Serial Comparisons in Small Domain Models: A Residual-based Approach*

Wesley Basel, Sam Hawala and David Powers
U.S. Census Bureau, Washington, D.C. 20233

Abstract

The U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program produces model-based estimates for small geographic areas using household survey data, administrative records, postcensal population estimates and decennial census data. This paper proposes and evaluates a method for making year-to-year statistical comparisons of poverty at the county level. The method uses aggregations of regression residuals in order to estimate the underlying serial correlation in SAIPE county-level estimates. Three residual-based estimators for the model error correlation are considered, with alternative weights used for each. The estimators are evaluated using simulations under the assumed error specification, and the effect of a heteroscedastic departure from these assumptions is discussed.

Key Words: small area, small domain, statistical comparison, census, SAIPE, poverty

1. Background

The SAIPE program produces model-based estimates of poverty that combine direct estimates from the American Community Survey (ACS) with regression predictions based on administrative records, postcensal population estimates and decennial census data. For both the survey data and the explanatory data, individual units are aggregated for the specified geographic area and year, producing inputs and estimates that are interpreted as single-year or annual data. The modeling techniques allow the SAIPE program to produce annual estimates of child poverty for all school districts and all counties, regardless of population size.

There is an interest to determine which areas in the SAIPE dataset have statistically significant changes in child poverty rate between pairs of years. In order to make these

* This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

statistical comparisons, the correlation in model errors between pairs of years must be estimated. This document discusses a method to estimate this year-to-year model error correlation for county-level estimates.

The layout of the paper is as follows. First, we introduce the structure of the model in Section 2. In Section 3, we motivate a general formula for the estimator of the year-to-year covariance and propose three alternative estimators using different weighting matrices. In Section 4, we present results from a simulation study of the three estimators. Section 5 reports empirical results for the years 2005-2008 and discusses the potential effect of one type of misspecification on the estimator. We give concluding remarks in Section 6.

2. Structure of the Model

In general, the SAIPE program's county poverty model follows a Fay-Herriot, or shrinkage, approach, by specifying both a sampling model and a regression model for the true value of log(poverty) (Fay and Herriot, 1979; Bell, 1999). The empirical-Bayes, best linear unbiased predictor (EBLUP) is then a weighted average of the direct estimate from the ACS sample and the predictions from the regression model. For a single year, the specification of this model is given in (1) below.

For $i = 1,2, \ldots, m$ areas,

$$\begin{aligned} \log(y_i) &= \log(Y_i) + e_i \\ \log(Y_i) &= \mathbf{x}_i'\beta + u_i \end{aligned} \tag{1}$$

where $\log(y_i)$ represents the direct survey estimate of log(poverty) from a single-year sample of the ACS, $\log(Y_i)$ is the logarithm of the unobservable true value of poverty, and $\mathbf{x}_i$ is a k x 1 vector of explanatory variables on the log scale. The model errors, u_i, are assumed i.i.d., and the sampling errors, e_i, are assumed independent. The ACS sampling variance for a given county is estimated directly from the sample using a successive difference replication method described in the ACS documentation (U.S. Census Bureau, 2009), and then are assumed known. For current SAIPE production, both the model and sampling error terms are assumed normally distributed in the log scale, and both the regression parameters and the model error variance are estimated by maximum likelihood.

For a final estimate of the SAIPE poverty in the native, or exponentiated scale, there are three more steps. First, the direct and indirect estimates are combined using an efficient weighting described in Bell (1999), producing the EBLUP on the log-scale. Then, the shrinkage estimates are transformed to the native scale, using the properties of the log-normal distribution to adjust the point estimates and associated standard errors. Finally,

the resulting estimates are controlled to state-level estimates produced by a separate model. Under the log-normal assumption, however, the standard errors and all cross-area and cross-time correlations are completely specified by the variance components for the errors in the log-scale equation (1). So for the purposes of this paper, we focus only on specifying these components in the log-scale and evaluating alternative estimators for them.

For estimation of the parameters in the current SAIPE production model, the two models in (1) are combined by substituting the regression specification into the sampling model. Also for notational convenience, we define $z_i \equiv \log(y_i)$. With this, the two-year specification of this model is given in (2) below.

