RESEARCH REPORT SERIES
(*Statistics #2003-08)*

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

Donald Malec

Statistical Research Division
U.S. Bureau of the Census
Washington D.C. 20233

Report Issued: November 24, 2003

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

Donald Malec

Abstract

A multivariate binomial/multinomial model is proposed for estimating poverty and housing-unit characteristics of small areas. It is demonstrated that the model is in concordance with the design in that the model can reproduce within small area design-based estimates of variance. The methodology for producing estimates is presented, along with several evaluations using data from the American Community Survey. 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 variance are fixed (as in most area-level models).

Keywords: Hierarchical Model, Logistic parameterization, Unit level Small Area Model, Full Bayesian Analysis

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. 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. 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.

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 (e.g. see Rao (1999), eq 2.2). A model 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. In addition, the direct estimates may

not have a distribution near normality. The effect of these assumptions on small area estimates has been of interest for a while. For example, the effect of mis-specification of composite estimator weights (which can be functions of variance components) was empirically evaluated by Schaible (1979). There, it was demonstrated that mean squared error was fairly robust to misspecification of the composite estimator weights. The effects of mis-specification of weights on estimates of variance was not evaluated, however. In order to reduce the variability of the within small area estimated variances, Isaki et Al. (1999) used models across small areas to reduce the variance. Variability of the actual variances across small areas was not accounted for in 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.

The model to be used is a unit level model of both individual and housing unit characteristics. 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 the ACS sample selection mechanism.

Estimates, and their estimated precision, are produced using Monte Carlo Markov Chain methods via a non-subjective Bayesian approach. The aim of this work is to use a unit-level model to produce a small area estimation method that acknowledges the fact that within tract-level sampling error is unknown and incorporates this extra source of error into the inferential framework. A secondary aim is to provide an approximate method that is not as computationally intensive to use. To this end, the difference between these estimates and estimates based on an approximation of a directly comparable aggregate-level model are assessed.

A hierarchical model of persons within housing units within tracts is proposed for making tract level estimates. Unlike Chand and Malec (2001), where hierarchical models based on the arcsin square-root transformation are used, logistic transformations are used here. It is shown that the model can provide estimates for within tract variances comparable to a standard jack-knife approach while additionally accounting for the model-based variability of these estimates. Estimates of tract-level poverty rate, persons per housing unit and occupancy rate are made from the model. In addition, these

estimates are contrasted with estimates arising from a comparable aggregate-level model. This comparison will document the effects of assuming large sample Normality of transformed proportions and using Taylor series linear approximations. Unlike Chand and Malec, where the competing aggregate-level model consisted of a tract-level aggregate with the estimated within-tract level variance treated as known, the aggregate model used here is picked to be close to the unit-level model. This aggregate level model does not substitute estimated variances as fixed numbers. Instead, variance stabilizing transformations relevant to the assumed model are used.

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. Both the arc-sine square root based model of Chand and Malec and the logistic-based model, used here, are generalizations of the model used by Chand and Alexander (1995) for making tract-level estimates of the percent of persons in poverty. The same data set, as used by Chand and Alexander, will be used here. It consists of a sample containing 163 Oregon census tracts, 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 sample and about 95% have at least 351.

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 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. 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 as part of the family and share family characteristics. In this application persons below poverty are of interest. Here, 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, etc. 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 their 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}:

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

where it is assumed that the T unique types of housing unit composition pairs, (g_k, u_k), k=1,...,T , are identifiable from the sample (or enough are identifiable to be used to approximate the collection of unique types). The distribution of housing unit composition within tract is:[1]

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

Conditional on the π_{ik}, $\sum_{h\in s} I_{[\delta_{ih}=k]} = a_{ik}$ 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:

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

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.:

$$p(\underline{a}_i|a_{i.}) \propto \prod_{k=1}^{T-1} \left\{ e^{\theta_{ik} a_{ik}} (1 + e^{\theta_{ik}})^{-\sum_{\ell=k}^{T} a_{i\ell}} \right\} \tag{2}$$

Completing the hierarchical model, specify:

$$\theta_{ik} \sim N(\mu_k, \gamma_k^2);\ \text{ind.}, i, k = 1 \ldots T-1. \tag{3}$$

[1]The Bayesian convention of using the "conditioning line" to show when model parameters are considered fixed is followed here.

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

To avoid problems with using improper priors that result in improper posteriors, the approximate method for obtaining overdispersed priors by Natarajan and Kass (2000) will be used. Specifically, a uniform improper prior will be used for μ_k and, given known constants ν_k, an approximate uniform shrinkage prior will be used for γ_k^2:

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

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

Natarajan and Kass (2000) show that the class of priors, in which this prior belongs, are all proper for any ν_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 ν_k. It is recognized that basing any prior on even part of the data in the likelihood precludes the direct use of Bayes theorem for posterior inference. However, Natarajan and Kass state that treating ν_k as if is 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 can be defined by the set of binary indicator variables, $\underline{x}_{ih} = (x_{Fih}, x_{Uih1} \ \ldots, \ x_{Uihu_{\delta_{ih}}})$ indicating presence or absence of poverty by either a 1 or a 0 for the family and each $u_{\delta_{ih}}$ unrelated housing unit member.

The distribution of family poverty is defined as independent Bernoulli:

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

Conditional on the family's poverty status, the poverty status of the unrelated individuals within the housing unit are also independent Bernoulli:

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

Completing the model between tracts, define the logits:

$$\begin{aligned}
\ln\left(\frac{p_{0i}}{1-p_{0i}}\right) &= \underline{z}_i'\underline{\beta} + t_i, \\
\ln\left(\frac{p_{Pi}}{1-p_{Pi}}\right) &= \underline{z}_i'\underline{\beta} + t_i + \nu_P, \text{ and} \\
\ln\left(\frac{p_{Ni}}{1-p_{Ni}}\right) &= \underline{z}_i'\underline{\beta} + t_i + \nu_N,
\end{aligned}$$

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,

$$\begin{aligned}
\tilde{p}_{0i} &= e^{\underline{z}_i'\tilde{\underline{\beta}}}/(1+e^{\underline{z}_i'\tilde{\underline{\beta}}}), \\
\tilde{p}_{Pi} &= e^{\underline{z}_i'\tilde{\underline{\beta}}+\tilde{\nu}_P}/(1+e^{\underline{z}_i'\tilde{\underline{\beta}}+\tilde{\nu}_P}) \text{ and} \\
\tilde{p}_{Ni} &= e^{\underline{z}_i'\tilde{\underline{\beta}}+\tilde{\nu}_N}/(1+e^{\underline{z}_i'\tilde{\underline{\beta}}+\tilde{\nu}_N})
\end{aligned}$$

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:

$$\begin{aligned}
m_{0i} &= \sum_{h\in s} x_{Fih}, \\
n_{Pi} &= \sum_{h\in s} x_{Fih} u_{\delta_{ih}}, \\
m_{Pi} &= \sum_{h\in s} x_{Fih} \sum_{j=1}^{u_{\delta_{ih}}} x_{Uihj}, \\
n_{Ni} &= \sum_{h\in s} (1 - x_{Fih}) u_{\delta_{ih}} \\
m_{Ni} &= \sum_{h\in s} (1 - x_{Fih}) \sum_{j=1}^{u_{\delta_{ih}}} x_{Uihj}.
\end{aligned}$$

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.:

$$p_{0i}^{m_{0i}}(1-p_{0i})^{n_{0i}-m_{0i}} p_{Pi}^{m_{Pi}}(1-p_{Pi})^{n_{Pi}-m_{Pi}} p_{Ni}^{m_{Ni}}(1-p_{Ni})^{n_{Ni}-m_{Ni}} \qquad (4)$$

The complete likelihood is the product of the two likelihoods in (2) and (4), since the distribution of person level outcomes was specified conditionally on the housing unit characteristic outcomes.

