Small Area Estimates of the Child Population and Poverty in School Districts Using Dirichlet-Multinomial Models *

Jerry Maples†

Abstract

The Small Area Income and Poverty Estimates (SAIPE) program produces estimates of the number of children, total and in poverty, for each school district across the United States. Currently, the estimation methodology for school district child population is based on the most recent decennial census school district to county shares that do not change between censuses. The methodology for school district poverty estimates is based on shares determined by the most recent ACS 5-year and Federal Tax data (Maples 2007). Neither method is built on a statistical model framework. Preliminary research has shown that the Dirichlet-Multinomial small area model is a promising model framework for modeling the subcounty to county population shares. We propose a pair of Dirichlet-Multinomial small area models to jointly estimate relevant school-aged child population and poverty. Data from the American Community Survey and Federal Tax records will be used to fit the models. An added improvement in switching to a stochastic model-based form is that prediction errors can now be quantified for both population and poverty estimates.

Key Words: Small Area, poverty, population estimates, administrative records, SAIPE

1. Introduction

The U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program provides annual estimates of income and poverty statistics for all states, counties, and school districts of the U.S. Additionally, SAIPE also produces population estimates of the number of age-relevant children for each school district. All of the estimates at the state and county levels of geography from the SAIPE program are produced through model-based methods. The school district estimates are the exception to this. Currently, the population and poverty estimates for school districts uses a purely synthetic estimation procedure. The estimates from the Population Estimates Program (PEP) are deemed to be accurate at the state level and assumed to be mostly accurate for county-level estimates (more likely to be true for larger counties, but less likely for the smaller counties). Sub-county estimates of population are not assumed to be error free, however, and obtaining any measure of uncertainty has been problematic. This is the motivation to create a model-based estimate of population for school districts (or any sub-county domain).

School districts have unique characteristics compared to other geographies such as counties, census tracts and blocks. School districts are nested within state but not necessarily within other geographies such as county and tract. They are arbitrary regions that can cross county boundaries. Some school districts are coterminous with counties (such as in the state of Maryland), some districts are properly contained within counties, and others cross over county lines. The SAIPE program handles all of these different cases by creating a geography called *school district*

*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 those of the U.S. Census Bureau.

†U.S. Census Bureau, 4600 Silver Hill Rd., Washington, DC 20233, jerry.j.maples@census.gov

piece which is the intersection of county and school district. One advantage for this geography is to be able to calibrate the school district piece estimates to the county-level totals for population or poverty. One unfortunate side effect of this process is that some school district pieces are extremely small and have very little data from sample surveys and administrative record sources.

The geography of school districts is crossed with a second dimension, grade range. Not all school districts cover all of the grade ranges K-12. Some areas are serviced by multiple school districts that cover different grade ranges. For example, an elementary school district and a secondary school district may contain a common set of housing units. School districts that cover the full grade range are called unified school districts. There are even some areas where housing units are in their own elementary school district but then the secondary grade ranges are covered by a neighboring unified school district. Surveys and administrative record sources rarely ever record the current school grade for children. We assume there is a one-to-one relationship between age and school grade, that 5 year olds are in kindergarten (K) and that 17 year olds are in 12th grade. This may not be true for all cases, but will be the working assumption.

The current estimation method for school age child population is based on the observed school district piece to county population share from the most recent decennial Census multiplied by the current year county population estimate $cpop_c$. For school district piece k in county c the estimate for the school age child population $sdpop_{ck}$ is

$$\text{sdpop}_{ck} = \text{Census Share}_{ck} \times cpop_c \tag{1}$$

This allocation method uses a fixed share through out the decade until the next census is performed. The population counts for school districts pieces are updated only through changes to the current year county population estimates. While the population shares may be approximately true, i.e. without error, for the census year, in the following years the underlying true population shares may change. This source of estimation error is not incorporated in estimates of child population for school districts or any other secondary use of these population estimates, such as school district poverty.

