Time-Series Cross-Sectional Approach for Small Area Poverty Models[1]

Jasen Taciak and Wesley Basel
U.S. Census Bureau, Washington, D.C. 20233

Abstract

Current production models of poverty utilized by the Census Bureau for the Small Area Income and Poverty Estimates (SAIPE) program generally incorporate only a single-year of inputs. These models produce parameter estimates by assuming some degree of homogeneity across areas. With six years of consistent inputs to the SAIPE model, including the American Community Survey, we find that additional precision is obtained by modifying assumptions of homogeneity in parameter estimates across years, and relaxing some cross-sectional assumptions. The impact of these alternative assumptions is then illustrated in terms of the change in precision to the estimate within population clusters and within the context of the current production SAIPE models.

Key Words: small area estimation, poverty, pooled time-series cross-sectional, SAIPE

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.

The purpose of this paper is to determine if the precision of the model-based estimate can be improved by pooling multiple years of the data, and to what extent alternate model restrictions affect this precision. This particular research application is part of a larger initiative within the US Census Bureau to investigate modeling techniques that may potentially increase the reliability of estimates. This initiative becomes increasingly relevant with ongoing federal budget constraints, where reductions to survey sample sizes are a substantive threat to the precision of sample data. The potential for diminishing precision to estimates is especially true of the smallest areas, which generally have lower sample size and higher variability. This is especially important for the SAIPE program, which has a mandate to produce single-year estimates of poverty for all school district and county domains throughout the US, to implement provisions of the No Child Left Behind Act of 2001. Hence, pooling multi-year data for modeling single-year estimates could potentially be a cost effective way of generating estimates with greater precision.

[1] 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.

While other approaches to modeling time-series and cross-sectional (TSCS) data exist (e.g. lagged dependent, stochastic regressors, etc.), the focus of this paper is limited to building off of previous SAIPE research investigating serial correlation in model error structures (Basel, et al., 2010).

The paper is organized as follows. First, we introduce the structure of the autoregressive models adapted for small area estimation along with alternate models with different restrictive assumptions to this structure for evaluation purposes, in Section 2. Second, we provide an overview of the estimation procedure and the empirical results in Section 3 and 4, respectively. Then, conclude with a discussion of results and potential areas for future research in Sections 5 and 6.

2. Structure of the Models

2.1 Current SAIPE Model

In general, the SAIPE program's county poverty model follows a Fay-Herriot, or shrinkage approach, by specifying both a sampling model (1.1) and a regression model (1.2) for the true value of log poverty (Fay & 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,

$$\log(y_i) = \log(Y_i) + \log(e_i) \quad (1.1)$$

$$\log(Y_i) = x'_i\beta + \log(u_i) \quad (1.2)$$

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 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, 2010), 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 lognormal 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 lognormal 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 equations (1.1) and (1.2). For the purposes of this paper, we focus only on specifying these components in the log-scale.

2.2 Proposed Pooled TSCS Model

Similar to the current SAIPE production model, to estimate parameters (1.1) and (1.2) are

combined by substituting the regression specification into the sampling model. For the sake of notational convenience, we define $z_i \equiv \log(y_i)$. With this, the multi-year[2] specification of this model is given in (2) below.

For $i,j = 1,2, \ldots, m$ areas, and t = 2005,2006, …, T years,

$$z_{it} = x'_{it}\beta_t + u_{it} + e_{it} \tag{2}$$

The sampling errors are assumed independent across both areas and time, while the model errors are assumed i.i.d. within a given year. Sampling errors are assumed independent from model errors, for any combination of areas (i, j) and years (t, s).

