A Bayesian Approach to Linking a Survey and a Census

Balgobin Nandram[*]

Abstract

 We predict the finite population proportion of a small area when individual-level data are available from a survey and more extensive household-level (not individual-level) data (covariates but not responses) are available from a census. The census and the survey consist of the same strata and primary sampling units (PSUs or wards) that are matched, but the households are not matched. There are some common covariates at the household level in the survey and the census and these covariates are used to link the households within wards. There are also covariates at the ward level, and the wards are the same in the survey and the census. Using a two-stage procedure, we study the multinomial counts in the sampled households within the wards and a projection method to infer about the nonsampled wards. This is accommodated by a three-stage hierarchical Bayesian model for multinomial counts as it is necessary to account for heterogeneity among the households. To proceed, we obtain samples from the distributions of the proportions for each multinomial cell, and then we use these samples to do projective inference for the finite population proportions. Using two projection procedures (parametric and nonparametric), we compare the heterogeneous model and a homogeneous model without household effects. An example on the second Nepal Living Standards Survey is presented.

Key Words: Data integration, Iterative re-weighted least squares, Metropolis sampler, Multinomial-Dirichlet model, Nested error regression model, Projective inference

1. Introduction

In a study on health, one might need to know how many people are in good heath, average health or poor health in different households within different counties in a state. The second Nepal Living Standards Survey has sparse counts of household members within wards (a district consists of wards) for four health status groups. We want to predict the finite population proportion of people in each health category in each ward based on a sample from the households within wards. Undoubtedly there is heterogeneity within wards, the small areas, and not taking this into consideration when inference is made about the finite population proportions within each ward, could lead to biased estimates and to incorrect variability (e.g., Rao and Molina 2015).

The World Bank has developed a method to link a sample survey to a Census using a clustered bootstrap method for poverty estimation (i.e., continuous income data). But this method is biased for several reasons; one of them is that survey and the Census are never done at the same time. Another problem is the method that is used by the World Bank is Empirical Bayes that typically underestimates variance and can also be biased. Recently Corral, Molina and Nguyen (2020) have improved this method considerably by incorporating survey weights and heteroscedasticity. Our

---

[*]Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609

method is different from theirs because it uses count data, not continuous income data, and it is fully Bayesian, not Empirical Bayes. Another major difference is that our method is computationally intensive.

Let the cell counts for the ℓ contingency tables of c cells be $n_{ijk}, k = 1, \ldots, c, j = 1, \ldots, m_i, i = 1, \ldots, \ell$. That is, there are ℓ one-way contingency tables each with m_i households having $n_{ij\cdot}$ individuals partitioned into c cells. There are M_i sub-areas (households) within the ith area (ward) in the population and m_i of these sub-areas are sampled. There are L areas in the population and ℓ of them are sampled. Here, we take $\underset{\sim}{n}'_{ij} = (n_{ijk}, k = 1, \ldots, c)$, where c is the number of health status. We use the dot notation (e.g., $n_{ij\cdot} = \sum_{k=1}^{c} n_{ijk}$). All individuals respond in a household.

We first consider a homogeneous area (ward) level model and we will write down the model for the sample (the model is assumed to hold for the whole population). We assume

$$\underset{\sim}{n}_{ij} \mid \underset{\sim}{p}_i \overset{ind}{\sim} \text{Multinomial}(n_{ij\cdot}, \underset{\sim}{p}_i), j = 1, \ldots, m_i.$$

Here $\underset{\sim}{n}_{i\cdot}$ are sufficient statistics and under this assumption, the hierarchical Bayesian model is

$$\underset{\sim}{n}_{i\cdot} \mid \underset{\sim}{p}_i \overset{ind}{\sim} \text{Multinomial}(n_{i\cdot\cdot}, \underset{\sim}{p}_i),$$

$$\underset{\sim}{p}_i \mid \underset{\sim}{\mu}, \tau \overset{iid}{\sim} \text{Dirichlet}(\underset{\sim}{\mu}\tau), i = 1, \ldots, \ell,$$

$$p(\underset{\sim}{\mu}, \tau) = \frac{(c-1)!}{(1+\tau)^2}, \tau > 0,$$

where, without any prior information, we have taken $\underset{\sim}{\mu}$ and τ to be independent; see Nandram (1998) for computations and Nandram and Sedransk (1993) for an interpretation of τ.

The conditional posterior density of $\underset{\sim}{p}_i \mid \underset{\sim}{\mu}, \tau, \underset{\sim}{n}_{i\cdot} \overset{ind}{\sim} \text{Dirichlet}(\underset{\sim}{n}_{i\cdot} + \underset{\sim}{\mu}\tau), i = 1, \ldots, \ell$. Thus, one can obtain Rao-Blackwellized density estimators of $\underset{\sim}{p}_i$ having obtained a sample of $(\underset{\sim}{\mu}, \tau)$ from their joint posterior density. It is easy to show that

$$\pi(\underset{\sim}{\mu}, \rho \mid \underset{\sim}{n}) \propto \prod_{i=1}^{\ell} \left\{ \frac{\prod_{k=1}^{c} \prod_{s=0}^{n_{i\cdot k}-1} \{\rho s + (1-\rho)\mu_k\}}{\prod_{s=0}^{n_{i\cdot\cdot}-1} \{\rho s + (1-\rho)\}} \right\}, 0 \leq \rho \leq 1, 0 < \mu_k < 1, \sum_{k=1}^{c} \mu_k = 1,$$

where we have transformed τ to $\rho = 1/(1+\tau)$, and any of the arguments must be set to unity if $n_{i\cdot k} = 0$ or $n_{i\cdot\cdot} = 0$, a very likely for some cells. So that this posterior density is well defined for all $\rho, 0 \leq \rho \leq 1$. It is easy to use the Gibbs sampler, not the Metropolis-Hastings sampler (Nandram 1998), to draw samples from $\pi(\underset{\sim}{\mu}, \tau \mid \underset{\sim}{n})$. We will call this model the homogeneous model or the area-level model. This model incorporates only the multinomial counts in the ℓ areas, and it does not take account of the sub-areas, hence the name homogeneous or area-level model. When inference about the sub-area is of interest, one can use the area-level model with the conditional posterior density, $\underset{\sim}{p}_{ij} \mid \underset{\sim}{\mu}, \rho, \underset{\sim}{n}_{ij} \overset{ind}{\sim} \text{Dirichlet}(\underset{\sim}{n}_{ij} + \underset{\sim}{\mu}\tau), j = 1, \ldots, m_i, i = 1, \ldots, \ell$ (i.e., draws of $\underset{\sim}{p}_{ij}$ can be made from it).

We will obtain a sub-area heterogeneous model for the multinomial counts in which there are both sub-area effects and area effects. The Nepal Living Standards Survey (NLSS) have covariates

that are different from those of the Census and they do not affect health directly, and we will call them indirect covariates. These are the nine covariates commonly used for some studies of these data. In this project, we have NLSS data in individual level and census data in household level. So one way to proceed is to ignore the covariates in the NLSS in order to proceed to inference about the finite population proportions, where we can use the covariates. We describe how to use these covariates in the concluding remarks. A further problem is that while the wards in the NLSS match the wards in the Census, the households do not (i.e., there are no labels to make this matching). So we have to match the households in the Census and the NLSS using a record-linkage procedure. There are six variables at the household level (including household sizes) and three variables at the ward level; these latter variables cannot be used for matching but they can be used for prediction. Most of the variables are either discrete or just proportions.

We matched the Census and the NLSS as follows (see Nandram, Chan and Manandhar 2018 for the the six covariates used for matching). We started by matching all six covariates, and this procedure provides a majority of the matches. For those that did not match, we went down to five variables at a time (five sets of matching), then four variables and so on. At the end there were some variables that did not match. For these we used a nearest neighbor matching procedure via a distance function. Then, the final data set consists of households within wards, sampled and nonsampled. The sampled households have responses (counts in different health status). There are also wards not sampled. There are 101 sampled wards (one lost by matching) and 12,133 nonsampled wards. Prediction is needed for the sampled households in the sampled wards and the households of the 101 sampled wards and all the households of the 12,133 nonsampled wards.

