Bayesian Analysis of Multinomial Counts from Small Areas and Sub-areas

Balgobin Nandram* Lu Chen† Binod Manandhar‡

Abstract

A standard problem in official statistics is to predict the finite population proportion of a small area when individual-level data are available from a survey and more extensive data (covariates but not responses) are available from a census. We can match the geographical location between sample and census but households are not matched and and the covariates in the sample and the census are different. The 2003-2004 Nepal Living Standards Survey, the second of its kind, and the 2001 census provide an example in which a PPS sample of the wards/sub-wards (PSUs) is selected and a systematic sample of the households within the wards is selected. In the largest stratum less than one percent of the wards and 12 households within the sampled wards are selected. We are interested in the health portion of the survey in which each individual in a household is categorized into one of four health status. Using a two-stage procedure, we study the counts in the households within the wards and a projection method to infer about the nonsampled households and wards. This is accommodated by a four-stage hierarchical Bayesian model for multinomial counts as it is necessary to accommodate heterogeneity (ie., differences in wards and households). To fit the model, we compare two computational methods, an approximate method and an exact method, that are used to obtain the distributions of the proportions in each health status, and then we use this distribution to do projective inference for the finite population proportions. In addition, we compare the heterogeneous model, with household effects, and a homogeneous model, without household effects, and two projection procedures (nonparametric and parametric).

Key Words: Approximation, Bayesian predictive inference, Dirichlet distribution, Hierarchical Bayesian model, Metropolis sampler, Numerical integration, Parallel computation

1. Introduction

One can observe counts in several areas. For example, in a study of 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 for four health status groups. We want to predict the finite population proportion of people in each 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 (see Rao and Molina 2015).

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 individuals partitioned into

*Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609
†Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609
‡Department of Mathematics, University of Houston, Houston, TX 77601

c cells. There are M_i sub-areas (households) within the i^{th} 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., $n_{ij\cdot} = \sum_{k=1}^{c} n_{ijk}$).

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. This model was originally used for cluster sampling generalizing the earlier work of Nandram and Sedransk (1993) for binomial data. Nowadays, we use this model as an area model in small area estimation with various applications. Nandram (1998) showed how to fit the area level model using a Metropolis-Hastings sampler, which is now considered a much too complicated algorithm for this problem.

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

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

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 scenario for some cells. So that this posterior density is well defined for all ρ, $0 \le \rho \le 1$. It is easy to use the Gibbs sampler, not the Metropolis-Hastings sampler, 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 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, $\underset{\sim}{p}_{ij} \mid \underset{\sim}{\mu}, \tau, \underset{\sim}{n}_{ij} \overset{ind}{\sim} \text{Dirichlet}(\underset{\sim}{n}_{ij} + \underset{\sim}{\mu}\tau), j = 1, \ldots, m_i, i = 1, \ldots, \ell$.

In the same manner, we can consider a sub-area homogeneous model for the sampled counts. Assuming that there are no area effects (i.e., the sub-areas are homogeneous), let

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

$$\underset{\sim}{p}_{ij} \mid \underset{\sim}{\psi}, \eta \overset{ind}{\sim} \text{Dirichlet}(\underset{\sim}{\psi}\eta),$$

$$\pi(\underset{\sim}{\psi}, \eta) = \frac{(c-1)!}{(1+\eta)^2}, \sum_{k=1}^{c} \psi_k = 1, \psi_k > 0, \eta \geq 0.$$

Now, the $\underset{\sim}{n}_{ij}$ are the sufficient statistics and there are no reductions. Under the model,

$$\pi(\underset{\sim}{\psi}, \eta \mid n) \propto \prod_{i=1}^{\ell} \prod_{j=1}^{m_i} \left\{ \frac{\prod_{k=1}^{c} \prod_{s=0}^{n_{ijk}-1} [\rho s + (1-\rho)\psi_k]}{\prod_{s=0}^{n_{ij\cdot}-1} [\rho s + 1 - \rho]} \right\}$$

where η has been transformed to $\rho = \frac{1}{1+\eta}$ and $\underset{\sim}{p}_{ij} \mid \underset{\sim}{\psi}, \eta, \underset{\sim}{n} \overset{ind}{\sim} \text{Dirichlet}(\underset{\sim}{n}_{ij} + \underset{\sim}{\psi}, \eta)$. Because there are numerous $\underset{\sim}{p}_{ij}$ (1212 in the NLSS data for the largest stratum), ($\underset{\sim}{\psi}$, η) will have very small variation and the $\underset{\sim}{p}_{ij}$ will be to close together (no ward effects), thereby loosing the practical import of the application.

We will obtain a sub-area heterogeneous model for the multinomial counts in which there both are sub-area effects and area effects. In the Nepal Living Standards Survey (NLSS), the covariates are different from those in the Census and they do not affect health directly, so that we will call them indirect covariates, nine covariates commonly used for any study 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. 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 Section 4.1 for these covariates). 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 three variables and so on. At the end there were some variables that did not match. For these we use a nearest neighbor matching procedure using 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 nonsampled households of the 101 sampled wards and all the households of the 12,133 nonsampled wards.

Our entire 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. Because for a large number of areas this procedure is time consuming, we also show how to make an approximation to 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 Reweighted 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 four-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. Section 5 has concluding remarks. Technical details are given in the appendices.

2. Four-Stage Hierarchical Bayesian Models

We extend the three-stage hierarchical Bayesian model of Nandram (1998) to four 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 use a single τ (i.e., not depending on area).

Note that if τ depends on area (i.e., τ_i is used) and we allow τ_i to go to infinity, the first homogeneous model occurs and if we allow η to go to infinity, we retrieve the second homogeneous model. For comparison, we need to fit the joint posterior densities under the first homogeneous (area-level) model and heterogeneous model separately. Inference about the $\underset{\sim}{p}_{ij}$ and the μ_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 $\frac{\prod_{k=1}^{c} \mu_{ik}^{\psi_k \eta - 1}}{D(\underset{\sim}{\psi}\eta)}$, the Dirichlet prior distribution. We can think of the posterior density $\pi(\underset{\sim}{\mu}_i, \tau \mid \underset{\sim}{n}_i, \underset{\sim}{\psi}, \eta)$ as the product of two densities, one proportional to $g(\underset{\sim}{\mu}_i, \tau \mid \underset{\sim}{n}_i)$, and $\frac{\prod_{k=1}^{c} \mu_{ik}^{\psi_k \eta - 1}}{D(\underset{\sim}{\psi}\eta)}$. In fact, we can think of $g(\underset{\sim}{\mu}_i, \tau \mid \underset{\sim}{n}_{ij})$, 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) obtained a Metropolis-Hastings sampler to draw samples from; see Appendix A.

