A Practical Guide to Small Area Estimation, Illustrated Using the Ohio Medicaid Assessment Survey

Rachel Harter[1], Akhil Vaish[1], Amang Sukasih[2], Jeniffer Iriondo-Perez[1], Kasey Jones[1], Bo Lu[3]

[1]RTI International, 3040 Cornwallis Road, Research Triangle Park, NC 27709

[2]RTI International, 701 13th St NW #750, Washington, DC 20005

[3]The Ohio State University, Cockins Hall, 1958 Neil Ave., Columbus OH, 43210

Abstract

Much literature has been written about the theory and statistical properties of small area estimators, but very little has been written about the practical aspects of producing small area estimates. This paper summarizes the basic steps for producing small area estimates. The steps involve identifying requirements such as dependent variables of interest and small areas or domains of interest; identifying and compiling auxiliary data and selecting significant predictors; determining an appropriate model, estimation method, and software for running the model; and producing, validating, and reporting the estimates. The steps are illustrated by the production of estimates of the proportion of adults without health insurance coverage, by county, using data from the Ohio Medicaid Assessment Survey.

Key Words: small area estimation, OMAS, hierarchical Bayes, auxiliary data, uninsured rates, OpenBugs, MCMC

1. Introduction

Small area estimation (SAE) refers to a set of tools for making estimates for domains or subpopulations smaller (in terms of sample size) than those for which a survey was designed. The unplanned domains or subpopulations often have too few sample observations for producing direct design-based estimates[1] with sufficient reliability. SAE combines the survey data for the outcome variable of interest in an area or subpopulation with correlated auxiliary data and assumptions regarding the relationship between the survey and auxiliary data. The estimates are said to "borrow strength" from the auxiliary data. Rao and Molina (2015) discuss small area estimation in greater breadth and depth.

Most discussions of SAE focus on the model relationships between the survey and auxiliary data and the structure of the resulting estimator. The literature rarely discusses the practical aspects of the entire process of SAE, as summarized in Figure 1. Successful implementation requires care and attention to quality at every step in the process. This paper describes the general steps involved in the SAE process as a guide for practitioners, who must customize the steps for their own applications.

Subsequent sections describe the steps summarized in Figure 1. Along with descriptions and explanations, we include illustrative details for one specific application for the Ohio Medicaid Assessment Survey (OMAS). OMAS (https://grc.osu.edu/OMAS) is a random digit dial telephone survey of non-institutional residential Ohioans. Sponsored by the Ohio

[1] Direct survey estimates such as the Horvitz-Thompson estimator involve only the observed survey data and their analytical weights.

Department of Medicaid and the Ohio Colleges of Medicine Government Resource Center, OMAS examines issues related to health status, health insurance status, and health care access and utilization. Each step in the SAE process is illustrated with brief summary information for the application of SAE in estimating county-level rates of adults without health care insurance (Sukasih et al. 2019).

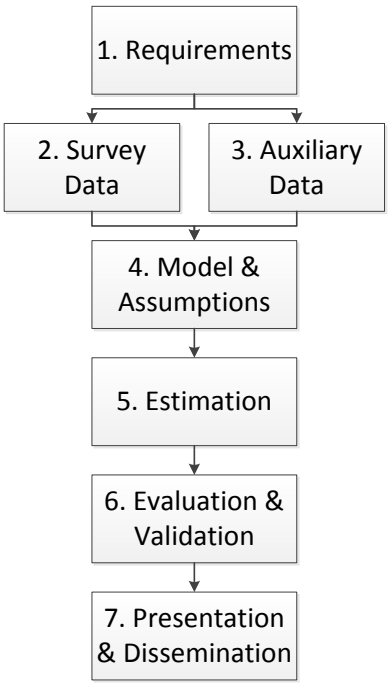

Figure 1: Steps in a small area estimation application

2. Step 1: Determine requirements

To avoid misunderstandings, the practitioner must be clear about the estimation requirements. Determining the requirements may mean iterative conversations with the client or stakeholders. It is good practice to document the requirements with the approval of all relevant parties.

