatial statistics aim to address issues that arise in practice, such as high dimensionality, complex dependence structures, incompatible spatial support, and non-random sampling. This dissertation follows this trend by presenting modeling frameworks that address practical issues of dependence, scale, and source in spatial and spatio-temporal data. We explore these concepts thoroughly through the course of these four projects, and make methodological and practical contributions that have relevance within and beyond the field of spatial statistics. Furthermore, they establish the groundwork for additional specialized and well-refined methodologies to be developed moving forward.

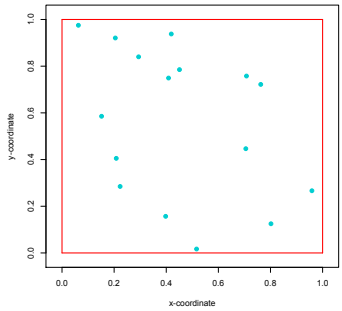
APPENDICES

APPENDIX A

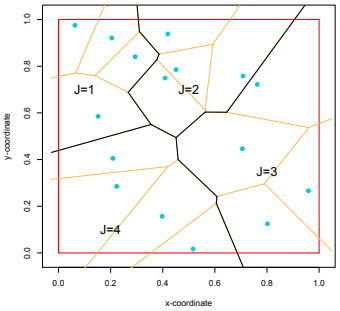
Additional Material for *Chapter II: Identifying Regions of Non-stationarity in Spatial Processes via Multi-resolution Approximation and Mixture Priors*

A.1 Example of knot placement using Voronoi tessellation

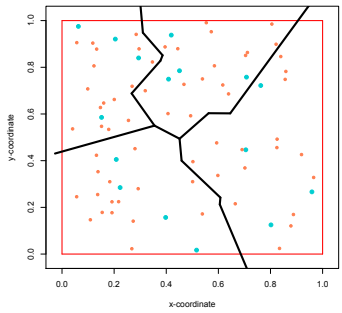
As a means of providing partitions and knots that are not strictly placed on a rectangular grid, a possibility for knot placement would be to do the following: first, at level $m = 0$, introduce r knots randomly located across the study domain; then, use these r knots to define a Voronoi tessellation of the domain. Using the Voronoi tessellation, define J non-overlapping subregions at level $m = 1$ as union of distinct, complete, Voronoi polygons. In each of the J subregions, introduce again r knots, randomly sampling them within each of the J subregions. These new set of knots will be used to define a new Voronoi tessellation within each of the J subregions. The procedure will the continue in a similar fashion for further levels. Figure A.1 shows an example of the Voronoi tessellation idea for the placement of knots in the Mixture M-RA model.



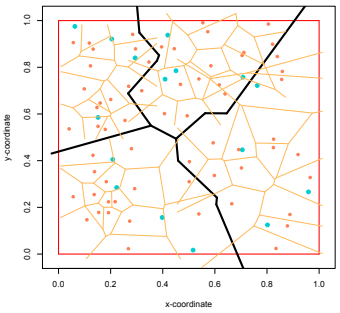
(a) Knots at level $m = 0$.



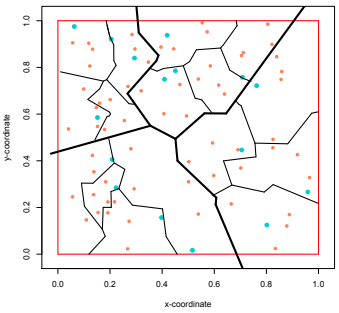
(b) Voronoi tessellation at level $m = 0$.



(c) Partitions and knots at level $m = 1$.



(d) Voronoi tessellation at level $m = 1$.



(e) Partitions at level $m = 2$.

Figure A.1: Illustration of Voronoi-defined partitions for the M-RA mixture model on the unit square: (a) r=16 knots introduced at level m=0; (b) Voronoi tessellation defined using the knots at level m=0; (c) J=4 partitions at level m=1 and r=16 knots introduced within each partition; (d) Voronoi tessellation defined by the r=16 knots introduced in (c); (e) partitions at level m=2.

A.2 Proofs

The following sections provide the form of the covariance between the M-level process $\tilde{w}_M(\mathbf{s}) = \sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(\mathbf{s})\boldsymbol{\eta}_{m,j}$ at two locations $\mathbf{s}$ and $\mathbf{s}' \in \mathcal{S}$ in the M-RA and the mixture M-RA model, respectively.

Covariance of $\tilde{w}_M(\mathbf{s}), \tilde{w}_M(\mathbf{s}')$ under the M-RA model

Under the M-RA model, at each point $\mathbf{s} \in \mathcal{S}$, $\tilde{w}_M(\mathbf{s}) = \sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(\mathbf{s})\boldsymbol{\eta}_{m,j}$ where $\boldsymbol{\eta}_{m,j} \sim N(0, \mathbf{K}_{m,j})$. Given these assumptions and the fact that basis function weights corresponding to different subregions are independent, it follows that:

$$\begin{aligned}
\mathrm{Cov}_{M-RA}(\tilde{w}_M(\mathbf{s}), \tilde{w}_M(\mathbf{s}')) &= E\left[\left(\sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(\mathbf{s})\boldsymbol{\eta}_{m,j}\right)\left(\sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(\mathbf{s}')\boldsymbol{\eta}_{m,j}\right)\right] \\
&= \sum_{m=0}^{M}\sum_{j=1}^{J^m} E\left[(\mathbf{b}_{m,j}(\mathbf{s})\boldsymbol{\eta}_{m,j})(\mathbf{b}_{m,j}(\mathbf{s}')\boldsymbol{\eta}_{m,j})\right] \\
&= \sum_{m=0}^{M}\sum_{j=1}^{J^m} E\left[(\mathbf{b}_{m,j}(\mathbf{s})\boldsymbol{\eta}_{m,j})(\boldsymbol{\eta}'_{m,j}\mathbf{b}_{m,j}(\mathbf{s}')')\right] \\
&= \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s}) E\left[\boldsymbol{\eta}_{m,j}\boldsymbol{\eta}'_{m,j}\right] \mathbf{b}_{m,j}(\mathbf{s}')' \\
&= \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s})\mathbf{K}_{m,j}\mathbf{b}_{m,j}(\mathbf{s}')
\end{aligned}$$

Covariance of $\tilde{w}(\mathbf{s}), \tilde{w}(\mathbf{s}')$ under the mixture M-RA model

Under the mixture M-RA model, the definition of $\tilde{w}_M(\mathbf{s})$ for $\mathbf{s} \in \mathcal{S}$ is still of the form $\tilde{w}_M(\mathbf{s}) = \sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(\mathbf{s})\boldsymbol{\eta}_{m,j}$, however the prior distribution for $\boldsymbol{\eta}_{m,j}$ is now of the form $\boldsymbol{\eta}_{m,j} \sim p_m N_r(0, \mathbf{K}_{m,j}) + (1-p_m)N_r(0, \mathbf{K}_{m,j}/L)$ with $p_m = \rho^m$ and

$\rho \sim \text{Beta}(\alpha_\rho, \beta_\rho)$.

Since

$$\begin{aligned} E(\boldsymbol{\eta}_{m,j}|p_m) &= 0 \\ \text{Var}(\boldsymbol{\eta}_{m,j}|p_m) &= E(\boldsymbol{\eta}_{m,j}\boldsymbol{\eta}'_{m,j}|p_m) \\ &= p_m \mathbf{K}_{m,j} + (1-p_m)\mathbf{K}_{m,j}/L \end{aligned}$$

it follows that the marginal variance $\text{Var}(\boldsymbol{\eta}_{m,j})$ is given by

$$\begin{aligned} \text{Var}(\boldsymbol{\eta}_{m,j}) &= E(\text{Var}(\boldsymbol{\eta}_{m,j}|p_m)) + \text{Var}(E(\boldsymbol{\eta}_{m,j}|p_m)) \\ &= E(p_m \mathbf{K}_{m,j} + (1-p_m)\mathbf{K}_{m,j}/L) \\ &= E(p_m)\mathbf{K}_{m,j} + (1-E(p_m))\mathbf{K}_{m,j}/L) \end{aligned} \tag{A.1}$$

Before proceeding further, we wish to examine more closely our definition of p_m. Recall that, when a set of basis function weights is shrunk towards zero, the basis function weights within that partition at all subsequent levels are also shrunk towards zero, i.e. for those subsequent levels, $p_m = 0$. Otherwise, $p_m = \rho^m$, with $\rho \sim \text{Beta}(\alpha_\rho, \beta_\rho)$.