3 An Approximate Model

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:

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

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 (1).

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

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

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:

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

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.

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

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

$$\begin{aligned} b_{0i} &= \frac{1}{2}\sqrt{p(\underline{z}_i'\tilde{\underline{\beta}})(1-p(\underline{z}_i'\tilde{\underline{\beta}}))} \\ c_{0i} &= \sin^{-1}\sqrt{p(\underline{z}_i'\tilde{\underline{\beta}})} - \underline{z}_i'\tilde{\underline{\beta}}b_{0i} \\ b_{Pi} &= \frac{1}{2}\sqrt{p(\underline{z}_i'\tilde{\underline{\beta}}+\tilde{\nu}_P)(1-p(\underline{z}_i'\tilde{\underline{\beta}}+\tilde{\nu}_P))} \\ c_{Pi} &= \sin^{-1}\sqrt{p(\underline{z}_i'\tilde{\underline{\beta}}+\tilde{\nu}_P)} - (\underline{z}_i'\tilde{\underline{\beta}}+\tilde{\nu}_P)b_{Pi} \\ b_{Ni} &= \frac{1}{2}\sqrt{p(\underline{z}_i'\tilde{\underline{\beta}}+\tilde{\nu}_N)(1-p(\underline{z}_i'\tilde{\underline{\beta}}+\tilde{\nu}_N))} \\ c_{Ni} &= sin^{-1}\sqrt{p(\underline{z}_i'\tilde{\underline{\beta}}+\tilde{\nu}_N)} - (\underline{z}_i'\tilde{\underline{\beta}}+\tilde{\nu}_N)b_{Pi} \end{aligned}$$

The distributions of the remaining parameters of the model are specified identically as in the exact model given in section 2.

4 Finite Population Parameters of Interest

For tract, i, estimates of the per-capita poverty rate, 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 above 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 to be:

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

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

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

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 to be:

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

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 a 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 computing estimates from the approximate model is much less that from using the exact model. For both models inference was obtained after discarding the first 500,000 iterations and using the next 700,000 for estimation. One long chain was run. Posterior means were estimated by averaging all iterations together after the burn-in period. To reduce the effects of correlated data, posterior variances were made 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 together to make the final estimate.

5.1 Assessment of Model Using 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 of 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 jack-knifed variance estimate of the arcsine square root transformation of the tract sample poverty rate can be empirically estimated. The jack-knife used is based on housing units to account for within housing unit correlation. Specifically, the variance of the arcsine square 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 jack-knife cells. The arcsine square 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 jack-knifed standard deviation with the actual jack-knifed standard deviations from the original sample. As can be seen, the model provides good coverage of the observed jack-knifed estimates, indicating that the model can replicate the within tract variances well. This figure also shows the degree of error of the jack-knifed estimate of variance, as measured with the model. The tracts are ordered by sample sizes and the increase in error as the within tract sample size decreases is apparent. (Note: since sample sizes tended to cluster, plots directly by sample size are difficult to read. Figure 11 provides the link between actual sample size and sample size order) 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. 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 (CVs) 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 increase, 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; illustrating 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 under-estimate the accuracy of the poverty rate but overestimate the accuracy for both average persons per housing unit and occupancy rate. It should not be a surprise 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 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 combination of errors it is possible that the approximate model actually improves on deficiencies in the exact model. Although models 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 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; 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:

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

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

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 %

Table 1: L-Criteria for Comparing Models

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

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

For estimating the within tract sampled standard error, $\widehat{SD}_i$, as measured by the square root of the tract-level sample size times the jack-knifed tract variance, it can be seen that the approximate model does do 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 approximation represents a reasonable approximation to the exact model but, as demonstrated, 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 12 lists the posterior CVs (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 Summary

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 make estimates of both level and accuracy for small areas, without making restrictive assumptions about the within small area variance, was demonstrated. The approximate model, while still requiring MCMC methods for estimation is much simpler to work 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. 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 lends itself 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.

References

[1] Besag, Julian, Peter Green, David Higdon and Kerrie Mengersen (1995). "Bayesian Computation and Stochastic Systems", *Statistical Science*, 10, 3-41.

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

[3] Chand, Nanak and Donald Malec (2001) "Small Area Estimates from the American Community Survey Using a Housing Unit Model".

Proceedings of the Federal Committee on Statistical Methodology Research Conference, Arlington VA. November 14, 2001.

[4] Chib, Siddhartha and Edward Greenberg (1995). "Understanding the Metropolis-Hastings Algorithm", *The American Statistician*, 49, 327–335.

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

[6] Isaki, Cary T., Elizabeth T. Huang and Julie H. Tsay (1991) "Smoothing Adjustment Factors from the 1990 Post Enumeration Survey". *Proceedings of the Social Statistics section, American Statistical Associations*. Pp. 338-343.

[7] Laud, Purushottam W. and Joseph G. Ibrahim (1995), "Predictive Model Selection", *Journal of the Royal Statistical Society, Series B.* 57, 1, 247-262.

[8] Natarajan, Ranjini and Robert E. Kass (2000), "Reference Bayesian Methods for Generalized Linear Mixed Models", *Journal of the American Statistical Association* 95, 450, 227-237.

[9] Otto, Mark C. and William R. Bell (1995) "Sampling Error Modeling of Poverty and Income Statistics for States". *Proceedings of the Government Statistics Section, American Statistical Association*. Pp. 160-165.

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

[11] Schaible, Wesley, A. . (1979). "A Composite Estimator for Small Area Statistics". in Synthetic Estimates for Small Areas, pp.36-53. National Institute on Drug Abuse Research Monograph Series 24. DHEW Publication No. (ADM)79-801. Chapman & Hall.

Figure 1: .