The first requirement is the survey outcome variables for which SAE is desired. That is, determine the statistics to be reported. Most often the outcome variables are continuous or binary, and the statistics to be reported are means or proportions. Categorical variables are more challenging; categorical variables may be collapsed into binary form for estimation. Different outcome variables may require different auxiliary variables and different model assumptions, so it is rarely the case that many outcome variables can be estimated as easily as one.

The second requirement is the set of domains for which the estimates will be reported. For examples of geographic domains, a national survey may require SAE for states because not all states have enough sample for state-level estimation, or a state survey may require SAE for counties because not all counties within the state have enough sample for county-

level estimation. Alternatively, the domains may be defined by subpopulations. A survey of adults, for example, may require SAE for specific age groups or racial/ethnic groups. It is possible that the requirements include sub-geographies and not subpopulations, subpopulations but not sub-geographies, or the cross classification of sub-geographies and subpopulations. For the remainder of this paper, we will refer to the sub-geographies and sub-populations for SAE as domains or areas. Domain identifiers must be available with the survey data, or other identifiers must be available for survey respondents to aggregate unambiguously to the domain level.

Sometimes the estimates have a precision requirement that cannot be met with direct survey estimates. Combining the survey data with auxiliary data and model assumptions can improve the precision of the estimates, if the assumptions reasonably fit the data. Many small area estimates are biased, but the hope is that the mean squared error (MSE) indicates an improvement over the unbiased direct survey estimate. Statisticians need a method of estimating the MSEs for small area estimates.

Another requirement may be to keep the costs of the work within the available budget. Yet another requirement may be the date the estimates are due. Requirements might include the software to be used, the level of security of the data, or any other consideration important to the stakeholders. All such requirements should be clearly understood and documented. Figure 2 summarizes some of the requirements for the OMAS.

- Primary objective: to estimate proportion of adults without health insurance in each county.
- Primary outcome variable: whether person has health insurance (binary variable).
- Geographical Areas: 88 Ohio counties
- Subpopulation: Adults age 19+
- Time period: 2017
- If n<100 (or other conditions not specified here) are met, use SAE; otherwise use the direct Horvitz-Thompson estimator.
- Use off-the-shelf software for SAE.
- Deliver a file of estimates by the end of the calendar year.
- Deliver a report by the end of the contract.
- Stay within budget.

Figure 2: Summary of requirements for OMAS (abbreviated for illustration)

3. Step 2: Survey Data

Often the survey data are provided by the client. Sometimes a public use file is available, but a public use file may or may not have all the details needed. SAE will require indicators for the small domains, various geographical (e.g. state, county, tract) identifiers, final analytical weights, and probably survey design variables such as stratum and PSU (primary sampling unit) indicators. Security requirements may accompany a restricted use file. These details should be worked out at the requirements stage and checked when the data are obtained.

Some data files are easier to work with than others. For example, some files may have household records as well as person records. Some have data about all household

members on the same record, and some have separate records for each person. It goes without saying that the file structure must be known to read the data properly.

Ideally, a codebook will accompany the data file. When exploring the file, it is standard quality practice to verify the correct number of records, missing observations, the necessary variables, and the distributions of values for the variables. Before proceeding with any analyses, detailed descriptive statistics should be produced for all the variables of interest to make sure the data are correct and the observed results are consistent with the expected results. The sample design variables and analytical weights should have no missing values, which could happen by incorrectly reading the data. If the small domain indicators or the outcome variables of interest have missing values, imputation may be required before proceeding. Because SAE methods are generally computationally expensive, an extract file with one record per analytical unit and only the complete, necessary variables will use far less space and process more quickly than the full survey file with hundreds of variables. Figure 3 summarizes the information extracted from the OMAS data file for county-level estimation.

