The Use of Sampling Weights in Bayesian Hierarchical Models for Small Area Estimation

Cici X. Chen
Department of Statistics, University of Washington Seattle, USA

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

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

Technical Report no. 583
Department of Statistics
University of Washington
Seattle, Washington, USA

Abstract

Empirical Bayes and Bayes hierarchical models have been used extensively for small area estimation. However, the sampling weights that are required to reflect complex surveys are rarely considered in these models. In this paper, we develop a method to incorporate the sampling weights for binary data when estimating, for example, small area proportions or predicting small area counts. We consider empirical Bayes beta-binomial models, and normal hierarchical models. The latter may include spatial random effects, with computation carried out using the integrated nested Laplace approximation, which is fast. A simulation study is presented, to demonstrate the performance of the proposed approaches, and to compare results from models with and without the sampling weights. The results show that estimation mean squared error can be greatly reduced using the proposed models, when compared with more standard approaches. Bias reduction occurs through the incorporation of sampling weights, with variance reduction being achieved through hierarchical smoothing. We also analyze data taken from the Washington 2006 Behavioral Risk Factor Surveillance System.

keywords:Integrated Nested Laplace Approximations; Poststratification; Sample Surveys; Spatial Smoothing.

1 Introduction

Small area estimation is a common enterprise in many disciplines, particularly in the social sciences and public health. Often such estimation is based on data arising from complex surveys, for which sampling weights are calculated to account for the disproportionate nature of the sample, by comparison with the target population of interest. Two common forms of bias that are controlled for by weights are non-response and non-coverage bias. In addition to bias control, a second important consideration in small area estimation is variance reduction. As the target areas of interest decrease in size, the uncertainty in estimation is increased, and so the search for smoothing techniques has been popular. From a statistical perspective there are a number of important issues that need consideration when faced with data from a survey. First, one needs to choose whether to use a design- or model-based approach to inference. Unfortunately, though a model-based approach is appealing, the practical difficulties of modeling complex survey design are currently problematic, see Gelman (2007) and the ensuing discussion. Second, one needs to decide on whether the target of inference is a characteristic of the population from which sampling has been carried out (for example an area total), or a characteristic of the superpopulation from which the population was hypothetically sampled (Graubard and Korn, 2002).

Many hierarchical approaches are available for reducing the mean squared error in estimation of small area proportions or counts, and here we provide a brief flavor of developments. Hierarchical models for small area estimation can be traced back to the Fay-Herriot model (Fay and Herriot, 1979), in which an adaptation of the James-Stein estimator was applied to sample estimates of income for small places (in their case, populations less than 1,000). Datta and Ghosh (1991) described a hierarchical Bayes model in which predictive distributions were derived for the unobserved non-sampled responses, in a normal linear model. Hierarchical Bayes models developed for binary survey data include Nandram and Sedransk (1993), in which Bayesian predictive inference was carried out for a finite population proportion from a two-stage cluster sample. A comprehensive treatment of the Bayesian predictive approach to binary survey was provided by Stroud (1994), and included simple random, stratified, cluster, and two-stage sampling, as well as two-stage sampling within strata. Ghosh et al. (1998) provide a general approach for small-area estimation based on hierarchical Bayes generalized linear models, with extension to a spatial correlation random effect structure. Farrell (2000) describes a logistic hierarchical model with computation via Markov chain Monte Carlo; the model compares favorably with an empirical Bayes technique (Farrell et al., 1997), though design bias is present in both procedures. Malec et al. (1997) describe a hierarchical Bayes model for binary survey data. They examined the use of sampling weights as a covariate in the model and did not find any improvement for their example of county-level data from the National Health Interview Survey. Review articles on small area estimation include Rao (1999) and Pfeffermann (2002), with Rao (2003) providing a comprehensive review of design-based, empirical Bayes, and hierarchical Bayesian methods. In this paper we add to this literature by providing a simple hierarchical approach that is appropriate for data with associated sampling weights.

The outline of this paper is as follows. We begin, in Section 2 with a motivating example that concerns diabetes prevalence in the Behavioral Risk Factor Surveillance System (BRFSS), and then describe traditional approaches to estimation in Section 3. Section 4 describes our modeling approach, which is then applied to simulated data in Section 5. We return to the BRFSS data in Section 6, and conclude the paper with a discussion in Section 7.

2 Motivating Example

BRFSS is an annual telephone health survey system 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. The objective of the BRFSS is to provide uniform, state-specific estimates of the prevalence of risk behaviors. In the BRFSS survey, interviewees (who are 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. In this paper we focus on the survey conducted in Washington State in 2006, and on the question, "Have you ever been told you have diabetes?", with interviewees responding with either a "Yes"or a "No". Therefore the response variable is a binary indicator of the presence of diabetes, and our objective is to estimate the number of 18 or older individuals with diabetes, by zip code, in Washington state. The CDC currently publishes coarser, county-level prevalence estimates using the model of Malec et al. (1997), at `http://apps.nccd.cdc.gov/DDTSTRS/`.

In 2006, the survey used land-lines only (from 2008, a small number of cell phones supplement the landlines), 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 then 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 then calculated as the product of four terms:

$$\texttt{Sample Wt} = \texttt{Strat Wt} \times \frac{1}{\texttt{No Telephones}} \times \texttt{No Adults} \times \texttt{Post Strat Wt} \tag{1}$$

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 residentual 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. The latter is given by the number of people in strata defined by gender and age and gender, with the 7 age groups 18–24, 25–34, 35–44, 45–54, 55–64, 65–74, 75+. The other source of data we use are population estimates for 2006.

Table 1: Summary statistics for population data, and 2006 Washington 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

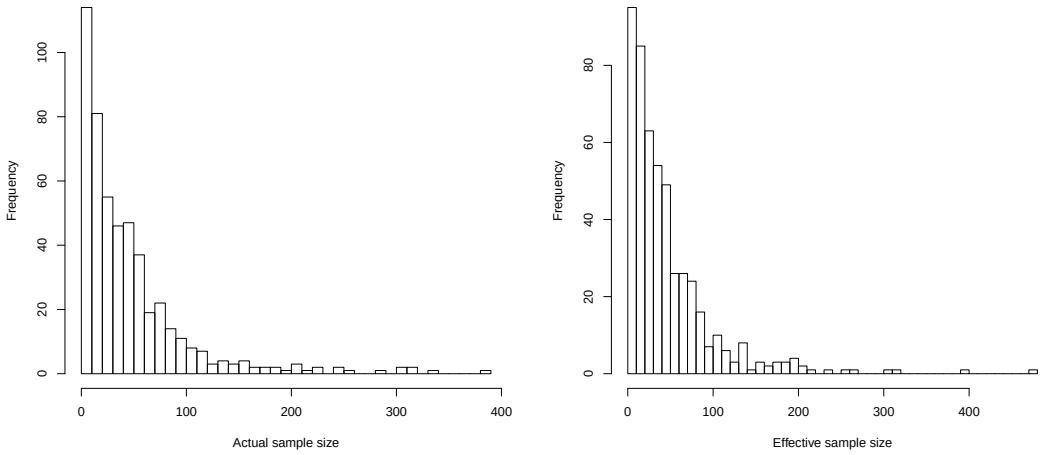

Figure 1: For 2006 Washington BRFSS data: histograms of *actual* sample sizes by zip code (left), and *effective* sample sizes (right).