For clarity, we present several cases, followed by a general form for p_m for

$m = 1, ..., M$:

$$
\begin{aligned}
p_0 &= 1 \\
p_1 &= \rho \\
p_2 &= \begin{cases} \rho^2 & \text{with probability } \rho \\ 0 & \text{with probability } 1-\rho \end{cases} \\
p_3 &= \begin{cases} \rho^3 & \text{with probability } \rho^2\rho \\ 0 & \text{with probability } 1-\rho^3 \end{cases} \\
p_4 &= \begin{cases} \rho^4 & \text{with probability } \rho^3\rho^2\rho \\ 0 & \text{with probability } 1-\rho^6 \end{cases} \\
p_m &= \begin{cases} \rho^m & \text{with probability } \rho^{m(m-1)/2} \\ 0 & \text{with probability } 1-\rho^{m(m-1)/2} \end{cases}
\end{aligned}
$$

Therefore, from $p_m = \rho^m$, with $\rho \sim \text{Beta}(\alpha_\rho, \beta_\rho)$, it follows:

$$
\begin{aligned}
E(p_m) &= E(\rho^{m(m-1)/2}\rho^m + (1-\rho^{m(m-1)/2})0) \\
&= E(\rho^{m(m+1)/2}) \\
&= \int_{\rho=0}^{1} \frac{\Gamma(\alpha_\rho+\beta_\rho)}{\Gamma(\alpha_\rho)\Gamma(\beta_\rho)} \rho^{m(m+1)/2}\rho^{\alpha_\rho-1}(1-\rho)^{\beta_\rho-1}d\rho \\
&= \frac{\Gamma(\alpha_\rho+\beta_\rho)}{\Gamma(\alpha_\rho)\Gamma(\beta_\rho)} \int_{\rho=0}^{1} \rho^{\alpha_\rho+m(m+1)/2-1}(1-\rho)^{\beta_\rho-1}d\rho \\
&= \frac{\Gamma(\alpha_\rho+\beta_\rho)}{\Gamma(\alpha_\rho)\Gamma(\beta_\rho)} \times \frac{\Gamma(\alpha_\rho+m(m+1)/2)\Gamma(\beta_\rho)}{\Gamma(\alpha_\rho+m(m+1)/2+\beta_\rho)} \\
&= \frac{\Gamma(\alpha_\rho+\beta_\rho)\Gamma(\alpha_\rho+m(m+1)/2)}{\Gamma(\alpha_\rho)\Gamma(\alpha_\rho+m(m+1)/2+\beta_\rho)} \qquad \text{(A.2)}
\end{aligned}
$$

Calling

$$G(\alpha_\rho, \beta_\rho, m) := \begin{cases} 1 & m = 0 \\ \frac{\Gamma(\alpha_\rho+\beta_\rho)\Gamma(\alpha_\rho+m(m+1)/2)}{\Gamma(\alpha_\rho)\Gamma(\alpha_\rho+m(m+1)/2+\beta_\rho)} & m = 1, ..., M \end{cases}$$

and substituting it into (A.2) and (A.1), it yields:

$$\begin{aligned} \mathrm{Var}(\boldsymbol{\eta}_{m,j}) &= E(p_m)\mathbf{K}_{m,j} + (1 - E(p_m))\mathbf{K}_{m,j}/L) \\ &= G(\alpha_\rho, \beta_\rho, m)\mathbf{K}_{m,j} + (1 - G(\alpha_\rho, \beta_\rho, m))\,\mathbf{K}_{m,j}/L \end{aligned} \tag{A.3}$$

From (A.3), it follows that for any pair of sites $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$,

$$\begin{aligned} \mathrm{Cov}_{mM-RA}(\tilde{w}_M(\mathbf{s}), \tilde{w}_M(\mathbf{s}')) &= E\left[\left(\sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s})\boldsymbol{\eta}_{m,j}\right)\left(\sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s}')\boldsymbol{\eta}_{m,j}\right)\right] \\ &= \sum_{m=0}^{M}\sum_{j=1}^{J^m} E\left[(\mathbf{b}_{m,j}(\mathbf{s})\boldsymbol{\eta}_{m,j})(\mathbf{b}_{m,j}(\mathbf{s}')\boldsymbol{\eta}_{m,j})\right] \\ &= \sum_{m=0}^{M}\sum_{j=1}^{J^m} E\left[(\mathbf{b}_{m,j}(\mathbf{s})\boldsymbol{\eta}_{m,j})(\boldsymbol{\eta}'_{m,j}\mathbf{b}_{m,j}(\mathbf{s}')')\right] \\ &= \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s}) E\left[\boldsymbol{\eta}_{m,j}\boldsymbol{\eta}'_{m,j}\right] \mathbf{b}_{m,j}(\mathbf{s}')' \\ &= \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s})\,[G(\alpha_\rho, \beta_\rho, m)\mathbf{K}_{m,j} + \\ &\qquad (1 - G(\alpha_\rho, \beta_\rho, m))\,\mathbf{K}_{m,j}/L]\,\mathbf{b}_{m,j}(\mathbf{s}')' \end{aligned}$$

A.3 Simulation study 5

In this simulation study, we aim to determine whether our model, which has been developed to detect inhomogeneities in the spatial dependence structure of a Gaussian process, can also be used with data that are a realization of a stationary Gaussian process. In particular, we are interested in determining whether

in this situation our model does indicate that there are no inhomogeneities in the spatial correlation structure. To assess this, we simulate 30 realizations of a spatial Gaussian process according to (1) at $1,012$ locations on $\mathcal{S}$ =[0, 1] $\times$ [0,1]. For the simulations we assume that $\mu(\mathbf{s}) \equiv 0$, $\forall \mathbf{s} \in \mathcal{S}$, $\tau^2 = 0.05$, while $w(\mathbf{s})$ is a mean-zero, stationary Gaussian process with Matérn covariance function with parameters $\sigma^2 = 1$, $\nu = 1$, and $\phi = 0.1$. The figure below presents one of the 30 realizations generated in this simulation study. As in simulations study 2 and simulation study 3, to assess the out-of-sample predictive performance of the mixture M-RA model, for each simulated dataset, we fit the mixture M-RA model to 756 randomly chosen locations, while we hold out data at 256 sites. Again, we compare the out-of-sample predictive performance of our model to that of a stationary Bayesian Kriging model, a standard model used for the analysis of stationary spatial data. For both the mixture M-RA model and the stationary Bayesian Kriging model, we employ a stationary Matèrn covariance function to define the basis functions and to characterize the spatial dependence of the spatial random effects.

We fit the mixture M-RA model using $M = 3$, $J = 4$, and $r = 16$. As for any simulation study except simulation study 1, also in this case, we estimate the covariance parameter vector $\boldsymbol{\theta}$. For each dataset, we run the MCMC algorithm for 10,000 iterations, and we discard the first 5,000 for burn-in. Looking at the posterior mean of the basis function weights, we notice that for each of the 30 simulated datasets, the basis function weights do not mix into the second component of the mixture prior, thus indicating that the number of M-RA levels needed, and hence

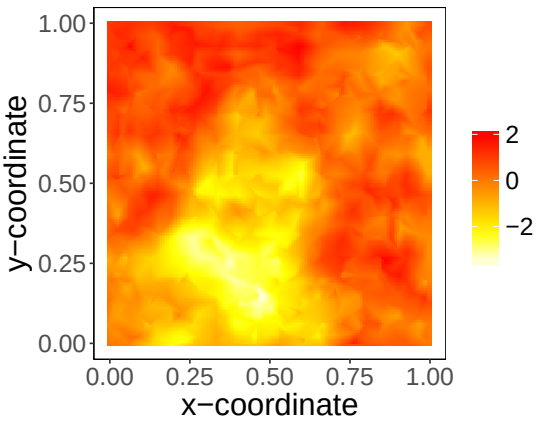


Figure A.2: Illustration of spatial data simulated using a stationary and isotropic spatial covariance function.

the strength of the spatial dependence, is constant across space: the average of the posterior means of the latent binary random variables $Z_{m,j}$'s, indicating the mixture component for each $\boldsymbol{\eta}_{m,j}$, range between 0.98 and 1.0 in the 30 simulation experiments. In addition, the standard deviation of the posterior means of the $Z_{m,j}$'s ranges from 0 to 0.03 across the 30 simulated datasets, indicating that in general there is very little variability in terms of which mixture component the basis function weights $\boldsymbol{\eta}_{m,j}$ are drawn from. In light of these results, we conclude that our model could also be used for stationary spatial data: in this case the model would indicate no inhomogeneities as the number of levels needed nor the strength of the spatial correlation does not change over space. The adequacy of the mixture M-RA as a model to analyze stationary spatial data is also confirmed by its out of sample predictive performance: while the average MSPE, averaged across the 30 simulations, for the stationary Bayesian Kriging model is 0.108, the

average MSPE for the mixture M-RA model is 0.103.

In summary, this simulation study suggests that not only the mixture M-RA model does not miss-identify stationary data as non-stationary, but it can also provide out-of-sample predictions that are comparable to those obtained using a stationary spatial model. We conclude this section noting that formal tests of non-stationarity by Li et al. (2008), Jun and Genton (2012), and Fuentes (2005) may be more straightforward if a practitioner's only goal is to determine whether or not the spatial process from which data are available is non-stationary.

APPENDIX B

Additional Material for *Chapter III: Accounting for Survey Design in Bayesian Disaggregation of Survey-based Estimates of Proportions*

B.1 Derivation of Covariance Under the Spatio-temporal Multi-resolution Approximation

In this Section, we present proofs that the ST-MRA provides an approximation for a separable space-time covariance function with an AR(1) dependence structure in time, and a spatial covariance function $C(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$ for $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$.

Recall the ST-MRA:

$$w_t(\mathbf{s}) \approx \sum_{m=0}^{M} \sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s}) \boldsymbol{\eta}_{t,m,j}$$

where the sum is taken over partitions (indexed by j) and levels (indexed by m). At time $t = 0$, $\boldsymbol{\eta}_{0,m,j} \sim N_r(0, \mathbf{K}_{m,j})$, as under the MRA construction of Katzfuss (2017). Assuming a stationary first-order autoregressive structure on the basis function weights Gelfand et al. (2005) for $t = 1, \ldots, T$ yields:

$$\begin{aligned} \boldsymbol{\eta}_{t,m,j} | \boldsymbol{\eta}_{t-1,m,j}, \boldsymbol{\eta}_{t-2,m,j}, \ldots, \boldsymbol{\eta}_{0,m,j} &\sim N_r(\alpha \boldsymbol{\eta}_{t-1,m,j}, \mathbf{U}_{m,j}) \\ \mathbf{U}_{m,j} &= (1 - \alpha^2) \mathbf{K}_{m,j} \end{aligned}$$

Alternatively, we can write $\boldsymbol{\eta}_{t,m,j}$ as $\boldsymbol{\eta}_{t,m,j} = \alpha\boldsymbol{\eta}_{t-1,m,j} + \mathbf{e}_{t,m,j}$, with $\mathbf{e}_{t,m,j} \sim N_r(0, \mathbf{U}_{m,j})$. $\mathbf{e}_{t,m,j}$ are mutually independent and independent of $\boldsymbol{\eta}_{t,m,j}$ for all m, j.

