Japanese Journal of Statistics and Data Science (2022) 5:407–430
https://doi.org/10.1007/s42081-022-00178-8

ORIGINAL PAPER (INVITED)

Bayesian finite-population inference with spatially correlated measurements

Alec Chan-Golston[1] · Sudipto Banerjee[2] · Thomas R. Belin[2] · Sarah E. Roth[3] · Michael L. Prelip[4]

Received: 22 March 2022 / Revised: 28 July 2022 / Accepted: 30 July 2022 /
Published online: 6 September 2022

Abstract

Community-based public health interventions often rely on representative, spatially referenced outcome data to draw conclusions about a finite population. To estimate finite-population parameters, we are posed with two challenges: to correctly account for spatial association among the sampled and nonsampled participants and to correctly model missingness in key covariates, which may be also spatially associated. To accomplish this, we take inspiration from the preferential sampling literature and develop a general Bayesian framework that can specifically account for preferential non-response. This framework is first applied to three missing data scenarios in a simulation study. It is then used to account for missing data patterns seen in reported annual household income in a corner-store intervention project. Through this, we are able to construct finite-population estimates of the percent of income spent on fruits and veg-

✉ Sudipto Banerjee
sudipto@ucla.edu

Alec Chan-Golston
achan-golston@ucmerced.edu

Thomas R. Belin
tbelin@mednet.ucla.edu

Sarah E. Roth
sarah.roth@providence.org

Michael L. Prelip
mprelip@ucla.edu

1 Department of Public Health, University of California, Merced, 5200 North Lake Rd., Merced, CA 95343, USA

2 Department of Biostatistics, UCLA, 650 Charles E. Young Drive South, Los Angeles, CA 90095-1772, USA

3 Center for Outcomes Research and Education, Providence, 5251 NE Glisan Street, Portland, OR 97213, USA

4 Department of Community Health Sciences, UCLA, 650 Charles E. Young Drive South, Los Angeles, CA 90095-1772, USA

etables. Such a framework provides a flexible way to account for spatial association and complex missing data structures in finite populations.

Keywords Spatial process · Missing data · Preferential sampling · Bayesian model · Public health

1 Introduction

Evaluations of prospective public health interventions can be strengthened by collecting outcomes from representative samples of the study population. It would not be surprising for there to be spatial patterns of association among individuals sampled from clusters of households in a neighborhood where an intervention was implemented, nor would it be surprising for there to be missing data on key study covariates. This manuscript outlines a strategy for evaluating community-based interventions that draws on frameworks for inference from finite-population sampling, Bayesian analysis of spatial process models, and methods for handling missing data.

Our primary application is provided by a study of a "corner store conversion" intervention implemented in under-resourced areas of Los Angeles (Ortega et al., 2015). In the realm of food purchasing, disparities by income exist in fruit and vegetable (FV) consumption (Grimm et al., 2012), nutrition (Casey et al., 2001), and overall food insecurity (Ribar & Hamrick, 2003; Rose, 1999). These problems are observed in "food swamps", communities with higher numbers of unhealthy establishment that serve fast-food or sell junk food (Rose et al., 2009) than stores with healthy food options. Corner store interventions are one public health strategy to change the food environment in the hope of improving eating behaviors at the individual and community level (Langellier et al., 2013). To facilitate improved FV sales, such interventions commonly increase the amount of fresh FVs sold in a store (Langellier et al., 2013), and may provide refrigeration units (Paek et al., 2014), store remodeling (Langellier et al., 2013), cooking demonstrations (Ortega et al., 2015), increased signage (Lawman et al., 2015), and business consulting (Ortega et al., 2015). Among these studies, findings regarding availability and sales of fruits, vegetables, and other healthy foods have been mixed (Albert et al., 2017; Lawman et al., 2015; Paek et al., 2014; Song et al., 2009; Thorndike et al., 2017).

The focus of analysis in these interventions have been the patrons of these corner stores while few studies have examined the effect of these interventions at the community level. Notably, in such an intervention in two low-income, predominantly Latino communities in California, East Los Angeles and Boyle Heights, Ortega et al. (2016) reported no significant improvements to FV purchasing or consumption. However, one variable of interest, the percentage of annual reported income spent on fruits and vegetables (PIFV), was not fully investigated in earlier reports due to complexities associated with the high rate of missing data on reported income. As many of the intervention components sought to influence the community context, it is important to assess the extent to which any intervention effect was discernible with attention to the potential for available data to be incomplete.

Non-response of household income is a common occurrence in survey research (Schenker et al., 2006; Watson & Starick, 2011; Yan et al., 2010), but any method for handling missing data must address two key challenges. First, there is evidence that reported income is spatially associated in neighborhoods (Breau et al., 2018; Chakravorty, 1996). One approach to account for this is to employ spatial process modeling (Banerjee et al., 2014; Cressie & Wikle, 2011; Ripley, 2004), embedded within a Bayesian inference framework, where inferences flow from averaging over (i.e., carrying out iterative-simulation-based numerical integration applicable to) joint distributions of observable values and unobserved parameters that encode conditional-independence assumptions in a generic framework such as

$$[\text{data, process, parameters}] = [\text{data} \mid \text{process}] \times [\text{process} \mid \text{parameters}] \times [\text{parameters}] \,. \quad (1)$$

Here, the data are assumed to be a partial realization of a Gaussian stochastic process, where the covariance between elements are defined by $C(d_{ab})$, a function of the distance, d_{ab} between any two locations $\boldsymbol{\ell}_a$ and $\boldsymbol{\ell}_b$. While there are many valid ways to represent such structure, a flexible choice is the Matérn family (Rasmussen & Williams, 2006) of functions, defined as $C(d_{ab}) = \sigma^2 + \delta^2$ if $d_{ab} = 0$ and $C(d_{ab}) = \frac{\delta^2}{2^{\nu-1}\Gamma(\nu)}\left(\sqrt{2\nu}d_{ab}\phi\right)^{\nu} K_{\nu}\left(\sqrt{2\nu}d_{ab}\phi\right)$ if $d_{ab} > 0$, where $K_{\nu}(\cdot)$ is the modified Bessel function. Here ν is a smoothness parameter, σ^2 describes the variation due to measurement error, δ^2 measures the spatial variance, ϕ is a decay parameter which determines the rate of decline in spatial association. The exponential function, $C(d_{ab}) = \delta^2 \exp(-\phi d_{ab})$ if $d_{ab} > 0$, is a special case when $\nu = 1/2$. Unlike the literature of small area estimation (Clayton & Kaldor, 1987; Ghosh & Rao, 1994; Rao, 2003), where the sampling units are regions such as counties, states or census-tracts, spatial process models consider quantities that, at least conceptually, exist in continuum over the entire domain.

A recent application of Bayesian spatial techniques to high dimensional survey is given by Bradley et al. (2015), who employed a multivariate spatio-temporal mixed effects model to examine differences in monthly income by gender, finding that men have larger average incomes in multiple industries, with the largest differences in the finance and insurance fields. Such models have effect dimension reduction by applying Moran's I basis functions to the spatio-temporal setting and outperform univariate spatial models in mean-square prediction error. Bradley et al. (2016) furthered this model by developing a hierarchical Bayesian approach to survey fusion, assuming a latent process shared by each survey dataset. Using data from the American Community Survey and Local Area Unemployment Statistics, they demonstrated higher precision in estimates of unemployment compared to analyzing each dataset separately. Bradley et al. (2016) also developed a Bayesian technique to account for a spatial change of support in count data and incorporate known survey variances.