Basic summary statistics for 2006 Washington BRFSS diabetes data, by 498 zip codes, is presented in Table 1. A total of $23,379$ individuals answered the diabetes question in the survey. The table shows there is a large variability in the population, sample and number of diabetes cases across zip codes. About 20% of the areas have sample sizes less than 10, so that the diabetes prevalence estimates are highly unstable in these areas in particular.

3 Notation and the Conventional Methods of Analysis

3.1 Notation

Let i index area, j the group classification by which post-stratification is carried out (which is age and gender in our motivating example) and k the individual, $i = 1, \ldots, I$, $j = 1, \ldots, J$ and $k = 1, \ldots, N_{ij}$, so that N_{ij} represents the population size in area i and group j. In this section, to simplify notation, we assume there are no strata beyond those used for post-stratification. Let Y_{ijk} denote the binary variable

indicating if the k^{th} individual in group j from area i has the outcome of interest ($Y_{ijk} = 1$) or not ($Y_{ijk} = 0$). Common small area characteristics of interest are the true proportions, $P_i = \frac{\sum_{j=1}^{J}\sum_{k=1}^{N_{ij}} Y_{ijk}}{\sum_{j=1}^{J} N_{ij}}$, or true counts, $T_i = \sum_{j=1}^{J}\sum_{k=1}^{N_{ij}} Y_{ijk}$, in area i, $i = 1, \ldots, I$. In the context of a complex survey, let S_{ijk} denote the binary variable indicating whether the k^{th} individual in group j from area i is sampled ($S_{ijk} = 1$) or not ($S_{ijk} = 0$). Given N_{ij} we may calculate selection probabilities as $\pi_{ij} = \frac{1}{N_{ij}}\sum_{k=1}^{N_{ij}} S_{ijk}$. In addition, let R_{ijk} be the binary variable indicating whether the sampled individual responds to the survey ($R_{ijk} = 1$) or not ($R_{ijk} = 0$). We let $m_{ij} = \sum_{k=1}^{N_{ij}} R_{ijk}$ denote the sample size of group j from area i. For brevity, we use $m_i = \sum_{j=1}^{J} m_{ij}$ to denote the total sample size from area i, and $N_j = \sum_{i=1}^{I} N_{ij}$ to denote the total population size of group j over the study area. To reflect the sampling design, weights w_{ijk} are attached to each respondent's outcome. For example, (1), provides the form of the weights in the BRFSS example. In general, the weights will reflect both the selection probability and post-stratification.

3.2 Conventional Methods

The most commonly used direct unbiased estimator of the area proportion in complex surveys is the (post-stratified) Horvitz–Thompson (Horvitz and Thompson, 1952; Särndal et al., 1992):

$$\widehat{p}_i = \frac{\sum_{j=1}^{J}\sum_{k=1}^{N_{ij}} R_{ijk} w_{ijk} y_{ijk}}{\sum_{j=1}^{J}\sum_{k=1}^{N_{ij}} R_{ijk} w_{ijk}}, \qquad i = 1, \ldots, I, \tag{2}$$

where y_{ijk} is the observed response, and w_{ijk} is the sampling weight assigned to the k^{th} person in area i and group j. A common strategy is to calculate weights as the product of the reciprocal of the sampling probability for selection (the design weight) and the post-stratification weights:

$$w_{ijk} = \pi_{ijk}^{-1} \times \frac{N_j}{\widehat{N}_j}, \tag{3}$$

for $k = 1, \ldots, N_{ij}$ where $\widehat{N}_j = \sum_{i=1}^{I}\sum_{k=1}^{N_{ij}} R_{ijk}\pi_{ijk}^{-1}$, so that $\sum_{i=1}^{I}\sum_{k=1}^{N_{ij}} w_{ijk} = N_j$, the known group totals in the population. Hence, the design weights adjust for the systematic sampling scheme used, while the post-stratification weights attempt to adjust for non-response, by rescaling each group j so that the estimated population total matches the known population total. The computations and the motivation are the same as direct standardization of proportions and rates in epidemiology, although direct standardization is typically used to reduce bias from confounding rather than from non-response. Non-response bias will be removed to the extent that it is predicted by group membership; post-stratification has no impact on differential non-response within groups.

The variance of $\widehat{p}_i$ can be expressed as

$$\widehat{\text{var}}(\widehat{p}_i) = \frac{1}{m_i}\left(\frac{N_i - m_i}{N_i}\right)\widehat{\sigma}_i^2, \qquad i = 1, \ldots, I, \tag{4}$$

where $\widehat{\sigma}_i^2 = \frac{1}{m_i - 1}\sum_{j=1}^{J}\sum_{k=1}^{N_{ij}} R_{ijk}(y_{ijk} - \widehat{p}_i)^2$ is the empirical variance of y_{ijk}, $j = 1, ..., J; k = 1, ..., N_{ij}$.

The estimator $\widehat{p}_i$ defined in (2) is unbiased in the absence of non-response, and approximately unbiased when post-stratification is used to correct non-response, but it is very imprecise when the sample size is small. For example, for rare events and a small sample size, the empirical variance $\hat{\sigma}_i^2$ can be 0, which results in a zero variance of the estimated proportion given in (4).

In this paper we show that the bias correction provided by sampling weights and non-response weights can be combined with the reduction in variance provided by Bayesian hierarchical models, to achieve more accurate estimation (in a mean squared error sense) than either technique alone.