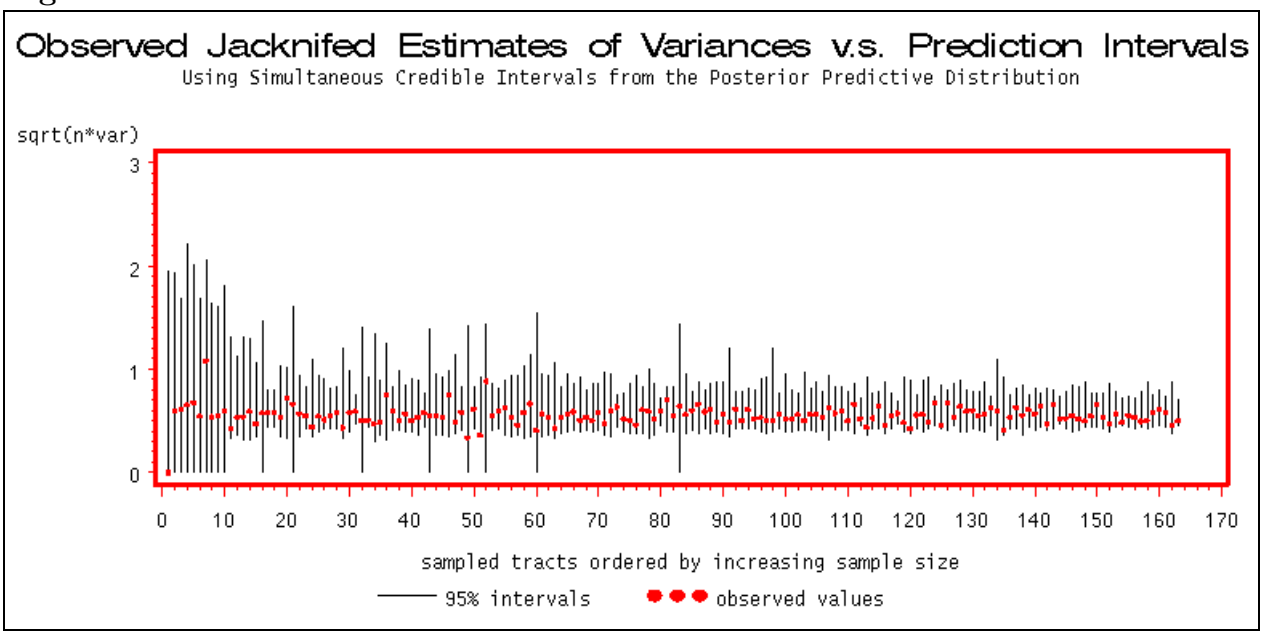


Figure 2: .

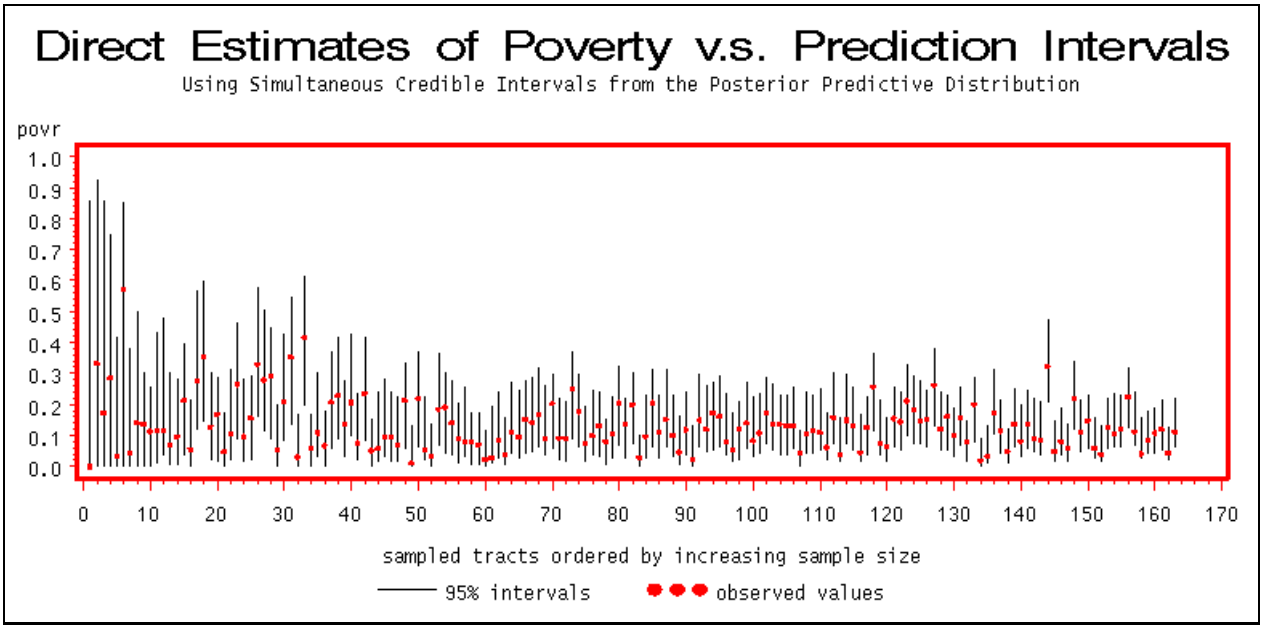


Figure 3: .

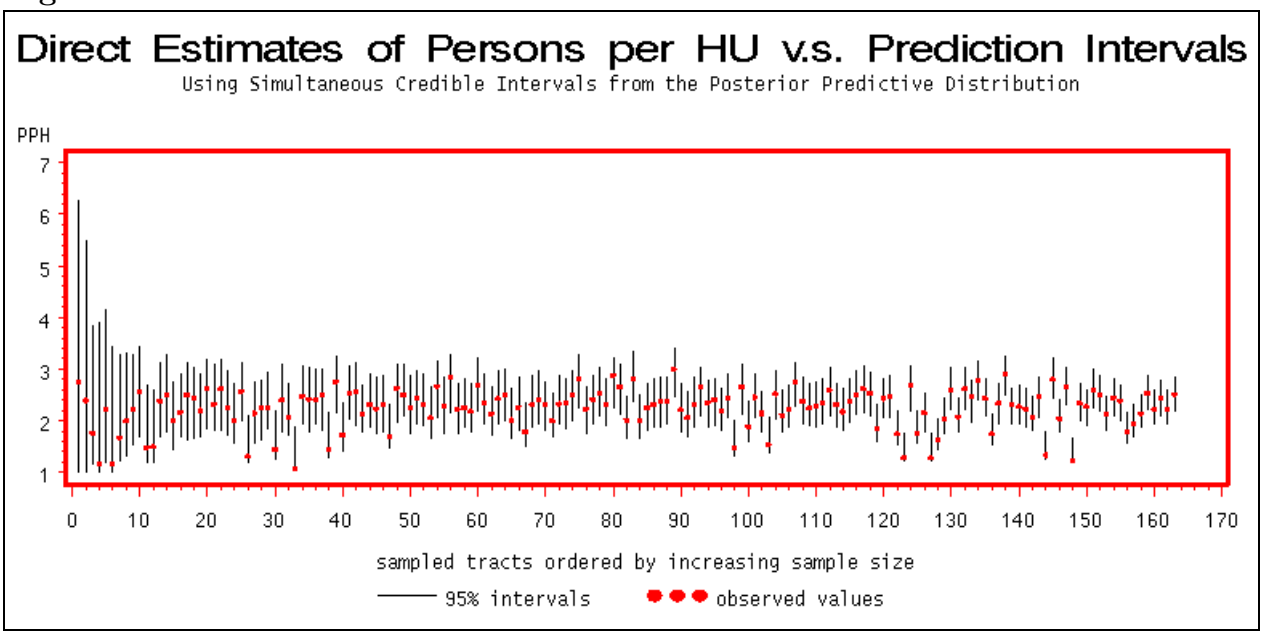


Figure 4: .