Our complete procedure is the following. First, we develop the sub-area model for the multinomial counts. We fit both the homogeneous and heterogeneous models using a Markov chain Monte Carlo sampler. For the small-area model, we use the Gibbs sampler and for the sub-area model we use the Metropolis-Hastings sampler. Nandram, Chen and Manandhar (2018) made approximation to do the posterior density and showed empirically that the approximation is reasonable. However, we use this approximation to construct a proposal density in the sub-area model. From these models, we generate the super-population proportions that are then linked to the Census data through the covariates. Noting that the proportions are the same for each household in a ward, we generated independent and identical proportions for the households within a ward. Then, we fit regression models to the logarithm of the proportions (iterates) to get the regression coefficients. We have used either iterative re-weighted least squares (IRLS) or the nested error regression (NER) model to get the regression coefficients that are then used to generate the counts for both the sampled and nonsampled households in both the sampled wards and the nonsampled wards. The rest is standard Monte Carlo procedures to infer a posteriori the finite population proportions for the wards.

The paper has four more sections. In Section 2, the three-stage hierarchical Bayesian model is described. In Section 3, we describe how to make inference about the finite population proportions. Specifically, we show how to connect the NLSS to the Census. In Section 4, we show how to analyze Nepal's data. We also describe how to calibrate to survey weights to the Census. Section 5 has concluding remarks and there are discussions on how to improve the current method. A technical detail is given in the appendix.

2. Three-Stage Hierarchical Bayesian Models

We extend the three-stage hierarchical Bayesian model of Nandram (1998) to three stages to capture heterogeneity within areas by modeling the sub-area counts.

2.1 General Model

Here, we write down the model for the sample $n_{ijk}, k = 1, \ldots, c, j = 1, \ldots, m_i, i = 1, \ldots, \ell$, but the model is assumed to hold for the entire population of areas and sub-areas, $n_{ijk}, k = 1, \ldots, c, j = 1, \ldots, M_i, i = 1, \ldots, L$.

Therefore, letting $\underset{\sim}{n}_{ij} = (n_{ij1}, \ldots, n_{ij\ell})'$, for the sample the multinomial-Dirichlet-Dirichlet model assumes that

$$\underset{\sim}{n}_{ij} \mid \underset{\sim}{p}_{ij} \overset{ind}{\sim} \text{Multinomial}(n_{ij\cdot}, \underset{\sim}{p}_{ij}), \quad \underset{\sim}{p}_{ij} \mid \underset{\sim}{\mu}_i, \tau \overset{ind}{\sim} \text{Dirichlet}(\underset{\sim}{\mu}_i \tau), j = 1, \ldots, m_i, i = 1, \ldots, \ell,$$

$$\underset{\sim}{\mu}_i \mid \underset{\sim}{\psi}, \eta \overset{iid}{\sim} \text{Dirichlet}(\underset{\sim}{\psi} \eta),$$

$$p(\underset{\sim}{\psi}, \eta, \tau) = \frac{(c-1)!}{(1+\tau)^2 (1+\eta)^2}, \eta > 0, \tau > 0,$$

where, without any prior information, we have taken $\underset{\sim}{\psi}$, η and τ to be independent. For simplicity, we have used a single τ (i.e., not depending on area). It is important to note the dependence of the $\underset{\sim}{p}_{ij}$ on j (i.e., household effects).

Inference about the $\underset{\sim}{p}_{ij}$ and the $\underset{\sim}{\mu}_i$ under the homogeneous model can be obtained by using samples from their joint posterior densities. This is an intermediate step to make inference about the finite population proportions.

It is straightforward to see that

$$\underset{\sim}{p}_{ij} \mid \underset{\sim}{\mu}_i, \tau, \underset{\sim}{n}_{ij} \overset{ind}{\sim} \text{Dirichet}(\underset{\sim}{n}_{ij} + \underset{\sim}{\mu}_i \tau), i = 1, \ldots, \ell.$$

Again, Rao-Blackwellized density estimators of $\underset{\sim}{p}_{ij}$ can be easily obtained. It is useful to note that this conditional posterior density depends only on $\underset{\sim}{\mu}_i$ and τ, but not on the other parameters. Then, integrating out the $\underset{\sim}{p}_{ij}$, we get the joint posterior density of $\underset{\sim}{\mu}, \underset{\sim}{\psi}, \eta, \tau \mid \underset{\sim}{n}$,

$$\pi(\underset{\sim}{\mu}, \underset{\sim}{\psi}, \eta, \tau \mid \underset{\sim}{n}) \propto \frac{1}{(1+\tau)^2} \frac{1}{(1+\eta)^2} \prod_{i=1}^{\ell} g(\underset{\sim}{\mu}_i, \tau \mid \underset{\sim}{n}_i) \{ \frac{\prod_{k=1}^{c} \mu_{ik}^{\psi_k \eta - 1}}{D(\underset{\sim}{\psi} \eta)} \},$$

where, $D(\cdot)$ is the Dirichlet function, and for convenience, we use

$$g(\underset{\sim}{\mu}_i, \tau \mid \underset{\sim}{n}_i) = \prod_{j=1}^{m_i} \frac{D(\underset{\sim}{n}_{ij} + \underset{\sim}{\mu}_i \tau)}{D(\underset{\sim}{\mu}_i \tau)}, i = 1, \ldots, \ell.$$

Note that the conditional posterior density of $(\underset{\sim}{\mu}_i, \tau)$ given $(\underset{\sim}{n}_i, \underset{\sim}{\psi}, \eta)$ is the product of two parts. The first part is $g(\underset{\sim}{\mu}_i, \tau \mid \underset{\sim}{n}_i)$, which is complicated, and the second part is $\pi(\underset{\sim}{\mu}_i, \tau \mid \underset{\sim}{n}, \underset{\sim}{\psi}, \eta) \propto$

$\frac{\Pi_{k=1}^{c} \mu_{ik}^{\psi_k \eta - 1}}{D(\psi \eta)}$, the Dirichlet prior distribution. We can think of the posterior density $\pi(\mu_i, \tau \mid n_i, \psi, \eta)$ as the product of two densities, one proportional to $g(\mu_i, \tau \mid n_i)$, and the other one is $\frac{\Pi_{k=1}^{c} \mu_{ik}^{\psi_k \eta - 1}}{D(\psi \eta)}$. In fact, we can think of $g(\mu_i, \tau \mid n_{ij})$, as the posterior density for each area (in a sub-area model) as in Nandram (1998). That is, this is just the posterior density that Nandram (1998) used to obtain a Metropolis-Hastings sampler to draw samples from; see Appendix A of Nandram, Chen and Manandhar (2018).

Note that ψ and η are not directly connected to the counts even after integrating out the p_{ij}. This indicates that there will be difficulties in running a Markov chain Monte Carlo sampler. Therefore, further integration is necessary. That is, we also need to integrate out the μ_i (i.e., need to draw the parameters simultaneously), but this is not possible analytically.

2.2 Sampling the Joint Posterior Density

We will describe a numerical method to to integrate out the μ_i. Nandram, Lu and Manandhar (2018) has an analytical approximation that we have emphasized therein. Our method to integrate out the μ_i is to use a representation of the Dirichlet distribution that is a product of beta distributions; see Darroch and Ratcliff (1971) and Connor and Mossimann (1969).

Let x be a vector with c components such that $\sum_{j=1}^{c} x_j = 1$ $x_j \geq 0, j = 1, \ldots, c$. Assume that $x \sim \text{Dirichlet}(a)$, where a is a vector with c known elements. Then,

$$p(x_1, \ldots, x_{c-1}) \propto (\prod_{j=1}^{c-1} x_j^{a_j - 1})(1 - \sum_{j=1}^{c-1} x_j)^{a_c - 1}.$$

Let $v_1 = x_1$, $v_j = x_j/(1 - \sum_{s=1}^{j-1} x_s), j = 2, \ldots, c-1$. Then, $v_j \overset{ind}{\sim} \text{Beta}(a_j, \sum_{s=j+1}^{c} a_s), j = 1, \ldots, c-1$.