Note that $\underset{\sim}{\psi}$ and η are not directly connected to the counts even after integrating out the $\underset{\sim}{p}_{ij}$. This indicates that there will be difficulties (weak mixing) in running a Markov chain Monte Carlo sampler. Therefore, further integration is necessary. That is, we also need to integrate out the $\underset{\sim}{\mu}_i$ (i.e., need to draw the parameters simultaneously), but this is not possible analytically.

2.2 Sampling the Joint Posterior Density

We will describe two methods for sampling the joint posterior density. In Section 2.2.1, we describe an analytical approximation based on Nandram (1998) and in Section 2.2.2, a numerical integration based on a characterization of the Dirichlet distribution.

2.2.1 Analytical Approximation

In Appendix A, we show how to approximate $g(\underset{\sim}{\mu}_i, \tau \mid \underset{\sim}{n}_i)$ as

$$g(\underset{\sim}{\mu}_i, \tau \mid \underset{\sim}{n}_i) \approx g_a(\underset{\sim}{\mu}_i \mid \tau, \underset{\sim}{n}_i) g_b(\tau \mid \underset{\sim}{n}_i)$$

where $g_a(\underset{\sim}{\mu}_i \mid \tau, \underset{\sim}{n}_i)$ is a Dirichlet distribution, depending on the cell counts and τ, and $g_b(\tau \mid \underset{\sim}{n}_i)$ is a gamma distribution, depending only on the cell counts. Note specifically that $g_a(\underset{\sim}{\mu}_i \mid \tau, \underset{\sim}{n}_i)$ is

Dirichlet($\underset{\sim}{\phi}_i\tau + \underset{\sim}{j}$), where $\underset{\sim}{j}$ is a vector or ones, $\underset{\sim}{\phi}_i$ depends on τ and the cell counts, and $g_b(\tau \mid \underset{\sim}{n}_i)$ is Gamma(η_i, ν_i), where η_i and ν_i depend on the cell counts. With this approximation, we can integrate out $\underset{\sim}{\mu}_i$, but not τ. So we will be left with an approximate posterior density $\pi_a(\underset{\sim}{\psi}, \eta, \tau \mid \underset{\sim}{n})$. The parameters $(\underset{\sim}{\psi}, \eta, \tau)$ are now connected to the cell counts, and this provides a better sampler (but still can be improved).

Using this approximation, the approximate joint posterior density of $\underset{\sim}{\mu}, i = 1, \ldots, \ell, \underset{\sim}{\psi}, \eta, \tau$ is

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

This approximation allows us to integrate out the $\underset{\sim}{\mu}_i$ to get

$$\pi_a(\underset{\sim}{\psi}, \eta, \tau \mid \underset{\sim}{n}) \propto \frac{1}{(1+\tau)^2} \frac{1}{(1+\eta)^2} \prod_{i=1}^{\ell} g_b(\tau \mid \underset{\sim}{n}_i) \frac{D(\underset{\sim}{\phi}_i \tau + \underset{\sim}{\psi}\eta)}{D(\underset{\sim}{\phi}_i \tau) D(\underset{\sim}{\psi}\eta)}.$$

Finally, we can now draw samples of $\underset{\sim}{\psi}, \eta, \tau$ using the griddy Gibbs sampler. It is convenient to transform η and τ to (0,1) and taking $\gamma_1 = 1/(1+\eta)$ and $\gamma_2 = 1/(1+\tau)$ to get

$$\pi_a(\underset{\sim}{\psi}, \gamma_1, \gamma_2 \mid \underset{\sim}{n}) \propto \left\{ \prod_{i=1}^{\ell} g_b(\tau \mid \underset{\sim}{n}_i) \frac{D(\underset{\sim}{\phi}_i \tau + \underset{\sim}{\psi}\eta)}{D(\underset{\sim}{\phi}_i \tau) D(\underset{\sim}{\psi}\eta)} \right\}_{\eta = \gamma_1^*, \tau = \gamma_2^*},$$

where $\gamma_1^* = \frac{1-\gamma_1}{\gamma_1}$ and $\gamma_2^* = \frac{1-\gamma_2}{\gamma_2}$.

2.2.2 Numerical Integration in the Sub-area Model

Another way to integrate out the $\underset{\sim}{\mu}_i$ 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 $\underset{\sim}{x}$ be a vector with c components such that $\sum_{j=1}^{c} x_i = 1$ $x_j \geq 0, j = 1, \ldots, c$. Assume that $\underset{\sim}{x} \sim$ Dirichlet($\underset{\sim}{a}$), where $\underset{\sim}{a}$ is a vector with c known elements. Then,

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

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

Let $\underset{\sim}{u}$ have a Dirichlet distribution, and $h(\underset{\sim}{u})$ be a function of $\underset{\sim}{u}$. Then, letting $\mathscr{C} = \{\underset{\sim}{x} : x_j \geq 0, j = 1, \ldots, c-1, \sum_{j=1}^{c-1} x_j < 1\}$, we require the integral,

$$I = \int_{\underset{\sim}{x} \in \mathscr{C}} h(\underset{\sim}{x}) p(\underset{\sim}{x}) d\underset{\sim}{x}.$$

Then, transforming $x_1, \ldots, x_{c-1}$ (as is done above), we have

$$I = \int_0^1 \ldots \int_0^1 \{\prod_{j=1}^{c-2}(1-u_j)^{c-j-1}\} h^*(\underset{\sim}{u}) p(\underset{\sim}{u}) d\underset{\sim}{u},$$

where $u_1, \ldots, u_{c-1}$ are independent beta random variables, $h(\underset{\sim}{x})$ becomes $h^*(\underset{\sim}{u})$ and $\mathscr{C}$ transforms to $[0,1]^{c-1}$.

