HHS Public Access

Author manuscript

Spat Spatiotemporal Epidemiol. Author manuscript; available in PMC 2015 April 01.

Published in final edited form as:

Spat Spatiotemporal Epidemiol. 2014 October ; 11: 33–43. doi:10.1016/j.sste.2014.07.002.

The use of sampling weights in Bayesian hierarchical models for small area estimation

Cici Chen,
Department of Biostatistics, Brown University, USA

Jon Wakefield, and
Departments of Statistics and Biostatistics, University of Washington, USA

Thomas Lumely
Department of Statistics, University of Auckland, New Zealand

Cici Chen: cici_bauer@brown.edu; Jon Wakefield: jonno@uw.edu

Abstract

Hierarchical modeling has been used extensively for small area estimation. However, design weights that are required to reflect complex surveys are rarely considered in these models. We develop computationally efficient, Bayesian spatial smoothing models that acknowledge the design weights. Computation is carried out using the integrated nested Laplace approximation, which is fast. A simulation study is presented that considers the effects of non-response and non-random selection of individuals. We examine the impact of ignoring the design weights and the benefits of spatial smoothing. The results show that, when compared with standard approaches, mean squared error can be greatly reduced with the proposed models. Bias reduction occurs through the inclusion of the design weights, with variance reduction being achieved through hierarchical smoothing. We analyze data from the Washington State 2006 Behavioral Risk Factor Surveillance System. The models are easily and quickly fitted within the R environment, using existing packages.

Keywords

Bayesian methods; integrated nested Laplace approximation; sample surveys; spatial statistics

1. Introduction

In this paper we consider spatial models for small area estimation (SAE). SAE is an important endeavor since many agencies require estimates of health, education and environmental measures in order to plan and allocate resources and target interventions. The data upon which SAE is based are often gathered via complex designs.

Supporting information
is available for this article. S1: effective sample size when the variance is undefined; S2: Example code for WinBUGS and INLA; S3: Comparison between MCMC and INLA; S4: Sensitivity to priors on the precision parameters; S5: Simulation results from models using unit-level data; S6: Additional summaries of the simulation studies; S7: Simulation study with non-response and selection bias; S8: Squared bias, variance and MSE versus sample size; S9: Comparison of point estimates and variance estimates with and without variance adjustment; S10: Additional summaries for the BRFSS example.

Standard model-based approaches to the analysis often ignore the sampling mechanism and are therefore subject to potentially large biases. Adjusting for the sampling scheme by including in the model the design variables upon which sampling was based (and which are associated with the outcome of interest) is often not possible, because the required variables are unavailable, or the required model would be overly complex (Gelman, 2007). In this paper we consider the situation in which it is not possible to model the sampling scheme. Weighted design-based approaches provide a common approach to bias removal but the resultant estimators can be highly variable for areas in which only small sample sizes are collected. Hierarchical models provide a method to reduce the variance, with Fay and Herriot (1979) providing an early example and, notably, acknowledging the sampling scheme. Since this influential paper many hierarchical modeling approaches have been suggested, see Rao (2003) for a comprehensive summary of the literature and Pfeffermann (2013) for a more recent account.

In terms of spatial smoothing techniques, a number of authors allow for spatial correlation between areas, see for example Singh et al. (2005), Pratesi and Salvati (2008) and Pereira and Coelho (2010). These models are subject to bias, however, since they do not adjust for the sampling scheme. Pseudo-likelihood (Skinner, 1989; Pfeffermann et al., 1998) has been used within a hierarchical modeling framework with the scaling of the weights being a major issue (Potthoff et al., 1992; Longford, 1996; Asparouhov, 2006; Rabe-Hesketh and Skrondal, 2006). Congdon and Lloyd (2010) use such an approach and introduce residual spatial random effects. In this paper we describe a range of models that can acknowledge the sampling scheme and allow spatial smoothing. We describe a new approach based on the concept of “effective sample size” and “effective number of cases”. A related Bayesian model has recently been suggested by Ghitza and Gelman (2013), while quite a different approach, based on a penalized spline model, is described in Zheng and Little (2003) and Zheng and Little (2005). A key feature of the models we describe is that computation is fast and can be carried out using existing packages within the R computing environment.

The outline of this paper is as follows. We begin with a motivating example that concerns diabetes prevalence in the Behavioral Risk Factor Surveillance System (BRFSS) in Section 2. In Section 3 we describe hierarchical spatial and non-spatial models, which we then compare with various approaches in Section 4, via an extensive simulation study. We return to the BRFSS data in Section 5 and conclude the paper with a discussion in Section 6. The supplementary materials contain more technical details and additional supporting information.

2. Motivating Example

The BRFSS is an annual telephone health survey conducted by the Centers for Disease Control and Prevention (CDC) that tracks health conditions and risk behaviors in the United States and its territories since 1984. In the BRFSS survey, interviewees (who must be 18 years or older) are asked a series of questions on their health behaviors and provide general demographic information, such as age, race, gender and the zip code in which they live. Here we focus on the survey conducted in Washington State in 2006 (http://www.doh.wa.gov/brfss), and on the question, “Have you ever been told you have

diabetes?", with interviewees responding with the binary response "Yes"/"No". Our objective is to estimate the number of 18 or over individuals with diabetes, by zip code, in Washington State. The CDC currently publishes coarser, county-level prevalence estimates (http://apps.nccd.cdc.gov/DDTSTRS/), using the model of Malec et al. (1997).

In 2006, the survey used land-lines only, and a disproportionate stratified random sample scheme with stratification by county and "phone likelihood". Under this scheme in each county, based on previous surveys, blocks of 100 telephone numbers are classified into strata that are either "likely" or "unlikely" to yield residential numbers. Telephone numbers in the "likely" strata are sampled at a higher rate than their "unlikely" counterparts. Once a number is reached the number of eligible adults (aged 18 or over) is determined, and one of these is randomly selected for interview. The sample weight, Sample Wt, is calculated as the product of four terms

$$\text{Sample}-\text{Wt}=\text{Strat}-\text{Wt}\times\frac{1}{\text{No}-\text{Telephones}}\times\text{No}-\text{Adults}\times\text{Post}-\text{Strat}-\text{Wt} \quad (1)$$

where Strat-Wt is the inverse probability of a "likely" or "unlikely" stratum being selected in a particular county, No-Telephones represents the number of residential telephones in the respondent's household, No-Adults is the number of adults in the household, and Post-Strat-Wt is the post-stratification correction factor, with the strata defined by the 7 age groups 18–24, 25–34, 35–44, 45–54, 55–64, 65–74, 75+ and gender. The other source of data we use are population estimates for 2006.

