SMALL AREA ESTIMATES FROM THE AMERICAN COMMUNITY SURVEY USING A HOUSING UNIT MODEL

Nanak Chand and Donald Malec
U.S. Bureau of the Census

Abstract

The American Community Survey (ACS) is designed to, ultimately, provide census long-form information on a continuous basis. Although the aim of providing current socio-economic data on the population can be realized, a yearly sample size equal to the traditional census long-form sample size would be prohibitively expensive.

The aim of this work is to produce a small area estimation method that accounts for the sample design and does not assume that the within tract variance is estimated without error. In future work, this model can be easily extended to incorporate more covariates, at any of the levels, or to include data collected at previous times.

Keywords: Hierarchical Model, Arcsine square root transformation, Unit level Small Area Model

Introduction

In an effort to provide estimates for census-type aggregations such as tracts, on a yearly basis, small area methods can be employed. Recent review articles on small area estimation methods include Marker (1999) and Rao (1999). We propose a hierarchical model of persons within housing units within tracts for making tract level estimates. Besides developing estimates from this model, we investigate possible gains of this approach over inferences from a standard model that assumes that the estimated within tract sampling variances are known. The purpose of modeling person characteristics, within housing units, within tracts is to be able to estimate and specify the variability of the within tract sampling error and resulting effects on small area estimates. Comparisons of estimates are made assuming that the, more complex, housing unit model is true. The amount of borrowing is also evaluated. In addition, predictions of design-based tract-level summaries are compared with the actual sampled data. Also, the fit of the model as a description of the within tract variability is evaluated graphically. Based on the above comparisons, the utility of using the housing unit model will be assessed.

Estimates, and their estimated precision, are produced using Monte Carlo Markov Chain methods via a non-subjective Bayesian approach. As an illustration of the method, we generalize the model used by Chand and Alexander (1995) for making tract-level estimates of the percent of persons in poverty. Their model specifies a tract-level linear relationship between the arc-sine square root of the proportion of persons in poverty and tract-level income characteristics from income tax returns. We incorporate this model into one that models persons in a housing unit via a family (who are either all in poverty or not) and unrelated persons (who have an individual poverty index) living in the same housing unit. Our 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, we include a hierarchical multinomial model of housing unit characteristics. The same data set, as used by Chand and Alexander, consisting of a sample containing 163 Oregon census tracts, collected in 1996 will be used. A sampling fraction of 15% was used for this sample. The median within-tract sample size is 192 housing units. About 5% of the sampled tracts have 47, or fewer, housing units in sample and about 95% have a sample size of at least 351

The Population Model

The American Community Survey is a systematic sample of housing units. Because a systematic sample of housing units is selected, it is assumed that there is no sample selection bias at the housing unit level. There may be a selection bias within housing units. We propose a model of persons within housing unit to account for a possible selection bias due to correlation within housing units.

Since person characteristics tend to cluster within household, 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. An alternative approach is to use estimates based on a simple random sample but adjust the variance to take into account the cluster sample. This latter approach is employed by Chand and Alexander (1995) who use a jackknife method to adjust variance. Although this latter approach provides an appropriate adjustment of variance, an empirical Bayes type approach is employed and the adjustments are treated as known. Any sampling error of these variance estimates is not accounted for in deriving an estimate. Since borrowing strength is directly related to the amount of within and between variance, not accounting for this error could bias the results. By contrast, the housing unit model will automatically adjust borrowing based on the uncertainty of the variance estimates. Estimates from the two approaches will be compared.

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 as part of the family and share family characteristics. In this application persons below poverty are of interest. Here, the salient features of the model are 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, etc. could be investigated in the future. As in Chand and Alexander, administrative records are employed to model tract variability of poverty rates.

Heuristically, the model for an individual's poverty status depends on whether he or she is a family member, or not:

$$P(\text{person is in poverty} \mid \text{in a family}) = \text{P(family is in poverty)}$$

$$P(\text{unrelated person is in poverty} \mid \text{family poverty status}) = \text{P(person is in poverty} \mid \text{family poverty status)} \times \text{P(family poverty status)}$$

In order to estimate the poverty rate or count at the individual level, a model for the number and composition of housing unit residents is needed. A multinomial model with probabilities of the form: $P(HU\ contains exacly\ f\ \ family\ members and\ u\ unrelated\ persons) = \text{P}_{fu}$ will be used.