To complete the specification, assumptions regarding the year-to-year relation of the model errors must be made. We explore two nested AR(1) models with similar structure:

$$u_{it} = \rho u_{it-1} + \epsilon_{it}\ , \quad |\rho| < 1 \tag{3}$$

Where current disturbance, u_{it}, is composed of prior period disturbances, u_{it-1}, factored by the autocorrelation between model error disturbances of adjacent year, ρ, and a error term, ϵ_{it}. The differences between the two TSCS models are in the innovation assumptions. The first model assumes constant parameters across a period of time and areas. That is, for any county *i* and any period *t*,

$$\epsilon_{it} \sim N(0, \sigma_\epsilon^2) \tag{3.1}$$

The second process differs slightly, where the model error variance is unrestricted across years and represented by the following.

$$\epsilon_{it} \sim N(0, \sigma_{\epsilon t}^2) \tag{3.2}$$

Henceforth, equation (2) with restricted auto-regressive process (3.1) will be called R-AR(1), while U-AR(1) refers to equation (2) with the unrestricted nested error structure represented by (3.2).

The model error variance can be re-parameterized as a function of the innovation variance to make it more analogous to the single-year model error variance and helps to minimize notation. Thus,

$$\begin{pmatrix} u_{it} \\ u_{is} \end{pmatrix} \sim N\left(0, \begin{cases} \frac{\sigma_{\epsilon t}^2}{1-\rho^2} & , \ i = j \ and \ t = s \\ \frac{\rho^{|t-s|}\sigma_{\epsilon t}\sigma_{\epsilon s}}{1-\rho^2} & , \ i = j \ and \ t \neq s \\ 0 & , \ otherwise \end{cases}\right) \tag{3.3}$$

[2] Throughout this paper, terms 'multi-year' and 'time-series cross-sectional' (TSCS) are used interchangeably.

So notationally, we use σ_u^2 when assume constant model error variance across time and $\sigma_t^2 = \sigma_{\epsilon t}^2/1-\rho^2$ when allowed to vary. The full covariance structures for equation (2), with nested autoregressive processes illustrated by (3.1) and (3.2) are block-diagonal matrices. Using an arbitrary three-year sequence represented as periods *r*, *s*, and *t* of observations for a given county, *i*, the blocks of total variance of the ACS sample observations are given by

$$\begin{pmatrix} y_{ir} \\ y_{is} \\ y_{it} \end{pmatrix} \sim N\left\{ \begin{pmatrix} Y_{ir} \\ Y_{is} \\ Y_{it} \end{pmatrix}, \begin{pmatrix} \sigma_u^2 + v_{ir} & \sigma_u^2\rho & \sigma_u^2\rho^2 \\ \sigma_u^2\rho & \sigma_u^2 + v_{is} & \sigma_u^2\rho^2 \\ \sigma_u^2\rho^2 & \sigma_u^2\rho & \sigma_u^2 + v_{it} \end{pmatrix} \right\} \tag{4.1}$$

$$\begin{pmatrix} y_{ir} \\ y_{is} \\ y_{it} \end{pmatrix} \sim N\left\{ \begin{pmatrix} Y_{ir} \\ Y_{is} \\ Y_{it} \end{pmatrix}, \begin{pmatrix} \sigma_r^2 + v_{ir} & \sigma_r\sigma_s\rho & \sigma_{rt}\sigma_t\rho^2 \\ \sigma_s\sigma_r\rho & \sigma_s^2 + v_{is} & \sigma_s\sigma_t\rho \\ \sigma_t\sigma_r\rho^2 & \sigma_t\sigma_s\rho & \sigma_t^2 + v_{it} \end{pmatrix} \right\} \tag{4.2}$$

Where, v_{it}, is the ACS sampling variance for each area, *i*, at time *t*. As discussed earlier, these sampling variances are estimated directly from the sample and are then assumed known.

For this paper, the parameter estimates for this model are obtained by maximum likelihood estimation using a grid-search method over values of σ_u^2 and ρ. The MLE for β in the model is the generalized least squares (GLS) estimator given the variance parameters. The derivations for optimal predictor for this model and associated prediction error, given the data, were taken from notes written by William R. Bell (2012). This estimator reduces to the usual Fay-Herriot shrinkage estimator when $T = 1$.