Basic summary statistics, across 498 zip codes, are presented in Table 1. There is large variability in the population, sample and number of diabetes cases, across zip codes. The left panel of Figure 1 gives a histogram of the survey sample sizes. About 20% of the areas have sample sizes of 9 or less, so that the diabetes prevalence estimates are highly unstable in these areas. The sample weights which have a large range of 0.82 to 4991; the coefficient of variation (CV) of the weights is 1.11 so that the inefficiency of using the sample weights under the assumption that the unweighted mean is unbiased is about 55%, as calculated by $100 \times CV^2/(CV^2 + 1)$ (Korn and Graubard 1999, Section 4.4). In Section 5 we will make inference for the proportion with diabetes at the zip code level using a hierarchical model that acknowledges the sampling scheme and leverages spatial smoothing.

3. Sample Weighted Bayesian Hierarchical Models

Hierarchical models have been used extensively for SAE. In this section, we first review some commonly used three-stage hierarchical models, including a spatial model, without considering the sampling weights. We refer to the resultant estimators as *unadjusted*; these estimators can be seriously biased in the event of non-random selection of individuals or non-response. Subsequently, we will describe our approach to incorporating the sampling weights for binary data using the same set of hierarchical models. Estimates that use the weights will be referred to as *adjusted*.

We begin by introducing notation. Let Y_{ik} denote the binary variable indicating if the k-th individual from area i has the outcome of interest ($Y_{ik} = 1$) or not ($Y_{ik} = 0$). Common small

area characteristics of interest are the true total count, $T_i = \sum_{k=1}^{N_i} Y_{ik}$, or the true proportion, $P_i = T_i/N_i$, where N_i is the total population in area i, $i = 1; \ldots; I$. We let s_i denote the set of individuals who are sampled from area i with y_{ik} the observed sample for $k \in s_i$ with $|s_i| = m_i$.

3.1. Conventional Three-Stage Bayesian Hierarchical Models

A conventional three-stage Bayesian hierarchical model for a binary outcome uses a binomial distribution at stage one

$$y_i | P_i \sim \text{Binomial}\,(m_i, P_i), \quad i = 1, \ldots, I, \quad (2)$$

where $y_i = \sum_{k \in s_i} y_{ik}$, and m_i is the sample size, for area i. At the second stage, we model between-area variation in P_i using a random effects model. Finally, the unknown hyperparameters at the second stage are assigned hyperprior distributions at the third stage. We consider two possible random effects models to account for between-area variation at the second stage.

Model 1: *Independent Random Effects Model*

In this model we assume that the log odds of the area proportion P_i are drawn, independently, from a normal distribution:

$$\log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \in_i, \quad \in_i | \sigma^2_{\in} \sim_{iid} \mathrm{N}(0, \sigma^2_{\in}), \quad (3)$$

where β_0 is the intercept and the random effects ε_i capture between-area variability in the residual log odds. This model was used by MacGibbon and Tomberlin (1989) in an SAE context.

Model 2: *Independent and Spatial Random Effects Model*

In general, we might expect that areas which are close to each other will share more similarities than areas that are far away, and we would like to exploit this information in order to provide more reliable estimates in each area. We adopt the spatial model introduced by Besag et al. (1991) that includes both non-spatial and spatial random effects and assigns the spatial random effects an intrinsic conditional autoregressive (ICAR) prior

$$\log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \in_i + U_i, \in_i | \sigma^2_{\in} \sim_{iid} \mathrm{N}(0, \sigma^2_{\in}), \quad (4)$$

$$U_i | U_j, j \in \mathrm{ne}(i), \sigma^2_u \sim \mathrm{N}\left(\overline{U}_i, \frac{\sigma^2_u}{n_i}\right), \quad (5)$$

where ne(i) is the set of neighbors of area i, n_i is the number of such neighbors and $\overline{U}_i$, is the mean of the neighboring spatial random effects. In this model the nature of the spatial dependency is defined by the neighborhood structure. For example, a common approach,

that we adopt, defines areas *i* and *i′* to be neighbors if they share a common boundary. We require priors for β_0 and the random effects variances. A normal hyperprior is typically assumed for the former, and inverse gamma distributions for the latter; we follow the prescription described in Wakefield (2009) in which the prior specifications are related to the sizes of the residual odds ratios.

3.2. Bayesian Hierarchical Models with Complex Survey Weights

3.2.1. A Definition of Effective Sample Size—Our approach to acknowledging the design is to construct a binomial likeli hood that is based on the direct estimate of P_i and its associated variance. Fay and Herriot (1979) used a direct estimate within a hierarchical model, and our approach is in the same spirit. The direct estimator is

$$\hat{P}_i = \frac{\sum_{k \in s_i} w_{ik} y_{ik}}{\sum_{k \in s_i} w_{ik}}, \quad i = 1, \ldots, I, \quad (6)$$

where w_{ik} is a weight which, in the simplest case, is given by $w_{ik} = \pi_{ik}^{-1}$ where π_{ik} is the probability that the *k*-th person in the *i*-th area is sampled. We let $\hat{V}_i = v\hat{a}r(\hat{P}_i)$ be the estimated variance of $\hat{P}_i$, see for example, Särndal et al. (1992). The estimator (6) is design unbiased, but Bayesian modeling requires more than bias correction, we need a full probability model for the data. Viewing the "data" as $\left\{\hat{P}_i, \hat{V}_i\right\}$, perhaps the first candidate for a likelihood would be the asymptotic normal distribution $\mathrm{N}(\hat{P}_i, \hat{V}_i)$. This distribution will be accurate for large samples, but in small samples will be inadequate, with one reason being that the range is not restricted to [0,1]. Our proposal is based on a binomial approximation to the distribution of the *effective number of cases*, which we define shortly. We begin by defining an *effective sample size* m_i^*. In a simple random sample, the estimated variance would be $\hat{P}_i(1 - \hat{P})/m_i$, where $\hat{P}_i$ is the estimator in (6). The effective sample size m_i^* is then obtained by solving $\hat{V}_i = \hat{P}_i(1 - \hat{P}_i)/m_i^*$ to give