For $i = 1,2, \ldots, m$ areas, and two arbitrary years, labeled s and t,

$$\begin{aligned} z_{is} &= \mathbf{x}'_{is}\beta_s + u_{is} + e_{is} \\ z_{it} &= \mathbf{x}'_{it}\beta_t + u_{it} + e_{it} \end{aligned} \tag{2}$$

The sampling errors are assumed independent across both areas and time, while the model errors are assumed i.i.d. for a given year and with a constant year-to-year covariance for given i. Sampling errors are assumed independent from model errors, for any combination of i, j, s and t. Thus, for the n x 1 vectors $\mathbf{e}_s$, $\mathbf{e}_t$, $\mathbf{u}_s$ and $\mathbf{u}_t$,

$$\begin{aligned} \begin{pmatrix} \mathbf{u}_s \\ \mathbf{u}_t \end{pmatrix} &\sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_s^2 I & \sigma_{st} I \\ \sigma_{st} I & \sigma_t^2 I \end{pmatrix} \right\} \\ \begin{pmatrix} \mathbf{e}_s \\ \mathbf{e}_t \end{pmatrix} &\sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V_s & 0 \\ 0 & V_t \end{pmatrix} \right\} \end{aligned} \tag{3}$$

where $V_s = \text{diag}\{v_{is}\}$ is the ACS sampling variance for each area, i. As discussed earlier, these sampling variances are estimated directly from the sample and are then assumed known.

For derivations and exposition, it can be simpler to stack the model into a single matrix form, keeping in mind that the serial structure is not utilized in estimating the single-year parameters. So we define $\mathbf{z} = \{\mathbf{z}_s', \mathbf{z}_t'\}'$, and similarly for β and $\mathbf{u} + \mathbf{e}$, and we define $\mathbf{X}$ as block diagonal $\{\mathbf{X}_s\ ,\ \mathbf{0}\ //\ \mathbf{0}\ ,\ \mathbf{X}_t\}$, where $\mathbf{X}_r = \text{col}_{1 \le i \le m} \mathbf{x}_{i,r}$ for $r = s$ or t. The framework then simplifies to:

$$\begin{aligned} &\mathbf{z} = \mathbf{X}\beta + \mathbf{u} + \mathbf{e}, \quad \text{where } \mathbf{u} + \mathbf{e} \sim \mathbf{N}\{\mathbf{0}, \Omega\} \\ &\Omega = \begin{pmatrix} \Sigma_s = V_s + \sigma_s^2 I & \sigma_{st} I \\ \sigma_{st} I & \Sigma_t = V_t + \sigma_t^2 I \end{pmatrix} \end{aligned} \tag{4}$$

This is a similar structure to the seemingly-unrelated regression (SUR) model originally proposed by Zellner; see for example, Amemiya (1985), p. 197. There are two primary differences between the usual estimation procedures for that model and the approach taken in this paper. Contemporaneous heteroscedasticity is a fairly uncommon assumption for that model. Furthermore, in SAIPE production, the parameter estimates for each year are obtained from single-year data only, with no consideration of the potential serial information contained in the above structure. The β and σ_s^2 are estimated by their marginal (i.e. single-year) likelihoods, and so are still consistent for the parameters of the multiyear model, but they are not asymptotically efficient. In this case, maximum likelihood estimates are obtained for each year. For year *s*, for example:

$$\hat{\beta}_s(\hat{\sigma}_s^2) = \left(\mathbf{X}_s'\hat{\Sigma}_s^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}_s'\hat{\Sigma}_s^{-1}\mathbf{y}_s$$
$$\text{where}\,\hat{\Sigma}_s = \text{diag}\{\hat{\sigma}_s^2 + v_i\} \qquad (5)$$