A single year of OMAS restricted data contained respondent-level records. The extract from the survey data file contained:

- Final analysis weight
- Sample design variables: Stratum, and PSU
- County identifier
- Edited outcome variables (rare missing values were imputed)

Figure 3: OMAS survey data for SAE

It is a good idea to summarize the number of observations per small domain. Some domains may have no sample whatsoever. In some applications, multiple cycles over the years of the survey are pooled together to increase the sample sizes per domain and improve coverage across areas. The tradeoff is that the small area estimates will be the averages across the pooled survey years. Domains with no sample will have SAE predictions that are purely model-based, where the model is estimated using the observed survey data in other domains. The assumption is that the domains with no sample data have the same model relationships as the observed domains. Without sample data, the prediction of that domain may be more biased, and the MSE of the prediction may be underestimated.

If possible, generate the direct survey estimates and standard errors for the small domains for later comparison with the SAE predictions. If SAE modeling is done at the domain level rather than the sample unit level (step 4), then the direct survey estimates are required for modeling.

4. Step 3: Auxiliary data

The key to successful SAE is finding good auxiliary variables. Almost any SAE method will work well if the auxiliary variables are closely related with the outcome variable of interest. For example, the Illinois Department of Employment Security wanted to produce monthly estimates of employment by county using survey data collected for the Bureau of Labor Statistic's Current Employment Statistics program. The ideal auxiliary variable was

a statewide quarterly census of employment (the Bureau of Labor Statistic's Quarterly Census of Employment and Wages) that lagged behind the survey data by several months. The same measures were collected from the same target population by the same organization; the census was available for the entire population, and the difference in time periods was relatively minor. A very simple model was possible with this auxiliary variable (Harter et al. 2003).

Time spent brainstorming possible auxiliary variables in collaboration with substantive experts is well worth the effort. Depending on the nature of the outcome variable and domains of interest, auxiliary may come from:

- Administrative records from federal, state, or local agencies
- Data from large surveys such as the American Community Survey (ACS) or censuses from government sponsors
- Commercial businesses that provide data (details about the data provenance and quality may not be readily available)

It is a general assumption that auxiliary variables are fixed and have no errors; otherwise, estimated MSEs will not be correct (Ybarra and Lohr 2008).

Good auxiliary variables should have the following characteristics:

- Be correlated with the outcome of interest
- Be available for all the areas in the population, either individually or in aggregate
- Be of high quality, with very little error, if any
- Have identifiers that enable linkage to the small domains of interest
- Be affordable
- Be available in the timeframe needed for the work
- Be defined consistently over time in case of periodic production of small area estimates

If the auxiliary variable is not correlated with the outcome of interest, it will likely be dropped during the model selection process. Calculating correlations or performing tests of association between outcome variable and auxiliary variables is a quick way to see which auxiliary variables are most promising.

If the objective is to produce small area estimates for all areas in the population, then the auxiliary variables must be available for all the areas in the population. If the objective is to improve the MSEs of small areas estimates of the areas observed in the sample, then auxiliary variables must be available for the observed sample areas.

The auxiliary variable should not have missing values. A small number of missing values can be imputed, but more extensive imputation will cause problems for SAE. Furthermore, the auxiliary data should have very little error. Variables from the ACS are often used as auxiliary data because the standard errors of the ACS estimates are substantially smaller than those of other surveys. Most SAE methods do not capture the variability in the auxiliary variables, and ignoring very small errors in auxiliary variables is common practice. However, the MSEs from SAE are often underestimated, especially if the auxiliary variables' errors are not trivial.

Auxiliary variables are rarely contemporaneous with the survey data. For outcome variables that do not change much over time, the time difference between data sources is not much of a concern. For more volatile outcomes, having nearly contemporaneous auxiliary data is critical.

To relate the auxiliary data to the outcome of interest, a linkage mechanism is required. For example, if the auxiliary variable is available at county level, the county identifier must be "matchable" to the survey data. Sometimes the link is not directly matchable, but the data may still be useful if a match can be constructed through careful mapping and collapsing. Auxiliary data by racial/ethnic group, for example, must map cleanly to racial/ethnic groups in the survey data, collapsing categories as necessary.