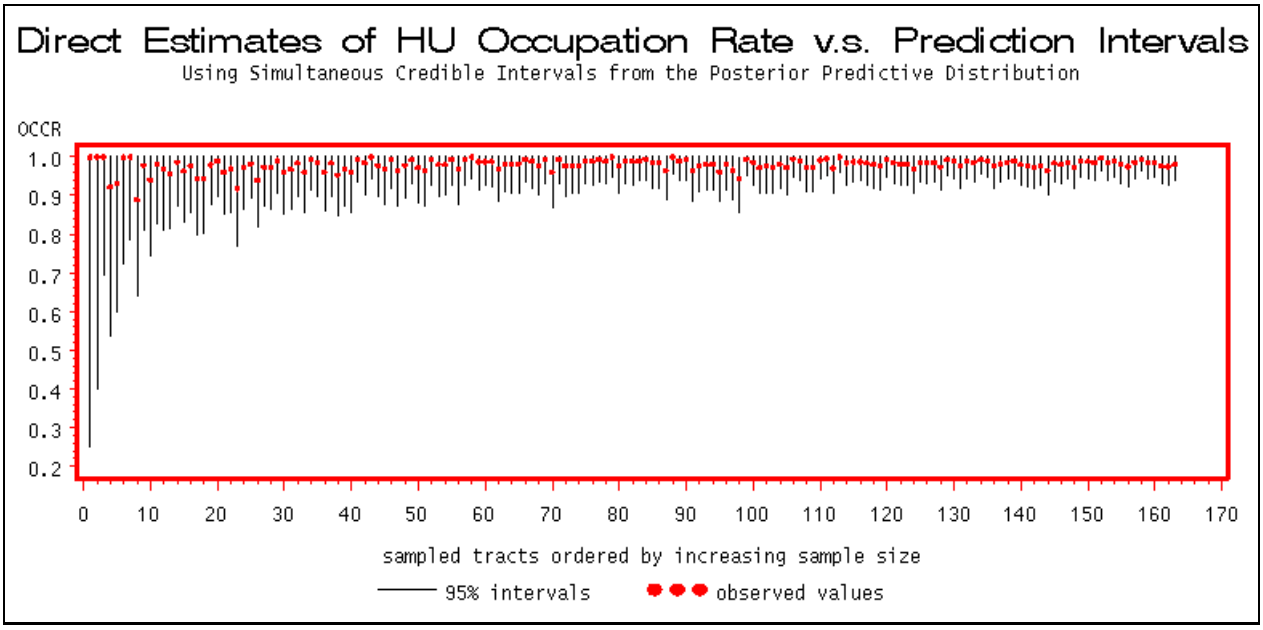


Figure 5: .

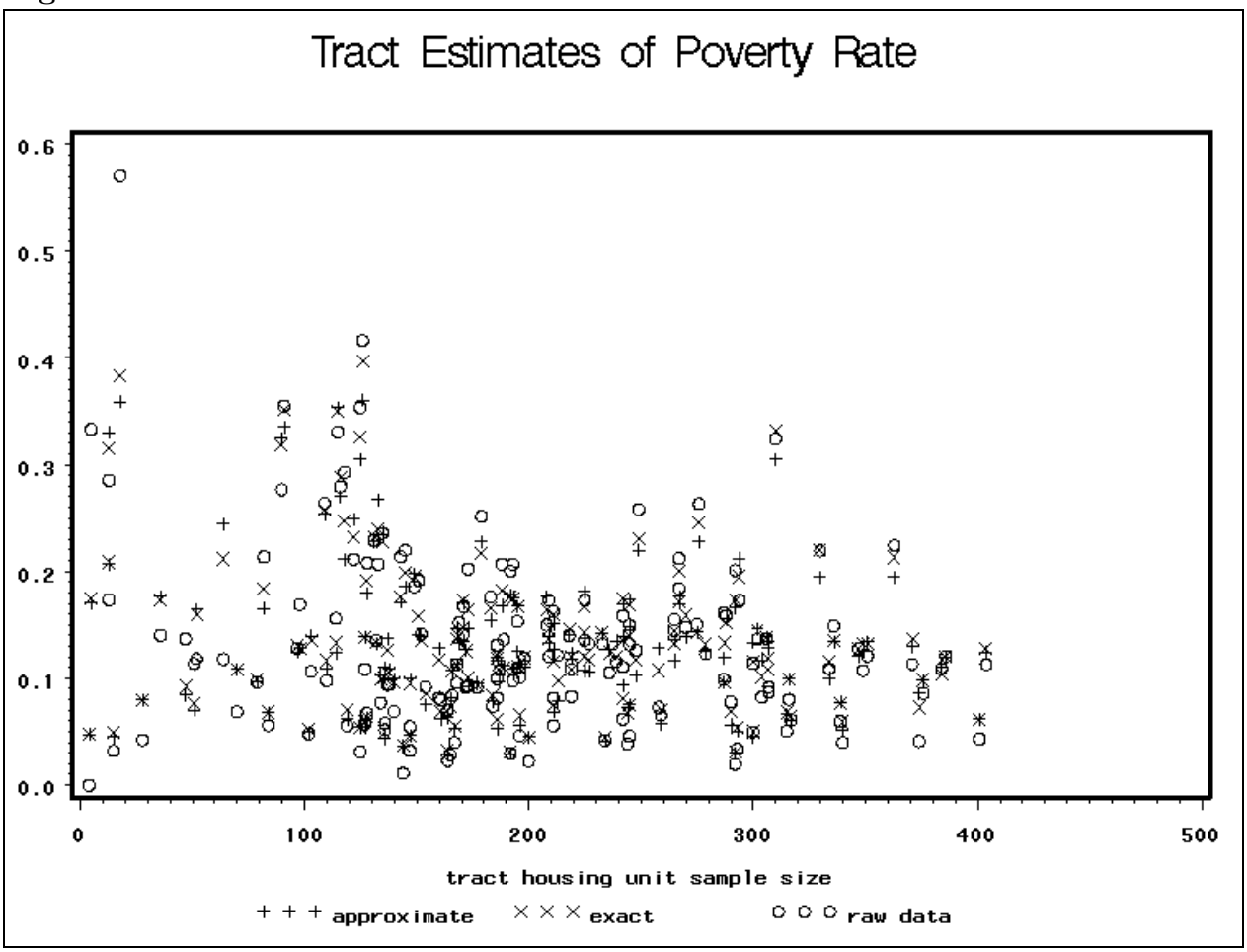

Tract Estimates of Poverty Rate
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
100
200
300
400
500
tract housing unit sample size
+ + + approximate
× × × exact
o o o raw data

Figure 6: .