The Within Tract-level Population Model

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. Further hierarchies, e.g., across states, can be included. However, only Oregon is used in this analysis, so State is not specified here.

The Within Tract-level Housing Unit Composition Model

Formally, within tract, i:

$a_{i1},\ldots,a_{iT}$ ~Multinomial $(\text{a}_{\text{i.}}, p_{i1},\ldots,p_{iT})$ where

T: the number of unique housing unit compositions, in sample.

- The "T" types consists of the unique pairs, (f,u), of family size, f, and number of unrelated persons,

u, in a housing unit. This includes vacant housing units k=(0,0). By convention, occupied housing units will have at least one family.

- It is assumed that this set of unique housing unit types is diverse enough to represent the population of unique housing unit types.

a_{ik} : The number of housing units in tract i who have composition of type "k".

p_{ik} : The associated probability that a housing unit in tract i is of composition type "k".

An alternative but equivalent specification of the housing unit model is to define a multivariate indicator random variable, $\underline{d}_{ih} = (d_{ih1}, \ldots, d_{ihT})$, such that

$d_{ihk} = \begin{cases} 1, \textit{ if hu composition is type } k = (f_k, u_k) \\ 0, \textit{ otherwise} \end{cases}$, and

$\underline{d}_{ih}$ ~ multivariate Bernoulli ($p_{ih1}, \ldots, p_{ihT}$), independent.

The Within Tract-level Poverty Status Model

Within tract, i, within an occupied housing unit, h, of type k=(f,u):

I_{ih}, $m_{i.h}$ ~Bernoulli(p_{i0}) binomial $(u,\ p_{ip}I_{ih} + p_{ip}(1-I_{ih})$, where,

I_{ih} : 1/0 indicator of whether family in housing unit h in tract i is/is not in poverty.

$m_{i.h}$: then number of unrelated persons in unit h in poverty

p_{i0} : tract level probability of family poverty status

p_{ip} : tract level probability of poverty status of unrelated persons in housing units with families in poverty.

p_{iN} : tract level probability of poverty status of unrelated persons in housing units with families not in poverty.

Note that the above model specifies two types of dependences within a housing unit; that family member poverty status is all or nothing and that the poverty status of an unrelated individual is dependent on the poverty status of the family residing in the same housing unit.

In summary, the likelihood within tract, i, is proportional to

$$p_{i0}^{m_{i0}}(1-p_{i0})^{n_{i0}-m_{i0}}\ p_{ip}^{m_{ip}}(1-p_{ip})^{n_{ip}-m_{ip}}\ p_{iN}^{m_{iN}}(1-p_{iN})^{n_{iN}-m_{iN}}\prod_{k=1}^{T} p_{ik}^{a_{ik}}$$, where

n_{io}= the number of occupied housing units in sample, in tract i

m_{io} = the number of families in poverty in sample, in tract i

n_{ip} = the number of unrelated persons living with families in poverty in sample, in tract i

m_{ip} = the number of unrelated persons in poverty living with families in poverty in sample, in tract i

n_{iN}=the number of unrelated persons living with families who are not in poverty, in sample, in tract i

m_{iN} = the number of unrelated persons in poverty living with families who are not in poverty in sample, in tract i.

The counts of housing unit types a_{ik} have been defined.

The Between Tract-level Population Model

The between tract-level model is specified as a distribution of the tract-level parameters:
p_{i0}, p_{ip}, p_{iN} and p_{ik} .

The Between Tract-level Housing Unit Composition Model

A hierarchical, multinomial distribution will be specified for the distribution of housing unit types within tract. A spherical transform (a generalization of the arcsine, square root transform) on the multinomial probabilities is used because covariates can be included, relatively easily (unlike multinomial / Dirichlet models) and because it can be generalized to accommodate probabilities with mass at zero or one (unlike logistic transforms).

