Published in final edited form as:

J R Stat Soc Ser A Stat Soc. 2015 October ; 178(4): 1009–1023. doi:10.1111/rssa.12103.

Bayesian hierarchical models for smoothing in two-phase studies, with application to small area estimation

Michelle Ross and

Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA USA

Jon Wakefield

Departments of Biostatistics and Statistics, University of Washington, Seattle, WA USA

Summary

Two-phase study designs are appealing since they allow for the oversampling of rare sub-populations which improves efficiency. In this paper we describe a Bayesian hierarchical model for the analysis of two-phase data. Such a model is particularly appealing in a spatial setting in which random effects are introduced to model between-area variability. In such a situation, one may be interested in estimating regression coefficients or, in the context of small area estimation, in reconstructing the population totals by strata. The efficiency gains of the two-phase sampling scheme are compared to standard approaches using 2011 birth data from the research triangle area of North Carolina. We show that the proposed method can overcome small sample difficulties and improve on existing techniques. We conclude that the two-phase design is an attractive approach for small area estimation.

Keywords

Small area estimation; Outcome-dependent sampling; Markov chain Monte Carlo; Bayesian hierarchical model

1. Introduction

It is often the case that outcome data, along with some exposure and/or confounder information is available on a large sample from a population of interest. For example, it is common for individuals with and without disease to be cross-classified by area, age and gender. However, due to high costs, missing data or feasibility, additional information is only available on a small subset of individuals. A two-phase study design is one in which limited information (including outcome) is known for a large “first-phase” population, and more extensive information is known for a “second-phase” subset of these individuals. Such studies offer large efficiency gains since they allow for the oversampling of rare subpopulations. We consider a two-phase design in which, at phase I, a simple random sample is drawn from a hypothetical population of size M and individuals are cross-classified with respect to a binary outcome variable, y , and to stratification variables, z . At phase II, subjects are randomly sampled from the cells of the cross-classified data and additional data are collected on discrete exposure variables, which we label x (though these

variables may be confounders in some contexts). Analyzing the second phase data using conventional methods leads to biased estimates of the parameters of interest since the phase II sample is not representative of the entire first-phase study population. Hence, specialized methods of analysis are required. Table 1 defines the notation that we adopt, where M_{yxz} denotes the (unobserved) number of individuals in the population with disease outcome y , exposure level x and strata level z , and M^{yxz} denotes the vector of such counts.

Two-phase studies were introduced by Neyman (1938), who described “double sampling” estimation methods. White (1982) discussed the use of a two-stage approach in studies of the relationship between a rare disease and a rare exposure in an effort to gain efficiency of odds ratio estimation over the standard case-control scheme. Approaches to the analysis of two-phase data have since become more complex, with a number of likelihood-based methods being proposed. The simplest of these have their origins in sampling theory, where weighted design-based estimators are often used (Flanders and Greenland, 1991; Reilly and Pepe, 1995). Data on auxiliary variables can be used to improve estimates by adjusting the sampling weights using calibration (Robins et al., 1994), modified and centered calibration (Saegusa and Wellner, 2013), and estimated weights (Deville and Särndal, 1992; Lumley et al., 2011). Other likelihood variants have been proposed in the two-phase context including pseudo-likelihood (Breslow and Cain, 1988; Scott and Wild, 1991; Schill et al., 1993) and non-parametric maximum likelihood (Breslow and Holubkov, 1997a,b; Scott and Wild, 1991, 1997). There is limited literature on Bayesian approaches to the analysis of two-phase studies. Ross and Wakefield (2013) present one approach which utilizes a log-linear model and Ahn et al. (2013) describe a similar approach to estimate gene-gene and/or gene-environment interactions.

In this paper, we consider the use of two-phase study designs in the context of small area estimation (SAE), where the goal is to reconstruct the unobserved population totals, M^{yxz} . One applicable context is a study of influenza where counts of mild and severe cases are obtained by area in a given year, and the phase II subset are sampled for pathology testing to determine which pathogens caused the disease. In this situation, we may wish to know the population counts of disease severity by area and pathogen type for vaccine implementation. While we concentrate on SAE, the hierarchical modeling approach we describe can also be used in a regression context in which we are explicitly interested in estimating associations between a response and exposures/confounders. We describe two possible examples. The first is in the context of a two-phase study described by Sharpe et al. (2000) to examine the effects of non-steroidal anti-inflammatory drugs (NSAIDs) on breast cancer risk and stage in Saskatchewan, Canada between 1981 and 1995. The phase I population consisted of women aged 45-79 years with known disease and exposure status. At phase II, information on potential confounders was collected through telephone interviews, where unexposed and highly exposed individuals were oversampled at this stage. Taking into account location information (through postal codes, for example) would have helped to control for unmeasured confounding such as socioeconomic status. We may also consider data from a case-control study which could be combined with census and cancer registry data at the zip code level, and could then be analyzed as a two-phase study. In this case, the case-control data would constitute the phase II sample and the phase I sample would consist of the census

and cancer registry data. Davis et al. (2002) describe a case-control study of breast cancer and magnetic field exposure in King and Snohomish counties in Washington that may benefit from such an approach.

