Chapter: Use of Administrative Records in Small Area Estimation*

Andreea L. Erciulescu, Carolina Franco, Partha Lahiri †

November 1, 2018

1 Introduction

The steadily increasing demand for various socio-economic and health statistics for small geographical areas or geo-demographic groups has led to the implementation of small area estimation programs at some government agencies as well as to a flurry of research on related statistical methods.

Small area estimation typically involves the use of auxiliary information to improve upon traditional design-based methods for inference from survey data. These design-based methods rely only on the survey data from the domain of interest for inference about that domain, rather than seeking to model and exploit relationships between the survey data from all domains and the auxiliary information available. They are typically referred to in the small area literature as *direct* methods. The auxiliary information for small area estimation methods is often drawn from administrative records and censuses. In the literature, the term *small area* is described as a domain for which the survey data alone cannot provide reliable direct estimates (Rao

*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

†Andreea L. Erciulescu is a Research Associate at the National Institute of Statistical Sciences. Carolina Franco is a Research Mathematical Statistician at the U.S. Census Bureau. Partha Lahiri is a professor in the Joint Program in Survey Methodology and Department of Mathematics, at the University of Maryland, College Park.

and Molina, 2015) either because the sample size is too small or because there is no sample at all for the domain. Small area estimation refers to techniques to "borrow strength" from auxiliary variables, typically through modeling. Though this definition of small area estimation links it directly to the use of data from surveys, it is not necessary to have survey data to develop techniques for estimation of population characteristics of small domains. Some techniques that derive estimates from administrative records without using survey data will be mentioned briefly. A broader definition of small area estimation also includes these techniques. In this chapter, we adopt this broader definition, though we focus primarily on techniques using survey data and administrative records.

Administrative records arise from the operation of government programs, not for the purpose of estimating population characteristics. Hence, their content, coverage, accuracy, reference period, definition of variables, etc., are determined by their use in program operation, not by their use for statistical purposes. Nonetheless, valuable information for statistical inference can often be extracted from administrative records. Due to advances in computing, government agencies can process administrative records and link them with sample survey and census records for statistical purposes in a fraction of the time and costs required for field data collection. Brackstone (1987) discussed potential uses of administrative records in the production of a wide range of official statistics and pointed out their merits and demerits.

There are various ways administrative records can be used to produce small area statistics. There are methods that are purely based on administrative registers, commonly referred to as register-based methods, and use no survey data at all. Such use of administrative records in small area estimation can be traced back to eleventh century England and seventeenth century Canada; see Brackstone (1987). Zhang and Fosen (2012) constructed register-based small area employment rates and evaluated the progressive measurement errors in the small area estimates, using historic data from the Norwegian Employer Employee Register (NEER). For a good review of register-based small area methods, the readers are referred to Zhang and Giusti (2016).

Demographers have been using administrative records such as birth, death, migration, housing records, etc., in conjunction with the population census for estimating population for small areas for a long time now. For a comprehensive review of demographic methods for small area estimation, see Ghosh and Rao (1987) and Rao (2003).

Zanutto and Zaslavsky (2002) developed a small area method using census data and administrative records in conjunction with the nonresponse follow-up survey to impute small area detail while constraining aggregate-level estimates to agree with unbiased survey estimates. They applied their method in the 1995 U.S. Decennial Test Census where small area estimation was necessary because nonresponse follow-up was conducted in only a sample of blocks, leaving the data incomplete in the remaining blocks.

The above methods do not directly model survey data. Area-level and unit level models, which will be discussed in more length Section 3, use modeling to capture the relationship between the auxiliary data and the survey data. Such models have been widely studied in the literature and have been applied by government agencies, as will be seen in Section 3.

The accessibility of different administrative data from different sources has brought new opportunities for statisticians to develop innovative SAE methods that can cut down costs and improve the quality of estimates. Hundreds of papers were written in the last three decades and conferences and workshops are now being organized every year to disseminate research in the small area estimation research community. A recent comprehensive review of various methods in small area estimation can be found in Rao and Molina (2015). In this chapter, we attempt to illustrate the benefit of model-based methods that extract information from administrative records using sample survey and census data. In practice, there may be a need to use complex hierarchical models to capture spatio-temporal variations. We, however, stay with more basic models for the sake of simplicity in exposition.

The chapter is organized as follows. Section 2 discusses the preparation of data, including the identification and processing of administrative records for use in small area estimation models. Section 3 discusses small area estimation models, including both area level models in subsection 3.1 and unit-level models in subsection 3.2. Section 4 illustrates the concepts discussed in the previous sections with an application involving 1993 county poverty rates for school-aged children, with covariates drawn from adminsitrative records. Section 5 concludes, with exercises provided in Section 6.