Define the spherical transformation of the multinomial probabilities

$$p_{i1} = \sin^2 q'_{i1}$$

$$p_{ij} = \sin^2 q'_{i1} \prod_{r=1}^{j-1} \cos^2 q'_{ir} \text{ , } 1 < j < T$$

$$p_{iT} = \prod_{r=1}^{T-1} \cos^2 q'_{ir}$$

Further, define $-\infty < q_{ij} < \infty$, such that

$$q'_{ij} = \begin{cases} 0 & q_{ij} \le 0 \\ q_{ij} & 0 < q_{ij} < \frac{p}{2} \\ \frac{p}{2} & \frac{p}{2} \le q_{ij} \end{cases}$$

Allowing q_{ij} to range over the real line enables one to model zero probabilities (and one's, too) with positive point mass. It is expected that many tracts will not have housing units of certain types. This type of model will be able to represent these cases.

There will be very little data in each tract to estimate the parameter of the multinomial model. A hierarchical model between tracts is utilized to borrow data by letting

$$q_{ij} \sim N(m_j, g_j^2), \qquad \text{ind., i, j = 1,.., T - 1.}$$

The specifications for the housing unit model are completed with the independent priors for the $m_j's$ and $g_j's$.

$m_j \sim N(m_{mj}, s^2_{mj})$

$g_j^{-2} \sim$ Gamma ($d_?$, $d_?$).

The parameters, s^2_{mj} and d? are chosen so that they have a negligible effect on estimation and other inference.

The Between Tract-level Poverty Status Model

Borrowing of information data on poverty status parameters across tracts will be achieved in two ways. First, a regression relationship across tracts based on available covariates will be postulated. Second, random tract effects will be included to capitalize on any remaining similarities of the parameters across tracts.

Define,

$p_{0i} = \sin^2(\underline{x}_i' \underline{b} + t_i)$, if $0 < d_{0i} = \underline{x}_i' \underline{b} + t_i < \frac{p}{2}$

$p_{pi} = \sin^2(\underline{x}_i' \underline{b} + t_i + n_p + z_{pi})$, if $0 < d_{pi} = \underline{x}_i' \underline{b} + t_i + n_p + z_{pi} < \frac{p}{2}$

and

$p_{Ni} = \sin^2(\underline{x}_i' \underline{b} + t_i + n_N + z_{Ni})$, if $0 < d_{Ni} = \underline{x}_i' \underline{b} + t_i + n_N + z_{Ni} < \frac{p}{2}$.

If $d_{zi} \notin (0, \frac{p}{2})$,

Define

$$p_{zi} = \begin{cases} 0, & if\ d_{zi} \le 0 \\ 1, & if d_{zi} \ge \frac{p}{2} \end{cases}.$$

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

$x_{i1} = 1$,
$x_{i2} = \ln$(median income)
$x_{i3} = \ln$(per capita income)
$x_{i4} = \ln(Q_L)$
$x_{i5} = \ln(Q_U)$
$x_{i6} = 2\sin^{-1}\sqrt{P_V}$, 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.

t_i is a random tract effect
$?_p$, $?_N$ are fixed effects denoting the influence of a family's poverty status on unrelated individuals in the housing unit
z_{pi}, z_{Ni} are the corresponding tract-level random effects of a family's poverty status influence on unrelated persons in the housing unit.

The hierarchical model for poverty status is completed by defining

$\underline{w}_i = (t_i, z_{pi}, z_{Ni})'$ and specifying that
$\underline{w}_i \sim N(\underline{0}, 3)$.

Independent priors for the location parameters, $\underline{a} = [\underline{ß}', ?_p, ?_N]$, and the scale parameters, 3^{-1}, are specified as

$\underline{a} \sim N(\underline{\mu}_a\ V_a)$, and

$3^{-1} \sim$ Wishart $\left(d_\Sigma, \frac{1}{d_\Sigma} M_\Sigma\right)$.

As with the housing unit composition model, the parameters, V_a and d_Σ, are chosen so that the prior has a negligible effect on the resulting inference.

Estimation

Ultimately, tract level estimates of the person-level poverty rate, with estimates of precision, are needed. Estimates are based on the availability of precise estimates of the total number of housing units, H_i, in each tract (available as adjusted counts from the sampling frame).