In small area estimation, only small survey samples are generally available to estimate counts of interest. Hierarchical models are used to borrow strength across areas using unstructured and/or spatially structured random effects (Fay and Herriot, 1979; Rao, 2003, Section 5.4.4). A weighted likelihood can be constructed to acknowledge the sampling design (Binder, 1983; Skinner, 1989) and random effects can be included in such models (Congdon and Lloyd, 2010; Rabe-Hesketh and Skrondal, 2006). By adopting the two-phase study design, we can improve estimates of the population counts of interest by using information known about the study population to efficiently sample the phase II data. In survey sampling, common approaches to utilize population information are post-stratification and raking (Lumley, 2010, Chapter 7). The conventional approach is to adjust the design weights so that the population totals are recovered. Further efficiency gains are possible by adopting a hierarchical Bayesian approach that smooths across areas.

We describe an approach that extends the model of Ross and Wakefield (2013) to include both spatial and non-spatial random effects to account for confounding by location allowing for greater modeling flexibility. The methods are illustrated using birth data from North Carolina where we wish to estimate the association between birth weight and race, controlling for gender and their interaction, and to obtain point and interval estimates for $M^{y,x,z}$, the population counts that we do not observe in a two-phase study. For these data, we construct several survey sampling schemes and two-phase designs in order to compare the efficiency of standard survey approaches with two-phase approaches in both parameter and small area estimation. Analyses of the complete individual-level data serve as a benchmark for the other analyses. We show that the count reconstruction can be carried out accurately based on a fraction of the original data.

2. Bayesian Hierarchical Model

In this section, we derive the likelihood for the observed data. It is convenient to adopt a log-linear model, since this may be used for either simple random or case-control sampling at phase I. However, we combine it with a logistic disease model, whose parameters are more interpretable.

2.1. Likelihood

In the context of spatial epidemiology, a two-phase study design may be adopted where one may wish to acknowledge confounding by location using spatial random effects (Wakefield and Haneuse, 2008). In this case, we assume that the stratification variable z represents area, of which there are m_z .

The observed data consist of the response stratum margin observed at phase I, $N^{y,z} = \{N_{y,z}, y = 0, 1; z = 1, \dots, m_z\}$, the phase II sample sizes $n^{y,z} = \{n_{y,z}, y = 0, 1; z = 1, \dots, m_z\}$ and the phase II outcomes $n^{y,x,z} = \{n_{y,x,z}, y = 0, 1; x = 1, \dots, m_x; z = 1, \dots, m_z\}$. At phase II, sample sizes $n^{y,z}$ are decided upon and individuals are drawn at random from the $N^{y,z}$ and exposure

information, x , is measured for each individual. Hence, $\mathbf{n}^{y \cdot z}$ is fixed by design and must be smaller than the corresponding (random) $N^{y \cdot z}$ values. However, this term does not depend on λ and so it need not be considered in the likelihood (Ross and Wakefield, 2013). Assuming simple random sampling at phase I, we condition on the grand total $N \dots$ and the likelihood is

$$pr(N^{y \cdot z}, \mathbf{n}^{y \cdot z}, \mathbf{n}^{y x z} | N \dots, \lambda) \propto pr(N^{y \cdot z} | N \dots, \lambda) \times pr(\mathbf{n}^{y x z} | \mathbf{n}^{y \cdot z}, \lambda), \quad (1)$$

where

$$N^{y \cdot z} | N \dots, \lambda \sim \text{Multinomial}_{2 \times m_z} \left(N \dots, \{pr(Y=y, Z=z | \lambda)\}_{y=0,1; 1 \leq z \leq m_z} \right), \quad (2)$$

and

$$\mathbf{n}^{y x z} | \mathbf{n}^{y \cdot z}, \lambda \sim_{\text{ind}} \text{Multinomial}_{m_x} \left(\mathbf{n}_{y \cdot z}, \{pr(X=x | Z=z, Y=y | \lambda)\}_{1 \leq x \leq m_x} \right), \quad y=0,1; 1 \leq z \leq m_z, \quad (3)$$

where λ is the vector of parameters in a log-linear model, which we describe shortly. Breslow and Chatterjee (1999) and Ross and Wakefield (2013) give further details of this derivation.

We model the association between covariates and the outcome using a linear logistic disease model, which includes structured and unstructured random effects for area:

$$\log \left[\frac{pr(Y=1 | X=x, Z=z)}{pr(Y=0 | X=x, Z=z)} \right] = \mathbf{W}\beta + \mathbf{U}_z + \mathbf{V}_z, \quad (4)$$

where $\mathbf{W}$ represents the design matrix, and U_z and V_z are random effects with and without spatial structure, respectively. Without loss of generality, we suppose there is simple random sampling of the phase I data. The case-control sampling at phase I scenario follows in a straightforward manner. To model the complete Y, X, Z data, we embed the disease model (4) within a log-linear model. In terms of the log-linear parameterization, we have

$$\log(\mu_{y x z}) = \mu + \lambda_y^Y + \lambda_x^X + \lambda_z^Z + \lambda_{yx}^{YX} + \lambda_{yz}^{YZ} + \lambda_{zx}^{XZ}, \quad (5)$$

where $\mu_{y x z} = E[N_{y x z} | N \dots, \lambda]$, the mean of the disease-stratum-confounder count in cell (y, x, z) of the $2 \times m_x \times m_z$ contingency table, $y = 0, 1$; $x = 1, \dots, m_x$; $z = 1, \dots, m_z$. We do not consider a saturated log-linear model, as this would correspond to including an interaction term between X and Z in the disease model (4), which is not of interest. Constraints need to be imposed on the log-linear model parameters for identifiability. Ultimately, our goal is to relate parameters in the log-linear model to those in (4) so that inference can be based on the coefficients from the disease model. Under the sum-to-zero constraints