Remark: *The estimated marginal variance under the ST-MRA is precisely the estimated variance under the MRA:*

$$\text{Var}(w_t(\mathbf{s})) \approx \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s})\mathbf{K}_{m,j}\mathbf{b}_{m,j}(\mathbf{s})^T$$

Proof.

$$\begin{aligned}
\text{Var}(w_t(\mathbf{s})) &\approx \text{Var}\left(\sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s})\boldsymbol{\eta}_{t,m,j}\right) \\
&= \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s})\left[\text{Var}\left(\boldsymbol{\eta}_{t,m,j}\right)\right]\mathbf{b}_{m,j}(\mathbf{s})^T \\
&= \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s})\left[\text{Var}\left(\alpha\boldsymbol{\eta}_{t-1,m,j} + \mathbf{e}_{t,m,j}\right)\right]\mathbf{b}_{m,j}(\mathbf{s})^T \\
&= \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s})\left[\text{Var}\left(\alpha\left(\alpha\boldsymbol{\eta}_{t-2,m,j} + \mathbf{e}_{t-1,m,j}\right) + \mathbf{e}_{t,m,j}\right)\right]\mathbf{b}_{m,j}(\mathbf{s})^T \\
&= \ldots \\
&= \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s})\left[\text{Var}\left(\alpha^t\boldsymbol{\eta}_{0,m,j} + \sum_{i=1}^{t-1}\alpha^i\mathbf{e}_{t-i,m,j}\right)\right]\mathbf{b}_{m,j}(\mathbf{s})^T \\
&= \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s})\left[\alpha^{2t}\text{Var}\left(\boldsymbol{\eta}_{0,m,j}\right) + \sum_{i=1}^{t-1}\alpha^{2i}\text{Var}\left(\mathbf{e}_{t-i,m,j}\right)\right]\mathbf{b}_{m,j}(\mathbf{s})^T \\
&= \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s})\left[\alpha^{2t}\mathbf{K}_{m,j} + \sum_{i=1}^{t-1}\alpha^{2i}\mathbf{U}_{m,j}\right]\mathbf{b}_{m,j}(\mathbf{s})^T \\
&= \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s})\left[\alpha^{2t}\mathbf{K}_{m,j} + \frac{\alpha^{2t}-1}{\alpha^2-1}(1-\alpha^2)\mathbf{K}_{m,j}\right]\mathbf{b}_{m,j}(\mathbf{s})^T \\
&= \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s})\mathbf{K}_{m,j}\mathbf{b}_{m,j}(\mathbf{s})^T
\end{aligned}$$

□

Remark: *Under the ST-MRA, the covariance of* $w_t(\mathbf{s})$ *and* $w_t(\mathbf{s}')$ *is:*

$$\text{(B.1)} \qquad \text{Cov}(w_t(\mathbf{s}), w_t(\mathbf{s}')) \approx \sum_{m=0}^{M} \sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s}) \mathbf{K}_{m,j} \mathbf{b}_{m,j}(\mathbf{s}')^T$$

Proof.

$$\begin{aligned}
\text{Cov}(w_t(\mathbf{s}), w_t(\mathbf{s}')) &= E(w_t(\mathbf{s}) w_t(\mathbf{s}')) - E(w_t(\mathbf{s})) E(w_t(\mathbf{s}')) \\
&\approx E\left[\left(\sum_{m=0}^{M} \sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s}) \boldsymbol{\eta}_{t,m,j} \right) \left(\sum_{m=0}^{M} \sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s}') \boldsymbol{\eta}_{t,m,j} \right) \right] \\
&= \sum_{m=0}^{M} \sum_{j=1}^{J^m} E\left[\left(\mathbf{b}_{m,j}(\mathbf{s}) \boldsymbol{\eta}_{t,m,j} \right) \left(\mathbf{b}_{m,j}(\mathbf{s}') \boldsymbol{\eta}_{t,m,j} \right) \right] \\
&= \sum_{m=0}^{M} \sum_{j=1}^{J^m} E\left[\mathbf{b}_{m,j}(\mathbf{s}) \boldsymbol{\eta}_{t,m,j} \boldsymbol{\eta}_{t,m,j}^T \mathbf{b}_{m,j}(\mathbf{s}')^T \right] \\
&= \sum_{m=0}^{M} \sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s}) \text{Var}(\boldsymbol{\eta}_{t,m,j}) \mathbf{b}_{m,j}(\mathbf{s}')^T \\
&= \sum_{m=0}^{M} \sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s}) \mathbf{K}_{m,j} \mathbf{b}_{m,j}(\mathbf{s}')^T
\end{aligned}$$

□

Remark: *Under the ST-MRA, the covariance between* $w_t(\mathbf{s})$ *and* $w_{t+j}(\mathbf{s})$ *is given by:*

$$\text{Cov}(w_t(\mathbf{s}), w_{t+j}(\mathbf{s})) \approx \alpha^j \sum_{m=0}^{M} \sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s}) \mathbf{K}_{m,j} \mathbf{b}_{m,j}(\mathbf{s})^T$$

Proof.

$$\begin{aligned}
\mathrm{Cov}(w_t(\mathbf{s}), w_{t+j}(\mathbf{s})) &= E(w_t(\mathbf{s})w_{t+j}(\mathbf{s})) - E(w_t(\mathbf{s}))E(w_{t+j}(\mathbf{s})) \\
&\approx E\left[\left(\sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(\mathbf{s})\boldsymbol{\eta}_{t,m,j}\right)\left(\sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(\mathbf{s})\boldsymbol{\eta}_{t+j,m,j}\right)\right] \\
&= E\left[\left(\sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(\mathbf{s})\boldsymbol{\eta}_{t,m,j}\right)\left(\sum_{m=0}^{M}\sum_{j=1}^{J^m}\left(\alpha^j\mathbf{b}_{m,j}(\mathbf{s})\boldsymbol{\eta}_{t,m,j} + \sum_{i=1}^{j}\alpha^i\mathbf{e}_{t,m,j}\right)\right)\right] \\
&= \alpha^j\sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(\mathbf{s})E(\boldsymbol{\eta}_{t,m,j}\boldsymbol{\eta}_{t,m,j}^T)\mathbf{b}_{m,j}(\mathbf{s})^T \\
&= \alpha^j\sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(\mathbf{s})\mathrm{Var}(\boldsymbol{\eta}_{t,m,j})\mathbf{b}_{m,j}(\mathbf{s})^T \\
&= \alpha^j\sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(\mathbf{s})\mathbf{K}_{m,j}\mathbf{b}_{m,j}(\mathbf{s})^T
\end{aligned}$$

□

Remark: *Under the ST-MRA, the covariance between* $w_t(\mathbf{s})$ *and* $w_{t+j}(\mathbf{s}')$ *is given by:*

$$\mathrm{Cov}(w_t(\mathbf{s}), w_{t+j}(\mathbf{s}')) \approx \alpha^j\sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(\mathbf{s})\mathbf{K}_{m,j}\mathbf{b}_{m,j}(\mathbf{s}')^T$$

Proof.

$$\begin{aligned}
\mathrm{Cov}(w_t(\mathbf{s}), w_{t+j}(\mathbf{s}')) &= E(w_t(\mathbf{s})w_{t+j}(\mathbf{s}')) - E(w_t(\mathbf{s}))E(w_{t+j}(\mathbf{s}')) \\
&\approx E\left[\left(\sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(\mathbf{s})\boldsymbol{\eta}_{t,m,j}\right)\left(\sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(\mathbf{s}')\boldsymbol{\eta}_{t+j,m,j}\right)\right] \\
&= E\left[\left(\sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(\mathbf{s})\boldsymbol{\eta}_{t,m,j}\right)\left(\sum_{m=0}^{M}\sum_{j=1}^{J^m}\left(\alpha^j\mathbf{b}_{m,j}(\mathbf{s}')\boldsymbol{\eta}_{t,m,j} + \sum_{i=1}^{j}\alpha^i\mathbf{e}_{t,m,j}\right)\right)\right] \\
&= \alpha^j\sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(\mathbf{s})E(\boldsymbol{\eta}_{t,m,j}\boldsymbol{\eta}_{t,m,j}^T)\mathbf{b}_{m,j}(\mathbf{s}')^T \\
&= \alpha^j\sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(\mathbf{s})\mathrm{Var}(\boldsymbol{\eta}_{t,m,j})\mathbf{b}_{m,j}(\mathbf{s}')^T \\
&= \alpha^j\sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(\mathbf{s})\mathbf{K}_{m,j}\mathbf{b}_{m,j}(\mathbf{s}')^T
\end{aligned}$$

□

B.2 Marginal Covariance of the Effective Number of Cases

We begin by briefly restating our modeling framework. After computing the effective sample sizes and effective number of cases, $m_t^{*(5)}(A_{ig})$ and $q_t^{*(5)}(A_{ig})$ respectively, we assume the following conditional distribution on the effective number of cases:

$$\begin{aligned} q_t^{*(5)}(A_{ig})|\ \pi_t^{(5)}(A_{ig}) &\overset{ind}{\sim} \text{Binomial}\left(m_t^{*(5)}(A_{ig}), \pi_t^{(5)}(A_{ig})\right) \\ q_t^{*(1)}(A_i)|\ \pi_t^{(1)}(A_i) &\overset{ind}{\sim} \text{Binomial}\left(m_t^{*(1)}(A_i), \pi_t^{(1)}(A_i)\right) \end{aligned} \tag{B.2}$$

with $g = 1, \ldots, G_i$ and $i = 1, \ldots, I$. We proceed by stating the $\pi_t^{(5)}(A_{ig})$ and $\pi_t^{(1)}(A_i)$ in terms of our desired spatio-temporal resolution, that is the 1-year census tract resolution or $\pi_t^{(1)}(A_{ig})$:

$$\begin{aligned} \pi_t^{(5)}(A_{ig}) &= \frac{1}{5}\sum_{k=t-4}^{t} \pi_k^{(1)}(A_{ig}) \\ \pi_t^{(1)}(A_i) &= \frac{1}{N_t(A_i)}\sum_{h=1}^{G_i} N_t(A_{ih})\pi_t^{(1)}(A_{ih}) \end{aligned}$$

with $g = 1, \ldots, G_i$ and $i = 1, \ldots, I$. Let the set of all $\pi_t^{(1)}(A_{ig})$ be denoted as

