Small Area Model Diagnostics and Validation with Applications to the Voting Rights Act Section 203

Robert Ashmead[1] and Eric Slud[1]

Abstract

We consider the dual problems of choosing between competing small area models and validating model assumptions in an area-level model. Many classes of small area models result in an estimate that is a convex combination of the direct and the marginal estimate for a given area. Therefore, competing models may share the same direct estimates, but give different marginal estimates as well as relative weight on the estimates. We discuss diagnostics to choose between competing models and parametric bootstrap methods to check for model validity and goodness of fit. We use the example of small area models related to the *Voting Rights Act Section 203(b)*, which are used to estimate the number of limited English proficient and illiterate persons in certain language minority groups within jurisdictions using 5-year data from the American Community Survey.

1. Introduction

Small area estimation (SAE) models are often employed when the direct survey estimates from small domains are unreliable, meaning that their standard errors are too large or the estimates themselves are based on extremely few observations. See Rao and Molina (2015) for an overview. SAE models can improve on the direct survey estimates by "borrowing strength" from similar small areas or by taking advantage of relationships between covariates and the variable of interest. In any modeling situation, model selection and model diagnostics are essential steps of the modeling process. This holds true for SAE models and more generally for mixed-effect models, which are typically used in SAE. Mixed-effect models present unique challenges, and as a result diagnostic methods are not as well developed as in other areas of statistics.

Model selection and diagnostics with mixed-effect models are challenging because the prediction target of interest often includes the random effect, which is generally unobservable. In addition, for many models, the best predictions are convex combinations of the direct estimator and a marginal mean regression estimator. As a result, models can be evaluated not only by their marginal mean regression estimator, but by the relative weight they give to the direct versus marginal mean part. For example, at one extreme if a model gives almost all its weight to the direct estimates, the predictions will be incredibly close to the direct estimate and when compared may look like a well-fitting model. Lastly, generally in SAE modeling we do not observe the true target values for small areas, only direct survey estimates. As a consequence, methods relying on out-of-sample predictions or validations must be carefully considered.

In this paper, we consider only area-level models. Our goal is to estimate a parameter of interest θ_i for each small area indexed $i = 1, \ldots, m$. Often small area models consist of two parts, the population model and the sampling model. Generically, we assume a class of models such that for small areas $i = 1, \ldots, m$

$$\begin{aligned} y_i|\theta_i &\sim f(\cdot\,, \theta_i) \ \text{ [Sampling Model];} \\ \theta_i|\tau, \psi_i &\sim h(\cdot\,, \tau, \psi_i) \ \text{ [Population Model];} \\ \psi_i &= g(\mathbf{x}_i'\beta); \ \text{ and} \\ E[y_i|\theta_i] &= \theta_i, \end{aligned} \tag{1}$$

[1]U.S. Census Bureau, Center for Statistical Research & Methodology, 4600 Silver Hill Road, Washington DC 20233. robert.douglas.ashmead@census.gov, eric.v.slud@census.gov. This report is released to inform interested parties of (ongoing) research and to encourage discussion (of work in progress). The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

where $\mathbf{x_i}$ are known covariates. The last equation specifies that the direct estimate y_i is an unbiased estimator of θ_i, and it implies that $E[y_i - \theta_i] = 0$.

This class of models is useful for small area estimation applications, but also could include other random effect data problems. In the SAE context, y_i is the unbiased survey direct estimate of θ_i. The most widely used area-level SAE model, the Fay-Herriot model, is a linear mixed model and can be written as

$$\begin{aligned} y_i|\theta_i, \sigma_i^2 &\sim \mathcal{N}(\theta_i, \sigma_i^2) \\ \theta_i|\psi, z_i, \tau &\sim \mathcal{N}(\psi_i, z_i^2\tau^2) \\ \psi_i &= \mathbf{x}_i'\beta, \end{aligned} \tag{2}$$