2 Data Preparation

Small area estimation models are most effective when good auxiliary information is available–that is, additional information that can be used in conjuction

with the survey data to improve inference. This information is often drawn from administrative records, though censuses, other surveys, or past estimates from the same survey can also be used in modeling. The challenges in the identification and preparation of covariates from administrative records are not always emphasized in the literature. This section highlights some of the practical considerations for implementing a small area estimation program using administrative records. We focus here on the issues surrounding preparation of data from administrative records for use in small area estimation models, and on the qualities that make for good covariates. Examples from small area estimation programs used to produce official statistics are used as illustrations.

Data from administrative records often have the advantages that they cover large parts of the population and are relatively inexpensive. The data are collected for other purposes, so no additional cost for data collection or additional respondent burden need be incurred, although there are some costs for obtaining the proper agreements to use the data and for preparing them. However, administrative records may not accurately represent the population for which inference is desired, or measure the quantity of interest directly. For example, about 88% of the U.S. population is included in tax records from the Internal Revenue Service (IRS), the agency that collects taxes in the U.S. Selected data items of from tax records are provided to the Census Bureau for its use in statistical purposes [1]. These data can be used, for instance, to tabulate covariates that can aid in the estimation of poverty rates accross different geographic groups. Federal tax records have the advantage that the laws governing them are uniform across the country and hence there is consistency in the definition of tabulations obtained from them across different geographic areas. However, these records alone cannot be used to generate reliable estimates of poverty for two reasons. First, low income households are not required to file tax returns in the U.S. Second, the pseudo poverty rate that can be derived from IRS records differs in definition from that which can be obtained by household surveys (National Research Council, 2000a). Nonetheless, covariates derived from IRS records serve as very valuable predictors for estimating poverty in the U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) Program, which esti-

[1] These data are kept in the strictest confidentiality under the requirements in IRS publication 1075,"Tax Information Security Guidance For Federal, State, and Local Agencies" (`http://www.irs.gov/pub/irs-pdf/p1075.pdf`) as well as with the Census Bureau's own confidentiality standards

mates poverty rates for several age groups for the U.S. at the state, county, and school district levels. This is because covariates for use in small area models do not need to cover the entire target population or directly measure the quantity of interest. For a covariate to be useful in a small area estimation model it need only be strongly related to the quantity of interest and consistently defined across domains.

Inconsistencies in what data represent from place to place can lead to measurement error and to incorrect inference. When such inconsistencies are severe and cannot be adequately adjusted for, data from administrative records may not be suited for inclusion in a small area model. For instance, SAIPE considered the use of data from the National School Lunch Program, which provides free and reduced-price lunches (FRPL) for children in schools, for estimating poverty rates at the school district level. However, due to a high level of incompleteness of the data for many states, and inconsistencies about what quantities were reported by each state, the data were not included in estimation at the school district level. Although there were particular states for which the FRPL data were more accurate and complete, they could not be used for those states because of concerns about using different methodologies for different states, both due to the difficulty in implementation and perceived fairness (National Research Council, 2000a). Additional reasons for not including FRPL in SAIPE modeling are cited in Cruse and Powers (2006).

A key consideration in preparing administrative records data for use in small area estimation models is to what level the auxiliary data can be obtained. This often depends on confidentiality concerns. It may be possible to obtain the data at the unit level (i.e., person, family, household, or firm level), or it may only be possible to obtain aggregate summaries. It should be noted that most unit-level small area level models require that the unit-level covariates be known for both the sampled and non-sampled units in the population, so that just having unit-level covariates for the sample will typically not suffice. In the area-level case, when the auxiliary variables come in the form of estimates from other surveys, care must be taken to account for their sampling error–see Section 3.1.

Linking or matching data from administrative records to the survey data poses challenges. Exact matching involves linking two records from the same unit, whereas statistical matching involves linking files based on similar characteristics. The error due to linking can be hard to ascertain, though a measure of error is available for many statistical and exact matching techniques

(Winkler, 2007). The error due to linkage is typically not incorporated in small area models.

For area-level models, which model aggregate data summaries for the domains of interest rather than unit level data, it is not necessary to match at the unit level. However, computing such summaries sometimes involves allocating units to domains. This can be challenging for smaller areas of aggregation. For instance, geocoding involves identifying addresses with geographic locations. Addresses from the IRS need to be geocoded so that they can be allocated to the small areas of interest to produce summary statistics at the state, county, or school-district levels for small area models. Geocoding of tax records from the IRS to sub county areas can be difficult, especially in some rural areas. This is primarily because some rural addresses are not in city-style format–i.e. street number and street name. At the inception of the SAIPE program, the Census Bureau had not yet developed accurate geocoding for sub-county areas, so the IRS administrative records were not used for the estimation of poverty at the school district level (National Research Council, 2000a). Starting with the 2005 data, SAIPE altered its methodology at the school-district level to incorporate IRS data (Bell et al., 2016). There were still many tax exemptions that could not be geocoded, so Maples and Bell (2007) developed an algorithm to allocate these among school districts. Maples (2008) developed a methodology for estimating associated coefficients of variation.