A second challenge is that non-response to income questions might depend on underlying income values and associated demographic characteristics. Greenlees et al. (1982), David et al. (1986) and Riphahn and Serfling (2005) all noted evidence from population-based surveys that individuals with higher incomes were less likely to respond, although in surveys of lower income communities, it is plausible that the

direction of the association between income and non-response would be reversed. As we suspect our outcome is spatially associated, however, we turn to the recent literature regarding preferential sampling to better understand this problem.

First described in Diggle et al. (2010), preferential sampling is a technique in which the probability of selection on a spatial domain increases as a function of intensity of the measurement. Diggle and colleagues present a joint model in which the selection sites and the measured values arise from the same spatial process. Pati et al. (2011) presents a model for preferential sampling in a fully Bayesian framework by including a function of intensity as a predictor of the outcome to account for informative sampling. In addition, preferential sampling has been shown to give biased predictions (Gelfand et al., 2012; Lee et al., 2015) and parameter estimation (Antonelli et al., 2016). In our corner-store scenario, we consider "preferential" response, in which the probability of a spatially associated variable being reported is dependent on the underlying value of that variable.

Finally, as we are interested in estimating average percentage of income spent on fruits and vegetables for all individuals in a community, we examine the problem from a finite-population perspective, considering those who reported income to be the sampled or observed cases. Finite-population survey sampling (Cochran, 1977; Hartley & Sielken, 1975; Horvitz & Thompson, 1952; Royall, 1970) considers sampling designs in the statistical modeling and inference on finite populations. Bayesian models (Ericson, 1969; Gelman, 2007; Ghosh & Meeden, 1997) can incorporate aspects of study design and often perform better with small datasets while yielding similar results to design-based results in large datasets (Little, 2004). Estimation of finite-population quantities within spatial process settings has not received much attention in the literature. Recent work includes a method for block kriging which connects geostatistical models and classical design-based sampling (Hoef, 2002), a spline-based estimator of the mean for samples drawn from a spatially-correlated population (Cicchitelli & Montanari, 2012), and the use of linear spatial interpolator to create a design-based predictor of values at unobserved locations (Bruno et al., 2013). Chan-Golston et al. (2020) demonstrated that accounting for both design and spatial association in a two-stage sampling context led to better model fit and better coverage of the finite-population parameters.

The rest of the paper is as follows: Sect. 2 elaborates on data collected during the corner-store intervention described in Ortega et al. (2016) and provides an in depth explanation of the income non-response by community, Sect. 3 presents a Bayesian framework that accommodates preferential non-response, and Sect. 4 examines a simulation study of the proposed framework. Section 5 presents a data analysis to assess the extent of any intervention effect on the percentage of income spent on fruits and vegetables utilizing model-based finite-population estimates of the outcome of interest both pre-intervention and post-intervention. The paper concludes with a discussion in Sect. 6.

2 Motivating application

Supported by NIH center-grant funding focused on reducing population-based health disparities, and with input from a consultant who had experience with previous corner-store conversions in Northern California, researchers identified 4 pairs of corner stores in the East Los Angeles and Boyle Heights communities of Los Angeles to compare the active intervention with a control intervention. The active intervention of "corner-store conversion" included a reorganization of store items to promote healthy food purchasing, an external transformation of the store, a social marketing campaign and cooking demonstrations put on by local youth, connections to local wholesale markets, and refrigeration units (Ortega et al., 2016). Both the active and control interventions provided training to improve bookkeeping and accounting. A more detailed review of the study design and implementation is provided by Ortega et al. (2015). To assess the potential effects of this intervention, a survey was given to residents within a specified radius of each of the eight corner stores. This community survey sought to extensively catalog the food purchasing of residents, including where they shopped, what types of food they bought, and who was being supported by their food purchases. As such, the survey was directed to adults who were identified as the main food purchaser of the family. Many other items were also collected, including demographic characteristics, health problems, family history of residency, and government food program participation (such as the Supplemental Nutrition Assistance Program and the Special Supplemental Program for Women, Infants, and Children). This survey was conducted in each of the eight communities surrounding the store (generally a 2–3 block radius) before the conversion and then again roughly one year after the conversion. There were 1035 observations collected at baseline and 1052 observations collected at follow-up, with approximately 60% of the individuals surveyed at baseline surveyed again at follow-up.

While there is a strong interest in describing PIFV in each community, the sample had high levels of missingness in income (one-third) at both baseline and follow-up, which are presented in Table 1. Noticeably, PIFV is highest on average at baseline in Communities 1 and 7, 26.0% and 46.5%, respectively, which also observed lower levels of response and income compared to the averages of the total. With high levels of non-response, it is important to know if this value is being inflated due to the missing values of income. In addition, while the number of sampled units ranged from 114 to 143, the percentage of missingness ranged widely from 4.9 to 66.6%. For this paper, we consider the sampled data to be the finite population of eight communities in East Los Angeles and Boyle Heights. This is a reasonable assumption, as the response rate of 80% and 71% at baseline and follow-up suggest that a majority of individuals in these communities are represented in this dataset. Amount spent on fruits and vegetables was reported on weekly, bi-weekly, or monthly scale. These values were multiplied by 52, 26, and 12, respectively, to reflect the annual amount spent on fruits and vegetables in a household. Reported yearly income is continuous and ranged from \$0 to \$300,000. Twenty-four individuals reported a higher amount spent on fruits and vegetables than their income, so their income was imputed to the value spent on FV, so that PIFV was no more than 100 and no annual income was equal to 0. Both annual income and annual amount spent on FV were log-transformed to produce a more normal distribution of

Table 1 Annual income and FV expenditures by site and time-point

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

Income and FV expenditure are presented in units of $1000
B baseline, *F* follow-up, *M* mean and *SD* standard deviation

the outcome. The analysis was restricted to cases with no missing covariates and with a recorded amount spent on FV. This resulted in a final dataset with 982 observations at baseline and 1033 at follow-up.

Other individual-level variables that were hypothesized to affect PIFV were age at time of interview, gender, household size, marital status (collapsed into a binary classification distinguishing other possibilities from being in a marriage or marriage-like relationship), and education level (collapsed into a binary classification of at least a high-school education or less than a high-school education). Due to the homogeneity of ethnicity in the sample, Latino ethnicity was not considered in the analyses. Summary statistics of these covariates by time-point are presented in Table 2.

