Journal of Official Statistics, Vol. 21, No. 3, 2005, pp. 411–432

Small Area Estimation from the American Community Survey Using a Hierarchical Logistic Model of Persons and Housing Units

Donald Malec[1]

A multivariate binomial/multinomial model is proposed for estimating poverty and housing-unit characteristics of small areas. The methodology for producing estimates is presented, along with several evaluations using data from the American Community Survey. In one of these evaluations, it is demonstrated that the model produces predicted samples whose within small area design-based estimates of variance are in concordance with the original design-based estimates. It is concluded that this approach can be a viable way to make small area estimates without needing to assume that the design-based estimates of within-small area variance are fixed (as in most area-level models) or that the design-based estimates themselves, are normally distributed. The model introduced proposes a way to incorporate both housing unit information and person level information and may be of use in similar contexts.

Key words: Hierarchical model; logistic parameterization; unit level small area model; full Bayesian analysis; MCMC; Metropolis/Hastings; Gibbs sampling.

1. Introduction

In an effort to provide accurate estimates for census-type aggregations such as tracts on a yearly basis from the American Community Survey (ACS), small area methods can be employed. A hierarchical logistic model of both persons and housing units within-tracts is proposed for making tract-level estimates. Besides providing a model that can be extended to include both person-level and housing-unit covariates, this approach directly accounts for the uncertainty of within-tract variability, a component whose estimate is often regarded as fixed in other small area estimation methods. The "borrowing strength" of an estimate, defined as the degree to which data outside of a small area is used to make a small area estimate, depends crucially on the within and between small area variability. Since within-tract variability is a component that affects "borrowing strength," this approach automatically accounts for the additional uncertainty of unknown within-tract variability in the magnitude of "borrowing strength."

[1] U.S. Bureau of the Census, Statistical Research Division, Room 3132-4, Washington, DC 20233, U.S.A. Email: Donald.J.Malec@census.gov

Acknowledgments: The author would like to express his appreciation to Eric Slud for useful discussions concerning the methods in this article. The author would also like to express his appreciation to the associate editor and three anonymous referees whose thoughtful comments greatly improved the article.

Disclaimer: This article reports the results of research and analysis undertaken by the U.S. Census Bureau staff. It has undergone a U.S. Census Bureau review more limited in scope than that given to official U.S. Census Bureau publications. This article is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

A typical assumption used in small area estimation is that the direct small area estimates are normally distributed with unknown mean but with the corresponding estimated variance treated as fixed and known (see e.g., Rao 1999, Eq. 2.2). This type of small area estimation falls under the more general approach of directly modeling design-based statistics aggregated up to the small area level. This approach has been termed the "aggregate level model" or "area level model" (see Rao 2003, p. 76). In general, any number of aggregate statistics, such as the sample mean and sample variance, could be modeled jointly. However, in practice, the sample variance is often treated as fixed and only the sample mean is modeled. An additional level of modeling linking the unknown small area means together is then used to produce final estimates.

Due to small sample sizes in a small area, variance estimates of the direct estimates may be imprecise and in that case should not be treated as known. The effect of the assumption of error-free direct estimates of small area variance has been of interest for a while. For example, the effect of misspecification of composite estimator weights (which can be functions of variance components) was empirically evaluated by Schaible (1979). There, it was demonstrated that the mean squared error was fairly robust to misspecification of the composite estimator weights. The effects of misspecification of weights on estimates of variance were not evaluated, however. In order to reduce the variability of the within small area estimated variances, Isaki et al. (1991) used models similar to variance curves across small areas to smooth the estimates. However, these smoothed estimates were then treated as fixed in the subsequent small area estimation, ignoring both the new estimate's variability and bias due to the model. By assuming the estimated within small area covariance matrices are distributed approximately Wishart and assuming that some of the parameters of the model can be estimated with negligible error, Otto and Bell (1995) smooth the small area variances using a model that includes a term accounting for small area variability of the true covariance. Instead of modeling the within small area variances with the aim of improving the small area estimates, Wang and Fuller (2003) start with the usual empirical best linear unbiased predictors (EBLUP) and their inherent assumption of fixed variance. Their aim, however, is to provide valid estimates of MSE for an EBLUP by subsequently accounting for the fact that the within small area variances use plug-in estimates.

In addition, the direct small area estimates may not have a distribution near normality. By definition, the small areas have small samples and the usual practice of appealing to normality through large sample asymptotics may not apply.

An alternative approach to accounting for the variance of the direct small area estimates is to model below the small area level. This approach is often referred to as using a "unit level model" (see, Rao (2003), p. 76). Due to the use of within small area models, individual-level data can be used, and error due to an unknown within small area variability can be accounted for and estimates of the finite population parameter with accompanying variances can be made. This last point can be contrasted with the aggregate level approach which assumes that small area finite populations parameters are fixed with no inherent internal variability. The unit level model approach has been shown to be readily applied when one can use hierarchical models of independent distributions. In a Bayesian context, models which correctly account for the unknown variability of within small areas are similar to that proposed by Gelfand et al. (1990). Although normal

hierarchical distributions are often used for location parameters and the inverse gamma distribution for scale, the unit level model allows one, for example, to incorporate generalized linear models to account for discrete data with any variety of hierarchical distributions. A difficulty with the unit level model approach is in selecting a model that accurately accounts for the sample selection process as well as the underlying population distribution. There have been few attempts of using unit level models when the design is informative. Arora and Lahiri (1997) model a sample weighted average of unit level measurements; however, the underlying model is based on the observations being independently, normally distributed.

The model to be used here is a unit level model of both individual and housing unit characteristics. It is a unit level model, devised to account for within housing unit clustering and the discrete, (nonnormal) nature of the observations. This type of model avoids the need to assume that within tract variance estimates are measured without error. This model also avoids making assumptions that tract-level summary statistics are normally distributed. In addition, this model automatically incorporates clustering effects due to the ACS sample selection of entire housing units and may be useful for other cluster-based surveys that have no other sample design components that are informative.

Estimates, and their estimated precision, are produced using Markov chain Monte Carlo (MCMC) methods via a nonsubjective Bayesian approach. The aim of this work is to use a unit level model to provide a small area estimation method that acknowledges the fact that within tract-level sampling error is unknown and is dependent on the clustered sample selection while incorporating this source of error into the inferential framework. A secondary aim is to compare the model and resulting estimates to a close competitor that is based on normal approximations. The resulting model is developed in Section 3. If the normal approximations used are benign, inferences should be the same between the two models. In this case, the approximate model would be preferred on computational grounds. While still requiring MCMC methods, estimation from the approximate model is not as computationally intensive; it requires only Gibbs sampling instead of Gibbs sampling with an additional proposal distribution.

The model used here accounts for possibly different poverty rates for family members in a housing unit (who are either all in poverty or not) and unrelated persons (who have an individual poverty index) living in the same housing unit. The model includes a provision that the poverty status of unrelated individuals may depend on the poverty status of the housing unit's family. In order to account for the sampling variability and to make estimates at the tract level, a hierarchical multinomial model of housing unit characteristics is used. It is shown that the model is good enough to reproduce estimates for within tract variances comparable to a standard jackknife approach while additionally accounting for the model-based variability of these estimates. Model-based (indirect) tract-level estimates of poverty rate, persons per housing unit and occupancy rate are of interest here.

The same ACS test site dataset, used by Chand and Alexander (1999), will be used here. It consists of a sample containing 163 census tracts in Multnomah County, Oregon, collected in 1996. A sampling fraction of 15% was used for this sample. The distribution of the number of sampled housing units varied considerably across tracts. The median

sample size is 192 housing units. About 5% of the sampled tracts have 47, or fewer, housing units in the sample and about 95% have at least 351.