with σ_i^2 and z_i known positive constants for $i = 1, \ldots, m$. The z_i's could be for example, a covariate that is proportional to the random effect variance. If the parameters of interest are proportions, one could use a logistic regression-type model with a normal random effect:

$$\begin{aligned} y_i = t_i/n_i;\ t_i|\theta_i, n_i &\sim Binomial(n_i, \theta_i) \\ \theta_i|\psi, z_i, \tau &\sim logitNormal(\psi_i, z_i^2\tau^2) \\ \psi_i &= \mathbf{x}_i'\beta, \end{aligned} \tag{3}$$

with n_i and $z_i > 0$ known for $i = 1, \ldots, m$. The logit-normal distribution is such that if $X \sim logitNormal(\mu, \sigma^2)$, then $logit(X) \sim \mathcal{N}(\mu, \sigma^2)$. Similarly, we could estimate proportions with a beta-binomial model with a logit link:

$$\begin{aligned} y_i = t_i/n_i;\ t_i|\theta_i, n_i &\sim Binomial(n_i, \theta_i) \\ \theta_i|\tau, \psi_i &\sim Beta\left(\tau\psi_i, \tau(1-\psi_i)\right) \\ \psi_i &= logit^{-1}(\mathbf{x}_i'\beta), \end{aligned} \tag{4}$$

with n_i known for $i = 1, \ldots, m$.

Define the marginal mean of small area i as $\theta_{i,marg} = E[y_i] = E[E[y_i|\theta_i]] = E[\theta_i] = \int \theta_i h(\cdot\,, \tau, \psi_i) d\theta_i$. In both the Fay-Herriot (2) and Beta-Binomial (4) models, the marginal mean is equal to ψ_i, whereas in the the normal logistic regression model (3), the marginal mean must be computed numerically as a function of the parameters. In the cases of models (2) and (4) this is a useful property of the EBLUP estimator $\tilde{\theta}_i$. For (2) and (4), the EBLUP estimator is a convex combination of the direct estimate y_i and the estimated marginal mean $\hat{\theta}_{i,marg} = \hat{\psi}_i$, i.e.,

$$\tilde{\theta}_i = \alpha_i y_i + (1 - \alpha_i)\hat{\theta}_{i,marg} \tag{5}$$

The parameter α_i is determined for each small area by the relative sizes of the sampling variance and the model variance. In the case of model (2), $\alpha_i = z_i^2\hat{\tau}^2/(\sigma_i^2 + z_i^2\hat{\tau}^2)$, and in the case of model (4), $\alpha_i = n/(n + \hat{\tau})$. When the sampling variance for area i is small, α_i will be close to 1.

The convex-combination property (5) motivates the use of the estimated marginal mean in the calculation of residuals for diagnostic purposes. While it might seem natural to calculate residuals comparing the predicted values $\tilde{\theta}_i$ with their direct estimates y_i, we argue that comparisons of the marginal mean estimator $\hat{\theta}_{i,marg}$ and y_i are more useful. Assume that our estimators have the convex-combination property in equation (5). Then the residual between the $\tilde{\theta}_i$ and y_i can be re-written as a weighted version of the direct minus marginal residuals:

$$\tilde{\theta}_i - y_i = (1 - \alpha_i)(\hat{\theta}_{i,marg} - y_i). \tag{6}$$

When α_i is near 1, the $\tilde{\theta}_i - y_i$ will be zero no matter how badly the marginal predictor estimates θ_i. Therefore, we found it more useful to consider residuals between $\hat{\theta}_{i,marg}$ and y_i.

The goal of this paper is to propose and illustrate a method for model selection based on cross-validation as well as a parametric bootstrap method for model validation and diagnostics. In the following sections we give an overview of existing SAE diagnostic methods, discuss the *Voting Rights Act* (VRA) *Section 203(b)* application, and propose the model selection and diagnostic methods. Lastly, we apply them to the VRA application as well as to a simulation.