Using this characterization on $\mu_i \mid \psi, \eta \overset{ind}{\sim} \text{Dirichlet}(\psi \eta)$ with $\psi_1 + \psi_2 + \psi_3 = 1$, we get

$$\pi(\psi, \eta, \tau \mid n) \propto \frac{1}{(1+\tau)^2} \frac{1}{(1+\eta)^2} \prod_{i=1}^{\ell} \left\{ \int_0^1 \int_0^1 g^*(v_i, \tau \mid n_i) \right.$$

$$\left. \times \frac{v_{i1}^{\psi_1 \eta - 1}(1 - v_{i1})^{(\psi_2 + \psi_3)\eta - 1}}{B\{\psi_1 \eta, (\psi_2 + \psi_3)\eta\}} \frac{v_{i2}^{\psi_2 \eta - 1}(1 - v_{i2})^{\psi_3 \eta - 1}}{B(\psi_2 \eta, \psi_3 \eta)} dv_{i1} dv_{i2} \right\}.$$

Here, $g(\mu_i, \tau \mid n_i)$ transforms to $g^*(v_i, \tau \mid n_i)$ and the middle Riemann sum for integration. The integration is easy to carry out by discretization over the range of the independent beta random variables, v_{i1} and v_{i2}.

We can now draw ψ, η, τ using a Metropolis sampler. The candidate generating density is the analytic approximation discussed in Nandram, Chen and Manandhar (2018). We draw a sample of $M = 10,000$ iterates using the Gibbs sampler (to allow a "burn in" and thinning to get 1,000 samples). We showed how to construct a proposal density for (ψ, η, τ) to run the Metropolis sampler (Nandram, Chen and Manandhar 2018). We used a grid method to draw ψ in this preliminary

sample. Thus, we have samples from the approximate posterior density of $(\underline{\psi}^{(h)}, \eta^{(h)}, \tau^{(h)}), h = 1, \ldots, M$. We transform these to $\beta^{(h)}, h = 1, \ldots, M$ where $\beta_i^{(h)} = \log\{\psi_i^{(h)}/(1 - \sum_{j=1}^{c-1} \psi_j^{(h)})\}, i = 1, \ldots, c-1, \beta_c^{(h)} = \log(\eta^{(h)}), \beta_{c+1} = \log(\tau^{(h)})$. Then, we fit a multivariate normal density to $\beta^{(h)}$, where $\hat{\theta}$ and $\hat{\Sigma}$ are the mean and covariance matrix of the samples, and $\kappa/\sigma^2 \sim \text{Gamma}(\kappa/2, \tilde{1}/2)$ to complete the $(p+1)$-variate Student's t density on κ degrees of freedom, where κ is a tuning constant. We restart the algorithm if it is necessary. The order restriction on $\underline{\psi}$ (see below) is enforced by rejection sampling. Both the Metropolis sampler for the exact computations and the Gibbs sampler show good performance as evident by the trace plots, auto-correlations, Geweke test of of stationarity, the effective sample sizes and the jumping rate of the Metropolis sampler.

To draw the μ_i, we use a Metropolis algorithm with the approximate Dirichlet distribution (Nandram 1998) as the proposal density to draw samples of μ_i independently given $\underline{\psi}, \eta, \tau$ and data. We ran each Metropolis sampler 100 times (500 times did not make a significant difference) and picked the last one. If the Metropolis step fails (jumping rate is not in $(.25, .75)$), we use the griddy Gibbs sampler within this Metropolis step. Parallel computing can also be used in this latter step. This is performed in the same manner for the exact method (i.e., numerical integration). For our application with 101 wards, this latter step runs very fast. Of course, with much larger number of wards, the computing time will be substantial, but now parallel (embarrassingly) computing is available.

Because the counts in the households are sparse for many households (see discussion later) it is necessary to adjust the heterogeneous model. The counts in the last cell are mostly zeros, so we decided to combine the last two cells. Even as such it is still sparse. However, we noticed an order restriction of the proportions of household members in the three cells. So we impose the order restriction $1 > \psi_1 > \psi_2 > \psi_3 > 0$. We apply this restriction to the homogeneous model also. The order restriction also helps to provide a better MCMC algorithm. Let us consider how this order restriction changes the conditional posterior densities (cpd's) of ψ_1 and ψ_2; notice $\psi_3 = 1 - \psi_1 - \psi_2$. Thus, the order restriction is really

$$1 > \psi_1 > \psi_2 > 1 - \psi_1 - \psi_2 > 0,$$

and this is a key inequality. Now, we need the support of the cpd of ψ_1 given the other parameters and the support of the cpd of ψ_2 given other parameters. In Appendix B of Nandram, Chen and Manandhar (2020), we showed that that given ψ_2, the support of ψ_1 is $\max\{\frac{1}{3}, \psi_2, 1 - 2\psi_2\} < \psi_1 < 1 - \psi_2$ and given ψ_1, the support of ψ_2 is $\frac{1}{2}(1 - \psi_1) < \psi_2 < \min\{\frac{1}{2}, \psi_1, 1 - \psi_1\}$. We have drawn samples from the cpds of ψ_1 and ψ_2 using the grid method.

3. Inference for Finite Population Proportions

We have now obtained samples $\underline{p}_{ij}^{(h)} = (p_{ijk}, k = 1, \ldots, c = 3), j = 1, \ldots, m_i, i = 1, \ldots, \ell, h = 1, \ldots, M$, say $M = 1000$. The next step is to link these $\underline{p}_{ij}^{(h)}$ to the census. The census has covariates $\underline{x}_{ij}$, a vector of $(r = 10)$ components, including an intercept, for all households, sampled or

nonsampled. There are three parts, the sampled households in the sampled wards, the nonsampled households in the sampled wards and the nonsampled households in the nonsampled wards.

We use the following steps to implement the projection procedure.

a. Model the y_{ijk} for each $k = 1, \ldots, c-1$ (independent regression-type analyzes)

$$y_{ijk} = \log\{p_{ijk}/(1 - \sum_{k=1}^{c-1} p_{ijk})\}, k = 1, \ldots, c-1,$$

see Agresti (2012) for the multinomial logit transformation;

b. Project y_{ijk} for each k using the model (entire census);

c. Define new p_{ijk} using the y_{ijk},

$$p_{ijk} = e^{y_{ijk}}/(1 + \sum_{k'=1}^{c-1} e^{y_{ijk'}}), \; k = 1, \ldots, c-1,$$

and $p_{ijc} = 1 - \sum_{k=1}^{c-1} p_{ijk}, \; j = 1, \ldots, M_i, i = 1, \ldots, L$;

d. Draw the n_{ijk} from multinomial models; obtain copies of the census counts.

This regression analysis will be performed for each of the M samples of $p_{ij}^{(h)}$. We can obtain all the nonsampled p_{ijk} for all M iterates in the same manner. Having obtained the cell probabilities, the multinomial counts can be generated for each of the M iterates, thereby obtaining a large sample (size M) of contingency tables for the entire census. We have two methods of doing this operation in a comprehensive manner.

3.1 Iterative Re-weighted Least Squares Method

We have used the ensemble M-estimation model in small area estimation; see Chambers and Tzavidis (2006), where the regression coefficients are estimated using iterative re-weighted least squares (IRLS). This is an attractive nonparametric procedure for small area estimation because there are no random effects to model, but random effects come out as summaries of q-scores. A good review paper is given by Dawber and Chambers (2019). This is directly related to the procedure mentioned above.

Let y_{ij} denote the responses and x_{ij} the covariates for unit j in area i, $j = 1, \ldots, n_i, i = 1, \ldots, \ell$. We string out these values and relabeling, the responses are y_i and the covariates are x_i for $i = 1, \ldots, n = \sum_{i=1}^{\ell} n_i$. In the IRLS, $\hat{\beta}_q$ for each q in $(0,1)$ is obtained by solving

$$n^{-1} \sum_{i=1}^{n} w_{iq} \psi_{qk}(\frac{y_i - x_i \hat{\beta}_q}{\hat{\sigma}_{qk}}) = 0,$$

where $\hat{\sigma}_{qk}$ are the median absolute deviation of $y_i - \underset{\sim}{x}_i\hat{\underset{\sim}{\beta}}_q$ and the weights are

$$w_{iqk} = \frac{\psi_{qk}\left(\frac{y_i - \underset{\sim}{x}_i\hat{\underset{\sim}{\beta}}_q}{\hat{\sigma}_{qk}}\right)}{\frac{y_i - \underset{\sim}{x}_i\hat{\underset{\sim}{\beta}}_q}{\hat{\sigma}_{qk}}}$$

with $\psi_{q\kappa}(u) = 2[(1-q)I_{u\leq 0} + qI_{u>0}][-\kappa I_{u\leq -\kappa} + uI_{-\kappa<u<\kappa} + \kappa I_{u\geq\kappa}]$. Here, I is the indicator function, κ is a tuning constant and the second term is the Huber influence function. In our analysis, we have set $\kappa = 2$. In the ensemble M-estimation model, the IRLS procedure is executed for every q on a fine grid in $(0,1)$. Then, the ith q-score solves the equation,

$$\underset{\sim}{x}_i\hat{\underset{\sim}{\beta}}_{q_i^*} = y_i, i = 1,\ldots,n.$$

The random effects are then obtained as a summary (e.g., median) of the q_i^* for area i, denoted by q^*. Now, $\hat{\underset{\sim}{\beta}}_{q^*}$ are the estimated regression coefficients for this area.