1.1. The American Community Survey (ACS)

The American Community Survey has been developed by the U.S. Census Bureau as a way to obtain information throughout the decade like that obtained in the once-a-decade decennial census long-form. The data collected on the decennial census long-form is in demand and used repeatedly throughout the decade, in spite of its degrading timeliness. The ACS is planned to fill the timeliness gap and provide more up-to-date information on social, economic, housing, and financial characteristics for both the nation and for small geographic areas, such as census tracts. The ACS will be fielded continuously throughout the year via monthly samples. Sampling will be spread, continuously, across the country via a systematic sample of housing unit addresses (with a group quarter component added) obtained from an up-to-date address list. The initial contact to a sampled unit is by mail. Nonresponse followup is by telephone, when a telephone number can be linked to an address. Otherwise, those nonrespondents without a linked telephone number or who do not respond by telephone are subsampled (to defray costs) and followed up by face-to-face interview. The basic sample design is relatively straightforward, resembling the long-form sample but on a smaller scale. As in the long-form, housing units are stratified by governmental units, with housing units in smaller governmental units being oversampled. A systematic sample of housing units, spread over the United States, is then taken. Full implementation for persons not living in group quarters began in January of 2005. At that time, the sampling rate was approximately three million addresses per year. Group Quarter residents will be added to the sampling frame in 2006. Test sites for the ACS have been fielded since 1996. In that first year, four test sites were chosen. This number was expanded to 31 test sites in 36 counties in 1999. In addition, a nationally representative sample, in about 1,200 counties has been taken since 2000, partly as a comparison with the 2000 decennial census long-form results. More details on the ACS design and implementation can be found at the U.S. Census Bureau's web site: http://www.census.gov/acs/. More details on the concept of a continuous measurement survey and its application are provided by Alexander (2002).

1.2. Small area estimation for the American Community Survey

Consisting of a sample of approximately three million housing units annually that are geographically spaced, the size and breadth of the ACS give a new meaning to small area estimation. By executing the continuous measurement concept of Leslie Kish's, as stated in Alexander (1998), the ACS can produce estimates made by accumulating samples over years or across domains. Direct state estimates can be made from the ACS for areas (or domains) that are usually thought of as small areas with respect to the sample sizes in other national surveys. For example, Alexander (1998) recommends that direct annual estimates can be made precisely for areas with a population of 65,000 or larger, and direct estimates can be made for areas with a population between 30,000 and 65,000 by averaging two years of data. He also states that areas with a population less that 15,000 such as census tracts, will require five years of data to make precise direct estimates. Although much can

be done using only direct estimates there may still be a need for indirect estimates of small areas, for example annual estimates of census tracts. To this end, the U.S. Census Bureau has supported a research effort of indirect small area estimation.

Methods for indirect, small area estimation from the ACS began about the same time as the survey began being field tested. As an aid to improve small area estimates of tract-level unemployment rates as measured in the CPS, Chand and Alexander (1995) apply a regression model with an additional random effect for each small area to the arcsine squared root transformation of CPS direct estimates of the unemployment rate. Here, the ACS measurement of the unemployment rate is used as one of the regression variables. (For this article, there was no ACS data available and simulated data was used.) The sampling variances are treated as fixed and known but actually estimated using the standard CPS variance curve formula. Estimates are compared and contrasted using several criteria. An adjustment is made that forces the small area estimates to sum up to one selected larger area design based estimate. Standard diagnostics were applied to the residuals from the model. This basic model and approach are continued in Chand and Alexander (1996) for estimation of ACS tract level estimates of the proportion of persons below poverty. A logistic link function is used in addition to the arcsine squared root link function. Regression variables used are tract-level data from the 1990 census and (simulated) tax filer data. Results using simulated data from the ACS in Alameda County, California are presented. In Chand and Alexander (1999), actual ACS data from three of the four 1996 test sites along with IRS tract level regression variables such as median income and per capita income are used to make estimates using an arcsine squared root link. The within small area variances are estimated by jackknifing the transformed sample proportion by placing one housing unit in its own jackknife cell. These variance estimates are subsequently treated as fixed. In addition to a comparison of four different estimates from the model, small area estimates are made from a subsample and compared with the more precise direct estimates obtained from the whole sample. Limited results of residuals are presented for Brevard County, Florida.

Chand and Malec (2001) subsequently modified the basic arcsine squared root model to include a unit level component in order to model the within tract-level variance. The aim of this work was to determine whether the model was adequate enough to account for the within tract variances, and whether use of the model produced estimates that were different from the estimates produced from the aggregate level model. Both approaches were implemented using Bayesian methods and posterior means were used as estimates. Estimates were made for tract-level poverty rates in Multnomah County, Oregon. Using one housing unit per jackknife cell, as in Chand and Alexander (1999), it was demonstrated that replicate samples produced entirely from the model resembled the actual sample, in terms of its within small area jackknife estimates of variance. However, there were differences between the resulting estimates when using the aggregate or unit level approach. For large sample sizes, the unit level model estimates tended to be closer to the direct estimates than to those from the aggregate model, exhibiting less borrowing from outside the areas but comparable posterior variance. Although there is no gold standard with which to compare the two estimates, the aggregate level model can be viewed as a special case of the unit model since the unit level model was able to reproduce the within variance estimates.

The present work extends Chand and Malec (2001) in several ways. First, a logistic model is used. The arcsine squared root model was difficult to implement because the transformed sample proportions (and transformed parameters) are constrained. A logistic model is much easier to implement. The model was used to produce new, predicted, samples used to help validate the model by comparing the predicted samples to the observed samples. For each predicted sample, direct estimates of the within-tract variance as well as tract level estimates of poverty rate can be made. By predicting a large number of these new samples, confidence intervals of these direct estimates can be made from the model. If these intervals are consistent with the observed data, there is no serious bias that can be observed relative to the amount of error present. To avoid the possibility that the model has just incorporated too much error so that the confidence intervals are large enough to include anything, the gain in the precision in the small area estimates is also evaluated. If the model does not exhibit any serious biases with respect to direct estimates of population moments and if the resulting small area estimates adequately reduce precision we consider the estimates as good candidates for use. Lastly, this work contains a more focused evaluation of the use of the normality assumption employed at the small area level. The model and procedures used are identical except for the use of normal likelihoods instead of binomial and multinomial likelihoods.

2. The Population Model

The American Community Survey is a systematic sample of housing units. It is assumed that the sample of housing units can adequately be approximated by a simple random sample. There may be a selection bias within housing units; e.g., persons within a housing unit may have correlated responses. An extreme example of this is in measuring poverty because poverty is assigned to an entire family, resulting in a perfect correlation among persons in the same family.

Since person characteristics tend to cluster within households, a model that treats individuals as independent observations is inappropriate. A model that can account for some degree of within-housing-unit correlation will be used here to circumvent this problem. Since "borrowing strength" is directly related to the amount of within and between variance, not accounting for this error could bias the results. The housing unit model will automatically adjust borrowing based on the uncertainty of the variance estimates.

Within a state, a two-stage model is employed. A model of housing unit characteristics is postulated. Then, within a housing unit, a model of individual characteristics within a housing unit is provided. In this preliminary development, housing unit size and composition into family members and unrelated housing unit residents are modeled. Subfamilies are considered part of the family and share family characteristics. In this application persons below poverty are of interest. The salient feature of the model is that all members of a family are either in or out of poverty. Unrelated individuals will have their own unique poverty status. However, a model is employed which will account for possible correlation between family poverty status and the poverty status of unrelated individuals within the same housing unit. Further modeling of family characteristics as a function of housing size, demographic characteristics, and so on could be investigated in

the future. As in Chand and Alexander (1995), administrative records are employed to model tract variability of poverty rates.

2.1. Notation and distributional assumptions

In order to utilize tract-level data to estimate possible unique tract-level features, the above models will all have tract-level-specific parameters. A hierarchical model across tracts, within a state, will be specified in order to increase the sample size while estimating common features across tracts.

2.1.1. The housing unit composition model

For each housing unit, h, in tract i, both the housing unit composition and the poverty status of all individuals within the housing unit can be measured. Housing unit composition consists of family size and the number of unrelated persons living in the housing unit. For housing unit (i,h), denote these two counts by c_{fih} and c_{uih}, respectively. This includes vacant housing units $(c_{fih}, c_{uih}) = (0, 0)$. By convention, occupied housing units will always have one, and only one, family. This definition of housing unit composition represents the most basic description of a housing unit's inhabitants needed to define person-level poverty, since an entire family is either in poverty or not and each unrelated individual has his or her own poverty status. The poverty status for all persons in housing unit (i,h), can be described by indicator variables of family composition and of poverty status.

Denote the type of composition of the household by the multiple-valued indicator δ_{ih}

$$\delta_{ih} = k \quad \text{if } (c_{fih}, c_{uih}) = (g_k, u_k)$$

where it is assumed that the T unique types of housing unit composition pairs, (g_k, u_k), $k = 1, \ldots, T$, are identifiable from the sample (or enough are identifiable to be used to approximate the collection of unique types). For the illustration using the 163 census tracts in Multnomah County, Oregon $T = 54$, $0 \le g_k \le 17$, and $0 \le u_k \le 9$.