3. Estimation of σ_{st}

A natural form for the estimator of σ_{st}, obtained from a method-of-moments approach, is:

$$\sigma_{st}^{(A)} = \frac{1}{m}(\mathbf{z_s} - \mathbf{X_s}\hat{\beta}_s)'\mathbf{W}^{(A)}(\mathbf{z}_t - \mathbf{X}_t\hat{\beta}_t) \qquad (6)$$

where $\mathbf{W}^{(A)}$ is an alternative weighting matrix. For this paper, all the choices of **W** are diagonal matrices. If one chooses **W** such that the cross-product terms for each individual area, *i*, are unbiased for σ_{st}, then the properties of the normal distribution assumed in (3) would imply finite variance for the sum of the cross-products over the sample. And thus convergence of $(1/m)\mathbf{W}^{(A)}$ would suffice to imply consistency and asymptotic normality for the general estimator for σ_{st} in (6). Without the normality assumption stated in (3), some additional assumptions regarding finiteness and convergence of the third and fourth moments would be needed. So the first criterion for a candidate W is the unbiasedness property. We note that:

$$\text{E}[(\mathbf{z_s} - \mathbf{X_s}\hat{\beta}_s)'(\mathbf{z}_t - \mathbf{X}_t\hat{\beta}_t)] = \text{E}[\mathbf{z_s}'(\mathbf{I} - \mathbf{M}_s)'(\mathbf{I} - \mathbf{M}_t)\mathbf{z}_t] = \sigma_{st}\text{trace}((\mathbf{I} - \mathbf{M}_s)(\mathbf{I} - \mathbf{M}_t)')$$
$$\mathbf{M_s} = \mathbf{X_s}(\mathbf{X_s}'\Sigma_s^{-1}\mathbf{X_s})^{-1}\mathbf{X_s}'\Sigma_s^{-1},\ \mathbf{M}_t = \mathbf{X}_t(\mathbf{X}_t'\Sigma_t^{-1}\mathbf{X}_t)^{-1}\mathbf{X}_t'\Sigma_t^{-1} \qquad (7)$$

where $\mathbf{M}_s$ is the generalized projection matrix for year *s*, and similarly for $\mathbf{M}_t$. So our first alternative is to set $(1/m)\mathbf{W}^{(A)}$ equal to the constant value defined by the inverse of the trace in (7). Note that the assumptions necessary to ensure convergence of $(1/m)\mathbf{W}^{(A)}$ essentially ensure that the trace of the projection matrices defined in (7) converge to *m*.

Thus, for large sample this simplest weighting converges to the simple average, $(1/m)$. For the SAIPE 2007-to-2008 covariance estimator described in Section 6, m = 2,845 and the trace defined in (7) equals 2,844.8.

This basic $\mathbf{W}^{(A)}$ is the weighting matrix used for one alternative that we consider, labeled Estimator III in the listing below (8.3). This is unlikely to be the most efficient estimator, however, since it weights all observations equally, even though we know there are substantial differences in their variances.

An intuitive weighting, labeled Estimator II in the listing (8.2), would be to pre-multiply each residual by the square-root of the error variance matrix, approximately standardizing the residuals. As a final weighting alternative, we define Estimator I in the listing (8.1) by pre-multiplying each residual by a full power of the error variance matrix, which allows for wider variation in the relative weights.

The three estimators are shown below. We will test these three estimators through simulations in the next section.

- **Estimator I**

$$\hat{\sigma}_{st} = \frac{\left(\hat{\Sigma}_s^{-1}\left(\mathbf{z}_s - \mathbf{X}_s\hat{\beta}_s\right)\right)' \left(\hat{\Sigma}_t^{-1}\left(\mathbf{z}_t - \mathbf{X}_t\hat{\beta}_t\right)\right)}{tr\left[\hat{\Sigma}_s^{-1}\left(\mathbf{I} - \hat{\mathbf{M}}_s\right)\left(\mathbf{I} - \hat{\mathbf{M}}_t\right)'\hat{\Sigma}_t^{-1}\right]} \tag{8.1}$$

- **Estimator II**

$$\hat{\sigma}_{st} = \frac{\left(\hat{\Sigma}_s^{-1/2}\left(\mathbf{z}_s - \mathbf{X}_s\hat{\beta}_s\right)\right)' \left(\hat{\Sigma}_t^{-1/2}\left(\mathbf{z}_t - \mathbf{X}_t\hat{\beta}_t\right)\right)}{tr\left[\hat{\Sigma}_s^{-1/2}\left(\mathbf{I} - \hat{\mathbf{M}}_s\right)\left(\mathbf{I} - \hat{\mathbf{M}}_t\right)'\hat{\Sigma}_t^{-1/2}\right]} \tag{8.2}$$

