RESEARCH REPORT SERIES
(Statistics #2008-07)

Exploration of the Use of Empirical Bayes Procedures for Estimating Changes in Occupancy Rate and Persons per Household

Lynn Weidman
Robert Creecy
Donald Malec
Julie Tsay

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

Report Issued: September 30, 2008

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Exploration of the Use of Empirical Bayes Procedures for Estimating Changes in Occupancy Rate and Persons per Household

Lynn Weidman, Robert Creecy, Donald Malec, Julie Tsay
Statistical Research Division, U.S. Census Bureau

Abstract The Census Bureau is carrying out research with the intent of improving the housing unit-based method for estimating population totals. One approach is to use the Decennial Census as the baseline for measures and update them annually using American Community Survey (ACS) data to estimate their change since the census year. This study looks at the possibility of using an Empirical Bayes approach to produce county estimates of change with smaller variances than direct ACS estimates. Results from national and state models are compared. Data from the 1990 and 2000 long form samples are used to represent ACS and data from the 1990 and 2000 short form are used as independent variables in the models. In the actual application, sources for the latter would be the Master Address file and Administrative Records.

1. Introduction

Task 3b of the Housing Unit Based Estimates Research Project Draft Research Agenda (HUBERT) of the Population Division (POP) asks the overall question, "Should statistically significant PPH and vacancy rate changes from the ACS be used" to estimate change over time in persons per household (PPH) and vacancy rate (VR)? In the corresponding Team Requirements Document, the team goals/objectives are:

1. Determine legitimacy of using estimates normally considered to be insignificant.
2. Determine if optimal frequentist significance level exists.
3. Determine structure(s) of Bayesian prior for use in estimating change. Produce estimates. Is this approach desirable?
4. Develop empirical Bayesian framework for estimating change. Produce estimates. Is this approach desirable?
5. If Bayesian or empirical Bayesian approaches are deemed desirable, compare to frequentist estimates.

SRD staff met with Charles Coleman of POP in May 2007 to learn more about this task and what type of research they could perform to help him meet the team goals. After discussing the problem, SRD staff suggested that they look at the empirical Bayes (EB) method for determining estimates of the two change variables from ACS data. This method does not require the use of prior distributions as does the full Bayes method. It only requires the availability of variables correlated with the characteristic being estimated. These variables are regressed on that characteristic and for each observation a regression estimate of the characteristic is calculated in addition to the direct estimate. The EB estimator is then a weighted average of the direct estimate and the linear regression estimate of the characteristic. The weights on the two estimates depend on the sampling variance of the direct estimate and how well the model fits the data. It was suggested that a full Bayesian approach not be carried out until we know the success of the EB approach, which will convey the benefits of using a statistical model. This approach does not require determining whether there has been a statistically significant change in the estimate of PPH or VR before using a current year estimate that differs from the previous year's, but determines the optimal estimator within a class of estimators.

The housing unit-based method estimates the number of housing units (HUs), percent of occupied housing units (%OCC), and persons per occupied housing unit (PPH), and uses their product (HUs*%OCC*PPH) to estimate the number of persons living in HUs. (%OCC will be used in the remainder of this report rather than its complement VR.) The American Community Survey (ACS) is a very large new household survey that provides updated information annually on %OCC, PPH, and related measures, as well as on a large number of demographic, economic, and social characteristics. All of these characteristics were formerly available only once a decade following the decennial census. The approach of the project we report on is to use the ACS as the source of direct estimates of change in %OCC and PPH between the current year and the previous census year, then apply Empirical Bayes (EB) methods to find estimates of change in %OCC and PPH that have smaller variances than these direct estimates. These change estimates are then combined with the previous census values to get the current year estimates. In addition, we want to look at how close the EB estimates are to the full Bayes estimates.

The sample size and continuous nature of the ACS make it an obvious source to use as the basis for producing current year estimates with relatively small sampling variability. The variables correlated with %OCC and PPH that are required for the EB approach would in practice be derived from administrative records and the Census Bureau's Master Address File (MAF). This paper reports on the initial results of this research project using decennial census data and introduces some of the practical and statistical issues that must be addressed before this approach can be applied.

We give a short presentation of the general approach and some estimation issues in the following section, summarize the EB method in section 3, describe the data used in section 4, give selected regression models in section 5, summarize results of the EB estimation in section 6, and present a brief conclusion in section 7. All tables and figures are given in Appendix 1.

2. General Approach

Consider the situation where the Census Bureau is producing housing unit-based estimates of county populations using %OCC and PPH in the years following the 2010 Census (C2010). This project looks at a method of estimating the changes in these variables between the 2010 ACS and a future year ACS. For each of the variables, the change estimate would be combined with the Census 2010 value to obtain an estimate of the current year value. Why don't we just use the current year value of ACS as our estimate? There are the following differences between the data collection methodologies used in the census and the ACS that result in their not estimating exactly the same %OCC and PPH parameters.

(1) The reference date for each ACS interview is the day on which it is begun, which can be any day throughout the year. The census determines its reference date in a similar manner, but it is always close to April 1.

(2) Because the ACS represents an average of characteristics that can change over time, the persons included as residents of a HU are generally only those who are staying there for longer than two months. We refer to this as the current residence rule. This differs from the census which includes all persons who are considered to stay at the HU 'most of the time.' (All sample persons currently staying at a non-HU residence, referred to as group quarters (GQ), are interviewed as in the census.)

(3) An ACS interview of the persons in a HU can take place during a three-month period. A HU that is initially vacant during this time can become occupied and an interview completed later, so the HU is measured as occupied although it is only occupied during part of the period. A HU can also be initially occupied and become vacant later, but if the interview was completed during the initial period then it also is measured as occupied although it was not occupied during the entire period. There are additional scenarios, but overall they result in the number of occupied houses being over-counted compared to the census.

As a result of these differences in data collection methods, changes between estimates from ACS in two different years will need to be scaled to correctly represent changes based on the Census 2010 residence rule.

There is an additional issue with using ACS county estimates that needs to be addressed. Single year estimates and their variances are currently only produced for counties (and other geographical entities) with a population larger than 65,000, due to larger variances for smaller populations. But it would be easy to produce them for all smaller counties because the necessary weighting has been done. And the larger sampling variances for %OCC and PPH in these smaller counties is not a concern for the EB approach, as they are just used in determination of the relative contributions of the direct and modeled estimates to the EB estimate.

3. Empirical Bayes Method

The EB method requires a direct estimate of the characteristic of interest and variables correlated with that characteristic to be available. The direct estimate of the characteristic is regressed on these correlated variables and for each observation a regression estimate of the characteristic is calculated in addition to the direct estimate. The EB estimator is then a weighted average of the direct estimate and the linear regression estimate of the characteristic. The weights on the two estimates depend on the sampling variance of the direct estimates and how well the model fits the data.

For purposes of illustration we will use change in %OCC as the variable being estimated.
Let

$$\theta_i = \text{(current year \%OCC under current residence rule for county i)} - \text{(2010 \%OCC under current residence rule for county i)}, \quad i=1,2,\ldots,k, \qquad (1)$$

and

$$Y_i = \theta_i + \eta_i \qquad (2)$$

be a direct estimate of θ_i from a survey, where the η_i's are independent sampling errors with $E(\eta_i|\theta_i) = 0$ and $V(\eta_i|\theta_i) = V_i$, so the Y_i's are design unbiased for θ_i (their expected values over all possible samples are equal to the quantities they are attempting to estimate).

Now suppose that we have a set of r model variables available for each county i, represented by the vector $\underline{z}_i' = (z_{i1}, z_{i2}, \ldots z_{ir})$ such that

$$\theta_i = \underline{z}_i'\underline{\beta} + \varepsilon_i , \tag{3}$$

is a model that represents %OCC change, where $\underline{\beta}$ is a vector of unknown coefficients, $E(\varepsilon_i) = 0$, and $V(\varepsilon_i) = A_i$ with the ε_i's independent.