The distribution of housing unit composition within tract is:

$$P(\delta_{ih} = k | \pi_{ik}) = \pi_{ik} \quad \{\delta_{ih}\}_h \text{ independent given } \{\pi_{ik}\}_k$$

The Bayesian convention of using the "conditioning line" to show when model parameters are considered fixed is followed here.) Conditional on the π_{ik},

$$\sum_{h \in s} I_{[\delta_{ih}=k]} = a_{ik} \tag{1}$$

form sufficient statistics.

The joint distribution of $\underline{a}_i = (a_{i1}, \ldots, a_{iT})$ is multinomial$(a_{i.}, \pi_{i1}, \ldots, \pi_{iT})$ Conditional on $a_{i.}$ the total number of sampled housing units in tract i $(a_{i.} = \sum_{k=1}^{T} a_{ik})$.

Define the transformations:

$$\theta_{ik} = \ln\left(\frac{\pi_{ik}}{\sum_{\ell=k+1}^{T} \pi_{i\ell}}\right); \quad k = 1. \ldots T - 1 \tag{2}$$

As a result, given the number of sampled housing units in tract i and, when the parameters $(\theta_{i1} \ldots, \theta_{i,T-1})$ are fixed, the likelihood of the housing unit parameters can be obtained

from the joint distribution of $\underline{a}_i$, i.e.:

$$p(\underline{a}_i|a_{i.}) \propto \prod_{k=1}^{T-1} \left\{ e^{\theta_{ik} a_{ik}} (1 + e^{\theta_{ik}})^{-\sum_{\ell=k}^{T} a_{i\ell}} \right\} \tag{3}$$

Completing the hierarchical model, specify:

$$\theta_{ik} \sim N(\mu_k, \gamma_k^2); \quad \text{independent}, i, k = 1. \ . \ . \ T-1 \tag{4}$$

Note that a model allowing for some dependence between the θ_{ik} may be more appropriate. Further modeling to discern this was not attempted here. However, as will be shown, the independence model employed appears to be adequate for predicting the attributes of interest in this article.

The Housing Unit Composition model is defined by (3) and (4). Once a prior is provided for $\mu_1, \gamma_1^2, \ldots, \mu_{T-1}, \gamma_{T-1}^2$, Bayesian methods can be employed.

As shown by Hobert and Casella (1996), the use of improper priors may result in improper posteriors when using hierarchical models. Strategies for avoiding this problem while still using diffuse priors have mostly centered on the use of uniform shrinkage priors as demonstrated by Strawderman (1971), Christiansen and Morris (1997) and others. Natarajan and Kass (2000) apply these techniques to a generalized mixed model, and hence, can be readily adopted to the setting here. Specifically, a uniform improper prior will be used for μ_k. In addition, an approximate uniform shrinkage prior will be used for γ_k^2 of the form:

$$pr(\gamma_k^2) \propto \left[1 + \gamma_k^2 \tilde{p}(\zeta_k)(1 - \tilde{p}(\zeta_k)) \sum_{i \in s} \sum_{\ell=k}^{T} a_{i\ell} / a_{i.} \right]^{-2}$$

where, the sample counts, $a_{i\ell}$, are defined in (1), ζ_k are any given known constants (to be specified shortly), and

$$\tilde{p}(\zeta) = e^{\zeta}/(1 + e^{\zeta})$$

Natarajan and Kass (2000) show that the class of priors, in which this prior belongs, are all proper for any ζ_k. In addition, they suggest substituting the MLE of μ_k, based on a fixed effect model (i.e., assuming $\gamma_1^2 = \ldots = \gamma_{T-1}^2 = 0$), into ζ_k. It is recognized that basing any prior on even a part of the data in the likelihood precludes the direct use of Bayes theorem for posterior inference. However, Natarajan and Kass state that treating ζ_k as if it does not depend on the data appears to only have a minor effect on the posterior. Their suggestion is followed here.

2.1.2. The poverty status model

Given the housing unit composition, poverty status (i.e., 1 for being in poverty and 0 otherwise) will be defined for the family as a whole using the indicator x_{Fih}. Specifically, x_{Fih} equals 1 if the family is in poverty and equals 0, otherwise. Similarly, all unrelated family members (if any) requires their own indicator, x_{Uih1}, x_{Uih2}, . . . to denote their poverty status. The set of indicators, $\underline{x}_{ih} = (x_{Fih}, x_{Uih1} \ \ldots, \ x_{Uihu_{\delta_{ih}}})$ denotes the poverty status of the entire housing unit.

The distribution of family poverty is defined as independent Bernoulli:

$P(x_{Fih} = 1|p_{0i}) = p_{0i}$

The poverty status of the unrelated individuals within a housing unit is also independent Bernoulli, given the associated family's poverty status. Specifically, the probability that an unrelated individual is in poverty, given that the housing unit's family is in poverty is denoted by p_{Pi}. Similarly, the probability that an unrelated individual is in poverty, given that the housing unit's family is not in poverty is denoted by p_{Ni}. Similarly,

$$P(x_{Uihj} = 1|x_{Fih}, p_{Pi}, p_{Ni}) = \begin{cases} p_{Pi}, & \text{if } x_{Fih} = 1 \\ p_{Ni}, & \text{if } x_{Fih} = 0 \end{cases}$$

Completing the model between tracts, define the logits:

$$\ln\left(\frac{p_{0i}}{1 - p_{0i}}\right) = \underline{z}'_i\underline{\beta} + t_i$$

$$\ln\left(\frac{p_{Pi}}{1 - p_{Pi}}\right) = \underline{z}'_i\underline{\beta} + t_i + \nu_P \quad \text{and}$$

$$\ln\left(\frac{p_{Ni}}{1 - p_{Ni}}\right) = \underline{z}'_i\underline{\beta} + t_i + \nu_N$$

where $\underline{z}_i$ are tract-level covariates available for all tracts and

$t_i \sim N(0, \sigma^2)$

The $\underline{z}_i$ are the known tract-level IRS covariates used by Chand and Alexander (1995) in modeling poverty status:

$z_{i1} = 1$
$z_{i2} = \ln$ (median income)
$z_{i3} = \ln$ (per capita income)
$z_{i4} = \ln(Q_L)$
$z_{i5} = \ln(Q_U)$
$z_{i6} = 2\sin^{-1}\sqrt{P_V}$, where Q_L, Q_U, and P_V are respectively, the lower quartile income, the upper quartile income, and the proportion of persons below poverty level in the tract.

Functionally independent uniform, improper priors are used for $\underline{\beta}$, ν_P, and ν_N. As with the housing unit model, an approximate uniform shrinkage prior, see Natarajan and Kass (2000), is used for σ^2. In this case,

$$pr(\sigma_k^2) \propto \left[1 + \sigma_k^2 \sum_{i \in s} \{n_{0i}\tilde{p}_{0i}(1 - \tilde{p}_{0i})\, n_{Pi}\tilde{p}_{Pi}(1 - \tilde{p}_{Pi})\, n_{Ni}\tilde{p}_{Ni}(1 - \tilde{p}_{Ni})\}/n_{H_i}\right]^{-2}$$

Here,

$$\tilde{p}_{0i} = e^{\underline{z}'i\underline{\beta}}/(1 + e^{\underline{z}'_i\underline{\beta}})$$

$$\tilde{p}_{Pi} = e^{\underline{z}_i\underline{\beta} + \tilde{\nu}_P}/(1 + e^{\underline{z}'_i\underline{\beta} + \tilde{\nu}_P}) \text{ and}$$

$$\tilde{p}_{Ni} = e^{\underline{z}'_i\underline{\beta} + \tilde{\nu}_N}/(1 + e^{\underline{z}'_i\underline{\beta} + \tilde{\nu}_N})$$

where $\tilde{\underline{\beta}}$, $\tilde{\nu}_p$, and $\tilde{\nu}_N$ are the MLE estimates of the person level model with all $t_i = 0$. Lastly, n_{0i}, n_{Pi}, and n_{Ni} are the number of sampled families, the number of unrelated persons in sampled housing units of families in poverty, and the number of unrelated persons in sampled housing units of families not in poverty, respectively.