- **Estimator III**

$$\hat{\sigma}_{st} = \frac{\left(\mathbf{z}_s - \mathbf{X}_s\hat{\beta}_s\right)' \left(\mathbf{z}_t - \mathbf{X}_t\hat{\beta}_t\right)}{tr\left[\left(\mathbf{I} - \hat{\mathbf{M}}_s\right)\left(\mathbf{I} - \hat{\mathbf{M}}_t\right)'\right]} \tag{8.3}$$

4. Simulation Results

The setup described in the prior section yields consistent estimators, under the assumed error structure, but we did not derive an efficient weighting matrix. So to evaluate the performance of the three estimators above, we generated simulations using the data process as defined in (3) and (4). All simulation and empirical results pertain to estimators for the log of the number of children, ages 5 to 17 in families, in poverty.

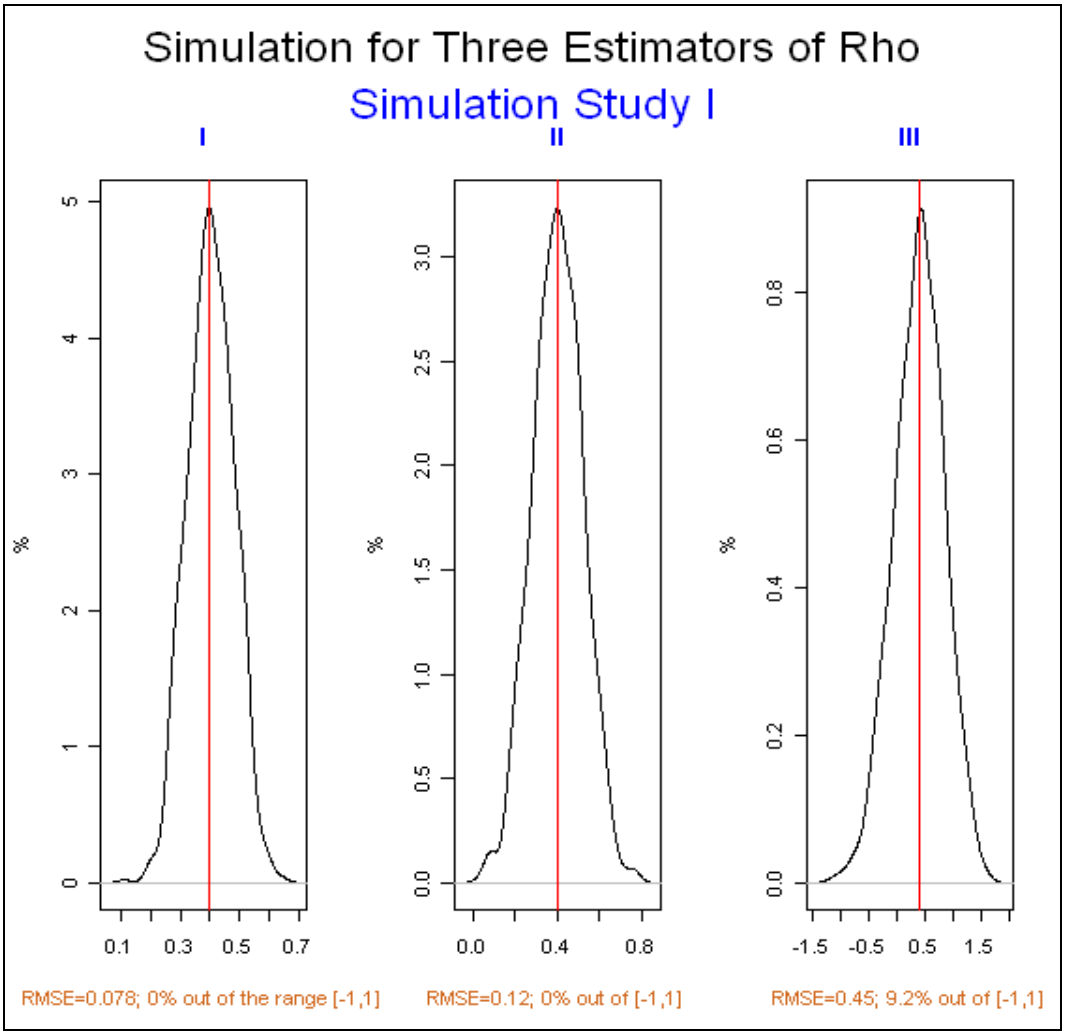

Figure 1: Estimated Density of $\hat{\rho}$, showing RMSE$(\hat{\rho})$ and Percentage of $|\hat{\rho}| > 1$ For Three Alternative Estimators, when the true ρ=0.4 (repetitions=1000).