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

Table 2 Description of the sample data by time-point

	Baseline $N = 982$	Follow-up $N = 1033$
Demographic	% or M (SD)	% or M (SD)
Intervention	50.7	50.5
Control	49.3	49.5
Gender		
Male	21.6	20.0
Female	78.4	80.0
Marital status		
Married/common law marriage	56.9	58.6
Single/separated/divorced/widowed[a]	43.1	41.4
Education level		
$\geq$ High school	50.5	49.9
$<$ High school[b]	49.5	50.1
Age	45.61 (16.0)	46.9 (15.5)
Household size	4.0 (1.9)	4.0 (2.0)

M mean and *SD* standard deviation
[a] This category also contains 6 responses of "Don't Know" and 1 refused response
[b] This category also contains 9 responses of "Don't Know" and 3 refused responses

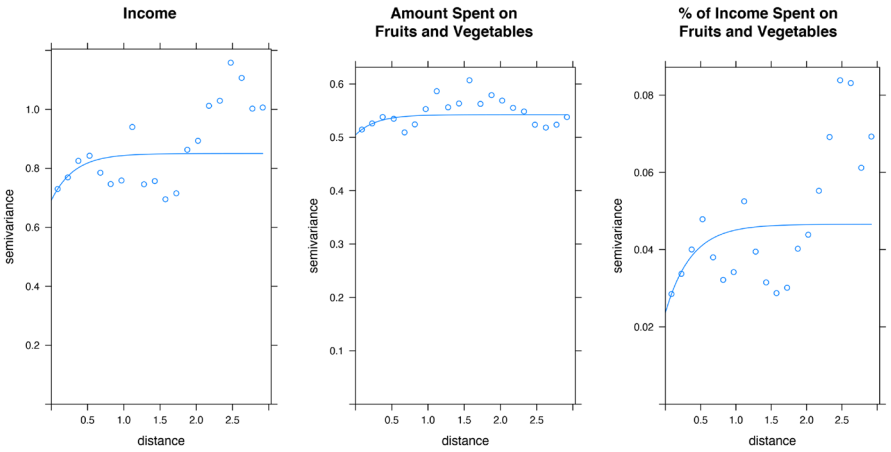

Fig. 1 Variograms of income (log-scale), amount spent on FV, and PIFV

Variograms of the PIFV outcome, amount spent on FV, and log-income were constructed. All variograms suggested evidence of spatial association, as shown in Fig. 2. To explore our primary outcome and determine if there is any evidence of preferential response in income, a linear model was first fit using the previously described covariates, as well as a indicators for time-point, intervention status, and the interaction of these two indicators to detect an interaction effect, predicting the log-percent of income spent on fruits and vegetables. For individuals who did not report income, predictions of this log-percent were made using the results of the linear model. By

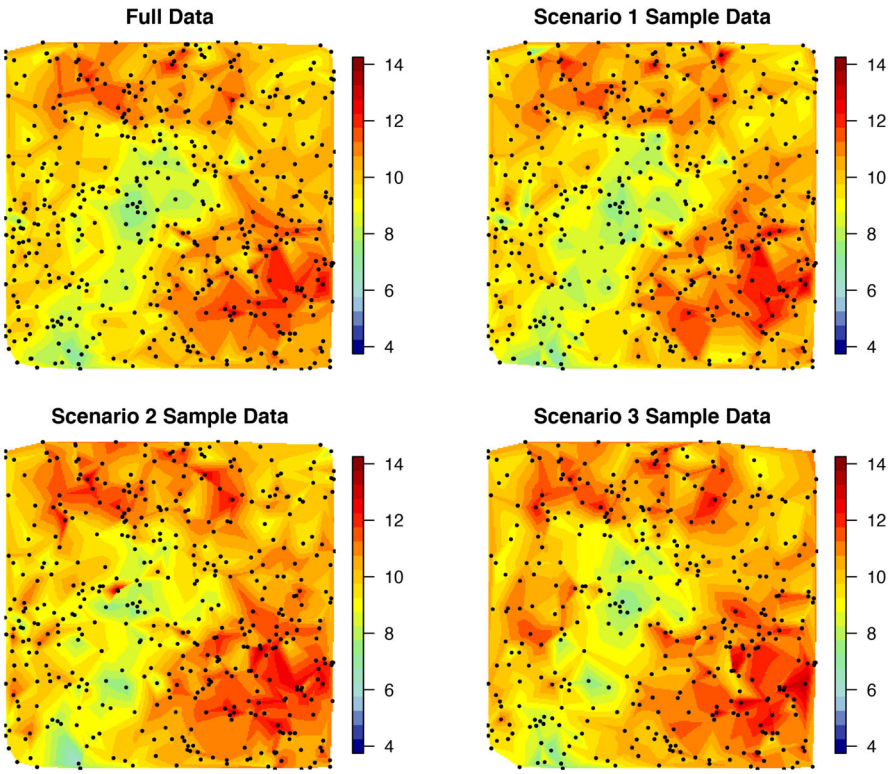

Fig. 2 Linear interpolation plots from full simulated data and 3 scenarios

dividing the reported amount spent on fruits and vegetables by this percent, we have constructed a prediction of income for the non-respondents. Then, a logistic regression model predicting the response of income was fit with an intercept term and income (either the reported value for those that responded or the predicted value from the linear model for those that did not respond). This model found income to be significantly associated with the probability of response. An observed coefficient estimate of 0.12 (SE = 0.05) suggests that individuals with higher values of income are more likely to report income, and, conversely, lower income in these communities are more likely to be under-reported. Further, a logistic regression model with random intercepts for location was fit and the standard deviation corresponding to the random intercept was 0.67.

3 Representing spatial structure in finite-population inference

3.1 A general framework

Formally, define a spatial domain $\mathscr{L} \subseteq \mathfrak{R}^2$, where a finite population of size T is located in N locations, $\mathscr{L}_{FP} = \{\ell_1, \ldots, \ell_N\}$, $T \geq N$. Suppose there are M_i units

at the i^{th} location, hence $T = \sum_{i=1}^{N} M_i$. Further, suppose that t, $t \leq T$, units are sampled from the finite population and thus n, $n \leq N$, locations are represented in this sample. Taking the first n locations to be sampled, define the sampled and nonsampled location sets as $\mathscr{L}_s = \{\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_n\}$ and $\mathscr{L}_{ns} = \{\boldsymbol{\ell}_{n+1}, \ldots, \boldsymbol{\ell}_N\}$, respectively. In addition, denoting m_i the number of sampled units at the i^{th} location, $i = 0, \ldots, N$, we have that $t = \sum_{i=1}^{N} m_i = \sum_{i=1}^{n} m_i$, as $m_i = 0$ for $i = n+1, \ldots, N$. In the context of our data, we have that $T = 2015$, $t = 1294$, $N = 635$, and $n = 555$. We are interested in measuring annual reported income on the natural log-scale, $\mathbf{y}$, which is a vector of sampled and nonsampled measurements, e.g., $\mathbf{y} = [\mathbf{y}_s^\top, \mathbf{y}_{ns}^\top]^\top$. Denoting $y_j(\boldsymbol{\ell}_i)$ as the annual income on the natural log scale of the j^{th} individual at the i^{th} location, let $\mathbf{y}_s = [y_1(\boldsymbol{\ell}_1), \ldots, y_{m_1}(\boldsymbol{\ell}_1), \ldots, y_1(\boldsymbol{\ell}_n), \ldots, y_{m_n}(\boldsymbol{\ell}_n)]^\top$ and $\mathbf{y}_{ns} = [y_{m_1+1}(\boldsymbol{\ell}_1), \ldots, y_{M_1}(\boldsymbol{\ell}_1), \ldots, y_{m_N+1}(\boldsymbol{\ell}_N), \ldots, y_{M_N}(\boldsymbol{\ell}_N)]^\top$. In addition, let $z_j(\boldsymbol{\ell}_i)$ be the reported amount spent on fruits and vegetables on the natural log-scale corresponding to $y_j(\boldsymbol{\ell}_i)$. This is measured for all members of the finite population and, therefore, vectors $\boldsymbol{z}_s$ and $\boldsymbol{z}_{ns}$, defined in the same manner as $\boldsymbol{y}_s$ and $\boldsymbol{y}_{ns}$, denote reported values of FV expenditures corresponding to individuals who reported and did not report income, respectively. We examine the log-percent of income spent on fruits and vegetables, which can be written as $\mathbf{z} - \mathbf{y}$, by modeling $\mathbf{y}$ with an offset term of $\mathbf{z}$. Assume that there is a Gaussian spatial process, $\omega(\cdot)$, defined on $\mathscr{L}$ with covariance function $K_\omega(d)$, and that $\mathbf{y}$ is a partial realization of this process. Finally, define the inclusion mechanism as a spatial process on $\mathscr{L}$, which is dependent on $\mathbf{y}$ and another Gaussian spatial process, $\upsilon(\cdot)$, defined on the same domain with covariance function $K_\upsilon(d)$. A joint model defined in the form of our generic spatial paradigm (1) is