Given n_{0i}, the number of sampled families in tract i (i.e., the number of occupied sampled housing units), sufficient statistics for the joint distribution of $\{\underline{x}_{ih}\}_{h\in s}$ are:

$$m_{0i} = \sum_{h\in s} x_{Fih}$$

$$n_{Pi} = \sum_{h\in s} x_{Fih} u_{\delta_{ih}}$$

$$m_{Pi} = \sum_{h\in s} x_{Fih} \sum_{j=1}^{u_{\delta_{ih}}} x_{Uihj}$$

$$n_{Ni} = \sum_{h\in s} (1 - x_{Fih}) u_{\delta_{ih}}$$

$$m_{Ni} = \sum_{h\in s} (1 - x_{Fih}) \sum_{j=1}^{u_{\delta_{ih}}} x_{Uihj}$$

The likelihood of the person level model parameters can be obtained from the joint distribution of m_{0i}, m_{Pi}, m_{Ni}, n_{Pi}, and n_{Ni}, i.e.:

$$p_{0i}^{m_{0i}}(1 - p_{0i})^{n_{0i}-m_{0i}} p_{Pi}^{m_{Pi}}(1 - p_{Pi})^{n_{Pi}-m_{Pi}} p_{Ni}^{m_{Ni}}(1 - p_{Ni})^{n_{Ni}-m_{Ni}} \tag{5}$$

The complete likelihood is the product of the two likelihoods in (3) and (5), since the distribution of person level outcomes was specified conditionally on the housing unit characteristic outcomes.

3. An Approximate Model

As stated in the introduction, a normality assumption based on an appeal to large sample size is often applied to the direct estimates in each small area, even though these sample sizes are small. To evaluate the effects of this assumption, empirically, for this example, the closest normal approximation to the model used here, will be employed and evaluated.

The following model uses the two approximations repeatedly.

Approximation 1. For a sample proportion $\hat{p} = m/n$, where $m \sim binomial(n, p)$, approximately

$$\sin^{-1}\sqrt{\hat{p}} \sim N(\sin^{-1}\sqrt{p}, 1/4n)$$

Approximation 2. When $p(\mu) = e^{\mu}/(1 + e^{\mu})$ and $\hat{\mu}$ is a consistent estimator of μ, the Taylor linearization of $p(\mu)$ provides an adequate approximation:

$$\sin^{-1}\sqrt{p(\mu)} \approx \sin^{-1}\sqrt{p(\hat{\mu})} - \hat{\mu}\tfrac{1}{2}\sqrt{p(\hat{\mu})(1 - p(\hat{\mu}))} + \mu\tfrac{1}{2}\sqrt{p(\hat{\mu})(1 - p(\hat{\mu}))}$$

Applying these approximations to the housing unit composition model, define $\tilde{\mu}_k$ to be the MLE from the fixed effect model specified by $\theta_{ik} = \mu_k$ instead of assuming θ_{ik} has a distribution as in (2).

Define $g_{ij} = \sin^{-1}\sqrt{a_{ij}/\sum_{\ell=j}^{T} a_{i\ell}}\; j = 1, \ldots, T-1$. Using the two approximations one may infer that

$$g_{ij} \sim N\left(c_j + b_j\theta_{ij}, 1 \Big/ \left(4\sum_{\ell=j}^{T} a_{i\ell}\right)\right)$$

where $b_j = \frac{1}{2}\sqrt{p(\tilde{\mu}_j)(1-p(\tilde{\mu}_j))}$, $c_j = \sin^{-1}\sqrt{p(\tilde{\mu}_j)} - \tilde{\mu}_j b_j$, and $p(\tilde{\mu}_j) = e^{\tilde{\mu}_j}/(1+e^{\tilde{\mu}_j})$.

The resulting housing unit component of the likelihood is approximated by the normal distribution:

$$p(g_{i1}, \ldots, g_{i(T-1)} | \theta_{i1}, \ldots, \theta_{i(T-1)}) \propto \prod_{j=1}^{T-1} e^{-2\left[\sum_{\ell=j}^{T} a_{i\ell}\right](g_{ij}-[c_j+b_j\theta_{ij}])^2}$$

Expanding around the MLE estimates $\tilde{\underline{\beta}}$, $\tilde{\nu}_p$, and $\tilde{\nu}_N$ of the person level model with all $t_i = 0$, one has the following approximation to the joint distribution of the person level model:

$$P(g_{0i}, g_{Pi}, g_{Ni}, n_{Pi}, n_{Ni}) \propto e^{-2n_{0i}(g_{0i}-[c_{0i}+b_{0i}(\underline{x}'_i\underline{\beta}+t_i)])^2} \times e^{-2n_{Pi}(g_{Pi}-[c_{Pi}+b_{Pi}(\underline{z}'_i\underline{\beta}+t_i+\nu_P)])^2} \times e^{-2n_{Ni}(g_{Ni}-[c_{Ni}+b_{Ni}(\underline{z}'_i\underline{\beta}+t_i+\nu_N)])^2}$$

where $g_{0i} = \sin^{-1}\sqrt{m_{0i}/n_{0i}}$, $g_{Pi} = \sin^{-1}\sqrt{m_{Pi}/n_{Pi}}$, $g_{Ni} = \sin^{-1}\sqrt{m_{Ni}/n_{Ni}}$ and

$$b_{0i} = \tfrac{1}{2}\sqrt{p(\underline{z}'_i\tilde{\underline{\beta}})(1-p(\underline{z}'_i\tilde{\underline{\beta}}))}$$

$$c_{0i} = \sin^{-1}\sqrt{p(\underline{z}'_i\tilde{\underline{\beta}})} - \underline{z}'_i\tilde{\underline{\beta}}b_{0i}$$

$$b_{Pi} = \tfrac{1}{2}\sqrt{p(\underline{z}'_i\tilde{\underline{\beta}}+\tilde{\nu}_P)(1-p(\underline{z}'_i\tilde{\underline{\beta}}+\tilde{\nu}_P))}$$

$$c_{Pi} = \sin^{-1}\sqrt{p(\underline{z}'_i\tilde{\underline{\beta}}+\tilde{\nu}_P)} - (\underline{z}'_i\tilde{\underline{\beta}}+\tilde{\nu}_P)b_{Pi}$$

$$b_{Ni} = \tfrac{1}{2}\sqrt{p(\underline{z}'_i\tilde{\underline{\beta}}+\tilde{\nu}_N)(1-p(\underline{z}'_i\tilde{\underline{\beta}}+\tilde{\nu}_N))}$$

$$c_{Ni} = \sin^{-1}\sqrt{p(\underline{z}'_i\tilde{\underline{\beta}}+\tilde{\nu}_N)} - (\underline{z}'_i\tilde{\underline{\beta}}+\tilde{\nu}_N)b_{Pi}$$

The distributions of the remaining parameters of the model are specified identical to the exact model given in Section 2.

4. Finite Population Parameters of Interest

For tract i, estimates of the per capita poverty rate, the average number of persons per household and the occupancy rate can be estimated using the model and accompanying data. These three population characteristics can be expressed in terms of the model in the previous section.

Let k_0 be the vacant household composition indicator (i.e., $(g_{k_0}, u_{k_0}) = (0,0)$). The population housing unit occupancy rate is defined as:

$$OCR_i = 1 - \frac{\sum_{h=1}^{H_i} I_{[\delta_{ih}=k_0]}}{H_i}$$

The number of persons per housing unit in tract i can be written as:

$$PPH_i = \frac{\sum_{h=1}^{H_i} (g_{\delta_{ih}} + u_{\delta_{ih}})}{H_i}$$

Lastly, the per capita poverty rate in tract i can be described as:

$$POVR_i = \frac{POV_i}{\sum_{h=1}^{H_i} (g_{\delta_{ih}} + u_{\delta_{ih}})}$$

where the total number of persons in poverty in tract i is defined as:

$$POV_i = \sum_{h=1}^{H_i} x_{Fih} g_{\delta_{ih}} + \sum_{j=1}^{u_{\delta_{ih}}} x_{Uihj}$$

5. Estimation

Estimates of both location and scale will be made using Bayesian predictive inference. Briefly, the predictive distribution of all unsampled indicators that make up OCR_i, PPH_i, and $POVR_i$ is obtained based on the model assumptions and the posterior distribution of the model parameters. The posterior distributions are obtained using one block at a time MCMC algorithm (Chib and Greenberg 1995) with either Metropolis/Hastings steps or Gibbs sampling steps within blocks.