The IRLS is performed sequentially as follows. Start up the process using standard least square estimators, obtain the residuals and the median absolute deviation of the residuals, and finally the weights. Then, use the weights and the median absolute deviation to get the first re-weighted least square estimates of the regression coefficients. Now iterate the process for each value of q in a fine grid in $(0,1)$; we have used $(.10,.90)$ because of computational instability.

3.2 Nested Error Regression Method

The second method is the nested error regression (NER) model; see Battese, Harter and Fuller (1988). We use the full Bayesian version of the NER model which was originally developed by Toto and Nandram (2010) and later applied to poverty estimation by Molina, Nandram and Rao (2014).

Letting, $y_{ij} = z_{ijk}$ for each k, $k = 1,\ldots,c-1$, the Bayesian NER model is

$$y_{ij} \overset{ind}{\sim} \text{Normal}(\underset{\sim}{x}'_{ij}\underset{\sim}{\beta} + v_i, \sigma^2), j = 1,\ldots,n_i,$$

$$v_i \mid \rho, \sigma^2 \overset{ind}{\sim} \text{Normal}(0, \frac{\rho}{1-\rho}\sigma^2), i = 1,\ldots,\ell,$$

$$\pi(\underset{\sim}{\beta}, \sigma^2, \rho) \propto 1/\sigma^2.$$

Again, the Bayesian NER model is applied $c-1$ times for each of the M samples.

The joint posterior density is proper provided the design matrix is full rank. Molina, Nandram and Rao (2014) need an additional condition for posterior propriety, and it turns out that this condition is not necessary. However, we note here that the NER model is not robust against outliers and non-normality (e.g., see Chakraborty, Datta and Mandal 2019 and Gershunskaya and Lahiri 2018). There is a clear advantage to use M-estimation, albeit it is more complicated and computationally intensive.

The posterior density of ρ can be obtained apart from the normalization constant, and all the other conditional posterior densities, in order to use the multiplication rule to get the joint posterior density, are in standard forms. Therefore, it is easy to get a sample from the joint posterior density using the composition method (i.e., the multiplication rule of probability). Draws from the posterior density of ρ is obtained using a fine grid on $(0, 1)$. This method is very fast, and much faster than the IRLS, because it simply uses random draws, not a Markov chain, and the IRLS has to be done on a fine grid. While the IRLS projection method took more than ten hours even with a parallel system with 32 processors, the NER method took about 5 hours on a single processor.

Of course, the ensemble M-quantile estimation model is nonparametric, but there may be some difficulties in finding the q-scores near 0 and 1, if these are actually needed.

3.3 Bayesian Projective Inference

We can perform either a Bayesian predictive inference or a Bayesian projective inference about the finite population proportions for each ward or even at a more micro level for each household. Predictive inference means that the nonsampled households are filled in while projective inference means that all households (sampled and non-sampled) are filled in. We have actually used the projective method for the NLSS sampled households and the Census because smoothed estimates of the household proportions may be needed for the sampled households.

Finally, in a projective method, we will obtain all the p_{ijk} for both the samples and the nonsamples. Then, we will draw the cell counts for all households. For each ward, sum all the cell counts of the households in a ward, and divide by the population size (known) to get the finite population proportions in the three cells. Then, draw all the cell counts. For each ward, we will add these counts to the already observed counts. For the nonsampled households, use the IRLS estimates at $\underset{\sim}{x}'_{ij}\underset{\sim}{\beta}_{.5}$. For the NER model, we would need to use $\underset{\sim}{x}'_{ij}\underset{\sim}{\beta} + \nu_i$ for the nonsampled households within sampled wards via the posterior samples of the ν_i (already drawn). For the nonsampled wards, we will need to use a draw from the the prior distribution of ν_i (no-data posterior density) and compute $\underset{\sim}{x}'_{ij}\underset{\sim}{\beta} + \nu_i$.

4. Analysis of Nepal's Data

We use data from the Nepal Living Standards Survey (NLSS II, Central Bureau of Statistics, 2003-2004) to illustrate our methods. In Section 4.1, we describe the 2001 Census and NLSS II. In Section 4.2, we apply our method to Terai Rural stratum, the narrow strip of rural areas in southern side of Nepal bordering India; see Figure 1 that shows the districts, not the lower levels of wards and households. We note that in the Terai Rural stratum the wards are scattered in the whole strip. [In this region, there are 29 Mertropolitan areas, each with at least 9 wards, roughly 300 wards in total, and these are obviously not included in the Terai Rural stratum.]

Figure 1: Map of Nepal showing the Terai Region

4.1 Census and NLSS II

The Census and NLSS II (simply NLSS) are the same as described in Nandram, Chen and Manandhar (2018). So we do not repeat the discussion here in detail, but below we give the key points.

NLSS is a national household survey in Nepal, actually population based (i.e., interviews are done for individual household members). NLSS follows the World Bank's Living Standards Measurement Survey methodology with a two-stage stratified sampling scheme, which has been successfully applied in many parts of the world. It is an integrated survey which covers samples from the whole country and runs throughout the year. The main objective of the NLSS is to collect data from Nepalese households and provide information to monitor progress in national living standards. The NLSS gathers information on a variety of subjects. It has collected data on demographics, housing, education, health, fertility, employment, income, agricultural activity, consumption, and various other subjects. We choose the polychotomous variable, health status, from the health section of the questionnaire.

Health status is covered in Section 8 of the questionnaire, which collected information on chronic and acute illnesses, uses medical facilities, expenditures on them and health status. The health status questionnaire is asked about every individual that was covered in the survey, and the questionnaire has four options (excellent, good, fair, poor), but because the fourth cell is overly sparse, we combine fair and poor into a single cell (renamed poor).

In the NLSS, Nepal is divided into wards/sub-areas (psu's) and within each ward/sub-ward there are a number of households. The sample design of the NLSS used two-stage stratified sampling. A sample of psu's was selected using PPS sampling and then twelve households were systematically selected from each ward. Thus, households have equal probability of selection. But while individuals in a household have equal probability of selection, the survey weights have various adjustments, so they vary with the size of the households and the individuals. For this project we will ignore the survey weights because this needs a multinomial logit model that we are currently studying. But we have discussed alternatives in the concluding remarks.

Nandram, Chen and Manandhar (2018) chose nine relevant covariates which can possibly influence health and they are available in the 2001 census data. Six of the covariates are at the household level and three at ward level. At the ward level the NLSS and the Census are matched, but the households are not matched. Therefore, only the household variables are used for matching, but all covariates are used for projection.

There are six strata and we study the Terai Rural stratum, the largest stratum in Nepal. It has 102 wards/sub-wards with 1,224 households in the sample of 12,239 wards in the population (sample frame) with 1,686,317 households with 9,744,810 people. After matching we ended up with 101 wards in the sample and 12,133 wards in the nonsampled part of Terai Rural stratum. The number of people in the sample is 6,979 with 3901 in the first cell, 2921 in the second cell and 157 in the third cell with percentages 55.9%, 41.9%, and 2.2%. The sample of 6,979 will speak for 9,744,810 people (i.e., a sample of just 0.07%). So we have imposed an order restriction over the three cells to assist the computations; specifically, we have taken $\psi_1 > \psi_2 > \psi_3$.

4.2 Comparisons

We compare the two models (Hom, Het) and the two projection methods (IRLS, NER). The NER method, based on the NER model, is not robust against outliers and non-normality, but M-estimation, based on the IRLS, is robust. Here, we compare the four scenarios: (Het, IRLS), (Het, NER), (Hom, IRLS) and (Hom, NER).