Timing plays a big role in the inclusion of administrative records into small area models. One important consideration is the frequency with which the administrative records are produced, and whether that meets the needs of the small area estimation program. Usually, administrative records are released periodically– often annually or even less frequently– and sometimes data are updated later to make additions and/or corrections. Ideally, data from administrative records should cover a similar time frame as the survey of interest. However, this is not always possible, since often the reference periods of both sources are different. For instance, for the current SAIPE methods for estimating poverty at the state, county, and school district levels, the primary data source, the American Community Survey (ACS), collects data over the course of a year and asks individuals about their income in the 12 months preceding the time of response, spanning 23 months of income overall. IRS records, on the other hand, refer to the income in a particular calendar year (Luery, 2010).

Time delays in the release of administrative records can affect the timing

of the release of the related small area estimates and/or the choice of which time period to use for the covariate from administrative records. In some cases, a decision is made to use lagged administrative records in order to expedite production of small area estimates. The Census Bureau's Small Area Health Insurance Estimates (SAHIE) Program estimates numbers and proportions of health insurance coverage by counties and states. The SAHIE program uses Medicaid records from the Centers for Medicare and Medicaid Services (CMS) as well as Childrens Insurance Program (CHIP) participation counts obtained from states and counties. Historically, SAHIE models had a one to two year time lag in their Medicaid/CHIP covariates. In 2013 and 2014, however, many states expanded their Medicaid eligibility due to the enactment of the Patient Protection and Affordable Care Act (ACA). In response, for its 2014 estimates SAHIE started projecting the administrative records to the year of estimation. For comparability, SAHIE re-released estimates for 2013 using the same methodology (see Powers et al., 2016 and Bauder et al., 2018).

Once some potential sources of administrative records have been identified for inclusion in small area models, relevant covariates can be derived, typically from aggregate summaries or transformations of the data. For instance, from the IRS data the SAIPE program computes a tax return pseudo-poverty rate for children for each state, computed as the number of child exemptions for returns determined to be in poverty, divided by the total number of child tax exemptions. It also computes an estimate of the proportion of people who do not file taxes and are under the age of 65 (i.e., the "non-filer rate").

When there are known inconsistencies for covariates among domains, whenever possible they should be addressed. The Supplemental Nutritional Assistance Program (SNAP, formerly known as the Food Stamp Program) provides subsidies for low-income households for food purchases. The SNAP eligibility criteria are broadly the same for all states except Hawaii and Alaska, which include some individuals with higher household incomes. For these two states the Census Bureau adjusts the data to exclude recipients who would not be eligible under the other states' criteria. It makes other adjustments for monthly outliers based on time series analysis. Some of these outliers arise from issuance of emergency SNAP benefits in response to natural disasters, particularly hurricanes. SNAP data are used to derive predictors for both the SAIPE state and county models and for SAHIE models.

In some cases, many alternative possible covariates from administrative

records are available, and extensive analysis is needed to find the covariates that will be used in the final model. This is usually done with model selection tools such as the Akaike Information Criterion (AIC), checks of significance of coefficients, residual analysis, and other model diagnostics. For poverty mapping for Chilean comunas, the Ministerio de Desarrollo (Ministry of Development) had many options to choose from for auxiliary variables. In order to select the variables included in the final model, the ministry used a stepwise procedure along with model diagnostics from the statistical software Stata (Casa-Cordero Valencia et al., 2016). Before the statistical analysis was performed, the initial pool of potential variables was first narrowed down by subject matter experts that helped determine the relevance and reliability of the auxiliary data. Timeliness of the records was also an important consideration. This illustrates an important point–it is important to ascertain the quality of the administrative records before considering them for model inclusion. Even a statistically significant covariate may have measurement error, and this can lead to errors in inference.

After covariates from administrative records have been identified and selected for use in a small area estimation model, their quality should be periodically re-evaluated. External factors such as changes in legislation or program administration may affect the predictive ability of a covariate over time. An example was mentioned above related to the effect of changes of healthcare legislation on the Medicaid/CHIP administrative records. Another example relates to the use of the SNAP covariate in SAIPE state models for poverty. In 1997, the Welfare Reform Act went into effect, which among other things, gave states more freedom to administer the SNAP program (which prior to 2008 was called the Food Stamp Program). This might have led to inconsistencies in the administration of the program that reduced the comparability of the SNAP data across states. That year, the SNAP covariate became statistically insignificant in the state poverty models. SAIPE continued to monitor the significance of the covariate, and after it was also insignificant for 1998, the covariate was removed. The covariate eventually regained its significance and was reintroduced for the 2004 estimates (Bell at al., 2016).