Specifically, variates from the full conditional posterior of the θ_{ik}'s and the t_i's are obtained one at a time using a normal proposal function with mean and variance corresponding to the posterior mode and Hessian of the posterior and a Metropolis/Hastings rejection step. Variates from the joint, full conditional posterior distribution of $\underline{\beta}$, ν_P, and ν_N are obtained in a similar manner using their posterior mode and corresponding Hessian. The conditional posterior distribution of μ_k is normal and can be sampled from directly. As in Natarajan and Kass, the conditional posterior of the variance components are sampled by using an inverse gamma proposal distribution obtained by replacing the approximate uniform shrinkage prior with Jeffreys' prior (i.e., where the prior of log of the variance component is constant). This is followed by a Metropolis/Hastings rejection step. Gibbs samplers are used for the new features in the approximate model. The computational burden on developing estimates from the approximate model is much less than that of the exact model. For both models inference was possible after discarding the first 500,000 iterations and using the next 700,000 for estimation. One long chain was run. The burn-in period (i.e., the iteration number in which the MCMC variates are treated as actual samples from the posterior) was assessed visually by plotting the posterior parameters versus the MCMC iteration number. Posterior means were estimated by averaging all iterations together after the burn-in period. To reduce the effects of correlated data, posterior variances were obtained by calculating the sample variance based on every 100-th observation. The resulting 100 sample variances, each based on 7,000 data values, were then averaged to make the final estimate. There is a variety of methods for assessing both burn-in time and the MCMC sample size needed; each has apparent limitations. In a production setting, a number of these methods (including assessing multiple runs, Gelman and Rubin (1992)) should be

presented. In addition, more efficient methods such as the use of "Rao-Blackwellization" could be employed to speed up the estimation.

Due to the priors used, specialized software in Fortran was developed. However, the procedure could have been implemented in SAS IML or Matlab. Using other priors, WinBugs could be used.

5.1. Assessment of the model using the observed and predicted sample

Since a novel model for within tract variance that incorporates both housing unit level and person level characteristics is being advocated, the first assessment is how well the model describes the within tract variances. One way of assessment is to examine how well the model can reproduce the original estimates derived from the observed data (e.g., see Gelman et al. (1995), Section 6.3) By using the model to generate a new set of sample data, the distribution of a jackknife estimate of the variance of the arcsine squared root transformation of the tract sample poverty rate can be empirically estimated. The jackknife used is based on housing units to account for within housing unit correlation. Specifically, the variance of the arcsine squared root of the sample proportion of persons in poverty, in a tract, is obtained for each sampled tract. This is accomplished by first randomly grouping housing units (the sampling unit) into jackknife cells. The arcsine squared root transformation is used because of its variance stabilizing property. Figure 1 compares 95% simultaneous coverage intervals, Besag et al. (1995), from the model-based predictive distribution of the jackknife standard deviation with the actual jackknife standard deviations from the original sample. (Note: since many tracts have nearly the same sample

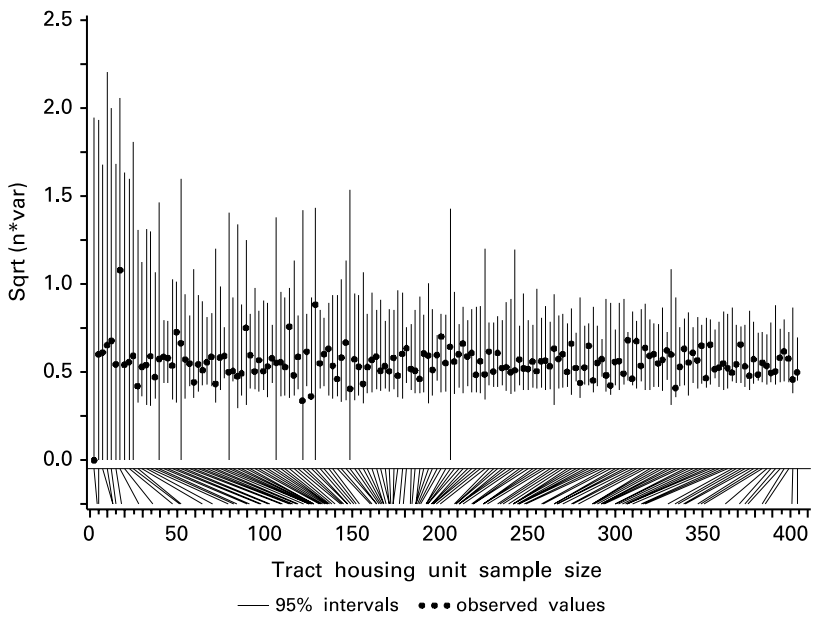

Fig. 1. Observed jackknifed estimates of variances vs. 95% simultaneous prediction intervals for each sampled tract (ordered uniformly by sample size)

size, their confidence sets become obscured. The following four plots are spread out to reveal each confidence set. An (identical) key, at the base of each plot, is included to enable the reader to assess each tract's sample size.) As can be seen, the model provides good coverage of the observed jackknife estimates, indicating that the model can replicate the within tract variances well. One can conclude that the model used is at least consistent with the within tract variation observed. This figure also shows the degree of error of the jackknife estimate of variance, as measured with the model. This is one example of error that is often completely ignored and unaccounted for in aggregate-level small area estimation modeling. In this example, this error can be sizable. The model presented here both accounts for this error and automatically incorporates it into the final small area estimates. The tracts are ordered by sample sizes and the increase in error as the within tract sample size decreases is apparent.

The design-based estimates of tract-level poverty rate per person (povr), number of persons per housing unit (pph) and occupied housing unit rate (occr) are similarly compared to their model-based predictive distribution in Figures 2, 3, and 4, respectively. As can be seen, each tract-level design-based estimate is, at least, comfortably covered by the 95% simultaneous coverage intervals. As can be seen in Figures 2 and 3 there are tracts which exhibit outlying estimates of poverty and of persons per housing unit, which the model is still able to predict. It should be noted that, under full implementation, the ACS will contain a nonzero sample size for every tract. Hence, the assessment in this section should be adequate. However, other surveys may require estimates for small areas with no sample size. In this situation, cross-validation leaving entire small areas out, should be carried out. By including cross validation, one can assess how well the model can predict for small areas that are outliers but have no sample size at all.

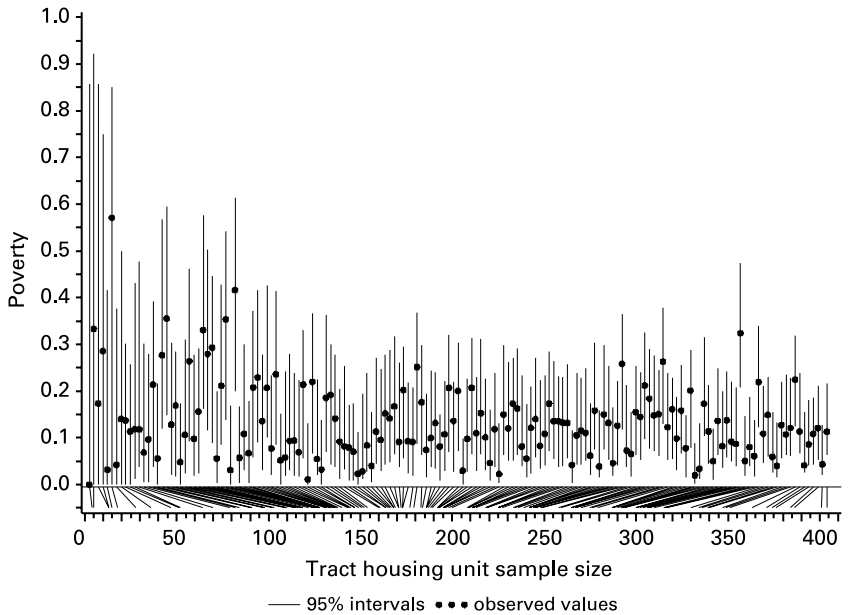

Fig. 2. Direct estimates of poverty vs. 95% simultaneous prediction intervals for each sampled tract (ordered uniformly by sample size)