If parameters $\psi = (\sigma_u^2, \rho)'$ and β are known, the minimum mean-squared predictor of Y_i is the expectation conditional on the data and parameters:

$$\ddot{Y}_i = E(\mathrm{Y_i}|\mathbf{y}, \psi, \beta) = E(\mathrm{Y_i}|\mathrm{y_i}, \psi, \beta) = \mathrm{X_i}\beta + \Sigma_u(\psi)\Sigma_{y_i}^{-1}(\mathrm{y_i} - \mathrm{X_i}\beta) \tag{5}$$

Y_i is the (T x 1) vector of unobservable true log poverty values for county *i* and all *T* years, **y** is a vector of all direct ACS log poverty estimates, which is replaced by the single-county vector, y_i, by using the assumption of independence across counties for the ACS estimates. The terms Σ_{e_i} and Σ_u within (5) are the diagonal matrix of sampling variances and the diagonal block matrix of model error variance, and can be represented as linear components of the block structures from (4.1) and (4.2):

$$\Sigma_{y_i} = \Sigma_u + \Sigma_{e_i} \tag{6}$$

In the case of a single year, (5) simplifies to the well-known Fay-Herriot shrinkage result. The mean-squared prediction error of this minimum mean-squared predictor is:

$$\mathrm{Var}\left(\mathrm{Y_i} - \ddot{Y}_i\right) = \left\{\Sigma_{e_i} - \Sigma_{e_i}\Sigma_{y_i}^{-1}\Sigma_{e_i}\right\} + \Sigma_{e_i}\Sigma_{y_i}^{-1}\mathrm{X_i}\mathrm{Var}(\beta)\mathrm{X_i}'\Sigma_{y_i}^{-1}\Sigma_{e_i} \tag{7}$$

In practice, we substitute the maximum-likelihood estimates of $\ddot{\psi} = (\hat{\sigma}_u^2, \hat{\rho})'$ and $\hat{\beta}$ into (5) and (7), providing the EBLUPs.

2.3 Alternate Models

In order to test the proposed AR(1) models, two additional models were constructed. The first being identical to that of R-AR(1), but with the serial correlation parameter restricted to zero (i.e. $\rho = 0$). Thus, the AR process incorporated in Σ_u would be a diagonal matrix, comprising $\sigma_u^2 I$. Where I is an (nT x nT) identity matrix, as opposed to the block-diagonal matrix illustrated in (4.1). Such a framework can be seen as a collection of "seemingly-unrelated regressions" (Zellner, 1962) where the p regressions and kp estimated coefficient parameters, are restricted to k parameter estimates.

The second alternate model, is a stacked single-year GLS regression model, with identical samples to both the R-AR(1), U-AR(1) and AR(0) models, but assumes uncorrelated model error across periods and permits regression coefficients to vary across years. This model would be the most comparable to the current SAIPE production model.

The motivation for these alternative specifications (henceforth referred to as AR(0) and SSY, respectively) being they measure the improvement that contemporaneous information provides to a single year model, and provide a means of testing inferences about the effectiveness of the R-AR(1) and U-AR(1) models.

3. Estimation Procedure

As mentioned above, GLS procedure is implemented to estimate β, given a specified σ_u^2 and ρ which maximize the following likelihood function,

$$\ln l\left(\hat{\beta}, \hat{\rho}, \hat{\sigma}_u^2 | y, X\right) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{nT} log\left|\hat{\Sigma}_{yi}\right| - \frac{1}{2}\sum_{i=1}^{nT}(y_i - {}' X_i'\hat{\beta})\hat{\Sigma}_{yi}^{-1}(y_i - X_i'\hat{\beta}) \quad (8)$$

Current SAIPE production estimates model variance, by implementing the first- and second-order partial derivatives of (6) with respect to σ_u^2 and the Newton-Raphson scoring algorithm:

$$\hat{\sigma}_u^{2\{i+1\}} = \hat{\sigma}_u^{2\{i\}} + [I\left(\hat{\sigma}_u^{2\{i\}}\right)]^{-1} S[\hat{\beta}\left(\hat{\sigma}_u^{2\{i\}}\right] \quad (9)$$

Using θ to represent all estimated parameters (β, σ_u^2, ρ) , then $S[\theta]$ is the score vector and $I[\theta]^{-1}$ the information matrix containing the first and second partial derivatives with respect to all θ.