In addition to periodically evaluating existing models, statisticians should also always be looking for new sources of covariates from administrative records. We should emphasize that administrative records for use in small area estimation will always have some error. The statistician should attempt to reduce the sources of error as much as possible, and exclude covariates from administrative records that are of poor quality and/or are not consistent in

what they measure across domains, as was illustrated in this section. Errors in covariates that can be estimated or modeled could be incorporated into small area estimation models. In most cases, it is hard to estimate the error inherent in administrative records, whether the error is considered random or deterministic. This will be discussed more in Section 3.

3 Small area estimation models for combining information

We can envision different possible situations related to the availability of administrative data. For example, data from the administrative records can be available in one or any combination of the following forms: (i) summaries available for geographical areas that contain one or more small areas, (ii) summaries for the small areas, (iii) summaries available for geographical areas fully contained in the small areas of interest, (iv) unit level records. Hierarchical modeling is not necessarily simple, but it is flexible in combining information from different sources such as surveys, administrative records and census data that are available at different levels of geography. The exact nature of hierarchical modeling depends on the nature and availability of data from different sources. In this section, we discuss the modeling and estimation, and we comment on possible issue(s) associated with the use of administrative data in small area estimation.

3.1 Area-level models

Area level models improve upon direct survey estimators by assuming a relationship between the small area parameters of interest and covariates via hierarchical models. They have the advantage that only area-level aggregate summaries of the covariates from administrative records are needed – unit level covariates either for the sample or the population need not be available. Some early applications of area level models in small area estimation can be found in Efron and Morris (1973, 1975) and Carter and Rolph (1974). These authors used area level covariates implicitly in forming groups of similar small areas. However, they did not explicitly use any area level covariate in modeling and did not use complex survey data.

Let θ_i be the population characteristic of interest for area i, and y_i the direct survey estimate of θ_i, with sampling variance D_i $(i = 1, \cdots, m)$. In

the context of estimating per-capita income for small places (population less than 1,000), Fay and Herriot (1979) considered the following generalization of the Efron-Morris and Carter-Rolph Bayesian models:

$$\text{Level 1 (Sampling Distribution): } y_i \overset{ind}{\sim} N(\theta_i, D_i), \tag{1}$$

$$\text{Level 2 (Prior Distribution): } \theta_i \overset{ind}{\sim} N(x_i'\beta, A), \tag{2}$$

where x_i is a $p \times 1$ vector of known auxiliary variables derived from administrative records or other sources; D_i is the known sampling variance of y_i; β is a $p \times 1$ vector of unknown regression coefficients and A is an unknown prior variance ($i = 1, \ldots, m$). The Fay-Herriot model is a special case of a multilevel or hierarchical model where the first level captures the errors of the direct survey estimates y_i due to sampling and the second level, often called the linking model, links the true small area parameters θ_i to a set of auxiliary variables. In practice, the D_i's need to be estimated, typically using survey data. This is indeed one of the most challenging problems in the application of an area level model. We will discuss estimation of D_i subsequently.

The Fay-Herriot model given by (1) and (2) is a special case of the general linear mixed model (see, e.g., equation (5.2.1), page 98, of Rao and Molina 2015), and can be expressed as

$$y_i = \theta_i + e_i = x_i'\beta + v_i + e_i, \tag{3}$$

where model or linking errors $\{v_i\}$ and sampling errors $\{e_i\}$ are independent with $v_i \overset{iid}{\sim} N(0, A)$ and $e_i \overset{iid}{\sim} N(0, D_i)$ $(i = 1, \cdots, m)$. Note that v_i can be viewed as a leftover random effect due to area i that is not explained by the auxiliary variables.

The Best Predictor (BP) of θ_i when the parameters are known is obtained by minimizing the mean squared prediction error (MSPE) defined as $\text{MSPE}(\hat{\theta}_i) = E(\hat{\theta}_i - \theta_i)^2$, where the expectation E is with respect to the linear mixed model (3). The BP of θ_i is given by

$$\hat{\theta}_i^{BP} = (1 - B_i)y_i + B_i x_i'\beta, \tag{4}$$

with

$$MSPE(\hat{\theta}_i^{BP}) = (1 - B_i)D_i = g_{1i}(A), \text{ say}, \tag{5}$$

where $B_i = D_i/(D_i + A)$ is known as the shrinkage factor. For domains or small areas with smaller sampling variances, more weight is placed on the

direct estimator y_i. Note that, under the squared error loss function, the Bayes estimator of θ_i, that is, the conditional mean of θ_i given y_i (known as the posterior mean of θ_i) is identical to the BP of θ_i. Moreover, the associated measure of uncertainty for the Bayes estimator, that is, the conditional variance of θ_i given y_i (known as the posterior variance of θ_i) is identical to $\text{MSPE}(\hat{\theta}_i^B)$.