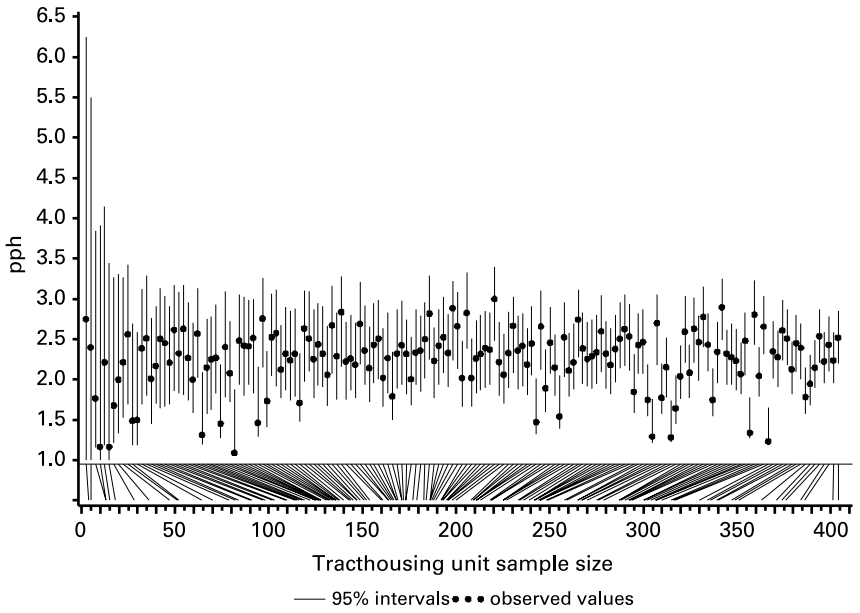

Fig. 3. Direct estimates of persons per HU vs. 95% simultaneous prediction intervals for each sampled tract (ordered uniformly by sample size)

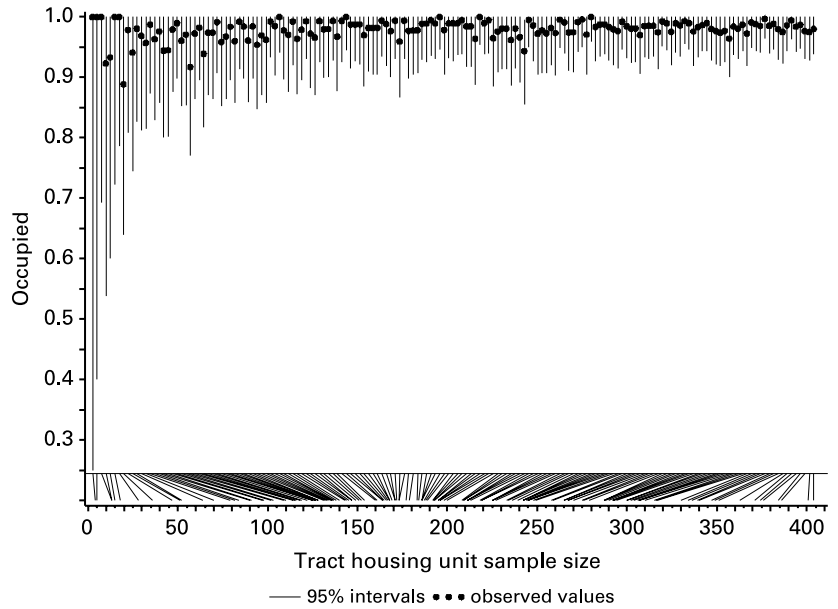

Fig. 4. Direct estimates of HU occupation rate vs. 95% simultaneous prediction intervals for each sampled tract (ordered uniformly by sample size)

Although this type of assessment does not rule out better models (with smaller confidence intervals), it is a way to rule out serious model failures. This assessment does not rule out the possibility that the model used is over-parameterized and produces large probability intervals due to a poorly estimated model. As will be seen in Section 6, this is

not the case; the posterior variances and posterior coefficients of variation (CV) accompanying the key small area estimates are reasonably small.

6. Small Area Estimates

The purpose of modeling this data is to provide small area estimates at the tract-level with accompanying precision. Figures 5, 6, and 7 provide posterior means of tract-level poverty rate, tract-level persons per housing unit, and tract-level occupancy rate, respectively. These tract-level estimates are ordered by tract housing unit sample size and sample estimates are included as a reference. As typically seen with hierarchical models, the model based estimates deviate less from the sampled estimates as the within tract sample size increases, a product of decreased borrowing as the sample size increases. In all three graphs, the exact model and the approximate model estimates are closer together for large sample sizes. This illustrates that the approximation holds well for the large tracts coupled with the fact that any differences due to borrowing outside of the tract from different models is minimized for large samples. Although differences are apparent between the two models for the smaller tracts, agreement is relatively close. The average absolute relative error due to using the approximate model for estimates are 6.2% for estimated poverty rate, 1.1% for estimated persons per housing unit and .2% for estimated occupancy rate.

The differences between the posterior variances from the approximate and the exact model are more apparent, as seen in Figures 8, 9, and 10. The approximate model tends to underestimate the accuracy of the poverty rate but overestimate the accuracy for both average persons per housing unit and occupancy rate. It is not surprising that the approximate model can overestimate the variance because the approximation, while based on large sample theory, does not ignore any source of error. Using the approximate model

