DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 1, 2016, pp. 187–208, http://dx.doi.org/10.1515/JOS-2016-0009

Bayesian Predictive Inference of a Proportion Under a Twofold Small-Area Model

Balgobin Nandram[1]

We extend the twofold small-area model of Stukel and Rao (1997; 1999) to accommodate binary data. An example is the Third International Mathematics and Science Study (TIMSS), in which pass-fail data for mathematics of students from US schools (clusters) are available at the third grade by regions and communities (small areas). We compare the finite population proportions of these small areas. We present a hierarchical Bayesian model in which the first-stage binary responses have independent Bernoulli distributions, and each subsequent stage is modeled using a beta distribution, which is parameterized by its mean and a correlation coefficient. This twofold small-area model has an intracluster correlation at the first stage and an intercluster correlation at the second stage. The final-stage mean and all correlations are assumed to be noninformative independent random variables. We show how to infer the finite population proportion of each area. We have applied our models to synthetic TIMSS data to show that the twofold model is preferred over a onefold small-area model that ignores the clustering within areas. We further compare these models using a simulation study, which shows that the intracluster correlation is particularly important.

Key words: Intracluster and intercluster correlations; credible intervals; goodness of fit; hierarchical model; simulation study.

1. Introduction

We assume that there are several small areas and each area consists of several clusters; each cluster consists of a number of units (individuals). A random sample of clusters is taken from each area and within each sampled cluster a random sample of units is taken. This is the twofold sample design. A hierarchical Bayesian model is used to make inference about the finite population proportion of each small-area. In this model we have an intracluster (between two units in the same cluster) correlation at the first stage and an intercluster (between two units in two different clusters in the same area) correlation at the second stage. We show that the intracluster correlation is important by comparing the

[1] Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609, U.S.A. Email: balnan@wpi.edu

Acknowledgments: Part of this work was first presented by Balgobin Nandram in honor of Jon Rao on his Seventy-fifth Birthday Conference at Carleton University, Canada. In addition, presentations were made at several universities and conferences worldwide. This work was supported by a grant from the Simons Foundation (#353953, Balgobin Nandram). The author is grateful to the three reviewers for their careful reading of the manuscript and their extensive reports. He is also grateful to the Co-Editor-in-Chief, Dr. Ingegerd Jansson, for encouragement. Further research on this topic is currently being pursued by my PhD students.

Dedication: This article is dedicated to my late parents, Chandramattie Boodram and Nandram Boodram, who have enormously influenced my thinking. Both parents passed away while I was doing this work.

twofold small-area model with a onefold small-area model (the intracluster correlation is ignored). The Third International Mathematics and Science Study (TIMSS) uses a similar design.

In Subsection 1.1 we describe the TIMSS data that we use to illustrate our methodology and we discuss its importance. In Subsection 1.2 we introduce pertinent literature to show what has been done in twofold modeling and related problems. In Subsection 1.3 we clearly identify the innovations in this paper. Finally, we show a plan of the entire article.

1.1. Description of TIMSS Data

TIMSS is sponsored by the International Association for the Evaluation of Education Achievement, an international organization of national research institutions and government research agencies, and it is used to compare the performance of primary school students in mathematics and science. TIMSS provides reliable and timely data on the mathematics and science achievement of third-grade US students compared to that of students in other countries. Of course, there are other studies used for this purpose with similar objectives (e.g., the Program for International Student Assessment, PISA). These studies provide information to “No Child Left Behind” and the “Race to the Top” programs in the US; to date, the US has spent more than ten billion dollars on the Race to the Top program since it was announced by President Barack Obama in 2009 (Hamilton 2009). Our study can potentially be used to suggest which regions and communities in the US need funding to improve the education systems (e.g., qualified teachers, improved equipment, parental participation, extramural programs, etc).

The basic sample design used in TIMSS for the population of third and fourth grade students was a two-stage stratified cluster design. The first stage consisted of a sample of schools; the second stage consisted of samples of one mathematics classroom from each eligible target grade in the sampled schools. The design required schools to be sampled using a probability proportional to size (PPS) systematic sampling (Foy et al. 1996), and classrooms to be sampled with equal probabilities. Different aspects of the design were adapted to national conditions and analytical needs. For example, many countries stratified the school sampling frame by variables of national interest. As another example, if geographic regions were an explicit stratification variable, then separate school sampling frames would be constructed for each region. The multistage stratified cluster design results in differential probabilities of selection and each student consequently has different weights. In a realistic analysis of the TIMSS data we would need to incorporate the survey weights into the analysis. However, because our main interest is to show how to handle the clustering within small areas, we have ignored the survey weights.

The data set, which we used and collected in 1999, consists of 2,477 students (135 schools) who participated in TIMSS (see Calsyn et al. 1999). Clusters are schools while the units within the clusters are the students. Areas are formed crossing region and community. There are four regions of the US (Northeast, South, Central, and West) and there are three communities (village or rural area, outskirts of a town or city, and close to the center of a town or city), which the students come from. Thus there are twelve areas (strata). The binary variable is whether a student’s mathematics score is below average. We use synthetic data to illustrate our methodology and we take roughly half of the sampled data (i.e., a simple random sample of half the number of schools and a simple random sample of half the number of third-grade students from each selected school) for analysis and we use the other half to assess the predictive power of our procedure. The finite population is the original sample. Our objective is to make inference about the finite population proportion of students who earned below average scores in mathematics for each small-area. This measure can be used to compare the regions and the communities in the US.

The data (half) on the mathematics test scores are shown in Table 1, where we define the twelve areas (e.g., NR is a village or rural area in the north east). There are some schools in which all students were either below average or above average, thereby creating some difficulties for estimation. Looking at Table 1, the numbers of schools sampled in the twelve areas are 2, 4, 5, 4, 8, 6, 1, 3, 7, 3, 6, 15 and the numbers of students sampled in the schools range from 4 to 13. Each area is too sparse for direct estimation even with the complete data set.

1.2. Pertinent Literature

Nandram and Sedransk (1993) described a hierarchical Bayesian model to make inference about the finite population proportion under two-stage cluster sampling, the design we have within each area in a twofold sample design. The model can be viewed as a discrete analogue of the model for two-stage cluster sampling with normal data (Scott and Smith 1969) that has been extended in many directions (e.g., Malec and Sedransk 1985). We note that the work of Nandram and Sedransk (1993) was extended by Nandram (1998) to multinomial data and this extension may be viewed as a Bayesian analogue of the Dirichlet-multinomial model for cluster sampling (Brier 1980). However, our onefold model is different because in this design a simple random sample is taken from each area, but in the twofold model a two-stage cluster sample is performed in each area.

When there is a clustering effect, the units in a cluster are, in general, positively correlated leading to a smaller effective sample size and therefore larger variability in the estimates of the cell probabilities (i.e., the design effect is larger than one for each area). For example, see Brier (1980), Bedrick (1983), Holt et al. (1980), and Scott and Holt (1982). There is a similar issue in hypothesis testing. Clustering will evidently result in larger p-values than what would be obtained under simple random sampling. Rao and Scott (1981; 1984) have studied this problem very carefully for contingency tables and obtained simple and familiar corrections to the standard chi-squared statistic for the test of independence for two-way contingency tables arising from two-stage cluster sampling and more generally. Nandram et al. (2013) have a Bayesian analogue of these works.

From a Bayesian perspective, a related problem is when data are fitted to a hierarchical model but actually follow a model with an additional unknown structure. This is like our problem in which a onefold model is fitted and the second-stage cluster sampling within each area is ignored. Using posterior predictive p-values, Yan and Sedransk (2007) studied the situation where the data follow a normal model with a two-stage (three-stage) hierarchical structure while the fitted model has a one-stage (two-stage) hierarchical structure.

Table 1. Number of US students below average in mathematics within schools by area (region by community)

Area			*m*	Schools														
NR	s	9	2	4	5													
	n	18		8	10													
NO	s	12	4	5	5	1	1											
	n	38		10	12	7	9											
NC	s	39	5	9	11	6	13	0										
	n	56		11	12	12	13	8										
SR	s	24	4	7	4	9	4											
	n	37		10	7	11	9											
SO	s	31	8	3	2	0	8	7	2	9	0							
	n	81		9	10	8	9	13	9	12	11							
SC	s	43	6	8	6	4	7	8	10									
	n	52		8	8	9	8	9	10									
CR	s	4	1	4														
	n	11		11														
CO	s	13	3	7	4	2												
	n	24		11	9	4												
CC	s	41	7	10	6	7	6	3	5	4								
	n	60		10	6	9	10	8	9	8								
WR	s	8	3	1	4	3												
	n	27		8	8	11												
WO	s	27	6	3	6	8	5	3	2									
	n	46		6	6	10	8	8	8									
WC	s	83	15	4	5	7	10	10	3	6	6	2	8	6	6	2	5	3
	n	118		9	5	7	11	11	5	13	6	8	9	7	8	5	7	7

NOTE: For each area the counts are in pairs (*s*,*n*) with *s* at top and *n* bottom. The areas are formed by crossing region (N: north, S: south, C: central, W: west) and community (R: rural, O: outskirt of a town or city, C: town or city). For example, in area NR there are $m = 2$ schools with a total of $n = 18$ students (8 from one school and 10 from the other) and $s = 9$ students scored below average (4 from one school and 5 from the other). This constitutes approximately half the number of schools and half the number of students.