3.3 Inference

For concreteness we focus on predicting counts. To obtain point and interval estimates of the total count we have

$$\widehat{Y}_i = \mathrm{E}(Y_i) = y_i + \widehat{p}_i \times (N_i - m_i), \qquad i = 1, \ldots, I. \tag{5}$$

with variance estimate

$$\widehat{\mathrm{var}}(\widehat{Y}_i) = \widehat{\mathrm{var}}(\hat{p}_i) \times (N_i - m_i)^2, \qquad i = 1, \ldots, I. \tag{6}$$

In most large-scale surveys the sample size is very small compared to the population size, and the overlap between sample and population is often ignored. In survey terminology, the data are analyzed as if they were sampled with replacement. This simplifies the point estimate of the population count to

$$\widehat{Y}_i = \mathrm{E}(Y_i) = \widehat{p}_i \times N_i, \qquad i = 1, \ldots, I. \tag{7}$$

with variance estimate

$$\widehat{\mathrm{var}}(\widehat{Y}_i) = \widehat{\mathrm{var}}(\widehat{p}_i) \times N_i^2, \qquad i = 1, \ldots, I. \tag{8}$$

This point estimate (6) is valid even with large sampling fractions when $\widehat{p}_i$ is the Horvitz–Thompson estimator (2), because this estimator is based on data from area i only. It is not valid with large sampling fractions for the Bayesian estimators that we describe in the next section, which borrow strength from other regions.

4 Sample Weighted Bayesian Hierarchical Models

4.1 A Definition of Effective Sample Size

As described in the previous section, the weights (3) will correct the mean of a population total for sampling bias, and for non-response to the extent that this is explained by the post-strata. Bayesian modelling requires more than bias correction; we need a likelihood that approximates the distribution of the data/weighted

probabilities. Following Korn and Graubard (1998), we model the weighted probability estimates as binomial proportions, with an "effective sample size" chosen to match the binomial variance to the sampling variance of the estimates. Using the effective sample size rather than the actual sample size allows for the varying information per observation 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, 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. 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.

To approximate the sampling distribution in the estimator (2) we express the sampling variances in terms of the effective sample sizes for simple random samples. In a simple random sample, the estimated variance would be $\widehat{p}_i(1-\widehat{p}_i)/m_i$. For our approach define

$$\widehat{p}_{\cdot j} = \frac{\sum_{i=1}^{I}\sum_{k=1}^{N_{ij}} R_{ijk}w_{ijk}y_{ijk}}{\sum_{i=1}^{I}\sum_{k=1}^{N_{ij}} R_{ijk}w_{ijk}},$$

and $e_{ijk} = y_{ijk} - \widehat{p}_{\cdot j}$. The estimated variance for the post-stratified mean is

$$\widehat{\text{var}}\,(\widehat{p}_i) = \frac{N_i - m_i}{N_i}\frac{1}{m_i(m_i-1)}\sum_{j=1}^{J}\sum_{k=1}^{N_{ij}} R_{ijk}e_{ijk}^2.$$

The "effective sample size" m_i^* is then obtained by solving

$$\frac{\widehat{p}_i(1-\widehat{p}_i)}{m_i^*} \quad = \quad \frac{N_i - m_i}{N_i}\frac{1}{m_i(m_i-1)}\sum_{j=1}^{J}\sum_{k=1}^{N_{ij}} R_{ijk}e_{ijk}^2$$

to give

$$m_i^* = \widehat{p}_i(1-\widehat{p}_i)/\left(\frac{N_i - m_i}{N_i}\frac{1}{m_i(m_i-1)}\sum_{j=1}^{J}\sum_{k=1}^{N_{ij}} R_{ijk}e_{ijk}^2\right). \tag{9}$$

We define the effective number of cases y_i^* as the product of the effective sample size m_i^* and the estimated proportion $\widehat{p}_i$:

$$y_i^* = m_i^* \times \widehat{p}_i. \tag{10}$$

For small sample sizes, the design-weighted estimator $\widehat{p}_i$ in (2) can be zero or unity; the estimated variance of the weighted estimator in (9) is then zero. To avoid this problem we first use an empirical Bayes model based on a beta-binomial model (described in greater detail in Section 4) to estimate the proportion, $\widetilde{p}_j$, using data from all areas where $\sum_j \sum_k Y_{ijk}$ is available to estimate the group proportion with the outcome of interest. We then define the weighted estimator of proportions for these area i as:

$$\widehat{p}_i = \frac{\sum_{j=1}^{J} m_{ij}\widetilde{p}_j}{m_i}.$$

The weighted estimator defined in this way is moved slightly away from 0 or 1. The calculation of the effective sample size for these areas remains as in (9).

4.2 Multiple Strata

In this section we extend the model to allow for $h = 1, ..., H$ strata. We begin with the simplest situation in which groups, j, and areas, i, are nested within strata, $i = 1, ..., I_h$. For example, if h indexes large sampling areas, i indexes smaller areas nested within h, and j age-gender post-stratified groups also within h. In this case we define $e_{hijk} = y_{hijk} - \widehat{p}_{h\cdot j}$ with

$$\widehat{p}_{h\cdot j} = \frac{\sum_{i=1}^{I_h} \sum_{k=1}^{N_{hij}} R_{hijk} w_{hijk} y_{hijk}}{\sum_{i=1}^{I_h} \sum_{k=1}^{N_{hij}} R_{hijk} w_{hijk}},$$

where R_{hijk} and w_{hijk} are the response indicators and sampling weights. In this nested case $\widehat{p}_i = \widehat{p}_{hi}$, $N_i = N_{hi}$ and $m_i = m_{hi}$. The variance is then

$$\widehat{\text{var}}\,(\widehat{p}_{hi}) = \frac{N_{hi} - m_{hi}}{N_{hi}} \frac{1}{m_{hi}(m_{hi} - 1)} \sum_{j=1}^{J} \sum_{k=1}^{N_{hij}} R_{hijk} e_{hijk}^2. \tag{11}$$

The effective sample size, $m_{hi}^{\star}$, is defined exactly as in the non-stratified case by setting $\widehat{\text{var}}\,(\widehat{p}_{hi}) = \widehat{p}_{hi}(1 - \widehat{p}_{hi})/m_{hi}^{\star}$ and solving for $m_{hi}^{\star}$. The National Health And Nutrition Examination Surveys (NHANES) are one example of a large survey where post-stratification groups are nested in strata (Mohadjer et al., 1996).