Note however that given the block diagonal structure of Σ_y, deriving the partial derivatives for the scoring measure (9) would be onerous, especially when considering the U-AR(1) specification. For this reason, we employed a grid search to obtain ML estimators. This was done in two stages. In the first stage, an estimate of, β, was derived using GLS procedure,

$$\hat{\beta}^{\{i\}} = \left(\sum_{i=1}^{N} X_i'\left[\Sigma_{yi}^{\{i\}}\right]^{-1} X_i'\right)^{-1} \left(\sum_{i=1}^{N} X_i'\left[\Sigma_{yi}^{\{i\}}\right]^{-1} y_i\right)$$

Where $\Sigma_{yi}^{\{i\}}$, is of form (6) parameterized with $\hat{\rho}^{\{i\}}$

This process was iterated over a range of values for $\hat{\sigma}_u^2$ and $\hat{\rho}$. For σ_u^2, a search range of 0.0015 to 0.05 at intervals of .0001, was deemed sufficient since previous years of SAIPE production have shown model error to average 0.02+/- 0.007. At each iteration, a value of $\hat{\sigma}_u^{2(i)}$ and $\hat{\rho}^{(i)}$ (search range = $0.1<|\hat{\rho}|<0.9$) was chosen to obtain $\hat{\beta}^{(i)}$. Then, given $\hat{\psi}^{(i)} = (\sigma_u^{2(i)}, \hat{\rho}^{(i)})$ and $\hat{\beta}^{(i)}$ the sum of the likelihood function (8) was derived and evaluated against the previous likelihood of highest value.

To provide additional precision to the estimated parameters, in the second stage, this process was repeated with condensed search ranges and higher resolution, given the values of $\ddot{\psi}^{(i)}$ and $\hat{\beta}^{(i)}$ which resulted in the highest likelihood from the first search. For example, if the first stage returned MLE values $\hat{\sigma}_u^2 = 0.02$ and $\hat{\rho} = 0.7$, then the second stage would iterate over ranges of $0.0155 \leq \hat{\sigma}_u^2 \leq 0.0355$, $0.655 \leq \hat{\rho} \leq 0.855$.

It should also be noted that the span of the search grids were selected based on previous research, historic SAIPE estimation results, and computation time. The latter constraint was especially influential in modeling the U-AR(1) specification, where model error varied by time (i.e. $\hat{\sigma}_{ut}^2$), since the number of iterations grew exponentially given the number of parameters being estimated and desired precision. While the authors of this paper believe that these ranges span all realistic possibilities for values, we monitored for convergence to boundary values and found none occurred. Additionally, we used starting positions for searches outside the seemingly realistic span in order to provide more confidence that the estimates obtained were global maximums.

4. Empirical Results

4.1 Variable Definitions

Table 1 provides a variable list of model inputs. Not listed in the table is the sample error variances for the log of poverty for ages 5-17 related, estimated directly using the successive differences replication (SDR) method (US Census Bureau 2010).