$$[y(\cdot) \mid \omega(\cdot)] \times [I(\cdot) \mid y(\cdot), \upsilon(\cdot)] \times [\omega(\cdot)] \times [\upsilon(\cdot)] \tag{2}$$

The first component of (2) is the conditional distribution of $\mathbf{y}$, $[y(\cdot) \mid \omega(\cdot)]$. Assuming $\mathbf{y}$ is a $T \times 1$ vector, this conditional distribution can be written as

$$y_j(\boldsymbol{\ell}_i) = z_j(\boldsymbol{\ell}_i) - \mathbf{x}_j(\boldsymbol{\ell}_i)^\top \boldsymbol{\beta} + \omega(\boldsymbol{\ell}_i) + \epsilon_j(\boldsymbol{\ell}_i)\,; \quad \epsilon_j(\boldsymbol{\ell}_i) \overset{iid}{\sim} N(0, \sigma^2)\,. \tag{3}$$

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

Spatial variation is accounted for with the 635×1 vector $\boldsymbol{\omega} \sim N(\mathbf{0}, \mathbf{6}_{\omega})$, where $\mathbf{6}_{\omega}$ is a 635×635 matrix defined by the covariance function $K_{\omega}(d)$. Finally, construct a 2015×635 site indicator matrix $\mathbf{A} = [\mathbf{A}_s^{\top}, \mathbf{A}_{ns}^{\top}]^{\top}$, where $\mathbf{A}_s = [\oplus_{i=1}^{555} \mathbf{1}_{m_i} : \mathbf{0}]$ and $\mathbf{A}_{ns} = [\oplus_{i=1}^{635} \mathbf{1}_{M_i - m_i}]$ ($\oplus$ denotes the Kronecker sum). Thus the row in $\mathbf{A}$ corresponding to measurement $y_j(\boldsymbol{\ell}_i)$ has value 1 in i^{th} column and 0 elsewhere. We then have that $\mathbf{y} \sim N(\mathbf{z} - \mathbf{X}\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\omega}, \boldsymbol{\Sigma}_{\epsilon})$.

The second component of (2), $[I(\cdot) \mid y(\cdot), \upsilon(\cdot)]$ describes the response mechanism. Here the $T \times 1$ vector $\mathbf{I}$ has element $I_j(\boldsymbol{\ell}_i) = 1$ if the corresponding jth individual in the ith location reported their income, e.g., $y_j(\boldsymbol{\ell}_i)$ is observed, and $I_j(\boldsymbol{\ell}_i) = 0$ if they did not report their income. This can be expressed as

$$I_j(\boldsymbol{\ell}_i) \sim \text{Ber}(\pi_j(\boldsymbol{\ell}_i)); \quad \text{logit}(\pi_j(\boldsymbol{\ell}_i)) = y_j(\boldsymbol{\ell}_i)\eta_y + \mathbf{q}_j(\boldsymbol{\ell}_i)^{\top}\boldsymbol{\eta} + \upsilon(\boldsymbol{\ell}_i). \tag{4}$$

The probability of response for each individual in the finite population is permitted to vary by its corresponding value of y, which is captured in the regression coefficient η_y, $\eta_y \sim N(0, \sigma^2_{\eta_y})$. Similar to our modeling of the outcome, $\mathbf{q}_j(\boldsymbol{\ell}_i)$ is a 2×1 vector composed of an intercept term and age, which corresponds to a 2×1 vector of coefficients $\boldsymbol{\eta}$, $\boldsymbol{\eta} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\eta})$. Additional spatial variability in the probability of inclusion is accounted for with $\boldsymbol{\upsilon}$, $\boldsymbol{\upsilon} \sim N(\mathbf{0}, \mathbf{6}_{\upsilon})$, where $\mathbf{6}_{\upsilon}$ is a 635×635 matrix and is defined by covariance function $K_{\upsilon}(d)$.

In addition, we take the two processes, ω and υ, to be independent. Collecting additional variance parameters in $\boldsymbol{\theta}$, the joint posterior distribution of (2) is proportional to

$$\begin{aligned} p(\boldsymbol{\omega}, \boldsymbol{\upsilon}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\eta}, \eta_y, \mathbf{y}_{ns} | \mathbf{y}_s, \mathbf{I}) \\ \propto p(\boldsymbol{\theta}) \times N(\boldsymbol{\omega} | \mathbf{0}, \boldsymbol{\Sigma}_{\omega}) \times N(\boldsymbol{\upsilon} | \mathbf{0}, \boldsymbol{\Sigma}_{\upsilon}) \times N(\boldsymbol{\beta} | \mathbf{0}, \boldsymbol{\Sigma}_{\beta}) \\ \times N(\boldsymbol{\eta} | \mathbf{0}, \boldsymbol{\Sigma}_{\eta}) \times N(\eta_y | 0, \sigma^2_{\eta_y}) \times \prod_{i=1}^{N} \prod_{j=1}^{M_i} \text{Ber}(I_j(\boldsymbol{\ell}_i) | \pi_j(\boldsymbol{\ell}_i)) \\ \times \prod_{i=1}^{N} \prod_{j=1}^{M_i} N(y_j(\boldsymbol{\ell}_i) | z_j(\boldsymbol{\ell}_i) - \mathbf{x}_j(\boldsymbol{\ell}_i)^{\top}\boldsymbol{\beta} + \omega(\boldsymbol{\ell}_i), \sigma^2), \end{aligned} \tag{5}$$

where

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

3.2 MCMC estimation strategy

Markov chain Monte Carlo must be used to sample from (5). A Gibbs update can be employed to sample the posterior distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$, which are

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

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

respectively. The conditional distributions for the remaining parameters are not available in closed form but can be sampled using a Metropolis–Hastings step. Specifically, we have

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

$$\times\prod_{i=1}^{N}\prod_{j=1}^{M_i}N(y_j(\boldsymbol{\ell}_i)|z_j(\boldsymbol{\ell}_i)-\mathbf{x}_j(\boldsymbol{\ell}_i)^{\top}\boldsymbol{\beta}+\omega(\boldsymbol{\ell}_i),\sigma^2)\ ,$$

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

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

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

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

3.3 Alternative models encompassing response mechanism and the extent of spatial variation in data

Four models are considered in the form of (2) and are described below. For these models, regression parameters are considered independent, e.g., $\boldsymbol{\Sigma}_{\beta}=\sigma^2_{\beta}\mathbf{I}_{10}$ and $\boldsymbol{\Sigma}_{\eta}=\sigma^2_{\eta}\mathbf{I}_2$, and their associated variance parameters, σ^2_{β} and σ^2_{η}, are fixed in both the simulation and data analysis. Similarly, σ_{η_y} is fixed. The spatial covariance functions are taken to be exponential, as described in Sect. 1.