For the school district poverty estimates, aggregates from the Federal Tax Data are used to compute a tax-based child poverty rate. When yearly income tax is filed, the number of child exemptions are recorded on the tax form. Tabulations at the school district piece level for the number of total child exemptions and number of child exemptions on forms below the poverty threshold are used to help predict the school aged child population and poverty. First, the school district piece to county share of child tax exemptions and child tax exemptions in poverty are calculated, called *taxchildshare* and *taxpoorshare* respectively using the Minimum Change method (Maples and Bell, 2007) to account for tax exemptions that cannot be geocoded to a school district piece within a county. Next the tax-based poverty rate is multipled by school district piece population estimate given in (1).

$$y_{ck}^{Pov} = \frac{\text{taxpoorshare}_{ck}\text{cpov}_c}{\text{taxchildshare}_{ck}\text{cpop}_c}\ \text{sdpop}_{ck} \tag{2}$$

The estimates from (2) do not generally add up to the county poverty estimate $cpov_c$ and must be ratio adjusted (raked). Full details of the current estimation procedures are given in Bell et. al. (2016). Note that the denominator of (2) could be used as an alternative school district population estimate, but to stay consistent with the rest of the Census Bureau's population estimates, the version in (1) is used.

Over the past few recent years, new methods have been implemented by the staff in the Census Bureau to improve the geocoding of tax returns to the school district pieces. This improvement is achieved through better processing and imputation using zip code information. This eliminates the need for the Minimum Change method and allows the direct tabulations of the child exemptions to be used directly in modeling. The goal is to create a model to jointly estimate school district poverty and population estimates for school aged children. The modeling strategy will be to separately estimate the school aged children in-poverty and not-in-poverty since those are non-overlapping groups. The population estimate will be the sum of the in-poverty and not-in-poverty estimates. In the section 2, a small area share model based on the Multinomial-Dirichlet will be developed to predict the school district piece to county share and the additional steps to create estimates and uncertainty measures for whole school districts. In Section 3, this methodology will be used to create estimates for the school districts in Nebraska.

2. Dirichlet-Multinomial Model

The goal is to jointly model the school aged child poverty and population for school districts. To achieve this, a model based on estimating the school district piece to county share will be developed for both the school aged children in poverty and the school aged children not in poverty. These share models will be based on the Dirichlet-Multinomial (DM) distribution, which is the multi-category generalization of the Beta-Binomial distribution. The DM model allows for key features needed to jointly estimate both of poverty and population. First, the model can be specified to handle a varying number of school district pieces per county. Second, the model should be invariant to the order that the school district pieces are labeled within a county. Finally, by estimating the shares rather than a counts, as long as the shares sum to 1 within county then the aggregate to the county will always match the given county total. Additionally, by breaking the population estimate into the poverty and not-in-poverty components, the model will never estimate more children in poverty than total children.

Let $\hat{y}_{c1}, ..., \hat{y}_{cK_c}$ be set of counts for the K_c school district pieces for county c. The relationship between the $\hat{y}_{ck}$ and the the survey weighted estimate of the shares $\hat{p}_{ck}$ will be given in Section 2.1. The sampling model, which characterizes the sampling distribution given the true underlying shares is a multinomial:

$$\hat{\mathbf{y}}_c = (\hat{y}_{c1}, ..., \hat{y}_{cK_c}) | \mathbf{p_c} \sim \text{Multinomial}(n_c, \mathbf{p_c}) \tag{3}$$

where $n_c = \sum_k \hat{y}_{ck}$, and $\mathbf{p_c}$ is the true share for piece k in county c. Note that ideally a direct distribution on the survey estimated shares would be desirable, but instead is defined through a multinomial count distribution.

The linking model describes the relationship between covariates and the true underlying share. This is also the part of the model that allows the 'borrowing of strength' between areas.

$$\begin{aligned} \mathbf{p}_c &\sim \text{Dirichlet}(\boldsymbol{\alpha}_c) = \text{Dirichlet}(\tau \times \boldsymbol{\pi}_c) \\ \pi_{ck} &= e^{X_{ck}^T \beta} / \sum_j e^{X_{cj}^T \beta} \end{aligned} \tag{4}$$

where $\sum_k \pi_{ck} = 1$, and τ is the model precision parameter. This parametric form of the mean function can be viewed as a generalized logistic function. It can also

be derivied from taking the shares from the model-based estimates of log-linear models noting that the intercept term cancels in the numerator and denominator. Additionally, predictor variables X_{ck} which do not vary within county, including the intercept term, factor out and cancel in (4) as they provide no information in differentiating the shares within county.