We simulate the observed **z** = log(**y**) using the design matrix **X** of the k = 5 explanatory variables (for a description of these variables, see Bell, et. al. (2007)) for the years 2005 and 2006, treating these design matrices as fixed. For these years, m = 2,886. For the parameters of the model, apart from σ_{st}, namely β_s, β_t, V_s, V_t and the model error

variances, we use those from the published SAIPE estimates from the corresponding years, treating them as "true" values. We then generate simulations for each of several values of $\rho = \sigma_{st} / \sigma_s \sigma_t$. Figure 1 displays the summary results for 1000 such simulations for the three alternative estimators, for a true value of $\rho = 0.4$.

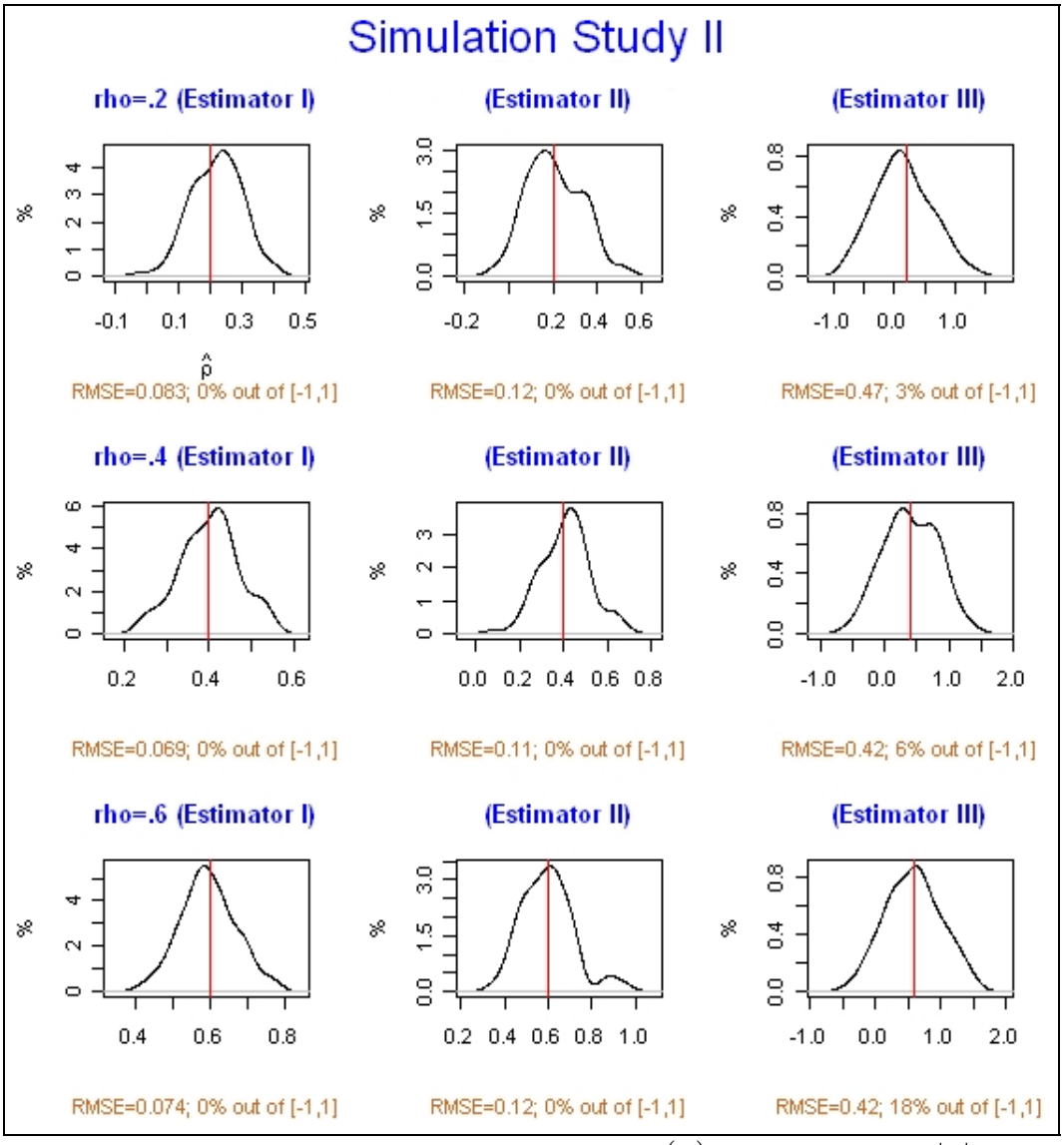

Figure 2: Estimated Density of $\hat{\rho}$, showing RMSE$(\hat{\rho})$ and Percentage of $|\hat{\rho}| > 1$
For Alternative True Values of ρ (repetitions = 100)

Despite the different shapes of the densities of the three estimators, the mean values are fairly similar to one another, confirming the theoretical consistency of the three estimators. The spread differs considerably, however. Estimator I, which uses a full power of the error variance matrix as its weights, has the lowest root mean-squared error

(RMSE) of 0.08. Estimator II, using the square-root of the error-variance matrix, displays somewhat higher dispersion with a RMSE of 0.12. Estimator III, which is un-weighted, is considerably worse, with an MSE of 0.45 and with 9% of the occurrences falling outside the parameter space of [-1, 1]. Under the assumption that (3) correctly specifies the error structure, the un-weighted Estimator III is a poor choice.

Figure 2 displays the simulated distributions for the three estimators under three alternative values of ρ, 0.2, 0.4 and 0.6, using 100 simulated repetitions. The estimators remain unbiased, with similar dispersion to one another, for all values of the true parameter near our prior expectations.

5. Empirical Results