Then, substituting (3) into (2), we can write a model for our direct estimate as

$$Y_i = \theta_i + \eta_i = \underline{z}_i'\underline{\beta} + \varepsilon_i + \eta_i . \tag{4}$$

This is called a mixed model for Y_i because it includes the fixed parameters $\underline{\beta}$ and the random term ε_i with $V(\varepsilon_i) = A_i$. (This use of a single variable ε_i is the simplest form for the random part of a mixed model. The random part could be a vector, a regression, or another function of random variables.) If all the $V_i + A_i$ are known, then we can compute a regression estimate $\underline{b}$ of the coefficient vector $\underline{\beta}$ via generalized least squares as

$$\underline{b} = (Z'DZ)^{-1}Z'D\underline{Y} ,$$

with Z the k x r matrix with rows z_i', $D^{-1} = \text{Diagonal}(V_1+A_1, \ldots, V_k+A_k)$, and $\underline{Y} = (Y_1, \ldots, Y_k)$.

If we make the assumption that both ε_i and η_i have normal distributions, then we have

$$Y_i | \theta_i \sim N(\theta_i, V_i) \text{ and } \theta_i | \underline{\beta}, \eta_i \sim N(\underline{z}_i'\underline{\beta}, A_i) . \tag{5}$$

(We can make other distributional assumptions but normality makes the problem simpler, so we will work with it for now.)

We then look at estimators of the θ_i, which are weighted averages of the Y_i and the $\underline{z}_i'\underline{b}$.

$$\hat{\theta}_i = C_i Y_i + (1 - C_i)\, \underline{z}_i'\underline{b} . \tag{6}$$

Then the current year %OCC is estimated as g(C2010 %OCC, $\hat{\theta}_i$), where $g(\cdot)$ is a function that adjusts for definitional and operational differences between the ACS and the decennial census. (If there were no such differences, then $g(\cdot)$ = (C2010 %OCC) + $\hat{\theta}_i$.) The best linear unbiased predictor (BLUP) in (6) when the A_i's and V_i's are known requires $C_i = A_i / (V_i+A_i)$. Since the A_i's and V_i's are not known, we substitute estimates $\hat{A}_i$ and $\hat{V}_i$ for them to get $\hat{C}_i$ and the empirical BLUP (EBLUP). Under our assumption of normality of the variance components, the

EBLUP estimators are the same as the EB estimators, where the posterior distributions of the $\hat{\theta}_i$ are normal with

$$\text{mean } \hat{C}_i Y_i + (1 - \hat{C}_i)\underline{z}_i'\underline{b}\text{ , and variance } \hat{C}_i V_i . \qquad (7)$$

$\hat{C}_i V_i$ is actually an underestimate of the variance because it does not take account of the variability in $\hat{A}_i$ and $\hat{V}_i$. It may be desirable to find a better estimator of the true posterior variance than $\hat{C}_i V_i$ and there are approaches in the literature for doing so (e.g., Rao, 2003). (See Morris, 1983 for additional details about the basic EB set-up.)

4. Data Used in the Study

Our task was to carry out an exploration of the feasibility of using the EB procedure with data similar to what would be available in actual implementation. The 2010 ACS data are represented by the C90 longform, and the 2010 shortform data are represented by the C90 shortform (100% data). Future year ACS data are represented by the C2K longform, while the MAF and administrative record variables for the future year are represented by the C2K shortform.

Model Variables

In the actual application of this methodology, the sources of the EB model variables should have very small sampling variances. If we tried to use variables from the ACS, they would have sampling variances of a magnitude, especially for the less populous geographies, that would need to be incorporated into the estimation procedure. We would prefer to select variables from the shortform of C2010, the MAF, and administrative records. Table 1 in the Appendix gives some suggested main effects variables and their sources, as well as the sources used in this study: the C90 shortform for the C2010 short form and the C2K shortform for the future year MAF and administrative records.

Estimation Variables

It is not appropriate to use the census longform estimates of %OCC and PPH as the estimation variables when shortform estimates are used as the model variables. This is because the final stage of longform weighting controls combinations of several characteristics to be equal to their shortform values, which results in (some of) the z_{ij} in the mixed model having higher correlation with the estimation variables than in the actual application. To avoid this, we create estimates of %OCC and PPH by using only the HU sampling rates and a single ‘nonresponse’ adjustment as follows. Both the Census 1990 and 2000 samples are selected from multiple strata. On the long form files there are sample occupied HUs that have data but no final weights. This is because they were determined not to meet the definition of a HU. We treat them as nonresponding units and apply a nonresponse ratio adjustment by stratum for each county.

The variance for the number of occupied HUs in a county for each census year is estimated using the appropriate formula for a stratified sample. It is divided by the square of the number of HUs to get variance(%OCC). (Our assumption is that for each county the total number of HUs in the sampling frame is equal or very close to the sum over strata of the number of units in sample times the inverse of the stratum sampling rate. This sum is used for the number of HUs when calculating %OCC and variance estimates.) The variance formula for Census 1990 is given by

$$\hat{V}(\%\hat{O}CC_{90}) = \left(\frac{1}{N_{90}^{2}}\right) \sum_{h,90} N_{h,90}^{2} \left(\frac{p_{h,90}\left(1-p_{h,90}\right)\left(1-f_{h,90}\right)}{\left(n_{h,90}-1\right)}\right) \quad , \qquad (8)$$

where

h denotes stratum,

N_{90} is the number of HUs in the county,

$N_{h,90}$ is the number of HUs in stratum h,

$n_{h,90}$ is the number of sample HUs in stratum h,

$p_{h,90}$ is the fraction of sample HUs in stratum h that are occupied, and

$f_{h,90}$ is the proportion of HUs sampled in stratum h

In full notation there would be additional subscripts for county and state but for simplicity we consider a given county and state. For Census 2000 the subscripts 90 are replaced by 00. The estimated variance of the difference in %OCC and PPH between Census 1990 and Census 2000 is the sum of their individual variance estimates.

The variance of the estimated number of persons in a county is also estimated based on the appropriate stratified sampling formula, with adjustment for the nonresponding occupied units and other assumptions required. Derivation of the following approximate variance formula is given in Appendix 2.

$$\hat{V}(pers\hat{o}ns_{90}) \approx \Sigma N_{h,90}^{2} \left\{ \frac{p_{h,90}\left(1-p_{h,90}\right)\left(1-f_{h,90}\right)}{\left(n_{h,90}-1\right)} \left[\frac{s_{hor,90}^{2}\left(1-\rho_{ho,90} f_{h,90}\right)}{\rho_{ho,90}\, n_{ho,90}} + \bar{x}_{hor,90}^{2} \right] + p_{ho,90}^{2} \frac{s_{hor,90}^{2}\left(1-\rho_{ho,90} f_{ho,90}\right)}{\rho_{ho,90}\, n_{h,90}} \right\} \qquad (9)$$

where

$n_{ho,90}$ is the number of occupied sample HUs in stratum h,

$\rho_{ho,90}$ is the response rate for occupied HUs in stratum h,

$\bar{x}_{hor,90}$ is the mean number of persons in occupied respondent HUs in stratum h, and

$s_{hor,90}^{2}$ is the estimated variance of the number of persons per occupied HU in stratum h based on the responding HUs (see Appendix 2).