In Table 1, we see that the nonsampled part of the population dominates the inference. There is a small, but important, selection bias. Considering the proportion of people in excellent health, the PMs for the four groups (IRLS, Het), (IRLS, Hom), (NER, Het) and (NER, Hom) are respectively .540, .608, .512, .574 for the sample and .453, .512, .378, .389 significantly larger for the nonsample. This selection bias can be mitigated when the survey weights are incorporated into the model.

Inference for the nonsampled part of the population is virtually the same as the whole population. (Compare (b) and (c) of Table 1.) This is not surprising because there are 101 wards in the sample and 12,133 wards in the nonsample (i.e., the sampling fraction, $101/12,234 \approx .0083$, is very small).

In Figures 2 - 7, we further compare the four groups. Each of these plots contains 36,702 points corresponding to the three health classes and 12,234 wards and the plots are for the two posterior means of the finite population mean.

In Figures 2 and 3, we compare the two models (Hom vs. Het). In Figure 2, we can see clearly the difference between the two models for the NER projection method. The points are scattered around the 45-degree straight line with smaller slope. In Figure 3, we can see clearly the difference between the two models for the IRLS projection method. The points are scattered around the 45-degree straight line with smaller slope, but much more concentrated than the NER projection method.

In Figures 4 and 5, we compare the two projection methods (IRLS vs. NER). In Figure 4, we can see clearly the difference between the two models for the NER projection method. The points are scattered around the 45-degree straight line with smaller slope, but much more spread out. In Figure 5, we can see clearly the difference between the two models for the IRLS projection method. The points appear to be in a rectangle with the 45-degree roughly along its diagonal.

In Figures 6 and 7, we look at the interaction of the two factors (model, projection methods). Again, the clouds of points cross the 45-degree straight line in both plots and there is large spread.

The heterogeneous model incorporates variation among the households but the homogeneous model does not. This is a major difference already. The IRLS method is robust against non-normality and outliers, but the NER method is not so robust. In fact, work is being done currently to allow for robustness in the NER projection method.

4.3   Calibration

It is necessary to calibrate the NLSS to the Census using survey weights and census covariates. We would need to do this for each sampled ward. There are nonsampled households in each ward. We have all the household covariates for all the sampled wards in the Census. So we have the totals for all the households within a ward. We do not know the household sizes for the sampled wards; we have used data fusion to fill these in by using the household sizes of the sample to donate to the nonsampled household within a ward. While we have not done this yet, we note that this will be an enormously useful addition to our procedure.

Let $x_j, j = 1, \ldots, n$, denote the covariates, including an intercept, in the sample survey and w_j denote the survey weights. Let $\tilde{x}_j, j = 1, \ldots, N$, where N is the number of households in the census. The method is described for households but it can be used for individuals. Then, $t = \sum_{j=1}^{N} \tilde{x}_j$. The survey weights w_j are calibrated when we can find $\tilde{w}_j$ such that $\sum_{j=1}^{n} \tilde{w}_j x_j = t$, where $\tilde{w}_j$ are the survey weights; a distance function is used to measure closeness between the w_j and the $\tilde{w}_j$. In a paper in Statistical Science, Haziza and Beaumont (2020) gave a very informative and clear pedagogical review of this method.

Let the original survey weights in the NLSS be $w_j, j = 1, \ldots, n$. We search for a calibrated weighting system $\tilde{w}_j, j = 1, \ldots, n$, such that $\sum_{j=1}^{n} \tilde{w}_j x_j = t$, the calibration equation. We want to the $\tilde{w}_j, j = 1, \ldots, n$, to be as close as possible to $w_j, j = 1, \ldots, n$. Haziza and Beaumont (2017) judge closeness by a distance function, $G(u)$, where

a. $G(u) \geq 0$ and $G(1) = 0$;

b. $G(u)$ is differentiable, $g(u) = G'(u), g(1) = 0$, and strictly convex.

One needs to minimize $\sum_{j=1}^{n} \frac{w_j}{q_j} G(\frac{\tilde{w}_j}{w_j})$ over $\tilde{w}_j, j = 1, \ldots, n$, where q_j denote the importance of unit j, subject to the constraint $\sum_{j=1}^{n} \tilde{w}_j x_j = t$. We present some details here on how to solve the Lagrangian system of equations for a general $G(u), -\infty < u < \infty$. (we have made some changes to Haziza and Beaumont, 2017).

We give the final step on how to numerically compute λ for a general $G(u)$. Consider the function,

$$\phi(\tilde{w}_1, \ldots, \tilde{w}_n, \lambda) = \sum_{j=1}^{n} \frac{\tilde{w}_j G(\frac{\tilde{w}_j}{w_j})}{q_j} - \lambda'(\sum_{j=1}^{n} \tilde{w}_j x_j - t),$$

where $\lambda = (\lambda_1, \ldots, \lambda_p)$ are Lagrangian multipliers.

Differentiating $\phi(\tilde{w}_1, \ldots, \tilde{w}_n, \lambda)$ with respect to $\tilde{w}_j$, we get

$$\tilde{w}_j = w_j g^{-1}(q_j \lambda' x_j), j = 1, \ldots, n.$$

Therefore,

$$\sum_{j=1}^{n} w_j g^{-1}(q_j \lambda' x_j) x_j = t.$$

Here, our method differs from Haziza and Beaumont (2020) who used the Newton-Ralphson method. We use the Nelder-Mead algorithm to minimize $\sum_{k=1}^{p} | \sum_{j=1}^{n} w_j g^{-1}(q_j \lambda' x_j) x_{jk} - t_k |$ over λ, forcing each component down to zero, to get $\hat{\lambda}$. The weights are then,

$$\tilde{w}_j = w_j g^{-1}(q_j \hat{\lambda}' x_j), j = 1, \ldots, n.$$

In our case, we use the simple Euclidean distance. This gives close form answers; apparently this was not recognized by Haziza and Beaumont (2017). We choose $q_j = 1, j = 1, \ldots, n$. For $G(u) = (u-1)^2, -\infty < u < \infty$, is a legitimate distance function because $G(1) = 0$, $g(u) = 2(u-1)$, $g(1) = 0$, $g'(u) = 2 > 0$ and so $G(u)$ is strictly convex. Also $g^{-1}(y) = 1 + \frac{y}{2}$. In this case, we have

$$\sum_{j=1}^{n} w_j(1 + \frac{\lambda' x_j}{2}) x_{jk} = t_k, k = 1, \ldots, p.$$

That is,

$$\sum_{k'=1}^{p} \sum_{j=1}^{n} \lambda_{k'} x_{ik'} x_{ik} = 2(t_k - \sum_{j=1}^{n} w_j x_{jk}), k = 1, \ldots, p,$$

and

$$A\lambda = b, \quad A = \left(\sum_{j=1}^{n} w_j x_{jk'} x_{ik} \right)_{(k',k)}, \quad b = 2(t - \sum_{j=1}^{n} w_j x_j).$$

Therefore, assuming A is invertible, $\hat{\lambda} = A^{-1}b$ and the calibrated weights are $\tilde{w}_j = w_j(1 + \frac{\hat{\lambda}' x_j}{2}), j = 1, \ldots, n.$

5. Concluding Remarks

We have shown how to obtain $M = 1000$ copies of the contingency tables of the $12,234$ wards in the Nepal Census. We have actually studied the largest stratum among six strata of Nepal, Terrai Rural. Our procedure can be applied to each of the six strata in much the same way. No new methodology is needed. However, this matching procedure has been a very extensive endeavor. One might ask about a spatial analysis. One would like to incorporate a spatial effect among the households; unfortunately, there are no lists with the neighborhood of each household, but there is a list of the neighborhoods of the wards and the districts. Incorporating spatial effects at the district or ward level is not beneficial (i.e., too coarse).