Using this characterization on the $\underset{\sim}{\mu}_i \mid \underset{\sim}{\psi}, \eta \overset{ind}{\sim} \text{Dirichlet}(\underset{\sim}{\psi}\eta)$, we get

$$\pi(\underset{\sim}{\psi}, \eta, \tau \mid \underset{\sim}{n}) \propto \frac{1}{(1+\tau)^2} \frac{1}{(1+\eta)^2} \prod_{i=1}^{\ell} \int g^*(\underset{\sim}{y}_i, \tau \mid \underset{\sim}{n}_{ij}) \prod_{k=1}^{c-2} (1-v_{ik})^{c-k-1} \pi(\underset{\sim}{y}_i) d\underset{\sim}{y}_i,$$

where $v_{ik}, k = 1, \ldots, c-2$, are independent beta random variables for $i = 1, \ldots, \ell$. The integration is easy to carry out by discretization over the range of the independent beta random variables.

We can now draw $\underset{\sim}{\psi}, \eta, \tau$ using a Metropolis sampler. The candidate generating is the analytic approximation discussed in the previous section. 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 show how to construct a proposal density for $(\underset{\sim}{\psi}, \eta, \tau)$ to run the Metropolis sampler. We have samples from the approximate posterior density of $(\underset{\sim}{\psi}^{(h)}, \eta^{(h)}, \tau^{(h)}), h = 1, \ldots, M$. We transform these to $\underset{\sim}{\beta}^{(h)}, h = 1, \ldots, M$ 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 $\underset{\sim}{\beta}^{(h)}$, where $\hat{\underset{\sim}{\theta}}$ and $\hat{\Sigma}$ are the mean and covariance matrix of the samples, and $\kappa/\sigma^2 \sim \text{Gamma}(\kappa/2, 1/2)$ 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. 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 $\underset{\sim}{\mu}_i$, we use a Metropolis algorithm with the approximate Dirichlet distribution (Nandram 1998) as the proposal density to draw samples of $\underset{\sim}{\mu}_i$ independently given $\underset{\sim}{\psi}, \eta, \tau$ and data. We run each Metropolis 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 $(.25, .50)$), 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 computing is available.

3. Inference for Finite Population Proportions

We have now obtained samples $\underset{\sim}{p}_{ij}^{(h)} = (p_{ijk}, k = 1, \ldots, c = 3), j = 1, \ldots, m_i, i = 1, \ldots, \ell, h = 1, \ldots, M$, say $M = 1000$. The next step is to link these $\underset{\sim}{p}_{ij}^{(h)}$ to the census. The census has covariates $\underset{\sim}{x}_{ij}$, 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},

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

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

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 $\underset{\sim}{p}_{ij}^{(h)}$. We can obtain all the nonsampled p_{ijk} for all M iterates in the same manner. Having obtained the 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 (2018). This is directly related to the procedure mentioned above.

Let y_{ij} denote the responses and $\underset{\sim}{x}_{ij}$ 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 $\underset{\sim}{x}_i$ for $i = 1, \ldots, n = \sum_{i=1}^{\ell} n_i$. In the IRLS, $\underset{\sim}{\hat{\beta}}_q$ for each q in $(0,1)$ is obtained by solving

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

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

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

with $\psi_{q\kappa}(u) = 2[(1-q)I_{u\leq 0} + qI_{u>0}][-\kappa I_{u\leq -\kappa} + uI_{-\kappa<u<\kappa} + \kappa I_{u\geq\kappa}]$, where 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 i^{th} 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 squares 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 squares 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, and $\varepsilon < \rho < 1-\varepsilon$, where ε is a very small number. For example, $\varepsilon \approx .0001$, and Molina, Nandram and Rao (2014) showed posterior inference about poverty parameters is not sensitive to a choice of a small ε.

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 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 for each ward or even at a more micro level for each household. Predictive inference means that the non-sampled 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 data 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 $\underset{\sim}{x}'_{ij}\underset{\sim}{\beta} + v_i$ for the nonsampled households within sampled wards via the posterior samples of the v_i (already drawn). For the nonsampled wards, we will need to use a draw from the the prior distribution of v_i (no-data posterior density) and compute $\underset{\sim}{x}'_{ij}\underset{\sim}{\beta} + v_i$.

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 the wards are scattered in the whole strip.

Figure 1: Map of Nepal showing the Terai Region

4.1 Census and NLSS

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 of medical facilities, expenditures on them and health status. Health status questionnaire is asked for every individual that was covered in the survey. The health status 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 II, 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 II 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.

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. These binary variables are nativity (Indigenous = 1, Non-indigenous = 0), religion ((Hindu = 1, Non-Hindu = 0), sex (Male = 1, Female = 0) and area (Urban = 1, Rural = 0). Older age and child age are more vulnerable than younger age. Indigenous people can have different health status from non-indigeneous people. Similarly, health status of urban and rural citizens could be different. Unfortunately, these covariates are at the individual level, whereas the available covariates in the census are at the household and ward level. This is one of the reasons that we did not use the covariates at the individual level for the sample.

We chose nine relevant covariates which can possibly influence health and they are available in the 2001 census data. They are (i) "Household size" (*hhsize*), (ii) "proportion of kids aged 0 - 6 in the household" (*skids6*), (iii) "proportion of kids aged 7 - 14 in the household" (*skids714*), (iv) "abroad migrant" (*remtab*), (v) "House temporary" (*hutype3*), (vi) "House owned" (*huown2*), (vii) "proportion of households with cooking fuel LP/gas in Ward" (*ckfuel3w*), (viii) "proportion of households with land-owning females in municipality/VDC" (*pflandv*), and (ix) "proportion of kids 6-16 attending school in municipality/VDC" (*pschv*) from NLSS-II, 2003–2004.