Table 1: Variable definitions for the poverty rate model

Short Name	Description
Dependent Variable	
Log(ACS poverty for 5-17)	Log of county poverty estimate from the single-year ACS samples, 2005-2010, ages 5-17 in families.
Regressors	
Log(IRS child tax-poor exemptions)	Log of number of county tax-poor child exemptions from IRS administrative records, where tax-poor is defined as Adjusted Gross Income (AGI) below the poverty level for a household size defined by the total number of exemptions on the return.
Log(SNAP participation count)	Log of the number of county SNAP participants reported in July (data from the USDA Food and Nutrition Service), raked to a control total obtained from state SNAP participant data.
Log(PEP population, Age 0-17)	County population, ages 0 - 17, as of July 1, 2010, from the Census Bureau's Population Estimates Program (PEP) of post-censal demographic estimates.
Log(IRS Child Exemptions)	Log of Total IRS child exemptions.
Further information about these input data is available on the SAIPE program's webpage, http://www.census.gov/did/www/saipe/data/model/info/index.html.	

4.2 Data

Data for the R-AR(1), U-AR(1) and AR(0) models consist of SAIPE county data spanning 6 consecutive periods, from 2005 to 2010. As mentioned previously, the dependent variable in all models is the log of the direct single year ACS county estimates of the number of related children ages 5-17 in poverty. The ACS provides estimates for every county; however, in some counties with small samples, the estimate is zero. Since log of zero is undefined, these observations are not included in the model. Over this 6-year period, the ACS released approximately 18,840 county estimates, or 3,140 counties each year. Thus, taking into account county/county-coincident changes (e.g., county consolidations, dissolution, individual counties splitting into multiple entities, etc.) and non-zero estimates, there were 2,633 distinct, *i* county records for any period *t*, or 15,798 consistent counties records over the 6 periods.

4.3 Coefficient and Autocorrelation Estimates

Table 2 reports on the regression prediction results. The coefficients for all three models are significant, maintain a similar degree of magnitude, and identical direction as historical SAIPE results.

Also provided in the table is an estimate for $\hat{\rho}$, which, for both AR(1) specifications was 0.8, and is consistent with previous findings (Basel et al., 2010). Under the assumption that $\hat{\rho} \sim N(\rho, \frac{1-\rho^2}{nT})$, the test statistic $z = (\hat{\rho} - \rho)/\sqrt{(1-\rho^2)/nT} = \sqrt{nT}\hat{\rho} \geq 1.96$ (α=0.05) provides a criterion for the existence of autocorrelation ($H_0: \rho = 0$) (Judge et al., 1985, p. 394). Given the result $z = 112.5$, there is adequate support for the inclusion of this first-order autoregressive process in the model.

Table 2: Regression Results for Model Specifications 2005-2010 (T-statistic in parentheses)

Regressor		Specifications			
		SSY	**AR(0)**	**R-AR(1)**	**U-AR(1)**
Intercept	β_0	-0.61 (-10.34)	-0.58 (-23.20)	-0.55 (-17.45)	-0.53 (-16.94)
Log(Child Tax-Poor Exemptions)	β_1	0.88 (26.56)	0.83 (55.63)	0.82 (43.99)	0.83 (44.12)
Log(SNAP Participation Counts)	β_2	0.20 (9.31)	0.22 (25.60)	0.22 (20.49)	0.22 (20.22)
Log(Population Estimate, Age 0-17)	β_3	0.84 (5.81)	1.21 (24.46)	1.25 (23.10)	1.23 (22.64)
Log(Child Tax Exemptions)	β_4	-0.92 (-6.66)	-1.25 (-27.34)	-1.29 (-25.26)	-1.28 (-24.83)
Model error variance	$\hat{\sigma}_u^2$	0.0227	0.0256	0.0235	0.0243
MLE of Auto-Regression	$\hat{\rho}$, $(\sqrt{nT}\hat{\rho})$	0	0	0.8 (112.5)	0.8 (112.5)
Maximum log Likelihood		-9564.51	-9634.23	-9395.34	-9389.38
Median standard error of shrinkage estimate		0.139	0.146	0.130	0.127

The log-likelihood totals (Table 2) provide additional support for the inclusion of contemporaneous correlation. Using the criterion of the maximum likelihood principle, a comparison of the likelihood totals suggests that the R-AR(1) and U-AR(1) models performs best overall, with the unrestricted model having a slightly higher likelihood (note that the likelihood value of the 'Single-Year' model is the sum of a single year maximum log-likelihood totals from 2005 to 2010). Given the differences in model specification, this conclusion is more of a general interpretation than a statistically significant test.