$\mathbf{\Pi} := \{\pi_t^{(1)}(A_{ig})\}_{t=1,\ldots,T, i=1,\ldots,I, g=1,\ldots,G_i}$. Then:

$$
\begin{aligned}
q_t^{*(5)}(A_{ig})|\,\mathbf{\Pi} &\overset{ind}{\sim} \text{Binomial}\left(m_t^{*(5)}(A_{ig}), \frac{1}{5}\sum_{k=t-4}^{t}\pi_k^{(1)}(A_{ig})\right) \\
q_t^{*(1)}(A_i)|\,\mathbf{\Pi} &\overset{ind}{\sim} \text{Binomial}\left(m_t^{*(1)}(A_i), \frac{1}{N_t(A_i)}\sum_{h=1}^{G_i}N_t(A_{ih})\pi_t^{(1)}(A_{ih})\right) \\
\Phi^{-1}\left(\pi_t^{(1)}(A_{ig})\right) &= \frac{1}{|A_{ig}|}\int_{\mathbf{s}\in A_{ig}}\left(\mu_t(\mathbf{s}) + w_t(\mathbf{s})\right)d\mathbf{s} + \xi(C_{A_{ig}}) \\
\Phi^{-1}\left(\pi_t^{(1)}(A_{ig})\right) &\approx \mu_t + \sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{t,m,j} + \xi(C_{A_{ig}}) + \epsilon_t(A_{ig}) \\
\epsilon_t(A_{ig}) &\overset{iid}{\sim} N(0,\tau^2) \\
\xi(C_{A_{ig}}) &\overset{iid}{\sim} N(0,\tau_C^2)
\end{aligned}
$$

The basis function weights $\boldsymbol{\eta}_{t,m,j}$ vary in time and we provide them with a stationary first-order autoregressive structure. At time $t = 1$ we assume $\boldsymbol{\eta}_{1,m,j} \sim N_r(\mathbf{0}, \mathbf{K}_{m,j})$ while for $t = 2, \ldots, T$:

$$
\begin{aligned}
\text{(B.3)} \qquad \boldsymbol{\eta}_{t,m,j}|\boldsymbol{\eta}_{t-1,m,j}, \boldsymbol{\eta}_{t-2,m,j}, \ldots, \boldsymbol{\eta}_{1,m,j} &\sim N_r(\alpha\boldsymbol{\eta}_{t-1,m,j}, \mathbf{U}_{m,j}) \\
\mathbf{U}_{m,j} &= (1-\alpha^2)\mathbf{K}_{m,j}
\end{aligned}
$$

In Appendix B.1, we showed that the covariance between $w_t(\mathbf{s})$ and $w_{t+j}(\mathbf{s}')$, that is the covariance between the spatio-temrpoal random effect at two locations $\mathbf{s}$ and $\mathbf{s}'$ at times t and $t+j$ respectively, is approximated as:

$$
\text{Cov}(w_t(\mathbf{s}), w_{t+j}(\mathbf{s}')) \approx \alpha^j \sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(\mathbf{s})\mathbf{K}_{m,j}\mathbf{b}_{m,j}(\mathbf{s}')^T
$$

In aggregating the model from the point-referenced spatial resolution ($\mathbf{s}$) to the areal resolution (A), we approximate numerically the average of the basis functions over all locations $\mathbf{s} \in A$. The basis function *weights* remain a point-referenced

quantity, corresponding to the knot locations, with the same covariance as in (B.3). It follows that the covariance of the spatial random effect approximation $w_t(A_{ig}) \approx \sum_{m=0}^{M} \sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{t,m,j}$ between two census tracts A_{ig} and A_{ih} at time t and $t+j$ respectively is:

$$\text{Cov}(w_t(A_{ig}), w_{t+j}(A_{ih})) \approx \alpha^j \sum_{m=0}^{M} \sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ig})\mathbf{K}_{m,j}\mathbf{b}_{m,j}(A_{ih})^T$$

Covariance between 1-year PUMA and 5-year census tract ENCs

We proceed by computing the covariance between the 1-year and 5-year effective number of cases given in (B.2), that is $\text{Cov}\left(q_t^{*(5)}(A_{ig}), q_t^{*(1)}(A_i)\right)$. From the law of total covariance, it follows that:

$$\begin{aligned}
\text{Cov}\left(q_t^{*(5)}(A_{ig}), q_t^{*(1)}(A_i)\right) &= E\left(\text{Cov}\left(q_t^{*(5)}(A_{ig}), q_t^{*(1)}(A_i)|\ \boldsymbol{\Pi}\right)\right) + \text{Cov}\left(E\left(q_t^{*(5)}(A_{ig})|\ \boldsymbol{\Pi}\right), E\left(q_t^{*(1)}(A_i)|\ \boldsymbol{\Pi}\right)\right) \\
&= \text{Cov}\left(m_t^{*(5)}(A_{ig}) \times \pi_t^{(5)}(A_{ig}), m_t^{*(1)}(A_i) \times \pi_t^{(1)}(A_i)\right) \\
&= \text{Cov}\left(m_t^{*(5)}(A_{ig}) \times \frac{1}{5}\sum_{k=t-4}^{t} \pi_k^{(1)}(A_{ig}), m_t^{*(1)}(A_i) \times \frac{1}{N_t(A_i)} \sum_{h=1}^{G_i} N_t(A_{ih})\pi_t^{(1)}(A_{ih})\right) \\
&= \frac{m_t^{*(5)}(A_{ig})}{5} \times \frac{m_t^{*(1)}(A_i)}{N_t(A_i)} \times \text{Cov}\left(\sum_{k=t-4}^{t} \pi_k^{(1)}(A_{ig}), \sum_{h=1}^{G_i} N_t(A_{ih})\pi_t^{(1)}(A_{ih})\right) \\
&\propto \sum_{k=t-4}^{t}\sum_{h=1}^{G_i} N_t(A_{ih}) \times \text{Cov}\left(\Phi\left(\mu_k + \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{k,m,j} + \xi(C_{A_{ig}}) + \epsilon_k(A_{ig})\right),\right. \\
&\qquad \left.\Phi\left(\mu_t + \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ih})\boldsymbol{\eta}_{t,m,j} + \xi(C_{A_{ig}}) + \epsilon_t(A_{ih})\right)\right)
\end{aligned}$$

In order to continue with our derivation, we take 1^{st} order Taylor expansion around zero of the terms within the covariance expression in (B.4). We note that (1) the probit function $\Phi(\cdot)$ is infinitely differentiable at 0 and (2) the second derivative of $\Phi(x)$ at $x = 0$ is 0, however subsequent derivatives need not be. The first order Taylor expansion of $\Phi\left(\mu_t + \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{t,m,j} + \xi(C_{A_{ig}}) + \epsilon_t(A_{ig})\right)$

is:

$$\begin{aligned}\Phi\left(\mu_t + \sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{t,m,j} + \xi(C_{A_{ig}}) + \epsilon_t(A_{ig})\right) &\approx \Phi(0) + \frac{\Phi'(0)}{1}\left(\mu_t + \sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{t,m,j} + \right.\\ &\qquad \left.\xi(C_{A_{ig}}) + \epsilon_t(A_{ig}) - 0\right)\\ &\propto \mu_t + \sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{t,m,j} + \xi(C_{A_{ig}}) + \epsilon_t(A_{ig})\end{aligned}$$

Continuing to derive the covariance between 1-year PUMA and 5-year census tract ENC, for all $A_{ig} \in C_{A_{ig}}$, we get:

$$\begin{aligned}\mathrm{Cov}\left(q_t^{*(5)}(A_{ig}), q_t^{*(1)}(A_i)\right) &\approx \sum_{k=t-4}^{t}\sum_{h=1}^{G_i} N_t(A_{ih}) \times \mathrm{Cov}\left(\mu_k + \sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{k,m,j} + \xi(C_{A_{ig}}) + \epsilon_k(A_{ig}),\right.\\ &\qquad \left.\mu_t + \sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(A_{ih})\boldsymbol{\eta}_{t,m,j} + \xi(C_{A_{ig}})I(A_{ih} \in C_{A_{ig}}) + \epsilon_t(A_{ih})\right)\\ &= \sum_{k=t-4}^{t}\sum_{h=1}^{G_i}\left(N_t(A_{ih}) \times \mathrm{Cov}\left(\sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{k,m,j}, \sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(A_{ih})\boldsymbol{\eta}_{t,m,j}\right)\right) +\\ &\qquad N_{\mathrm{tract}}(A_{ig})\tau_C^2 + \tau^2\\ &= \sum_{r=0}^{4}\sum_{h=1}^{G_i}\left(N_t(A_{ih}) \times \left(\alpha^r \times \sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(A_{ig})\mathbf{K}_{m,j}\mathbf{b}_{m,j}(A_{ih})^T\right)\right) +\\ &\qquad N_{\mathrm{tract}}(A_{ig})\tau_C^2 + \tau^2\\ &\neq 0\end{aligned}$$

We note that the single τ^2 term corresponds to the single instance across the sums when $k = t$ and $h = g$, that is the time and census tract are the same.