An estimated variance for PPH is obtained using the standard formula for the approximate variance of the ratio of two random variables,

$$\hat{V}(PP\hat{H}_{90}) \approx \frac{\hat{V}(pers\hat{o}ns_{90})}{N_{90}^{2}\,(\%\hat{O}CC_{90})^{2}} + \frac{(pers\hat{o}ns_{90})^{2}}{N_{90}^{4}\,(\%\hat{O}CC_{90})^{4}} N_{90}^{2}\, \hat{V}(\%\,\hat{O}CC_{90}) \quad , \qquad (10)$$

and substituting $\hat{V}(\%\hat{O}CC_{90})$ from (8) and $\hat{V}(\text{per}\hat{s}\text{ons}_{90})$ from (9).

The estimated variance of a difference between estimates from separate census years is the sum of the individual year variances, due to the independence of their samples.

5. Regression Models

The initial step in development of mixed models was to identify variables correlated with change in %OCC and PPH via forward selection stepwise linear regression. Table 2 lists the full set of variables used. Table 3 shows the number of variables selected and the model R^2 by estimation variable for most states and the nation. (DC, DE, HI, and RI are excluded from the state models because they have five or fewer counties and no degrees of freedom for estimating the random county effect in the mixed model. In future work we would pursue how to handle these states, perhaps by grouping them together or combining them individually with neighboring states.) In most cases only a few variables are included in the regression model and R^2 is of a reasonable size. These results suggest that modeling change in %OCC and PPH with independent variables similar to those used in this task should provide some improvement in estimates via the EB procedure.

6. Empirical Bayes Modeling Results

We attempted to fit mixed models with a single random county term for %OCC and PPH for the nation and most states, using the model variables selected in the stepwise regressions. The county sampling variances V_i of the estimation variables were treated as fixed at their estimated values. The variances A_i for the random county effects in a given model were assumed to have the same value A. We used both the SAS procedure MIXED and a custom-written R program using an E-M algorithm (Creecy, 2008) to find maximum likelihood estimates of the parameters in the mixed models. For the national models SAS gave the message that there was not enough memory to estimate them. Even consultation with SAS staff did not lead to a solution for this problem. We were able to fit these models using R. For %OCC, both programs obtained solutions for all states. For PPH, complete convergence was not obtained for 17 states. For some of these states, MIXED was not able to start the solution procedure and for others $\hat{A} = 0$ but the final Hessian was not positive definite. For these states, plus states 22 and 25, R was not able to obtain convergence but $\hat{A}$ was close to zero when it stopped. As a double check on these results, we searched the likelihood surface for a few of the 17 states using a Fortran program and $\hat{A}$ was equal to zero, so we use $\hat{A} = 0$ as the estimate for these 17 states. (Because of the variability in the distributions of the estimates $\hat{A}$, in a given state the true value that is being estimated may not be close to zero. Seven of these states have more than 50 counties, so the variability is probably fairly small and the true values are likely to be near 0. The remaining 10 states have fewer than 40 counties and their true values are more likely to not be near 0.) None of the states for which both programs converged to an estimate without any warning messages found $\hat{A} = 0$.

Table 4 gives the estimated variance $\hat{A}$ of the single random county variance term for each mixed model. Tables 5 and 6 show percent reductions in variance for the EB estimates compared to the direct sample estimates, using the estimated coefficient $\hat{C}_i = \hat{A} \;/\; (\hat{V}_i + \hat{A})$. The reductions are calculated for the counties with the minimum, median, and maximum estimated sampling variances in each state, to give an idea of the range of reductions. These same values are plotted in Figures 1 and 2 to more easily see the relationships. As mentioned previously, state models were not estimated for the four states with the fewest counties and their results from the national model are shown at the bottom of the table.

%OCC variance reduction summary. There is very little reduction for the minimum variance counties using either model, except for one state. This is not surprising since these variances are so small. For the median variance counties most of the national model reductions are less than 20% and about half the states have additional improvemement of at least 10% for the state model. The national model gives reductions of at least 30% for most states and the state models usually show substantial additional reduction – more than 40% for some states.

PPH variance reduction summary. The variance reductions are in general much larger for PPH than for %OCC. Twenty-two states show more than 10% reduction for the minimum variance counties with the national model, and many show an additional 10% or more reduction with the state model. For two states this additional improvement is more than 70%, so the national model is not appropriate for them. Most states show at least 40% reduction for their median variance counties with noticable additional reductions for the state models. There are large reductions from the national model in most states, so the state models do not usually offer substantial additional reduction.

Note two things about these comparisons. First, care should be used when interpreting the amount of improvement of a state model over the national model. Estimates for the state models are based on many fewer degrees of freedom than are those for the national model, so the variances of the variance component estimates are larger. But 30 of the 47 states have more than 50 counties and these variances are are probably suitably small to allow valid comparison of the two values. Secondly, there are a few states for which the national model gives more variance reduction than the state model. Based on the observed estimates, for these states the national model provides more information than do their state models. We might expect that this would happen in states with smaller numbers of counties, where the national model would supply more data points for the modeling and more degrees of freedom for estimation. Upon examination of the number of counties we see that this is not the case and the state with the most counties, Texas, is in this group.

Figures 3 and 4 show, as an example from a single state, how the estimates of change in %OCC and PPH differ across the sample, the state regression model, the EB procedure, and the full Bayes procedure in Mississippi. The counties are ordered from smallest to largest sampling variance of the direct estimate. For each county the EB estimate lies between the sample and model estimate since it is a linear combination of them. The most important thing to note is that the Full Bayes and EB estimates are almost always very close together, so that in these cases the EB estimator obtains most of the benefits of the full Bayes estimator without the additional assumptions it requires about prior distributions. However, there are some counties where the full Bayes estimator does not lie between the sample and model estimates, so the EB estimate is not close to it.

7. Conclusion

The purpose of this paper is to present research into the feasibility of using the EB approach to reduce the variablity of direct estimates of change in %OCC and PPH across years. Overall we see that the EB approach can give noticable reductions in variance from the direct sample estimates, especially as sampling errors get larger, even with the simple models used. For most states, using state-specific information in the models gives additional improvement over using just national information. The size of the variance reductions shown in Tables 5 and 6 (Figures 1 and 2) and the closeness of EB estimates to the full Bayes estimates in Figures 3 and 4 suggest that further research into the application of the EB methodology to estimating %OCC and PPH from the ACS with auxiliary data from the MAF and administrative records would be worthwhile.

Before this methodology can be applied to the situation introduced at the beginning of this paper, there are multiple avenues of investigation that would need to be pursued. Several issues concerning the relationship of estimates between the decennnial census and the ACS were introduced but not pursued, as well as the issue of how to handle states with few counties. In addition, we have not attempted to look at more complex mixed models to find additional reductions in variances of EB estimates.

Of course, the EB approach can be applied with any sample estimator, not just the ones used here. So it may be possible to use it with estimators investigated in other HUBERT projects after determining an appropriate set of correlated auxiliary variables.

References

Cochran, William (1977). *Sampling Techniques*, New York, John Wiley and Sons.

Creecy, Robert (2008). “Computation of Empirical Bayes Estimates Using Single Level Mixed Models,” *Research Report - Statistics #2008-2*, Statistical Research Division, U.S. Census Bureau.

Morris, Carl (1983). Parametric Empirical Bayes Inference: Theory and Applications, *Journal of the American Statistical Association*, 73, pp. 47-55.

Rao, J.N.K. (2003), *Small Area Estimation*, John Wiley and Sons.