However, as indicated in Section 3, evaluating the AR(0) model against the SSY model is

possible, since both models can be interpreted as TSCS models, with the AR(0) being the more restrictive TSCS model due to the constraints placed on the autocorrelation coefficients. In similar fashion the unrestricted and restricted AR(1) models can be assessed. Thus, a log-likelihood ratio test is appropriate:

$$\lambda = -2\left[\ln l\left(\hat{\beta}^R, \hat{\rho}^R, \hat{\sigma}_u^{2R} \middle| y, X\right) - \ln l\left(\hat{\beta}^U, \hat{\rho}^U, \hat{\sigma}_u^{2U} \middle| y, X\right)\right] \sim \chi^2_{(J)}.$$

Where superscript 'R' and 'U' within the log-likelihood functions indicate the restricted and unrestricted models, and value λ, is distributed χ^2 with J degrees of freedom, equal to the difference in estimated parameters between the unrestricted and restricted model. Table 3 provides the results, which, in all three cases, lead us to reject the null hypotheses that the two specifications are equal.

Table 3: Log-Likelihood Ratio Tests

Specifications	Sum of Log-likelihood		Degrees of Freedom	Likelihood Ratio λ	Critical Values	
					(α=0.05)	(α=0.10)
R-AR(1) : AR(0)	-9395.34	-9634.23	1	477.78	3.84	2.71
U-AR(1) : R-AR(1)	-9389.38	-9395.34	5	11.924	11.07	9.236
SSY : AR(0)	-9564.51	-9634.23	30	139.44	43.773	40.256

4.4 Analysis of Shrinkage Estimates and Precision

Previous tests have only suggested that the inclusion of past information is valid, but we have not been able to distinguish whether this addition data is beneficial. Again, due to the variation of model specification, no commensurate statistical test is adequate to evaluate the overall performance of all four specifications. However, since the purpose of the SAIPE program is to model small domains—areas that contain high sampling variability and thus generally contain high error—evaluations based on standard errors and variation by population size are most appropriate. Additionally, the underlying question this investigation seeks to answer is whether past information, within model disturbances, can be exploited to increase the accuracy of a single-year estimate, thus Table 4 reports the median standard errors and point estimates, and variability encompassing period 2010 only.

Table 4: Median Shrinkage Standard Error and Coefficient of Variation (CV) for 2010, by County Resident Population[3]

Pop	Cnty	SSY			AR(0)			R-AR(1)			U-AR(1)		
Size	N	Est	SE	CV	Est	SE	CV	Est	SE	CV	Est	SE	CV
All	2633	7.13	(0.139)	1.95%	7.14	(0.146)	2.05%	7.14	(0.130)	1.82%	7.14	(0.127)	1.77%
<10k	320	5.51	(0.149)	2.70%	5.54	(0.157)	2.83%	5.54	(0.145)	2.62%	5.54	(0.141)	2.54%
10-20k	528	6.40	(0.146)	2.28%	6.42	(0.154)	2.39%	6.43	(0.141)	2.19%	6.43	(0.136)	2.12%
20-65k	993	7.16	(0.139)	1.95%	7.18	(0.146)	2.04%	7.17	(0.130)	1.81%	7.16	(0.126)	1.76%
65-250k	537	8.22	(0.126)	1.53%	8.23	(0.131)	1.59%	8.20	(0.112)	1.37%	8.20	(0.110)	1.34%
>250k	255	9.56	(0.094)	0.98%	9.55	(0.095)	1.00%	9.56	(0.084)	0.87%	9.56	(0.082)	0.86%

[3] Comparisons between model specifications based on standard errors / CVs are for illustration purposes and may not be statistically significant.