$$\sum_y \lambda_y^Y = \sum_x \lambda_x^X = \sum_z \lambda_z^Z = \sum_y \lambda_{yx}^{YX} = \sum_x \lambda_{yx}^{YX} = \sum_x \lambda_{xz}^{XZ} = \sum_z \lambda_{xz}^{XZ} = 0,$$

the β and λ parameters are related through

$$\begin{aligned}\beta_0 &= -2 \left(\lambda_0^Y + \lambda_{01}^{YX} \right) \\ \beta_x &= -2 \left(\lambda_{0x}^{YX} - \lambda_{01}^{YX} \right)\end{aligned}$$

for $x = 2, \dots, m_x$. The λ_{yz}^{YZ} are related to the random effects parameters U_z and V_z via $U_z + V_z = \lambda_{1z}^{YZ} - \lambda_{0z}^{YZ}$, $z = 1, \dots, m_z$. We will place priors on $\mathbf{U} = (U_1, \dots, U_{m_z})$ and $\mathbf{V} = (V_1, \dots, V_{m_z})$, without imposing constraints on these random effects. Hence, we do not impose any constraints on the z margin of λ_{yz}^{YZ} . In addition, we want to assume symmetric effects in the $Y = 0$ and $Y = 1$ groups, so we constrain the y margin such that $\sum_y \lambda_{yz}^{YZ} = 0$. Then, $U_z + V_z = -2\lambda_{0z}^{YZ}$, for $z = 1, \dots, m_z$.

The probabilities in (2) and (3) are calculated using the log-linear model (5):

$$\begin{aligned}pr(Y=y, Z=z|\boldsymbol{\lambda}) &= \sum_{x=1}^{m_x} \left(\frac{\mu_{y x z}}{\sum_u \sum_v \sum_w \mu_{u v w}} \right) \\ pr(X=x|Z=z, Y=y|\boldsymbol{\lambda}) &= \frac{\sum_u \sum_v \sum_w \mu_{u v w}}{\sum_{v=1}^{m_x} \left(\frac{\mu_{y v z}}{\sum_u \sum_{v'} \sum_w \mu_{u v' w}} \right)}.\end{aligned}$$

2.2. Prior Specification

With respect to the log-linear model (5), let $\boldsymbol{\lambda}^Y = \lambda_0^Y$, $\boldsymbol{\lambda}^X = (\lambda_1^X, \dots, \lambda_{m_x-1}^X)$, $\boldsymbol{\lambda}^Z = (\lambda_1^Z, \dots, \lambda_{m_z-1}^Z)$, $\boldsymbol{\lambda}^{YX} = (\lambda_{01}^{YX}, \dots, \lambda_{0, m_x-1}^{YX})$, $\boldsymbol{\lambda}^{XZ} = (\lambda_{11}^{XZ}, \dots, \lambda_{m_x-1, m_z-1}^{XZ})$ and let $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}^Y, \boldsymbol{\lambda}^{YX})$. The relationship between $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ is described via $\boldsymbol{\beta} = \mathbf{C} \boldsymbol{\Lambda}^Y$, where the invertible matrix $\mathbf{C}$ is an $m_z \times m_z$ square matrix. Computations are performed for $\boldsymbol{\lambda}$ first, and we then transform to $\boldsymbol{\beta}$ using $\mathbf{C}$.

We specify prior distributions for $\boldsymbol{\Lambda}^Y, \boldsymbol{\lambda}^X, \boldsymbol{\lambda}^Z, \boldsymbol{\lambda}^{XZ}, \mathbf{U}$ and $\mathbf{V}$, as well as τ_u and τ_v , the precision of the spatial and non-spatial random effects, respectively, where we assume

$$\pi(\boldsymbol{\Lambda}^Y, \boldsymbol{\lambda}^X, \boldsymbol{\lambda}^Z, \boldsymbol{\lambda}^{XZ}, \mathbf{U}, \mathbf{V} | \tau_u, \tau_v) = \pi(\boldsymbol{\Lambda}^Y) \pi(\boldsymbol{\lambda}^X) \pi(\boldsymbol{\lambda}^Z) \pi(\boldsymbol{\lambda}^{XZ}) \pi(\mathbf{U} | \tau_u) \pi(\mathbf{V} | \tau_v).$$

We assume a multivariate normal prior distribution for $\boldsymbol{\beta}$ with mean $\mathbf{m}$ and variance-covariance matrix $\boldsymbol{\Sigma}_\beta$, where $\mathbf{m}$ and $\boldsymbol{\Sigma}_\beta$ are chosen based on the context. Then, the induced multivariate normal prior for $\boldsymbol{\Lambda}^Y$ has mean $\mathbf{C}^{-1}\boldsymbol{\mu}$ and variance-covariance matrix $\mathbf{C}^{-1}\boldsymbol{\Sigma}_\beta(\mathbf{C}^{-1})^T$.

For $\boldsymbol{\lambda}^X, \boldsymbol{\lambda}^Z$ and $\boldsymbol{\lambda}^{XZ}$, which are usually nuisance parameters, we use the specification proposed by Knuiman and Speed (1988), and refined by Dellaportas and Forster (1999) as described in Ross and Wakefield (2013). In this approach, independent, mean zero

multivariate normal priors are placed on the collections of main effects and interactions with variance $m_x m_z \Omega_a$, $a \in \{X, Z, XZ\}$, where

$$\Omega_a = \frac{1}{m_x m_z} \prod_{\gamma \in a} |I_\gamma| \otimes \left(I_{|I_\gamma|-1} - \frac{1}{|I_\gamma|} \mathbf{J}_{|I_\gamma|-1} \right),$$

I_γ is the set of levels of factor γ , $\mathbf{I}_{|I_\gamma|}$ is the $|I_\gamma| \times |I_\gamma|$ identity matrix and $\mathbf{J}_{|I_\gamma|}$ is the $|I_\gamma| \times |I_\gamma|$ matrix containing all ones. This specification incorporates the constraints on main effects and interaction parameters of the log-linear model directly into the prior distribution.