2. SAE Model Selection, Diagnostics and Validation Methods

When model error distributions are specified, methods based on information-based criteria (AIC, BIC, etc.) are available for model selection. In the case of linear mixed models such as (2), extensions to information criteria have been developed (Müller et al., 2013) as well as a fence method (Jiang et al., 2008) which utilizes a lack-of-fit measure rather than a log-likelihood function. Residual analysis methods are also available for some methods. With the Fay-Herriot model (2), it is possible to transform the model into a standard linear regression model and use the typical residuals based on the transformed data (Calvin and Sedransk, 1991). So-called BLUP residuals (Calvin and Sedransk, 1991) $\tilde{e}_i = y_i - \tilde{\theta}_i$ can also be used investigate the distribution of model errors; however, the BLUP residuals are correlated, making interpretation more difficult. Skinner (2007) proposes a cross-validation method for linear mixed models in which predictions from small areas left out of fitting are compared with the direct estimate, and a Wald test statistic is formed and compared with a chi-squared distribution in order to test for departures from the true model. Additionally, Tang et al. (2014) propose a class of goodness-of-fit tests for the mean structure of linear mixed models based on observations and expectations restricted by where the covariates lie.

Many of these methods apply only to linear mixed models, and therefore do not apply to other types of models such as (3) and (4) above. While motivated by longitudinal repeated measures data, not small area estimation, Pan and Lin (2005) present goodness-of-fit methods for generalized linear mixed models based on the cumulative sums of residuals. In their method, residuals are defined as the difference between observed responses and the marginal means of the observed responses $r_i = y_i - \hat{\theta}_{i,marg}$. Sums of residuals are calculated over covariates or predicted values and compared with realizations of zero-mean Gaussian processes, which approximate the asymptotic distributions under the assumed model. This resembles our parametric bootstrap method in Section 2.3 in that we utilize residuals with respect to the estimated marginal mean and are comparing a function of residuals to their approximate distribution under the assumed model. However, in our method we calculate the reference distribution from a parametric bootstrap and utilize sums of squared residuals.

2.1 Voting Rights Act Section 203(b)

According to *Section 203(b)* of the *U.S. Voting Rights Act*, the Census Bureau Director determines subdivisions that are required to provide language assistance during elections for designated language-minority groups of citizens who are unable to speak or understand English well enough to participate in the electoral process. The criteria for determinations are based on estimates of counts and ratios of language minority group voting-age persons (VOT), voting-age citizens (CIT), limited English proficiency voting-age citizens (LEP), illiterate limited English proficiency voting-age citizens (ILL), and total voting-age persons in each specific political jurisdiction. Political subdivisions include States, counties or Minor Civil Divisions, and American Indian areas, with specific determination criteria for each.

In order to improve the reliability of estimates from the American Community Survey used to estimate the quantities needed to make the determinations, it was decided in 2011 [see Joyce et al. (2014)] and again for 2016 to make use of small area estimation models. Multiple models are necessary because estimates are needed across geographies for each of 68 non-exclusive language minority groups (LMGs). The LMGs consist of 16 Asian groups, 51 American Indian groups, and a Hispanic group. Additionally, more than one outcome needs to be estimated in the models. Estimates for each of the four nested categories above (VOT, CIT, LEP, ILL), are needed for the determination criteria.

This setup makes for a difficult small area estimation problem. We not only need to make predictions for multiple nested quantities, but we need to create many models for many different partitions of the sampled population into areas (e.g., jurisdictions). The general model form chosen for this problem was a Dirichlet-Multinomial model with 4 categories. This model is a generalization of the beta-binomial model (4). The four estimation categories (VOT, CIT, LEP, ILL) needed for determinations are nested, not mutually exclusive. Therefore, we translated these into mutually-exclusive categories that were used to fit the model, then we translated them back. Let $\hat{N}_i^V$, $\hat{N}_i^C$, $\hat{N}_i^L$, and $\hat{N}_i^I$ be the directly estimated totals of VOT, CIT, LEP, and ILL persons, respectively, in jurisdiction i for a given LMG. We write the model as:

$$\mathbf{T}_i = \frac{n_i}{\hat{N}_i^V}(\hat{N}_i^V - \hat{N}_i^C, \hat{N}_i^C - \hat{N}_i^L, \hat{N}_i^L - \hat{N}_i^I, \hat{N}_i^I)$$

$$\mathbf{Y}_i = \mathbf{T}_i/n_i;\ \mathbf{T}_i \sim \text{Multinomial}(n_i, \boldsymbol{\omega}_i),$$

$$\boldsymbol{\omega}_i = (\omega_{i1}, \omega_{i2}, \omega_{i3}, \omega_{i4}) \sim \text{Dirichlet}(\tau\sqrt{n_i}, 1-\mu_i, \mu_i(1-\nu_i), \mu_i\nu_i(1-\rho), \mu_i\nu_i\rho) \quad (7)$$

$$\mu_i = exp(\boldsymbol{\beta}'\mathbf{X}_i^C)/exp(1+\boldsymbol{\beta}'\mathbf{X}_i^C)$$

$$\nu_i = exp(\boldsymbol{\gamma}'\mathbf{X}_i^L)/(1+(\boldsymbol{\gamma}'\mathbf{X}_i^L)$$

$$\text{for } i = 1, \ldots, m \text{ small areas (jurisdictions)},$$

where n_i is the sample size (number of voting age persons in the LMG sampled in small area i), μ_i represents the citizenship proportion, ν represents the LEP proportion among citizens, and ρ represents the illiteracy proportion among LEP citizens. $\mathbf{X}_i^C$ and $\mathbf{X}_i^L$ are sets of covariates that correspond to citizenship and limited English proficiency respectively. The data $\mathbf{Y}_i$ are directly estimated proportions of the four mutually exclusive categories. One slightly unusual aspect of the model is the use of a non-constant precision parameter $\tau\sqrt{n_i}$. The change from a constant precision was made in an effort to improve the overall fit of the model after looking a several diagnostics across LMGs. The covariates considered in the models were the corresponding rates directly estimated from the ACS at the level of the state containing the domain, as well as several other covariates such as educational level, age, proportion foreign born, and average time in US, separately calculated for all adults in the domain and also for the adults in the LMG within the domain.

While we estimate $\boldsymbol{\omega}_i$ in the model, we actually care about certain linear combinations of $\boldsymbol{\omega}_i$ instead: $\theta_i^C = 1-\omega_1, \theta_i^L = \omega_3+\omega_4$, and $\theta_i^I = \omega_4$. These represent the proportion of CIT, LEP, and ILL voting-age persons respectively among the voting-age population. Maximum likelihood estimation was used to fit the models. Marginal mean estimates of $(\theta_i^C, \theta_i^L, \theta_i^I)$ are given by

$$\hat{\theta}_{i,marg}^C = \hat{\mu}_i = exp(\hat{\boldsymbol{\beta}}'\mathbf{X}_i^C)/exp(1+\hat{\boldsymbol{\beta}}'\mathbf{X}_i^C)$$

$$\hat{\theta}_{i,marg}^L = \hat{\nu}_i = exp(\hat{\boldsymbol{\gamma}}'\mathbf{X}_i^L)/(1+(\hat{\boldsymbol{\gamma}}'\mathbf{X}_i^L))$$

$$\hat{\theta}_{i,marg}^I = \hat{\rho}$$

where $\hat{\beta}, \hat{\nu}, \hat{\rho}$ are the respective maximum likelihood estimates. EBLUP predictions of $(\theta_i^C, \theta_i^L, \theta_i^I)$ are made using Equation (5) with $\alpha_i = n_i/(n_i+\hat{\tau}\sqrt{n_i})$. The mean-squared prediction error of modeled estimates were made using a novel method containing balanced repeated replication and parametric-bootstrapping (Slud and Ashmead, 2017).