The Dirichlet-Multinomial model has the following marginal (observed data) likelihood after integrating out the unobserved shares $\mathbf{p_c}$ from the hierarchical specification given in (3) and (4):

$$L(\theta; \hat{\mathbf{y}}_\mathbf{c}) = \prod_{c=1}^{C} \left(\frac{\Gamma(n_c+1)\Gamma(\tau)}{\Gamma(n_c+\tau)} \prod_k \frac{\Gamma(\hat{y}_{ck}+\alpha_{ck})}{\Gamma(\hat{y}_{ck}+1)\Gamma(\alpha_{ck})} \right) \tag{5}$$

where $\theta = (\beta, \tau)$. From the likelihood in (5), the parameters can be estimated by maximum likelihood.

2.1 Estimation of effective sample size

Generally, one should not directly use the survey weighted totals from each school district piece for the multinomial distribution in (3). Ignoring the complex design from the survey will typically over estimate the precision of the survey data in model-based methods. One solution to this problem is to adjust the survey weighted counts so that they sum up to an effective sample size which conveys the correct amount of statistical information, e.g. number of independent observations, for model inference. Before computing the effective sample size, first the shares $\hat{p}_{ck}$ and their estimated design-based variance $\widehat{Var}(\hat{p}_{ck}) = \hat{V}_{ck}$ from the survey data are computed.

One could compute the effective sample size for each category, comparing one category to all of the others, using the standard formula for the binomial distribution:

$$n_{eff,ck} = \frac{\hat{p}_{ck}(1-\hat{p}_{ck})}{\hat{V}_{ck}}$$

The problem with the above approach is that one obtains a different effective sample size estimate, n_{ck}, for each category of the multinomial. A suggested modification of the binomial effective sample size formula was given by McAllister and Ianelli (1997) as follows:

$$\begin{aligned} n_c &\approx \frac{\sum_{k=1}^{K_c} \hat{p}_{ck}(1-\hat{p}_{ck})}{\sum_{k=1}^{K_c} \hat{V}_{ck}} \\ &= \sum_{k=1}^{K_c} w_{ck} n_{ck}, \qquad w_{ck} = \hat{V}_{ck} / \sum_{j=1}^{K_c} \hat{V}_{cj} \end{aligned} \tag{6}$$

This approximation can be viewed as a weighted average of the component effective sample sizes. This formula is also robust to one or more categories having a zero count, as long as one category does not have 100%.

Once the effective sample size for county c is calculated, the survey weighted shares can be converted into counts for use in the multinomial distribution specified in (3). Let $\hat{y}_{ck} = \hat{p}_{ck} \times n_c$. Note that to keep the distribution properly specified, the $\hat{y}_{ck}$'s and n_c will need to be rounded to integers.

2.2 Prediction and mean squared prediction error

The model-based prediction for the share S_{ck} from the DM model is:

$$\begin{aligned} S_{ck} = S_{ck}(\hat{\theta}) = E(p_{ck}|\hat{y}_{ck}; \theta = \hat{\theta}) &= \frac{\hat{y}_{ck} + \hat{\tau}\hat{\pi}_{ck}}{n_c + \hat{\tau}} \\ &= \frac{n_c}{n_c + \hat{\tau}}\hat{p}_{ck} + \frac{\hat{\tau}}{n_c + \hat{\tau}}\hat{\pi}_{ck} \end{aligned} \tag{7}$$

where $\hat{\pi}_{ck} = \exp(X_{ck}^T\hat{\beta})/\sum_j \exp(X_{cj}^T\beta)$ This is the empirical Bayes, also the linear Bayes, predictor using the estimated parameters $\hat{\theta}$ in place of the true parameters. Note that the predictors of the shares, S_{ck}, sum up to 1 within county. The predicted share is a weighted average of the design-based and model-based estimate.