Appendix 1 – Tables and Figures

Table 1. Main Effects Variables for Empirical Bayes Regression Models

Variable	Source	Name
from 2010		
%single units	C2010 longform	si9
%multi unit 10+	C2010 longform	mu9
% urban units	C2010 longform	ur9
% Hispanic	C2010 longform	his9
% non-Hispanic white	C2010 longform	nhw9
from change between current year and 2010		
change in % single units	current MAF – C2010 longform	sic
change in % multi unit 10+	current MAF – C2010 longform	muc
change in %urban units	current MAF – C2010 longform	urc
change in % Hispanic	current ARs – C2010 longform	hisc
change in % non-Hispanic whites	current ARs – C2010 longform	nhwc

Table 2. Regression Model Variables

Main Effects	**C90 Interactions**	**C90 by Change Interactions**		**Change Interactions**
si9	ur9*si9	si9*sic	ur9*hisc	urc*sic
mu9	ur9*mu9	si9*hisc	ur9*nhwc	urc*muc
ur9	ur9*his9	si9*nhwc	his9*hisc	urc*hisc
his9	ur9*nhw9	si9*urc	his9*urc	urc*nhwc
nhw9	si9*his9	mu9*muc	his9*sic	sic*hisc
sic	si9*nhw9	mu9*hisc	his9*muc	sic*nhwc
muc	mu9*his9	mu9*nhwc	nhw9*nhwc	sic*hisc
urc	mu9*nhw9	mu9*urc	nhw9*urc	sic*nhwc
hisc		ur9*urc	nhw9*sic	
nhwc		ur9*sic	nhw9*muc	
		ur9*muc		

Table 3. Number of Variables Selected and R^2 for 1st order Interaction Regression Models

State	Number of Counties	%OCC		PPH	
		Number of Variables	R^2	Number of Variables	R^2
Nation	3137	16	0.306	13	0.380
AL	67	4	0.265	5	0.726
AK	24	1	0.348	6	0.762
AZ	15	4	0.934	6	0.994
AR	75	5	0.335	3	0.574
CA	58	5	0.531	5	0.796
CO	63	3	0.439	4	0.508
CN	8	1	0.492	2	0.786
FL	66	2	0.127	6	0.741
GA	159	3	0.245	4	0.441
ID	44	3	0.710	3	0.536
IL	102	8	0.501	2	0.166
IN	92	7	0.753	2	0.109
IA	99	3	0.172	1	0.069
KS	105	5	0.374	4	0.248
KY	120	3	0.152	4	0.377
LA	64	10	0.694	4	0.276
ME	16	5	0.928	7	0.962
MD	24	1	0.311	9	0.896
MA	14	3	0.842	3	0.964
MI	83	3	0.743	4	0.210
MN	87	3	0.579	2	0.331
MS	82	5	0.350	4	0.430
MO	115	10	0.519	2	0.103
MT	56	5	0.347	3	0.180
NE	93	2	0.123	3	0.293
NV	17	1	0.380	3	0.576
NH	10	1	0.591	6	0.999
NJ	21	3	0.762	3	0.741
NM	33	4	0.716	1	0.475
NY	62	12	0.804	7	0.745
NC	100	4	0.206	6	0.498
ND	53	1	0.382	6	0.415
OH	88	2	0.131	4	0.260
OK	77	6	0.409	3	0.324
OR	36	6	0.744	2	0.598
PA	67	5	0.681	4	0.478
SC	46	2	0.499	5	0.698
SD	66	3	0.290	6	0.376
TN	95	3	0.160	5	0.396
TX	254	4	0.130	4	0.232
UT	29	5	0.729	3	0.441
VT	14	4	0.831	5	0.943
VA	135	8	0.441	4	0.254
WA	39	5	0.720	6	0.776
WV	55	6	0.514	4	0.330
WI	72	5	0.756	1	0.116
WY	23	2	0.566	1	0.117

Table 4. Mixed model variance component estimates

State	%OCC	PPH
Nation	0.000992	0.001495
AL	0.000857	0.000341
AK	0.010336	0.000000
AZ	0.000243	0.000000
AR	0.000296	0.000219
CA	0.000309	0.000853
CO	0.002190	0.000963
CN	0.000042	0.000000
FL	0.001629	0.000602
GA	0.000763	0.000659
ID	0.000310	0.000803
IL	0.000123	0.000343
IN	0.000199	0.000236
IA	0.000160	0.000705
KS	0.000423	0.000000
KY	0.000374	0.000618
LA	0.000174	0.000003
ME	0.000029	0.000000
MD	0.000354	0.000000
MA	0.000139	0.000006
MI	0.000618	0.000532
MN	0.000881	0.000958
MS	0.000561	0.000952
MO	0.000169	0.000919
MT	0.000501	0.000000
NE	0.000792	0.000000
NV	0.001043	0.000000
NH	0.000222	0.000000
NJ	0.000076	0.000357
NM	0.000505	0.000000
NY	0.000169	0.000287
NC	0.000446	0.000692
ND	0.000645	0.000000
OH	0.000119	0.000326
OK	0.000346	0.000038
OR	0.000201	0.000246
PA	0.000665	0.000456
SC	0.000332	0.000188
SD	0.001126	0.000000
TN	0.000388	0.000000
TX	0.001359	0.000992
UT	0.000286	0.005332
VT	0.000087	0.000000
VA	0.000249	0.002753
WA	0.000271	0.000000
WV	0.000303	0.000000
WI	0.000408	0.000992
WY	0.000784	0.002561

Table 5. County variance reductions for %OCC with minimum, median, and maximum sampling variances