Federal data are usually available for free, other than the labor to download datafiles from the internet. Data from commercial sources come at a cost, which may or may not fit within the project's budget. Restricted data may require data use agreements and other security measures, which take time and planning. Do not underestimate the time it takes to obtain restricted data.

The auxiliary variables, when obtained, should be examined much the same way the survey data were examined. Checking the data structure, variable values, data completeness, data dependence, and quality is paramount to producing good quality estimates.

- X_1 = total adult population (ACS 5-year, 2012–2016)
- X_2 = proportion of Hispanic (ACS 5-year, 2012–2016)
- X_3 = proportion of American Indian and Alaskan Native (ACS 5-year, 2012–2016)
- X_4 = proportion of household with person age 65 and over (ACS 5-year, 2012–2016)
- X_5 = proportion of non-citizen (ACS 5-year, 2012–2016)
- X_6 = proportion of housing units that are resident-owned (ACS 5-year, 2012–2016)
- X_7 = proportion of adults with less than high school (ACS 5-year, 2012–2016)
- X_8 = proportion of male (ACS 5-year, 2012–2016)
- X_9 = proportion of adults employed by non-retail firms (County Business Patterns)
- X_{10} = proportion of housing units that are rural (2010 Census)
- X_{11} = average unemployment rate (Bureau of Labor Statistics)
- X_{12} = median adjusted gross income or median household income (SAIPE covariate)
- X_{13} = percent in poverty, all age (SAIPE covariate)
- X_{14} = total Supplemental Nutrition Assistance Program benefit recipients (SAIPE covariate)
- X_{15} = per capita income (Bureau of Economic Analysis)

Figure 4: Potential county-level auxiliary variables for OMAS SAE

Figure 4 lists the auxiliary variables that were initially considered and examined for the OMAS SAE.

The goal of SAE models is the predictions for the small domains, not the parsimoniousness of the model. However, too many auxiliary variables may cause the model estimation to fail to converge, and the accumulation of errors in the auxiliary variables add uncertainty to the small domain predictions. Therefore, when many auxiliary variables are in consideration, some method of reducing the set of variables is required. In practice, the method is often a matter of person choice—stepwise regressions, decision trees, principal components, residual analysis, etc. Prioritize the remaining variables, in case further reductions are needed when fitting the models.

Figure 5 shows the reduced set of variables for the OMAS model. The variables were selected by first centering the variables on their respective means. A standard linear regression model was estimated in R software using the ***lm*** function. The function ***stepAIC*** in R was used to select the predictors of the transformed adult uninsured rates.

- per capita income
- percent of population in poverty
- proportion of housing units that are resident-owned
- proportion of adults with less than high school education

Figure 5: Reduced set of auxiliary variables selected for OMAS SAE of adult uninsured rates

5. Step 4: Model & Assumptions

SAE-based estimators are based on an assumed relationship between the outcome variable of interest and the auxiliary data. The assumed relationship provides the foundation for the structure of the estimation model. Even methods such as synthetic estimation, which is not explicitly model-based, nevertheless can be expressed in model terms. With explicit models, the survey outcome variable (or a transformation of it) is the dependent variable, and the auxiliary variables are the covariates. It is also possible to set up multivariate models for multiple survey outcome variables simultaneously.

Models may be fit at the individual reporting unit level (unit-level model) or at the small domain level of aggregation (area-level model). Under the latter approach, the dependent variable may be the direct survey estimate for the small domains. In addition to auxiliary data, models may contain random error components corresponding to the small domains and possibly higher aggregations, and model error. In recent years, many of the published models are hierarchical in structure, and the parameters are estimated by Bayesian methods. Much of the SAE literature focuses on the models and the forms of the estimators. Rao and Molina (2015) provide an overview of many models. Covering all possible models

and corresponding assumptions is beyond the scope of this paper, but thought must be given to models that are reasonable (structurally and computationally) given the survey and auxiliary data.