They used several diagnostic procedures to help detect this additional structure. Yan and Sedransk (2010) studied the ability to detect a three-stage model when a two-stage model is actually fitted, and using Bayesian standardized residuals concluded that it is due to the magnitude of correlation induced by the additional structure. This is the key point of our work.

For twofold modeling, there have been some activities for continuous response variables, not binary response variables, and most of this work has been within the empirical Bayes framework. Onefold and twofold nested error regression models were introduced by Fuller and Battese (1973) in which transformations to uncorrelated errors with constant variance are obtained starting with a general error covariance matrix. Transformations permit the calculation of generalized least-squares estimators and their covariance matrices by ordinary least-squares regression. They have made an analogy between survey sampling and experimental design via subsampling of primary, secondary, and tertiary sampling units, and split-split-plot experiments. Ghosh and Lahiri (1988) studied multistage sampling under posterior linearity using Bayes and empirical Bayes methods. Estimation of regression models with nested error structure and unequal error variances were further studied by Stukel and Rao (1997) under two-stage and three-stage cluster sampling. Small-area models under twofold nested error regression models were also studied by Stukel and Rao (1999); see also Rao (2003, sec. 5.5.3) and Datta and Ghosh (1991).

1.3. Innovations

This is mainly a methodological article on twofold small-area modeling, and in attempting to analyze the TIMSS data, we have made the following significant innovations.

1. Our models are for categorical data (binary). As can be seen from the literature, twofold modeling has been done for continuous data. While the categorical data models are related to the continuous data models, they pose additional difficulties for methodology and model fitting.
2. We have a new reparameterization of the beta distribution in terms of correlation (intracluster and intercluster). This permits modeling these correlations directly. In fact, this opens up a new avenue for the analysis of data collected using a twofold sample design and further analysis of more complex categorical (e.g., polychotomous) data.
3. With these reparameterizations we develop two hierarchical Bayesian models, a onefold and a twofold model for binary data.
4. The computations pose some difficulties for the Gibbs sampler and we have overcome these difficulties using random samples instead of the Gibbs sampler. In our twofold model there are two weakly identified parameters, thereby causing long-range dependence in the Gibbs sampler.
5. The TIMSS data will be analyzed using both our onefold and twofold models. We demonstrate that the intracluster correlation creates an important difference between the two models and provides additional insight to the analysis of these data. A simulation study demonstrates the importance of the twofold model for TIMSS data as well.

In Section 2 we describe the onefold and twofold models and we describe how to fit them. In Appendix A we describe how to perform the computation for the twofold model without using the Gibbs sampler. In a technical report, Nandram (2014) now called TRN14, we compare our sampling-based method with the Gibbs sampler. In Section 3, we analyze the TIMSS data and we also compare the onefold and twofold models. We also present a simulation study to compare the onefold and the twofold small-area model even further. Section 4 contains concluding remarks, and some additional problems are discussed. In Appendix B we briefly describe a multifold model.

2. Bayesian Small-Area Models

We make two simple observations. Let $y_i|p \overset{iid}{\sim} \text{Bernoulli}(p)$, $p \sim \text{Beta}\{\mu\tau,(1-\mu)\tau\}$, where $0 < \mu < 1$ is the mean of the beta random variable and τ is the sum of the parameters of the standard beta distribution.

First, the y_i are exchangeable and the correlation between y_i and y_j is $\rho = (1+\tau)^{-1}$ with $\tau = (1-\rho)/\rho$. Thus, we can write the model as $y_i|p \overset{iid}{\sim} \text{Bernoulli}(p)$, $p \sim \text{Beta}\left\{\mu\frac{1-\rho}{\rho}, (1-\mu)\frac{1-\rho}{\rho}\right\}$.

Second, considering a single observation, y_1 say, the posterior mean of p given y_1 is

$$\text{E}(p|\rho,\mu,y_1) = \rho y_1 + (1-\rho)\mu.$$

The prior density, $\rho \sim \text{Uniform}(0,1)$, is called a shrinkage prior. Shrinkage priors have good frequentist properties (see Natarajan and Kass 2000; Molina et al. 2014; Toto and Nandram 2010). These observations motivate the construction of our small-area model for binary data.

We have a population of ℓ small areas and within the i^{th} area there are M_i clusters. Within the j^{th} cluster there are N_{ij} individuals. The binary responses are $y_{ijk}, k = 1, \ldots, N_{ij}, j = 1, \ldots, M_i, i = 1, \ldots, \ell$. A simple random sample of m_i clusters is taken from the i^{th} area and a simple random sample of n_{ij} individuals is taken from the j^{th} cluster. Let $n_i = \sum_{j=1}^{m_i} n_{ij}$, $s_{ij} = \sum_{k=1}^{n_{ij}} y_{ijk}$, $s_i = \sum_{j=1}^{m_i} s_{ij}$.

Letting $N_i = \sum_{j=1}^{M_i} N_{ij}$, the finite population proportion for the i^{th} area is

$$P_i = \sum_{j=1}^{M_i}\sum_{k=1}^{N_{ij}} y_{ijk}/N_i, i = 1, \ldots, \ell.$$

Let $T_{ij}^{(1)} = \sum_{k=n_{ij}+1}^{N_{ij}} y_{ijk}$, $j = 1, \ldots, m_i$, denote the nonsampled total of the j^{th} sampled clusters and $T_{ij}^{(2)} = \sum_{k=1}^{N_{ij}} y_{ijk}, j = m_i+1, \ldots, M_i$, the total of the j^{th} nonsampled cluster. Letting $n_i = \sum_{j=1}^{m_i} n_{ij}$, $\hat{p}_i = \sum_{j=1}^{m_i}\sum_{k=1}^{n_{ij}} y_{ijk}/n_i$, it is convenient to express P_i as

$$P_i = \left\{n_i\hat{p}_i + \sum_{j=1}^{m_i} T_{ij}^{(1)} + \sum_{j=m_i+1}^{M_i} T_{ij}^{(2)}\right\}/N_i, i = 1, \ldots, \ell, \tag{1}$$

where the $\hat{p}_i$ are observed. Bayesian predictive inference is required for $T_{ij}^{(1)}$ and $T_{ij}^{(2)}$. There is an expression similar to (1) for the finite population mean for each area (Stukel and Rao 1999).

2.1. A Onefold Model

We first construct the small-area onefold Bayesian model,

$$y_{ijk}|p_i \overset{ind}{\sim} \text{Bernoulli}(p_i), j = 1, \ldots, M_i, k = 1, \ldots, N_{ij},$$
$$p_i|\theta, \gamma \overset{iid}{\sim} \text{Beta}\left\{\theta\frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\right\}, i = 1, \ldots, \ell,$$

where in a standard Beta(α, β), $\theta = \alpha/(\alpha + \beta)$ and $\gamma = (\alpha + \beta + 1)^{-1}$. Note that the cluster effects are dropped (i.e., the p_i do not have subscript j). Noting that θ and γ are really probabilities, a priori

$$\theta, \gamma \overset{iid}{\sim} \text{Beta}(\alpha_o, \beta_o),$$

where $\alpha_o = \beta_o$ for a noninformative prior with small values (e.g., $\alpha_o = 1$ for a uniform prior and $\alpha_o = .5$ for Jeffreys prior). Here, $0 < \gamma < 1$ strictly, and the uniform prior on γ is a shrinkage prior.

The model of Nandram and Sedransk (1993) for two-stage cluster sampling with binary responses is similar to the current one. One important difference is in the prior specification of θ and the reparametrization of γ, which unlike Nandram and Sedransk (1993) is stochastic here. Furthermore, we predict the finite population proportion of each area, not the overall finite population proportion.

The onefold model can be fitted easily by making random draws from the joint posterior density of θ and γ, and samples of p_i can be obtained using the multiplication rule. Specifically,

$$p_i|s_i, \theta, \gamma \overset{ind}{\sim} \text{Beta}\left\{s_i + \theta\frac{1-\gamma}{\gamma}, n_i - s_i + (1-\theta)\frac{1-\gamma}{\gamma}\right\},$$

and

$$\pi(\theta, \gamma|\underset{\sim}{y}) \propto \prod_{i=1}^{\ell} \frac{B\{s_i + \theta(1-\gamma)/\gamma, n_i - s_i + (1-\theta)(1-\gamma)/\gamma\}}{B\{\theta(1-\gamma)/\gamma, (1-\theta)(1-\gamma)/\gamma\}} \times \theta^{\alpha_o - 1}$$
$$(1-\theta)^{\beta_o - 1}\gamma^{\alpha_o - 1}(1-\gamma)^{\beta_o - 1}, 0 < \theta, \gamma < 1, \qquad (2)$$

where $B(\cdot, \cdot)$ is the beta function.

Because the posterior density of (θ,γ) is not in a simple form, we use a one-dimensional grid method and numerical integration via Gaussian quadrature to draw samples from it. We first integrate out θ to get $\pi(\gamma|\underset{\sim}{y}) \approx \sum_{g=1}^{G} w_g \pi(x_g, \gamma|\underset{\sim}{y})$, where $x_g, g = 1, \ldots, G$, are the G roots of a Legendre orthogonal polynomial with weights $w_g, g = 1, \ldots, G$; $G = 20$ or so provides a very accurate and fast procedure. Then, we use a one-dimensional grid to draw γ from $\pi(\gamma|\underset{\sim}{y})$. The unit interval is simply divided into 100 subintervals of equal width, and the joint posterior density is approximated by a discrete distribution with probabilities proportional to the heights of the continuous distribution at the midpoints of these subintervals. Now, it is easy to draw a sample from this univariate discrete distribution of $\pi(\gamma|\underset{\sim}{y})$. It is efficient to remove subintervals with small probabilities (smaller than 10^{-6}); we call the others probable subintervals. To draw a single deviate, we first draw one of the probable subintervals. After we have obtained this subinterval, a uniform random variable is drawn within this subinterval. This is a standard jittering procedure and it provides different deviates with probability one. We call this random number generator the univariate grid sampler that is also used to fit the twofold model.