We place independent normal prior distributions on the non-spatial components $\mathbf{V}$,

$V_z | \tau_v \sim N(0, \tau_v^{-1})$, for $z = 1, \dots, m_z$, and we adopt the intrinsic conditional autoregression (ICAR) model for $\mathbf{U}$:

$$\pi(\mathbf{U} | \tau_u) \propto \tau_u^{(m_z-1)/2} \exp\left(-\frac{\tau_u}{2} \mathbf{U}^T \mathbf{K} \mathbf{U}\right),$$

where $\mathbf{K}$ is an $m_z \times m_z$ matrix representing the neighbourhood structure of the areas in which the (i, j) th off-diagonal element is -1 if areas i and j are neighbors and 0 otherwise, and the i th diagonal element is the number of neighbors of area i (Besag et al., 1991). For the precision parameters τ_v and τ_u , we adopt conjugate Gamma priors, $\tau_v \sim G(a_v, b_v)$, $\tau_u \sim G(a_u, b_u)$, where the shape parameters, a_v and a_u , and rate parameters, b_v and b_u , are chosen based on the context. See Sørbye and Rue (2014) for a detailed discussion on the selection of hyperprior parameters.

We construct a Markov chain and update the parameters using separate Metropolis-Hastings steps for $\boldsymbol{\lambda} = (\boldsymbol{\lambda}^Y, \boldsymbol{\lambda}^X, \boldsymbol{\lambda}^Z, \boldsymbol{\lambda}^{YX}, \boldsymbol{\lambda}^{XZ})$, $\mathbf{U}$ and $\mathbf{V}$, and Gibbs steps for τ_u and τ_v . Details are provided in Appendix A. Convergence was assessed through visual inspection of trace plots by running three chains from different starting points.

2.3. Predicting Population Counts

In the Bayesian paradigm, it is straightforward to embed the prediction of the population counts in each disease-exposure-stratum group within the MCMC scheme. It is less straightforward to obtain the population count predictions using frequentist models, as the X - Z marginal distribution is left unspecified (see Appendix B.1). We implement an auxiliary variable scheme which requires sampling both the phase I internal cells, N_{yxz} , and the population counts, M_{yxz} . Details are provided in Appendix B.2.

The predictions are evaluated using the bias, variance and mean squared error (MSE) with respect to the posterior distribution. For T draws from the posterior distribution, we have

$$\begin{aligned}\text{Bias}_{yxz} &= \left(\widehat{M}_{yxz} - M_{yxz} \right), \quad \widehat{M}_{yxz} = \frac{1}{T} \sum_{t=1}^T \hat{M}_{yxz}^{(t)}, \\ \text{Variance}_{yxz} &= \left(\frac{1}{T} \sum_{t=1}^T \left(\hat{M}_{yxz}^{(t)} - \widehat{M}_{yxz} \right)^2 \right), \\ \text{MSE} &= \sum_{y=0}^1 \sum_{x=1}^{m_x} \sum_{z=1}^{m_z} \left(\text{Bias}_{yxz}^2 + \text{Variance}_{yxz} \right).\end{aligned}$$

Estimates with small MSE will be generally preferred.

3. North Carolina Low Birth Weight Example

Vital statistics are collected annually for all North Carolina residents by the North Carolina State Center for Health Services (SCHS). We consider all births from 2011 and focus on the 13 counties making up the research triangle area of North Carolina (Chatham, Durham, Franklin, Granville, Harnett, Johnston, Lee, Moore, Orange, Person, Vance, Wake and Warren). Data were collected on an infant's race (white, non-white), gender and birth weight status (low, normal), where a low birth weight is defined to be a birth weight of less than 2,500 g. There were a total of 26,837 births and the proportion of infants that were of low birth weight ranged from 7% to 15% across the study region (Figure 1). Recall that our aims here are two-fold. First, we wish to estimate the association between birth weight and race, controlling for gender and their interaction. Second, we wish to obtain point and interval estimates for M^{yxz} , the population counts that we do not observe in a two-phase study.

We first consider an analysis of the full individual level data keeping the two aims in mind. We fit a Bayesian spatial random effects model to the data using the INLA package in R (Rue et al., 2009; Martins et al., 2012). R code to implement this model is available in Appendix C and at <http://faculty.washington.edu/jonno/software.html>. Point estimates and 95% intervals for the regression coefficients are displayed in Figure 2. Based on this analysis, we would conclude that white race is statistically significantly associated with decreased odds of low birth weight status compared to non-white race, and that sex significantly modifies this association. For the small area estimation goal, we focus on the counties with the largest numbers of low birth weight infants. Wake and Durham counties have the biggest disease burden, with the counts being largest for non-white females and non-white males. Hence, special attention should be paid to these areas with respect to resource allocation and public health intervention strategies.

In the remainder of this section, we sample from the full North Carolina birth data in order to create an artificial two-phase study. We then analyze the subsequent data (fitting three Bayesian models), and compare our conclusions with those reached from the analysis of the full data. We then consider two simulated datasets, both of which are based on the North Carolina birth data in order to ensure a realistic setting: one with spatial structure to highlight the efficiency of the spatial model over the alternatives that do not include spatial aspects, the other without spatial structure to illustrate that we do not lose efficiency by fitting a spatial model to data with no spatial structure.