Given the total number of housing units in tract i, the poverty rate in tract i is:

$$\text{POVR}_i = \sum_{h=1}^{H_i}\left[\left(\sum_{k=1}^{T} f_k d_{ihk}\right) I_{ih} + m_{ih}\right] \Big/ \sum_{h=1}^{H_i}\left[\left(\sum_{k=1}^{T} f_k d_{ihk}\right) + u_k\right], \text{ where}$$

f_k and u_k are, respectively, the family size and number of unrelated persons contained in a household with composition of type, k.

The distribution of $POVR_i$ is completely specified from the model of section 2. The posterior predictive distribution will also be proper since only proper priors are used. The posterior mean and variance of $POVR_i$ will be determined and used as estimates of location and scale. Although an analytical equation is not available, these estimates are made numerically.

Inference of all model parameters will be made from their posterior distribution. This will be accomplished using Markov Chain Monte Carlo methods successively applied to the conditional posterior distributions of the parameters. In particular, the adaptive rejection algorithm will be used on the conditional posterior of the parameters, q_{ij}, $\underline{b}$, n_N, n_p, t_i, z_{Ni} and z_{pi}. The conditional posterior distributions of the remaining parameters are all either Normal or Wishart distributions or the Gibbs sampler is used.

A Tract-Level Model

As mentioned in the introduction Chand and Alexander used a model of data aggregated at the tract level to make estimates of poverty. Unlike the housing unit model, the tract-level model does not account for the uncertainty of the within tract variance when estimating poverty rates and associated precisions. This feature that may affect the amount of borrowing and may affect the total precision of the resulting estimates. However, the extra effort of using the housing unit model may not be necessary and estimates from both models are compared to each other to assess whether is are any practical differences between the two.

For the purposes of comparison, we will use the following tract-level model:

$$g_i \sim N(\underline{x}_i' \underline{b} + d_i, \hat{v}_i^2)$$
$$d_i \sim N(0, t^2), \text{ independent,}$$

where

$g_i = 2\sin^{-1}(\sqrt{\hat{p}_i})$,

$\hat{p}_i$, is the weighted estimate of person-level poverty rate, in tract i.

$\hat{v}_i^2$, is estimated sample variance of $\hat{p}_i$ obtained using a jackknife on housing units. Although this is an estimate based only on data from tract, i, it is assumed to be fixed and have no variability.

Note that the tract level model implies that $E(\hat{p}_i | \underline{b}, d_i) = \sin^2(\underline{x}_i' \underline{b} + d_i) = p_{0i}$.

For comparison purposes, a Bayesian analysis will be used for this model, also. Keeping the type of inference the same for both models will provide a more even comparison. As with the housing-unit level model, over-dispersed priors are assigned to the remaining parameters:

$\underline{b} \sim N(\underline{\mu}_b, V_b)$, and
$\tau^{-2} \sim \text{Gamma}(a, b)$.

By definition, the tract-level model is not specified below the tract level. Hence, estimates poverty rate based a predictive distribution of unsampled housing units cannot be obtained. Instead, the posterior mean and variance of p_{0i} will be used as estimates of the location and scale of poverty rate

Note that the tract-level model, as specified, cannot provide tract-level estimates of the total number of persons in poverty because the population size has not been included in the model. The housing unit model includes a model for population and can also be used to estimate the total number of persons in poverty. In order to estimate poverty counts from a tract-level model an additional model of population counts will need to be fit.

Model comparison

The housing unit model has been formulated to model the within tract variance. Since the tract-level model variances are obtained empirically, there is no one-to-one corresponds between the two models at the tract level. To evaluate the adequacy of the housing unit model, predictive samples are generated from the model and resulting predictive jack-knifed estimates of variances are obtained. If the predictive distribution contains the sampled jack-knifed estimates with high probability, the model will be deemed adequate.