The mean squared error of prediction for the share estimate is

$$\begin{aligned} MSE(S_{ck}) &= E_{\hat{y}_c}(S_{ck}(\hat{\theta}) - p_{ck})^2 = E_{\hat{y}_c}(S_{ck}((\hat{\theta}) - S_{ck}(\theta)) + (S_{ck}(\theta) - p_{ck}))^2 \\ &\approx (1 - w_c)\hat{\sigma}_{ck}^2 + \\ &\quad w_c\hat{\sigma}_{ck}^2 \mathrm{Var}(\hat{\tau})/(n_c + \hat{\tau})^2 + \\ &\quad (1 - w_c)^2(\pi_{ck})^2(X_{ck} - X_{cw}^*)^T \mathrm{Var}(\hat{\beta})(X_{ck} - X_{cw}^*) \end{aligned} \tag{8}$$

$$\begin{aligned} \hat{\sigma}_{ck}^2 &= \hat{\pi}_{ck}(1 - \hat{\pi}_{ck})/(1 + \hat{\tau}) \\ w_c &= n_c/(n_c + \hat{\tau}) \\ X_{cw}^* &= \sum_j \exp(X_{cj}^T\hat{\beta})X_{cj}/\sum_j \exp(X_{cj}^T\hat{\beta}) \end{aligned}$$

The first term in (8) is the mean squared prediction error when the parameters are known. The second and third terms are the additional mean squared error due to estimating τ and β respectively. The MSPE was written in terms of the variance of the Dirichlet (linking) distribution, i.e. model error variance σ^2, and the shrinkage weight. This is similar to how the MSPE is written for a linear Fay-Herriot style small area model.

Estimation of the within county shares for school aged children in- and not-in-poverty is only an intermediate step in predicting school district population and poverty. The next step is to convert the predicted shares into counts for the school district pieces (sdp). Let C_c^{pop} denote the demographic population of the number of school age children in county c and C_c^{pov} denote the SAIPE estimate of number of school aged children in poverty for county c. By subtraction, $C_c^{nonpov} = C_c^{pop} - C_c^{pov}$, the number of school age children not in poverty in county c is obtained. Since the shares are the within county allocation of the total, the estimate of the number of children in-poverty and not-in-poverty for school district piece k is the product of the share and the appropriate county total. The school district piece population estimate is the sum of the in-poverty and not-in-poverty estimates,

$$\begin{aligned} sdp_{ck}^{pov} &= S^{pov} \times C_c^{pov} \\ sdp_{ck}^{pop} &= sdp^{pov} + S_{ck}^{nonpov} C_c^{nonpov} \end{aligned} \tag{9}$$

The county population estimates from the Census Bureau's Population Estimates Program is assumed to be without error. The SAIPE county estimate for number of school age children in poverty has a variance of $Var(C_c^{pov}) = V_c$. Since the in-poverty and not-in-poverty sum to a fixed total and that total is without

error, then $Var(C_c^{nonpov}) = V_c$ and $Corr(C_c^{pov}, C_c^{nonpov}) = -1$. The MSE of the in-poverty and not-in-poverty counts for school district piece k is calculated based on the conditional variance formula to take into account the uncertainty in both the share and the county count estimates.

$$\begin{aligned} MSE(sdp_{ck}^{pov}) &= (S_{ck}^{pov})^2 V_c + MSE(S_{ck}^{pov})((C_c^{pov})^2 + V_c) && (10) \\ MSE(sdp_{ck}^{nonpov}) &= (S_{ck}^{nonpov})^2 V_c + MSE(S_{ck}^{nonpov})((C_c^{nonpov})^2 + V_c) && (11) \\ MSE(sdp_{ck}^{pop}) &= MSE(sdp_{ck}^{pov}) + MSE(sdp_{ck}^{nonpov}) - 2S_{ck}^{pov} S_{ck}^{nonpov} V_c && (12) \end{aligned}$$

The last step is to create the estimates and MSEs for the whole school districts (sd). The models assume that pieces in one county are independent of the pieces in another county, therefore the different pieces of a given school district are independent from each other. The count estimate and the MSEs for both the poverty and population of school age children for the school district is the sum of the pieces. Let d index the school district and $ck \in d$ denotes the set of school district pieces for school district d.

$$\begin{aligned} sd_d^{pov} &= \sum_{ck \in d} sdp_{ck}^{pov} \\ MSE(sd_d^{pov}) &= \sum_{ck \in d} MSE(sdp_{ck}^{pov}) \\ sd_d^{pop} &= \sum_{ck \in d} sdp_{ck}^{pop} \\ MSE(sd_d^{pop}) &= \sum_{ck \in d} MSE(sdp_{ck}^{pop}) \end{aligned} \quad (13)$$