In the non-nested case, the centering by stratum mean and group mean must be done separately. We define $d_{hijk} = y_{hijk} - \widehat{p}_{\cdot\cdot j}$ and $e_{hijk} = d_{hijk} - \widehat{d}_h$ where

$$\widehat{p}_{\cdot\cdot j} = \frac{\sum_h \sum_i \sum_k R_{hijk} w_{hijk} y_{hijk}}{\sum_h \sum_i \sum_k R_{hijk} w_{hijk}}, \tag{12}$$

and

$$\widehat{d}_h = \frac{\sum_i \sum_j \sum_k R_{hijk} w_{hijk} d_{hijk}}{\sum_i \sum_j \sum_k R_{hijk} w_{hijk}} \tag{13}$$

and the variance is

$$\widehat{\text{var}}\,(\widehat{p}_i) = \frac{N_i - m_i}{N_i} \frac{1}{m_i(m_i - 1)} \sum_h \sum_j \sum_k R_{hijk} e_{hijk}^2. \tag{14}$$

where the sums are taken over all (h, j, k) combinations that exist in the population. In (12) and (13) the summations are over all combinations of indices that occur in the population.

4.3 Hierarchical Models

We employ Bayesian hierarchical models that involve three stages for inference. At the first stage, we approximate the sampling distribution using the design-based variance for the survey-weighted estimator, as

defined in Section 4.1. At the second stage, we model between-area variation using random effect models. Finally, the unknown hyperparameters in the second stage are assigned proper hyperprior distributions at Stage 3.

In the first stage, the data distribution is assumed to be:

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

where y_i^* and m_i^* are as defined (10) and (9), respectively. By construction, the sampling distribution of the commonly used estimator, y_i^*/m_i^*, is unbiased for the population prevalence (under the same conditions as the estimator (2) is unbiased, as detailed in Section 3) p_i 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. As this data distribution is a good approximation to the mean, variance, and range of the actual data distribution, it should give a reasonable approximation to the likelihood for Bayesian inference. At the second stage, we describe three possible random effect models to account for between-area variation.

Model 1: *Independent beta random effects model, with empirical Bayes estimation*
The small area proportions p_i follow a beta distribution:

$$p_i | \alpha, \beta \sim_{iid} \text{Beta}(\alpha, \beta), \qquad i = 1, \ldots, I. \tag{15}$$

The marginal distribution is

$$\begin{aligned} \Pr(y_i^* | \alpha, \beta) &= \int \Pr(y_i^* | p_i) f(p_i | \alpha, \beta)\, dp \\ &= \binom{m_i^*}{y_i^*} \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+y_i^*)\Gamma(\beta+m_i-y_i^*)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+m_i^*)}. \end{aligned} \tag{16}$$

Under a full Bayesian approach a prior is placed on α, β, and these parameters are subsequently integrated over. Unfortunately this integration is not analytically tractable, and so as a simple empirical Bayes alternative α and β can be estimated from the marginal distribution of the data by maximum likelihood to give estimates $\widehat{\alpha}, \widehat{\beta}$. Inference for p_i can then be made based on the posterior mean:

$$\widehat{p}_i = v_i \times \frac{y_i^*}{m_i^*} + (1 - v_i) \times \left(\frac{\widehat{\alpha}}{\widehat{\alpha}+\widehat{\beta}} \right), \qquad \text{where } v_i = \frac{m_i^*}{\widehat{\alpha}+\widehat{\beta}+m_i^*} \tag{17}$$

The estimated p_i is a familiar form, the weighted combination of the maximum likelihood estimator y_i^*/m_i^* and the mean of the Beta prior distribution $\widehat{\alpha}/(\widehat{\alpha}+\widehat{\beta})$. Hence an outlying estimate based on a small (effective) sample size will be shrunk towards the posterior mean. This shrinkage of random effect estimators inherently induces bias in estimation but reduces the estimated variance.

Model 2: *Independent normal random effects model, full Bayes estimation*

A commonly generalized linear mixed model is

$$\begin{aligned} \log\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + V_i \\ V_i|\sigma_v^2 &\sim_{iid} N(0,\sigma_v^2), \qquad i = 1,\ldots,I, \end{aligned} \tag{18}$$

where β_0 is the overall effect and the random effects V_i capture the unexplained log odds ratio of the prevalence in the residuals in area i. When area-level covariates are available, the model can be extended to:

$$\text{logit}(p_i) = \beta_0 + \boldsymbol{x}_i{}^T\boldsymbol{\beta} + V_i,$$

where $\boldsymbol{x}_i$ is a vector of length p associated with area i and $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects.

Model 3: *Intrinsic conditional autoregressive (ICAR) model, full Bayes estimation*
In general, we might expect areal units that are close to each other tend to share more similarities then units that are far away and we would like to exploit this information in order to provide more reliable estimates in each area. Here we adopt the spatial model introduced by Besag et al. (1991) that supplements the independent normal random effects with spatial terms. Specifically, the model is

$$\begin{aligned} \text{logit}(p_i) &= \beta_0 + U_i + V_i \\ U_i|U_j, j \neq i &\sim N\left(\frac{1}{k_i}\sum_{j\in\delta_i} U_j, \frac{\sigma_u^2}{k_i}\right) \\ V_i &\sim_{iid} N(0,\sigma_v^2), \qquad i = 1,\cdots,I, \end{aligned} \tag{19}$$

where δ_i denotes the set of neighbors of area i and k_i is the number of its neighbors. This model includes a non-spatial normal random effect component V_i, and a spatial random effect component U_i to which an intrinsic conditional autoregressive (ICAR) prior is assigned. The spatial random effect U_i has a normal distribution with conditional mean given by the average of its neighbors and conditional variance inversely proportional to the number of its neighbors. The parameter σ_u^2 is a *conditional* variance and it determines the contribution of spatial variation. The variance parameters σ_v^2 and σ_u^2 are on different scales, and therefore cannot be directly compared. However, the amount of variation that can be explained by the spatial component can be estimated as the empirical variance, var(U_i). Again, area-level covariates can be added to the model in a straightforward manner.

In this model the nature of the spatial dependency is defined by the neighborhood structure. For example, a common approach defines areas i and j to be neighbors if they share a common boundary. Other neighborhood schemes are possible, for example, Cressie and Chan (1989) define the neighborhood structure as a function of the distance between centroids. For the independent normal and spatial models we require priors for β_0 and the random effects variances. A normal hyperprior is typically assumed for the former, and gamma distributions for the latter.

A related model has been suggested by Raghunathan et al. (2007), in the context of combining data from

multiple sources. In the context of estimating a proportion, they assume the model

$$\sin^{-1}\sqrt{\widehat{p}_i}|p_i \sim_{ind} N\left(\sin^{-1}\sqrt{p_i}, \frac{1}{4\widetilde{m}_i}\right).$$

The arcsine-squared root transformation stabilizes the variance but may be deficient for areas with small sample sizes.