Above the tract level, the aggregate-level model is a special case of the housing unit model. This can be seen as follows. First, define the sample poverty rate as:

$\hat{P}_{\mathrm{i}} = \sum_{h \in s} [f_{ih} I_{\mathrm{i}h} + m_{ih}] \Big/ \sum_{h \in s} [f_{ih} + u_{ih}]$, where f_{ih} and u_{ih} are the corresponding housing unit family size and number of unrelated persons.

By definition, $E_A(\hat{P}_{\mathrm{i}} \mid p_{0i}, \ f_{ih}, u_{ih}) = p_{0i}$ for the aggregate model.

For the unit-level model:

$$E_U(\hat{P}_{\mathrm{i}} \mid p_{0i}, p_{pi}, p_{Ni}, \ f_{ih}, u_{ih}) = \sum_{h \in s} \left[f_k p_{0i} + (p_{0i} p_{pi} + (1 - p_{0i}) p_{Ni}) u_k \right] \Big/ \sum_{h \in s} [f_k + u_k] .$$

If poverty status is homogenous within housing unit model (i.e., $n_p = n_N = 0$ and $\Sigma = 0$) then

$p_{0i} = p_{pi} = p_{Ni}$. In this case

$E_U(\hat{P}_{\mathrm{i}} \mid p_{0i}, p_{pi}, p_{Ni}, \ f_{ih}, u_{ih}) = p_{0i}$.

Results

All plots are for sampled tracts, only. In addition, each plot presents tract results sorted by the size of the tract sample. To see clearly the results for each tract, without using a lot of space, the results are presented by the tract sample size order. Figure 6 show the correspondence between actual sample size and the order presented in the other figures.

Figure 1 show the differences between estimates of poverty rate from the two models. Taking the housing unit model as the truth, 95% posterior probability intervals are calculated for each tract and compared with the posterior mean of poverty rate, using the tract-level model. As can be seen, using the posterior mean from the tract-level model can be far off from the posterior distribution based on the housing unit model. It matters which model is used. Figure 2 plots the posterior means of the poverty rate from both models along with the tract level sample mean. This figure illustrates the importance of sample size, in that the three estimate's values become close as the tract sample size increases and less borrowing takes place. For most of the tracts the housing unit model takes values closer to the sample means than the tract level model showing that, in general, the housing model borrows less. Even though the housing unit model borrows less and includes more parameters, the coefficients of variation are comparable between the two models. The housing unit model does have larger CV's when the sample size gets very small.

Figures 4 and 5 look more closely at the usefulness of modeling the within tract variability. Based on the housing-unit model a new sample can be predicted, the arcsine square root of the sample tract-level mean and jackknifed estimate of variances can be calculated. Figure 4 provides 95% probability intervals for the jackknifed standard deviations (i.e., the square root of the jackknifed variances). As can be seen, the sample standard deviations may be very imprecise for small sample sizes. Although this is not a major problem for this data set, the actual ACS is expected to only take a 2-3% sample (instead of the 15% target

taken here). Figure 5 is an informal check on the adequacy of the housing unit model of within tract variability. Here, 95% predictive probability intervals for jackknifed standard deviation multiplied by the square root of sample size is presented. If the model is any good, it should at least be a good predictor of the actual within-tract sample standard deviation (see, e.g. Gelman, et al. (1995), section 6.3). As shown in Figure 5, the predictions are fairly good.

Discussion

The housing unit model has been specified in order to account for the within tract level error of tract level sampling variances, an important error to measure since it a major factor in setting the borrowing strength of small area estimates. A Bayesian implementation has been presented here but a frequentist analysis, such as Maximum Likelihood Estimation, could have been carried out using the same model. The housing unit model can also be expanded to include other terms or be applied to other situations. It could easily be applied at the county or higher level. Additional hierarchical models, such as state effects, could be added. A housing unit model, of this type, could also incorporate housing unit composition rates from the decennial census, relying on the ACS to update changes from the census.

It has been shown that the tract-level model is a special case of the housing unit model at and above the tract level and it has been demonstrated, empirically, the housing unit model does an adequate job of modeling within tract variability. However, more model refinement could be made. First, dependence of poverty or other outcomes on housing characteristics such as size, demographic composition, etc. could be refined. The utility of using transformations other than the arcsine square root could also be evaluated. Also, related structure among types of housing unit characteristic may simplify the model.