$$m_i^* = \frac{\hat{P}_i(1 - \hat{P}_i)}{\hat{V}_i}. \quad (7)$$

Using the effective sample size rather than the actual sample size acknowledges the variable information that each individual supplies under complex sampling. The precision of an estimate from a complex sample can be higher than for a simple random sample, because of the better use of population data, obtained via stratification and post-stratification. However, the precision can also be lower, either because of correlation within clusters (which reduces information), or because the design was optimized for estimating a specific quantity which is not well correlated with the quantity of interest (see the right panel of Figure 1 for the effective versus observed sample sizes in the BRFSS example). The ratio of the effective sample to the actual sample size is the reciprocal of Kish's "design effect" (Kish 1995), a standard summary of the efficiency of a sampling design.

Our likelihood derivation goes one step further by defining an effective number of cases y_i^* as the product of the effective sample size m_i^* and the weighted proportion $\hat{P}_i$ estimated via (6):

$$y_i^* = m_i^* \times \hat{P}_i. \quad (8)$$

In the supplementary materials we describe how we overcome the difficulties in using (7) when $m_i \neq N_i$ and $\hat{P}_i = 0/1$ or $\hat{V}_i = 0$. In the case of $m_i = N_i$ then P_i is known and the likelihood is a point mass at this value. Since the likelihood corresponds to a point there will be no shrinkage from the prior (as is desirable).

3.2.2. Sample Weighted Bayesian Hierarchical Models—To incorporate sample weights in a three-stage hierarchical model we define the first stage likelihood as

$$y_i^* | P_i \sim \text{Binomial}(m_i^*, P_i), \quad i = 1, \ldots, I,$$

where y_i^* and m_i^* are as defined in Section 3.2.1. By construction, the sampling distribution of the commonly used estimator y_i^*/m_i^* is unbiased for the population prevalence P_i (under the same conditions as the estimator (6) is unbiased) and the reciprocal of the Fisher information is equal to the design-based variance estimate, giving an appropriate indication of precision. As a binomial distribution, it also respects the [0; 1] bounds on P_i.

A related approach was suggested by Raghunathan et al. (2007), in the context of combining data from multiple sources. For estimating a proportion, they assume the model $\sin^{-1}\hat{P}_i^{1/2} | P_i \sim_{ind} \text{N}\left(\sin^{-1}P_i^{1/2}, \frac{1}{4m_i^*}\right)$, where m_i^* is again the effective sample size. The arcsine square root transformation stabilizes the variance but may be deficient for areas with small sample sizes. In addition, the model does not constrain the target proportions of interest to lie within [0, 1].

In terms of inference, for concreteness we focus on predicting the total count T_i for small area i. The point estimate of the population count is

$$\hat{T}_i = \hat{P}_i \times N_i, \quad (9)$$

where $\hat{P}_i$ is the direct estimator (6) and the variance is

$$\hat{var}(\hat{T}_i) = \hat{V}_i \times N_i^2. \quad (10)$$

In a Bayesian analysis one may summarize the posterior distribution for T_i using quantiles. If a point estimate is required then it is given by (9) with $\hat{P}_i$ replaced by the posterior mean or median. The posterior variance var (T_i|y) is given by (10) with $\hat{V}_i$ replaced by the posterior variance, var(P_i|y).

3.3. Implementation

The usual implementation of Bayesian hierarchical models is via Markov chain Monte Carlo (MCMC). However, the large computational burden can impede the application of Bayesian hierarchical models in practice, and there is a need for convergence assessment which makes the approach difficult to automate. For these reasons, we employ the integrated nested Laplace approximation (INLA) which has recently been proposed as a computationally convenient alternative to MCMC (Rue et al., 2009). This method carries out fully Bayesian inference by combining Laplace approximations and numerical integration in a very efficient manner, see Rue et al. (2009) for details. Fong et al. (2010) provide a comprehensive review of implementing Bayesian GLMMs using INLA, including a comparison with MCMC. These authors illustrate the accuracy of INLA in a range of examples, and this accuracy has now been borne out in multiple publications in a variety of application areas, see for example Rue et al. (2009); Paul et al. (2010); Schrodle et al. (2011) and Riebler et al. (2012). For the approach that we advocate, based on the effective sample size, we use the survey package (Lumley, 2010) (to obtain the required variance estimate) in combination with the R implementation of INLA. Example code can be found at http://faculty.washington.edu/jonno/cv.html. The supplementary materials contain example INLA and WinBUGS code, along with comparisons for a typical simulation.

INLA is particularly useful for simulation studies, as we demonstrate in the next section. In our simulations the gain in speed is substantial, for example, on an Intel Quad CPU Q6700 with 2.66 GHz Process and 4.00 GB memory computer using Windows 7 it takes 1439s to run 205,000 iterations of MCMC while in the R implementation of INLA, the total running time is 5.2s.

4. Simulation Study

In this section we report the results of simulation studies, under a variety of scenarios, to evaluate the performance of:

Direct Estimates: using either the observed counts y_i and sample sizes m_i (Unadjusted) or the design-based estimator defined in (6) along with the appropriate variance estimate $\hat{V}_i$ (Adjusted).

Independent Normal Random Effects: The hierarchical model with independent normal random effects given by (3) along with a binomial first stage model based on (y_i, m_i) (Unadjusted) or (y_i^*, m_i^*) (Adjusted).

Spatial Normal Random Effects: The hierarchical model with both independent normal random effects and spatial ICAR random effects given by (5) along with a binomial first stage model based on (y_i, m_i) (Unadjusted) or (y_i^*, m_i^*) (Adjusted).