3. Application: School District Estimates for Nebraska

The joint Dirichlet-Multinomial model from the previous section will be used to make predictions for the school districts in Nebraska for 2012. School district piece to county shares are calculated from the 5-year 2012 American Community Survey (ACS) data. The axillary variable X_{ck} is the share based on the tabulations of child exemptions in-poverty and not-in-poverty. A child exemption is determined to be in poverty if it is on a tax return form that reports income below the poverty threshold based on family size which is determined by the total number of exemptions. Tabulations of the federal tax data is used at multiple levels of models (state, county, and school district) because of the high correlation between poverty based on the tax data and survey estimates of poverty. It is also one of the few auxiliary datasets that can be broken down to sub-county estimates. In Nebraska, there are 250 school districts. The state has 93 counties which, when intersected with the school districts, creates 581 school district pieces. Figure 1 shows the scatter plot of the design-based estimates of the within county shares versus the tax data based shares for both in-poverty and not-in-poverty. The relationship between the not-in-poverty shares is stronger than for the poverty shares. The poverty shares from the ACS had a much higher sampling variance.

The data for the shares was fitted to the Dirichlet-Multinomial models using maximum likelihood. The parameter estimates are given in Table 1. Also given is the median county-level effective sample size. Unlike other regression style models, the null model is when $\beta = 1$ which is the basic share model, sometimes referred

Figure 1: Plot of the ACS in-poverty and not-in-poverty shares to the Federal Tax data shares