Model 1. Non-spatial in outcome with ignorable response This model is a standard linear regression model and, therefore, spatial effects ($\boldsymbol{\omega}$ and $\boldsymbol{\upsilon}$) are fixed at 0.

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

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

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

Model 4. Spatial association in outcome and probability of inclusion with preferential response This model expands upon Model 3 by permitting spatial association in the probability of response. We take $\boldsymbol{\theta} = [\sigma^2, \delta_\omega^2, \phi_\omega, \delta_\upsilon^2, \phi_\upsilon]^\top$ and $p(\boldsymbol{\theta}) = IG(\sigma^2|a, b) \times IG(\delta_\omega^2|a_\omega, b_\omega) \times Unif(\phi_\omega|c_\omega, d_\omega) \times IG(\delta_\upsilon^2|a_\upsilon, b_\upsilon) \times Unif(\phi_\upsilon|c_\upsilon, d_\upsilon)$.

3.4 Model comparison and assessment strategy

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

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

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

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

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

4 Simulation

To examine the ability of the proposed models to capture various sampling schemes, a simplified dataset was simulated and three response scenarios were implemented. For simplicity, in this simulation study, we predict income (on the log-scale) with only the

covariates gender and household size for a finite population of size 2000, e.g., $z_j(\boldsymbol{\ell}_i)$ is fixed at 0 and $-\mathbf{x}_j(\boldsymbol{\ell}_i)$ is replaced by $\mathbf{x}_j(\boldsymbol{\ell}_i)$ for all i and j in (3). For each unit of the population, gender was drawn from a bernoulli distribution with the probability of female set to 0.8 and household size was drawn from a Poisson distribution with a mean of 4. To induce spatial correlation, a 5×5 square was created and 500 locations were randomly assigned within the square and distance matrix was constructed from these locations. The spatial process parameters were fixed at $\sigma^2 = 1, \delta_\omega^2 = 1$, and $\phi_\omega = 0.5$. Each unit of the population was randomly assigned to a location, with the requirement that at least one unit was located at each location. Regression parameters were fixed at $\boldsymbol{\beta} = [\beta_0, \beta_{\text{fem}}, \beta_{hhs}]^\top = [10, -0.2, 0.1]^\top$, to reflect an average income of $\exp(10) =$ \$22,000 in the reference group, a small average reduction in income for females, and a small average increase in income for larger household sizes. Log-income values were generated from (3).

Three scenarios were considered to reflect possible response scenarios in which there is spatial association in the outcome. In Scenario 1, income is from a spatial process but there is no preferential response. This arises from Model 3, fixing $\eta_y = 0$ and $\mathbf{q} = [1, \ldots, 1]^\top$. The probability of inclusion was set at 0.5, which is equivalent to fixing $\eta = 0$. This resulted in a selection of 54% of the simulated data. In the second scenario, income is from a spatial process which is reported preferentially, as described in Model 3. Here, η_y was set to 0.5 and $\boldsymbol{\eta} = [\eta_0, \eta_{\text{fem}}] = [-4, -1]^\top$, to reflect higher odds of response for larger values of income and lower odds of response for women. The choice of these coefficients resulted in 54.15% of the simulated data having income responses. The third scenario considers income as coming from a spatial process whose response in preferential and whose inclusion probability is dependent on another spatial process, which is described in Model 4. To reflect this, we set $\phi_\upsilon = 1.5$ and $\delta_\upsilon^2 = 1$; this resulted in responses in 48.1% of the simulated data. All data generation and analyses were performed using R version 3.6.1 (R Core Team, 2018).

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

Models were run for 10,000 iterations with 1000 burn-in, as examination of individual trace plots suggested sufficient mixing and convergence of the non-spatial parameters. At each iteration g, estimates of the nonsampled units were drawn and estimates for the population mean, $\bar{y}^{(g)} = \frac{1}{T}\Big(\sum_{i=1}^{n}\sum_{j=1}^{m_i}\exp[y_j(\boldsymbol{\ell}_i)] + \sum_{i=1}^{N}\sum_{j=m_i+1}^{M_i}\exp[y_j(\boldsymbol{\ell}_i)^{(g)}]\Big)$ were calculated. The variance parameter σ_β^2 was fixed at 1000 to reflect an uninformative prior, while the σ_η^2 and σ_{η_y} terms were fixed at 10 as a weakly informative prior restricting the range of the logistic regression coefficients. The non-spatial, σ^2, and spatial, δ_ω^2 and δ_υ^2, variance components were assigned prior

Table 3 Simulation results of scenario 1: spatial outcome, random response

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

distributions of IG(2,10), to reflect a small point mass centered at 10. The spatial range parameters, ϕ_ω and ϕ_υ, were assigned prior distributions of Unif(0.1, 2), to reflect a spatial range of 1.5 (3/2) to 30 (3/0.1). MCMC sampling was performed using the computer program JAGS (Plummer, 2017) in R.

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

The results of Scenario 2 are given in Table 4 and examines a preferential response of a spatially associated outcome. Importantly, unlike Scenario 1, the two non-spatial models fail to capture the true finite-population mean of 9.94 within their 95% credible intervals. This is also true of the intercept term, β_0, and non-spatial variance, σ^2, although we expect σ^2 to be larger, as it absorbing the variability in the outcome attributed to spatial association. Model 1 also incorrectly provides a positive estimative of β_{fem} whose credible interval does not contain the true value of -0.2. Moreover, while Models 2–4 provide similar estimates of η_{fem}, Model 2 fails to capture the true values of η_0 and η_y in its credible intervals, unlike the two spatial models. Possibly due to the poor modeling of income, Model 2 spuriously concludes that there is no evidence of preferential sampling. Finally, as in Scenario 1, both spatial models have

Table 4 Simulation results of scenario 2: spatial outcome, preferential sampling

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

similar estimates and correctly capture the spatial parameters ϕ_ω and δ^2_ω. In Model 4, even though ϕ_υ varies, it estimates very small values of δ^2_υ, which correctly suggests little evidence of spatial association in the probability of response. The model fit statistics both slightly favor Model 3 to Model 4, due to the lack of response-level spatial association, and prefer the spatial to non-spatial models.