At the third stage of the hierarchical models, we assume an improper uniform prior for β_0, and assign Gamma (0.5, 0.008) distributions to the precision parameters σ_v^{-2} and σ_u^{-2}. This prior gives a 95% range for the residuals odds of (0.5,2.0) (Wakefield, 2009). In both the simulation study and the BRFSS example there is very little sensitivity to the priors on the

variance components, since the number of areas is large (an example is given in the supplementary materials).

In the paper we compute and report three statistics to evaluate the estimates: the squared bias, the variance and the mean squared error (MSE). In the supplementary materials we give more extensive summaries. Let S denote the total number of simulations, and T_i the "true" diabetes count in area i (which for each scenario is kept constant across simulations). The summary statistics are calculated as

$$\text{Bias}=\frac{1}{I}\sum_{i=1}^{I}(\overline{\hat{T}}_i - T_i), \text{ where } \overline{\hat{T}}_i=\frac{1}{S}\sum_{s=1}^{S}\hat{T}_i^{(s)},$$
$$\text{Variance}=\frac{1}{I}\sum_{i=1}^{I}\left(\frac{1}{S-1}\sum_{s=1}^{S}(\hat{T}_i^{(s)} - \overline{\hat{T}}_i)^2\right),$$
$$\text{MSE}=\text{Bias}^2+\text{Variance}.$$

In both the simulations and the BRFSS example we use (9) as the point estimate, using the posterior median for the Bayesian approaches. So far as the variance is concerned we can use $\hat{V}_i$ for the direct estimates and the posterior variance var($P_i|y$) for the Bayesian approaches. Estimators with small MSE are considered superior, although amongst estimators with comparable MSE those with lower bias are preferred because they lead to interval estimates with improved calibration. We examine two types of bias that are commonly seen in complex surveys, selection bias and non-response bias. The former bias occurs when the variables upon which selection are based are associated with the outcome. Non-response bias occurs when sub-populations respond to the survey at different rates.

In all simulation studies, we take the geography of Washington State at the zip code level and focus on the prediction of a total count for each zip code, with the outcome of interest being labeled "Diabetes" since we base the prevalences on this variable. The design weights are the inverse of the sampling probabilities, and the post-stratification weights are based on age-gender population data for Washington State. The sample size m_i is chosen to be the actual number of individuals who responded in the Washington 2006 BRFSS survey. For those zip codes that result in $\hat{P}_i=0/1$ or $\hat{V}_i=0$ we use the technique described in the supplementary materials to obtain a variance estimate, and therefore an effective sample size. There were 98 such areas for the 2006 BRFSS application. For the simulation study, the number of such areas varies across simulations.

4.1. Non-Response Bias

We examine five scenarios with different response probabilities. We consider stratified random sampling with stratification based on gender and three age bands, so that there are J = 6 groups. In all five scenarios, individuals are selected randomly within each area. However, a selected individual in group j and area i responds to the survey with probability q_{ij}.

Scenario 1: the ideal situation where every selected individual responds to the survey. The prevalences of diabetes p_{ij} we use across the six gender-age groups j and in area i are given

in Table 2. The median values are based on National Surveillance Data from the CDC (http://www.cdc.gov/diabetes/statistics/prev/national/menuage.htm).

Scenario 2: a more realistic sampling situation in which not every selected individual responds to the survey, with the response rates being different for each group *j*, but constant across areas. Table 3 gives the response rates; the groups with older people have slightly higher response rates, which is generally the case.

Scenario 3: we allow the response rates for each group to vary between areas via

$$\text{logit}(q_{ij}) = \text{logit}(q_j) + b \times \varepsilon_i, \quad i = 1, \ldots, I,$$

where $\varepsilon_i \sim_{iid} \text{N}(0, 1)$. The response rates in this scenario are controlled (via *b)* in such a way that the median response rates for each group correspond to those in scenario 2, with 95% ranges given in Table 3.

Scenario 4: the prevalence rates include spatial dependency induced by adding a spatially correlated area-level covariate

$$\text{logit}(p_{ij}) = \text{logit}(p_j) + b \times x_i, \quad i = 1, \ldots, I.$$

The spatial covariate x_i is simulated from a zero mean, unit variance ICAR model, using the method described in Rue and Held (2005). We choose *b* to give the ranges in Table 2. The purpose of this scenario is to investigate the effect of spatial dependency in the prevalence when the underlying cause of the dependency is unobserved. In this case, the spatial random effects are being used as a surrogate for the unmeasured covariates x_i.

Scenario 5: the response rates for each group vary between areas by adding a spatial component

$$\text{logit}(q_{ij}) = \text{logit}(q_j) + b \times x_i, \quad i = 1, \ldots, I,$$

where x_i again is simulated from a zero mean, unit variance ICAR model with *b* chosen to give a range for the response rates q_{ij} as given in Table 3.

For scenarios 1, 2, 3 and 5 the diabetes status of each individual in the total population is simulated using the prevalence rates described in scenario 1. These population outcomes are then considered the “truth” in these scenarios 5. In scenario 4, the true prevalence rates exhibit spatial dependency and so a second population is generated.

4.2. Selection Bias

To investigate the effects of selection bias, let Z_{ijk} represent a binary design variable that is used to dictate whether the *k*-th individual in group *j* and area *i* will be sampled. We use the population simulated from scenario 1 in the simulation study for non-response, and assume the model

$$Pr(Z_{ijk}=1|y_{ijk}=1)=s,\ \ Pr(Z_{ijk}=1|y_{ijk}=0)=0.1.$$

If $s \neq 0.1$ a correlation is induced between the design and outcome variables Z and Y. We examine the extent of the correlation by assigning s values of 0.1 (no selection bias), 0.3, 0.5 and 0.8. Let r_i denote the percentage of the population with $Z = 1$ in area i. We set the sample size $m_i = m \times r_i/\sum r_i$, and within each area take half of the samples with $Z = 1$ and the remainder with $Z = 0$. Hence, we are carrying out stratified random sampling with allocation proportional to r_i. Oversampling individuals with certain characteristics is a common technique in surveys. For the simulations, the information on the variable Z is used only when conducting the survey (and when calculating the sample weights) and is considered unavailable at the time of analysis. In the supplementary material we report additional results in which the data on the design variables are available, and we fit various models using these data.