3.1. Substantive Analysis of a Simulated Two-phase Study

To mimic a two-phase study, we will assume that, at phase I, we have observed the birth weight status counts by county for a random sample of 5,000 infants. In this instance, the phase I stratification variable is county. At phase II, low and normal birth weight infants are sampled from within each county and we assume that additional data are collected on an infant's race and gender. In many situations, the total size of the phase II sample will be fixed in advance (for example, due to cost). In this case, it is natural to consider the best way to sample these individuals. To explore this idea further, we examine two different sampling designs for obtaining the phase II data and compare the corresponding predictions for the population counts of interest using the MSE. We aim for a total phase II sample size of 900 infants. Hence, the phase I sample contains 18.6% of the total data with full information being gathered on just 3.3% of the children.

The first sampling scheme we consider is one where an equal number of cases (low birth weight) and non-cases (normal birth weight) are sampled from each county, which leads to 34 low and normal birth weight infants being sampled from each county. We refer to this scheme as the “uniform sampling scheme”. If a county has fewer low birth weight infants than the required number, all low birth weight infants from that county are sampled and we correspondingly oversample the normal birth weight infants so the same number of total infants are sampled from each county. Across counties, this results in sampling between 16% and 100% of the low birth weight infants and between 2% and 100% of the normal birth weight infants in the phase I sample. In the second sampling scheme, we sample all available low birth weight infants and then sample an equal number of normal birth weight infants from each county (“case-based sampling scheme”). The number of low birth weight infants ranged from 4 to 212 and under this scheme, between 6% and 18% of the normal birth weight infants in the phase I sample were sampled across counties. Table 1 in Appendix C displays the population data, while Tables 2 through 4 display the phase I and phase II data sets, respectively.

For each sampling design, we compared three Bayesian two-phase analyses of the data: one fixed effects analysis, and two random effects analyses. In the fixed effects approach, county was included as a fixed effect in the disease model. In the random effects analyses, county was included as a random effect where we considered only non-spatial random effects in the first model, and both spatial and non-spatial random effects in the second. The results of these analyses were compared to the full data results.

Figure 2 compares the parameter estimates and 95% intervals from the three two-phase analyses with those obtained from analyzing the full population data. The parameter estimates and intervals from the two-phase analyses are quite similar. The intercept parameter is somewhat different in the random and fixed effects analyses, however this is expected since the interpretation of this parameter is different under the two approaches. In the fixed effects approach, the intercept is interpreted as the log odds of being a low birth weight infant for a non-white male born in Chatham county. However in the random effects approach, the intercept is interpreted as the log odds of being a low birth weight infant for a non-white male born in *an average county* in the research triangle area of North Carolina.

The two-phase analyses slightly overestimate the effects of race and sex in the uniform sampling scheme, and also slightly overestimate the interaction, but this is just due to the particular sample that was selected. The case-based sampling scheme analyses also slightly overestimate the effects of race and sex, and there is a corresponding underestimation of the interaction for the random effects analyses. Both sampling schemes result in very similar point estimates and credible intervals, regardless of which analysis approach is taken.

In general, the two-phase results agree well with the results from the complete data analysis, however the inferences under the two-phase and complete data analyses differ. From the random effects analyses, we would conclude that there was no statistically significant association between race and birth weight status, and that sex does not significantly modify this association. The main effect due to gender is significant, however. The reason for the lack of significant associations is the reduced power in the two-phase study as compared to the complete data (recall we have full data on just 3.3% of the sample). In this example, however, we are primarily interested in the reconstruction of the population counts, rather than the estimation of the regression coefficients.

Figure 3 compares maps of the posterior mean estimates for the spatial (U) and non-spatial (V) random effects from the random effects two-phase analyses with the complete data analysis of the North Carolina birth data set. The larger estimates tend to be concentrated in the upper right corner of the map. We see similar patterns comparing the full data analysis with the two-phase analyses indicating that the two-phase analyses are adequately capturing the residual variability. In the case of uniform sampling, we estimate τ_v to be 64.0 (95% CI: 15.4, 157.9) and τ_u to be 15.0 (95% CI: 3.4, 40.5), and the proportion of the residual variability that is spatial is 0.56. Under case-based sampling, we estimate τ_v to be 78.2 (95% CI: 17.9, 195.6) and τ_u to be 12.6 (95% CI: 2.8, 33.2), and the proportion of the residual variability that is spatial is 0.64. These estimates are consistent with the results from the full data analysis ($\tau_v = 76.7$ (95% CI: 19.5, 201.6) and $\tau_u = 48.3$ (95% CI: 8.1, 154.8)), where the proportion of the residual spatial variability that is spatial in nature is 0.36.

Table 2 shows the bias, variance and MSE from the resulting predictions of the population counts M^{yxz} , all based on the posterior distribution. Noticeable gains in efficiency were obtained when random effects were included in the two-phase analysis. Similar gains were obtained when comparing the bias, variance and MSE from predictions of the joint population probabilities $p_{yxz} = P(Y = y, X = x, Z = z)$ (Table 5, Appendix C). The improvement in estimation of the spatial model is because of the significant spatial contribution to the residual risk, as seen in Figure 3. In addition, we notice that the case-based sampling scheme produces lower variance estimates of the population counts than the uniform sampling scheme.