When incorporating spatial association into the probability of income response, seen in Table 5, Model 4 outperforms the other three models in terms of model fit by correctly accounting for this additional association in the logistic regression component of the model. As before, non-spatial models have poorer model fit and larger estimates of the non-spatial variance term. Unlike Models 2–4, Model 1 fails to include the true finite-population mean in its credible interval, which may be attributable to a disregard for the preferential response. As in Scenario 2, Model 1 incorrectly provides a positive estimate of β_{fem}, and all models except Model 2 contain the true intercept in their credible intervals. Models 2–4 each correctly capture the logistic regression coefficients, η_0, η_{fem}, and η_y. In addition, the spatial models provide reasonable estimates of ϕ_ω and δ^2_ω, and in the case of Model 4, ϕ_υ and δ^2_υ.

5 Data analysis

5.1 Implementation

As before, Models 1–4 were implemented using JAGS (Plummer, 2017) in R and run for 10,000 iterations with 1000 burn-in, as examination of individual trace plots suggested sufficient mixing and convergence of the non-spatial parameters. At each iteration g, the finite-population mean income, $\bar{y}^{(g)} =$

Table 5 Simulation results of scenario 3: spatial outcome, non-ignorable sampling, spatial inclusion

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

$\exp\left[\frac{1}{T}\left(\sum_{i=1}^{n}\sum_{j=1}^{m_i} y_j(\boldsymbol{\ell}_i) + \sum_{i=1}^{N}\sum_{j=m_i+1}^{M_i} y_j(\boldsymbol{\ell}_i)^{(g)}\right)\right]$, and the finite-population mean percentage of income spent on fruits and vegetables, $\bar{y}^{(g)} = \frac{1}{T}\left(\sum_{i=1}^{n}\sum_{j=1}^{m_i} \exp\left[z_j(\boldsymbol{\ell}_i) - y_j(\boldsymbol{\ell}_i)\right] + \sum_{i=1}^{N}\sum_{j=m_i+1}^{M_i} \exp\left[z_j(\boldsymbol{\ell}_i) - y_j(\boldsymbol{\ell}_i)^{(g)}\right]\right)$, were calculated using estimates of the nonsampled units drawn at that iteration. The variance parameter σ^2_β was fixed at 1000 to reflect an diffuse prior, while the σ^2_η and σ_{η_y} terms were fixed at 0.68 as a weakly informative prior restricting the range of the exponentiated logistic regression coefficients to $\frac{1}{5}$ and 5. The non-spatial σ^2, and spatial, δ^2_ω, variance components were assigned prior distributions of IG(2,10) and IG(2,2), respectively, to reflect small point masses centered at 10 and 2. The prior for δ_υ was assigned to be uniform distribution ranging from 0 to 0.75, so that the standard deviation reported in Sect. 2 is included in this range. This tight prior was found to improve convergence in the other logistic regression parameters. The spatial range parameters, ϕ_ω and ϕ_υ, were assigned prior distributions of Unif (0.1, 2), to reflect a spatial range of 1.5 (3/2) to 30 (3/0.1). Computation times on a 2018 MacBook Pro laptop were negligible for Models 1 and 2 but were approximately 15 h for Model 3 and 24 h for Model 4.

5.2 Results

The results of this analysis are presented in Table 6. Notably, there is no evidence of an intervention effect on PIFV in any of the models, denoted by the coefficient $\beta_{\text{treat}*\text{follow}}$ being small and all credible intervals containing 0. An improvement in the intervention effect would have seen a larger positive coefficient. This finding supports previous findings of no community-level changes as reported in Ortega et al. (2016). The four models yield comparable estimates of all β regression coefficients,

Table 6 Results of regression models predicting percentage of income spent on fruits and vegetables (log-scale)

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

The finite-population mean income is presented in units of $ 10,000

so the following interpretations are based on Model 4. All else equal, the amount of spending on fruits and vegetables by men was estimated to have occurred with a multiple of 58% ($\exp(-0.54)$) applied to the corresponding spending by women. Larger reported households were associated with higher amounts of household income spent on fruits and vegetables, with PIFV multiplicatively increasing by 15% for every additional household member. Spending on fruits and vegetables by food purchasers who reported having less than a high-school education was estimated to have occurred with a multiple of 1.8 times the spending of those with a high-school diploma or more education. There was a small negative linear effect of age on the outcome, as well as a small positive quadratic term. PIFV was also lower at follow-up, which is consistent with the raw percentages presented in Table 1. There were no differences were detected for partner status.

Confirming the preliminary analyses discussed in Sect. 2, all three models that account for preferential response conclude that larger incomes are more likely to provide their income. Models 2–4 agree that age is not associated with the probability of response. Accounting for association in the probability of response appears to also best fits the data, as evidenced by the lowest value of D and highest GRS value. Interestingly, the model fit for Model 2 is poorest (on the GRS scale), suggesting that accounting for preferential sampling while not accounting for spatial association (either at the outcome or response levels) leads to poorer fit. In addition, Model 3 fits poorer than Model 1 (and Model 2 on the D scale), which suggests that spatial association at the outcome level may have been accounted for with the inclusion of additional covariates.

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

In addition, as posterior samples are drawn for all individuals with non-response, finite-population estimates can be constructed for each community at both time-points, which are presented for each model in Table 7. Bolded estimates represent instances where the 95% credible intervals do not include the raw average reported in Table 1.

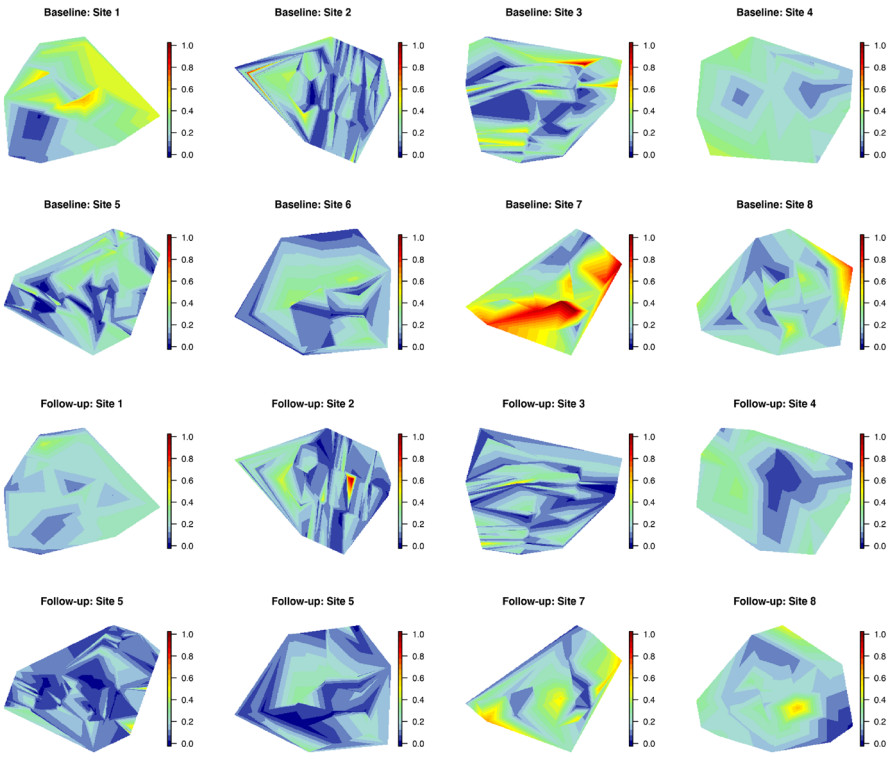

Fig. 3 Linear interpolation plots of estimated PIFV by site and time-point from model 4