It is possible to incorporate survey weights into our homogeneous and heterogeneous models. Nandram, Chen and Manandhar (2018) described one method to incorporate survey weights. Here, we suggest a simpler method, where our procedure described in this proposal can be used almost immediately. Let w_{ijk} denote the standardized survey weights of the kth individual in the jth household in the ith ward, $j = 1, \ldots, m_{i \cdot}, k = 1, \ldots, n_{ij}, i = 1, \ldots, \ell$. Potthoff, Woodbury and Manton (1992) showed how to obtain the standardized weights under the assumption that the individual responses are independent but not identically distributed (each individual has a different mean and variance); so we apply the standardization within ward but the responses of the household members might be correlated. This result can be generalized easily to a cluster (i.e., intra-class correlation). Let I_{ijkt} denote which cell the kth individual falls in the jth household within i^{th} ward (i.e., $I_{ijkt} = 1$ if it falls in cell t and 0 for all other cells. Then, the adjusted cell counts are $\tilde{n}_{ijt} = \sum_{k=1}^{n_{ij}} I_{ijkt} w_{ijk}$. Our procedure can simply be applied immediately to these adjusted cell counts.

There are five relevant covariates that are normally used to study health status from the same NLSS survey. They are age, nativity, sex, area and religion and they are mostly binary. These health covariates are studied in Nandram, Chen, Fu and Manandhar (2018). Here, our procedure does not apply directly. But the $p_{ijk}^{(h)}$ iterates can be obtained by fitting a multinomial logit model similar to the homogeneous and the heterogeneous models. But this is not easy. Another way is to partition the age variable, say young and old (binary), thereby making six categorical variables to form $2^6 = 64$ cells and one can fit the heterogeneous model using the procedure in this paper. However, one danger is the sparseness of the table. But one can certainly eliminate some cells that are structural zeros.

Another problem is how to obtain a model that takes care of district estimates, a kind of benchmarking. Because the districts are very large, it easy to obtain a reliable estimate for each district. One can therefore benchmark the ward estimates or household estimates to the districts to obtain a more robust model. It is also possible to do a sub-sub-area model.

APPENDIX A: Bayesian BHF Model with Survey Weights

We state the conditional posterior densities from Molina, Nandram and Rao (2014). We find that the condition, $\varepsilon \leq \rho \leq 1 - \varepsilon$, where ε is a small positive constant, is not necessary. This eliminates the sensitivity analysis on ε.

Defining $w_i = \sum_{j=1}^{n_i} w_{ij}$,

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} w_{ij}\underset{\sim}{x}_{ij}}{w_i}, \quad \bar{y}_i = \frac{\sum_{j=1}^{n_i} w_{ij}y_{ij}}{w_i}.$$

If the w_{ij} are survey weights, $\sum_{j=1}^{n_i} w_{ij} = N_i$, the population size, and these w_{ij} must be normalized so that $w_i = \hat{n}_i$, the effective sample size; see Potthoff, Woodbury and Manton (1992). But note that these w_{ij} do not exist (not defined) for nonsamples. Therefore, for prediction something like surrogate sampling (Nandram 2007) has to be done. [In our application here, the survey weights are assumed to be unity.]

First,

$$v_i \mid \underset{\sim}{\beta}, \sigma^2, \rho, \underset{\sim}{y}_s \overset{ind}{\sim} \text{Normal}\left\{ \frac{\rho w_i}{\rho w_i + 1 - \rho}(\bar{y}_i - \bar{x}_i'\underset{\sim}{\beta}), \frac{\rho}{\rho w_i + 1 - \rho}\sigma^2 \right\}, i = 1, \ldots, \ell. \quad (\text{A.1})$$

Observe that the mean and variance of v_i are defined for all values of ρ in $[0,1]$. When $\rho = 0$, $v_i \mid \underset{\sim}{\beta}, \sigma^2, \rho, \underset{\sim}{y}_s$ is a point mass at 0 and when $\rho = 1$, $v_i \mid \underset{\sim}{\beta}, \sigma^2, \rho, \underset{\sim}{y}_s \overset{ind}{\sim} \text{Normal}(\bar{y}_i - \bar{x}_i'\underset{\sim}{\beta}, \sigma^2/w_i)$.

Second, letting

$$B = \sum_{i=1}^{\ell}\sum_{j=1}^{n_i} w_{ij}(\underset{\sim}{x}_{ij} - \bar{x}_i)(\underset{\sim}{x}_{ij} - \bar{x}_i)' + (1 - \rho)\sum_{i=1}^{\ell} \frac{w_i}{\rho w_i + 1 - \rho}\bar{x}_i\bar{x}_i'$$

$$\underset{\sim}{b} = \sum_{i=1}^{\ell}\sum_{j=1}^{n_i} w_{ij}(\underset{\sim}{x}_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i) + (1 - \rho)\sum_{i=1}^{\ell} \frac{w_i}{\rho w_i + 1 - \rho}\bar{x}_i\bar{y}_i,$$

$$\underset{\sim}{\beta} \mid \sigma^2, \rho, \underset{\sim}{y}_s \sim \text{Normal}(\hat{\underset{\sim}{\beta}}, \sigma^2 B^{-1}), \quad (\text{A.2})$$

where $\hat{\underset{\sim}{\beta}} = B^{-1}\underset{\sim}{b}$. Again, observe that the mean and variance of $\underset{\sim}{\beta}$ are defined for all values of ρ in $[0,1]$.

Third, letting

$$G = \sum_{i=1}^{\ell}\sum_{j=1}^{n_i} w_{ij}\{y_{ij} - \bar{y}_i - (\underset{\sim}{x}_{ij} - \bar{x}_i)'\hat{\underset{\sim}{\beta}}\}^2 + (1 - \rho)\sum_{i=1}^{\ell} \frac{w_i}{\rho w_i + 1 - \rho}(\bar{y}_i - \bar{x}_i'\hat{\underset{\sim}{\beta}})^2,$$

$$\sigma^{-2} \mid \rho, \underset{\sim}{y}_s \sim \text{Gamma}\left(\frac{n - p}{2}, \frac{G}{2}\right). \quad (\text{A.3})$$

Again, observe that $\sigma^{-2} \mid \rho, \underset{\sim}{y}_s$ is defined for all values of ρ in $[0,1]$.

Finally,

$$\pi(\rho \mid \underset{\sim}{y}_s) \propto (1 - \rho)^{\ell/2} \mid B \mid^{-1/2} G^{-(n-p)/2} \prod_{i=1}^{\ell}\left\{ \frac{w_i}{\rho w_i + 1 - \rho} \right\}^{1/2}, \quad (\text{A.4})$$

and again $\pi(\rho \mid y_s)$ is defined for all values of ρ in $[0,1]$. Therefore, the joint posterior density, $\pi(\underset{\sim}{\nu},\underset{\sim}{\beta},\sigma^2,\rho \mid y_s)$, obtained by multiplying (A.1), (A.2), (A.3) and (A.4), is defined for all values of $\underset{\sim}{\rho}$ in $[0,1]$.

Acknowledgment

REFERENCES

Agresti, A. (2012), *Categorical Data Analysis*, 3rd Edition, New York: Wiley.

Central Bureau of Statistics Thapathali, Nepal living Standards Survey 2003/04, *Statistical Report Volume 1*, Kathmandu, Nepal.

Chakraborty, A., Datta, G. S. and Mandal, A. (2019), A Robust Hierarchical Bayes Small Area Estimation for Nested Error Linear Regression Model. *International Statistical Reviews*, 87, S1, S158-S156.

Chambers, R. and Tzavidis, N. (2006), "M-Qunatile Models for Small Area Estimation," *Biometrika*, 93, 255-268.

Connor, R. J. and Mosimann, J. E. (1969), "Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution," *Journal of the American Statistical Association*, 64, 194-206.

Corral, P., Molina, I. and Nguyen, M. (2020), "Pull Your Small Area Estimates up by the Bootstraps," *Policy Research Working Paper, World Bank Group, Poverty and Equity Global Practice*, pg 1-52.

Darroch, J. N. and Ratcliff, D. (1971), "A Characterization of the Dirichlet Distribution," *Journal of the American Statistical Association*, 66, 641-643.

Dawber, J. and Chambers, R. (2019), "Modelling Group Heterogeneity for small Area Estimation Using M-quantiles," *International Statistical Review*, 87, S1, S50-S63.

Gershunskaya, J. and Lahiri, P. (2018). Robust empirical best small area finite population mean estimation using a mixture model. *Calcutta Statistical Association Bulletin*, 69 (2), 183-204.

Haziza, D. and Beaumont, J-F. (2017), Construction of Weights in Surveys: A Review, *Statistical Science*, 32 (2), 206-226.