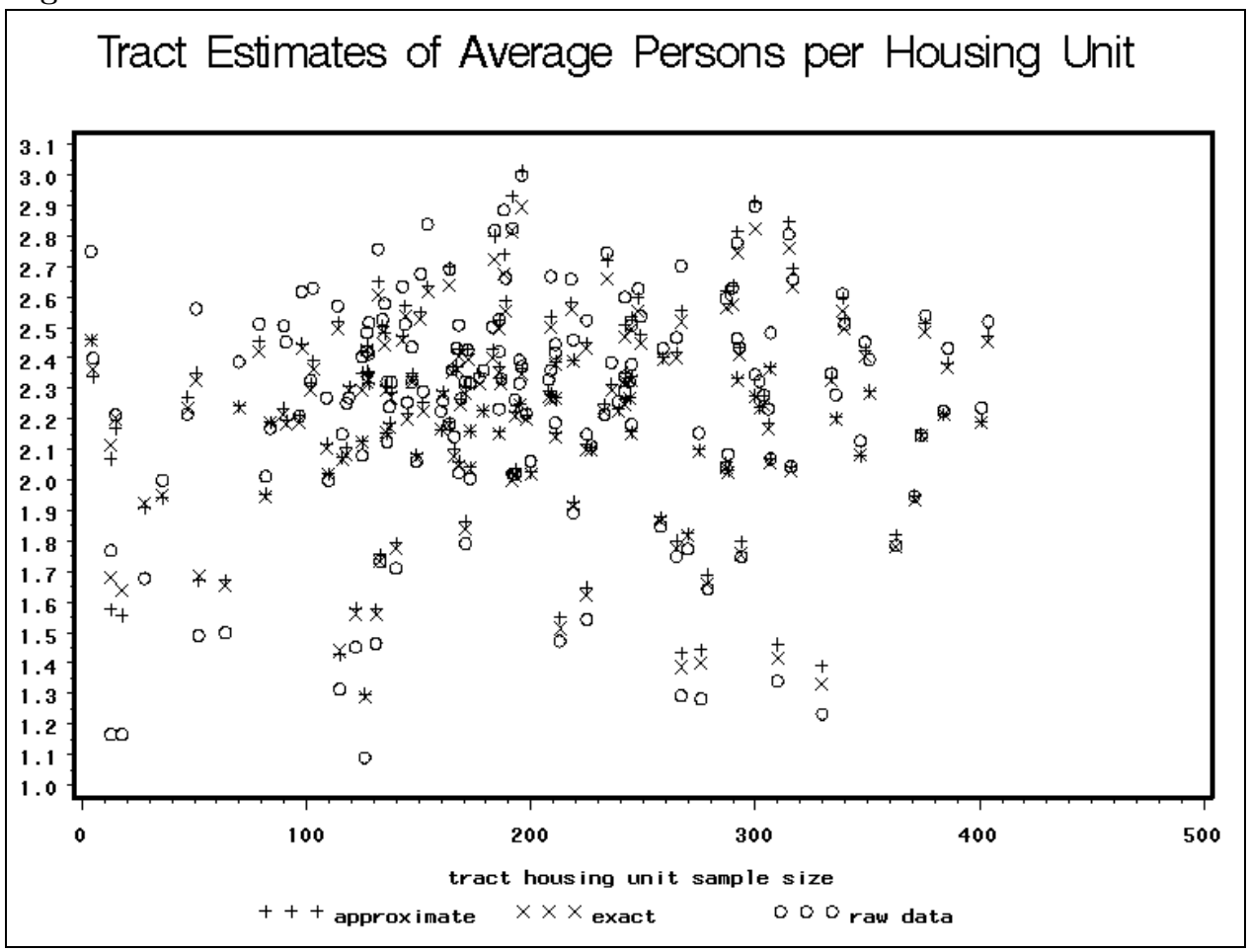

Tract Estimates of Average Persons per Housing Unit
3.1
3.0
2.9
2.8
2.7
2.6
2.5
2.4
2.3
2.2
2.1
2.0
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1.0
0
100
200
300
400
500
tract housing unit sample size
+ + + approximate
× × × exact
o o o raw data

Figure 7: .

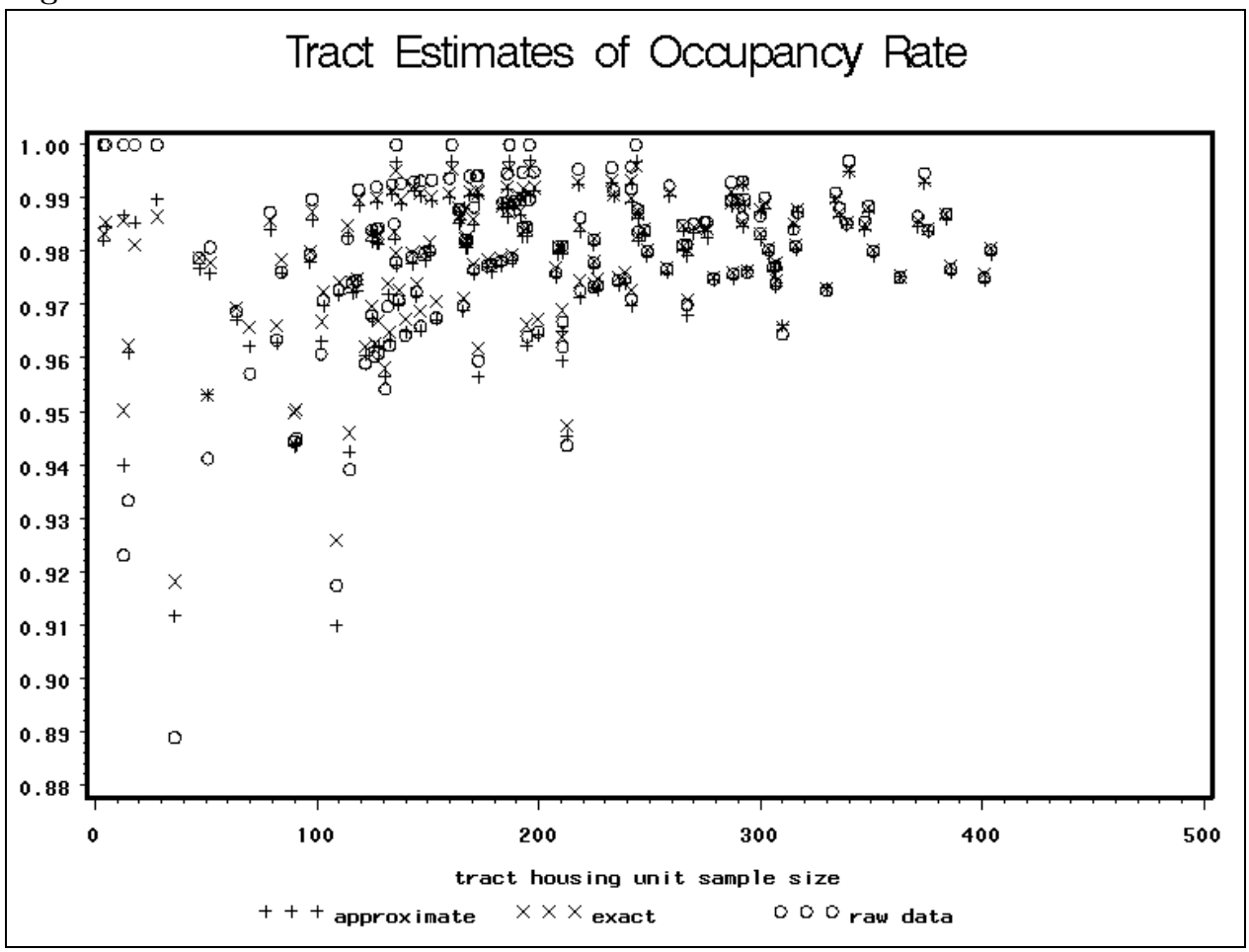

Tract Estimates of Occupancy Rate
1.00
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0.90
0.89
0.88
0
100
200
300
400
500
tract housing unit sample size
+ + + approximate
× × × exact
o o o raw data

Figure 8: .