When β is unknown but A is known, the Best Linear Unbiased Predictor (BLUP) of θ_i is obtained by minimizing the $\text{MSPE}(\hat{\theta}_i)$ among all linear unbiased predictors, that is, predictors of the form $\hat{\theta}_i = \sum_{j=1}^m l_{ij} y_j$ that satisfy the unbiasedness condition: $\text{E}(\hat{\theta}_i - \theta_i) = 0$, where the expectation is with respect to the linear mixed model (3). Using Henderson's theory (Henderson 1953), the BLUP of θ_i can be obtained as

$$\hat{\theta}_i^{BLUP} = (1 - B_i) y_i + B_i x_i' \hat{\beta}^{WLS} \tag{6}$$

where $\hat{\beta}$ is the weighted least square estimator of β given by

$$\hat{\beta}^{WLS} \equiv \hat{\beta}^{WLS}(A) = \left(\sum_{j=1}^m (1 - B_j) x_j x_j' \right)^{-1} \sum_{j=1}^m (1 - B_j) x_j y_j.$$

Note that (6) is identical to (4) except with β replaced by $\hat{\beta}^{WLS}$. It is straightforward to show that

$$\text{MSPE}(\hat{\theta}_i^{BLUP}) = g_{1i}(A) + g_{2i}(A), \tag{7}$$

where $g_{2i}(A) = B_i^2 x_i' \left(\sum_{j=1}^m (1 - B_j) x_j x_j' \right)^{-1} x_i$ is the additional variability in BLUP that is due to the estimation of β. Under standard regularity conditions, $g_{1i}(A) = O(1)$, but $g_{2i}(A) = O(m^{-1})$, for large m. Thus, in the standard small area higher-order asymptotic sense, $g_{1i}(A)$ contributes more to $\text{MSPE}(\hat{\theta}_i^{BLUP})$ than $g_{2i}(A)$ does.

Note that the normality of the linking and sampling models is not needed to justify (6) as the BLUP of θ_i, though it is required to justify (4) as the BP. Under normality and assuming a non-informative flat prior on β, $\hat{\theta}_i^{BLUP}$ and $\text{MSPE}(\hat{\theta}_i^{BLUP})$ are identical to the hierarchical Bayes estimator of θ_i and the corresponding posterior variance, respectively.

In practice, A is unknown. In small area applications, different estimators of A such as ANOVA, the Fay-Herriot method of moments, maximum likelihood and residual maximum likelihood methods have been considered. An

Empirical Best Linear Unbiased Predictor (EBLUP), say $\hat{\theta}_i^{EBLUP}$, is obtained when A is replaced by an estimator, say $\hat{A}$. Notice that $\hat{\theta}_i^{EBLUP}$ is a weighted average of the direct estimator y_i and the regression synthetic estimator $x_i'\hat{\beta}$, where $\hat{\beta} = \hat{\beta}^{WLS}(\hat{A})$. Different well-known estimators of A are equivalent in terms of mean squared error (MSE), up to the order $O(m^{-1})$. However, in the higher-order asymptotic sense, the REML estimator of A has the least bias ($o(m^{-1})$), compared to the bias of other standard estimators ($O(m^{-1})$).

It is well-known that the REML, ML and method-of-moments estimators of A can yield zero estimates, in which case the EBLUP reduces to the regression synthetic estimate for all areas; in other words the direct estimate y_i does not get any weight in the EBLUP formula, even for the largest area. In order to get around this problem, Yoshimori and Lahiri (2014a), building on earlier papers by Lahiri and Li (2009) and Li and Lahiri (2010), proposed the following general class of adjusted maximum likelihood estimators of A that includes most of the standard likelihood-based estimators of A available to date:

$$\hat{A}_h = \text{argmax}_{A\in[0,\infty]} h(A) L_{RE}(A),$$

where $L_{RE}(A)$ is the residual likelihood of A and $h(A)$ is a general adjustment term. They suggested a choice of $h(A)$ that produces a strictly positive estimate of A while maintaining the same bias and mean squared error properties of REML, up to the order $O(m^{-1})$.

The problem of finding an accurate estimator of the MSPE of EBLUP that captures additional variability due to the estimation of A is a challenging problem. An estimator $\widehat{\text{MSPE}}_i$ of MSPE_i is called second-order unbiased or nearly unbiased if $E(\widehat{\text{MSPE}}_i - \text{MSPE}_i) = o(m^{-1})$, under regularity conditions. For the ANOVA estimator, $\hat{A}_{ANOVA}$ of A, Prasad and Rao (1990) showed that

$$\text{MSPE}(\hat{\theta}_i^{EBLUP}) = g_{1i}(A) + g_{2i}(A) + g_{3i}(A) + o(m^{-1}), \tag{8}$$