Once samples of the p_i are obtained, Bayesian predictive inference follows easily because $T_{ij}^{(1)}|p_i \overset{ind}{\sim} \text{Binomial}(N_{ij} - n_{ij}, p_i)$ and $T_{ij}^{(2)}|p_i \overset{ind}{\sim} \text{Binomial}(N_{ij}, p_i)$ and, given p_i, $T_{ij}^{(1)}$ and $T_{ij}^{(2)}$ are independent. It follows easily that $\sum_{j=1}^{m_i} T_{ij}^{(1)} + \sum_{j=m_i+1}^{M_i} T_{ij}^{(2)}|p_i \sim \text{Binomial}(N_i - n_i, p_i)$. Thus it is easy to make inference about P_i by using data augmentation. For each iterate p_i, we simply draw $\sum_{j=1}^{m_i} T_{ij}^{(1)} + \sum_{j=m_i+1}^{M_i} T_{ij}^{(2)}$. We use 1,000 samples; convergence monitoring is not required.

2.2. A Twofold Model

The twofold small-area model adds one layer to the onefold model. For a twofold Bayesian model,

$$y_{ijk}|p_{ij} \overset{ind}{\sim} \text{Bernoulli}(p_{ij}), k = 1, \ldots, N_{ij},$$

$$p_{ij}|\mu_i, \rho \overset{ind}{\sim} \text{Beta}\left\{\mu_i \frac{1-\rho}{\rho}, (1-\mu_i)\frac{1-\rho}{\rho}\right\}, j = 1, \ldots, M_i,$$

$$\mu_i|\theta, \gamma \overset{iid}{\sim} \text{Beta}\left\{\theta \frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\right\}, i = 1, \ldots, \ell,$$

and a priori

$$\rho, \theta, \gamma \overset{iid}{\sim} \text{Beta}(\alpha_o, \beta_o)$$

with the same comments about this prior as for the onefold model. We assume that $0 < \theta, \rho, \gamma < 1$ strictly. This can be achieved by taking $\epsilon \leq \theta, \rho, \gamma \leq 1 - \epsilon$, where ϵ is a small positive quantity (e.g., $\epsilon = 10^{-6}$).

If we allow ρ to go to zero, then the p_{ij} almost surely go to the μ_i and the twofold model becomes the onefold model. (In the limit, the μ_i in the twofold model become the p_i in the onefold model.) That is, if ρ is small, we anticipate very little difference between the two models. Thus it is ρ that distinguishes the onefold and twofold models.

In Subsection 2.1 we stated that $\text{cor}(y_{ijk}, y_{ijk'}|\mu_i, \rho) = \rho, k \neq k'$. That is, within the same area, the correlation between two units in the same cluster (intracluster) is ρ. Clearly, $\text{cor}(y_{ijk}, y_{ij'k'}|\mu_i, \rho) = 0$ and within the same area the actual correlation between two units in two different clusters (intercluster) is 0. It is also easy to show that

$$\text{cor}(y_{ijk}, y_{ij'k'}|\theta, \rho, \gamma) = \gamma, j \neq j', k \neq k'.$$

That is, one can interpret γ as the intercluster correlation between two units in two different clusters in the same area. Finally, note that $\text{cor}(y_{ijk}, y_{ijk'}|\theta, \rho, \gamma) = \gamma + (1-\gamma)\rho \geq \max(\rho, \gamma)$.

Using Bayes' theorem and letting $s_{ij} = \sum_{k=1}^{n_{ij}} y_{ijk}$, $\underset{\sim}{p} = (p_{ij}, j = 1, \ldots, m_i, i = 1, \ldots, \ell)'$, and $\underset{\sim}{\mu} = (\mu_i, i = 1, \ldots, \ell)'$, the joint posterior density is

$$\pi(\underset{\sim}{p}, \underset{\sim}{\mu}, \underset{\sim}{\theta}, \rho, \gamma | \underset{\sim}{y}) \propto \prod_{i=1}^{\ell} \prod_{j=1}^{m_i} p_{ij}^{s_{ij}} (1 - p_{ij})^{n_{ij} - s_{ij}} \frac{p_{ij}^{\mu_i(1-\rho)/\rho - 1} (1 - p_{ij})^{(1-\mu_i)(1-\rho)/\rho - 1}}{B\{\mu_i(1-\rho)/\rho, (1-\mu_i)(1-\rho)/\rho\}}$$

$$\times \left\{ \prod_{i=1}^{\ell} \frac{\mu_i^{\theta(1-\gamma)/\gamma - 1} (1 - \mu_i)^{(1-\theta)(1-\gamma)/\gamma - 1}}{B\{\theta(1-\gamma)/\gamma, (1-\theta)(1-\gamma)/\gamma\}} \right\} \theta^{\alpha_o - 1} (1 - \theta)^{\beta_o - 1} \rho^{\alpha_o - 1} (1 - \rho)^{\beta_o - 1} \gamma^{\alpha_o - 1}$$

$$(1 - \gamma)^{\beta_o - 1}, 0 < p_{ij}, \mu_i, \theta, \rho, \gamma < 1, j = 1, \ldots, m_i, i = 1, \ldots, \ell.$$

We use both the Gibbs sampler and a random sampler to fit the model. The Gibbs sampler is used after collapsing over the p_{ij} and then samples are obtained from the posterior densities of the p_{ij} using the composition method (i.e., multiplication rule). Once samples of the p_{ij} are obtained, Bayesian predictive inference follows easily because $T_{ij}^{(1)} | p_{ij} \overset{ind}{\sim} \text{Binomial}(N_{ij} - n_{ij}, p_{ij}), j = 1, \ldots, m_i$, for the sampled clusters and $T_{ij}^{(2)} | p_{ij} \overset{ind}{\sim} \text{Binomial}(N_{ij}, p_{ij}), j = 1, \ldots, M_i$, for the nonsampled clusters. Given p_{ij}, $T_{ij}^{(1)}$ and $T_{ij}^{(2)}$ are independent. However, the Gibbs sampler is not easy to use because there are weakly identified parameters and this needs special attention. See TRN14 for the technical details and convergence monitoring.

For the random sampler, first note that conditionally a posteriori the p_{ij} are independent and

$$p_{ij} | s_{ij}, \mu_i, \rho \overset{ind}{\sim} \text{Beta}\{s_{ij} + \mu_i(1-\rho)/\rho, n_{ij} - s_{ij} + (1-\mu_i)(1-\rho)/\rho\}.$$

Accordingly, once samples are obtained from the joint posterior density of $\underset{\sim}{\mu}, \theta, \rho, \gamma | \underset{\sim}{s}$, a sample of p_{ij} is easy to obtain. Then, after integrating out the p_{ij}, we have

$$\pi(\underset{\sim}{\mu}, \theta, \rho, \gamma | \underset{\sim}{y}) \propto \prod_{i=1}^{\ell} \prod_{j=1}^{m_i} \frac{B\{s_{ij} + \mu_i(1-\rho)/\rho, n_{ij} - s_{ij} + (1-\mu_i)(1-\rho)/\rho\}}{B\{\mu_i(1-\rho)/\rho, (1-\mu_i)(1-\rho)/\rho\}}$$

$$\times \prod_{i=1}^{\ell} \frac{\mu_i^{\theta(1-\gamma)/\gamma - 1} (1 - \mu_i)^{(1-\theta)(1-\gamma)/\gamma - 1}}{B\{\theta(1-\gamma)/\gamma, (1-\theta)(1-\gamma)/\gamma\}} \times \theta^{\alpha_o - 1} (1 - \theta)^{\beta_o - 1} \rho^{\alpha_o - 1} (1 - \rho)^{\beta_o - 1}$$

$$\gamma^{\alpha_o - 1} (1 - \gamma)^{\beta_o - 1}, 0 < \mu_i, \theta, \rho, \gamma < 1, i = 1, \ldots, \ell. \quad (3)$$

See Appendix A for the more detailed computations using the random sampler. For the TIMSS data, the results from the Gibbs sampler and the random sample are similar (see TRN 14).

3. Numerical Analysis

We discuss an illustrative example using data from the Third International Mathematics and Science Study (TIMSS) and we perform a simulation study to confirm the superiority of the twofold small-area model. This section has three subsections.

In Subsection 3.1 we describe the model diagnostic procedures used for analysis. In Subsection 3.2 we analyze the TIMSS data. We compare the onefold and twofold models.

We have used the posterior mean (PM), posterior standard deviation (PSD), and 95% highest posterior density (HPD) interval to summarize the distributions. We also computed the numerical standard error (NSE), which is based on the batch means method; NSE is a measure of the repeatability of the entire sampling. In Subsection 3.3 we describe a simulation study.