Six of these covariates are directly related to the household: *hhsize, skids6, skids714, remtab, hutype3,* and *huown2*. Size of household and proportion of children in different age group have influence on expenditure and consumption. Any household member as abroad migrant indicates the sources of remittance. Covariate *"hutype3"* indicates a temporary type of house; there are three types of houses according to the construction material of the outside walls of the house: permanent, semi-permanent, and temporary. The house owned variable is a binary variable that indicates whether the household has its own house to live in or not. Covariate *"huown2"* indicates that the household has their own house to live in.

The other three area level variables are *ckfuel3w, pflandv,* and *pschv*. Around 2003–2004 in Nepal very few rural areas used LP/gas as cooking fuel while urban households did. LP/gas is expensive compared to other sources of fuel. Covariate *"pflandv"* indicates the proportion of households with a female owning land in the Municipality/VDC. The proportion of children aged 6–16, *pschv*, who are supposed to be in school indicates the awareness of the community and strength of the future.

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. 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.

4.2 Data Analysis

Because the counts in the households are sparse for many households, 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 densitied (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. See Appendix B, where we show 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.

In Figures 2, 3 and 4, we compare the approximate method with the exact method. We can see that the posterior means are very close, the posterior standard deviations and posterior coefficients of variation are close but a little less close than the posterior means. As we can see, R^2 for the three plots (PM2 vs. PM1, .9993; PSD2 vs, PSD1, .9749; PCV2 vs. PCV1, .9942) are very high, very close to the 45^o straight lines.

Figure 2: Posterior means - Approximate versus exact for all sampled households

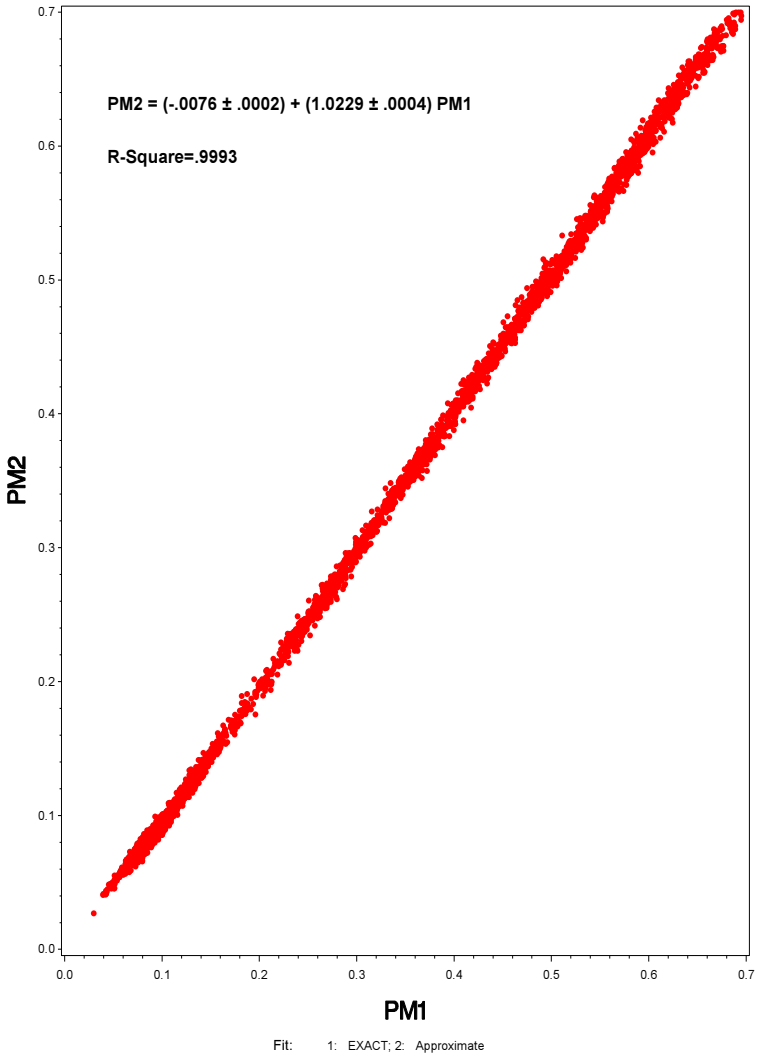


Figure 3: Posterior standard deviations - Approximate versus exact for all sampled households