For those analyses in which Z_i is unavailable at the time of analysis, the pairwise sampling probabilities $\pi_{ikk'}$ are also unavailable, and it is necessary to use an approximation to the Horvitz-Thompson variance formula to obtain $\hat{V}_i$. In public-use survey data, where sampling fractions are typically small, it is standard to approximate the variance by pretending the sample is taken with replacement; this approximation is typically slightly conservative.

4.3. Results

The simulation results for non-response bias are presented in Table 4. In scenarios 1 and 4, in which everyone responds to the survey, the unadjusted direct estimator is approximately unbiased by construction and therefore has the smallest squared bias. Nothing is gained by adjustment and there is a slight increase in the bias. However, in scenarios 2, 3, and 5 when non-response exists, the unadjusted estimator is highly biased. This bias can be reduced by post-stratification, as seen in the smaller squared bias in the adjusted direct estimates; this is the main purpose of post-stratification in large surveys. The reduction in bias carries over to the hierarchical estimators based on adjusted data; these estimators exhibit bias due to the shrinkage. Moving from scenario 2 to 3, the results show an increase in both the bias and variance under all models, due to the increased variation in the response rates in the simulated data. However, the effective number of cases (adjusted) approach provides a substantial reduction in MSE as compared to the direct estimates.

In scenarios 4 and 5, we impose spatial dependency in the data (in the prevalence rates and response rates, respectively) but pretend the source of the dependency is unknown to us. The spatial models produce estimates with the smallest MSE when compared to the other estimation methods. The spatial random effects can serve as a surrogate for the variables responsible for the dependency in the underlying prevalence. We note that in the other simulation scenarios where no spatial dependency is imposed, the spatial model still gives the minimum MSE. We speculate that less bias is imposed by the local smoothing of the Bayesian spatial model, compared to the global smoothing of the independent random effects approach. We would not expect this phenomenon to always occur, however.

Results from the selection bias simulations are summarized in Table 5. It is clear that ignoring the design variable used in the sampling procedure results in substantial bias, with the bias being higher if the correlation between the design variable and the outcome variable is stronger, as expected. If the design variable is available to the analyst, then the results in the supplementary material show there are substantial gains in reducing MSE. Again the spatial model performs well, with the minimum MSE for all situations apart from s = 0.8.

The supplementary materials include an examination of the behavior of squared bias, variance and MSE as a function of sample size.

5. Motivating Example Revisited

We apply the unadjusted and adjusted Bayesian hierarchical models we developed in Section 3 to the Washington State 2006 BRFSS data introduced in Section 2. Sampling weights are taken to be the final weight used in the BRFSS survey, as summarized in equation (1). We emphasize that the design variables are unavailable so that the weights are the only available means for adjusting for selection bias. For those nine areas with only $m_i = 1$ observation, the effective sample size and effective number of observation are based on the variance estimation procedure described in the supplementary materials. The right panel of Figure 1 plots the resultant effective sample sizes versus the observed sample sizes. We highlight that, for some areas, the effective sample size is larger than the observed sample size; we will provide further explanation shortly.

Figure 2 presents the boxplots of logit-transformed estimated diabetes prevalence by zip code under different approaches. For the direct unadjusted approach, we employ the empirical logit transformation, i.e. $\log[(y_i + 0.5)/(m_i + 0.5)]$. There is a large amount of variation in the unadjusted direct estimates due to large sampling variability, with the variability of the adjusted estimates being only slightly reduced. The variability of the estimates is significantly reduced under the hierarchical models. The location of the adjusted estimates is reduced relative to the unadjusted estimates in all models. The spatial random effects model gives estimates with slightly increased variation compared to the independent random effects model. In the simulation studies, the Bayesian spatial random effects models perform well in a range of circumstances. Further, we only see a small loss of performance when the weights are incorporated. Hence, we report inference based on the adjusted Bayesian spatial model. In a more comprehensive analysis we would identify areas with relatively large samples and then reduce the number of samples in this area. We would then compare the predictive performance of the candidate models in these areas, as compared to the “gold standard” of the full sample (Srebotnjak et al., 2010).

For the adjusted Bayesian spatial model 11% of the total residual variation is spatial. The standard deviation of the non-spatial and spatial random effects are 0.18 and 0.12 and 95% intervals for the non-spatial and spatial residual odds of diabetes are (0.70, 1.42) and (0.79, 1.27), respectively. Hence, for the BRFSS diabetes outcome there is significant excess-binomial variation, with the major component being non-spatial. Figure 3 gives the map that we would report, based on the adjusted spatial model. There are higher diabetes counts around the Puget Sound area (the channel running north-south with many small, highly

populated, zip codes to the east) and the central south area. These areas correspond to King, Snohomish and Spokane counties and the Yakima valley, which are the most populated counties in Washington State. The supplementary materials include a map of the estimated uncertainty (posterior standard deviation) of the predicted diabetes counts using the adjusted spatial model.

Figure 4 shows the standardized differences in the total count estimates

$$\frac{\hat{T}_i^{\text{spat,adj}} - \hat{T}_i^{\text{direct,adj}}}{\hat{T}_i^{\text{spat,adj}}}$$

for the adjusted spatial model and the adjusted direct approach. We see lots of differences, with a magnitude that is important; the totals in Figure 3 have a 10–90% range of (30, 1591). There is clear spatial structure in the differences, as we might expect. The supplementary materials includes the analogous plot for the adjusted and unadjusted spatial models.