Covariance between two census tracts' 5-year ENCs

For the following covariance, we will assume that $q_t^{*(5)}(A_{ig})$ and $q_t^{*(5)}(A_{ih})$ correspond to the 5-year effective number of cases in two different census tracts, that is $g \neq h$. Furthermore, we assume that tracts A_{ig} and A_{ih} are in the same county, denoted $C_{A_{ig}}$. For ease of notation, we assume that the tracts are in the same

PUMA, however this need not be the case. Applying the law of total covariance and a first order Taylor expansion of $\Phi(\cdot)$ around 0, the covariance is:

$$\begin{aligned}
\text{Cov}\left(q_t^{*(5)}(A_{ig}), q_t^{*(5)}(A_{ih})\right) &= E\left(\text{Cov}\left(q_t^{*(5)}(A_{ig}), q_t^{*(5)}(A_{ih})|\mathbf{\Pi}\right)\right) + \\
&\quad \text{Cov}\left(E\left(q_t^{*(5)}(A_{ig})|\mathbf{\Pi}\right), E\left(q_t^{*(5)}(A_{ih})|\mathbf{\Pi}\right)\right) \\
&= \text{Cov}\left(m_t^{*(5)}(A_{ig}) \times \pi_t^{(5)}(A_{ig}), m_t^{*(5)}(A_{ih}) \times \pi_t^{(5)}(A_{ih})\right) \\
&= \text{Cov}\left(m_t^{*(5)}(A_{ig}) \times \frac{1}{5}\sum_{k=t-4}^{t} \pi_k^{(1)}(A_{ig}), m_t^{*(5)}(A_{ih}) \times \frac{1}{5}\sum_{r=t-4}^{t} \pi_r^{(1)}(A_{ih})\right) \\
&= \frac{m_t^{*(5)}(A_{ig}) m_t^{*(5)}(A_{ih})}{25} \sum_{k=t-4}^{t}\sum_{r=t-4}^{t} \text{Cov}\left(\pi_k^{(1)}(A_{ig}), \pi_r^{(1)}(A_{ih})\right) \\
&\approx \sum_{k=t-4}^{t}\sum_{r=t-4}^{t} \text{Cov}\left(\mu_k + \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{k,m,j} + \xi(C_{A_{ig}}) + \epsilon_k(A_{ig}) \right. , \\
&\qquad \left. \mu_t + \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ih})\boldsymbol{\eta}_{t,m,j} + \xi(C_{A_{ig}}) + \epsilon_t(A_{ih})\right) \\
&= \sum_{d=0}^{4}\sum_{a=0}^{4} \text{Cov}\left(\sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{k,m,j} + \xi(C_{A_{ig}}), \right. \\
&\qquad \left. \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ih})\boldsymbol{\eta}_{t,m,j} + \xi(C_{A_{ig}})\right) \\
&= \sum_{d=0}^{4}\sum_{a=0}^{4}\left[\alpha^d \times \alpha^a \times \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ig})\mathbf{K}_{m,j}\mathbf{b}_{m,j}(A_{ih})^T + \tau_C^2\right]
\end{aligned}$$

When A_{ig} and A_{ih} are not in the same county, the covariance is:

$$\text{Cov}\left(q_t^{*(5)}(A_{ig}), q_t^{*(5)}(A_{ih})\right) = \sum_{d=0}^{4}\sum_{a=0}^{4}\left[\alpha^d \times \alpha^a \times \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ig})\mathbf{K}_{m,j}\mathbf{b}_{m,j}(A_{ih})^T\right]$$

APPENDIX C

Additional Material for *Chapter IV: A Spatio-temporal Change of Support Model for Survey-based Estimates of Births in Michigan*

As in Appendix B.2, we will begin by briefly restating our modeling framework. We then assume the following conditional distributions on the effective number of cases:

$$\begin{aligned} y_t^{*(5)}(A_{ig})|\ \lambda_t^{(5)}(A_{ig}) &\overset{ind}{\sim} \text{Poisson}\left(m_t^{*(5)}(A_{ig})\lambda_t^{(5)}(A_{ig})\right) \\ y_t^{*(1)}(A_i)|\ \lambda_t^{(1)}(A_i) &\overset{ind}{\sim} \text{Poisson}\left(m_t^{*(1)}(A_i)\lambda_t^{(1)}(A_i)\right) \end{aligned} \tag{C.1}$$

with $g = 1, \ldots, G_i$ and $i = 1, \ldots, I$. Stating $\lambda_t^{(5)}(A_{ig})$ and $\lambda_t^{(1)}(A_i)$ in terms of our desired spatio-temporal resolution, that is the 1-year census tract resolution or $\lambda_t^{(1)}(A_{ig})$, we get:

$$\begin{aligned} \lambda_t^{(5)}(A_{ig}) &= \frac{1}{5}\sum_{k=t-4}^{t}\lambda_k^{(1)}(A_{ig}) \\ \lambda_t^{(1)}(A_i) &= \frac{1}{N_t(A_i)}\sum_{h=1}^{G_i}N_t(A_{ih})\lambda_t^{(1)}(A_{ih}) \end{aligned}$$

with $g = 1, \ldots, G_i$ and $i = 1, \ldots, I$. The set of all $\lambda_t^{(1)}(A_{ig})$ is denoted as

$\mathbf{\Lambda} := \{\lambda_t^{(1)}(A_{ig})\}_{t=1,\ldots,T, i=1,\ldots,I, g=1,\ldots,G_i}$. Then:

$$
\begin{aligned}
y_t^{*(5)}(A_{ig})|\ \mathbf{\Lambda} &\overset{ind}{\sim} \text{Poisson}\left(m_t^{*(5)}(A_{ig})\frac{1}{5}\sum_{k=t-4}^{t}\lambda_k^{(1)}(A_{ig})\right) \\
y_t^{*(1)}(A_i)|\ \mathbf{\Lambda} &\overset{ind}{\sim} \text{Poisson}\left(m_t^{*(1)}(A_i)\frac{1}{N_t(A_i)}\sum_{h=1}^{G_i}N_t(A_{ih})\lambda_t^{(1)}(A_{ih})\right) \\
\text{(C.2)}\ \log\left(\lambda_t^{(1)}(A_{ig})\right) &= \frac{1}{|A_{ig}|}\int_{\mathbf{s}\in A_{ig}}(\mu_t(\mathbf{s}) + w_t(\mathbf{s}))\,d\mathbf{s}
\end{aligned}
$$

Aggregating the spatio-temporal process in (C.2), we introduce the aggregation error and zero-inflation term $\epsilon_t(A_{ig})$:

$$
\begin{aligned}
\log\left(\lambda_t^{(1)}(A_{ig})\right) &\approx \mu_t + \sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{t,m,j} + \epsilon_t(A_{ig}) \\
\text{(C.3)}\qquad \epsilon_t(A_{ig}) &\sim \gamma_t(A_{ig})N(0,\tau_1^2) + (1-\gamma_t(A_{ig}))N(c,\tau_2^2) \\
\gamma_t(A_{ig}) &\overset{iid}{\sim} \text{Beta}(1,1)
\end{aligned}
$$

Under the ST-MRA, at time $t = 1$ we assume $\boldsymbol{\eta}_{1,m,j} \sim N_r(\mathbf{0}, \mathbf{K}_{m,j})$. Then, for $t = 2, \ldots, T$:

$$
\begin{aligned}
\boldsymbol{\eta}_{t,m,j}|\boldsymbol{\eta}_{t-1,m,j}, \boldsymbol{\eta}_{t-2,m,j}, \ldots, \boldsymbol{\eta}_{1,m,j} &\sim N_r(\alpha\boldsymbol{\eta}_{t-1,m,j}, \mathbf{U}_{m,j}) \\
\mathbf{U}_{m,j} &= (1-\alpha^2)\mathbf{K}_{m,j}
\end{aligned}
$$

It follows from Appendix B.1 and later Appendix B.2 that the covariance of the spatial random effect approximation $w_t(A_{ig}) \approx \sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{t,m,j}$ between two census tracts A_{ig} and A_{ih} at time t and $t+j$ respectively is:

$$
\text{Cov}(w_t(A_{ig}), w_{t+j}(A_{ih})) \approx \alpha^j \sum_{m=0}^{M}\sum_{j=1}^{J^m}\mathbf{b}_{m,j}(A_{ig})\mathbf{K}_{m,j}\mathbf{b}_{m,j}(A_{ih})^T
$$

Covariance between 1-year PUMA and 5-year census tract ENCs

To derive the covariance between the 1-year and 5-year effective number of cases given in (C.1), that is $\text{Cov}\left(y_t^{*(5)}(A_{ig}), y_t^{*(1)}(A_i)\right)$, we proceed as in Appendix B.2, applying first the law of total covariance:

$$\begin{aligned}
\text{Cov}\left(y_t^{*(5)}(A_{ig}), y_t^{*(1)}(A_i)\right) &= E\left(\text{Cov}\left(y_t^{*(5)}(A_{ig}), y_t^{*(1)}(A_i)|\,\boldsymbol{\Lambda}\right)\right) + \\
&\quad \text{Cov}\left(E\left(y_t^{*(5)}(A_{ig})|\,\boldsymbol{\Lambda}\right), E\left(y_t^{*(1)}(A_i)|\,\boldsymbol{\Lambda}\right)\right) \\
&= \text{Cov}\left(m_t^{*(5)}(A_{ig}) \times \lambda_t^{(5)}(A_{ig}), m_t^{*(1)}(A_i) \times \lambda_t^{(1)}(A_i)\right) \\
&= \text{Cov}\left(m_t^{*(5)}(A_{ig}) \times \frac{1}{5}\sum_{k=t-4}^{t} \lambda_k^{(1)}(A_{ig}),\right. \\
&\quad \left. m_t^{*(1)}(A_i) \times \frac{1}{N_t(A_i)}\sum_{h=1}^{G_i} N_t(A_{ih})\lambda_t^{(1)}(A_{ih})\right) \\
&= \frac{m_t^{*(5)}(A_{ig})}{5} \times \frac{m_t^{*(1)}(A_i)}{N_t(A_i)} \times \text{Cov}\left(\sum_{k=t-4}^{t} \lambda_k^{(1)}(A_{ig}), \sum_{h=1}^{G_i} N_t(A_{ih})\lambda_t^{(1)}(A_{ih})\right) \\
&\propto \sum_{k=t-4}^{t}\sum_{h=1}^{G_i} N_t(A_{ih}) \times \text{Cov}\left(\exp\left(\mu_k + \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{k,m,j} + \epsilon_k(A_{ig})\right),\right. \\
&\quad \left. \exp\left(\mu_t + \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ih})\boldsymbol{\eta}_{t,m,j} + \epsilon_t(A_{ih})\right)\right) \qquad \text{(C.4)}
\end{aligned}$$