In terms of model selection, diagnostics, and validation methods this problem is particularly challenging because the sample sizes vary from very small to very large and our goal is to develop a model that fits well over all of them. Also, the LMGs are quite diverse in terms not only of their characteristics but also of the number of areas included in each. Instead of trying to optimize each individual model, we searched for classes of models that worked well over similar groups of LMGs. In this context similar means that the LMGs are related in data richness, i.e the numbers of jurisdictions with sample size over a certain small threshold. As a result, Asian LMGs, large American Indian, and small American Indian LMGs each used different covariate sets. Additionally, the discreteness of the data makes any sort of residual analysis difficult, especially for small sample sizes. Lastly, the multiple outcomes/levels of the model make evaluations multidimensional.

2.2 Model Selection Using Cross-Validation

Our first proposed technique is a cross-validation method to aid in model selection. The general idea is to compare the predictive properties of multiple possibly non-nested models. Consider the following procedure for a selected model:

1. Randomly order the m small areas;
2. Divide the small areas to K approximately equal sized groups;
3. For each group k in $1, \ldots, K$;

a. Leave out all the observations in group k and fit the model;

b. Estimate $\hat{\theta}_{i,marg}^{(-k)}$ for each small area in group k;

4. Repeat steps 1 through 3, L times;
5. Compute a statistic comparing the predictions to the direct estimates:

$$\Gamma = \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{m}\left((\hat{\theta}_{i,marg}^{(-k,l)} - y_i)\right)^2 \tag{8}$$

where $\hat{\theta}_{i,marg}^{(-k,l)}$ indicates the marginal mean estimate for small area i after being left out of the model fitting in iteration l.

A variation of the statistic (8) is to multiply the residuals by a function of the sample size.

$$\Gamma_{g(\cdot)} = \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{m}\left((\hat{\theta}_{i,marg}^{(-k,l)} - y_i)g(n_i)\right)^2 \tag{9}$$

In the VRA application, we used $g(\cdot) = \sqrt{n_i}$. We found this to be useful because the sample size varied by such extremes, and it was possible that many small jurisdictions might overwhelm the influence from medium and large jurisdictions. Using the $\sqrt{n_i}$ term also strikes a balance in the statistic between a sum of squared differences of proportions and totals.

In order to compare different models, we compute Γ for each model, ideally using the same leave-out groups. This allows for quantitative comparisons of different models and speaks to how predictive the marginal model is without the influence of the small areas themselves contributing to the model fit.

We have not yet developed reference distributions with which to compare (8) or (9). It may be possible do so using a parametric bootstrap; however, this would be computationally burdensome. Otherwise, it may be possible to derive large-sample limiting distributions for these statistics given specific models or classes of models. This will be the subject of future work.

The following examples are from the VRA application. In this exercise we considered four different models:

A. The model chosen for the actual production of the small area estimates which uses 5 covariates total (2 for the Citizenship level and 3 for the LEP level)

B. The model with the same covariates as a) but a constant precision parameter τ for all areas (as opposed to $\tau_i = \tau_o\sqrt{n_i}$)

C. A "lower" model with only 2 covariates (one for each Citizenship and LEP)

D. An "upper" model with 7 covariates total (2 and 5 respectively)

In Tables 1, 2, and 3 we show examples of Γ_g from three different LMGs. The first is a LMG with a moderate number of areas. The second and third both have a large number of areas. In addition to the Γ_g statistic for each outcome, we show the AIC for each model. In all three cases, we find that the AIC correlates strongly with the Γ_g statistic.

In all three tables we observe that model A does better than the "lower" model (C) by each of the metrics and generally by large margins. This indicates that the additional covariates are worthwhile. Comparing model A and B, we observe that model A does at least marginally better in all all cases and in some cases much better (AIC in Example 2 and Γ_g - LEP in Example 3). Interestingly, those large differences are not uniformly seen in both metrics. Comparing the "upper" model (D), which has two additional covariates, to model A, we find while it is best in each of the examples, only in Example 3 (Table 3) with the Γ_g - LEP category is the difference large (compared with differences between A and B or A and C). Still, without a reference distribution for Γ_g we cannot say for sure what is a meaningful difference.