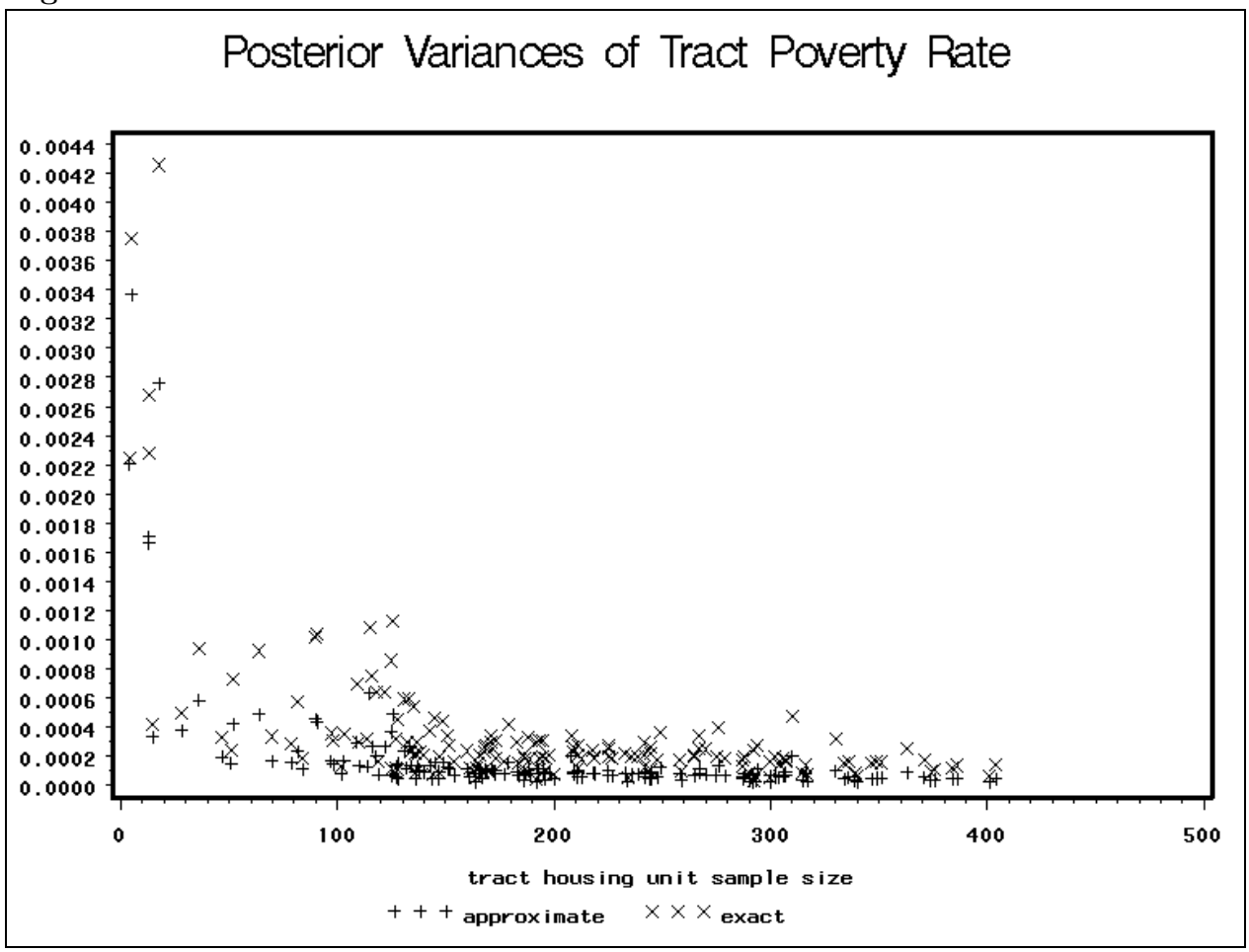

Posterior Variances of Tract Poverty Rate
0.0044
0.0042
0.0040
0.0038
0.0036
0.0034
0.0032
0.0030
0.0028
0.0026
0.0024
0.0022
0.0020
0.0018
0.0016
0.0014
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
0
100
200
300
400
500
tract housing unit sample size
approximate
exact

Figure 9: .