We take 1^{st} order Taylor expansion around zero of the terms within the covariance expression in (C.4).

$$\begin{aligned}
\exp\left(\mu_t + \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{t,m,j} + \epsilon_t(A_{ig})\right) &\approx \exp(0) + \frac{\exp(0)}{1}\left(\mu_t + \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{t,m,j} + \epsilon_t(A_{ig}) - 0\right) \\
&\propto \mu_t + \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{t,m,j} + \epsilon_t(A_{ig})
\end{aligned}$$

Continuing to derive the covariance between 1-year PUMA and 5-year census tract ENC, we get:

$$\begin{aligned}
\mathrm{Cov}\left(y_t^{*(5)}(A_{ig}), y_t^{*(1)}(A_i)\right) &\approx \sum_{k=t-4}^{t}\sum_{h=1}^{G_i} N_t(A_{ih}) \times \mathrm{Cov}\left(\mu_k + \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{k,m,j} + \epsilon_k(A_{ig})\ ,\right. \\
&\qquad \left. \mu_t + \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ih})\boldsymbol{\eta}_{t,m,j} + \epsilon_t(A_{ih})\right) \\
&= \sum_{k=t-4}^{t}\sum_{h=1}^{G_i}\left(N_t(A_{ih}) \times \mathrm{Cov}\left(\sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{k,m,j},\right.\right. \\
&\qquad \left.\left. \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ih})\boldsymbol{\eta}_{t,m,j}\right)\right) + \mathrm{Var}(\epsilon_t(A_{ig}))
\end{aligned}$$

Due to the mixture prior we have specified on $\epsilon_t(A_{ig})$ in (C.3), its marginal variance is:

$$\begin{aligned}
\mathrm{Var}(\epsilon_t(A_{ig})) &= E(\mathrm{Var}(\epsilon_t(A_{ig})|\gamma_t(A_{ig}))) + \mathrm{Var}(E(\epsilon_t(A_{ig})|\gamma_t(A_{ig}))) \\
&= E\left(\gamma_t(A_{ig})(\tau_1^2 - [c(1-\gamma_t(A_{ig}))]^2) + (1-\gamma_t(A_{ig}))(c^2 + \tau_2^2 - [c(1-\gamma_t(A_{ig}))]^2)\right) + \\
&\quad + \mathrm{Var}\left((1-\gamma_t(A_{ig}))c\right) \\
&= E\left(\gamma_t(A_{ig})\tau_1^2 - \gamma_t(A_{ig})[c(1-\gamma_t(A_{ig}))]^2 + \right. \\
&\quad \left. (1-\gamma_t(A_{ig}))c^2 + (1-\gamma_t(A_{ig}))\tau_2^2 - [c(1-\gamma_t(A_{ig}))]^2 + \gamma_t(A_{ig})[c(1-\gamma_t(A_{ig}))]^2\right) + \\
&\quad + c^2\mathrm{Var}\left((1-\gamma_t(A_{ig}))\right) \\
&= E\left(\gamma_t(A_{ig})\tau_1^2 + (1-\gamma_t(A_{ig}))\tau_2^2 + (1-\gamma_t(A_{ig}))c^2 - [c(1-\gamma_t(A_{ig}))]^2\right) + \\
&\quad + c^2\mathrm{Var}\left(\gamma_t(A_{ig})\right) \\
&= E\left(\gamma_t(A_{ig})\tau_1^2 + (1-\gamma_t(A_{ig}))\tau_2^2 + \gamma_t(A_{ig})(1-\gamma_t(A_{ig}))c^2\right) + \frac{c^2}{12} \\
&= \frac{\tau_1^2}{2} + \frac{\tau_2^2}{2} + \frac{c^2}{2} - E(\gamma_t(A_{ig})^2)c^2 + \frac{c^2}{12} \\
&= \frac{\tau_1^2}{2} + \frac{\tau_2^2}{2} + \frac{c^2}{2} - \frac{c^2}{3} + \frac{c^2}{12} \\
&= \frac{\tau_1^2}{2} + \frac{\tau_2^2}{2} + \frac{c^2}{4}
\end{aligned}$$

Returning to the covariance of the expected number of cases, our final result is:

$$\begin{aligned}
&= \sum_{r=0}^{4}\sum_{h=1}^{G_i}\left(N_t(A_{ih}) \times \left(\alpha^r \times \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ig})\mathbf{K}_{m,j}\mathbf{b}_{m,j}(A_{ih})^T\right)\right) + \frac{\tau_1^2}{2} + \frac{\tau_2^2}{2} + \frac{c^2}{4} \\
&\neq 0
\end{aligned}$$

Covariance between two census tracts' 5-year ENCs

Assume that $y_t^{*(5)}(A_{ig})$ and $y_t^{*(5)}(A_{ih})$ correspond to the 5-year effective number of cases in two different census tracts, that is $g \neq h$. For ease of notation, we assume that the tracts are in the same PUMA, however this need not be the case. Applying the law of total covariance and a first order Taylor expansion of $\exp(\cdot)$ around 0, the covariance is:

$$\begin{aligned}
\text{Cov}\left(y_t^{*(5)}(A_{ig}), y_t^{*(5)}(A_{ih})\right) &= E\left(\text{Cov}\left(y_t^{*(5)}(A_{ig}), y_t^{*(5)}(A_{ih})|\boldsymbol{\Lambda}\right)\right) + \\
&\quad \text{Cov}\left(E\left(y_t^{*(5)}(A_{ig})|\boldsymbol{\Lambda}\right), E\left(y_t^{*(5)}(A_{ih})|\boldsymbol{\Lambda}\right)\right) \\
&= \text{Cov}\left(m_t^{*(5)}(A_{ig}) \times \lambda_t^{(5)}(A_{ig}), m_t^{*(5)}(A_{ih}) \times \lambda_t^{(5)}(A_{ih})\right) \\
&= \text{Cov}\left(m_t^{*(5)}(A_{ig}) \times \frac{1}{5}\sum_{k=t-4}^{t} \lambda_k^{(1)}(A_{ig}), m_t^{*(5)}(A_{ih}) \times \frac{1}{5}\sum_{r=t-4}^{t} \lambda_r^{(1)}(A_{ih})\right) \\
&= \frac{m_t^{*(5)}(A_{ig}) m_t^{*(5)}(A_{ih})}{25} \sum_{k=t-4}^{t}\sum_{r=t-4}^{t} \text{Cov}\left(\lambda_k^{(1)}(A_{ig}), \lambda_r^{(1)}(A_{ih})\right) \\
&\approx \sum_{k=t-4}^{t}\sum_{r=t-4}^{t} \text{Cov}\left(\mu_k + \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{k,m,j} + \epsilon_k(A_{ig}) \,, \right. \\
&\qquad \left. \mu_t + \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ih})\boldsymbol{\eta}_{t,m,j} + \epsilon_t(A_{ih})\right) \\
&= \sum_{d=0}^{4}\sum_{a=0}^{4} \text{Cov}\left(\sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ig})\boldsymbol{\eta}_{k,m,j}, \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ih})\boldsymbol{\eta}_{t,m,j}\right) \\
&= \sum_{d=0}^{4}\sum_{a=0}^{4} \left[\alpha^d \times \alpha^a \times \sum_{m=0}^{M}\sum_{j=1}^{J^m} \mathbf{b}_{m,j}(A_{ig})\mathbf{K}_{m,j}\mathbf{b}_{m,j}(A_{ih})^T\right]
\end{aligned}$$

APPENDIX D

Additional Material for *Chapter V: Correcting Sampling Bias in Electronic Health Records Using an Inhomogeneous Poisson Process and Publicly Available Aggregate Data*

Below, we itemize diagnosis descriptions and codes of our definition of lung cancer. We believe that this is a fairly comprehensive definition, although future work may further refine or expand it.

- Malignant neoplasm of right main bronchus [C34.01]
- Malignant neoplasm of left main bronchus [C34.02]
- Malignant neoplasm of unspecified main bronchus [C34.00]
- Malignant neoplasm of unspecified part of unspecified bronchus lung [C34.90]
- Malignant neoplasm of unspecified part of right bronchus lung [C34.91]
- Malignant neoplasm of unspecified part of left bronchus lung [C34.92]
- Malignant neoplasm of trachea [C33]
- Malignant neoplasm of bronchus and lung [C34]
- Malignant neoplasm of heart, mediastinum and pleura [C38]

- Malignant neoplasm of thymus [C37]
- Malignant neoplasm of accessy sinuses [C31]
- Malignant neoplasm of larynx [C32]
- Malignant neoplasm of other and ill-defined sites in the respiratory system and intrathacic gans [C39]
- Malignant neoplasm of upper lobe, left bronchus lung [C34.12]
- Malignant neoplasm of upper lobe, right bronchus lung [C34.11]
- Malignant neoplasm of upper lobe, unspecified bronchus lung [C34.10]
- Malignant neoplasm of middle lobe, bronchus lung [C34.2]
- Malignant neoplasm of lower lobe, bronchus lung [C34.3]
- Malignant neoplasm of overlapping sites of bronchus and lung [C34.8]
- Malignant neoplasm of unspecified part of bronchus lung [C34.9]
- Malignant neoplasm of main bronchus [C34.0]
- Malignant neoplasm of upper lobe, bronchus lung [C34.1]
- Malignant neoplasm of lower lobe, right bronchus lung [C34.31]
- Malignant neoplasm of lower lobe, left bronchus lung [C34.32]
- Malignant neoplasm of lower lobe, unspecified bronchus lung [C34.30]
- Malignant neoplasm of overlapping sites of left bronchus and lung [C34.82]
- Malignant neoplasm of overlapping sites of unspecified bronchus and lung [C34.80]
- Malignant neoplasm of overlapping sites of right bronchus and lung [C34.81]

- Malignant neoplasm of bronchus and lung, unspecified [162.9]
- Malignant neoplasm of main bronchus [162.2]
- Malignant neoplasm of other parts of bronchus lung [162.8]
- Benign carcinoid tum of the bronchus and lung [209.61]
- Malignant neoplasm of trachea [162.0]
- Malignant neoplasm of thymus [164.0]
- Malignant neoplasm of upper lobe, bronchus lung [162.3]
- Malignant neoplasm of middle lobe, bronchus lung [162.4]
- Malignant neoplasm of lower lobe, bronchus lung [162.5]

BIBLIOGRAPHY

L. Aguilar. Detroit's Cass Corridor makes way for new era, April 2015.

C. L. Avery, K. L. Monda, and K. E. North. Genetic association studies and the effect of misclassification and selection bias in putative confounders. *BMC Proceedings*, 3:S48, 2009.