Importantly, these results emphasize the importance of imputation. Models 2–4 show remarkable similarity in these estimates and conclude that raw data of 6 of the 8 sites underestimates the percentage of income spent on fruits and vegetables at baseline and all but 1 of the 8 underestimate at follow-up. Model 3 additionally identifies site 1 at baseline, but this is not supported by the rest of the models. Even in the case of Model 1, at baseline 3 were found to underestimate the percentage and 1 suggested overestimation, and at follow-up, 2 communities were found to underestimate as well. Encouragingly, in all but one case (site 7 at baseline) of the disagreements with the raw data that Model 1 identified, Models 2–4 also identified these cases. In addition, Models 2–4 suggest that the baseline total is underestimating the true average and all models agree that the follow-up total is underestimated. Interpolated maps corresponding to these finite-population estimates from Model 4 are presented in Fig. 3.

6 Conclusion

This paper presents a new framework to account for data whose outcome is spatially associated and where the probability of response is assumed to be associated with the value of the outcome. We examine the implications of this data on finite-population

Table 7 Finite-population estimates and 95% CI of percentage of income spent on fruits and vegetables by community and timepoint

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

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

quantities and demonstrate how to perform Bayesian estimation on these values. This works builds on an existing literature in spatial statistics, Bayesian finite-population estimation, and missing data and has a wide range of applications in health, economics, and environmental work.

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

the level of income). This work can also assist in more accurate needs assessments of local communities and, therefore, improve the allocation of health resources. Further, as there is interest in estimating intervention effects, new approaches described in the casual inference literature which can account spatial association (Akbari et al., 2021) may be appropriate.

The literature of Bayesian finite-population estimation in the presence of spatial association is limited and future extensions to the work presented in this paper are numerous. While this model draws on the preferential sampling framework described by Diggle et al. (2010), we examined a missing data case that had similar evidence of preferential response. However, a data analysis implementing this technique on a dataset with preferential sampling from a finite population would be a strong addition to the literature. The authors view the framework discussed in Sect. 3.1 to be flexible enough to allow for other, more complicated sampling schemes as well, although more simulation work would be needed to fully understand the implications of these on finite-population quantities, especially if spatial association is assumed. Further, while a linear relationship between the log-odds of response and income was assumed, other relationships may be considered in future works.

In addition, while the sample size presented in the data analysis of this paper was small, this framework can be extended to account for massive sample sizes. The problem of spatial modeling for big data stems from the inversion of dense covariance matrices, but modern work in covariance approximation has made this feasible. Such techniques include low-rank models, sparsity-inducing processes, and map reducing approaches (Banerjee, 2017; Heaton et al., 2018; Guhaniyogi & Banerjee, 2018; Banerjee, 2020), see, e.g., and references therein.

Further, while the authors have only considered a Gaussian process to describe the outcome variable, this framework could be extended to other processes, such as mixtures of Gaussian processes (Neelon et al., 2014), a generalized Gaussian process (Chan & Dong, 2011), or a spatial Dirichlet process (Gelfand et al., 2005). Extensions to multivariate responses and spatio-temporal data may also serve useful, particularly when examining health outcomes. Finally, learning about spatial difference boundaries (Gao et al., 2022) from finite population estimates for regionally aggregated health outcomes is witnessing growing interest among public health researchers and will comprise future investigations.

Funding Funding for Chan-Golston and Banerjee was provided by National Institute of Environmental Health Sciences (Grant nos. R01ES027027 and R01ES030210), Division of Mathematical Sciences (Grant nos. 1916349 and 2113778).

Declarations

Conflict of interest On behalf of all the authors, the corresponding author states that there is no conflict of interest.

References

Akbari, K., Winter, S., & Tomko, M. (2021). Spatial causality: A systematic review on spatial causal inference. *Geographical Analysis*. https://doi.org/10.1111/gean.12312.

Albert, S. L., Langellier, B. A., Sharif, M. Z., Chan-Golston, A. M., Prelip, M. L., Garcia, R. E., et al. (2017). A corner store intervention to improve access to fruits and vegetables in two Latino communities. *Public Health Nutrition, 20*(12), 2249–2259.

Antonelli, J., Cefalu, M., & Bornn, L. (2016). The positive effects of population-based preferential sampling in environmental epidemiology. *Biostatistics, 17*(4), 764–778.

Banerjee, S. (2017). High-dimensional Bayesian geostatistics. *Bayesian Analysis, 12,* 583–614.

Banerjee, S. (2020). Modeling massive spatial datasets using a conjugate Bayesian linear modeling framework. *Spatial Statistics,37,* 100417. https://doi.org/10.1016/j.spasta.2020.100417 (**Frontiers in Spatial and Spatio-temporal Research**).

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data* (2nd ed.). Chapman & Hall/CRC.

Bradley, J. R., Holan, S. H., & Wikle, C. K. (2015). Multivariate spatio-temporal models for high-dimensional areal data with application to longitudinal employer-household dynamics. *The Annals of Applied Statistics, 9*(4), 1761–1791. https://doi.org/10.1214/15-AOAS862.

Bradley, J. R., Holan, S. H., & Wikle, C. K. (2016). Multivariate spatio-temporal survey fusion with application to the American community survey and local area unemployment statistics. *Stat, 5*(1), 224–233.

Bradley, J. R., Wikle, C. K., & Holan, S. H. (2016). Bayesian spatial change of support for count-valued survey data with application to the American community survey. *Journal of the American Statistical Association, 111*(514), 472–487. https://doi.org/10.1080/01621459.2015.1117471.

Breau, S., Shin, M., & Burkhart, N. (2018). Pulling apart: New perspectives on the spatial dimensions of neighbourhood income disparities in Canadian cities. *Journal of Geographical Systems, 20*(1), 1–25.

Bruno, F., Cocchi, D., & Vagheggini, A. (2013). Finite population properties of individual predictors based on spatial pattern. *Environmental and Ecological Statistics, 20*(3), 467–494.

Casey, P. H., Szeto, K., Lensing, S., Bogle, M., & Weber, J. (2001). Children in food-insufficient, low-income families: Prevalence, health, and nutrition status. *Archives of Pediatrics and Adolescent Medicine, 155*(4), 508–514. https://doi.org/10.1001/archpedi.155.4.508.

Chakravorty, S. (1996). A measurement of spatial disparity: The case of income inequality. *Urban Studies, 33*(9), 1671–1686.

Chan, A.B., & Dong, D. (2011). Generalized Gaussian process models. In *CVPR 2011* (pp. 2681–2688). doi:https://doi.org/10.1109/CVPR.2011.5995688

Chan-Golston, A. M., Banerjee, S., & Handcock, M. S. (2020). Bayesian inference for finite populations under spatial process settings. *Environmetrics, 31*(3), 2606. https://doi.org/10.1002/env.2606.