A modified Fay-Herriot (1979) hierarchical Bayes model, which contains two parts, was selected for OMAS. The sampling model relates the survey outcome of interest, which is the uninsured rate (proportion), and the true proportion. This sampling model includes a sampling error component for direct survey estimation. The second model is a linking model between the parameter of interest and the auxiliary variables. This model also includes a random error component. When the parameter of interest is a proportion, a common linking model is a logistic regression that regresses the logit of the dependent variable on the independent auxiliary variables as predictors. In this case, the OMAS weighted sample proportions were transformed using the logit transformation prior to fitting the Fay-Herriot model.

Off-the-shelf software is readily available for the Fay-Herriot model, and it is a standard approach for many applications. A main assumption with this model in this context is that the model relationship holds throughout the state, other than county-level random error components.

For Bayesian estimation of model parameters, prior distributions for model parameters need to be specified. Options for priors include information from previous empirical data/surveys or simply non-informative priors. Because no existing data were available for priors for OMAS, we chose to use flat/non-informative priors. The model and assumptions selected for OMAS are given in Figure 6.

- Transformed Fay-Herriot model for county i:
  - (p_i,π_i) is the sample and population proportions
  - Use logit transformation $y_i = ln\left(\frac{p_i}{1-p_i}\right)$ and $\theta_i = ln\left(\frac{\pi_i}{1-\pi_i}\right)$
  - Sampling model: $y_i = \theta_i + e_i$
  - Linking model: $\theta_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + v_i$
    where v_i is a county-level random effect, i=1,…,88

- Assumptions and priors:

$$\text{Sampling error } e_i \sim N(0, \varphi_i)$$

$$\text{where } \varphi_i \text{ is empirically estimated by } var[ln\left(\frac{p_i}{1-p_i}\right)] = \frac{var(p_i)}{[p_i(1-p_i)]^2}$$

$$\text{Model error } v_i \sim N(0, \sigma_v^2)$$

$$\sigma_v^2 \sim U(0,100)$$

$$\boldsymbol{\beta} \sim N\big(0, I(1000)\big)$$

Figure 6: Model and assumptions for OMAS estimation of the proportion of uninsured adults by county

6. Step 5: Estimation

For model-based SAE, the next step is to estimate the model parameters. The frequentist approach typically uses generalized linear models, but first the variance components must be estimated. The Bayesian approach usually uses Gibbs sampling/Markov Chain Monte Carlo (MCMC) samples, and the estimated parameters are the means of the posterior distributions. For more details on these estimation methods, see Rao and Molina (2015) and references cited there.

Estimated models should always be checked. If Bayesian methods are used, check that the MCMC process has converged using a visual approach or some test statistics for checking the convergence (Brooks & Gelman 1997; Gelman and Rubin 1992). If the model did not converge, the SAE model may need to be modified; for example, the set of auxiliary variables may have to be reduced or the number of areas may have to be reduced by pooling neighboring areas. Regardless of the estimation method, check that the model assumptions are not violated using tests such as residual analysis. For more information about model diagnostics, see Rao and Molina (2015).

Implicit SAE modeling requires similar care and review of the estimation. For example, an unexpected estimate may occur if the formula does not allow some values in the survey data (e.g., a very small value or zero proportion).

The small domain estimates are the predictions from the estimated model using the small domains' auxiliary variable values. If the survey outcome variable was transformed prior to modeling, the predictions must be transformed back.

With the explicit model-based SAE computed using MCMC replicate samples, the MSE of the estimates can be calculated as the variation computed across MCMC point estimates. The credible interval (analog to the confidence interval in frequentist calculations) can be calculated using percentile values of the MCMC posterior distribution. For some frequentist methods, MSEs are specified in closed form, but resampling methods can be used to estimate MSEs otherwise (Rao & Molina 2015).

7. Step 6: Evaluation & Validation

After the fitted model has been thoroughly checked, it is wise to evaluate the resulting small domain estimates and validate them against other sources of data where possible. For starters, the small domain estimates should closely align with the direct survey estimates in domains with larger sample sizes. The root MSEs of the small domain estimates should be better than those of the direct survey estimates, or else there is no point to doing SAE.