Molina, I., Nandram, B. and Rao, J. N. K. (2014), "Small Area Estimation of General Parameters with Application to Poverty Estimators: A Hierarchical Bayesian Approach," *Annals of Applied Statistics*, 8, 852-885.

Nandram, B., Choi, J. W. (2010), "A Bayesian Analysis of Body Mass Index Data From Small Domains Under Nonignorable Nonresponse and Selection," *Journal of the American Statistical Association*, 105, 120-135.

Nandram, B. (1998), "A Bayesian Analysis of the Three-stage Hierarchical Multinomial Model," *Journal of Statistical Computation and Simulation*, 61, 97-126.

Nandram, B. and Sedransk, J. (1993), "Bayesian Predictive Inference for a Finite Population Proportion: Two-stage Cluster Sampling," *Journal of the Royal Statistical Society*, Series B, 55, 399-488.

Nandram, B., Chen, L., Fu, S. and Manandhar, B. (2018), "Bayesian Logistic Regression for Small Areas with Numerous Households," *Statistics and Application* (In honor of JNK Rao), 16, 171-205.

Nandram, B., Chen, L. and Manandhar, B. (2018), Bayesian Analysis of Multinomial Counts from Small Areas and Sub-areas. *In JSM Proceedings, Section on Bayesian Statistical Science*, Alexandria, VA: American Statistical Association, 1140-1162.

Potthoff, R. F., Woodbury, M. A. and Manton, K. G. (1992), "Equivalent Sample Size" and "Equivalent Degrees of Freedom" Refinements for Inference Using Survey Weights Under Superpopulation Models. *Journal of the American Statistical Association*, 87 (418), 383-396.

Toto, M. C. S. and Nandram, B. (2010), "A Bayesian Predictive Inference for Small Area Means Incorporating Covariates and Sampling Weights," *Journal of Statistical Planning and Inference*, 140, 2963-2979.

Figure 2: Comparison of the homogeneous model and the heterogeneous model under the NER method for prediction of the finite population proportions (posterior means) for all wards
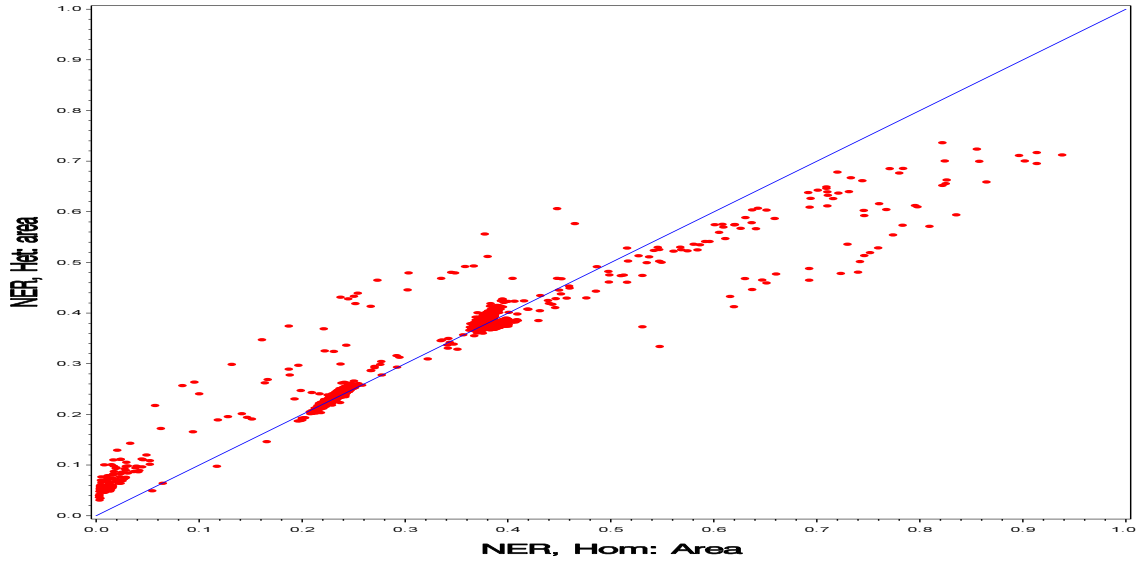

Figure 3: Comparison of the homogeneous model and the heterogeneous model for prediction of the finite population proportions (posterior means) for all wards under the IRLS method
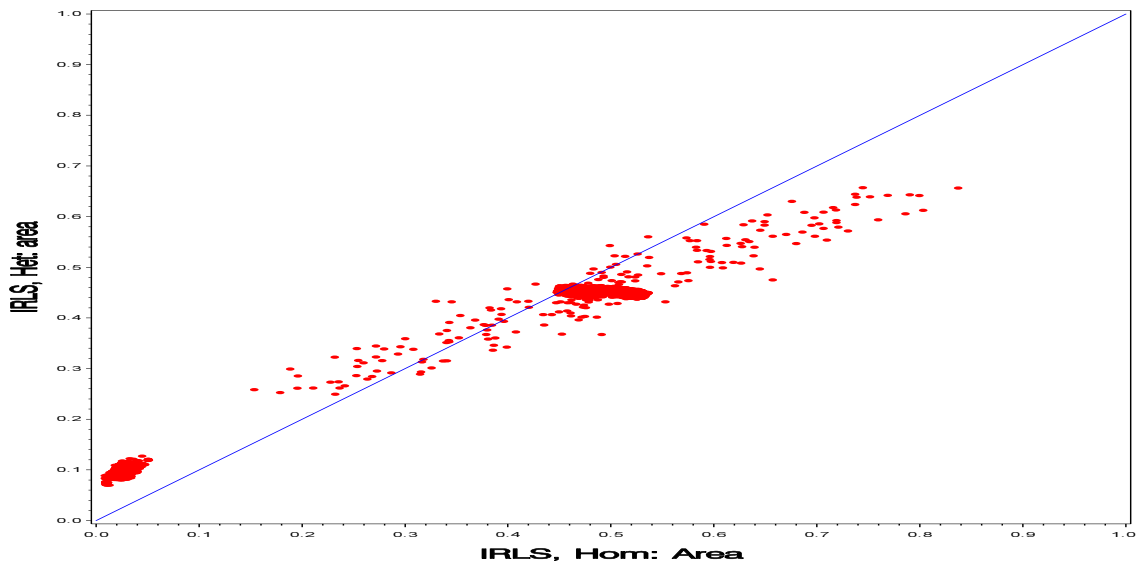
Figure 4: Comparison of the NER and IRLS projection methods under the homogeneous model prediction of the finite population proportions (posterior means) for all wards
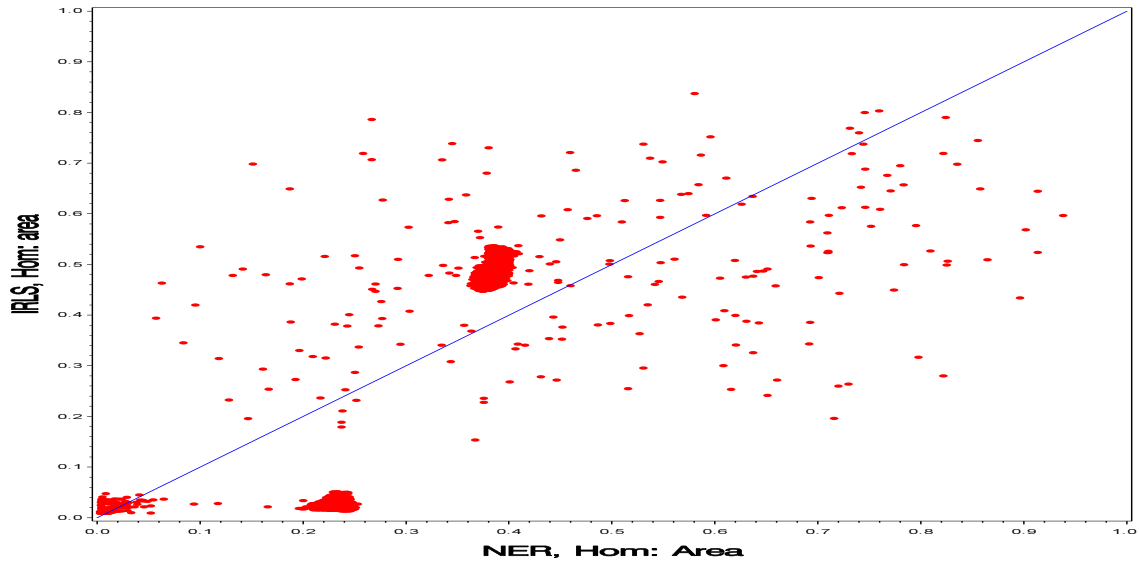