Considering the 'All' population category (first row), the differences between shrinkage estimates and associated errors produced by the various specifications are minimal, but on the basis of overall error dispersion (CV) the AR(1) specifications outperform others. Comparing models over all sub-population categories, we see that the U-AR(1) specification exhibits the lowest variation in the smallest population categories (counties with populations less than 10,000 and between 10,000-20,000) and larger counties (65,000-250,000). The AR(0) specification generated the highest SEs and CVs of all models. Approximately 1% higher errors when compared with the single-year model, and 2%-3% more than the U-AR(1) specification. Further discussion on these results provided in Section 5.

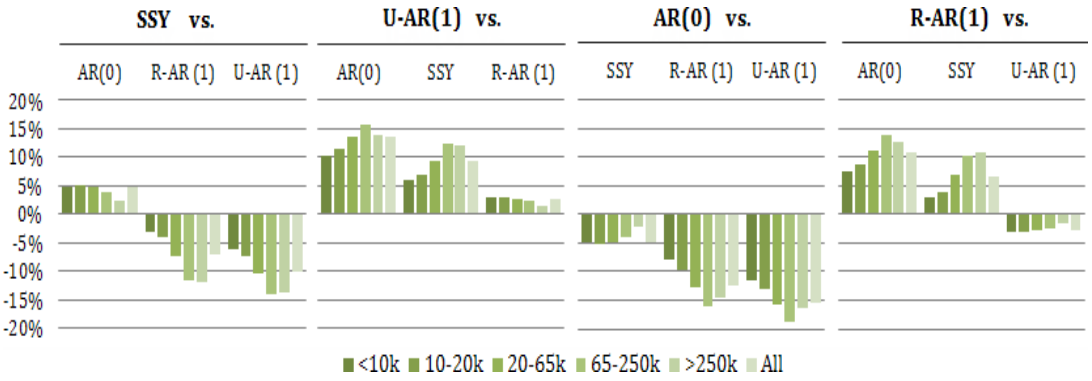

Chart 1: Estimated Percent Change in Precision (CP), by County Resident Population[4]

Using the median SEs of each specification (Table 4), we construct a measure to determine the precision gained/lost by implementing one specification over another:

Change in Precision $(CP) = 1 - (Median\ CV\ of\ Model\ A / Median\ CV\ of\ Model\ B)$

The results, depicted in Chart 1, are composed of four main sub-charts where each depicts the performance of the specification (bolded model at the top of graph) against each of the remaining three (un-bolded models below). For example, the first six bars on the left-hand-side represent the extent to which the estimate's precision changes, over six population clusters, when comparing SSY specification, against AR(0). In this case, there is a gain in precision for all clusters. Thus, lower CVs result in an efficiency gain (positive percentage) while higher CVs result in a loss (negative)).

As expected, Table 4 illustrates the largest gains in precision occur with the AR(1) specifications. In fact, over all counties, the U-AR(1) increases precision by 9% and 13% when compared to the single-year and AR(0) specifications. The R-AR(1) performs equally well, with gains of 7% and 11% when compared to the same alternate specifications. This positive trend continues throughout all clusters including the

[4] Comparisons between model specifications based on standard errors / CVs are for illustration purposes and may not be statistically significant.

smallest—although to a lesser extent: For areas with populations less than 10,000, U-AR(1) generated gains of 6% and 11% and R-AR(1) gains of 3% and 8% against the respective alternate models.

5. Discussion of Results

The inference tests, described in Section 4.3, indicate a difference between the single-year and AR(0) model, as well as support of the existence of a serial-correlated error structure in the data. Although no statistical test was derived, the comparisons of errors and precision (Section 4.4) also highlights the effectiveness of including contemporaneous information in the county poverty model. As stated previously, AR(0) consistently performed worst amongst all other specification. Chart 1 illustrates that replacing the SSY model (which has the most similarity to the current SAIPE model) with AR(0) results in decreased precision of 5% overall, and 5% in small areas.