To illustrate the effect of our proposed method we now provide some examples that compare the observed sample size and the effective sample sizes. For zip code areas with moderate sample size and somewhat balanced samples in each age/gender group, the effective sample sizes and effective number of cases defined in our approach should be close to the raw data values, as shown in the right panel of Figure 1. We describe two circumstances when the effective and observed samples sizes can be quite different. The first circumstance is when the usual design-weighted estimate is 0 (i.e., 0 observed cases). For example, in zip code 98008 the observed sample size is 41 with 0 cases. The sample size for each age/gender group is: Female, 18–44, 12; Female, 45–74, 9; Female, 75+, 3; Male, 18–44, 8; Male, 45–74, 7; and Male and 75+, 2. Using the Bayes smoothing method based on a beta-binomial model, as described in the supplementary materials, the estimated prevalence is modified to 0.022 which is significantly different from the raw estimate of 0 and far more reasonable. In this case, the traditional design-based estimation approach therefore fails even when the sample size is moderate. A second circumstance in which the effective and observed sample sizes can differ is when the samples are highly unbalanced (i.e., most of the samples are from a particular age/gender group, with no/small samples from other groups). As an example we take zip code 98294. The observed sample size is 16 with 1 case. However, of the total sample size of 16, 4 are Female 18–44, 10 are Female 45–74, and 2 are Male with 45–74 with no individuals from the remaining age/gender groups. After the adjustment under our proposed method, the effective sample size for this area is estimated as 42.2 (the final weights for the 14 individuals range between 78 and 495) with an effective number of cases of 0.9. The observed ratio of effective number of cases to sample size gives a naive estimate of the prevalence as 0.02, which is quite different to the raw estimate of 0.06.

6. Conclusion

In this paper we have described a pragmatic approach to SAE that allows spatial smoothing, and incorporates sample weights to acknowledge the design. We have assumed that the

design variables are unavailable so that directly modeling of the sampling mechanism is not possible. By using the sample weights to adjust the data before estimation we separate the design-based survey computations and the model-based Bayesian shrinkage, allowing both components to be modified as the situation requires. The simulation study demonstrates the potential of the approach for bias reduction relative to an approach that ignores the weights and variance reduction relative to a non-hierarchical approach, under a number of difference scenarios. We have utilized INLA for computation due to its fast computation time and convenient R implementation. For the simulations and application of this paper INLA is very accurate but in general one should be careful since the algorithm can produce inaccurate inference in some situations, most notably for rare binary events (Fong et al., 2010).

Rao and Wu (2010) have recently proposed another way of combining survey design information and Bayesian models, through a version of empirical likelihood with a similar rescaling by effective sample size. They considered only complete population mean estimation, but an extension of their approach to SAE would be of interest.

Often there will be areas within which no samples are collected. In such situations, under the Bayesian approach that we have followed, the unknown count can be treated as a missing value. A prediction for this count can be carried out as part of the model fitting, and is easily implemented in INLA.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institutes of Health grant R01 AI029168. The authors are grateful to Laina Mercer for helpful discussions and aid in the computation of the variance adjustment.

References

Asparouhov T. General multi-level modeling with sampling weights. Communications in Statistics – Theory and Methods. 2006; 35:439–460.

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.

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.

Fay R, Herriot R. Estimates of income for small places: an application of James–Stein procedure to census data. Journal of the American Statistical Association. 1979; 74:269–277.

Fong Y, Rue H, Wakefield J. Bayesian inference for generalized linear mixed models. Biostatistics. 2010; 11:397–412. [PubMed: 19966070]

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

Ghitza Y, Gelman A. Deep interactions with mrp: Election turnout and voting patterns among small electoral subgroups. American Journal of Political Science. 2013; 57:762–776.

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

Korn, E.; Graubard, B. Analysis of Health Surveys. John Wiley and Sons; New York: 1999.

Longford N. Model-based variance estimation in surveys with stratified clustered design. Australian Journal of Statistics. 1996; 38:333–352.

Lumley, T. Complex Surveys: A Guide to Analysis using R. John Wiley and Sons; Hoboken, Jersey: 2010.

MacGibbon B, Tomberlin T. Small area estimates of proportions via empirical Bayes techniques. Survey Methodology. 1989; 15:237–252.

Malec D, Sedransk J, Moriarity CL, LeClere FB. Small area inference for binary variables in the National Health Interview Survey. Journal of the American Statistical Association. 1997; 92:815–826.

Paul M, Riebler A, Bachmann L, Rue H, Held L. Bayesian bivariate meta-analysis of diagnostic test studies using integrated nested laplace approximations. Statistics in Medicine. 2010; 29:1325–1339. [PubMed: 20101670]

Pereira LN, Coelho P. Small area estimation of mean price of habitation transaction using timeseries and cross-sectional area-level models. Journal of Applied Statistics. 2010; 37:651–666.

Pfeffermann D. New important developments in small area estimation. Statistical Science. 2013; 28:40–68.

Pfeffermann D, Skinner C, Holmes D, Goldstein H, Rasbash J. Weighting for unequal selection probabilities in multilevel models. Journal of the Royal Statistical Society, Series B. 1998; 60:23–40.

Potthoff R, Woodbury M, Manton K. ”Equivalent sample size” and ”equivalent degrees of freedom” refinements for inference using survey weights under superpopulation models. Journal of the American Statistical Association. 1992; 87:383–396.

Pratesi M, Salvati N. Small area estimation: the EBLUP estimator based on spatially correlated random area effects. Statistical Methods and Applications. 2008; 17:113–141.

Rabe-Hesketh S, Skrondal A. Multilevel modelling of complex survey data. Journal of the Royal Statistical Society, Series A. 2006; 169:805–827.

Raghunathan T, Xie D, Schenker N, Parsons V, Davis W, Dood K, Feuer E. Combining information from two surveys to estimate countylevel prevalence rates of cancer risk factos and screening. Journal of the American Statistical Association. 2007; 102:474–486.

Rao, J. Small Area Estimation. John Wiley; New York: 2003.

Rao J, Wu C. Bayesian pseudo-empirical-likelihood intervals for complex surveys. Journal of the Royal Statistical Society, Series B. 2010; 72:533–544.

Riebler A, Held L, Rue H. Estimation and extrapolation of time trends in registry data – borrowing strength from related populations. Annals of Applied Statistics. 2012; 6:304–333.

Rue, H.; Held, L. Gaussian Markov Random Fields: Theory and Application. Chapman and Hall/CRC Press; Boca Raton: 2005.