References

Chand, Nanak and Alexander, Charles H. (1995). “Indirect Estimation of Rates and Proportions for Small Areas With Continuous Measurement”. ASA Proceedings of the Section on Survey Research Methods, 549-54.

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

Marker, David A. (1999). “Organization of Small Area Estimators Using a Generalized Linear Regression Framework”, Journal of Official Statistics, 15, 1-24.

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

Figure 1. Comparison of Estimates of Tract Poverty Rate

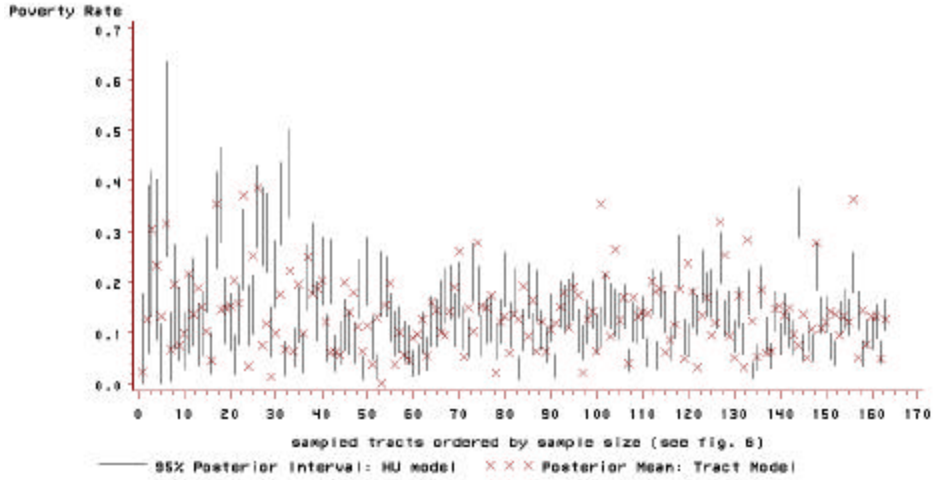

Figure 2. Comparison of Estimates with Tract–level Sample Proportions