For inference, formulas (5)–(8) are all still relevant, but with y_i replaced by y_i^* and m_i replaced by m_i^*, and $\widehat{p}_i$ a suitable location estimate such as the posterior median. In the simulations and the BRFSS example we used (7) as the point estimate, since the survey sample sizes in each zip code were small compared to the populations. So far as the variance is concerned we can use the the posterior, $\text{var}(p_i|y)$. A more precise summary is the complete posterior distribution of Y_i, but this is more computationally complex to obtain.

4.4 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. For this reason, we employ the integrated nested Laplace approximation (INLA) which has recently been proposed as a computationally convenient alternative to MCMC. The method combines 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 models using INLA, including a comparison with MCMC. 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, the analysis of the BRFSS diabetes data in Washington State takes about 5 seconds using INLA, and hours using MCMC.

5 Simulation Study

5.1 Simulation Scenarios

We examine five sampling scenarios with different response probabilities. For simplicity we consider the single stratum case. In order to investigate the bias introduced by non-response, we let q_{ijk} denote the response probability associated with the k^{th} individual in group j and area i. Therefore, given that the k^{th} individual is sampled, the response indicator R_{ijk} follows a Bernoulli distribution, $R_{ijk}|S_{ijk} = 1, q_{ijk} \sim Bern(q_{ijk})$. We consider five data generating scenarios:

Scenario 1:

In scenario 1 we assume there is no non-response in the survey. In other words, $q_{ij} = 1$ for all sampled individuals. This is the ideal situation. The prevalence of diabetes we use across six gender-age groups are: Female, 18–44, 0.017; Female, 45–74, 0.15; Female, 75+, 0.17; Male, 18–44, 0.014; Male, 45–74, 0.16; Male, 75+, 0.19. These values are chosen based on the National Surveillance Data from the CDC, `http://www.cdc.gov/diabetes/statistics/prev/national/menuage.htm`

Scenario 2:

In scenario 2 we consider a more practical sampling situation. We assume that not every sampled individual will respond to the survey and the response rate is different for each group j. However, the response rate in each group is the same for each area. The response rates are: Female, 18–44, 0.55; Female, 45–74, 0.65; Female, 75+, 0.80; Male, 18–44, 0.50; Male, 45–74, 0.60; Male, 75+, 0.75. The groups with older people have slightly higher response rates, which is generally considered to be true in surveys.

Scenario 3:

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

$$\text{logit}(q_{ij}) = \text{logit}(q_j) + b \times \epsilon_i, \qquad i = 1, \cdots, I,$$

where $\epsilon_i \sim_{iid} N(0, 1)$. The response rates in this scenario are formulated such that the average response rate for each group are the same as in scenario 2. We set $b = 0.35$ to give response rates with 10% and 90% quantiles of 0.46 and 0.81.

Scenario 4:

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

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

We choose $b = 0.2$ to allow sufficient variation in the prevalence rates between area. Across areas the 10% and 90% quantiles for prevalence p_{ij} are 0.013 and 0.20. The setup in this simulation scenario induces far greater variability in group prevalence than in the other scenarios. To simulate spatially correlated covariates x_i, we employ an ICAR model with mean 0 and conditional variance 1. Details on how to simulate from ICAR models can be found in Fong et al. (2010). The purpose of this simulation is to investigate the effect of the underlying spatial dependency on small area estimation when the underlying cause of the dependency is not observed. In this case, spatial models can be used as a surrogate for the unmeasured covariates. For the analysis using this scenario, we use the spatial model introduced in Section 4 but omit any area-level covariates to estimate the true counts.

Scenario 5:

In Scenario 5 we allow the response rate for each group to vary between areas by adding a spatial component

to the variation:

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

where x_i again is simulated from an ICAR model with mean 0 and conditional variance 1. We let $b = 0.3$ to give 10% and 90% quantiles for the response rate q_{ij} across areas as 0.49 and 0.79.

In Scenarios 1, 2, 3 and 5, the underlying diabetes prevalence p_i is considered to be the same for each area. In Scenario 4, the true prevalence rates exhibit spatial dependency and so a second set of prevalences are needed. The population sample sizes are the same in all scenarios.

To draw samples from the population, the sampling strategy we take is, for a particular zip code, to randomly draw individuals from the population. The sample size m_i is chosen to be the actual number of individuals who responded in the Washington 2006 BRFSS survey. For 9 areas with 1 sample only, we change the sample size to 2 in order to provide variance estimates.

5.2 Simulation Results

We analyze each simulated dataset using the empirical Bayes model, and the independent and spatial full Bayesian models. As a baseline, "conventional" estimates are also calculated with the unadjusted version based on y_i/m_i, and the adjusted version being (2). At the third stage of the full Bayesian model, 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 the 95% of residuals odds in the range of $(0.5, 2.0)$ (Wakefield, 2009). This gives a more prudent prior than the commonly used Gamma(0.001, 0.001). We denote the estimated diabetes counts for each zip code using our proposed method as the "adjusted" estimates. The diabetes counts are also estimated using observed number of diabetes, y_i and observed sample size, m_i for comparison, which we denote as the "unadjusted" estimates.

We compute three statistics to evaluate the estimates in the simulation study: the estimated squared bias, the estimated variance and the estimated mean squared error (MSE). Denote by S the total number of simulations, and y_i the "true" diabetes count in area i (which is the same across simulations). The summary statistics are calculated as:

$$\begin{aligned}
\text{Bias} &= \frac{1}{I}\sum_{i=1}^{I}\left(\widehat{\overline{y}}_i - y_i\right), \text{ where } \widehat{\overline{y}}_i = \frac{1}{S}\sum_{s=1}^{S}\widehat{y}_i^{(s)}, \\
\text{Variance} &= \frac{1}{I}\sum_{i=1}^{I}\left(\sum_{s=1}^{S}(\widehat{y}_i^{(s)} - \widehat{\overline{y}}_i)^2\right), \\
\text{MSE} &= \text{Bias}^2 + \text{Variance}.
\end{aligned}$$

Estimators with small MSE are considered superior, although among estimators with comparable MSE those with lower bias are preferred because they lead to interval estimates with improved calibration.

The results are presented in Table 2 with all results based on 100 simulations. In scenario 1, the unadjusted conventional estimator is approximately unbiased by construction and therefore has the smallest squared bias. The adjusted conventional estimator has slightly larger estimated bias. This is as expected because when non-response does not occur in the survey, nothing is gained by the adjusted estimator. Estimates from the random effect models (i.e., the empirical Bayes model, and the independent and spatial full Bayesian models) all have substantial bias due to the shrinkage towards the overall population prevalence, but reduced variance and MSE. The variance in the estimates is higher for the adjusted estimators than their unadjusted counterparts in the random effect models, due to the information loss in estimating the additional population group mean $\widehat{p}_{\cdot j}$ during post-stratification. This is true in general for all simulation scenarios.