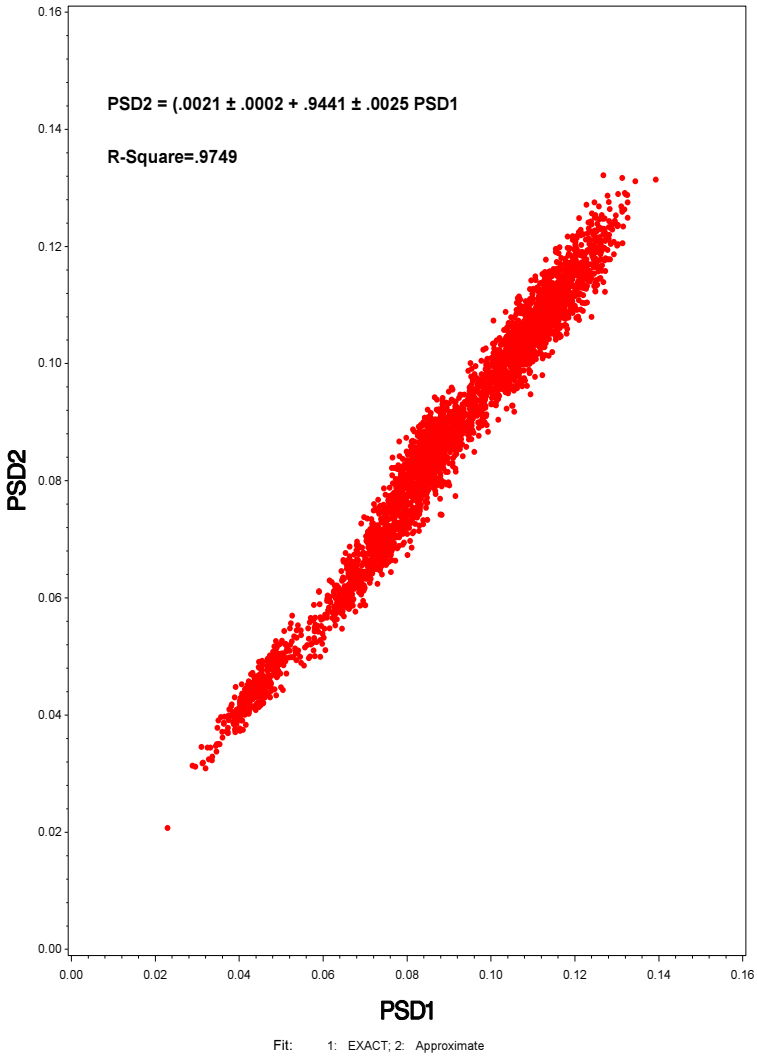
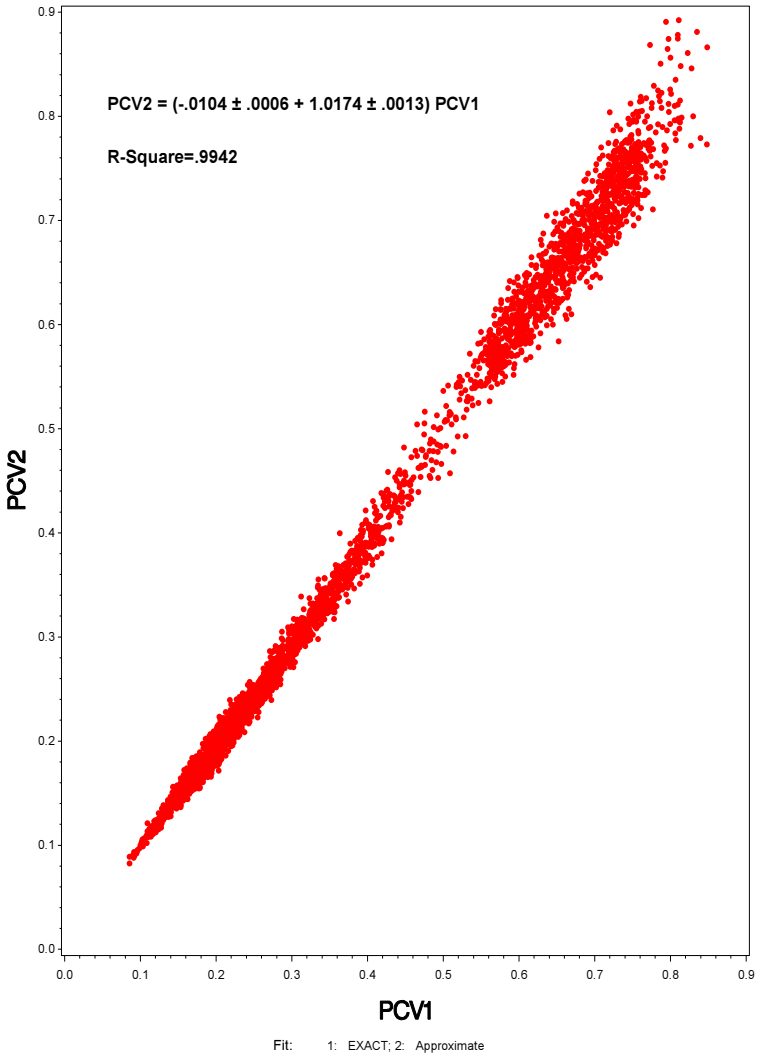


Figure 4: Posterior coefficients of variation - Approximate versus exact for sample households



In Figures 5 and 6, we compare the homogeneous model and the heterogeneous model under the NER method and the IRLS method respectively. We can see that the points do not fall reasonably well on the 45^o straight line. This indicates, everything being equal, the homogeneous model is inadequate. A shortcoming of the homogeneous model is that it does not model sub-areas, an additional source of variability that is accommodated by the sub-area model.