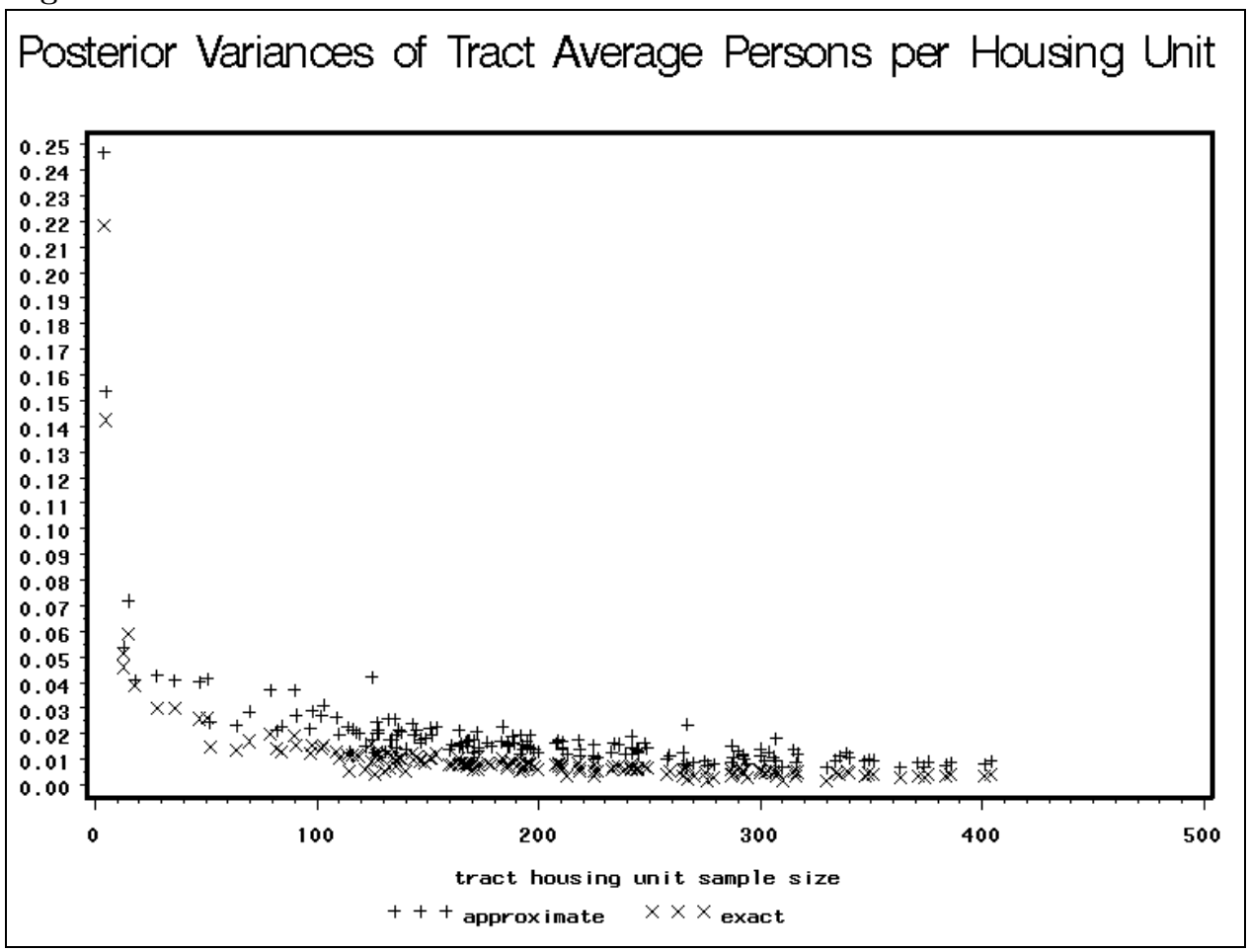

Posterior Variances of Tract Average Persons per Housing Unit
0.25
0.24
0.23
0.22
0.21
0.20
0.19
0.18
0.17
0.16
0.15
0.14
0.13
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
100
200
300
400
500
tract housing unit sample size
approximate
exact

Figure 10: .

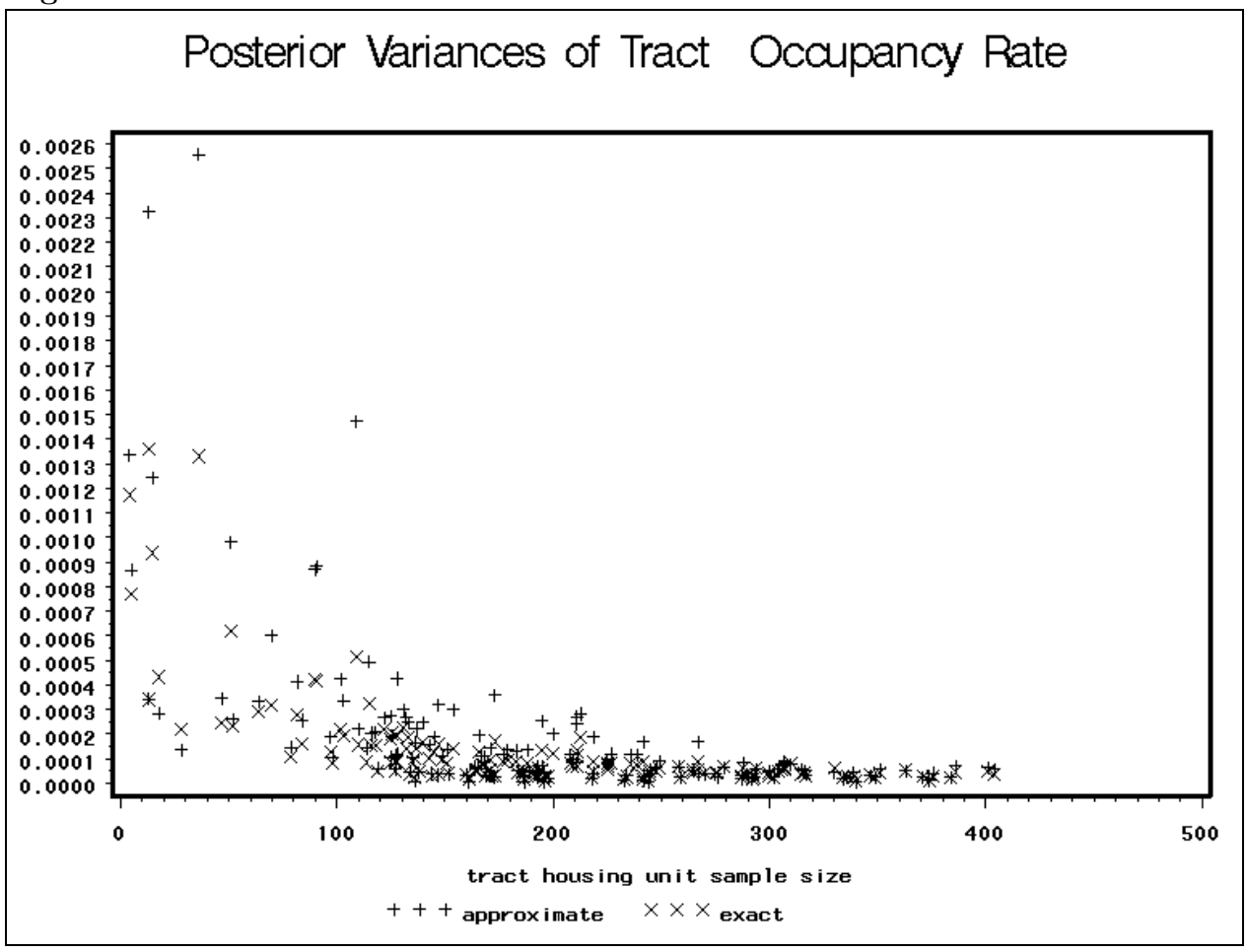

Posterior Variances of Tract Occupancy Rate
0.0026
0.0025
0.0024
0.0023
0.0022
0.0021
0.0020
0.0019
0.0018
0.0017
0.0016
0.0015
0.0014
0.0013
0.0012
0.0011
0.0010
0.0009
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0
100
200
300
400
500
tract housing unit sample size
+ + + approximate
× × × exact

Figure 11: .