The single-year ACS-based SAIPE estimator defined in (5) was applied at the county level, for children ages 5 to 17 in families, to four ACS survey years: 2005, 2006, 2007 and 2008. The year-to-year covariance Estimators I, II and III defined in Section 3 were then calculated directly for each of the six intervals defined by these four single-year estimates: 2005 to 2006, 2006 to 2007, 2007 to 2008, 2005 to 2007, 2006 to 2008, 2005 to 2008. The explanatory variables in the SAIPE county model are the logarithms of tax-based child exemptions and child exemptions in poverty, SNAP (Supplemental Nutrition Assistance Program) participants, child population estimates, and the decennial census estimate of poverty for the age group. Additional information about these variables is available in Bell, et. al. (2007).

The resulting estimates for year-to-year correlation ($\rho = \sigma_{st} / \sigma_s \sigma_t$) in the model error for log-poverty of related children ages 5 to 17 are reported for Estimator I in Table 1. The assumed error structure for these estimates is as defined in (3) and as used for the simulations reported in the previous section. Note that the correlations reported for intervals longer than one year were estimated directly from the residual-based approach defined by Estimator I, and not calculated from the one-year intervals estimates using any assumed time-series model.

	2005	2006	2007	2008
2005	1	0.34	0.27	0.26
2006	0.34	1	0.46	0.40
2007	0.27	0.46	1	0.40
2008	0.26	0.40	0.40	1

Table 1: Year-to-year Correlation Coefficient Estimates of the SAIPE Model Error Log-poverty of Related Children, ages 5 to 17 (from Estimator I)

All three of the one-year interval estimates for the correlation coefficient were between 0.3 and 0.5. The multi-year interval estimates did not quickly decay, indicating some persistence in model errors over time. We did expect fairly high year-to-year correlation and some degree of persistence in the model errors, due to the relative stability of the explanatory variables in the model, particularly when compared to the apparent trends in true poverty levels during this period.

For any given county, the difference between the log-linear fitted value based on these variables and the true poverty value will be due either to failure of the model structure (due to non-linearity, variability of the relation over counties, etc.), or to county-specific variations in the explanatory variables (tax-filing behavior, SNAP participation, or the relations of other variables). This second factor would likely persist over time since the explanatory variables are relatively stable over time.