Table 1: Prediction Statistics for VRA Example 1

Model	Γ_g - CIT	Γ_g - LEP	Γ_g - ILL	AIC
A	19.617	9.234	1.265	547.5
B	19.819	9.378	1.267	550.1
C	40.984	11.273	1.300	606.1
D	19.590	8.790	1.260	547.8

Table 2: Prediction Statistics for VRA Example 2

Model	Γ_g - CIT	Γ_g - LEP	Γ_g - ILL	AIC
A	551.278	367.065	13.783	13,524.3
B	552.939	367.650	13.842	13,727.9
C	2064.005	457.580	15.915	15,117.6
D	551.194	360.276	13.664	13,512.7

Table 3: Prediction Statistics for VRA Example 3

Model	Γ_g - CIT	Γ_g - LEP	Γ_g - ILL	AIC
A	489.144	1097.937	80.067	15,674.7
B	494.988	1433.494	92.360	15,678.3
C	2987.981	3275.775	91.765	17,832.6
D	488.370	769.515	78.284	15,563.9

2.3 Model Validation and Diagnostics using a Parametric Bootstrap

In general, bootstrapping is a useful method and is often used to estimate the variance of an estimator. In a parametric bootstrap, after selecting a fitted model of interest, we assume that model and its fitted parameters are true and we take random draws from the model corresponding to each observation. Next, we re-fit the model using the bootstrapped observations and create bootstrapped estimates. Repeating this process many times allows us to generate distributions of bootstrapped estimates and therefore estimate variance of those estimates under the assumption that the model is true. Instead of using the bootstrapped estimates to estimate variances, we can use them for model validation and diagnostics by comparing quantities of interest (e.g. sums of squares between y_i and $\hat{\theta}_{i,marg}$) from the actual data with their bootstrapped equivalents. When the assumed fitted model is true, the quantities of interest should generally fall in the central quantiles of the bootstrapped equivalents (at least at a rate corresponding to the quantile). However, if the assumed fitted model departs strongly from the true model, the quantities of interest will correspond to extreme quantiles of the bootstrapped equivalents, implying that the observed data could not have come from the given model. Additionally, the direction of the extreme quantile may give additional diagnostic information for the model. As in Section 2.2, we use statistics based on the squared residuals. Again, consider residuals based on the difference of the marginal mean and direct estimates. For the bootstrap diagnostic do the following:

1. For our observed data and fitted model, compute:

$$\widehat{SS} = \sum_{i \in \mathcal{A}} (\hat{\theta}_{i,marg} - y_i)^2 \tag{10}$$

   where $\mathcal{A}$ represents all or some subset of the areas.

2. Compute the reference distribution

$$\widehat{SS}^{(b)} = \sum_{i \in \mathcal{A}} (\hat{\theta}^{(b)}_{i,marg} - y^{(b)}_i)^2 \tag{11}$$

   for parametric bootstraps $b = 1, \ldots, B$ from the fitted model.

3. Compare the observed statistic $\widehat{SS}$ or to the reference distribution $\widehat{SS}^{(b)}$ by calculating the bootstrap quantile

$$Q = Pr(\widehat{SS} \geq \widehat{SS}^{(b)}).$$

Extreme values of Q (near 0 or 1), indicate that the amount of variation in the observed data is respectively less than or greater than the parametrically bootstrapped data. The choice of $\mathcal{A}$ depends on how we think the true model deviates from the fitted model. For example, if we think it possible that the random effect variance is not constant across all areas, then we may choose $\mathcal{A}$ to subdivide the areas (by size or by a covariate) into several groups and calculate the bootstrap diagnostic for each of the groups separately. When the variable used to form the groups is closely related to the model misspecification, the power of the method is improved.