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

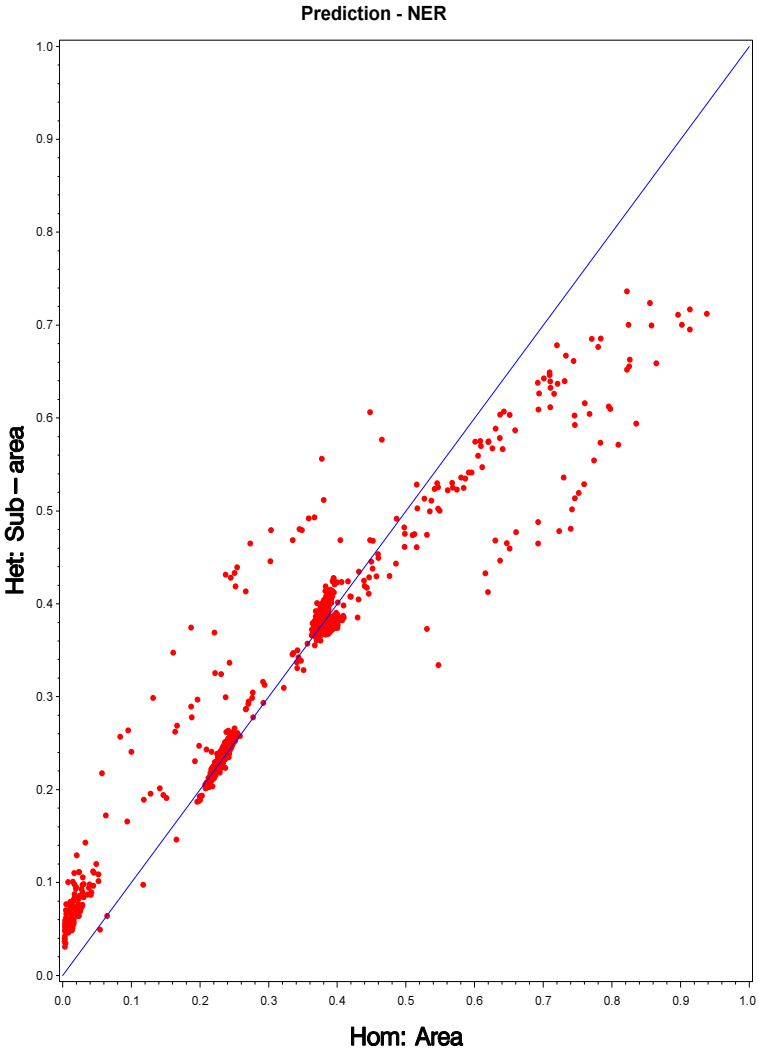
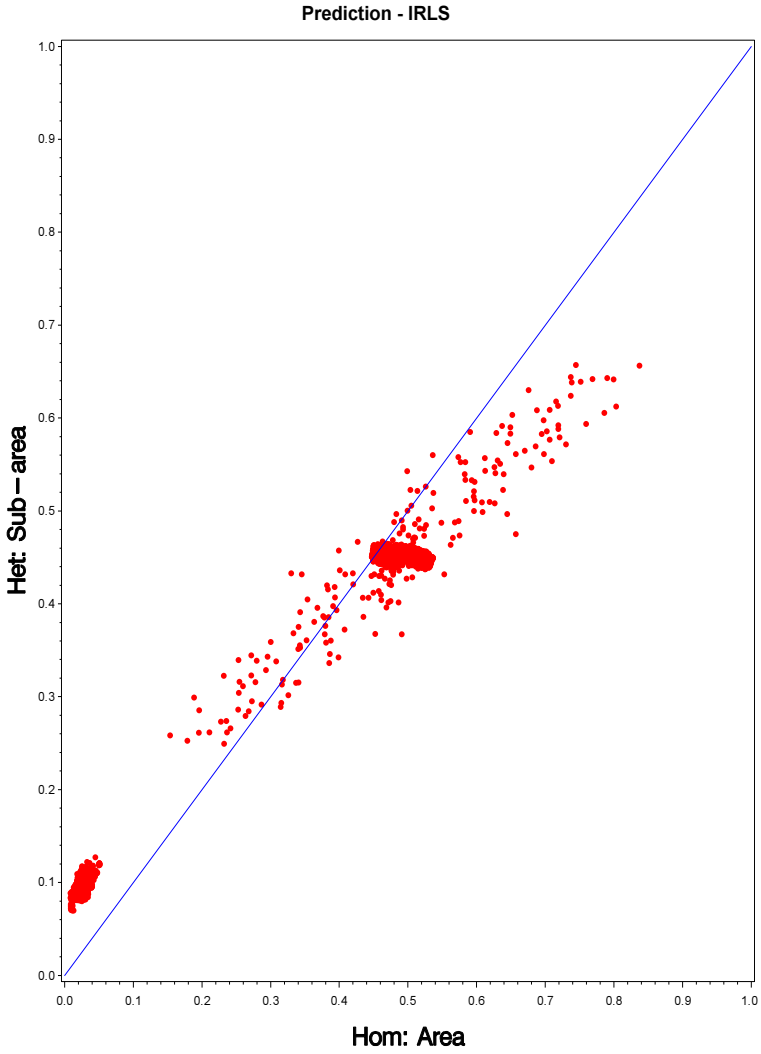


Figure 6: Comparison of the homogeneous model and the heterogeneous model for prediction of the finite population proportions (posterior means) for all wards under the IRLS method



Finally, in Table 1, we compare the IRLS and the NER model under the sub-area model (Het) and the area-level model (Hom) with respect to posterior inference about the finite population proportion by health class. We have used the posterior means and the posterior standard deviations for this exercise, and we have summarized the 12,334 by their medians and inter-quartile range for each of the three finite population proportions. If we consider only the proportion of people in excellent health, we can see some interesting things. For the sample, the medians of the proportions under Hom are always larger regardless of whether IRLS or NER is used; the IQR of the proportions follow the same pattern. [Note the large IQRs under NER (Hom).] The same results hold

for the nonsamples; here the IQRs are much smaller perhaps because the nonsampled covariates are more homogeneous. We can see also that these proportions are larger for the samples than the nonsamples perhaps reflecting a possible selection bias. Also, for the sample (nsample), the PSDs under Hom are always smaller (larger) than the corresponding ones under Het; these differences are small.

Table 1: Comparison of IRLS method and the NER projective method, the sub-area model (Het) and the area-level model (Hom) with respect to posterior inference about the finite population proportions by health status (three proportions)

	IRLS, Het		IRLS, Hom		NER, Het		NER, Hom	
	Med	IQR	Med	IQR	Med	IQR	Med	IQR
a. Sample								
PM	.540	.098	.608	.164	.512	.185	.574	.361
	.375	.105	.379	.159	.425	.208	.406	.361
	.095	.016	.022	.018	.069	.034	.015	.021
PSD	.078	.020	.073	.029	.077	.039	.031	.011
	.071	.011	.068	.027	.074	.037	.031	.010
	.038	.010	.008	.004	.038	.030	.006	.004
b. Nonsample								
PM	.453	.005	.512	.016	.378	.003	.389	.004
	.454	.004	.466	.012	.382	.008	.375	.004
	.093	.008	.022	.005	.241	.011	.237	.007
PSD	.138	.017	.185	.051	.046	.014	.053	.017
	.138	.017	.184	.051	.045	.014	.052	.017
	.040	.006	.013	.002	.036	.009	.042	.013

NOTE: Entries are the medians and IQRs over the 101 sampled wards and the 12,133 nonsampled wards for the three health status.

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.