In Appendix D, we consider the North Carolina 2011 birth data in the context of survey sampling, where we compare the Horvitz-Thompson estimator to the two-phase NPML estimator in their ability to predict the population counts M^{yxz} across six different sampling schemes. Since the NPML approach is computationally faster than the equivalent Bayesian approach and the two approaches yield comparable results (Ross and Wakefield, 2013), we elected to use the NPML method for comparison purposes. Under all sampling designs, the

two-phase estimator had lower MSE than the Horvitz-Thompson estimator owing to much smaller variance of the estimates. Obtaining point estimates and intervals for the predicted counts requires estimating the X - Z marginal distribution using an empirical approach, and relies on asymptotics. By contrast, using a Bayesian approach, point estimates and intervals are immediate and the estimated distributions of the predicted counts are readily available from the posterior distribution. As an example, Figure 1 in Appendix C displays histograms for the predicted population counts for normal birth weight infants from each of the four race-sex categories from Franklin county from the Bayesian spatial analysis, where uniform sampling was used to obtain the phase II sample. Across all counties, there were only five instances out of 104 where the true M_{yxz} did not fall in the 95% credible interval.

Based on our predictions of the population counts, we identified Wake and Durham counties as having the highest disease burdens. Further, non-white males and females in these counties appeared to be the more vulnerable population and so resource allocation should focus on this population in these two counties. These conclusions agree with those from the full data analysis. Figures 2 and 3 in Appendix C display histograms for the predicted counts of low birth weight infants by race and sex categories in Wake and Durham counties, respectively. From the histograms, we see that our predictions agree well with the true counts, and are fairly precise.

3.2. Simulation Study

We now consider a simulated data set where spatial effects were included in the data generating process. We base the simulation on the North Carolina dataset to ensure a realistic setting. We use estimates from two models fit to the individual-level data to generate the simulated data: the posterior medians of the random effects from fitting a spatial random effects model using INLA in R, and the maximum likelihood estimates from fitting the log-linear model:

$$\log(\mu_{yxz}) = \mu + \lambda_y^Y + \lambda_x^X + \lambda_z^Z + \lambda_{yx}^{YX} + \lambda_{xz}^{XZ}.$$

We plug these estimates into (5) assuming $\lambda_{0z}^{YZ} = -(U_z + V_z)/2 = -\lambda_{1z}^{YZ}$, and we calculate the joint Y - X - Z probabilities using

$$P(Y=y, X=x, Z=z|\boldsymbol{\lambda}) = \frac{\mu_{yxz}}{\sum_u \sum_v \sum_w \mu_{uvw}}.$$

We assume the total population size is $M = 30,000$ and M^{yxz} is drawn from a multinomial distribution of size M with probabilities $P(Y=y, X=x, Z=z|\boldsymbol{\lambda})$, $y = 0, 1$, $x = 1, \dots, m_x$, $z = 1, \dots, m_z$. The phase I data are obtained by randomly sampling 5,000 infants who are then cross-classified by county and birth weight status. The phase II sample is obtained using the case-based sampling scheme described above. We considered only the case-based sampling scheme here due to the noticeable gain in efficiency over the uniform sampling scheme seen in the previous section. The population, phase I and phase II data are displayed in Tables 8-10 in Appendix E.

Table 3 compares the efficiency of three Bayesian two-phase approaches to predicting the population counts. As expected, there are substantial gains when county is included as a random effect in the modeling approach, and further gains are obtained when including spatial random effects over a model including only unstructured random effects.

Finally, we consider a simulated data set which is not spatially correlated to examine whether anything is lost by fitting a spatial model to non-spatial data. For this simulation, we permute the area labels of the NC birth data, which will reduce the spatial dependence in the data set while maintaining the race-gender and birth weight associations. The phase I and phase II data are obtained in the same manner as in the previous simulation study. Tables 11-13 in Appendix E display the full population, phase I and phase II data.

The random effects approaches were more efficient than the fixed effects approach, as expected (Table 4). The non-spatial approach was only slightly more efficient than the spatial approach. This is because any residual spatial variability that happens to occur in the dataset is picked up by the flexible spatial model, resulting in reduced bias, however, the added complexity of the model results in increased variability.

4. Discussion

In this paper, we have described a hierarchical Bayesian approach to the analysis of two-phase data. One major drawback of the Bayesian approach is that the estimation of the parameters in the exposure-stratification model is required. The motivation for this added complexity is that random effects may be included in the model in a straightforward manner, as we have described here. The inclusion of random effects allows for greater modeling flexibility than the previously described fixed effects approaches. We have described the random effects model in the context of the response model for smoothing across geographic areas, but random effects may also be included in the modeling of the x and z data to smooth the fitted cell probabilities in contingency tables. This will be a useful endeavor particularly in the situation in which the number of levels of x and z is large and the data are sparse. In Appendix F of the supplementary material, we present detail of this model and an example.

We have described two-phase studies as a means of improving estimates in the context of small area estimation. Using the 2011 birth data from the research triangle area of North Carolina, we demonstrated that the Bayesian random effects approach yielded less biased and more efficient estimates of the population counts of interest compared to standard approaches. In addition, point estimates and 95% intervals are immediate and distributions of the estimated counts are readily available. The spatial model will be particularly useful in situations in which no data are collected in particular areas but population totals are known. In this case, one can leverage spatial dependence and use information from neighboring areas to impute the unknown counts. Through a simulation study, we also showed that there is no loss of efficiency by fitting a spatial model to non-spatial data.