3.1. Model Diagnostics

We discuss three goodness-of-fit procedures, the deviance information criterion (DIC) together with the complexity or effective number of parameters (PD), the conditional predictive ordinate (CPO) along with the logarithm of the pseudomarginal likelihood (LPML), and the Bayesian predictive p-value (BPP). The DIC, LPML, and BPP look at the overall fit of the model; see Gelman et al. (2013) for further discussions of these measures. We give expressions for the twofold model because it is easy to write down similar ones for the onefold model.

In the twofold model $s_{ij}|p_{ij}\overset{ind}{\sim}\text{Binomial}(n_{ij}, p_{ij})$, $p_{ij}|\mu_i\overset{ind}{\sim}\text{Beta}\{\mu_i(1-\rho)/\rho, (1-\mu_i)(1-\rho)/\rho\}$. Thus, integrating out the p_{ij} we get a product of beta-binomial probability mass functions,

$$p(\underset{\sim}{s}|\underset{\sim}{\mu}, \rho) = \prod_{i=1}^{\ell}\prod_{j=1}^{m_i}\binom{n_{ij}}{s_{ij}}\frac{B\{s_{ij}+\mu_i(1-\rho)/\rho, n_{ij}-s_{ij}+(1-\mu_i)(1-\rho)/\rho\}}{B\{\mu_i(1-\rho)/\rho, (1-\mu_i)(1-\rho)/\rho\}}.$$

It is also true that $\mathrm{E}(\mathrm{s}_{ij}|\mu_i, \rho) = n_{ij}\mu_i$ and $\mathrm{Var}(s_{ij}|\mu_i, \rho) = n_{ij}\{1+(n_{ij}-1)\rho\}\mu_i(1-\mu_i)$.

Let

$$PD = \bar{D} - D(\underset{\sim}{\bar{\theta}}, \bar{\gamma}), \quad DIC = \bar{D} + PD$$

respectively be the complexity of the model and the deviance information criterion, where $\bar{D}$ and $D(\underset{\sim}{\bar{\theta}}, \bar{\gamma})$ are defined below for the onefold and twofold models.

Let $\mu_i^{(h)}, i = 1, \ldots, \ell, \rho^{(h)}, h = 1, \ldots, M$, denote the iterates of Gibbs sampling from the twofold model, $\bar{\mu}_i = \sum_{h=1}^{M}\mu_i^{(h)}/M, i = 1, \ldots, \ell$, and $\bar{\rho} = \sum_{h=1}^{M}\rho^{(h)}/M$. Then, $D(\underset{\sim}{\bar{\mu}}, \bar{\rho}) = -2log\{\mathrm{p}(\underset{\sim}{\mathrm{s}}|\underset{\sim}{\bar{\mu}}, \bar{\rho})\}$ and $\bar{D} = -2\sum_{h=1}^{M}log\{p(\underset{\sim}{s}|\underset{\sim}{\mu}^{(h)}, \rho^{(h)})\}/M$.

Models with smaller DIC are preferred over models with larger DIC. Models are penalized both by the value of $\bar{D}$, which favors a good fit, and PD. Since $\bar{D}$ will decrease as the number of parameters in a model increases, PD compensates for this effect by favoring models with a smaller number of parameters. However, DIC tends to select overfitted models. The Bayesian predictive information criterion (BPIC) can protect against this effect but it is difficult to compute, it is not meant for dependent data, and consistency (as the sample size increases) is needed (see Ando 2007). The inconsistency problem can be overcome by integrating out the p_{ij} and the μ_i, but this creates dependent data.

Similar to the DIC, the second measure is the LPML. Both measures are based on the same cross-validation (leave-one-out) procedure. A summary statistic for CPO values is LPML; unlike the DIC, larger values of LPML indicate better fitting models (e.g., Geisser and Eddy 1979). For the twofold model the CPO is given by

$$\widehat{CPO}_{ij} = \left\{ \frac{1}{M} \sum_{h=1}^{M} \frac{1}{f\left(s_{ij} | p_{ij}^{(h)}\right)} \right\}^{-1}, j = 1, \ldots, m_i, i = 1, \ldots, \ell,$$

where $p_{ij}^{(h)}, h = 1, \ldots, M$, are the samples from $p_{ij} | s_{ij}, \mu_i, \rho$ and $s_{ij} | p_{ij} \overset{ind}{\sim} \text{Binomial}(n_{ij}, p_{ij})$. Again, it is interesting to note that for each (ij), $\widehat{CPO}_{ij}$ is the harmonic mean of the likelihoods $f(s_{ij} | p_{ij}^{(h)}), h = 1, \ldots, M$. Then,

$$LPML = \sum_{i=1}^{\ell} \sum_{j=1}^{m_i} log(\widehat{CPO}_{ij}).$$

The LPML, like the DIC, can discriminate between the onefold and the twofold models. We compute the CPO and the LPML at the cluster level, the LPML being preferable (easy to use).

Our third measure is the BPP for the two models. For the twofold model, the discrepancy function is

$$T_2(\underset{\sim}{s}; \underset{\sim}{\mu}, \rho) = \sum_{i=1}^{\ell} \sum_{j=1}^{m_i} \frac{\{s_{ij} - E(s_{ij} | \mu_i, \rho)\}^2}{Var(s_{ij} | \mu_i, \rho)}.$$

Then the BPP is $P\{T_2(\underset{\sim}{s}^{(rep)}; \underset{\sim}{\mu}, \rho) \geq T_2(\underset{\sim}{s}^{(obs)}; \underset{\sim}{\mu}, \rho) | \underset{\sim}{s}\}$, where probability is calculated over the iterates $(\underset{\sim}{\mu}^{(h)}, \rho^{(h)})$, $h = 1, \ldots, M$. Extremely small (near 0) or extremely large (near 1) values of this probability indicate that the model does not fit well.

3.2. *Illustrative Example*

First, we compare the two models using the three measures. For the onefold (twofold) model, $PD = 6.70$ ($PD = 7.98$), $DIC = 313$ ($DIC = 282$), $LPML = -609$ ($LPML = -575$), and $BPP = .000$ ($BPP = .467$). The BPP tells us that while the twofold model fits the TIMSS data reasonably well, the onefold model does not. The other two measures, *DIC* and *LPML*, tell us that the twofold model provides a better fit to the TIMSS data.

Using the onefold model, for θ $PM = .556$, $PSD = .052$, $NSE = .002$, and the 95% HPD interval is (.448,.654); for γ $PM = .112$, $PSD = .053$, $NSE = .001$, and the 95% HPD interval is (.034,.215). Using the twofold model, for θ $PM = .566$, $PSD = .055$, and $NSE = .002$, the 95% HPD interval is (.443,.662); for γ $PM = .078$, $PSD = .056$, $NSE = .002$, and the 95% HPD interval is (.001,.187). Thus inferences about θ and γ are very similar under the onefold and twofold models.

More importantly, the posterior mean of ρ is .217 with a standard deviation of .050, $NSE = .001$, and the 95% HPD interval of (.122,.309). This also shows that the twofold model, which accommodates the two-stage cluster sampling via the intracluster correlation, ρ, may be preferred.

In Table 2 we present posterior inference about the finite population proportions for the mathematics scores. We see that the posterior means of the onefold model can be larger or smaller than the posterior means of the twofold model. However, the posterior standard deviations for the twofold model are always larger than those of the onefold model. This clearly shows how the twofold model accommodates the clustering effect. In Table 2 we

Table 2. Comparison of posterior inference from the onefold and twofold models for the finite population proportions by areas for US students below average in mathematics

		Onefold			Twofold		
Area	Direct	PM	PSD	95% HPD	PM	PSD	95% HPD
NR	$.500_{.104}$	.515	.087	(.367, .696)	.528	.116	(.316, .747)
NO	$.316_{.067}$	.355	.063	(.234, .480)	.396	.093	(.234, .594)
NC	$.696_{.054}$	.682	.052	(.587, .785)	.667	.075	(.523, .806)
SR	$.649_{.068}$	.636	.061	(.527, .760)	.618	.087	(.440, .773)
SO	$.383_{.047}$	.395	.047	(.303, .483)	.410	.067	(.288, .539)
SC	$.827_{.047}$	.795	.048	(.694, .877)	.757	.071	(.617, .889)
CR	$.364_{.127}$	.427	.108	(.217, .609)	.459	.152	(.196, .717)
CO	$.542_{.093}$	.548	.080	(.386, .707)	.549	.111	(.336, .757)
CC	$.683_{.052}$	.667	.051	(.573, .766)	.660	.068	(.516, .778)
WR	$.296_{.078}$	.345	.076	(.203, .484)	.403	.107	(.203, .602)
WO	$.587_{.065}$	.582	.058	(.452, .683)	.583	.080	(.448, .751)
WC	$.703_{.037}$	.694	.036	(.618, .760)	.685	.051	(.591, .783)

NOTE: PM is the posterior mean, PSD is the posterior standard deviation and HPD is highest posterior density interval. The Monte Carlo errors of the posterior means are smaller than .004 in all cases, and in most cases are substantially smaller than .004. The direct estimate and its standard error are written as a_b where a is the direct estimate and b is its standard error.

have also presented the direct estimates. The direct estimates and their standard errors seem to be closer to the PMs and PSDs of the onefold model, but there are some differences (e.g., areas CR and WR).

In Figure 1 we present plots of the empirical posterior densities of the finite population proportions. These are obtained using the Parzen-Rosenblatt normal kernel density estimator with an optimal window width (e.g., Silverman 1986). In both pictures (onefold and twofold models) we observe a clear difference between the onefold and twofold models. The distributions under the twofold model are more spread out than those of the corresponding onefold model.