Sometimes a separate survey or census provides estimates that can be used for comparison to the small area estimates. For example, the American Community Survey generates estimates of many outcome variables which can be useful for validating SAE. Most likely an outside source such as ACS will not exactly match the time period(s), the geographies, or the subpopulations, but nevertheless an outside source can provide a check of reasonableness.

Figure 7 displays the results of county SAE for OMAS with both the direct survey estimates and the corresponding ACS estimates of adult uninsured rates. Notice that the ACS estimates tend to be larger, on average; the ACS defines adults as age 18+, whereas the OMAS defines adults as 19+. The direct and model-based estimates are in a similar range, and the model-based estimates tend to have a narrower spread.

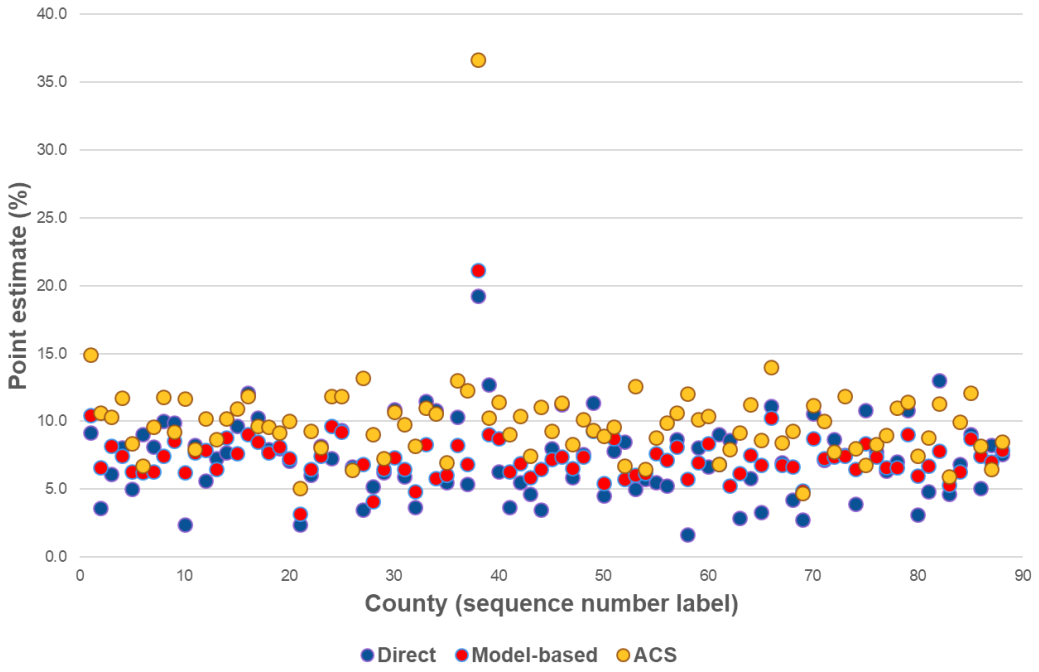

Figure 7: Comparison of estimates of adult uninsured rates obtained using direct design-based estimation, model-based estimation, and the ACS

Figure 7 displays a significant outlier in each set of estimates. Anomalies of this nature should be thoroughly investigated. In this case, Holmes County contains a large population of Old Order Amish, many of whom have no health insurance. The model used for SAE may not be appropriate for Holmes County, but both the direct OMAS estimates and ACS estimates also show Holmes County to be different from the rest of the state. The difference between the direct OMAS estimate and the ACS estimate may be related to the fact that OMAS was strictly a telephone survey. When such outliers exist, discarding them and refitting the model may improve the fit of the model. Sinha and Rao (2009) presented a robust modification of estimated best linear unbiased predictor (EBLUP) SAE methodology in the presence of outliers.

Figure 8 displays relative precision of the direct survey estimates and the model-based estimates for OMAS. The small area estimates have considerably smaller variability than the direct survey estimates, especially in counties with smaller sample sizes.