The empirical results for Estimator II corresponding to the results reported in Table I were similar, but they yielded higher actual values of the correlation in all cases. For example, the value of Estimator II for the 2007-2008 model error correlation was 0.59, compared to the value of 0.40 reported in Table 1 for Estimator I. This difference is more than what was expected from the simulations. For Estimator III, the resulting values for the estimates were all higher than one, some by a large amount. From the simulation results, we expected a higher probability of out-of-range results for Estimator III, but the actual empirical results were beyond the extremes noted in the simulations. Furthermore, this wide range of results was repeated for all the other single-year intervals. These results indicate likely misspecification of the model, with the departure substantially affecting the estimates of year-to-year correlation.

The homoscedasticity assumption for the model error implied by (3) is known not to hold for the SAIPE model. Most tests of heteroscedasticity remaining after adjusting for the assumed structure fail to support the structure specified in (3). A modeled adjustment for the remaining heteroscedasticity has not been applied to the production version of the SAIPE model in the past for two primary reasons. First, the regression coefficients, and thus the resulting predictions of $\mathbf{Z}$, remain consistent and asymptotically normal despite misspecification of the error structure. So the impact on the final SAIPE estimates is small. The second reason is that the weights of the Fay-Herriot shrinkage estimate require a precise decomposition of the overall error variance into sampling error variance and model error variance. There is some evidence that the replicate weight method for estimating the sampling variance may yield inconsistent estimates for small, unpublished, areas on the log scale. So it is not known whether the evidence of remaining heteroscedasticity is a result of non-constant model error variance or of poor estimation

of the sampling variance. For this reason, ongoing research efforts have concentrated on developing a generalized variance function to improve the precision of the sampling variance estimates (Maples, Bell and Huang, 2009).

6. Conclusions and Further Research

In this work we examined three estimators for the year-to-year SAIPE model error correlation, ρ_{st}. We determined the most efficient among three alternatives, within the assumed error structure, using simulations. For the most efficient among the three (Estimator I), we empirically estimated ρ for all combinations of years between 2005 and 2008. We observed that ρ declines with greater time lag, but we do not yet have a sufficient number of time periods to confidently model a structure to the serial behavior. However, these empirical estimates of ρ can potentially be used to support statistical comparisons of SAIPE data between pairs of years for individual counties.

The likely misspecification of the error variance structure may have a substantial impact on the estimated year-to-year model error correlation estimates. This might be indicated by the large deviation between the simulation results and the actual empirical estimates of the three alternatives. One potential type of misspecification of the error variance structure was discussed in more detail, namely heteroscedasticity. In theory, such misspecification has little impact on the Fay-Herriot model predictions, but a large impact on estimates of the standard errors and inter-year covariance.

Ongoing research is currently focused on specifying a model in the rate scale, rather than the log-level scale. This approach should yield more reliable sampling variance estimates for the smallest, unpublished counties since rate transformations of variables tend to be more stable across population size categories. Use of a generalized variance function for the sampling variance in the rate scale could help further. With more reliable sampling variance estimates, a more realistic error structure could be assumed for the model errors. As this research continues, the possibility of non-constant serial correlation will also be considered.

References

Amemiya, Takeshi. *Advanced Econometrics*, 1985. Harvard University Press.

Bell, William. "Accounting for Uncertainty About Variances in Small Area Estimation," 1999. U.S. Census Bureau. <http://www.census.gov/did/www/saipe/publications/files/Bell99.pdf>

Bell, William, Wesley Basel, Craig Cruse, Lucinda Dalzell, Jerry Maples, Brett O'Hara, and David Powers. "Use of ACS Data to Produce SAIPE Model-Based Estimates of Poverty for Counties," prepared in September 2007, <http://www.census.gov/did/www/saipe/publications/files/report.pdf>

Hansen, Lars Peter. "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica.* Vol. 50, No. 4 (July, 1982), pp. 1029-1054.

Fay, Robert E., III, and Roger A. Herriot. "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data," *Journal of the American Statistical Association.* Vol. 74, No. 366 (Jun., 1979), pp. 269-277.

Maples, Jerry J., William Bell, and Elizabeth Huang. "Small-Area Variance Modeling with Application to County Poverty Estimates from the American Community Survey," 2009. *JSM Proceedings, Section on Survey Research Methods.* Alexandria, VA: American Statistical Association.

U.S. Census Bureau. 2009. "Chapter 12: Variance Estimation." *ACS Design and Methodology*. <http://www.census.gov/acs/www/methodology/methodology_main>