Using the TIMSS data (half sample) we perform two small empirical studies. First, we study the quality of the Bayesian predictive inference. Then the ‘true values’ of the finite population proportions (original sample) for the areas are .541, .347, .608, .600, .550, .667, .436, .421, .560, .458, .522, .643. Under the twofold model the 95% HPD interval of the finite population proportion of area SO misses the true value. But under the onefold model the 95% HPD intervals for areas SO, SC, CC miss the true value (see Table 1 for abbreviations). Thus, once again the twofold model provides a better fit than the onefold model.

Second, we investigate the effect of a larger number of areas. As our half-sample dataset has only twelve areas, we have artificially increased the number of areas. Specifically, we have bootstrapped the twelve areas in the half sample to fill in the additional number of areas to get 25, 50, 75, and 100 areas. Detailed comparisons between random sampling and Gibbs sampling are given in TRN14. For example, in the computations random sampling is twice as fast as Gibbs sampling, but the measures (e.g., DIC, LPML, and BPP) are similar.

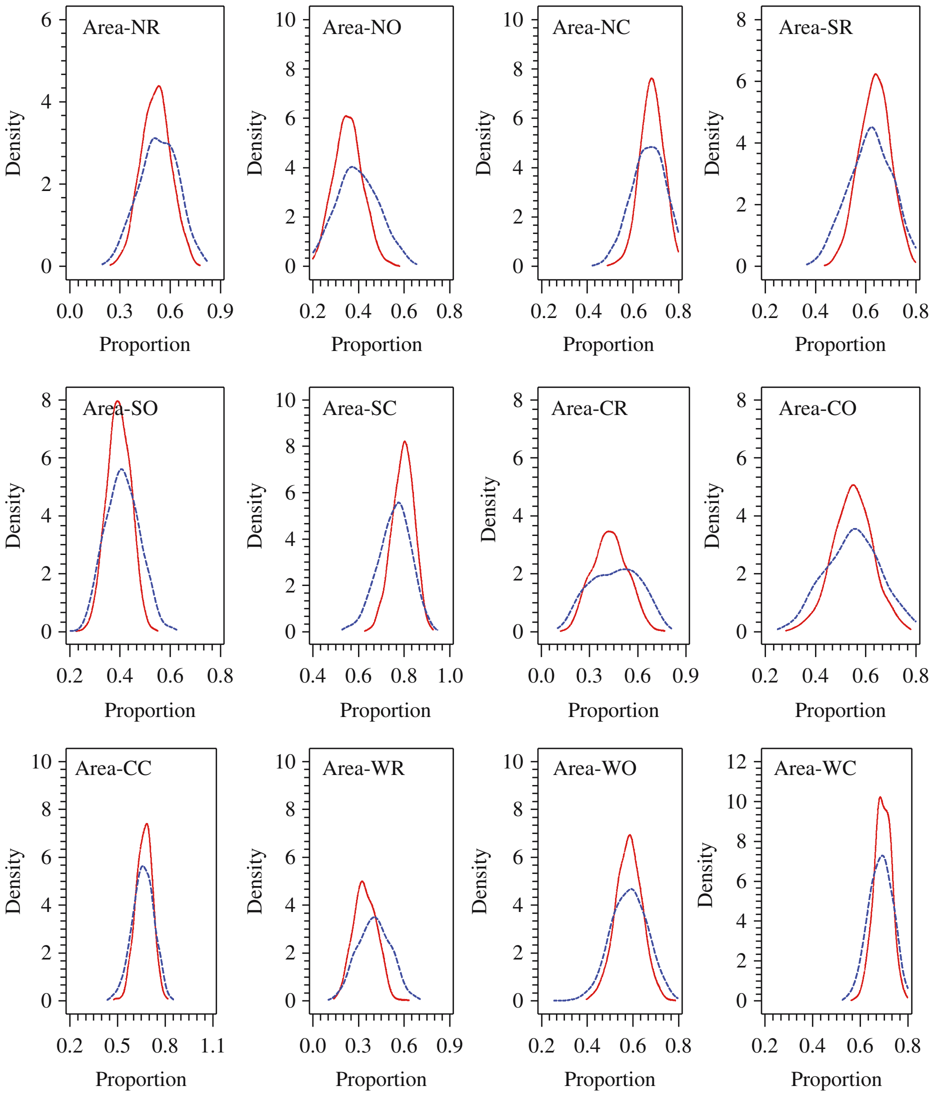

Fig. 1. Comparison of the onefold (solid) and twofold (dotted) models via posterior inferences of the finite population proportions of the empirical densities of finite population proportions by area

3.3. Simulation Study

We have performed a small simulation study to help understand how inferences about the finite population proportions change with the intracluster correlation coefficient (ρ) and the number (ℓ) of small areas. We have studied $\rho = .01, .10, .25, .50, .75$ and $\ell = 12, 25, 50, 75, 100$. Thus, there are twenty-five design points in our simulation study.

We have set the number of schools in each area to be 100 and the number of students within each school to be 15 (i.e., $N_{ij} = 15, j = 1, \ldots, M_i, \; M_i = 100, i = 1, \ldots, \ell$). We also hold $\theta = .60$ and $\gamma = .05$, near the posterior means calculated for the real data. We have taken a simple random sample of five schools from the 100 generated for the population, and a simple random sample of ten students from each selected school (i.e., $m_i = 5$ schools and $n_{ij} = 10$ students). So there are up to 100 areas each having 100 schools and each school having up to 15 students. So we have up to 10,000 schools and 150,000 students. The number of areas can be as large as current computing facilities allow because the area effects can be drawn using parallel computing via our method of random sampling (not Gibbs sampling).

We have simulated binary data from the twofold small-area model,

$$\mu_i|\theta,\gamma \overset{iid}{\sim} \text{Beta}\left\{\theta\frac{1-\gamma}{\gamma},(1-\theta)\frac{1-\gamma}{\gamma}\right\},\ i=1,\ldots,\ell,$$

$$p_{ij}|\mu_i,\rho \overset{ind}{\sim} \text{Beta}\left\{\mu_i\frac{1-\rho}{\rho},(1-\mu_i)\frac{1-\rho}{\rho}\right\},\ j=1,\ldots,M_i,$$

$$y_{ijk}|p_{ij} \overset{ind}{\sim} \text{Bernoulli}(p_{ij}),\ k=1,\ldots,N_{ij}.$$

Thus, we have the true value of $P_i = \sum_{j=1}^{M_i}\sum_{k=1}^{N_{ij}} y_{ijk}/\sum_{j=1}^{M_i} N_{ij},\ i=1,\ldots,\ell$. We have taken 1,000 samples at each of the 25 design points.

In a similar way, we have generated data from the onefold model,

$$p_i|\theta,\gamma \overset{iid}{\sim} \text{Beta}\left\{\theta\frac{1-\gamma}{\gamma},(1-\theta)\frac{1-\gamma}{\gamma}\right\},\ j=1,\ldots,M_i,$$

$$y_{ijk}|p_i \overset{ind}{\sim} \text{Bernoulli}(p_i),\ k=1,\ldots,N_{ij},$$

with a subset of the same design points (i.e., $\rho = 0$).

For all generated data sets we fit the onefold and twofold models using random draws, as described for the computations. We have used parallel computing to fit the models. Note that we need to fit 25,000 simulated data sets.

Here, we have also studied the frequentist properties of our procedure. We compute the absolute bias (AB), relative absolute bias (RAB), and root posterior mean squared error (RPMSE). Specifically, we obtain $AB_{ih} = |PM_{ih} - P_{ih}|$, $RAB_{ih} = AB_{ih}/P_{ih}$ and $RPMSE_{ih} = \sqrt{PSD_{ih}^2 + AB_{ih}^2}$, $i = 1,\ldots,\ell$, $h = 1,\ldots,1{,}000$. We have also computed the 95% HPD interval for each of the 1,000 simulated runs. We have looked at the width (W_{ih}) and the credible incidence (I_{ih}). Here $I_{ih} = 1$ if the 95% HPD interval contains the true value P_i and $I_{ih} = 0$ if the 95% credible interval does not contain the true value P_i. For each area and each design point we have taken the average of these quantities. For example, the estimated probability content of the 95% HPD interval for the i^{th} area is $C_i = \sum_{h=1}^{1000} I_{ih}/1{,}000$.