State	Number of Counties	County Sampling Variances			National Model % Variance Reductions			State Model % Variance Reductions		
		Minimum	Median	Maximum	Minimum	Median	Maximum	Minimum	Median	Maximum
AL	67	0.000004	0.000069	0.000327	0.36	6.49	24.79	0.41	7.44	27.62
AK	24	0.000012	0.000200	0.000681	1.15	16.76	40.71	0.11	1.90	6.18
AZ	15	0.000002	0.000047	0.000343	0.16	4.53	25.69	0.65	16.23	58.56
AR	75	0.000007	0.000120	0.000447	0.70	10.82	31.07	2.31	28.92	60.18
CA	58	0.000000	0.000016	0.000903	0.02	1.58	47.65	0.07	4.91	74.51
CO	63	0.000003	0.000209	0.001073	0.34	17.43	51.96	0.15	8.72	32.88
CN	8	0.000002	0.000011	0.000019	0.18	1.09	1.88	4.11	20.72	31.15
FL	66	0.000003	0.000060	0.000592	0.30	5.69	37.38	0.18	3.55	26.65
GA	159	0.000004	0.000150	0.000839	0.43	13.14	45.82	0.56	16.44	52.38
ID	44	0.000006	0.000158	0.000986	0.60	13.77	49.85	1.89	33.79	76.07
IL	102	0.000000	0.000059	0.000709	0.04	5.63	41.68	0.32	32.42	85.18
IN	92	0.000003	0.000053	0.000414	0.30	5.06	29.45	1.49	21.00	67.53
IA	99	0.000004	0.000082	0.000351	0.39	7.65	26.14	2.39	33.87	68.64
KS	105	0.000004	0.000225	0.000680	0.39	18.46	40.67	0.92	34.66	61.63
KY	120	0.000002	0.000133	0.000642	0.24	11.80	39.29	0.63	26.18	63.19
LA	64	0.000005	0.000087	0.000561	0.51	8.09	36.13	2.86	33.36	76.28
ME	16	0.000012	0.000032	0.000075	1.17	3.11	7.03	28.83	52.33	72.12
MD	24	0.000002	0.000030	0.000169	0.16	2.94	14.56	0.45	7.84	32.33
MA	14	0.000001	0.000007	0.000321	0.09	0.66	24.45	0.61	4.53	69.78
MI	83	0.000001	0.000038	0.000205	0.09	3.67	17.13	0.14	5.76	24.89
MN	87	0.000001	0.000037	0.000581	0.11	3.64	36.94	0.12	4.07	39.74
MS	82	0.000013	0.000109	0.000419	1.26	9.86	29.70	2.21	16.22	42.77
MO	115	0.000002	0.000120	0.000473	0.16	10.79	32.29	0.93	41.48	73.65
MT	56	0.000013	0.000298	0.001585	1.34	23.08	61.51	2.62	37.26	75.97
NE	93	0.000004	0.000177	0.001776	0.37	15.12	64.16	0.46	18.24	69.15
NV	17	0.000003	0.000207	0.000905	0.35	17.28	47.71	0.33	16.58	46.47
NH	10	0.000005	0.000026	0.000076	0.49	2.59	7.12	2.15	10.62	25.50
NJ	21	0.000002	0.000005	0.000034	0.16	0.49	3.31	2.09	6.02	31.00
NM	33	0.000004	0.000132	0.001133	0.45	11.78	53.32	0.88	20.77	69.17
NY	62	0.000001	0.000022	0.000128	0.07	2.12	11.43	0.44	11.28	43.05
NC	100	0.000004	0.000055	0.000977	0.38	5.22	49.62	0.84	10.92	68.67
ND	53	0.000013	0.000269	0.001130	1.33	21.34	53.25	2.02	29.45	63.67
OH	88	0.000001	0.000028	0.000244	0.13	2.73	19.74	1.09	18.92	67.20
OK	77	0.000005	0.000124	0.000820	0.46	11.09	45.26	1.30	26.31	70.30
OR	36	0.000003	0.000078	0.000483	0.26	7.33	32.75	1.25	28.05	70.58
PA	67	0.000001	0.000017	0.000242	0.14	1.67	19.61	0.21	2.47	26.67
SC	46	0.000006	0.000059	0.000328	0.64	5.63	24.85	1.87	15.12	49.69
SD	66	0.000008	0.000250	0.000758	0.82	20.11	43.32	0.72	18.15	40.23
TN	95	0.000003	0.000085	0.000675	0.27	7.88	40.49	0.70	17.94	63.50
TX	254	0.000001	0.000182	0.007850	0.11	15.54	88.78	0.08	11.84	85.24
UT	29	0.000002	0.000118	0.000661	0.25	10.66	39.99	0.86	29.31	69.83
VT	14	0.000011	0.000047	0.000119	1.07	4.55	10.71	10.95	35.07	57.63
VA	135	0.000002	0.000099	0.001099	0.16	9.10	52.56	0.63	28.54	81.56
WA	39	0.000001	0.000049	0.000645	0.09	4.66	39.40	0.34	15.21	70.45
WV	55	0.000010	0.000111	0.000926	0.99	10.09	48.28	3.18	26.89	75.36
WI	72	0.000001	0.000024	0.000541	0.14	2.34	35.29	0.35	5.49	57.00
WY	23	0.000026	0.000183	0.000518	2.54	15.58	34.31	3.20	18.94	39.80
States with 5 or fewer counties										
DE	3	0.000003	0.000013	0.000026	0.35	1.29	2.55			
DC	1		0.000004			0.44				
HI	5	0.000003	0.000045	0.022305	0.33	4.36	95.74			
RI	5	0.000003	0.000031	0.000038	0.32	2.99	3.69			

Table 6. County variance reductions for PPH with minimum, median, and maximum sampling variances

State	Number of Counties	County Sampling Variances			National Model % Variance Reductions			State Model % Variance Reductions		
		Minimum	Median	Maximum	Minimum	Median	Maximum	Minimum	Median	Maximum
AL	67	0.000157	0.002713	0.014745	9.49	64.47	90.80	31.48	88.83	97.74
AK	24	0.000519	0.013020	0.061369	25.75	89.70	97.62	100.00	100.00	100.00
AZ	15	0.000066	0.002794	0.017086	4.25	65.14	91.96	100.00	100.00	100.00
AR	75	0.000289	0.004469	0.015420	16.20	74.93	91.16	56.90	95.33	98.60
CA	58	0.000020	0.001030	0.071084	1.31	40.80	97.94	2.28	54.71	98.81
CO	63	0.000176	0.009335	0.084704	10.56	86.20	98.27	15.49	90.65	98.88
CN	8	0.000097	0.000517	0.000890	6.10	25.69	37.32	100.00	100.00	100.00
FL	66	0.000099	0.001899	0.024534	6.21	55.96	94.26	14.12	75.93	97.60
GA	159	0.000180	0.006544	0.035329	10.74	81.40	95.94	21.45	90.85	98.17
ID	44	0.000392	0.007322	0.071696	20.76	83.04	97.96	32.79	90.12	98.89
IL	102	0.000027	0.002938	0.023318	1.76	66.28	93.98	7.26	89.55	98.55
IN	92	0.000124	0.002642	0.017670	7.67	63.87	92.20	34.48	91.80	98.68
IA	99	0.000225	0.003622	0.009495	13.08	70.78	86.39	24.19	83.71	93.08
KS	105	0.000216	0.007403	0.023455	12.61	83.20	94.01	100.00	100.00	100.00
KY	120	0.000108	0.005129	0.023073	6.76	77.43	93.91	14.92	89.25	97.39
LA	64	0.000255	0.003955	0.033185	14.55	72.57	95.69	98.84	99.92	99.99
ME	16	0.000408	0.001452	0.005165	21.42	49.27	77.57	100.00	100.00	100.00
MD	24	0.000102	0.001055	0.005853	6.41	41.38	79.65	100.00	100.00	100.00
MA	14	0.000063	0.000313	0.039656	4.02	17.32	96.37	91.27	98.12	99.98
MI	83	0.000054	0.001719	0.020491	3.48	53.49	93.20	9.21	76.37	97.47
MN	87	0.000067	0.001905	0.033714	4.31	56.03	95.75	6.56	66.54	97.24
MS	82	0.000604	0.005396	0.023891	28.79	78.30	94.11	38.83	85.00	96.17
MO	115	0.000087	0.004717	0.016943	5.49	75.93	91.89	8.64	83.69	94.85
MT	56	0.000651	0.012311	0.061474	30.34	89.17	97.63	100.00	100.00	100.00
NE	93	0.000208	0.006397	0.068969	12.21	81.06	97.88	100.00	100.00	100.00
NV	17	0.000151	0.010466	0.043772	9.18	87.50	96.70	100.00	100.00	100.00
NH	10	0.000264	0.001216	0.003519	15.03	44.85	70.19	100.00	100.00	100.00
NJ	21	0.000110	0.000321	0.002693	6.87	17.69	64.28	23.59	47.37	88.28
NM	33	0.000200	0.007252	0.066009	11.79	82.91	97.79	100.00	100.00	100.00
NY	62	0.000062	0.001015	0.010799	3.97	40.43	87.84	17.71	77.95	97.41
NC	100	0.000167	0.002282	0.041561	10.07	60.42	96.53	19.48	76.73	98.36
ND	53	0.000640	0.009678	0.050056	29.96	86.62	97.10	100.00	100.00	100.00
OH	88	0.000060	0.001396	0.009011	3.87	48.29	85.77	15.57	81.07	96.51
OK	77	0.000168	0.004687	0.027789	10.13	75.82	94.89	81.59	99.20	99.86
OR	36	0.000133	0.002618	0.015191	8.15	63.65	91.04	35.02	91.41	98.41
PA	67	0.000057	0.000713	0.017423	3.68	32.28	92.10	11.14	60.98	97.45
SC	46	0.000261	0.002644	0.013071	14.87	63.88	89.74	58.15	93.36	98.58
SD	66	0.000536	0.010491	0.067097	26.39	87.53	97.82	100.00	100.00	100.00
TN	95	0.000144	0.003216	0.023021	8.78	68.26	93.90	100.00	100.00	100.00
TX	254	0.000056	0.007945	0.651198	3.61	84.16	99.77	5.35	88.90	99.85
UT	29	0.000207	0.008147	0.093894	12.15	84.50	98.43	3.73	60.44	94.63
VT	14	0.000558	0.001971	0.007816	27.18	56.87	83.95	100.00	100.00	100.00
VA	135	0.000116	0.004020	0.036676	7.18	72.89	96.08	4.03	59.35	93.02
WA	39	0.000052	0.002294	0.022549	3.34	60.55	93.78	100.00	100.00	100.00
WV	55	0.000364	0.004028	0.038480	19.59	72.93	96.26	100.00	100.00	100.00
WI	72	0.000090	0.001304	0.034511	5.65	46.59	95.85	8.28	56.80	97.21
WY	23	0.001153	0.008623	0.017455	43.55	85.22	92.11	31.05	77.10	87.20
States with 5 or fewer counties										
DE	3	0.000195	0.000751	0.001409	11.52	33.44	48.52			
DC	1	0.000179	0.000179	0.000179		10.68				
HI	5	0.000214	0.002612	0.172595	12.50	63.60	99.14			
RI	5	0.000169	0.001329	0.001658	10.14	47.06	52.61			