Scenario 2 is a more practical situation, where there is non-response in the survey and response rates are different by age-and-gender group (but the response rate for each group remains constant across areas). In this case, the unadjusted conventional estimator is highly biased due to non-response and this bias can be reduced by post-stratification; this is the main purpose of post-stratification in large surveys. The reduction in bias carries over to the empirical Bayes and Bayesian estimators based on adjusted data, and outweighs the increase in variance. The same message is shown again from the simulation results in scenario 3 and 5.

The simulation setup in scenario 4 is similar to that in scenario 1 but allows more variability in the underlying prevalence. The results show an increase in both the bias and variance estimation under all models, due to the increased variation in the simulated data. However, our proposed method again provides a substantial reduction in estimated bias and also in MSE.

In scenario 4 and 5, we impose spatial dependency in the data but pretend the source of the dependency is unknown to us. The spatial model produces estimates with the smallest MSE. This is because the spatial models can serve as a surrogate for the dependency in the underlying prevalence, especially when the dependency can not be accounted for by adding the area-level covariates. When we include the spatially-varying covariate, the difference in the estimates between independent and spatial full Bayesian models diminishes (results not shown). In general, ignoring the sampling weights produces very poor inference, illustrating that using unadjusted hierarchical models with complex sampling schemes is not a good idea.

6 Motivating Example Revisited

We apply the sample weighted Bayesian hierarchical models we developed in Section 4 to the Washington 2006 BRFSS data introduced in Section 2. Sampling weights w_{ijk} are taken to be the final weight used in the BRFSS survey. For those 9 areas with only 1 observation, the effective sample size and effective number of observation are taken to be the same as the corresponding observed values.

Table 2: Simulation summaries to compare estimated squared bias, variance and mean squared error for five different data generating scenarios, with four different models.

Bias^2($\times 10^3$)		*Conventional*	*Emp Bayes*	*Indept Normal*	*Spatial Normal*
Scenario 1	Unadjusted	2	25	25	20
	Adjusted	5	19	19	15
Scenario 2	Unadjusted	16	58	59	49
	Adjusted	6	28	29	22
Scenario 3	Unadjusted	20	56	57	48
	Adjusted	10	26	26	21
Scenario 4	Unadjusted	3	39	40	30
	Adjusted	7	32	33	26
Scenario 5	Unadjusted	13	55	56	45
	Adjusted	6	26	27	21
Variance($\times 10^3$)					
Scenario 1	Unadjusted	227	6	5	7
	Adjusted	204	14	14	13
Scenario 2	Unadjusted	252	6	5	6
	Adjusted	201	8	8	8
Scenario 3	Unadjusted	250	6	6	7
	Adjusted	199	9	8	9
Scenario 4	Unadjusted	235	8	8	8
	Adjusted	212	15	15	14
Scenario 5	Unadjusted	247	5	5	6
	Adjusted	197	8	8	8
MSE($\times 10^3$)					
Scenario 1	Unadjusted	229	31	30	27
	Adjusted	209	33	33	28
Scenario 2	Unadjusted	268	64	64	55
	Adjusted	207	36	37	31
Scenario 3	Unadjusted	270	63	63	54
	Adjusted	208	35	35	30
Scenario 4	Unadjusted	238	47	48	38
	Adjusted	219	47	48	40
Scenario 5	Unadjusted	260	60	61	51
	Adjusted	203	34	35	29

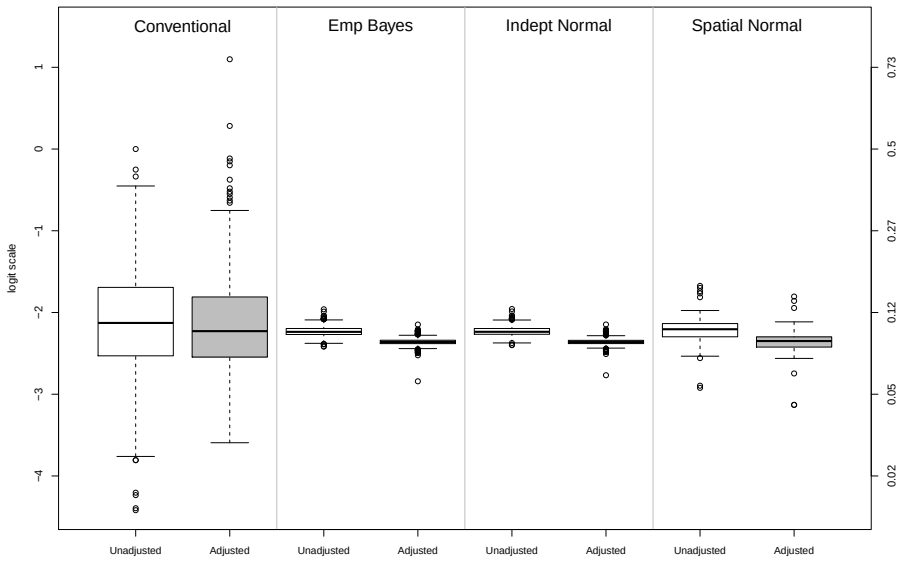

Figure 2: Estimated diabetes prevalence by zip code using models in Section 4: the left axis is on the logit scale and the right is on the $[0, 1]$ scale.

Figure 2 presents the boxplots of logit-transformed estimated diabetes prevalence by zip code under different approaches. For the conventional unadjusted model, we employ the empirical logit transformation, i.e. $\log[(y_i + 0.5/(m_i + 0.5)]$. The figure shows a large amount of variation in the unadjusted conventional estimates due to large sampling variability, with the variability of the adjusted estimates being only slightly reduced. With our proposed adjustment, the variability of the estimates is reduced. Boxplots of the prevalence estimates under the empirical Bayes model and the independent normal random effect are very similar in terms of both the mean and spread. The spatial model gives estimates with increased variation compared to other approaches. The take-home message here is that random effect models can greatly reduce the variance in estimation.

Figure 3 gives the map that we would report, based on the adjusted spatial model. There is higher diabetes population around Puget Sound area (the channel running north-south with with many small, highly populated, zip codes to the east) and the central south area. These areas correspond to the King county, Snohomish county, Spokane county and Yakima valley, which are the most populated counties in Washington State.