It is not obvious how to incorporate survey weights in this procedure. A possible procedure is to use a multinomial logit model for the cell counts, and incorporate the survey weights in this multinomial regression using the individual health covariates. Then, we can proceed with the cell probabilities and the census covariates as we have done. This is not really easy for the sub-area model. To be specific, for unit k within sub-area j and area i, the model is

$$\underset{\sim}{y}_{ijk} \mid \underset{\sim}{\beta}, \nu_i, \omega_{ij} \overset{ind}{\sim} \text{Multinomial}(1, \underset{\sim}{\phi}_{ijk}),$$

where $\underset{\sim}{\beta} = (\underset{\sim}{\beta}_1', \ldots, \underset{\sim}{\beta}_{c-1}')'$ and $\underset{\sim}{\phi}_{ijk} = (\phi_{ijk1}, \ldots, \phi_{ijkc})'$ with

$$\phi_{ijkg} = \frac{e^{\underset{\sim}{x}_{ijk}'\underset{\sim}{\beta}_g + \nu_i + \omega_{ij}}}{1 + \sum_{g'=1}^{c-1} e^{\underset{\sim}{x}_{ijk}'\underset{\sim}{\beta}_{g'} + \nu_i + \omega_{ij}}}, g = 1, \ldots, c-1, \phi_{ijkc} = 1 - \sum_{g'=1}^{c-1} \phi_{ijkg'}.$$

Then, the random effects have

$$\nu_i \mid \sigma^2 \overset{ind}{\sim} \text{Normal}(0, \sigma^2), \;\; \omega_{ij} \mid \delta^2 \overset{ind}{\sim} \text{Normal}(0, \delta^2),$$

$$\pi(\underset{\sim}{\beta}) = 1, \;\; \pi(\sigma^2, \delta^2) = \frac{1}{(1+\sigma^2)^2} \frac{1}{(1+\delta^2)^2}.$$

One can pitch the survey weights to form a composite likelihood (preferably normalized); see Nandram, Chen, Fu and Manandhar (2018) for an individual area-level model without sub-area effects. Letting the standardized weights be λ_{ijk} and $\mathscr{C} = \{(a_1, \ldots, a_c) : a_g = 0, 1, \sum_{g=1}^{c} a_g = 1\}$, the normalized composite likelihood is

$$f(\underset{\sim}{y}_{ijk} \mid \underset{\sim}{\phi}_{ijk}) = \frac{\prod_{g=1}^{c-1} \phi_{ijkg}^{y_{ijkg}} (1 - \sum_{g'=1}^{c-1} \phi_{ijkg'})^{y_{ijkg}}}{\sum_{\underset{\sim}{y}_{ijk} \in \mathscr{C}} \prod_{g=1}^{c-1} \phi_{ijkg}^{y_{ijkg}} (1 - \sum_{g'=1}^{c-1} \phi_{ijkg'})^{y_{ijkg}}}, y_{ijkg} = 0, 1, \sum_{g=1}^{c} y_{ijkg} = 1.$$

It is, indeed, a challenge to fit this model to the NLSS data with the individual covariates and survey weights.

There are many problems like the one we just discussed. Surveys generally use stratified sampling with PPS sampling of households, schools or farms within each stratum and systematic sampling of equal size within each household, school or farm. While this leads to equal probability sampling, there are sampling adjustments that are usually made, thereby producing unequal survey weights per individual. One would need to add desperate administrative data, typically obtained from a census. Such large surveys are common. The covariates in the survey may be different from the census covariates, as in the Nepal application, poising additional challenges.

APPENDIX A: Approximation of Posterior Density of $(\underset{\sim}{\mu}, \tau)$

We show how to approximate

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

by a Dirichlet-Gamma distribution motivated by Nandram (1998). For ease of exposition, we will drop the subscript *i*, and we consider

$$g(\underset{\sim}{\mu}, \tau \mid \underset{\sim}{n}) = \prod_{s=1}^{S} \frac{D(\underset{\sim}{n}_s + \underset{\sim}{\mu} \tau)}{D(\underset{\sim}{\mu} \tau)},$$

where S is the number of households within a ward and $\underset{\sim}{\mu}$ and $\underset{\sim}{n}_s$ are vectors of length r. It also convenient to write $g(\underset{\sim}{\mu}, \tau \mid \underset{\sim}{n})$ as

$$g(\underset{\sim}{\mu}, \tau \mid \underset{\sim}{n}) = \prod_{s=1}^{S} \left\{ \left(\prod_{j=1}^{r} \frac{\Gamma(n_{sj} + \mu_j \tau)}{\Gamma(\mu_j \tau)} \right) / \frac{\Gamma(n_s + \tau)}{\Gamma(\tau)} \right\}.$$

Note that $g(\underset{\sim}{\mu}, \tau \mid \underset{\sim}{n})$ does not contain prior information about $\underset{\sim}{\mu}$ and τ(ie., prior distributions). We seek a convenient approximation of the posterior density $g(\underset{\sim}{\mu}, \tau \mid \underset{\sim}{n})$. Nandram (1998) obtained an approximation for the conditional density $\underset{\sim}{\mu} \mid \tau, \underset{\sim}{n}$ and one for $\tau \mid \underset{\sim}{\mu}, \underset{\sim}{n}$ which he used to obtain candidate generating densities to facilitate the execution of a Metropolis-Hastings sampler. Here, our purpose is different and we adapt the approximation of Nandram (1998) to facilitate a composition method.

First, we obtain the approximation for the posterior density of τ by $p_a(\tau \mid \underset{\sim}{n})$ starting with conditional posterior density of $\tau \mid \underset{\sim}{\mu}, \underset{\sim}{n}$. This approximation is a gamma density. Define $\hat{\theta}_j = \frac{n_{\cdot j}+1}{\sum_{j=1}^{r}(n_{\cdot j}+1)}, j = 1, \ldots, r$, where $n_{\cdot j} = \sum_{s=1}^{S} n_{sj}$. Then, an approximation to the posterior density, $\pi(\tau \mid \underset{\sim}{n})$, is

$$p(\tau \mid \underset{\sim}{\mu} = \underset{\sim}{\hat{\theta}}, \underset{\sim}{n}) \propto \prod_{s=1}^{S} \left\{ \prod_{j=1}^{r} \frac{\Gamma(n_{sj} + \hat{\theta}_j)}{\Gamma(\hat{\theta}_j \tau)} \right\} / \{\Gamma(n_{s\cdot} + \tau)/\Gamma(\tau)\}, \tau > 0.$$

Let τ_* denote the posterior mode of $p(\tau \mid \underset{\sim}{\mu} = \underset{\sim}{\hat{\theta}}, \underset{\sim}{n})$ or some reasonable estimator (e.g., the posterior median if the posterior mode does not exist). Nandram (1998) obtained the posterior mode using the Nelder-Mead algorithm, and we have used a similar approach here. Now, define

$$\sigma_*^2 = \left[\sum_{s=1}^{S} \left\{ (\frac{1}{\tau_*} - \frac{1}{\tau_* + n_{s\cdot}}) + \sum_{j=1}^{r} \hat{\theta}_j^2 (\frac{1}{\hat{\theta}_j \tau_*} - \frac{1}{n_{sj} + \hat{\theta}_j \tau_*}) \right\} \right]^{-1}.$$