First, we discuss the simulations when data are generated from the twofold model. In Table 3 we present a comparison of the onefold and twofold models using these measures. The coverages for the twofold model are much closer to the nominal value of 95% than those from the onefold model. In some cases the coverages from the onefold model are much too small. However, the 95% HPD intervals from the twofold model are wider than those from the onefold model. These effects are much larger as ρ increases for each ℓ, thereby clearly showing how the twofold model takes care of the clustering effect. All measures (AB, RAB, RPMSE) for the twofold model are smaller than those for the

Table 3. Simulation for data drawn from the twofold model: Comparison of coverage and widths of 95% HPD intervals and absolute bias, relative absolute bias and, root posterior mean squared error at twelve design points

ℓ	ρ	Model	C-HPD	W-HPD	AB	RAB	RPMSE
25	.10	TFM	$.940_{.0015}$	$.276_{.0002}$	$.056_{.0003}$	$.098_{.0005}$	$.096_{.0002}$
		OFM	$.860_{.0022}$	$.227_{.0001}$	$.061_{.0003}$	$.106_{.0005}$	$.090_{.0002}$
	.25	TFM	$.938_{.0016}$	$.318_{.0002}$	$.068_{.0003}$	$.121_{.0007}$	$.113_{.0002}$
		OFM	$.732_{.0029}$	$.227_{.0001}$	$.081_{.0004}$	$.142_{.0008}$	$.107_{.0003}$
	.50	TFM	$.918_{.0018}$	$.355_{.0003}$	$.077_{.0004}$	$.136_{.0008}$	$.127_{.0003}$
		OFM	$.612_{.0031}$	$.227_{.0002}$	$.107_{.0005}$	$.184_{.0009}$	$.129_{.0004}$
	.75	TFM	$.944_{.0015}$	$.417_{.0003}$	$.083_{.0004}$	$.147_{.0009}$	$.145_{.0003}$
		OFM	$.495_{.0032}$	$.222_{.0003}$	$.137_{.0006}$	$.239_{.0012}$	$.157_{.0006}$
50	.10	TFM	$.940_{.0011}$	$.273_{.0001}$	$.058_{.0002}$	$.104_{.0004}$	$.097_{.0001}$
		OFM	$.857_{.0016}$	$.225_{.0001}$	$.060_{.0002}$	$.105_{.0004}$	$.090_{.0002}$
	.25	TFM	$.935_{.0011}$	$.314_{.0002}$	$.067_{.0002}$	$.119_{.0005}$	$.112_{.0002}$
		OFM	$.727_{.0020}$	$.228_{.0001}$	$.082_{.0003}$	$.143_{.0005}$	$.108_{.0002}$
	.50	TFM	$.936_{.0011}$	$.350_{.0002}$	$.074_{.0002}$	$.133_{.0005}$	$.124_{.0002}$
		OFM	$.607_{.0022}$	$.229_{.0001}$	$.108_{.0004}$	$.190_{.0007}$	$.131_{.0003}$
	.75	TFM	$.942_{.0010}$	$.386_{.0002}$	$.080_{.0003}$	$.143_{.0006}$	$.135_{.0002}$
		OFM	$.492_{.0022}$	$.222_{.0002}$	$.137_{.0004}$	$.240_{.0008}$	$.157_{.0004}$
100	.10	TFM	$.946_{.0007}$	$.275_{.0001}$	$.056_{.0001}$	$.100_{.0003}$	$.096_{.0001}$
		OFM	$.862_{.0011}$	$.225_{.0001}$	$.060_{.0001}$	$.105_{.0003}$	$.089_{.0001}$
	.25	TFM	$.939_{.0008}$	$.311_{.0001}$	$.066_{.0002}$	$.117_{.0003}$	$.110_{.0001}$
		OFM	$.752_{.0014}$	$.229_{.0001}$	$.080_{.0002}$	$.140_{.0004}$	$.106_{.0002}$
	.50	TFM	$.930_{.0008}$	$.340_{.0001}$	$.075_{.0002}$	$.136_{.0004}$	$.122_{.0001}$
		OFM	$.602_{.0015}$	$.227_{.0001}$	$.109_{.0003}$	$.192_{.0005}$	$.131_{.0002}$
	.75	TFM	$.936_{.0008}$	$.385_{.0001}$	$.082_{.0002}$	$.150_{.0005}$	$.137_{.0001}$
		OFM	$.485_{.0016}$	$.222_{.0001}$	$.140_{.0003}$	$.248_{.0006}$	$.160_{.0003}$

NOTE: TFM is the twofold model and OFM is the onefold model. W-HPD and C-HPD are respectively the width and the probability content of a HPD interval. A, AB, and RPMSE are the absolute bias, relative absolute bias and root posterior mean square error. The notation a_b means that a is the estimate and b is the standard error.

onefold model. These effects become more intense for larger ρ. Again this shows the superiority of the twofold model over the onefold model.

In Table 4 we present summaries of PD, DIC, LPML, and BPP. As expected, the PDs for the twofold model should be larger than those for the onefold model. All the DICs for the twofold model are smaller than the corresponding ones for the onefold model, and this disparity becomes larger as ℓ and ρ increase. The results are the same for the LPML. Under the onefold model most of the BPPs are near 0, but under the twofold model the corresponding BPPs are around 0.5. These measures show that while the twofold model is more complex, it is superior to the onefold model. In TRN14 we compare plots of the sample distributions of the negative LPML under the onefold and twofold models over the 1,000 runs by ℓ and ρ. The negative LPML under the twofold model are smaller than under the onefold model and this discrepancy increases with both ρ and ℓ. There are overlaps of distributions when $\rho = .10$ but not for other values of ρ.

Second, we discuss the simulations when data are generated from the onefold model. In Table 5 we present comparisons of the onefold and twofold models. As expected, the onefold model is slightly better than the twofold model. AB, RAB, RPMSE are only

Table 4. Summaries of the 1,000 simulation runs with (data drawn from the twofold model) for the complexity, deviance information criterion, log pseudomarginal likelihood, and the Bayesian predictive p-value by ℓ, ρ and model

		Onefold				Twofold			
ℓ	ρ	PD	DIC	LPML	BPP	PD	DIC	LPML	BPP
25	.10	4.998	553	−1107	.010	15.61	521	−1086	.462
	.25	6.795	625	−1190	.000	12.09	557	−1087	.467
	.50	9.289	717	−1333	.000	9.38	530	−1038	.478
	.75	10.970	710	−1465	.000	9.27	419	−953	.479
50	.10	4.980	1141	−2303	.000	30.86	1081	−2261	.461
	.25	6.981	1291	−2475	.000	23.60	1156	−2260	.473
	.50	9.566	1450	−2773	.000	17.29	1107	−2162	.479
	.75	11.220	1465	−3050	.000	15.06	870	−1983	.495
100	.10	5.015	2278	−4606	.000	61.67	2162	−4522	.472
	.25	6.889	2574	−4943	.000	45.68	2316	−4524	.476
	.50	9.649	2873	−5526	.000	31.90	2200	−4320	.485
	.75	11.420	2873	−6104	.000	28.39	1720	−3957	.496

NOTE: PD is the effective number of parameters, DIC is the deviance information criterion, LPML is the log pseudomarginal likelihood and BPP is the Bayesian predictive p-value based on the chi-squared measure. The standard errors are negligible.

slightly smaller under the onefold model. However, the coverages of the HPD intervals under the twofold model are closer to the nominal value of 95%, with those from the onefold model being slightly smaller. This is due to the phenomenon that the intervals under the onefold model are narrower.

In Table 6 we present summaries of PD, DIC, LPML, and the BPP. These measures are very similar for the two models. While the BPPs are different, they show that the two models fit equally well. However, the main difference is in PD, the complexity of the model. While the twofold model is more complex than the onefold model, they fit equally well when the onefold model is expected to hold.

Table 5. Simulation for data drawn from the onefold model: Comparison of coverage and widths of 95% HPD intervals and absolute bias, relative absolute bias, and root posterior mean squared error

ℓ	Model	C-HPD	W-HPD	AB	RAB	RPMSE
12	TFM	$.953_{.0019}$	$.244_{.0002}$	$.048_{.0003}$	$.085_{.0007}$	$.084_{.0002}$
	OFM	$.933_{.0023}$	$.223_{.0002}$	$.047_{.0003}$	$.084_{.0006}$	$.079_{.0002}$
25	TFM	$.949_{.0014}$	$.234_{.0001}$	$.049_{.0002}$	$.088_{.0005}$	$.082_{.0002}$
	OFM	$.929_{.0017}$	$.219_{.0001}$	$.046_{.0002}$	$.083_{.0005}$	$.078_{.0002}$
50	TFM	$.948_{.0010}$	$.233_{.0001}$	$.048_{.0002}$	$.086_{.0003}$	$.082_{.0001}$
	OFM	$.943_{.0010}$	$.220_{.0001}$	$.047_{.0002}$	$.083_{.0003}$	$.078_{.0001}$
75	TFM	$.954_{.0008}$	$.236_{.0001}$	$.046_{.0001}$	$.082_{.0003}$	$.081_{.0001}$
	OFM	$.945_{.0008}$	$.221_{.0000}$	$.045_{.0001}$	$.080_{.0002}$	$.077_{.0001}$
100	TFM	$.959_{.0006}$	$.236_{.0001}$	$.046_{.0001}$	$.081_{.0002}$	$.081_{.0001}$
	OFM	$.948_{.0007}$	$.221_{.0000}$	$.045_{.0001}$	$.079_{.0002}$	$.077_{.0001}$

NOTE: TFM is the twofold model and OFM is the onefold model. W-HPD and C-HPD are respectively the width and probability content of a HPD interval. A, AB and RPMSE are the absolute bias, relative absolute bias, and root posterior mean square error. The notation a_b means that a is the estimate and b is the standard error.

Table 6. Summaries of the 1,000 simulation runs (data are drawn from the onefold model) for the complexity, deviance information criterion, log pseudomarginal likelihood, and the Bayesian predictive p-value by ℓ and model

	Onefold				Twofold			
ℓ	PD	DIC	LPML	BPP	PD	DIC	LPML	BPP
12	3.673	250	−530	.445	8.94	233	−531	.661
25	3.362	486	−1055	.553	16.57	460	−1059	.781
50	3.587	1023	−2203	.479	34.43	962	−2208	.772
75	3.574	1527	−3301	.526	52.81	1438	−3308	.852
100	3.692	2043	−4401	.538	70.73	1914	−4410	.882

NOTE: PD is the effective number of parameters, DIC is the deviance information criterion, LPML is the log pseudomarginal likelihood, and BPP is the Bayesian predictive p-value based on the chi-squared measure. The standard errors are negligible.