As in in Section 2.2, we can weight the residuals by a function of the sample size. Define

$$\widehat{SS}_{g(\cdot)} = \sum_{i \in \mathcal{A}} [(\hat{\theta}_{i,marg} - y_i) g(\cdot)]^2, \text{ and} \tag{12}$$

$$\widehat{SS}^{(b)}_{g(\cdot)} = \sum_{i \in \mathcal{A}} [(\hat{\theta}^{(b)}_{i,marg} - y^{(b)}_i) g(\cdot)]^2. \tag{13}$$

As before, in our VRA application we use $g(\cdot) = \sqrt{n_i}$. Again, the parameters of interest in our VRA application are proportions. Instead of looking at squared differences of counts which would give much more weight to the large areas, or squared differences of proportions which give equal weight to all areas, multiplying by the $\sqrt{n_i}$ term splits the difference.

Consider a small simulation example with $m = 200$ small areas. Motivated by the VRA application, we use a beta-binomial model with two independent standard normal covariates. Assume that $\mu = exp(\boldsymbol{\beta}'\mathbf{X}_i)/(1+exp(\boldsymbol{\beta}'\mathbf{X}_i))$ and $\boldsymbol{\beta} = (0.1, -0.5, 0.5)$, where $\mathbf{X}_i = (1, X_{1i}, X_{2i})$ and X_{1i} and X_{2i} are independent draws from a standard normal distribution. Let $\theta_i \sim Beta(\tau_i\mu, \tau_i(1-\mu))$ where μ is the mean and τ the precision variable and $Y_i \sim Binomial(n_i, \theta_i)$ with n_i being generated from a Poisson random variable with rate parameters 15, 50, 100, and 200 for 50 small areas each. We first simulate the model using a constant τ value, $\tau_i = 4$, and fit that same model. Then, we simulate the data using $\tau_i = \sqrt{n_i}$, but fit a constant τ (misspecified) model in an attempt to see if the parametric bootstrap diagnostic $\hat{SS}_g$ can identify the model misspecification. In both cases we will use 1000 parametric bootstraps drawn from the fitted model. We repeat the simulation 100 times in each case.

First, we calculate $\hat{SS}_g$ including all 200 areas in the set $\mathcal{A}$. Using the correctly specified model, most of the bootstrap quantiles were near 0.5 and the median was 0.565. Only one out of one hundred were extreme (greater than 0.95 or less than 0.05), whereas under the misspecified model most of the bootstrap quantiles were small (median 0.0955) and 26 out of 100 were extreme. In this simulation the method was not always successful at identifying misspecified models; however, it can be improved by a more precise lack of fit statistic. If we are able to specify a variable to group areas by that is related to the misspecification, we can increase the power of the test. For example, in the VRA application, we were concerned that the model does not fit well for areas with a small population. Instead of summing over all areas to calculate $\hat{SS}_g$ and $\hat{SS}^{(b)}_g$, we will restrict ourselves to those with sample size less than or equal to 30. When we use this subset, we find over the 100 simulations, 65 gave extreme quantile values and the median quantile was 0.973. Note that the power of detecting lack of fit depends on the number of small areas. When we increase the number of small areas to 400, the rate of finding an extreme quantile under the misspecified model increases to 83 and 94 respectively out of 100 for the overall statistic and the statistic restricting to sample size less than or equal to 30.

Next we apply our proposed methodology to a few examples from the VRA application. In Table 4 we show the parametric bootstrap quantiles for lack of fit statistic for a LMG with a moderate number of areas that shows good fit properties. In the VRA application, we were particularly concerned with the fit of the model across all sample size groupings. As a result, we decided to calculate the fit statistics across sample size classes.