Figure 1. Percent Variance Reduction in %OCC by State

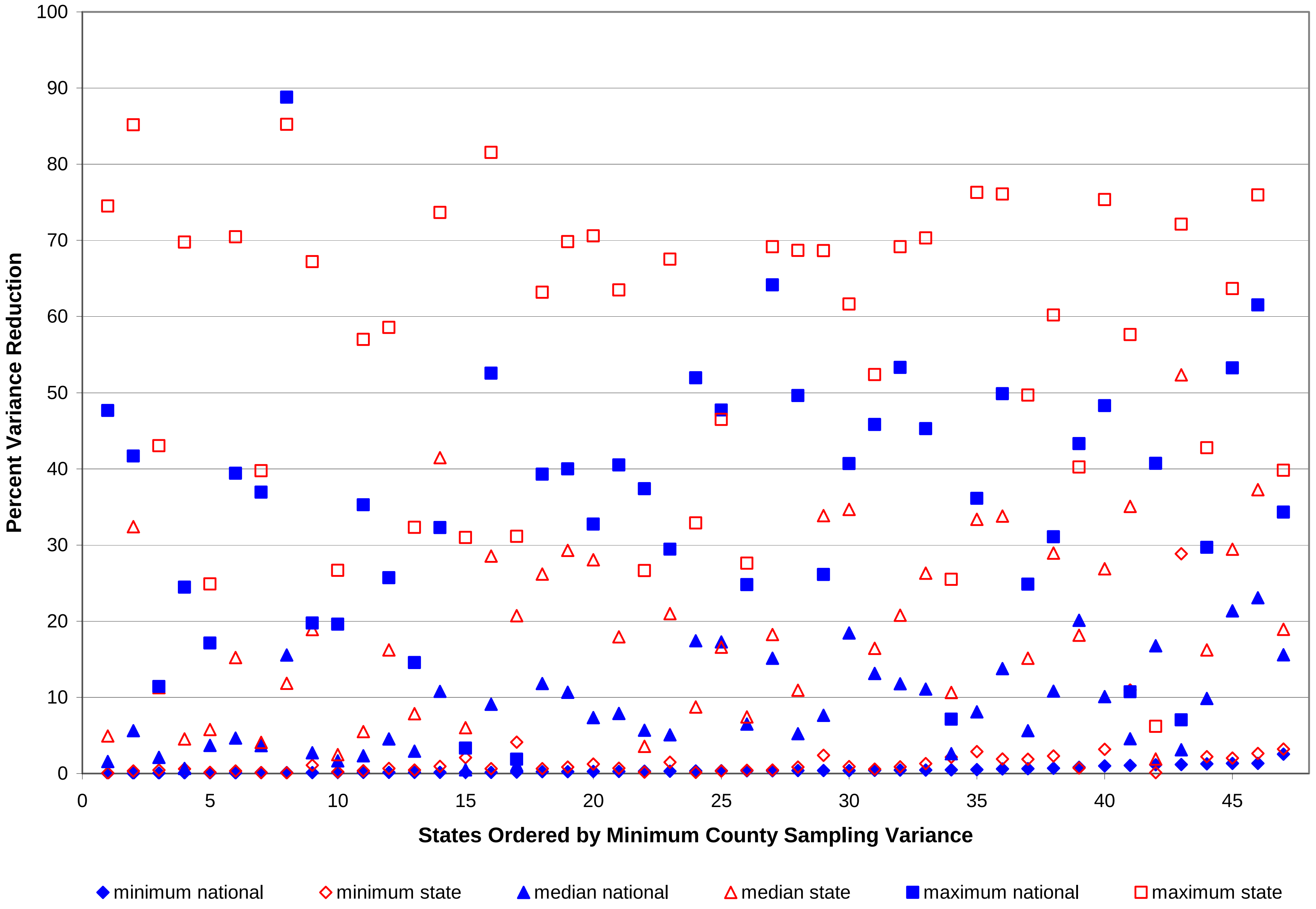


Figure 2. Percent Variance Reduction in PPH by State

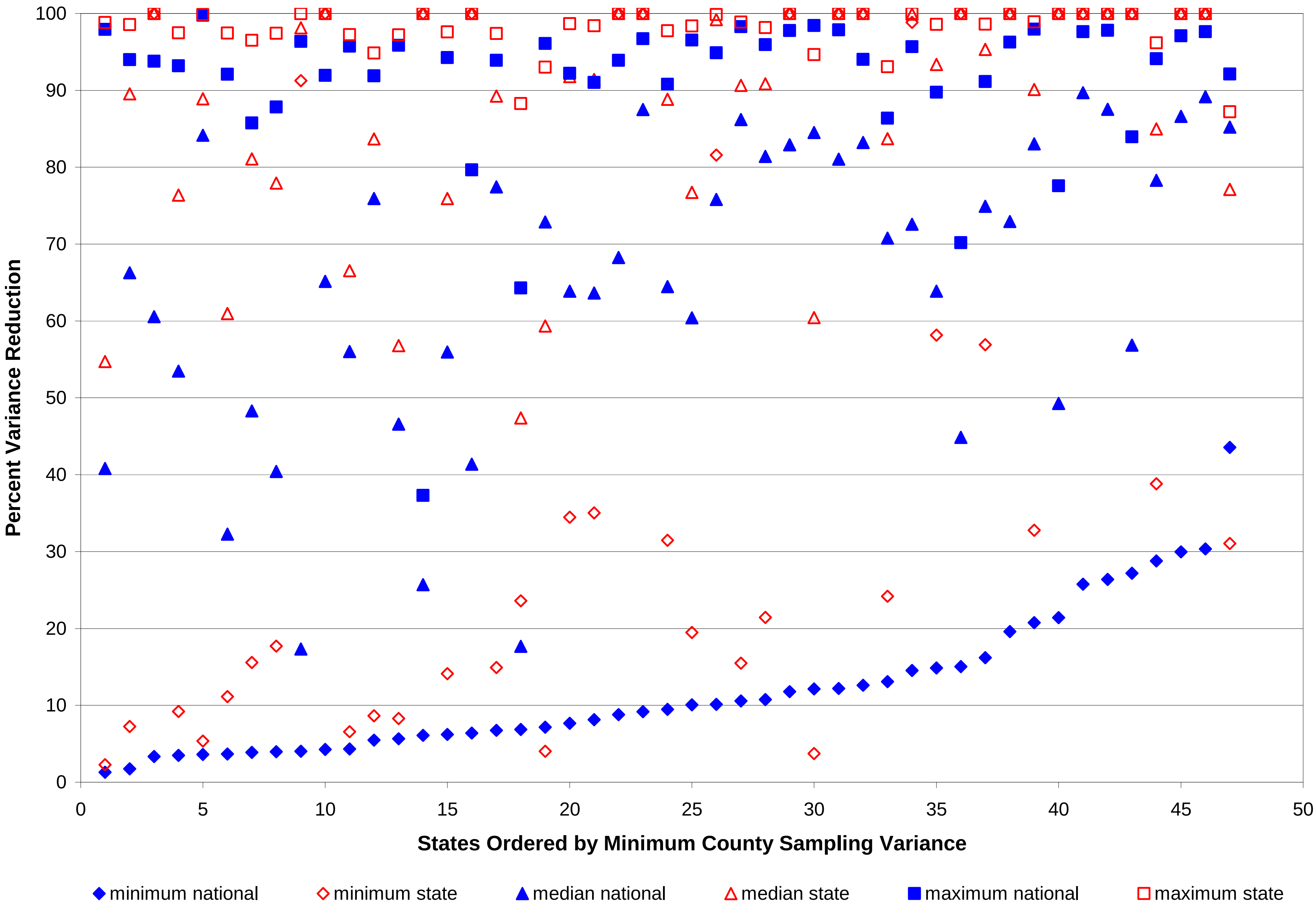