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.

Särndal, CE.; Swensson, B.; Wretman, J. Model Assisted Survey Sampling. Springer; New York: 1992.

Schrodle B, Held L, Riebler A, Danuser J. Using inla for the evaluation of veterinary surveillance data from switzerland: A case study. Journal of the Royal Statistical Society, Series C. 2011; 60:261–279.

Singh B, Shukla G, Kundu D. Spatio-temporal models in small-area estimation. Survey Methodology. 2005; 31:183–195.

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.

Srebotnjak T, Mokdad A, Murray C. A novel framework for validating and applying standardized small area measurement strategies. Population Health Metrics. 2010; 8:1–13. [PubMed: 20181218]

Wakefield J. Multi-level modelling, the ecologic fallacy, and hybrid study designs. International Journal of Epidemiology. 2009; 38:330–336. [PubMed: 19339258]

Zheng H, Little R. Penalized spline model-based estimation of the finite population total from probability-proportional-to-size samples. Journal of Official Statistics. 2003; 19:99–17.

Zheng H, Little R. Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline non parametric model. Journal of Official Statistics. 2005; 21:1–20.

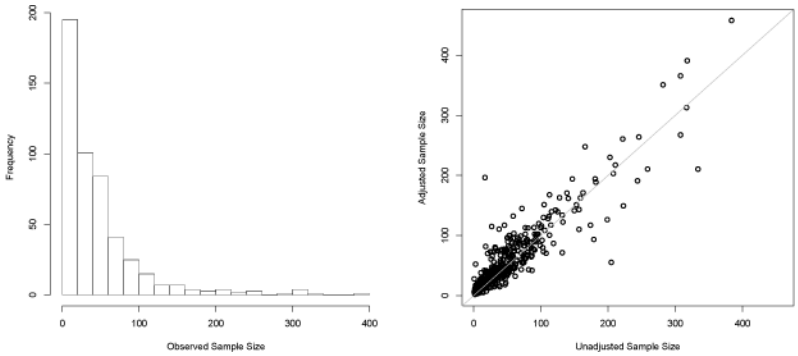

Figure 1.

For 2006 Washington State BRFSS data: histogram of *observed* sample sizes by zip code (left), and *effective* sample sizes versus *observed* sample sizes (right).

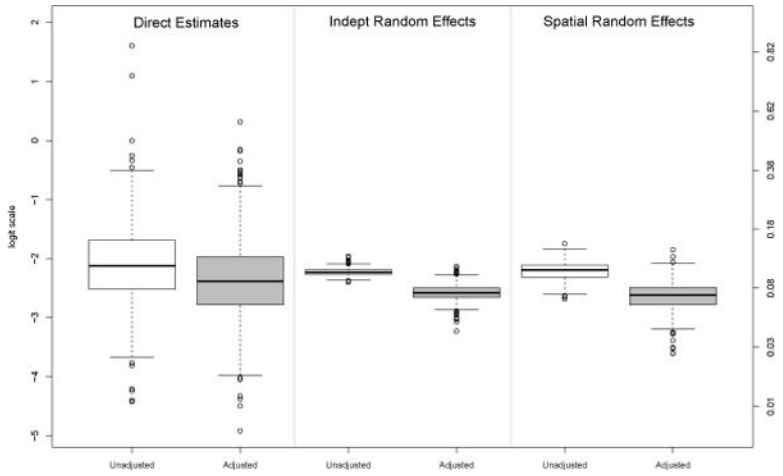

Figure 2.

Estimated diabetes prevalence by zip code under various unadjusted and adjusted models for the 2006 BRFSS data. The left axis is on the logit scale and the right axis is on the [0, 1] scale.

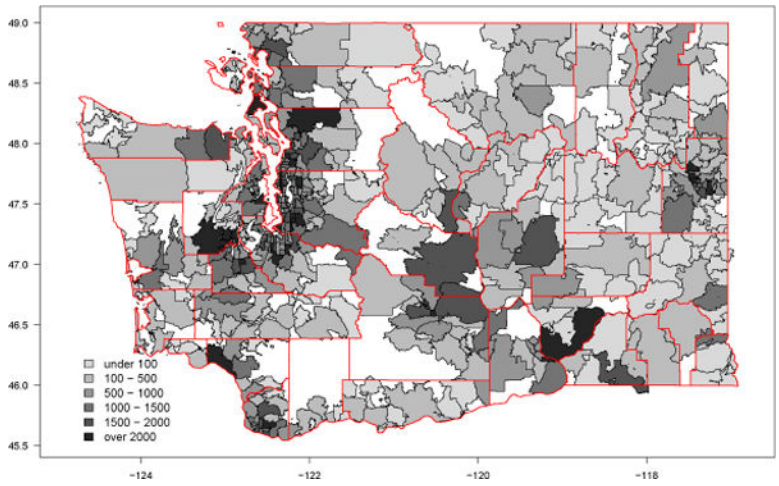

Figure 3.

The adjusted estimates of the total diabetes counts by zip code in Washington State under the spatial model. The red lines denote county boundaries.

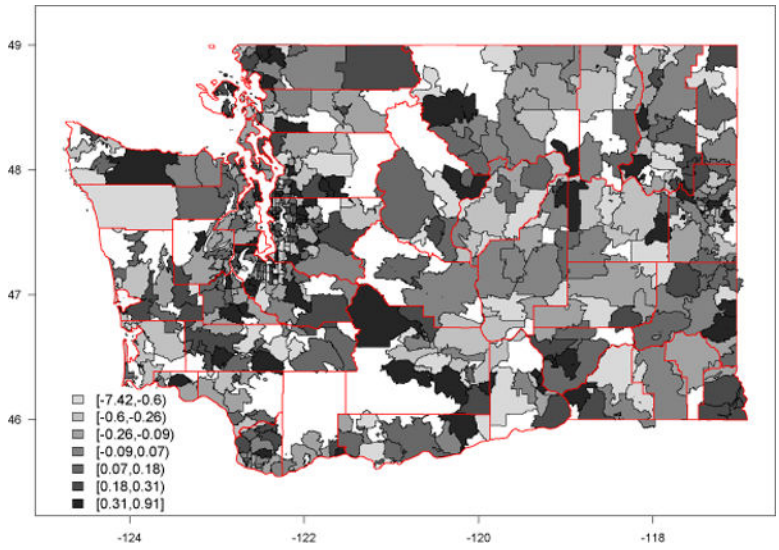

Figure 4.
Map of the standardized difference between the adjusted spatial model and the adjusted direct approach.