where $g_{3i}(A) = \frac{2D_i^2}{(A+D_i)^3}\text{AVar}(\hat{A}_{ANOVA})$, where $\text{AVar}(\hat{A}_{ANOVA})$ is the asymptotic variance of $\hat{A}_{ANOVA}$, up to the order $O(m^{-1})$. Interestingly, (8) holds for the general class of adjusted maximum likelihood estimators of A proposed by Yoshimori and Lahiri (2014a), under regularity conditions. Prasad and Rao (1990) noticed that a second-order unbiased estimator of $\text{MSPE}(\hat{\theta}_i^{EBLUP})$ is not obtained if we substitute $\hat{A}$ for A. By correcting the bias of $g_{1i}(\hat{A}_{ANOVA})$, up to the order $O(m^{-1})$, they obtained the following second-order unbiased

estimator of MSPE($\hat{\theta}_i^{EBLUP}$):

$$\widehat{\text{MSPE}}(\hat{\theta}_i^{EBLUP}) = g_{1i}(\hat{A}) + g_{2i}(\hat{A}) + 2g_{3i}(\hat{A}) + o(m^{-1}), \tag{9}$$

The same formula works for the REML estimator of A and the adjusted maximum likelihood estimator of Yoshimori and Lahiri (2014a), but we need additional bias corrections for the ML and adjusted maximum likelihood method of Li and Lahiri (2010). Second-order unbiased estimators based on jackknife and parametric bootstrap methods have been also proposed; see Jiang and Lahiri (2006) and Rao and Molina (2015), for a comprehensive review and comparison of these MSPE estimators.

Yoshimori and Lahiri (2014b) broadened the class of adjusted maximum likelihood estimators by introducing an area specific adjustment term $h_i(A)$, and proposed a simple second-order efficient prediction interval for θ_i of the form: $\hat{\theta}_i^{EBLUP}(\hat{A}_{h_i}) \pm z_{\alpha/2}\sqrt{g_{1i}(\hat{A}_{h_i})}$, where $z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal deviate, and $\hat{\theta}_i^{EBLUP}(\hat{A}_{h_i})$ is the EBLUP obtained from the BLUP when A is replaced by $\hat{A}_{h_i}$. The average length of this prediction interval is always smaller than that of the confidence interval based on the direct estimator and the coverage error is of the order $o(m^{-1})$, lower than the empirical Bayes confidence interval proposed by Cox (1976). Parametric bootstrap prediction intervals of θ_i based on EBLUP are also available; see Chatterjee et al. (2008) and Li and Lahiri (2010). Such prediction intervals have the same order of coverage error, but they are computationally intensive. Moreover, it is not known if the average length of a parametric bootstrap prediction interval is smaller than that of the direct confidence interval. Hall and Maiti (2006) also proposed a parametric bootstrap confidence interval, but it is based on a synthetic estimator.

One can assign prior distributions to β and A for a hierarchical Bayesian approach. If no prior information is available, then these parameters would typically be assigned non-informative priors (see, for instance, Berger 1985). Datta et al. (2005) and Ganesh and Lahiri (2008) considered non-informative priors for A with good frequentist properties that yield a proper posterior distribution of θ_i. Note that the posterior mean of A, an estimator of A under the hierarchical Bayesian approach, is strictly positive like the adjusted maximum likelihood estimators. One advantage of the hierarchical Bayesian approach is that it can capture different sources of uncertainties. Hierarchical Bayes implementations for the Fay-Herriot model can be implemented by numerical integration, Monte Carlo, Markov Chain Monte Carlo (MCMC)

or by certain approximations such as Laplace approximation or Adjustment for Density maximization (ADM); see Datta (2009), Morris and Tang (2011), Rao and Molina (2015).

Although many developments in area-level models have occurred since Fay and Herriot's seminal 1979 paper, the Fay-Herriot model is still quite useful in practice and in fact it is used for the production of official small area statistics by government agencies. For instance, SAIPE uses a hierarchical Bayes implementation for the production of state level poverty estimates. SAIPE switched from a frequentist method to a Bayesian method because estimation of the model variance, which was previously done via maximum likelihood, sometimes resulted in estimates of zero, which would imply that all the weight in EBLUP is placed on the synthetic estimate for all areas, an undesirable result. For estimation at the county level, SAIPE uses a Fay-Herriot model on log-transformed estimates of poverty counts, where the estimation of the parameters is done by maximum likelihood via an iterative method which alternates between estimating A via maximum likelihood and β via weighted least squares. Poverty mapping of Chilean comunas by the Ministerio de Desarrollo is also done via a Fay-Herriot model, featuring a variance stabilizing transformation. The issue of the potential for zero estimates for the model variance can be handled by using the adjusted maximum likelihood estimator of Li and Lahiri (2010).