Cicchitelli, G., & Montanari, G. E. (2012). Model-assisted estimation of a spatial population mean. *International Statistical Review, 80*(1), 111–126.

Clayton, D., & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics, 43*(3), 671–681.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). Wiley.

Cressie, N., & Wikle, C. K. (2011). *Statistics for spatio-temporal data*. Wiley.

David, M., Little, R. J. A., Samuhel, M. E., & Triest, R. K. (1986). Alternative methods for cps income imputation. *Journal of the American Statistical Association, 81*(393), 29–41.

Diggle, P. J., Menezes, R., & Su, T.-L. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C, 59*(2), 191–232.

Ericson, W. A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society, Series B, 31*(2), 195–233.

Gao, L., Banerjee, S., & Ritz, B. (2022). Spatial difference boundary detection for multiple outcomes using Bayesian disease mapping. *Biostatistics*. https://doi.org/10.1093/biostatistics/kxac013.

Gelfand, A. E., & Ghosh, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika, 85*(1), 1–11.
Gelfand, A. E., Kottas, A., & MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association, 100*(471), 1021–1035.
Gelfand, A. E., Sahu, S. K., & Holland, D. M. (2012). On the effect of preferential sampling in spatial prediction. *Environmetrics, 23*(7), 565–578.
Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science, 22*(2), 153–164.
Ghosh, M., & Meeden, G. (1997). *Bayesian methods for finite population sampling*. Chapman & Hall.
Ghosh, M., & Rao, J. N. K. (1994). Small area estimation: An appraisal. *Statistical Science, 9*(1), 55–93.
Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102*(447), 359–378.
Greenlees, J. S., Reece, W. S., & Zieschang, K. D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association, 77*(378), 251–261.
Grimm, K. A., Foltz, J. L., Blanck, H. M., & Scanlon, K. S. (2012). Household income disparities in fruit and vegetable consumption by state and territory: Results of the 2009 behavioral risk factor surveillance system. *Journal of the Academy of Nutrition and Dietetics, 112*(12), 2014–2021.
Guhaniyogi, R., & Banerjee, S. (2018). Meta-kriging: Scalable Bayesian modeling and inference for massive spatial datasets. *Technometrics, 60*(4), 430–444. https://doi.org/10.1080/00401706.2018.1437474.
Hartley, H. O., & Sielken, R. L., Jr. (1975). A "Super-Population Viewpoint" for finite population sampling. *Biometrics, 31*(2), 411–422.
Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., et al. (2018). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*. https://doi.org/10.1007/s13253-018-00348-w.
Hoef, J. V. (2002). Sampling and geostatistics for spatial data. *Écoscience, 9*(2), 152–161.
Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*(260), 663–685.
Langellier, B. A., Garza, J. R., Prelip, M. L., Glik, D., Brookmeyer, R., & Ortega, A. N. (2013). Corner store inventories, purchases, and strategies for intervention: A review of the literature. *Californian Journal of Health Promotion, 11*(3), 1–13.
Lawman, H. G., Veur, S. V., Mallya, G., McCoy, T. A., Wojtanowski, A., Colby, L., et al. (2015). Changes in quantity, spending, and nutritional characteristics of adult, adolescent and child urban corner store purchases after an environmental intervention. *Preventive Medicine, 74,* 81–85.
Lee, A., Szpiro, A., Kim, S. Y., & Sheppard, L. (2015). Impact of preferential sampling on exposure prediction and health effect inference in the context of air pollution epidemiology. *Environmetrics, 26*(4), 255–267.
Little, R. J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association, 99*(466), 546–556.
Neelon, B., Gelfand, A. E., & Miranda, M. L. (2014). A multivariate spatial mixture model for areal data: Examining regional differences in standardized test scores. *Journal of the Royal Statistical Society Series C (Applied Statistics), 63*(5), 737–761.
Ortega, A. N., Albert, S. L., Chan-Golston, A. M., Langellier, B. A., Glik, D. C., Belin, T. R., et al. (2016). Substantial improvements not seen in health behaviors following corner store conversions in two Latino food swamps. *BMC Public Health, 16*(389), 1–10.
Ortega, A. N., Albert, S. L., Sharif, M. Z., Langellier, B. A., Garcia, R. E., Glik, D. C., et al. (2015). A multi-level, community-engaged corner store intervention in East Los Angeles and Boyle Heights. *Journal of Community Health, 40,* 347–356.
Paek, H.-J., Oh, H. J., Jung, Y., Thompson, T., Alaimo, K., Risley, J., & Mayfield, K. (2014). Assessment of a healthy corner store program (fit store) in low-income, urban, and ethnically diverse neighborhoods in Michigan. *Family & Community Health, 37*(1), 86–99.
Pati, D., Reich, B. J., & Dunson, D. B. (2011). Bayesian geostatistical modelling with informative sampling locations. *Biometrika, 98*(1), 35–48.
Plummer, M. (2017). JAGS Version 4.3.0 User Manual. International Agency for Research on Cancer, Lyon, France. International Agency for Research on Cancer

R Core Team. (2018). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. R Foundation for Statistical Computing. https://www.R-project.org/

Rao, J. N. K. (2003). *Small area estimation*. Wiley.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. The MIT Press.

Ribar, D. C., & Hamrick, K. S. (2003). Dynamics of poverty and food sufficiency. *30*

Riphahn, R. T., & Serfling, O. (2005). Item non-response on income and wealth questions. *Empirical Economics, 30*(2), 521–538.

Ripley, B. D. (2004). *Spatial statistics*. Wiley.

Rose, D. (1999). Economic determinants and dietary consequences of food insecurity in the United States. *The Journal of Nutrition, 129*(2), 517–520.

Rose, D., Bodor, J. N., Swalm, C. M., Rice, J. C., Farley, T. A., & Hutchinson, P. L. (2009). *Deserts in new orleans?* Illustrations of urban food access and implications for policy: University of Michigan National Poverty Center/USDA Economic Research Service Research.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika, 57*(2), 377–387.

Schenker, N., Raghunathan, T. E., Chiu, P.-L., Makuc, D. M., Zhang, G., & Cohen, A. J. (2006). Multiple imputation of missing income data in the national health interview survey. *Journal of the American Statistical Association, 101*(475), 924–933.

Song, H.-J., Gittelsohn, J., Kim, M., Suratkar, S., Sharma, S., & Anliker, J. (2009). A corner store intervention in a low-income urban community is associated with increased availability and sales of some healthy foods. *Public Health Nutrition, 12*(11), 2060–2067. https://doi.org/10.1017/S1368980009005242.

Thorndike, A. N., Bright, O.-J.M., Dimond, M. A., Fishman, R., & Levy, D. E. (2017). Choice architecture to promote fruit and vegetable purchases by families participating in the special supplemental program for women, infants, and children (wic): randomized corner store pilot study. *Public Health Nutrition, 20*(7), 1297–1305. https://doi.org/10.1017/S1368980016003074.

Watson, N., & Starick, R. (2011). Evaluation of alternative income imputation methods for a longitudinal survey. *Journal of Official Statistics, 27*(4), 693.

Yan, T., Curtin, R., & Jans, M. (2010). Trends in income nonresponse over two decades. *Journal of Official Statistics, 26*(1), 145.