Then, essentially equating moments,

$$\eta = \left\{ \frac{\tau_*}{2\sigma_*} + \sqrt{(\frac{\tau_*}{2\sigma_*})^2 + 1} \right\}^2 \text{ and } \nu = \sqrt{\eta}/\sigma_*, \tag{A.1}$$

and the gamma approximation is $\tau \mid \underset{\sim}{n} \sim \text{Gamma}(\eta, \nu)$. Once a deviate of τ is obtained, we can draw $\underset{\sim}{\mu}$ from $\underset{\sim}{\mu} \mid \tau, \underset{\sim}{n} \sim \text{Dirichlet}(\underset{\sim}{\phi}\tau)$. Next, our departure from Nandram (1998) is in the specification of $\underset{\sim}{\phi}$.

Let

$$A_j = \tau \sum_{s=1}^{S} \ln(1 + n_{sj}/\tau\hat{\theta}_j), \;\; B_j = \tau \sum_{s=1}^{S} \left\{ \hat{\theta}_j^{-1} - (\hat{\theta}_j + n_{sj}/\tau)^{-1} \right\}, j = 1, \ldots, r,$$

where we assume that $n_{sj} \geq 1$ for each j and at least one s (i.e., $B_j > 0$). Specifically, we are assuming that there is at least one positive entry across all tables. Next, we define

$$\bar{A} = \sum_{j=1}^{r} B_j^{-1} A_j / \sum_{j=1}^{r} B_j^{-1}$$

and for $j = 1, \ldots, r$, using Nandram (1998), we have

$$\hat{\tilde{\mu}}_j = \hat{\theta}_j + B_j^{-1}(A_j - \bar{A})$$

$$\tilde{\mu}_j = \begin{cases} \hat{\tilde{\mu}}_j, & 0 < \hat{\tilde{\mu}}_j < 1 \\ \hat{\theta}_j, & \hat{\tilde{\mu}}_j \leq 0 \text{ or } \hat{\tilde{\mu}}_j \geq 1. \end{cases}$$

Finally, we take

$$\phi_j = \tilde{\mu}_j / \sum_{j=1}^{r} \tilde{\mu}_j, j = 1, \ldots, r. \tag{A.2}$$

Observe that the ϕ_j are functions of τ. These posterior distributions are reasonable to use as importance functions.

As a summary, our approximation to $g(\underset{\sim}{\mu}, \tau \mid \underset{\sim}{n})$ is

$$g(\underset{\sim}{\mu}, \tau \mid \underset{\sim}{n}) \approx \pi_a(\underset{\sim}{\mu} \mid \tau, \underset{\sim}{n}) \pi_b(\tau \mid \underset{\sim}{n}),$$

where in $\pi_a(\underset{\sim}{\mu} \mid \tau, \underset{\sim}{n})$, $\underset{\sim}{\mu} \mid \tau, \underset{\sim}{n} \sim \text{Dirichlet}(\underset{\sim}{\phi}\tau + \underset{\sim}{j})$, $jtil$ a vector of ones and $\underset{\sim}{\phi}$ comes from (A.2), and $\tau \mid \underset{\sim}{n} \sim \text{Gamma}(\eta, \nu)$ and η and ν comes from (A.1). Note that, as an adjustment to Nandram (1998), it is convenient to add unity to each component of $\underset{\sim}{\phi}\tau$.

APPENDIX B: Order Restriction on the Prior Distribution of $\underset{\sim}{\psi}$

We need the supports of the cpds of ψ_1 and ψ_2 in their joint prior distribution.

First, consider the support of the cpd of ψ_1. We have the following inequalities, $\psi_1 > \psi_2$; $\psi_2 > 1-\psi_1-\psi_2$, so that $\psi_1 > 1-2\psi_2$ (giving $\psi_2 < \frac{1}{2}$). Finally, $1-\psi_1-\psi_2 > 0$ gives $\psi_1 < 1-\psi_2$.

Second, consider the support of the cpd of ψ_2. We have the following inequalities, $\psi_1 > \psi_2$; $\psi_2 > 1-\psi_1-\psi_2$ gives $\psi_2 > \frac{1}{2}(1-\psi_1)$. But, of course, $\psi_1 > \frac{1}{2}(1-\psi_1)$, that gives $\psi_1 > \frac{1}{3}$. Finally, $1-\psi_1-\psi_2 > 0$ gives $\psi_2 < 1-\psi_1$.

Putting these together, we get that given ψ_2, the support of ψ_1 is

$$\max\{\frac{1}{3}, \psi_2, 1-2\psi_2\} < \psi_1 < 1-\psi_2, \tag{B.1}$$

and given ψ_1, the support of ψ_2 is

$$\frac{1}{2}(1-\psi_1) < \psi_2 < \min\{\frac{1}{2}, \psi_1, 1-\psi_1\}. \tag{B.2}$$

REFERENCES

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

Bhatta, D. R., Nandram, B. and Sedransk, J. (2018), "Bayesian Testing for Independence of Two Categorical Variables Under Two-Stage Cluster Sampling with Covariates," *Journal of Applied Statistics*, 45, 2365-2393.

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.

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. (2018), "Modelling Group Heterogeneity for small Area Estimation Using M-quantiles," *International Statistical Review* (submitted).

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., Bhatta, D. R. and Bhadra, D. (2013), "A Likelihood Ratio Test of Quasi-independence for Sparse Two-way Contingency Tables," *Journal of Statistical Computation and Simulation*, 85 (2), 284-304.

Nandram, B., Bhatta, D. R., Sedransk, J. and Bhadra, D. (2013), "A Bayesian Test of Independence in a Two-way Contingency Table Using Surrogate Sampling," *Journal of Statistical Planning and Inference*, 143, 1392-1408.

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*, Ser. 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.

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.