4. Concluding Remarks

We have developed a twofold hierarchical Bayesian model to analyze binary data arising from a twofold sample design for small areas. This model incorporates an intracluster correlation, and it is an extension of the two-stage hierarchical Bayesian model of Nandram and Sedransk (1993) and, more importantly, the twofold model of Stukel and Rao (1997; 1999) for binary data. A onefold model ignores the intracluster correlation. We have performed a Bayesian predictive inference for the finite population proportion of each area. We have discussed how to study the onefold and twofold small-area models in detail. As an illustrated example, we have used synthetic data from TIMSS, a study of the performance of US students at the third grade in mathematics. We have also performed a simulation study to compare the onefold and twofold models. We have shown how to overcome a difficulty in running the Gibbs sampler that we initially used to fit the twofold model (see TRN14).

We have shown that when there is clustering within each area, the onefold model gives poor performance, and the twofold model is much more preferable. The onefold model can lead to estimators that differ from the twofold model in terms of both location and spread. Our simulation study provides strong evidence that the twofold model is to be preferred when there is a two-stage cluster sampling design within each area. This is a direct consequence of the effect of the intracluster correlation. The Bayesian measures (deviance information criterion, log pseudomarginal likelihood, Bayesian predictive p-value) and frequentist measures (bias, mean squared error, coverage) show that the twofold model is better than the onefold model. While we have demonstrated that the twofold model is preferred when data are available from a twofold sampling design with cluster sampling, other sampling designs (e.g., stratification) in each area will give different results, and these need to be investigated separately.

We have shown that the twofold model is preferable to the onefold model for the TIMSS data. Although the two models give similar results, we have better point and interval estimates from the twofold model. We can see from Table 2 that there are some possibly interesting findings for TIMSS data even though we have not used all features of the data.

Apparently a school in a western rural (WR) area is the best and city schools (NC, SC, CC, WC) are not so good.

This research has opened up many avenues for future work on twofold small-area models. First, for a more realistic analysis of the TIMSS data, it is possible to incorporate the survey weights into our analysis. Second, it may be desirable to have the intracluster correlation to vary with area. It is expected that the computation will be challenging because with a single intracluster correlation there is long-range dependence among the iterates from the Gibbs sampler. Third, it is desirable to study threefold models (states within regions and counties within states). Fourth, we can look at polychotomous data instead of binary data; in TIMSS one can use three levels for mathematics score (below average, average, above average). Fifth, we can consider multivariate binary data; in TIMSS there are both mathematics and science scores. This will lead naturally to consider test of independence for two categorical variables. Sixth, benchmarking for small areas is also an important problem (states within regions and counties within states). Seventh, we can look at covariates via logistic regression; in TIMMS there are covariates. Eight, we can use nonparametric models (e.g., Dirichlet process mixtures and mixture of finite Polya trees) to help robustify our twofold model.

APPENDIX A: Computation Without Gibbs Sampling

Long-range dependence is a general problem for the hierarchical Bayesian model when Markov chain Monte Carlo methods are used to fit it. Typically long-range dependence is due to weak identifiability in some parameters and/or indirect functional relation among the parameters, and this causes poor mixing in the Gibbs sampler. The solution of thinning the iterates, used in practice, is not really efficient. These problems occur when the twofold model is fitted, and so it is pertinent to present an alternative algorithm that uses just random samples.

Our strategy is to use the composition method (i.e., multiplication rule) to draw random samples from the posterior density $\pi(\underset{\sim}{\mu}, \theta, \rho, \gamma|\underset{\sim}{y})$. That is,

$$\pi(\underset{\sim}{\mu}, \theta, \rho, \gamma|\underset{\sim}{y}) = \left\{\prod_{i=1}^{\ell} \pi(\mu_i|\theta, \rho, \gamma, \underset{\sim}{y})\right\} g(\theta, \rho, \gamma|\underset{\sim}{y}).$$

Integrating out $\mu_i, i = 1, \ldots, \ell$, the joint posterior density of $\theta, \rho, \gamma|\underset{\sim}{y}$ is

$$\pi(\theta, \rho, \gamma|\underset{\sim}{y}) = A\left[\prod_{i=1}^{\ell}\left\{\int_0^1 g_i(\mu_i) f(\mu_i) d\mu_i\right\}\right] \theta^{\alpha_o - 1}(1-\theta)^{\beta_o - 1}\rho^{\alpha_o - 1}(1-\rho)^{\beta_o - 1}\gamma^{\alpha_o - 1}$$

$$(1-\gamma)^{\beta_o - 1},$$

where A is a normalization constant hence forth omitted,

$$g_i(\mu_i) = \prod_{j=1}^{m_i} \frac{B\{s_{ij} + \mu_i(1-\rho)/\rho, n_{ij} - s_{ij} + (1-\mu_i)(1-\rho)/\rho\}}{B\{\mu_i(1-\rho)/\rho, (1-\mu_i)(1-\rho)/\rho\}},$$

and

$$f(\mu_i) = \frac{\mu_i^{\theta(1-\gamma)/\gamma-1}(1-\mu_i)^{(1-\theta)(1-\gamma)/\gamma-1}}{B\{\theta(1-\gamma)/\gamma, (1-\theta)(1-\gamma)/\gamma\}}.$$

Note that while $g_i(\mu_i)$ is the ratio of two beta functions (computations discussed earlier) both of which are functions of ρ but not θ and $\gamma, f(\mu_i)$ is a function of θ and γ but not ρ. More importantly, $f(\mu_i)$ is a density function of a beta random variable. We can integrate out the μ_i, one at a time, and form their product to obtain the complete integral. Thus, we only need to discuss how to compute $\int_0^1 g_i(\mu_i)f(\mu_i)d\mu_i$, $i = 1, \ldots, \ell$, for one area. Also, note that $f(\mu_i)$ does not depend on i under the integral sign. While this integral can be computed using Monte Carlo methods, it is much more efficient to use numerical integration in the following way.

Let $F(\cdot)$ denote the cdf corresponding to $f(\cdot)$. Partition the interval (0,1) into a mesh of G subintervals $[a_0, a_1], [a_1, a_2], \ldots, [a_{G-1}, a_G]$ where $a_0 = 0$, $a_i = i/G, i = 1, \ldots, G$. Then, using the Riemann middle sum, it is easy to show that

$$\underset{G\to\infty}{lim} \sum_{v=1}^{G} g_i\left(\frac{a_{v-1}+a_v}{2}\right)\{F(a_v) - F(a_{v-1})\} = \int_0^1 g_i(x)f(x)dx, \ i = 1, \ldots, \ell.$$

Thus, for reasonably large G, $\sum_{v=1}^{G} g_i\left(\frac{a_{v-1}+a_v}{2}\right)\{F(a_v) - F(a_{v-1})\} \approx \int_0^1 g_i(x)f(x)\,dx$, $i = 1, \ldots, \ell$.

Together with integrating out the μ_i, we have also integrated out θ, ρ, where we use Gaussian quadrature via Legendre orthogonal polynomials,

$$p(\gamma|\underset{\sim}{y}) \approx \sum_{g_1=1}^{G}\sum_{g_2=1}^{G} w_{g_1}w_{g_2}\left\{\prod_{i=1}^{\ell}\int_0^1 \pi(\mu_i, x_{g_1}, x_{g_2}, \gamma|\underset{\sim}{y})d\mu_i\right\},$$

where $w_g, g = 1, \ldots, G$, are the weights and $x_g, g = 1, \ldots, G$, are roots of the Legendre polynomial with x_{g_1} and x_{g_2} corresponding to θ and ρ respectively. Note that the single integral over each μ_i is done as described above and the whole procedure is a three-dimensional integral. Now, using univariate grids, samples of the posterior density of γ are obtained in exactly the same manner as described for the onefold model using the univariate grid sampler.

Then, conditional on γ, the posterior density of ρ is

$$p(\rho|\gamma, \underset{\sim}{y}) \approx \sum_{g=1}^{G} w_g\left\{\prod_{i=1}^{\ell}\int_0^1 \pi(\mu_i, x_g, \rho|\gamma, \underset{\sim}{y})d\mu_i\right\}.$$

Again using the univariate grid sampler, samples are drawn from the posterior density of ρ.

Next, conditional on (ρ, γ), the posterior density of θ is

$$p(\theta|\rho, \gamma, \underset{\sim}{y}) \approx \left\{\prod_{i=1}^{\ell}\int_0^1 \pi(\mu_i, \theta|\rho, \gamma, \underset{\sim}{y})d\mu_i\right\}.$$

Again using the univariate grid sampler, samples are drawn from the posterior density of θ.

Finally, conditional on (θ, ρ, γ), the μ_i are independent and samples are again obtained from $\pi(\mu_i | \theta, \rho, \gamma, \underset{\sim}{y})$ using the univariate grid sampler. We have always used 100 grids for the μ_i, θ, ρ and γ.