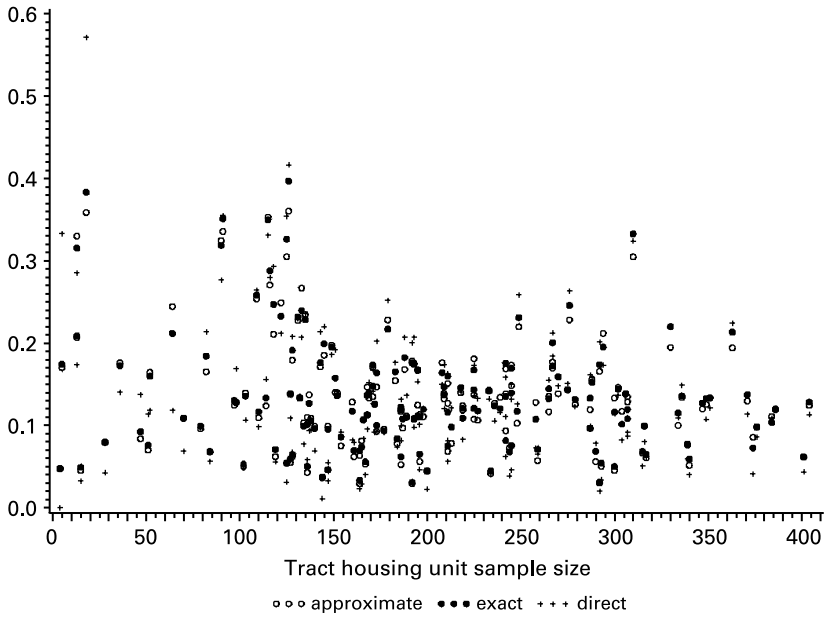

Fig. 5. Tract estimates of poverty rate

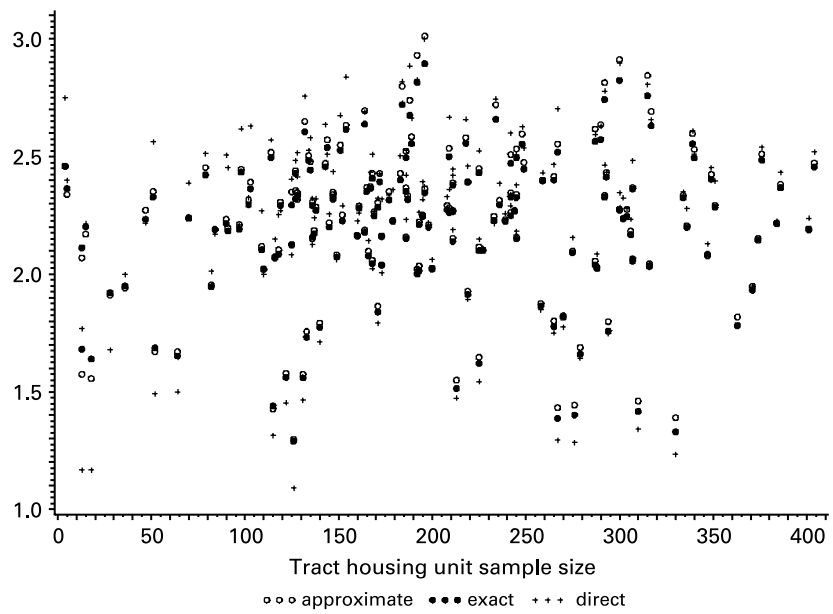

Fig. 6. Tract estimates of the average persons per housing unit

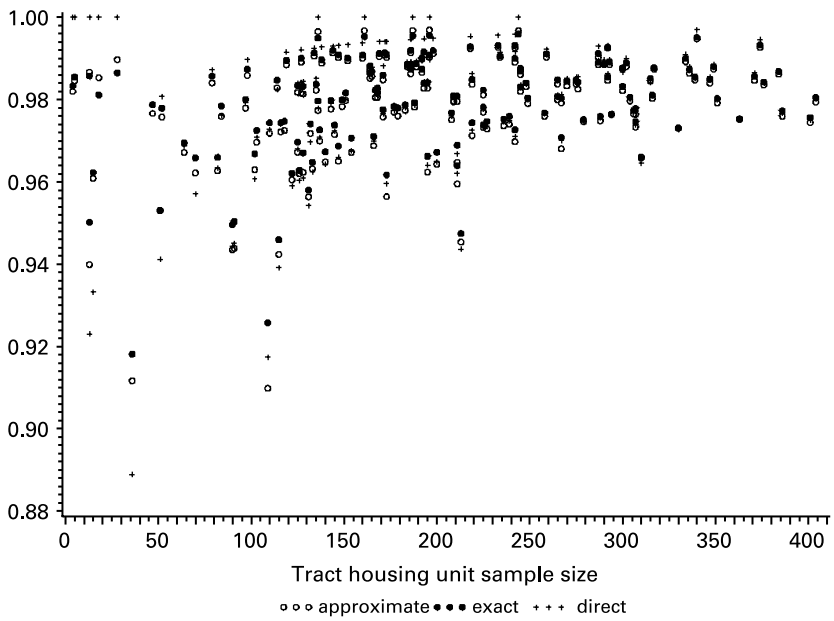

Fig. 7. Tract estimates of occupancy rate

to provide estimates of accuracy can be problematic, as evidenced in this example. The average absolute relative error due to using the approximate model for variance estimates are 57.9% for estimated poverty rate variance, 109.6% for estimated persons per housing unit variance, and 36.3% for estimated occupancy rate variance.

Now that it is seen that estimates of variability are different between the exact and approximate model, there is still the question of which model is better. By some lucky

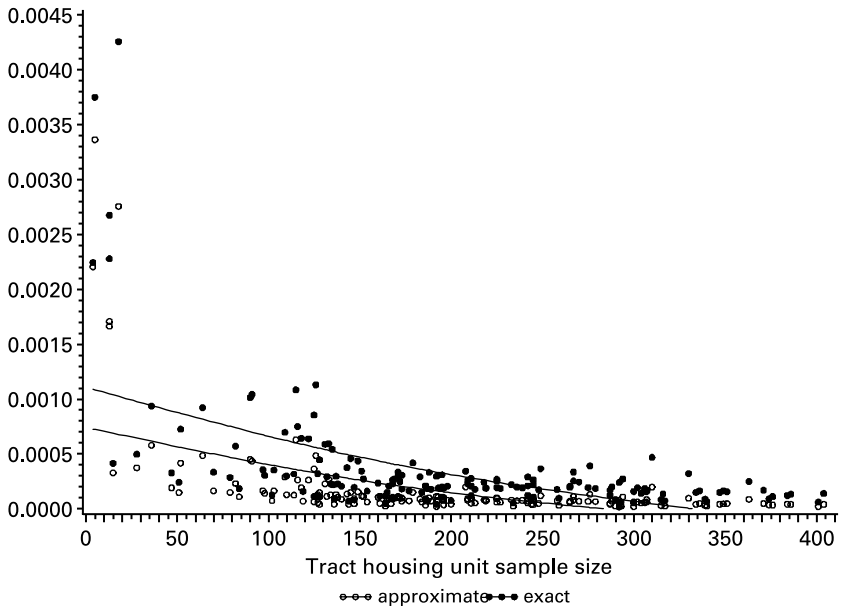

Fig. 8. Posterior variances of tract poverty rate

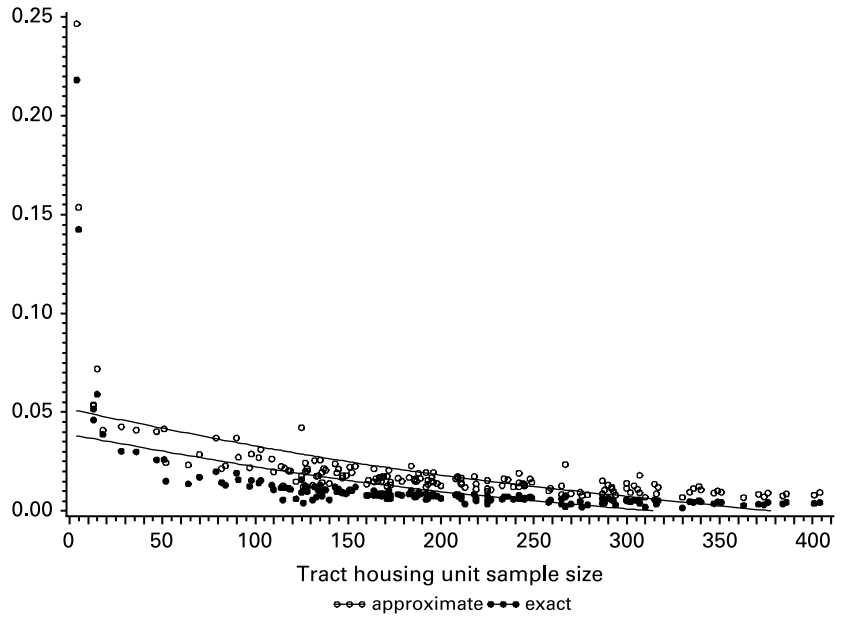

Fig. 9. Posterior variances of average persons per housing unit

combination of errors it is possible that the approximate model actually improves on deficiencies in the exact model. Although models that are better than either the exact or approximate model outlined here are possible, a comparison of the fit of these two models will still be informative. As a model fitting criterion, the Bayesian predictive model selection approach of Laud and Ibrahim (1995) is used. In particular, their "L-criterion" is

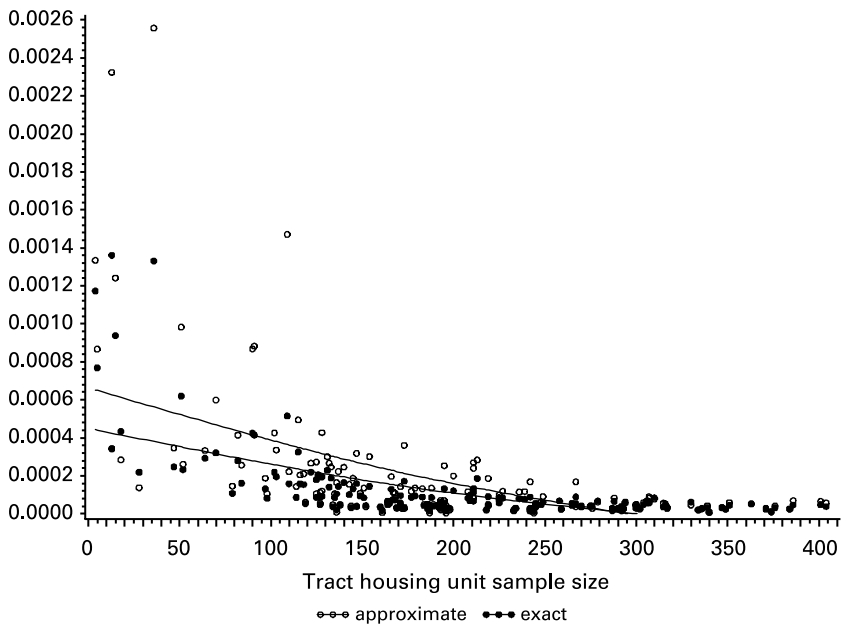

Fig. 10. Posterior variances of occupancy rate

used which is a measurement of the squared root of the expected sum of squared differences between observed tract-level sample statistics and their predictions from the respective models. This criterion reflects the mean squared error of the predictions but other criteria can be used. As seen in Table 1, the exact model provides a better fit than the approximate model for sampled estimates of $\widehat{OCR}_i$, $\widehat{PPH}_i$ and $\widehat{POVR}_i$, which are defined as the sample-based counterparts to the finite population parameters of Section 4:

$$\widehat{OCR}_i = 1 - \frac{\sum_{h\in s} I_{[\delta_{ih}=k_0]}}{n_{H_i}}$$

$$\widehat{PPH}_i = \frac{\sum_{h\in s} g_{\delta_{ih}} + u_{\delta_{ih}}}{n_{H_i}} \quad \text{and}$$

$$\widehat{POVR}_i = \frac{POV_i}{\sum_{h\in s} g_{\delta_{ih}} + u_{\delta_{ih}}} \quad \text{where}$$

$$\widehat{POV}i = \sum_{h\in s}^{H_i} x_{Fih} g_{\delta_{ih}} + \sum_{j=1}^{u_{\delta_{ih}}} x_{Uihj}$$