Figure 3. Comparison of County Estimates of Change in %OCC for Mississippi

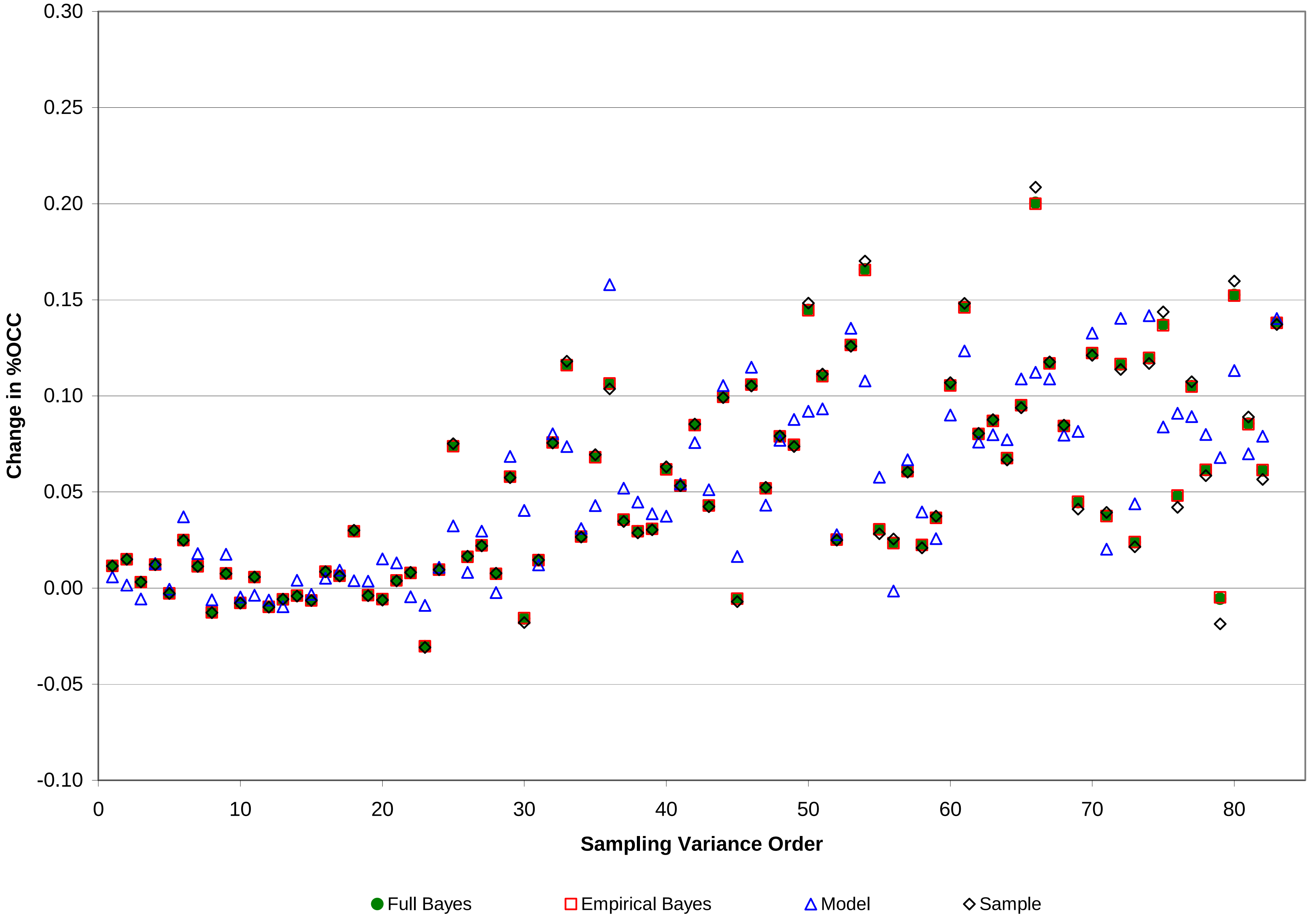
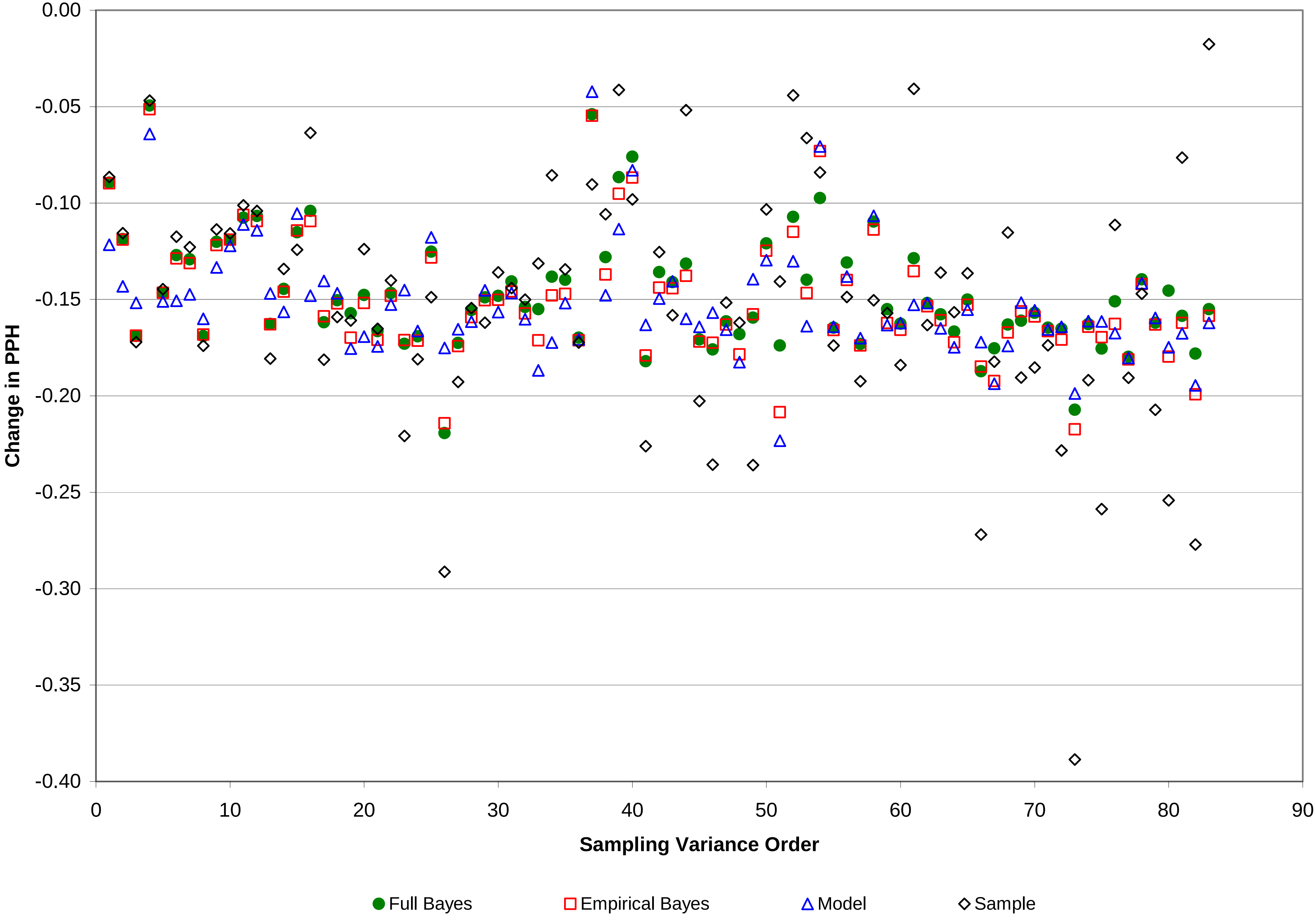