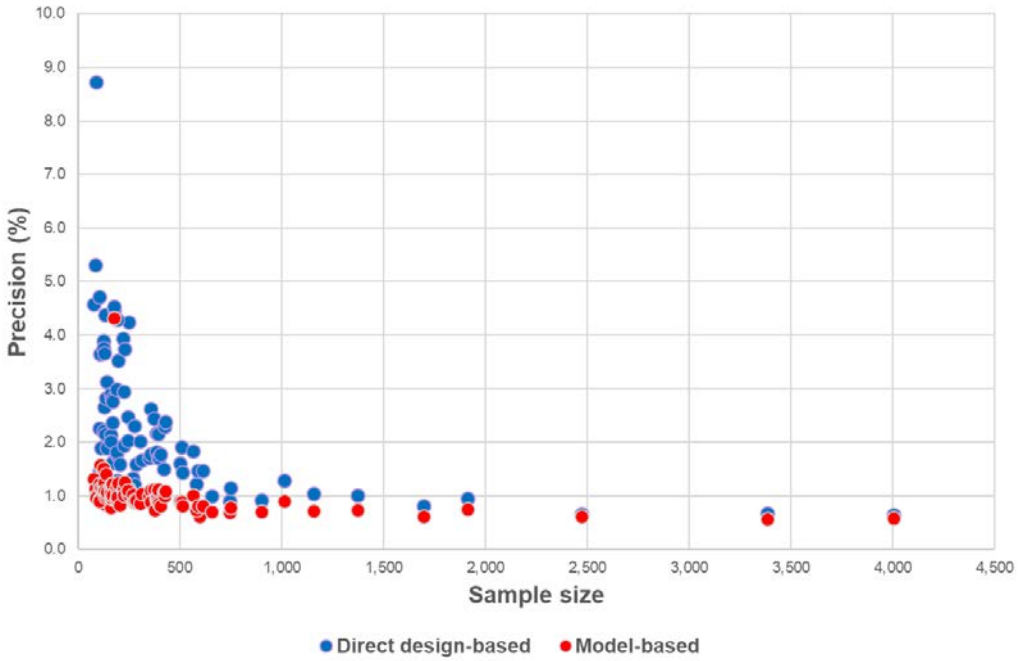

Figure 8: Sample size vs. precision of the adult uninsured rates obtained using direct design-based estimation and model-based estimation

In some applications, small domain estimates are benchmarked to be consistent with estimates at higher levels (e.g., benchmark county estimates to be consistent with direct state estimates) or with an alternative data source thought to be more precise at higher levels. Benchmark totals are not always available, however, and benchmarking may not be required or advisable, depending on the circumstances.

8. Step 7: Presentation & Dissemination

The completed, checked, and validated estimates are delivered to the customer in the form(s) requested by the client. Common delivery products include Excel spreadsheets, SAS® files, printed reports, graphical displays, and presentations, although many forms of electronic files are possible. Each client's request is driven by that agency's plans for dissemination and available tools. In most cases, the client owns the estimates and is responsible for the public presentation of the results. Sometimes the client will permit the statistician to present either the methodology or the results in a public forum, but typically in such cases the client reserves the right to review and approve the material. The deliverables for the OMAS project are summarized in Figure 9.

- Excel spreadsheets with county rows containing direct survey estimates with standard errors and confidence intervals (CIs), the small domain estimates with root MSEs and credible intervals (CIs), and quantities for intermediate calculations.
- Interactive maps of Ohio counties with percent uninsured adults colored in quintiles, where the user could hover on individual counties for specific values and CIs. See Figure 10.
- Detailed report with statistical explanations for methods used and summary tables of estimates (Sukasih et al. 2019).