Table 1. L-criteria for comparing models

Sample statistic	Exact model	Approximate model	Percent difference
$\widehat{OCR}_i$	.26175	.27553	−5.3%
$\widehat{PPH}_i$	2.59695	2.81266	−8.3%
$\widehat{POVR}_i$	.68389	.75263	−10.1%
$\widehat{SD}_i$	1.55288	1.53563	1.1%

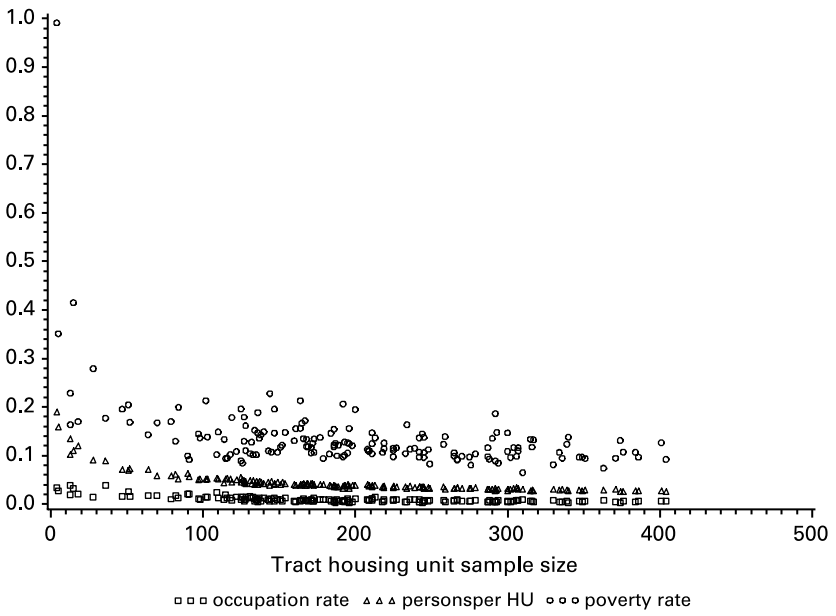

Fig. 11. Posterior CV of estimates

For estimating the within tract sampled standard error, $S\hat{D}_i$, as measured by the squared root of the tract-level sample size times the jackknife tract variance, it can be seen that the approximate model does slightly better but that the percent difference is small relative to the other comparisons. Overall, the exact model appears to provide a much better fit to the observed data. Other models (and even other approximations) may provide a better fit than the two used here. However, the exact model used here does appear to capture the salient features of the data. Also, the approximate model may produce estimates of precision, which are very different from the exact model.

For most tracts and most estimates, the exact model provides estimates with adequate precision for many purposes. Figure 11 lists the posterior CV's (i.e., the squared root of the posterior variance divided by the posterior mean) for the key estimates and tracts. Most poverty rate estimates have a CV between 20% and 30% with a few exceptions. Estimates of occupation rate and of persons per housing unit generally have a CV below 10%.

7. Conclusions

In general, a model and a method of validating the model that does not require knowing the actual population values of the small areas is presented. Modeling allows one to make more efficient estimates, if the model is correct. Here, the correctness of the model is assessed by comparing predicted sampled quantities with those actually observed. If developed further, more evaluations could be made, however. For example, through simulation from known population models, the frequentist properties of the estimates should be reasonable. In addition, the model used here has assumed that data is missing at random within a census tract.

A model describing housing unit composition and person level outcomes was formulated using a joint multinomial/binomial model. The primary goal of providing a methodology to provide estimates of both level and accuracy for small areas, without making restrictive assumptions about the within small area variance, was achieved. The approximate model, while still requiring MCMC methods for estimation, is much simpler to work with and estimates can be made via Gibbs sampling, as opposed to the Metropolis/Hastings proposal for the complete model. As demonstrated, the approximation provides relatively accurate estimates of location but poor estimates of scale. In general, the exact model also provides a better fit to the sampled data.

The multinomial/binomial logistic hierarchical model used here could be adapted to many of the outcomes from the American Community Survey. This model framework can be extended to include more covariates. For example, more dependence between household composition and person-level outcomes could be built in. Housing unit composition such as demographics could be included. Such models could provide for small area estimates of demographic groups. As mentioned in Section 1, correlation among household composition within a tract could be included. Because of the relatively simple design of the ACS, the only major deviation of the sample collection from simple random sampling has been accounted for in the model. In addition the multinomial and binomial models with logistic link functions lend themselves to data modeling due to the variety of software available.

As demonstrated, the exact model provides an adequate fit to the observed data (based on the posterior predictions of sampled statistics), and generally provides precise small area estimates (based on posterior CV's). Satisfying both of these requirements suggests that the model and methodology may be developed to produce defensible small area estimates.

8. References

Alexander, C.H. (1998). Recent Developments in the American Community Survey. Proceedings of the American Statistical Association, Section on Survey Research Methods, 92–100.

Alexander, C.H. (2002). Still Rolling: Leslie Kish's "Rolling Samples" and The American Community Survey. Survey Methodology, 28, 35–41.

Arora, V. and Lahiri, P. (1997). On the Superiority of the Bayesian Method over the BLUP in Small Area Estimation Problems. Statistica Sinica, 7, 1053–1064.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian Computation and Stochastic Systems. Statistical Science, 10, 3–41.

Chand, N. and Alexander, C.H. (1995). Indirect Estimation of Rates and Proportions for Small Areas With Continuous Measurement. Proceedings of the American Statistical Association, Section on Survey Research Methods, 549–554.

Chand, N. and Alexander, C.H. (1996). Small Area Estimation with Administrative Records and Continuous Measurement. Proceedings of the American Statistical Association, Section on Survey Research Methods, 870–875.

Chand, N. and Alexander, C.H. (1999). Indirect Estimation Based on Administrative Records and the American Community Survey. Proceedings of the American Statistical Association, Section on Survey Research Methods, 871–876.

Chand, N. and Malec, D. (2001). Small Area Estimates from the American Community Survey Using a Housing Unit Model. Proceedings of the Federal Committee on Statistical Methodology Research Conference.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. The American Statistician, 49, 327–335.

Christiansen, C.L. and Morris, C.N. (1997). Hierarchical Poisson Regression Modeling. Journal of the American Statistical Association, 92, 618–632.

Gelfand, A.E., Hills, S.E., Racine-Poon, A., and Smith, A.F.M. (1990). Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling. Journal of the American Statistical Association, 85, 972–985.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). Bayesian Data Analysis. Chapman and Hall.

Gelman, A. and Rubin, D.B. (1992). Inference from Iterative Simulation Using Multiple Sequences. Statistical Science, 7, 457–472.

Hobert, J.P. and Casella, G. (1996). The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models. Journal of the American Statistical Association, 91, 1461–1473.

Isaki, C.T., Huang, E.T., and Julie, H.T. (1991). Smoothing Adjustment Factors from the 1990 Post Enumeration Survey. Proceedings of the American Statistical Association, Section on Social Statistics, 338–343.

Laud, P.W. and Ibrahim, J.G. (1995). Predictive Model Selection. Journal of the Royal Statistical Society, Series B, 57, 247–262.

Natarajan, R. and Kass, R.E. (2000). Reference Bayesian Methods for Generalized Linear Mixed Models. Journal of the American Statistical Association, 95, 227–237.

Otto, M.C. and Bell, W.R. (1995). Sampling Error Modeling of Poverty and Income Statistics for States. Proceedings of the American Statistical Association, Government Statistics Section, 160–165.

Rao, J.N.K. (1999). Some Recent Advances in Model-based Small Area Estimation. Survey Methodology, 25, 175–186.

Rao, J.N.K. (2003). Small Area Estimation. Hoboken, New Jersey: John Wiley and Sons, Inc.

Schaible, W.A. (1979). A Composite Estimator for Small Area Statistics. In Synthetic Estimates for Small Areas: National Institute on Drug Abuse Research Monograph Series 24. DHEW Publication No. (ADM)79-801, Chapman and Hall.

Strawderman, W.E. (1971). Proper Bayes Minimax Estimators of the Multivariate Normal Mean. The Annals of Mathematical Statistics, 42, 385–388.

Wang, J. and Fuller, W.A. (2003). The Mean Squared Error of Small Area Predictors Constructed with Estimated Area Variances. Journal of the American Statistical Association, 98, 716–723.

Received October 2003
Revised February 2005