APPENDIX B: A Multistage Hierarchical Bayesian Model

In TIMSS the countries can be compared, a task beyond the scope of the current article. The small areas (regions and communities) are clustered within the countries and the schools are clustered within these small areas. This is a generalization of the twofold design, which we have discussed in detail in Section 2, to a threefold design. Thus we describe the multistage model mainly for reasons of theoretical interest.

The multifold hierarchical Bayesian model is

$$y_{ij_1,\ldots,j_k} | \mu_{ij_1,\ldots,j_{k-1}} \overset{ind}{\sim} \text{Bernoulli}(\mu_{ij_1,\ldots,j_{k-1}}).$$

For $s = 1, \ldots, k-1$,

$$\mu_{ij_1,\ldots,j_{k-s}} | \mu_{ij_1,\ldots,j_{k-(s+1)}}, \gamma_1 \overset{ind}{\sim} \text{Beta}\left\{ \mu_{ij_1,\ldots,j_{k-(s+1)}} \frac{1-\gamma_1}{\gamma_1}, \ (1-\mu_{ij_1,\ldots,j_{k-(s+1)}}) \frac{1-\gamma_1}{\gamma_1} \right\}.$$

and

$$\mu_i | \theta, \gamma_k \overset{iid}{\sim} \text{Beta}\left\{ \theta \frac{1-\gamma_k}{\gamma_k}, \ (1-\theta) \frac{1-\gamma_k}{\gamma_k} \right\}.$$

Finally, a priori

$$\theta, \gamma_1, \ldots, \gamma_k \overset{iid}{\sim} \text{Uniform}(0,1).$$

Note that in this hierarchical Bayesian model, the first two stages are conjugate and the other stages are nonconjugate. More importantly, the correlation between two units at the first stage is γ_1. Furthermore, when the first-stage means are integrated out, the correlation between two units in two different clusters is γ_2, and so on. It is expected that the correlations will decay as we go down the hierarchical structure of the model. That is, the correlation between two units at the area level is expected to be the smallest while the correlation at the last stage of the multistage cluster sampling design is expected to be the largest.

While the multistage model is of practical importance, it would need significant research to develop it into a useful methodology and it is expected that the computation will be challenging.

5. References

Ando, T. 2007. "Bayesian Predictive Information Criterion for the Evaluation of Hierarchical Bayesian and Empirical Bayes Models." *Biometrika* 94: 443–458. Doi: http://dx.doi.org/10.1093/biomet/asm017.

Bedrick, E.J. 1983. "Adjusted Chi-Squared Tests for Cross-Classified Tables of Survey Data." *Biometrika* 70: 591–595. Doi: http://dx.doi.org/10.1093/biomet/70.3.591.

Brier, S.S. 1980. "Analysis of Contingency Tables Under Cluster Sampling." *Biometrika* 67: 591–596. Doi: http://dx.doi.org/10.1093/biomet/67.3.591.

Calsyn, C., P. Gonzales, and M. Frase. 1999. "Highlights from TIMSS." National Center for Education Statistics, Washington, DC. Doi: http://mces.ed.gov/timss.

Datta, G.S. and M. Ghosh. 1991. "Bayesian Prediction in Linear Models: Applications to Small Area Estimation." *Annals of Statistics* 19: 1748–1770.

Foy, P., K. Rust, and A. Schleicher. 1996. "Sample Design." In *TIMMS Technical Report, Volume I: Design and Development*, edited by M.O. Martin and D.L. Kelly, pagenumber. Chestnut Hill, MA: Boston College.

Fuller, W.A. and G.E. Battese. 1973. "Transformations for Estimation of Linear Models with Nested-Error Structure." *Journal of the American Statistical Association* 68: 626–632. Doi: http://dx.doi.org/10.1080/01621459.1973.10481396.

Gelfand, A., D. Dey, and H. Chang. 1992. "Model Determination using Predictive Distributions with Implementation via Sampling-based Methods." In *Bayesian Statistics* 4, 147–168. New York: Oxford University Press.

Geisser, S. and W. Eddy. 1979. "A Predictive Approach to Model Selection." *Journal of the American Statistical Association* 74: 153–160. Doi: http://dx.doi.org/10.1080/01621459.1979.10481632.

Gelman, A., J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. 2013. *Bayesian Data Analysis*, 3rd ed. New York: Chapman & Hall/CRC.

Ghosh, M. and P. Lahiri. 1988. "Bayes and Empirical Bayes Analysis in Multistage Sampling." In *Statistical Decision Theory and Related Topics IV*, Vol. 1, edited by S.S. Gupta and J.O. Berger. 195–212. New York: Springer.

Hamilton, J. 2009. *President Obama, U.S. Secretary of Education Duncan Announce National Competition to Advance School Reform*. U.S. Department of Education: Available at: http://www.ed.gov/news/pressreleases/2009/07/07242009.html.

Holt, D., A.J. Scott, and P.D. Ewings. 1980. "Chi-Squared Tests with Survey Data." *Journal of the Royal Statistical Society, Series A* 143: 303–320. Doi: http://dx.doi.org/10.2307/2982131.

Malec, D. and J. Sedransk. 1985. "Bayesian Inference for Finite Population Parameters in Multistage Cluster Sampling." *Journal of the American Statistical Association* 80: 897–902. Doi: http://dx.doi.org/10.1080/01621459.1985.10478200.

Molina, I., B. Nandram, and J.N.K. Rao. 2014. "Small Area Estimation of General Parameters with Application to Poverty Indicators: A Hierarchical Bayes Approach." *Annals of Applied Statistics* 8: 852–885. Doi: http://dx.doi.org/10.1214/13-AOAS702.

Nandram, B. 2014. *Bayesian Predictive Inference for a Proportion Under a Two-Fold Small Area Model*. Technical Report, Department of Mathematical Sciences, Worcester Polytechnic Institute, 1–43. (Available on request.)

Nandram, B., D.R. Bhatta, J. Sedransk, and D. Bhadra. 2013. "A Bayesian Test of Independence in a Two-Way Contingency Table Using Surrogate Sampling." *Journal of Statistical Planning and Inference* 143: 1392–1408. Doi: http://dx.doi.org/10.1016/j.jspi.2013.03.011.

Nandram, B. 1998. "A Bayesian Analysis of the Three-Stage Hierarchical Multinomial Model." *Journal of Statistical Computation and Simulation* 61: 97–126. Doi: http://dx.doi.org/10.1080/00949659808811904.

Nandram, B. and J. Sedransk. 1993. “Bayesian Predictive Inference for a Finite Population Proportion: Two-Stage Cluster Sampling.” *Journal of the Royal Statistical Society, Series B* 55: 399–408.

Natarajan, R. and R.E. Kass. 2000. “Reference Bayesian Methods for Generalized Linear Mixed Models.” *Journal of the American Statistical Association* 95: 227–237. Doi: http://dx.doi.org/10.1080/01621459.2000.10473916.

Rao, J.N.K. 2003. *Small Area Estimation*. New York: Wiley.

Rao, J.N.K. and A.J. Scott. 1981. “The Analysis of Categorical Data from Complex Sample Surveys: Chi-squared Tests for Goodness of Fit and Independence in Two-Way Tables.” *Journal of the American Statistical Association* 76: 221–230. Doi: http://dx.doi.org/10.1080/01621459.1981.10477633.

Rao, J.N.K. and A.J. Scott. 1984. “On Chi-Squared Tests for Multi-way Tables with Cell Proportions Estimated from Survey Data.” *Annals of Statistics* 12: 46–60.

Scott, A.J. and D. Holt. 1982. “The Effect of Two-Stage Sampling on Ordinary Least Squares Methods.” *Journal of the American Statistical Association* 77: 848–854. Doi: http://dx.doi.org/10.1080/01621459.1982.10477897.

Scott, A. and T.M.F. Smith. 1969. “Estimation in Multi-Stage Surveys.” *Journal of the American Statistical Association* 101: 1387–1397. Doi: http://dx.doi.org/10.1080/01621459.1969.10501015.

Silverman, B.W. 1986. *Density Estimation for Statistics and Data Analysis*. New York: Chapman & Hall.

Stukel, D.M. and J.N.K. Rao. 1997. “Estimation of Regression Models with Nested Error Regression Structure and Unequal Error Variances Under Two and Three Stage Cluster Sampling.” *Statistics & Probability Letters* 35: 401–407. Doi: http://dx.doi.org/10.1016/S0167-7152(97)86602-3.

Stukel, D.M. and J.N.K. Rao. 1999. “On Small-Area Estimation Under Two-Fold Nested Error Regression Models.” *Journal of Statistical Planning and Inference* 78: 131–147. Doi: http://dx.doi.org/10.1016/S0378-3758(98)00211-0.

Toto, M.C.S. and B. Nandram. 2010. “A Bayesian Predictive Inference for Small Area Means Incorporating Covariates and Sampling Weights.” *Journal of Statistical Planning and Inference* 140: 2963–2979. Doi: http://dx.doi.org/10.1016/j.jspi.2010.03.043.

Yan, G. and J. Sedransk. 2007. “Bayesian Diagnostic Techniques for Detecting Hierarchical Structure.” *Bayesian Analysis* 2: 735–760. Doi: http://dx.doi.org/10.1214/07-BA230.

Yan, G. and J. Sedransk. 2010. “A Note on Bayesian Residuals as a Hierarchical Model Diagnostic Technique.” *Statistical Papers* 51: 1–10. Doi: http://dx.doi.org/10.1007/s00362-007-0111-2.

Received December 2012
Revised September 2015
Accepted October 2015