Thus, in terms of reducing total error, given the Fay-Harriet weighting structure from (7), the mere inclusion of past information does not add value, whereas a reduction in error is generated by both AR(1) models. This suggest that incorporating past data also incorporates the associated variance; however, without the proper auto-regression decay across time, the total error does not include the covariance terms of the block diagonal structure of $\boldsymbol{\Sigma}_{\mathbf{u}}$, which provide an added dampening effect to the current period's error and increase overall precision.

Continuing with this reasoning, one might conclude that there may be a point where the inclusion of past information adds more error than the auto-regressive structure dampens. However, as noted above, the two AR(1) models had the most gains in estimated efficiency, including over the SSY model at counties with lowest populations (less than 10,000). This result is insightful, since smaller domains have higher volatility (Table 4). Although a 3% gain is comparatively small, it does suggest that even in small areas, the benefits of including the contemporaneous information outweigh the impact of the higher volatility.

6. Future Research

This research was the first step in exploring the benefits incorporating time-series and cross-sectional in a county-level poverty model, and was limited to the investigation of correlated errors. Other well-known approaches to TSCS modeling could prove advantageous as well (distributed lag models, stochastic regressors, etc.). Additionally, other model variations of the specification could be considered to evaluate these results: AR(p) models; Toeplitz covariance model; random coefficient models; completely unrestricted regression coefficient model.

This analysis could also be expanded to include other data sources correlated with poverty (e.g. regional employment characteristics (Basel & Albert, 2012)).

Moreover, this paper explored the impact of alternative assumptions regarding homogeneous parameters across time, however relaxing these assumptions further—perhaps across time and demographic, or employment characteristics—might also be insightful.

Additionally, this paper's conclusions lack direct comparability to current production SAIPE models. As was mentioned earlier, the survey estimates of zeros are not included in the log model. This is problematic for smaller areas due to higher sampling variability, and especially when modeling pooled data since counties must be present across all periods. Hence, it would be beneficial to explore alternative modeling techniques easily adaptable to the presence of zero estimates, such as with a generalized linear models with linking functions (Wieczorek et. al., 2012) and censored models (Nugent, et. al 2012), within a time-series cross-sectional framework.

References

Wesley B., S. Hawala and D. Powers. 2010. Serial Comparisons in Small Domain Models: A Residual-Based Approach, U.S. Census Bureau. <http://www.census.gov/did/www/saipe/publications/files/BaselHawalaPowers2010asa.pdf>

Wesley B., N. Albert. 2012. Use of Labor Market Indicators in Small Area Poverty Models In *JSM Proceedings. Forthcoming*

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

Bell, W. 2012. Notes on a Multivariate Fay-Herriot Model with AR(1) Model Errors. Unpublished manuscript.

Fay, Robert E., III, and Roger A. Herriot. Jun., 1979. 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, 269-277.

Judge, G., W. Griffiths, C. Hill, and T. Lee. 1985. *The Theory and Practice of Econometrics*. New York: John Wiley and Sons,

Nugent, C., S. Hawal, 2012. Research and Development for Methods of Estimating Poverty for School-Age Children. In *JSM Proceedings. Forthcoming*

U.S. Census Bureau. 2010. Chapter 12: Variance Estimation. *ACS Design and Methodology*. <http://www.census.gov/acs/www/methodology/methodology_main>

Wieczorek, J., C. Nugent, S. Hawal, 2012. A Bayesian Zero-One Inflated Beta Model for Small Area Shrinkage Estimation. In *JSM Proceedings. Forthcoming*

Zellner, Arnold 1962. An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests of Aggregation Bias, Journal of American Statistical Association. 57: 348-368