Figure 9: Deliverables for OMAS

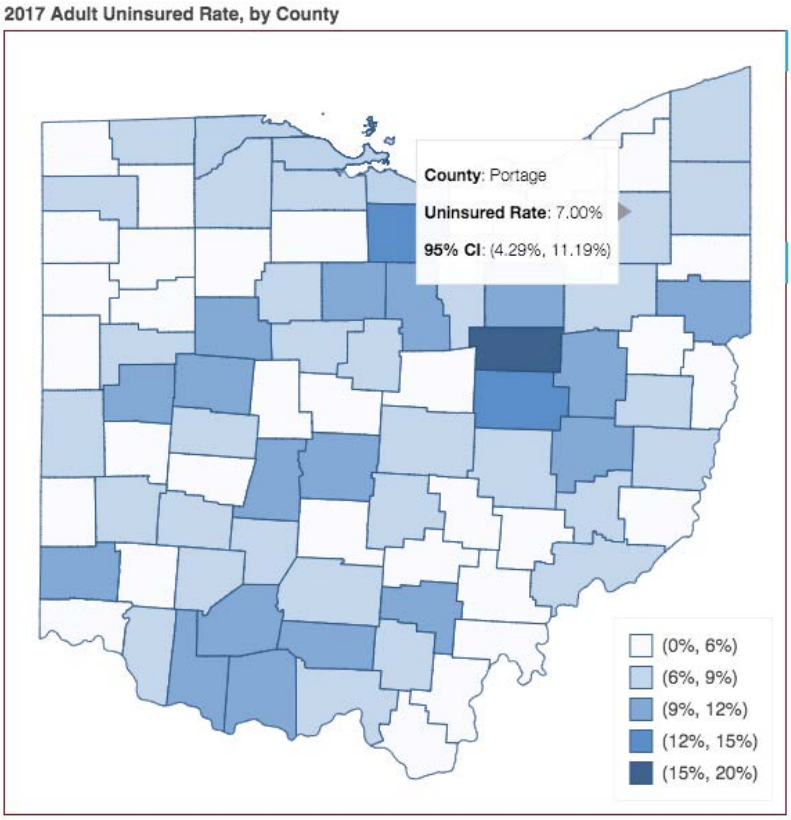

Figure 10: Static example of interactive map of Ohio counties

Some clients will request copies of the programs used to generate the estimates. Others will request that all components of the estimates be delivered along with the final estimates. Many clients use these materials to verify the quality of the work, to learn to produce the estimates themselves, or to archive the materials for a future contract. It is wise to maintain clean and well-documented programs and supporting materials (e.g., *.log* and *.lst* files for SAS programs) for internal review as well as for possible delivery to the client.

When preparing reports, presentations, and other dissemination materials, intended audience and the purpose of the materials should be considered. Design materials to be clear and easy to comprehend, which will be appreciated by technical and nontechnical

audiences alike. Avoid unnecessary technical jargon, but do not patronize. Statistical authors need to understand their methods well in order to write about them clearly. Technical detail may be appropriate in the main text or in an appendix, depending on the audience.

Acknowledgements

The SAE examples using OMAS data were completed under contract with the Ohio Department of Medicaid and the Ohio Colleges of Medicine Government Resource Center. This paper was completed with partial support from RTI International. The opinions expressed in this paper are those of the authors and not necessarily those of the supporting agencies.

References

Brooks, S. P., and A. Gelman. 1997. General Methods for Monitoring Convergence of Iterative Simulations. Journal of Computational and Graphical Statistics 7: 434–455.

Fay III, R. E., & Herriot, R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to Census data. *Journal of the American Statistical Association, 74*(366a), 269–277.

Gelman, A., and D. B. Rubin. 1992. Inference from Iterative Simulation Using Multiple Sequences. Statistical Science 7: 457–511.

Harter, R., Macaluso, M., & Wolter, K. (2003). Evaluating the fundamentals of Illinois' small domain estimator. *Survey Methodology, 29*, 63–70.

Rao, J.N.K. and Molina, I. (2015). *Small Area Estimation, 2nd Ed.* Hoboken, NJ: John Wiley & Sons.

Sinha, S. K. and Rao, J. N. K. (2009). Robust small area estimation. *The Canadian Journal of Statistics*, 37, 381-399.

Sukasih, A., Iriondo-Perez, J., & Harter, R. (2019). Small Area Estimation for the 2017 OMAS: Methodology Report. Report prepared for the Ohio Colleges of Medicine Government Resource Center.

Ybarra, L. & Lohr, S. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika* 95(4), 919-931.