Area level summaries from administrative records could be potentially useful in a situation where no survey data are available for some of the areas. For example, the CPS sample design generally does not produce any data for a majority of U.S. counties, and so EBLUP methodology as described above cannot be used to draw inference for these counties. However, from the EBLUP methodology one may derive synthetic estimates for counties with no data, using only the regression term of (6). Back in the 1990's, the U.S. Census Bureau found using administrative record covariates useful in small area estimation models for poverty. This permitted estimates for counties with no survey data. This implicitly assumes that regression coefficients for the linking model do not change across counties. A possible alternative solution might be to incorporate a spatial correlation into the Fay-Herriot linking model (see Vogt 2010). An estimate derived from such a model not only uses administrative data summaries but also uses survey estimates from neighboring areas. However, with this appeal of the spatial models comes the complexity in defining spatial neighborhood and in estimation of spatial correlation. More general spatial models have been used in the small area

literature (see, e.g., Rao and Molina 2015). However, the potential utility of spatial models to improve small area estimation for areas with no survey data needs further evaluation.

The sampling variance D_i is assumed to be known, but must be estimated in practice. Often, direct estimates of the sampling variance are available from the survey used as the primary data source. However, these direct estimates may be unreliable, or in some cases unavailable. In such cases alternative estimates must be explored. Some approaches for this are the use of generalized variance functions (GVF's, see Wolter, 2007), or simpler approaches that make assumptions about uniformity over larger levels of aggregation to obtain "smoothed" sampling variance estimates for small areas. As an example of the former, at the inception of SAIPE, the primary data source for annual poverty estimates was the Current Population Survey (CPS), later to be replaced by the American Community Survey (ACS), a survey with much larger sample size. Sampling variances of CPS estimates at the state-level were produced but were based on small sample sizes, and at the county-level direct sampling variance estimates were not produced. For this reason, SAIPE obtained sampling variances estimates via GVF's – for the state level, a GVF with random effects was developed by Otto and Bell (1995), and for the county level, a much simpler GVF was developed (Bell, 2016). For poverty mapping for Chilean comunas, the smallest territorial entity in Chile, the use of a variance-stabilizing transformation eliminated the need for sampling variance estimates, but estimates of the design effect were needed to compute the effective sample sizes for each small area. These were computed at the regional level to avoid unstable estimates at smaller levels of aggregation (Casa-Cordero Valencia et al., 2016), with the underlying assumption that these estimates were also valid under lower levels of aggregation.

We note that administrative data could be potentially used in estimating D_i for small areas. For example, one can use area level summaries from administrative records as covariates in the synthetic estimators of sampling variances of small area proportions; see Liu (2009). The method first fits the following logistic model using data only from large areas to obtain stable estimates of model parameters:

$$\text{logit}(p_{iw}) = x_i'\beta + \epsilon_i,\ i = 1, \cdots, I, \tag{10}$$

where for the ith large area p_{iw} is a direct estimate of proportion P_i and x_i is a vector of known covariates available in area i and $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ ($i =$

$1, \cdots, I$). Using an estimate $\hat{\beta}$ of β using data from the large areas, synthetic estimates of P_i for all areas are obtained as:

$$\tilde{p}_{isyn} = \frac{\exp(x_i'\hat{\beta})}{1 + \exp(x_i'\hat{\beta})},\ i = 1, \cdots, m. \tag{11}$$

Finally, smoothed estimates of D_i's are obtained as:

$$\hat{D}_{i;syn} = \frac{\tilde{p}_{isyn}(1 - \tilde{p}_{isyn})}{n_i}\text{deff}_{iw}, \tag{12}$$

where deff_{iw} is a suitable approximation of the design effect for area i (e.g., design effect estimate for a large area containing the small area i).

In order to capture variation due to the estimation of sampling variances D_i and at the same time to obtain smoothed sampling variances, Liu (2009) considered a number of integrated hierarchical models for proportions, including non-normal models. In the context of estimating small area proportions, one such integrated Bayesian model – an extension of the Fay-Herriot model – is given by:

$$\text{Level 1 (Sampling Distribution):}\ y_i \overset{ind}{\sim} N\left(\theta_i, \frac{\theta_i(1-\theta_i)}{n_i}\text{deff}_i\right) \tag{13}$$

$$\text{Level 2 (Prior Distribution):}\ \theta_i \overset{ind}{\sim} N(x_i'\beta, A). \tag{14}$$

Assuming a non-informative prior on the hyperparameters β and A and using Markov Chain Monte Carlo (MCMC), one can easily produce the hierarchical Bayesian estimate of P_i and sampling variance of y_i as the posterior means of P_i and $\frac{\theta_i(1-\theta_i)}{n_i}\text{deff}_i$, respectively. Liu (2009) (also see Liu et al. 2014) evaluated design-based properties of such hierarchical Bayesian methods using Monte Carlo simulations. Ha et al. (2014) compared a number of integrated hierarchical models in the context of estimating smoking prevalences in the U.S. states. An early example of an integrated hierarchical modeling approach in small area estimation for continuous variables can be found in Arora and Lahiri (1997). For a review of variance modeling, readers are referred to Hawala and Lahiri (2018).