We have focused on simple random sampling of the phase I data, however the extension to case-control sampling at phase I is straightforward. In addition, we have only considered the case where covariates are all discrete. In the case of continuous or mixed continuous and

discrete covariates, the methods of Sinha et al. (2004) and Dunson and Xing (2009) could be applied, though this extension is not trivial. Computation is expensive for our model and we are currently investigating approaches to improving this aspect.

The two-phase design is especially useful in ecologic settings as it overcomes ecologic bias (Wakefield and Haneuse, 2008). The aggregate-level data of phase I provides high statistical power and a framework for efficiently sampling the phase II data. The individual-level data of phase II characterizes the within-area variability in exposures and confounders thereby providing a control for ecological bias. The simulation results of Wakefield and Haneuse (2008) demonstrate the benefits associated with combining these two sources of data, in terms of both bias and efficiency.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This research was supported by grants R01 CA095994 and R01 CA125081 from the National Institutes of Health, and by the Biostatistics, Epidemiology and Bioinformatic Training in Environmental Health Training Grant T32 ES015459. The authors would like to thank the joint editor, associate editor and referees for constructive comments that greatly helped in the presentation of the material.

References

- Ahn J, Mukherjee B, Gruber S, Ghosh M. Bayesian semiparametric analysis for two-phase studies of gene-environment interaction. *The Annals of Applied Statistics*. 2013; 7:543–569. [PubMed: 24587840]
- Besag J, York J, Mollié A. Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistics and Mathematics*. 1991; 43:1–59.
- Binder D. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*. 1983; 51:279–292.
- Breslow N, Cain K. Logistic regression for two-stage case-control data. *Biometrika*. 1988; 75:11–20.
- Breslow N, Chatterjee N. Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Applied Statistics*. 1999; 48:457–468.
- Breslow N, Holubkov R. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society, Series B*. 1997a; 59:447–461.
- Breslow N, Holubkov R. Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Statistics in Medicine*. 1997b; 16:103–116. [PubMed: 9004386]
- Congdon P, Lloyd P. Estimating small area diabetes prevalence in the US using the behavioral risk factor surveillance system. *Journal of Data Science*. 2010; 8:235–252.
- Davis S, Mirick D, Stevens R. Residential magnetic fields and the risk of breast cancer. *American Journal of Epidemiology*. 2002; 155:446–454. [PubMed: 11867356]
- Dellaportas P, Forster J. Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika*. 1999; 86:615–633.
- Deville J-C, Särndal C-E. Calibration estimators in survey sampling. *Journal of the American Statistical Association*. 1992; 87:376–382.
- Dunson D, Xing C. Non-parametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*. 2009; 104:1042–1051. [PubMed: 23606777]

- Fay R III, Herriot R. Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*. 1979; 74:269–277.
- Flanders W, Greenland S. Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine*. 1991; 10:739–747. [PubMed: 2068427]
- Knuiman M, Speed T. Incorporating prior knowledge into the analysis of contingency tables. *Biometrics*. 1988; 44:1061–1071. [PubMed: 3233246]
- Lumley, T. *Complex Surveys: A Guide to Analysis Using R* (Wiley Series in Survey Methodology). Wiley; 2010.
- Lumley T, Shaw P, Dai J. Connections between survey calibration estimators and semi parametric models for incomplete data. *International Statistical Review*. 2011; 79:200–220. [PubMed: 23833390]
- Martins T, Simpson D, Lindgren F, Rue H. Bayesian computing with INLA: new features. *Computational Statistics and Data Analysis*. 2012; 67:68–83.
- Neyman J. Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*. 1938; 33:101–116.
- Rabe-Hesketh S, Skrondal A. Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society, Series A*. 2006; 169:805–827.
- Rao, J. *Small Area Estimation*. John Wiley; New Jersey: 2003.
- Reilly M, Pepe M. A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*. 1995; 82:299–314.
- Robins J, Rotnitzky A, Zhao L. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*. 1994; 89:846–866.
- Ross M, Wakefield J. Bayesian inference for two-phase studies with categorical covariates. *Biometrics*. 2013; 69:469–477. [PubMed: 23607570]
- Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*. 2009; 71:319–392.
- Saegusa T, Wellner J. Weighted likelihood estimation under two-phase sampling. *Annals of Statistics*. 2013; 41:269–295. [PubMed: 24563559]
- Schill J, Jockel JH, Drescher K, Timm J. Logistic analysis in case-control studies under validation sampling. *Biometrika*. 1993; 84:57–71.
- Scott A, Wild C. Fitting logistic regression in stratified case-control studies. *Biometrics*. 1991; 47:497–510.
- Scott A, Wild C. Fitting regression models to case-control data by maximum likelihood. *Biometrika*. 1997; 51:54–71.
- Sharpe C, Collet J-P, McNutt M, Belzile E, Boivin J-F, Hanley J. Nested case-control study of the effects of non-steroidal anti-inflammatory drugs on breast cancer risk and stage. *British Journal of Cancer*. 2000; 83:112–120. [PubMed: 10883678]
- Sinha S, Mukherjee B, Ghosh M. Bayesian semiparametric modeling for matched case-control studies with multiple disease states. *Biometrics*. 2004; 60:41–49. [PubMed: 15032772]
- Skinner, C. Domain means, regression and multivariate analysis.. In: Skinner, C.; Holt, D.; Smith, T., editors. *Analysis of Complex Surveys*. Wiley; Chichester: 1989. p. 59-87.
- Sørbye SH, Rue H. Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*. 2014; 8:39–51.
- Wakefield J, Haneuse S. Overcoming ecological bias using the two-phase study design. *American Journal of Epidemiology*. 2008; 167:908–916. [PubMed: 18270370]
- White E. A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*. 1982; 115:119–128. [PubMed: 7055123]

Fig. 1.