S. Bandyopadhyay and S. Subba Rao. A test for stationarity for irregularly spaced spatial data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79:95–123, 2017.

S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical modeling and analysis for spatial data*. Chapman & Hall/CRC, 2004. Boca Raton, FL.

S. Banerjee, A. E. Gelfand, A. O. Finley, and H. Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70:825–848, 2008.

L. J. Beesley, L. G. Fritsche, and B. Mukherjee. A modeling framework for exploring sampling and observation process biases in genome and phenome-wide association studies using electronic health records. *bioRxiv*, 2018.

J. Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2:47–53, 1946.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 36:192–236, 1974.

J. Besag, J. York, and A. Mollié. A Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43:1–20, 1991.

J. K. Bower, C. E. Bollinger, R. E. Foraker, Darryl B. Hood, A. B. Shoben, and A. M. Lai. Active Use of Electronic Health Records (EHRs) and Personal Health Records (PHRs) for Epidemiologic Research: Sample Representativeness and Nonresponse Bias in a Study of Women During Pregnancy. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 5, 2017.

J. R. Bradley, C. K. Wikle, and S. H. Holan. Spatio-temporal change of support with application to American Community Survey multi-year period estimates. *Stat*, 4:255–270, 2015.

J. R. Bradley, S. H. Holan, and C. K. Wikle. Multivariate spatio-temporal survey fusion with application to the American Community Survey and local area unemployment statistics. *Stat*, 5:224–233, 2016a.

J. R. Bradley, C. K. Wikle, and S. H. Holan. Bayesian spatial change of support

for count-valued survey data with application to the American Community Survey. *Journal of the American Statistical Association*, 111:472–487, 2016b.

D. Card, A. Mas, and J. Rothstein. Tipping and the dynamics of segregation. *The Quarterly Journal of Economics*, 123:177–218, 2008.

C. Chen, J. Wakefield, and T. Lumely. The use of sampling weights in Bayesian hierarchical models for small area estimation. *Spatial and Spatio-temporal Epidemiology*, 11:33 – 43, 2014.

H. Chipman, E. George, and R. McCulloch. Bayesian CART model search. *Journal of the American Statistical Association*, 93:935–960, 1998.

N. Cressie. *Statistics for Spatial Data*. Wiley, 1993. 2nd edition.

N. Cressie and G. Johannesson. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70:209–226, 2008.

N. Cressie and C.K. Wikle. *Statistics for Spatio-Temporal Data*. Wiley, 2011.

M. J. Daniels and H. W. Hogan. *Missing Data in Longitudinal Studies: strategies for Bayesian modeling and sensitivity analysis.* Chapman & Hall/CRC, 2008. New York.

A. Datta, S. Banerjee, A. O. Finley, and A. E. Gelfand. Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111:800–812, 2016.

E. D. de Leeuw, J. J. Hox, and D. A. Dillman. *International Handbook of Survey Methodology*. Taylor and Francis Group, 2008. New York.

M. Denis and N. Molinari. Free knot splines with rjmcmc in survival data analysis. *Communications in Statistics - Theory and Methods*, 39:2617–2629, 2010.

D. Denison, B. Mallick, and A. Smith. A Bayesian CART algorithm. *Biometrika*, 85:363–377, 1998.

P. J. Diggle, J. A. Tawn, and R. A. Moyeed. Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47:29–9–350, 1998.

P. J. Diggle, R. Menezes, and T. Su. Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59:191–232, 2010.

N. J. Durst. Racial gerrymandering of municipal borders: Direct democracy, participatory democracy, and voting rights in the united states. *Annals of the American Association of Geographers*, 108:938–954, 2018.

D. Eddelbuettel and R. François. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40:1–18, 2011.

I. Gould Ellen and K. M. O'Regan. How low income neighborhoods change: Entry, exit, and enhancement. *Regional Science and Urban Economics*, 41:89 – 97, 2011.

M. R. Elliott. Model averageing methods for weight trimming in generalized linear regression models. *Journal of Official Statistics*, 25:1 – 20, 2009.

A. Finley and S. Banerjee. The SAGE Handbook of Multilevel Modeling. pages 559–580. SAGE Publications Ltd, London, 2013.

F. J. Fowler. *Survey Research Methods*. SAGE Publications, 2014. Los Angeles.

M. Fuentes. A new high frequency kriging approach for non-stationary environmental processes. *Environmetrics*, 12:469–483, 2001.

M. Fuentes. A formal test for nonstationarity of spatial stochastic processes. *Journal of Multivariate Analysis*, 96:30–54, 2005.

M. Fuentes and R. L. Smith. A new class of nonstationary spatial models. Technical report, Department of Statistics, North Carolina State University, Raleigh, NC, 2001.

G. Fuglstad, D. Simpson, F. Lindgren, and H. Rue. Does non-stationary spatial data always require non-stationary random fields? *Spatial Statistics*, 14: 505–531, 2015.

A. E. Gelfand and S. K. Ghosh. Model choice: A minimum posterior predictive loss approach. *Biometrika*, 85:1–11, 1998.

A. E. Gelfand, L. Zhu, and B. P. Carlin. On the change of support problem for spatio-temporal data. *Biostatistics*, 2:31–45, 2001.

A. E. Gelfand, H. Kim, C.F. Sirmans, and S. Banerjee. Spatial modeling with

spatially varying coefficient processes. *Journal of the American Statistical Association*, 98:387–396, 2003.

A. E. Gelfand, S. Banerjee, and D. Gamerman. Spatial process modelling for univariate and multivariate dynamic spatial data. *Environmetrics*, 16:465–479, 2005.

A.E. Gelfand, P. Diggle, P. Guttorp, and M. Fuentes. *Handbook of Spatial Statistics*. CRC Press, 2010. Boca Raton, FL.

A. Gelman. Struggles with survey weighting and regression modeling. *Statistical Science*, 22:153–164, 2007.

E. I. George and D. P. Foster. Calibration and empirical Bayesian variable selection. *Biometrika*, 87:731–747, 2000.

E. I. George and R. E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88:881–889, 1993.

E. I. George and R. E. McCulloch. Approaches for Bayesian variable selection. *Statistica Sinica*, 7:339–374, 1997.

J. Geweke. Evaluating the accuracy of sampling-based approaches to calculate posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 169–193. Clarendon Press, 1992.

E. Goidts and B. Van Wesemael. Regional assessment of soil organic carbon

changes under agriculture in Southern Belgium, 1955-2005. *Geoderma*, 141: 341–354, 2007.

B.A. Goldstein, N. A. Bhavsar, M. Phelan, and M. J. Pencina. Controlling for informed presence bias due to the number of health encounters in an electronic health record. *American Journal of Epidemiology*, 184:847–855, 2016.

P. Goovaerts. Monitoring the aftermath of Flint drinking water contamination crisis: Another case of sampling bias? *Science of The Total Environment*, 590-591:139 – 153, 2017a.

P. Goovaerts. The drinking water contamination crisis in Flint: Modeling temporal trends of lead level since returning to detroit water system. *Science of The Total Environment*, 581-582:66 – 79, 2017b.

C. A Gotway and L. J Young. Combining incompatible spatial data. *Journal of the American Statistical Association*, 97:632–648, 2002.

R. B. Gramacy and H. K. H. Lee. Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103:1119–1130, 2008.

R. M. Groves, F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. *Survey Methodology*. Jon Wiley & Sons Inc., 2009. Hoboken, NJ.

Y. Guan, M. Sherman, and J. A. Calvin. A nonparametric test for spatial isotropy

using subsampling. *Journal of the American Statistical Association*, 99: 810–821, 2004.

J. D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

C. P. Heidkamp and S. Lucas. Finding the gentrification frontier using census data: The case of portland, maine. *Urban Geography*, 27:101–125, 2006.

H. Ishwaran and J. S. Rao. Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33:730–773, 2005.

G. Johannesson, N. Cressie, and H.-C. Huang. Dynamic multi-resolution spatial models. *Environmental and Ecological Statistics*, 14:5–25, 2007.

V. E. Johnson and D. Rossell. Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107:649–660, 2012.

M. Jun and M. G. Genton. A test for stationarity of spatio-temporal random fields on planar and spherical domains. *Statistica Sinica*, 22:1737–1764, 2012.

E. Kang and N. Cressie. Bayesian inference for the spatial random effects model. *Journal of the American Statistical Association (Theory and Methods)*, 106: 975–983, 2011.