Figure 4. Comparison of County Estimates of Change in PPH for Mississippi



Appendix 2 – Approximate Variance of Estimate of Total Persons in a County

If there is complete response, then the estimated variance of total persons is

$$\hat{V}(\text{persons}_{90}) = \sum_{h,90} N_{h,90}^{2}\left(\frac{s_{h,90}^{2}\left(1 - f_{h,90}\right)}{n_{h,90}}\right) , \qquad \text{(A2.1)}$$

where $s_{h,90}^{2} = \dfrac{\sum_{i}\left(x_{h,i,90} - \bar{x}_{h,90}\right)^{2}}{(n_{h,90} - 1)}$ is the estimated variance in the number of persons per HU, occupied and vacant,

$x_{h,i,90}$ is the number of persons in HU i in stratum h, and

$\bar{x}_{h,90}$ is the mean number of persons in a HU in stratum h.

But occupied units have nonresponse and vacant units don't, so we can't use $s_{h,90}^{2}$ and can't simply change it to a formula that uses the responding units, because the proportions of occupied and vacant responding units are not the same as in the full population. If instead we base the variance on only the occupied units and know how many there are, then the variance estimation formula is (A2.2), where the subscripts o and r respectively denote occupied and responding units.

$$\hat{V}(\text{persons}_{90}) = \sum_{h,90} N_{ho,90}^{2}\left(\frac{s_{hor,90}^{2}\left(1 - \rho_{ho,90} f_{h,90}\right)}{\rho_{ho,90}\, n_{ho,90}}\right) , \qquad \text{(A2.2)}$$

where $s_{hor,90}^{2} = \dfrac{\sum_{i}\left(x_{hor,i,90} - \bar{x}_{hor,90}\right)^{2}}{(\rho_{ho,90}\, n_{ho,90} - 1)}$ is the estimated variance of the number of persons per occupied HU in stratum h based on the responding HUs,

$N_{ho,90}$ is the number of occupied units in stratum h,

$x_{hor,i,90}$ is the number of persons in occupied responding HU i in stratum h,

$\bar{x}_{hor,90}$ is the mean number of persons in occupied responding HUs in stratum h,

$n_{ho,90}$ is the number of sample occupied units in stratum h, and

$\rho_{ho,90}$ is the response rate for occupied HUs in stratum h.

However, we don't know $N_{ho,90}$. An estimate is $N_{h,90}\, p_{h,90}$ but then we must account for $V(p_{h,90})$ and (A2.2) is no longer the appropriate formula.

Instead, we develop an approximation by letting $T_{h,90}$ be the total population in stratum h and writing it as

$$T_{h,90} = N_{h,90}\left(\frac{N_{hor,90}}{N_{h,90}}\bar{x}_{hor,90} + \frac{N_{honr,90}}{N_{h,90}}\bar{x}_{honr,90} + \frac{N_{hv,90}}{N_{h,90}}\bar{x}_{hv,90}\right), \tag{A2.3}$$

where the subscripts nr and v denote nonresponding and vacant units,
$N_{hor,90}$ is the number of occupied respondent units in stratum h,
$N_{honr,90}$ is the number of occupied nonrespondent units in stratum h,
$N_{hv,90}$ is the number of vacant units in stratum h,
$\bar{x}_{honr,90}$ is the mean number of persons in occupied respondent HUs in stratum h, and
$\bar{x}_{hv,90}$ is the mean number of persons in vacant respondent HUs in stratum h.

We know that $\bar{x}_{hv,90} = 0$. If we assume that $\bar{x}_{hor,90} = \bar{x}_{honr,90} = \bar{x}_{ho,90}$ and note that $\frac{N_{hor,90}}{N_{h,90}} + \frac{N_{honr,90}}{N_{h,90}} = P_{ho,90}$, the proportion of occupied units in stratum h, then we can write $T_{h,90} = N_{h,90} P_{ho,90} \bar{x}_{ho,90}$. This can be estimated as

$$\hat{T}_{h,90} = N_{h,90} p_{ho,90} \bar{x}_{hor,90} \, . \tag{A2.4}$$

For simplicity in developing an approximation for $\hat{V}(\hat{T}_{h,90})$, we drop the subscripts h and 90 and consider a single stratum. This leaves us with $\hat{V}(\hat{T}) = \hat{V}(N p_o \bar{x}_{or})$. Now use the relationship

$$V(a) = E(V(a|b)) + V(E(a|b)), \tag{A2.5}$$

where a and b are random variables and a is a function of b. In this case a = $N p_o \bar{x}_{or}$ and b = p_o. For the first term on the right hand side of (A2.5),

$$\begin{aligned} E(V(N p_o \bar{x}_{or})) &= E(N^2 p_o^2 V(\bar{x}_{or})) \\ &= N^2 E(p_o^2) V(\bar{x}_{or}) \\ &= N^2 [V(p_o) + E^2(p_o)] V(\bar{x}_{or}) \; . \end{aligned} \tag{A2.6}$$

For the second term on the right hand side of (A2.5),

$$\begin{aligned} V(E(N p_o \bar{x}_{or})) &= V(N p_o E(\bar{x}_{or})) \\ &= N^2 E^2(\bar{x}_{or}) V(p_o) \end{aligned} \tag{A2.7}$$

Now approximate $E(p_o)$ with p_o to get

$$\hat{V}(N\,p_o\,\overline{x}_{or}) \approx N^2\{V(p_o)[V(\overline{x}_{or})+\overline{x}_{or}^2]+p_o^2\,V(\overline{x}_{or})\}\,. \tag{A2.8}$$

In addition, substitute the estimates $\hat{V}(p_o)=\dfrac{p(1-p)(1-f)}{(n-1)}$ from (8) for $V(p_o)$ and $\hat{V}(\overline{x}_{or})=\dfrac{s_{or}^2(1-\rho_o f)}{\rho_o n_o}$ for $V(\overline{x}_{or})$, from the term in parentheses on the right hand side (A2.2), to get the overall stratum estimate of variance used in (9).

$$\hat{V}(N\,p_o\,\overline{x}_{or}) \approx N^2\left\{\frac{p(1-p)(1-f)}{(n-1)}\left[\frac{s_{or}^2(1-\rho_o f)}{\rho_o n_o}+\overline{x}_{or}^2\right]+p_o^2\frac{s_{or}^2(1-\rho_o f)}{\rho_o n_o}\right\}. \tag{A2.9}$$