Map of the proportion of low birth weight infants born in the research triangle area of North Carolina in 2011.

Fig. 2. Parameter estimates and 95% credible intervals from three Bayesian two-phase approaches to analyzing the 2011 North Carolina birth data set, where the phase II data is sampled using two different sampling schemes: uniform sampling, and case-based sampling.

Fig. 3.

Posterior mean estimates for the spatial (U) and non-spatial (V) random effects from three analyses of the 2011 North Carolina birth data set: (a) full population data (top row), (b) Bayesian two-phase analysis with phase II data sampled using uniform sampling scheme (middle row), and (c) Bayesian two-phase analysis with phase II data sampled using case-based sampling scheme (bottom row).

Table 1

Two-phase study design: the observed data are $N_{y \cdot z}$ at phase I, and $n_{y \cdot z}$ and n_{yxz} at phase II; the N_{yxz} and M_{yxz} entries are unobserved in a two-phase study.

Population Data							
	$Z = 1$		$Z = 2$		...	$Z = m_z$	
	$Y = 0$	$Y = 1$	$Y = 0$	$Y = 1$	...	$Y = 0$	$Y = 1$
$X = 1$	M_{011}	M_{111}	M_{012}	M_{112}	...	M_{01m_z}	M_{11m_z}
$X = 2$	M_{021}	M_{121}	M_{022}	M_{122}	...	M_{02m_z}	M_{12m_z}
...	...	...	...	...	...	...	...
$X = m_x$	M_{0m_x1}	M_{1m_x1}	M_{0m_x2}	M_{1m_x2}	...	$M_{0m_xm_z}$	$M_{1m_xm_z}$
	$M_{0 \cdot 1}$	$M_{1 \cdot 1}$	$M_{0 \cdot 2}$	$M_{1 \cdot 2}$	...	$M_{0 \cdot m_z}$	$M_{1 \cdot m_z}$
Phase I Data							
	$Z = 1$		$Z = 2$		...	$Z = m_z$	
	$Y = 0$	$Y = 1$	$Y = 0$	$Y = 1$	...	$Y = 0$	$Y = 1$
$X = 1$	N_{011}	N_{111}	N_{012}	N_{112}	...	N_{01m_z}	N_{11m_z}
$X = 2$	N_{021}	N_{121}	N_{022}	N_{122}	...	N_{02m_z}	N_{12m_z}
...	...	...	...	...	...	...	...
$X = m_x$	N_{0m_x1}	N_{1m_x1}	N_{0m_x2}	N_{1m_x2}	...	$N_{0m_xm_z}$	$N_{1m_xm_z}$
	$N_{0 \cdot 1}$	$N_{1 \cdot 1}$	$N_{0 \cdot 2}$	$N_{1 \cdot 2}$	...	$N_{0 \cdot m_z}$	$N_{1 \cdot m_z}$
Phase II Data							
	$Z = 1$		$Z = 2$		...	$Z = m_z$	
	$Y = 0$	$Y = 1$	$Y = 0$	$Y = 1$	...	$Y = 0$	$Y = 1$
$X = 1$	n_{011}	n_{111}	n_{012}	n_{112}	...	n_{01m_z}	n_{11m_z}
$X = 2$	n_{021}	n_{121}	n_{022}	n_{122}	...	n_{02m_z}	n_{12m_z}
...	...	...	...	...	...	...	...
$X = m_x$	n_{0m_x1}	n_{1m_x0}	n_{0m_x1}	n_{1m_x1}	...	$n_{0m_xm_z}$	$n_{1m_xm_z}$
	$n_{0 \cdot 1}$	$n_{1 \cdot 1}$	$n_{0 \cdot 2}$	$n_{1 \cdot 2}$	...	$n_{0 \cdot m_z}$	$n_{1 \cdot m_z}$

Table 2

Comparing bias, variance and MSE for estimating $M^{y_{xz}}$ from three Bayesian two-phase analyses of the NC 2011 birth data set.

Uniform Sampling Scheme			
	Bias ($\times 10^2$)	Variance ($\times 10^5$)	MSE ($\times 10^5$)
Fixed	7.6	22.8	28.6
Non-spatial	4.2	13.5	15.2
Spatial	4.0	12.7	14.3

Case-based Sampling Scheme			
	Bias ($\times 10^2$)	Variance ($\times 10^5$)	MSE ($\times 10^5$)
Fixed	8.5	6.6	13.8
Non-spatial	7.0	5.4	10.3
Spatial	7.0	5.2	10.1

Table 3

Comparing bias, variance and MSE for estimating $M^{y_{xz}}$ from three Bayesian two-phase analyses of a spatially simulated dataset, using case-based sampling to obtain the phase II sample.

	Bias ($\times 10^2$)	Variance ($\times 10^5$)	MSE ($\times 10^5$)
Fixed	9.9	8.3	16.2
Non-spatial	7.8	7.0	13.1
Spatial	7.3	6.9	12.2

Table 4

Comparing bias, variance and MSE for estimating $M^{y_{xz}}$ from three Bayesian two-phase analyses of a non-spatially simulated dataset, using case-based sampling to obtain the phase II sample.

	Bias ($\times 10^2$)	Variance ($\times 10^5$)	MSE ($\times 10^5$)
Fixed	8.3	7.4	14.3
Non-spatial	7.6	5.8	11.6
Spatial	7.5	6.1	11.8