The next example (Table 5) shows the parametric bootstrap quantiles for lack of fit statistic for a LMG with a large number of areas and shows poor fit properties. Overall, across all sample

Table 4: Estimated Parametric Bootstrap Quantiles (Q) for the Lack of Fit Statistic (SS_g), VRA Example 1

Sample Size Grouping	CIT	LEP	ILL
Overall	0.420	0.244	0.104
(1,4)	0.868	0.222	0.047
(5,9)	0.164	0.553	0.831
(10,24)	0.240	0.145	0.635
25+	0.101	0.850	0.236

size categories, the quantile in the CIT column does not seem to show a lack of fit; however, when broken down by sample size we observe a serious lack of fit for each sample size. The lack of fit in different directions cancels out in the overall statistic. This is a good example why it would be important to break the lack of fit statistic down by sample size.

Table 5: Estimated Parametric Bootstrap Quantiles (Q) for the Lack of Fit Statistic (SS_g), VRA Example 2

Sample Size Grouping	CIT	LEP	ILL
Overall	0.419	0.996	0.000
(1,4)	1.000	0.012	0.019
(5,9)	0.002	0.943	0.3631
(10,24)	0.017	0.961	0.017
25+	0.003	0.998	0.000

3. Discussion

Small area estimation relies on correctly specified models, and the tools to check for model assumptions in such models are limited compared to other types of models, especially when the model is not a mixed-linear effect model. Misspecification in small area models can cause estimates that are severely biased or estimated prediction errors that are incorrect, ascribing either too much or too little uncertainty to the estimate.

In this paper we have tried to develop tools for model selection and model diagnostics for model validation specifically for models that are not of the form of a mixed-linear type. We utilize residuals defined as the difference between the observed direct estimates and the marginal mean estimate. These methods are a work-in-progress and further research is needed to make them more broadly applicable. Specifically, we would like to further compare the model selection cross validation method to other selection methods. Additionally, we would like to explore reference distributions for the cross-validation statistic through either a parametric bootstrap or asymptotic distribution. In both methods, we would like to explore different specifications for the weight function applied to the residuals $g(\cdot)$. The relative weight α_i from equation 5 or some function thereof might be useful. For the parametric bootstrap evaluation we would like to understand further the power of the test to find a lack of fit in order to help us refine the approach. Additionally, we would like to explore its use in a variety of distributional models. For additional information on the VRA application and estimates see `https://www.census.gov/rdo/data/voting_rights_determination_file.html`.

References

J. A. Calvin and J. Sedransk. Bayesian and Frequentist Predictive Inference for the Patterns of Care Studies. *Journal of the American Statistical Association*, 86(413):36–48, 1991.

J. Jiang, J. S. Rao, Z. Gu, and T. Nguyen. Fence Methods for Mixed Model Selection. *The Annals of Statistics*, pages 1669–1692, 2008.

P. M. Joyce, D. Malec, R. J. Little, A. Gilary, A. Navarro, and M. E. Asiala. Statistical Modeling Methodology for the Voting Rights Act Section 203 Language Assistance Determinations. *Journal of the American Statistical Association*, 109(505):36–47, 2014.

S. Müller, J. Scealy, and A. Welsh. Model Selection in Linear Mixed Models. *Statistical Science*, pages 135–167, 2013.

Z. Pan and D. Lin. Goodness-of-Fit Methods for Generalized Linear Mixed Models. *Biometrics*, 61(4):1000–1009, 2005.

J. N. Rao and I. Molina. *Small Area Estimation*. John Wiley & Sons, 2015.

C. Skinner. Cross Validation in Small Area Estimation: ESRC Full Research Report. *RES-000-22-1798*, 2007.

E. Slud and R. Ashmead. Hybrid BRR and Parametric-Bootstrap Variance Estimates for Small Domains in Large Surveys. In JSM Proceedings, Survey Research and Methodology Section. Alexandria, VA: American Statistical Association, 2017.

M. Tang, E. V. Slud, and R. M. Pfeiffer. Goodness of Fit Tests for Linear Mixed Models. *Journal of Multivariate Analysis*, 130:176–193, 2014.