Figure 4 shows the difference in square root transformed total diabetes counts between the adjusted spatial model and the adjusted conventional model. We choose the square root transformation because it will approximately stabilize the variance for binomial counts when the prevalences are relatively small, and it is more interpretable than the arcsine-square root transformation which is variance stabilizing for binomial

counts. We see lots of differences, with a magnitude that is important; the totals in Figure 3 have a 10-90% range of (36,2121). There is clear spatial structure in the differences, as we might expect. Figure 5 shows the differences in square root transformed total diabetes counts between the adjusted and unadjusted spatial models. The diabetes prevalence estimate is lower for the adjusted spatial model than the unadjusted in almost all areas, as we saw to a greater or lesser extent with all models in Figure 2.

7 Discussion

In this paper we have described a pragmatic approach to small area estimation, that allows spatial smoothing, and incorporates sample weights to acknowledge the design. 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 problem requires. The simulation study demonstrated much better performance in bias and variance reduction using our proposed approach under a number of difference scenarios. We used the `R` package `INLA` for all our simulations; our code can be found at `http://faculty.washington.edu/jonno/cv.html`

To illustrate the effect of post-stratification we adopt in our method, we now provide some examples that compare the observed sample size and the effective sample size in the analysis. For zip code areas with moderate sample size and somewhat balanced samples in each age-by-gender group, the effective sample sizes and effective number of cases defined in our approach should be close. Take zip code 98022 for example, the sample size for each age-by-gender group is Female, 18–44, 6; Female, 45–74, 6; Female, 75+, 2; Male, 18–44, 10; Male, 45–74, 4 and Male and 75+, 1. The effective sample size is 25 while the observed sample size is 29. The consequent effective number of cases is 1.7 while the observed is 2. There are other areas which show large differences. For example, zip code 98433 has an observed sample size 17. However, all these samples are from two age-by-gender groups with the lowest diabetes prevalence. The number of observed diabetes cases in this zip code is zero. After the adjustment with our proposed method, the effective sample size for this area is estimated as 475.2 with an effective number of cases of 12.24. The observed ratio of effective number of cases to sample size gives a naive estimate of the prevalence as 0.026, which is quite different to 0, which is the estimate based on the observed values.

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 whole-population mean estimation, but an extension of their approach to small-area estimation would be of interest.

References

Besag, J., J. York, and A. Mollie (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Ann. Inst. Statist. Math. 43*, 1–59.

Cressie, N. and N. Chan (1989). Spatial modeling of regional variables. *Journal of the American Statistical Association 84*, 393–401.

Datta, G. and M. Ghosh (1991). Bayesian prediction in linear models: applications to small area estimation. *The Annals of Statistics 19*, 1748–1770.

Farrell, P. (2000). Bayesian inference for small area proportions. *Sankhyā, Series B 62*, 402–416.

Farrell, P., B. MacGibbon, and T. Tomberlin (1997). Empirical Bayes estimation of small area proportions in multistage designs. *Statistical Sinica 7*, 1065–1083.

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

Fong, Y., H. Rue., and J. Wakefield (2010). Bayesian inference for generalized linear mixed models. *Biostatistics 11*(3), 397–412.

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

Ghosh, M., K. Natarajan, T. Stroud, and B. Carlin (1998). Generalized linear models for small-area estimation. *Journal of American Statistical Association 93*(441), 55–93.

Graubard, B. and E. Korn (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science 17*, 73–96.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc. 47*, 663–685.

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

Korn, E. and B. Graubard (1998). Confidence intervals for proportions with small expected number of positive counts estimated from survey data. *Survey Methodology 23*, 192–201.

Malec, D., J. Sedransk, C. L. Moriarity, and F. B. LeClere (1997). Small area inference for binary variables in the national health interview survey. *J Amer Statist Assoc 92*(439), 815–826.

Mohadjer, L., J. Montaquila, J. Waksberg, B. Bell, P. James, I. Flores-Cervantes, and M. Montes (1996). Nhanes iii weighting and estimation methodology. Technical report, National Center for Health Statistics.

Nandram, B. and J. Sedransk (1993). Bayesian predictive inference for a finite population proportion: two-stage cluster sampling. *Journal of the Royal Statistical Society, Series B 55*, 399–408.

Pfeffermann, D. (2002). Small area estimation — new developments and directions. *International Statistical Review 70*(1), 125–143.

Raghunathan, T., D. Xie, N. Schenker, V. Parsons, W. Davis, K. Dodd, and E. Feuer (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *Journal of the American Statistical Association 102*, 474–486.

Rao, J. (1999). Some recent advances in model-based small area estimation. *Survey Methodology 25*, 12–001.

Rao, J. (2003). *Small Area Estimation*. New York: John Wiley and Sons.

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

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Statist. Soc. B 71*, 1–35.

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. Springer-Verlag Inc., New York.

Stroud, T. (1994). Bayesian inference from categorical survey data. *Canadian Journal of Statistics 22*, 33–45.

Wakefield, J. (2009). Multi-level modelling, the ecologic fallacy, and hybrid study designs. *International Journal of Epidemiology 38*, 330–336.