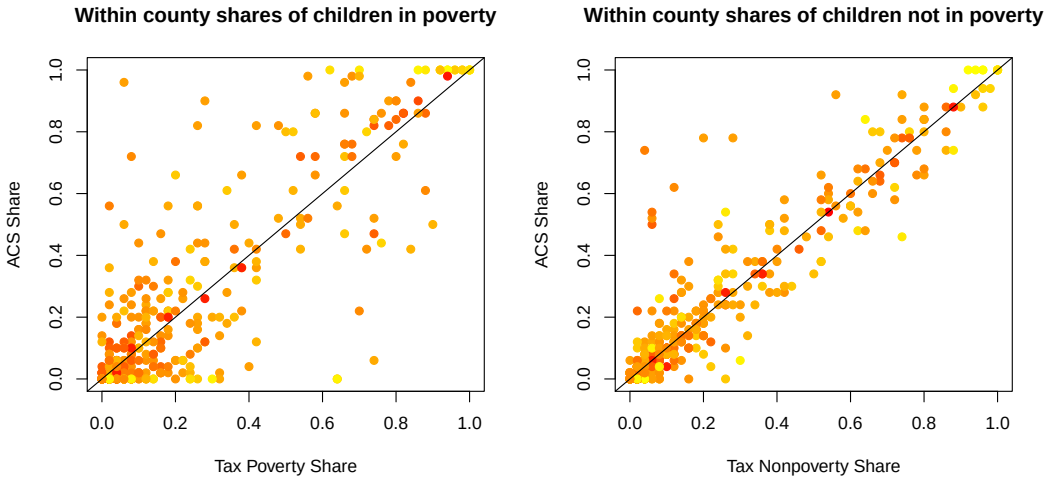


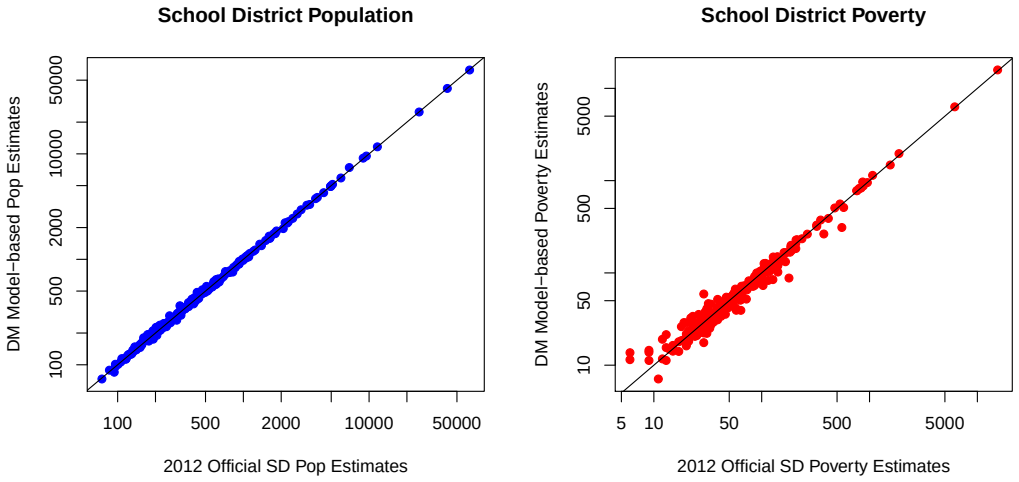
Table 1: Model Parameter Estimates

	β (S.E.)	τ (S.E.)	Median n
Poverty Shares	1.04	50.85	10.66
	(.04)	(19.32)	
Non Poverty Shares	1.00	35.35	94.95
	(.01)	(2208)	

to as the synthetic ratio method (Rao and Molina, 2015). The model formulation generalizes the share model, allowing the data to determine the best fit. For both the in-poverty and not-in-poverty shares, the basic share model could not be rejected. The τ precision parameter can be viewed as an effective sample size for the model-based estimates. The τ was about 5 times higher than the median effective sample size for the poverty shares, but only 2.7 times higher for the not-in-poverty shares.

The official SAIPE estimates for Nebraska school districts in 2012 did not include measures of uncertainty for the population counts. The poverty estimates do give guidance on an approximate coefficient of variation (CV) for the estimates based on the population size of the county (based on a moment based method proposed in Maples 2008). These measures are crude and seem to give artificially large estimates of the CV. Rather than using using the CVs from Maples (2008), the DM model estimates will be compared to using the ACS direct share estimates. The population estimates for school age children went from a median CV (across school districts) of 26.6% to 21.2%. For poverty estimates, the median CV decreased from 45.2% to 26.7%. Lastly, Figure 2 shows the plot of the model-based estimates for school age child population and poverty estimates to the official released 2012 SAIPE estimates for Nebraska. The population estimates match the official estimates more closely than do the poverty estimates.

Figure 2: Plot of the ACS in-poverty and not-in-poverty shares to the Federal Tax data shares



4. Discussion

A small area model based on the Dirichlet-Multinomial (DM) distribution was presented to model the school district piece to county share of the number of school age children in poverty and not in poverty. One of the key differences in this application of the DM model to others is that the number of categories varies between observations and the category labels themselves are not meaningful. However, this setup of the DM model could be highly useful in other applications for modeling within county counts when these counts will be calibrated to the county total. Prediction and mean squared estimates are given that also account for uncertainty in the county level totals. Additionally, the two estimates of interest, number of school age children in poverty and total number of school aged children, are nested and splitting the total count into in-poverty and not-in-poverty models made keeping the relationship between poverty and population proper. The federal tax data showed to be a highly predictive covariate whose simple shares were very close to the ideal model-based estimators.

This model was only used on a single state, Nebraska, which only had unified school districts. Unified school districts contain the full grade range K-12. Other states contain a mixture of unified, elementary and secondary school districts. The elementary districts cover the lower grade ranges and the secondary districts cover the higher grade ranges (the exact grade ranges can vary). The next step is to augment the model to fit the entire nation of school districts. The fixed effects may need to include interactions with the school district type (unified, elementary and secondary) due to potential interactions with household poverty status and age of children present in household. Another enhancement to the model is on the data processing side where obtaining the age information for the child exemptions could allow for more informative tabulations of the tax data.

REFERENCES

Bell, W., Basel, W., and Maples, J., (2016). "An Overview of the U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) Program", in *Analysis of Poverty Data by Small Area Methods*, Monica Pratesi (ed.), London: Wiley.

Maples, J. and Bell, W. (2007) "Small Area Estimation of School District Child Population and Poverty: Studying the Use of IRS Income Tax Data", *SRD Research Report Series RRS2007/11.*

Maples, J. (2008) "Calculating Coefficient of Variation for the Minimum Change School District Poverty Estimates and the Assessment of the Impact of Nongeocoded Tax Returns ", *SRD Research Report Series RRS2008/10.*

McAllister, M. and Ianelli, J. (1997) "Bayesian stock assessment using catch-at-age data and the sampling-importance resampling algorithm", *Canadian Journal of Fisheries and Aquatic Science*, **54(2)**, 284-300.

Rao, J. and Molina, I. (2015), Small Area Estimation 2nd Edition, Hoboken, NJ, John Wiley and Sons, Inc.