M. Katzfuss. A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, 112:201–214, 2017.

M. Katzfuss and J. Guinness. A general framework for Vecchia approximations of Gaussian processes, 2017. arXiv:1708.06302 [stat.ME].

M. Katzfuss and D. Hammerling. Parallel inference for massive distributed spatial data using low-rank models. *Statistics and Computing*, 27:363–375, 2017.

C. Kennedy, E. Yard, T. Dignam, S. Buchannan, S. Condon, M. J. Brown, J. Raymond, H. Schurz Rogers, J. Sarisky, R. Decastro, I. Arias, and P. Breysse. Blood lead levels among children aged <6 years - Flint, Michigan, 2013-2016. *MMWR. Morbidity and Mortality Weekly Report*, 65:650–654, 2016.

H. Kim, B. K. Mallick, and C. C. Holmes. Analyzing nonstationary spatial data using piecewise Gaussian processes. *Journal of the American Statistical Association*, 100:653–668, 2005.

L. Kish. Methods for design effects. *Journal of Official Statistics*, 11:55–77, 1995.

B. A. Konomi, H. Sang, and B. K. Mallick. Adaptive Bayesian nonstationary modeling for large spatial datasets using covariance approximations. *Journal of Computational and Graphical Statistics*, 3:802–829, 2014.

E. L. Korn and B. I. Graubard. Confidence intervals for proportions with small expected number of positive counts estimated from survey data. *Survey Methodology*, 24:193–201, 1998.

Diane Lambert. Zero-inflated Poisson regression, with and application to defects in manufacturing. *Technometrics*, 34:1–14, 1992.

L. Lees, T. Slater, and E. Wyly. *Gentrification*. Taylor & Francis Group, LLC, 2008.

C. Lefèvre, F. Rekik, V. Alcantara, and L. Wiese. Soil Organic Carbon: the hidden potential, 2017. Techinical Report of the Food and Agriculture Organization of the United Nations.

L. G. Leon-Novelo and T. D. Savitsky. Fully bayesian estimation under informative sampling, 2017. arXiv:1710.00019 [stat.ME].

B. Li, M. G. Genton, and M. Sherman. Testing the covariance structure of multivariate random fields. *Biometrika*, 95:813–829, 2008.

J. J. A. Little. Inference with survey weights. *Journal of Official Statistics*, 7:405 – 424, 1991.

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. Hoboken, NJ.

J. Mack. Number of Michigan births hits 76-year low; see numbers for your county, January 2019.

K. C. Martis. The original gerrymander. *Political Geography*, 2:833 – 839, 2008.

T. Matsuo, D. W. Nychka, and D. Paul. Nonstationary covariance modeling for incomplete data: Monte Carlo EM approach. *Computational Statistics and Data Analysis*, 55:2059–2073, 2011.

MDHHS. Crude birth rates, 2017. URL `https://www.mdch.state.mi.us/osr/natality/tab1.1.asp`.

L. Mercer, J. Wakefield, C. Chen, and T. Lumley. A comparison of spatial smooth-

ing methods for small area estimation with sampling weights. *Spatial Statistics*, 8:69–85, 2014.

U. Mishra, R. Lal, B. Slater, F. Calhoun, D. Liu, and M. Van Meirvenne. Predicting soil organic carbon stock using profile depth distribution functions and ordinary kriging. *Soil Science Society of America Journal*, 73:614–621, 2009.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83:1023–1032, 1988.

L. Moehlman and M. Robins-Somerville. The New Detroit: How gentrification has changed Detroit's economic landscape, September 2016.

John Mullahy. Specification and testing of some modified count data models. *Journal of Econometrics*, 33:341–365, 1986.

L. Murembya and E. Guthrie. Demographic and labor market profile: City of Flint. Technical report, State of Michigan: Department of Technology, Management, and Budget, April 2016.

N. N. Narisetty and X. He. Bayesian variable selection with shriking and diffusing priors. *The Annals of Statistics*, 42:789–817, 2014.

National Conference of State Legislatures. Redistricting and the Supreme Court: the most significant cases., 2018.

J. H. V. Neto, A. M. Schmidt, and P. Guttorp. Accounting for spatially varying

directional effects in spatial covariance structures. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63:103–122, 2014.

D. J. Nott and W. T. M. Dunsmuir. Estimation of nonstationary spatial covariance structure. *Biometrika*, 89:819–829, 2002.

D. Nychka, C. Wikle, and J. A. Royle. Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling*, 2:315–331, 2002.

D. W. Nychka, S. Bandyopadhyay, D. Hammerling, F. Lindgren, and S. Sain. A multi-resolution Gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics*, 24:579–599, 2016.

C. J. Paciorek and M. J. Schervish. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17:483–506, 2006.

D. Pati, B. J. Reich, and D. B. Dunson. Bayesian geostatistical modelling with informative sampling locations. *Biometrika*, 98:35–48, 2011.

C. A. Pope, J. P. Gosling, S. Barber, T. Yamaguchi, and P. G. Blackwell. Modelling spatial heterogeneity and discontinuities using Voronoi tessellations, 2018. arXiv:1802.05530 [stat.ME].

A. E. Raftery and S. M. Lewis. Practical markov chain monte carlo: Comment: One long run with diagnostics: Implementation strategies for markov chain monte carlo. *Statistical Science*, 7:493–497, 1992.

S. Rathbun. Discussion on the Paper by Diggle, Menezes and Su. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59:191–232, 2010.

M. Reibel and M. Regelson. Neighborhood racial and ethnic change: The time dimension in segregation. *Urban Geography*, 32:360–382, 05 2013.

A. Riebler, S. H Sørbye, D. Simpson, and H. Rue. An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, 25:1145–1165, 2016.

M. D. Risser and C. A. Calder. Local likelihood estimation for covariance functions with spatially-varying parameters: The convoSPAT package for R. *Journal of Statistical Software*, 81:1–32, 2017.

M. D. Risser, C. A. Calder, V. J. Berrocal, and C. Berrett. Nonstationary spatial prediction of soil organic carbon: Implications for stock assessment decision making. *Annals of Applied Statistics*, 2018. In press.

G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, 7:110–120, 1997.

P. D. Sampson. Constructions for nonstationary spatial processes. In A. E. Gelfand, P. Diggle, M. Fuentes, and P. Guttorp, editors, *Handbook of Spatial Statistics*, pages 119–130. CRC Press, 2010.

T. D. Savitsky. Bayesian nonparametric multiresolution estimation for the american community survey. *Annals of Applied Statistics*, 10:2157–2181, 2016.

T. D. Savitsky and D. Toth. Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 10:1677–1708, 2016.

T. C. Schelling. Dynamic models of segregation. *The Journal of Mathematical Sociology*, 1:143–186, 1971.

A. M. Schmidt, P. Guttorp, and O'Hagan A. Considering covariates in the covariance structure of spatial processes. *Environmetrics*, 22:487–500, 2011.

P. H. Schuck. The thickest thicket: Partisan gerrymandering and judicial regulation of politics. *Columbia Law Review*, 87:1325–1384, 1987.

M. Simpson, S. H. Holan, C. K. Wikle, and J. R. Bradley. Interpolating distributions for populations in nested geographies using public-use data with application to the american community survey, 2019. arXiv:1802.02626.

S. Sleutel, S. De Neve, G. Hofman, P. Boeckx, D. Beheydt, O. Van Cleemput, I. Mestdagh, P. Lootens, L. Carlier, N. Van Camp, H. Verbeeck, I. Vande Walle, R. Samson, N. Lust, and R. Lemeur. Carbon stock changes and carbon sequestration potential of Flemish cropland soils. *Global Change Biology*, 9:1193–1203, 2003.

M. L. Stein. A modeling approach for large spatial datasets. *Journal of the Korean Statistical Society*, 37:3–10, 2008.

Z. Tyson. The 2018 midterm vote: Divisions by race, gender, education, 2018.

U. S. Centers for Disease Control & Prevention. Behavioral Risk Factor Surveillance System, 2019a. https://www.cdc.gov/brfss/index.html.

U. S. Centers for Disease Control & Prevention. National Health Interview Survey, 2019b. https://www.cdc.gov/nchs/nhis/index.htm.

U. S. Centers for Disease Control & Prevention. State Cancer Profiles, 2019c. https://statecancerprofiles.cancer.gov/index.html.

U.S. Census Bureau. A compass for understanding and using American Community Survey data: What general data users need to know, 2008. U.S. Government Printing Office, Washington, DC.

U.S. Census Bureau. American Community Survey Design and Methodology, 2014.

U.S. Centers for Disease Control & Prevention. Electronic medical records/electronic health records (emrs/ehrs), 2019. https://www.cdc.gov/nchs/fastats/electronic-medical-records.htm.

M. van Lieshout. *Markov Point Porcesses and their Applications*. Imperial College Press, London, UK, 2000.

A. V. Vecchia. Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 50:297–312, 1988.

L.A. Waller and C.A. Gotway. *Applied Spatial Statistics for Public Health Data*. Wiley, 2004. Hoboken, NJ.

F. Wei and P. L. Knox. Neighborhood change in metropolitan America, 1990 to 2010. *Urban Affairs Review*, 50:459–489, 2014.

N. G. Weiskopf, A. Rusanov, and C. Weng. Sick patients have more data: The non-random completeness of electronic health records. *AMIA Annual Symposium Proceedings*, pages 1472–1477, 2013.

Z. D. Weller and J. A. Hoeting. A review of nonparametric hypothesis tests of isotropy properties of spatial data. *Statistical Science*, 31:305–324, 2016.

D. Westreich. Berkson's bias, selection bias, and missing data. *Epidemiology*, 23: 159–164, 2012.

E. Winowiecki. 5 things to know about the ball proposal to end gerrymandering in Michigan, 2018.

M. Zuk, A. H. Bierbaum, K. Chapple, K. Gorska, and A. Loukaitou-Sideris. Gentrification, displacement, and the role of public investment. *Journal of Planning Literature*, 33:31–44, 2018.