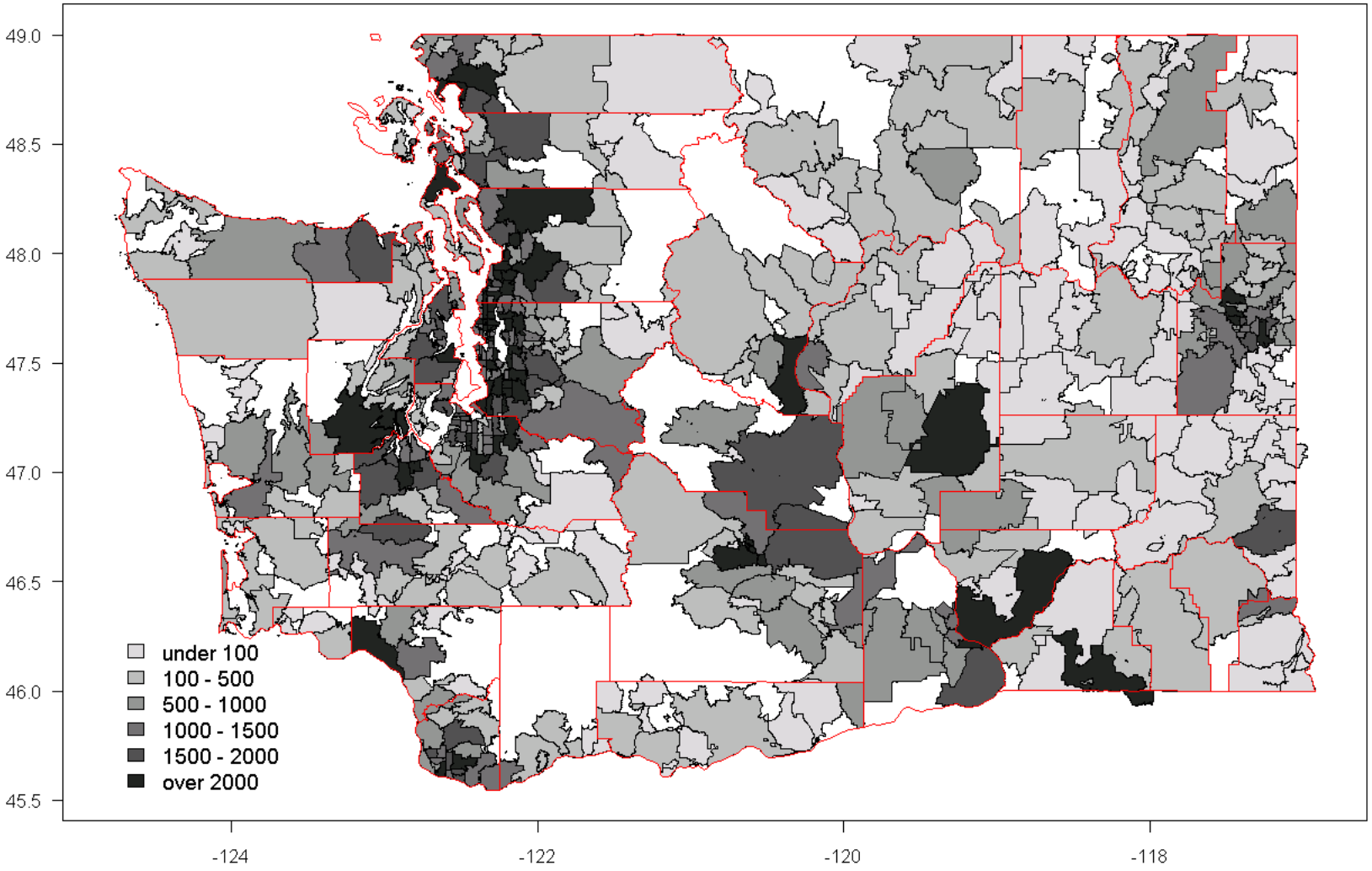

Figure 3: The adjusted estimates of the total diabetes counts by zip code in Washington State under the spatial model.

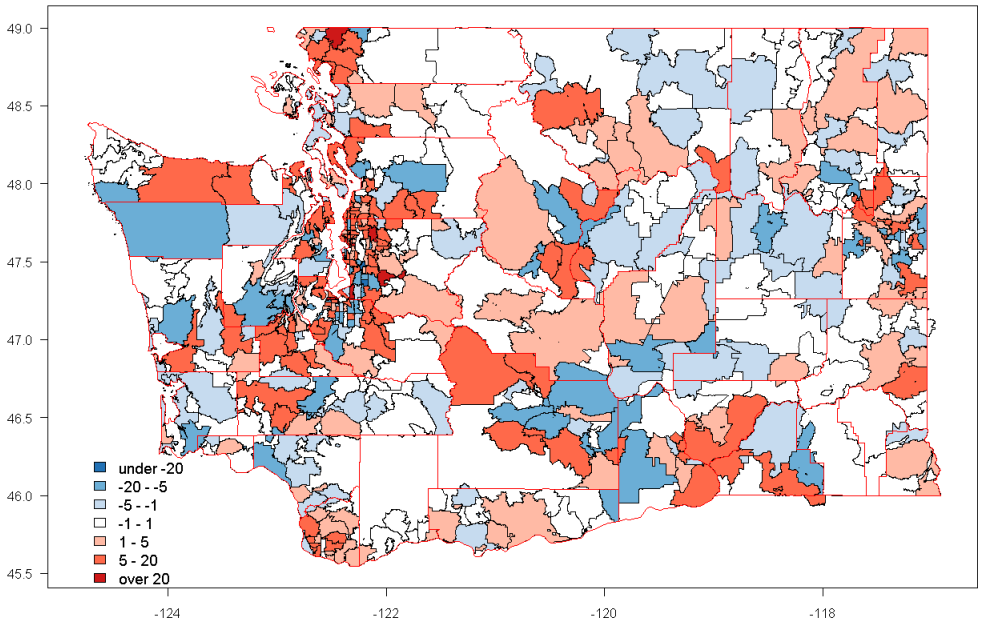

Figure 4: Map of the difference in the square root transformed estimated total diabetes counts between the adjusted spatial model and the adjusted conventional model.

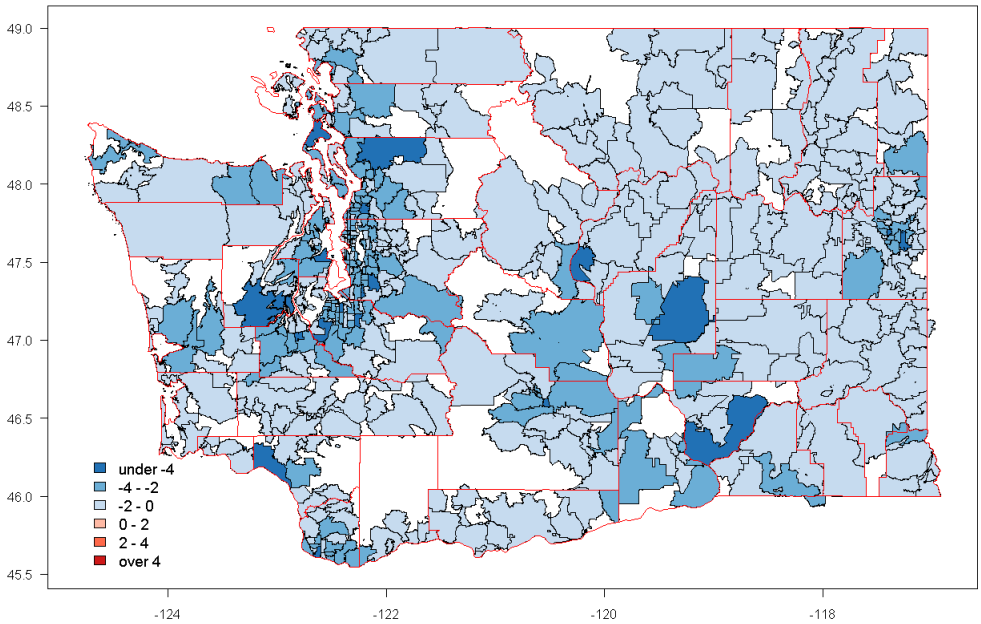

Figure 5: Map of the difference in the square root transformed total diabetes counts between adjusted and unadjusted spatial model.