Table 1

Summary statistics for population data, and for the 2006 Washington State BRFSS diabetes data, across zip codes.

	Mean	Std. Dev.	Median	Min	Max
Population	12570	12931	7208	11	55700
Sample Sizes	46.9	55	30	1	384
Diabetes Cases	4.6	6	3	0	38
Sample Weights	203.5	237.5	134	0.82	4992

Table 2

Diabetes prevalence rates p_{ij} in area i, $i = 1, \ldots, I$, and by age and gender group, $j = 1, \ldots, 6$. In scenarios 1, 2, 3 and 5 the rates are fixed across areas. In scenario 4 the values vary, with spatial structure, across areas, with the first figure in each cell denoting the median rate, and the figures in parentheses a 95% range.

		Age		
	Scenario	**18–44**	**45–74**	**75+**
Female	1, 2, 3, 5	0.017	0.15	0.17
	4	0.017 (0, 0.034)	0.15 (0.085, 0.21)	0.17 (0, 0.32)
Male	1, 2, 3, 5	0.014	0.16	0.19
	4	0.014 (0, 0.027)	0.16 (0.089, 0.23)	0.19 (0, 0.33)

Table 3

Response rates q_{ij} in area i, $i = 1, \ldots, I$ and by age and gender group, $j = 1, \ldots, 6$. In scenarios 1 and 4 there is full response. In scenario 2 the response rates are fixed across areas but vary by group. In scenario 3 the response rates vary, without spatial structure, across areas, with the first figure denoting the median rate, and the figures in parentheses a 95% range. In scenario 5 the response rates vary, with spatial structure, across areas, with the first figure in each cell denoting the median rate, and the figures in parentheses a 95% range.

		Age		
	Scenario	**18–44**	**45–74**	**75+**
Female	1, 4	1	1	1
	2	0.55	0.65	0.8
	3	0.55 (0.38, 0.70)	0.65 (0.48, 0.79)	0.80 (0.67, 0.89)
	5	0.55 (0.46, 0.65)	0.65 (0.57, 0.74)	0.80 (0.74, 0.86)
Male	1, 4	1	1	1
	2	0.50	0.60	0.75
	3	0.50 (0.34, 0.66)	0.60 (0.43, 0.75)	0.75 (0.60, 0.86)
	5	0.50 (0.41, 0.60)	0.60 (0.51, 0.69)	0.75 (0.68, 0.82)

Table 4

Results for the non-response bias simulation, based on 100 simulations. Unadjusted models use the observed y_i and m_i, adjusted use y_i^* and m_i^*. Inference is made on the total counts T_i. Figures in bold denote the row minimum.

(×10^3)	Direct Estimates		Bayes Indept Normal		Bayes Spatial Normal	
Bias^2	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
Scenario 1	**2.2**	2.9	26.3	15.5	18.9	12.4
Scenario 2	12.8	**3.1**	55.7	17.7	41.9	13.6
Scenario 3	17.4	**7.6**	55.8	19.3	41.6	14.6
Scenario 4	**2.2**	2.9	26.3	15.5	18.9	12.4
Scenario 5	14.4	**3.9**	56.1	18.5	41.7	14.0
Variance	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
Scenario 1	236.9	223.0	**5.6**	14.2	7.0	14.2
Scenario 2	253.1	209.4	**5.4**	11.3	7.0	11.5
Scenario 3	249.9	208.1	**5.7**	11.4	7.4	11.8
Scenario 4	236.9	223.0	**5.6**	14.2	7.0	14.2
Scenario 5	253.6	209.4	**4.8**	10.0	7.0	10.6
MSE	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
Scenario 1	239.1	225.9	31.9	29.8	**25.9**	26.5
Scenario 2	265.9	212.5	61.2	29.0	48.9	**25.1**
Scenario 3	267.3	215.7	61.4	30.7	49.0	**26.4**
Scenario 4	239.1	225.9	31.9	29.8	**25.9**	26.5
Scenario 5	267.9	213.3	60.9	28.5	48.7	**24.6**

Table 5

Results for the selection bias simulation, based on 100 simulations. Unadjusted models use the observed y_i and m_i, adjusted use y_i^* and m_i^*. Inference is made on the total counts T_i. The selection bias is determined by $s = \Pr(Z_{ijk} = 1 | y_{ijk} = 1)$ where Z is an unmeasured binary character; $s = 0.1$ corresponds to no selection bias. Figures in bold denote the row minimum.

(×10^3)	Direct Estimates		Bayes Indept Normal		Bayes Spatial Normal	
Bias^2	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
$s = 0.1$	**3.9**	5.1	29.2	6.2	20.3	6.1
$s = 0.3$	613.4	**4.3**	879.9	8.4	764.9	10.4
$s = 0.5$	1795.0	**2.9**	2282.2	12.2	2062.4	19.3
$s = 0.8$	3549.5	**1.6**	4309.3	19.4	4016.6	19.5
Variance	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
$s = 0.1$	319.5	512.1	**6.2**	210.7	7.9	206.3
$s = 0.3$	472.6	414.6	**10.0**	142.1	12.6	134.9
$s = 0.5$	531.7	313.3	**4.2**	61.6	10.4	52.7
$s = 0.8$	548.8	171.6	**1.9**	**1.9**	7.5	3.4
MSE	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
$s = 0.1$	323.5	517.1	35.4	216.9	**28.2**	212.4
$s = 0.3$	1086.0	418.9	889.9	150.5	777.5	**145.3**
$s = 0.5$	2326.6	316.3	2286.4	73.8	2072.8	**71.9**
$s = 0.8$	4098.4	173.2	4311.1	**21.3**	4024.1	22.9