Figure 5: Comparison of the NER and IRLS projection methods under the heterogeneous modelfor prediction of the finite population proportions (posterior means) for all wards
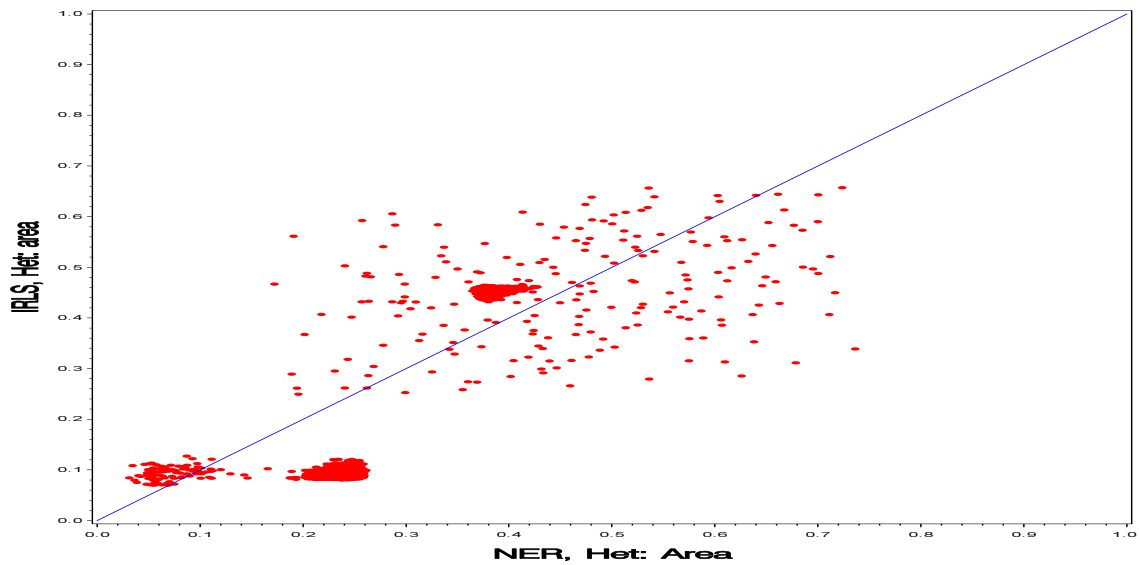
Figure 6: Comparison of the homogeneous model (NER projection method) and the heterogeneous model (IRLS projection method) for prediction of the finite population proportions (posterior means) for all wards
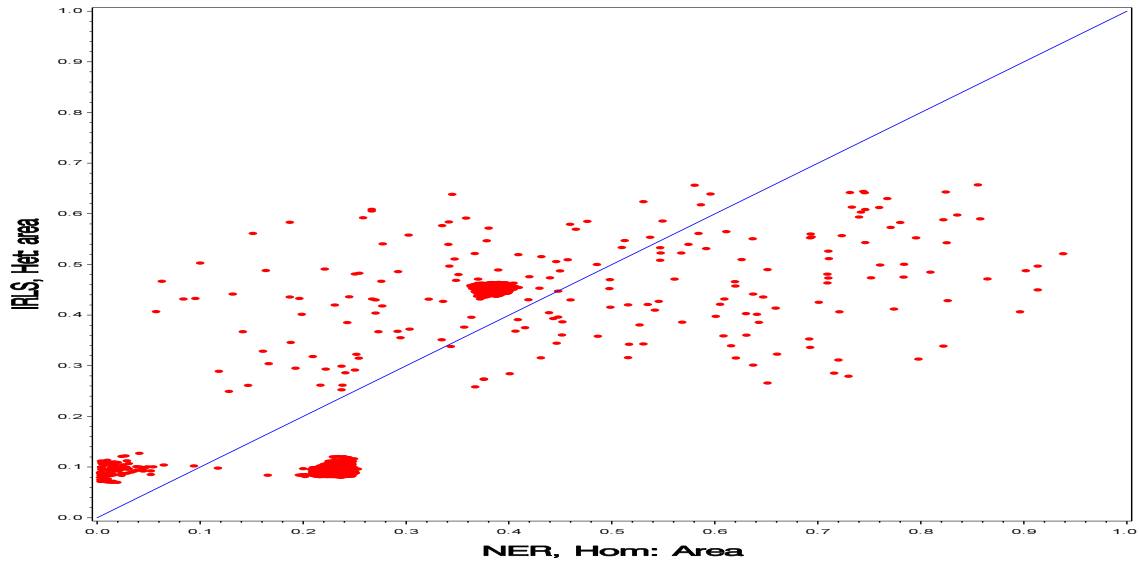

Figure 7: Comparison of the heterogeneous model (NER projection method) and the homogeneous model (IRLS projection method) for prediction of the finite population proportions (posterior means) for all wards
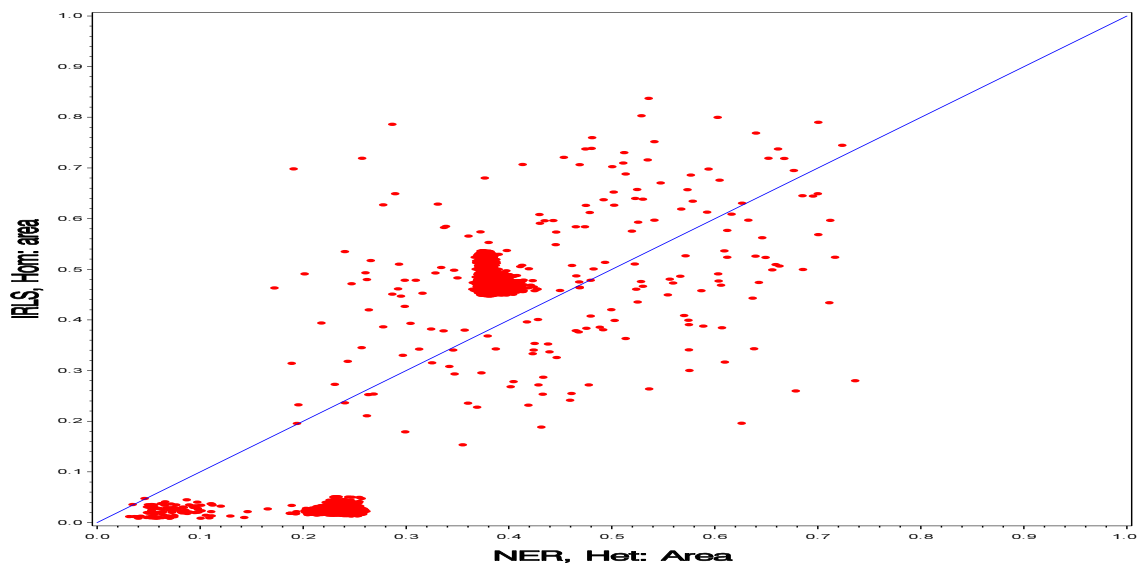
Table 1: Comparison of IRLS and the NER projective methods, the sub-area (Het) and the area-level (Hom) models with respect to posterior inference about the finite population proportions by health classes (three proportions)

	IRLS, Het	IRLS, Hom	NER, Het	NER, Hom
a. Sample only				
PM	.540	.608	.512	.574
	.375	.379	.425	.406
	.095	.022	.069	.015
PSD	.078	.073	.077	.031
	.071	.068	.074	.031
	.038	.008	.038	.006
b. Nonsample only				
PM	.453	.512	.378	.389
	.454	.466	.382	.375
	.093	.022	.241	.237
PSD	.138	.185	.046	.053
	.138	.184	.045	.052
	.040	.013	.036	.042
c. Population (both sample and nonsample)				
PM	.453	.511	.377	.388
	.454	.466	.382	.375
	.093	.023	.241	.236
PSD	.138	.185	.046	.053
	.138	.185	.045	.052
	.040	.013	.036	.042

NOTE: Entries are the medians over the 101 sampled wards in (a), 12,133 nonsampled wards in (b), and 12,234 wards in (c) for the three health classes.