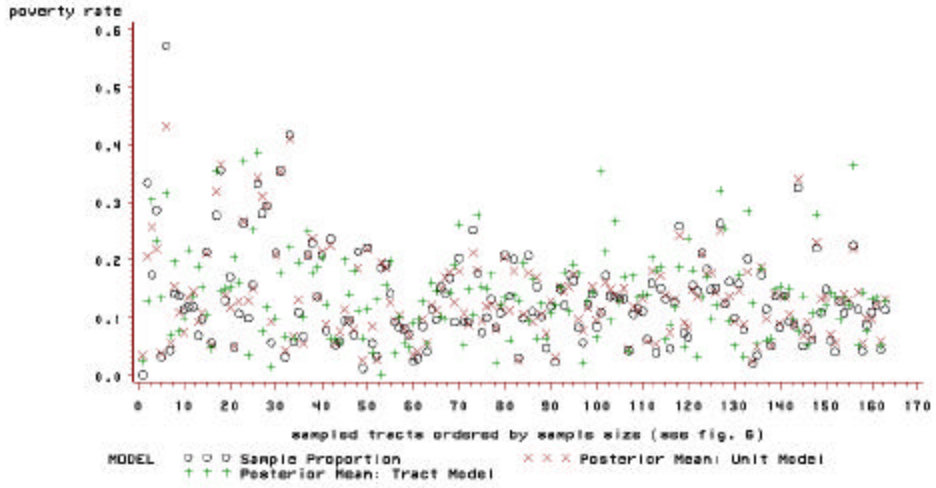

Figure 3. Posterior CV of Estimated Tract Poverty Rate

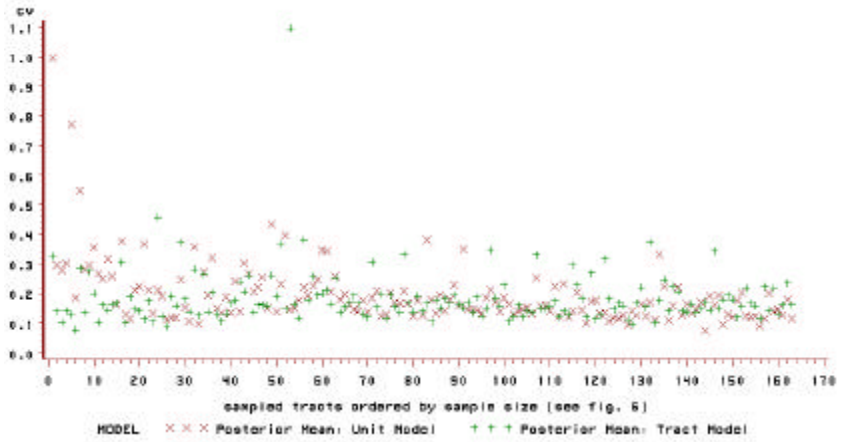

Figure 4. Predicited 95% Probability Intervals for Jacknifed Estimates

of the sampled standard deviation (s.d.) of the arcsine square root of sampled proportions

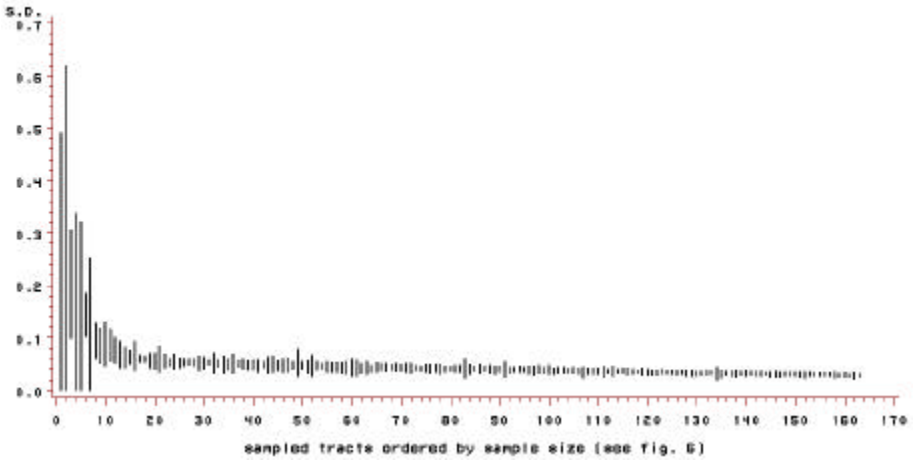

Figure 5. Comparison of Actual Jacknifed Variances v.s. Predicted 95% Intervals

using the posterior predictive distribution from the housing unit model

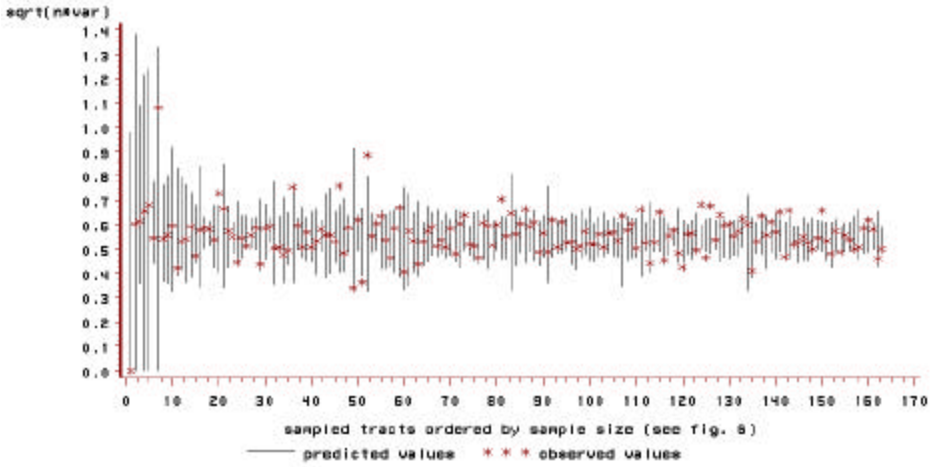

Figure 6. Relationship Between Sample Size and Order