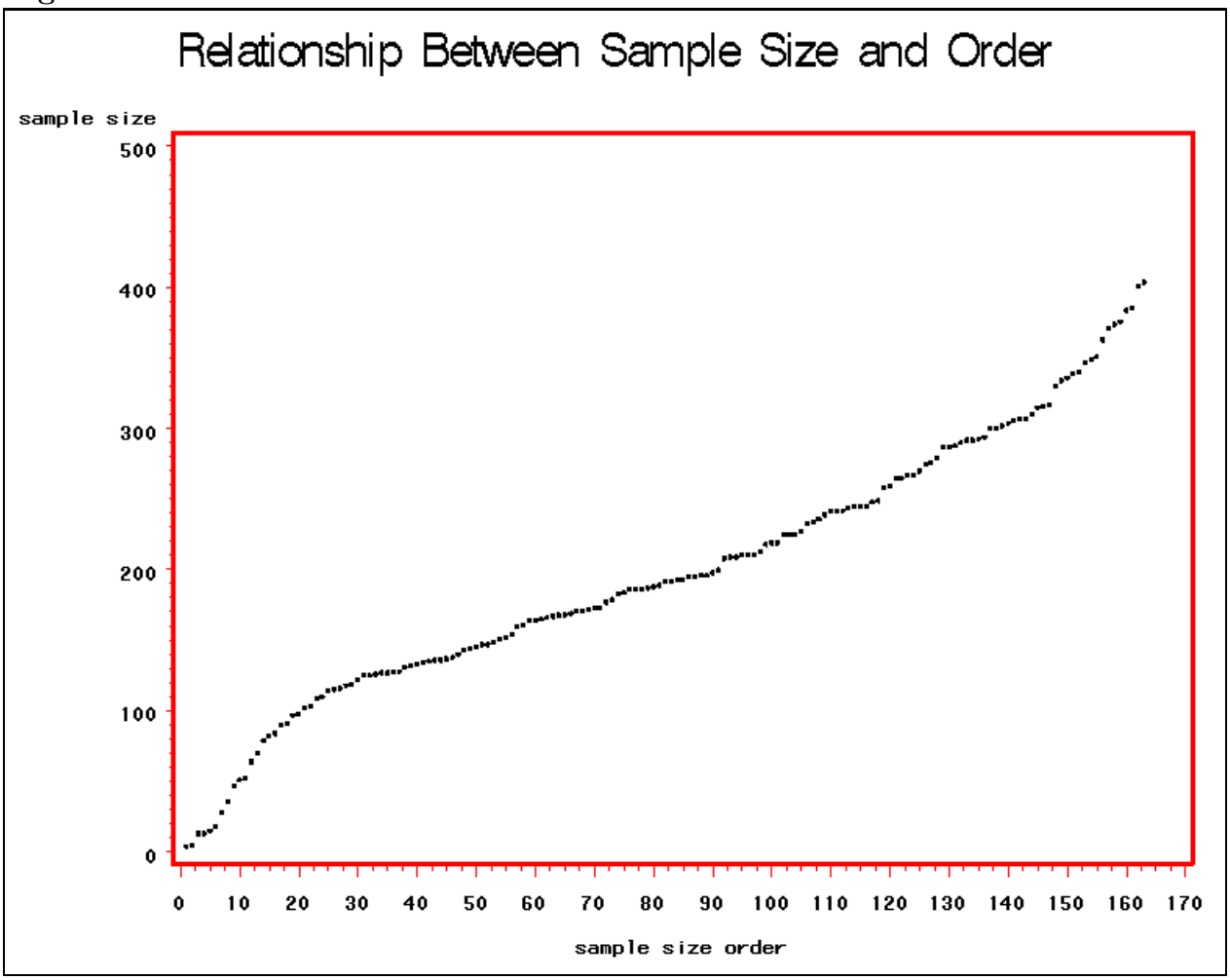
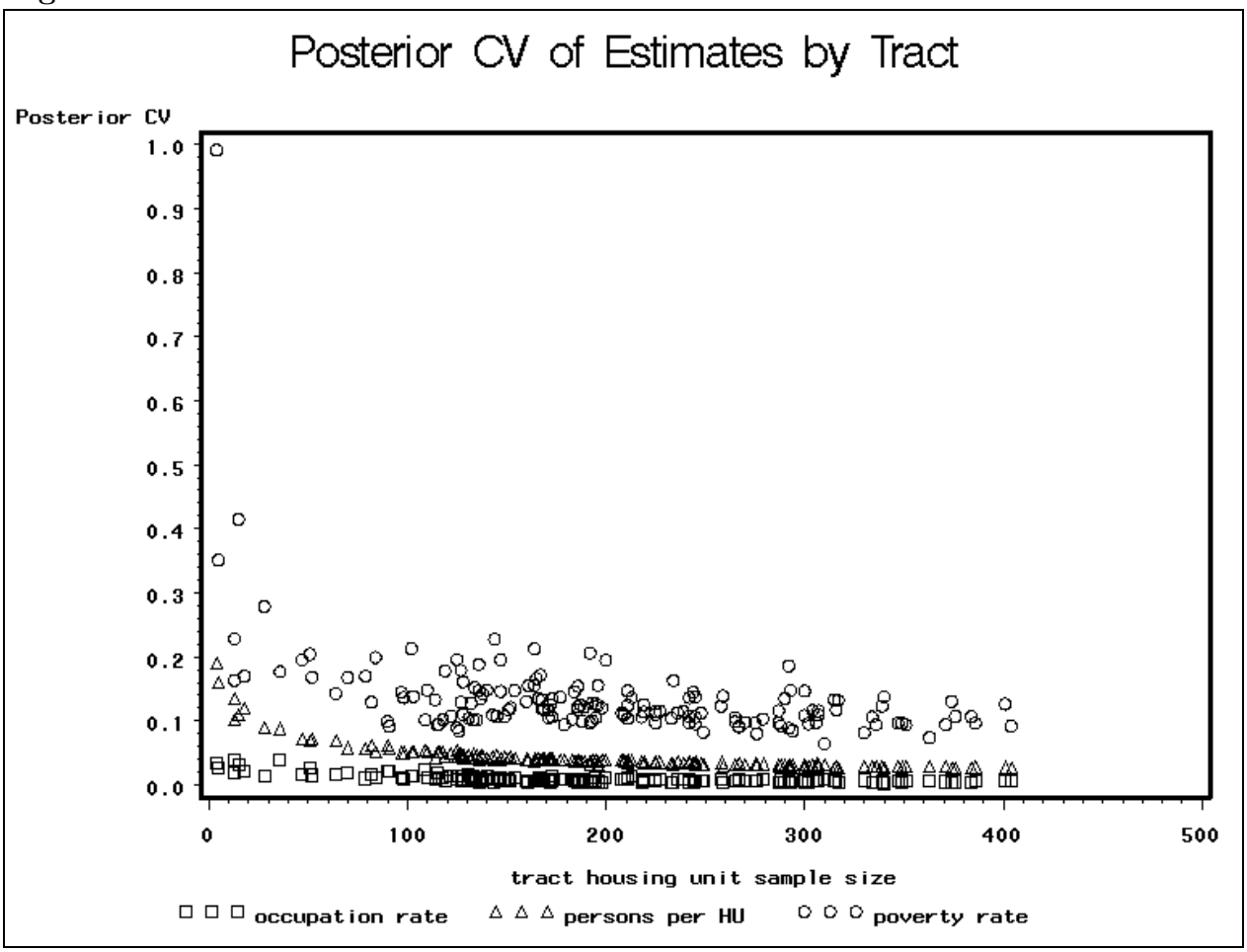

Relationship Between Sample Size and Order
sample size
500
400
300
200
100
0
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
sample size order

Figure 12: .


Posterior CV of Estimates by Tract
Posterior CV
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
100
200
300
400
500
tract housing unit sample size
occupation rate
persons per HU
poverty rate