The research on area-level models has been extensive. We describe here a few extensions, though many others have been proposed in the literature. For more information on area-level models, see also Molina and Rao (2015)

or Pfeffermann (2013). Some extensions to area-level modeling are those that use Generalized Linear Mixed Models, which can depart from the assumptions of normality in the original Fay-Herriot model. These can be appropriate in cases where the data are discrete, skewed, or when the model errors are thought to be heteroscedastic. Small area estimation models via GLMMs have been discussed by Ghosh et al., (1998) and Rao and Molina (2015), among others. In the context of estimating poverty rates for Chilean comunas, Ha (2013) developed an ADM approximation to the posterior distribution of small area proportions using a hierarchical Binomial-Beta model that combines information from area level summary statistics derived from different administrative records and complex survey data. Franco and Bell (2015) discuss a binomial-logit normal model with multivariate and time series extensions.

Another type of extension is subarea multi-level models, where each area is divided into subareas, and the interest is in prediction both at the area and subarea level. Fuller and Goyeneche (1998) introduced subarea-level models with an application to SAIPE, where the subareas were the counties, nested within states (areas). Subarea-level models have also been studied by Torabi and Rao (2014), Rao and Molina (2015), and Kim et al. (2018). Erciulescu et. al (2018) developed a hierarchical Bayes subarea-level model for harvested acreage of crop commodities in U.S. counties (sub-areas) and agricultural statistics districts (areas), with choices of covariates that included administrative acreage data.

One straightforward but useful extension is the multivariate version of the Fay-Herriot model. Multivariate Fay-Herriot models have been explored, for instance, by Fay (1987), Datta, Fay, and Ghosh (1991), and Bell and Huang (2012), among others. Multivariate area-level models allow for the joint modeling of related characteristics. These could be different estimates from the same survey or from different surveys. By jointly modeling related characteristics, it is possible to improve the estimation by exploiting the correlation among them. Moreover, when attempting to estimate functions of more than one of the responses, jointly modeling properly accounts for the correlations among model errors. For instance, one may want to estimate the year to year change in a poverty rate, or some other characteristic based on estimates of the same survey for two consecutive years (e.g., Arima et al. 2017).

When jointly modeling consecutive estimates from several years of data collection, one may also choose to use time series extensions to area-level

models. Time series extensions of the Fay-Herriot model are discussed in Rao and Yu (1994), Ghosh et al. (1996), Datta et al. (1999), and Rao and Molina (2015). Another growing area of research within small area estimation is spatial and spatio-temporal models, which exploit the dependence among data points across space. Note that space does not necessarily need to be defined as geographic distance. See, for instance, Esteban et. al (2016), Rao et. al (2016). These models may be explored to find alternative solutions to the situation when survey data are unavailable for some areas for some time points, as mentioned earlier.

In many area-level applications, summaries from different administrative databases are used as covariates. These are likely to be subject to measurement error, where such error may be defined as the difference between the summary statistic and the value it is intended to measure. Ybarra and Lohr (2008) stimulated a flurry of research activity on extending small area methodologies to account for errors in the covariates when these errors are random, as may arise when the covariates come from other surveys. In view of this recent research productivity, one may naturally ask whether these ideas can be applied to account for measurement error in covariates arising from administrative records. To discuss this question, we assume that the Fay-Herriot model (3) is the true model, but that we do not observe x_i, but rather a noisy estimate X_i of x_i. Suppose that $X_i = x_i + \eta_i$, $\eta_i \sim N(0, C_i)$. Under these assumptions, two types of measurement error have been discussed in the literature: *functional* (e.g., Ybarra and Lohr, 2007) and *structural* (e.g., Bell et al 2017). The former assumes that the true x_i is a fixed but unknown quantity. The latter assumes that x_i is random and follows a model.

A naive model would simply ignore measurement error and treat the X_is as if they were the true x_is. The resulting predictor would be of the form

$$\hat{\theta}_i^{NV} = (1 - \hat{B}_{i,NV})y_i + \hat{B}_{i,NV}X_i'\hat{\beta}_{NV}. \tag{15}$$

where B_i is the same as in (4). The estimators of β and B_i are given the subscripts NV because the model parameter estimates obtained assuming the naive model is true do not converge to the true β and A in (3) .

The implied model in expression (15) models the relationship between the true quantities θ_i and the noisy estimates X_i of the true quantity x_i, rather than modeling the relationship between the two true quantities θ_i and x_i, which is more reasonable. Prediction results will differ between the naive models and measurement error models, except in special cases. The naive

model also assumes homoscedasticity of the residuals $y_i - X_i'\beta$, which will not hold when the C_i's differ accross domains. Bell et. al (2017) compare the effects on MSPE's from using the naive, functional, or structural models, assuming either the structural or functional models are true.

Expressions for Empirical Best Linear Unbiased Predictors (EBLUPs) for functional and structural measurement error models can be found, for instance, in Ybarra and Lohr (2008) and Bell et. al. (2017). We now point out a few challenges in implementing these predictors when the X_i are drawn from adminstrative records. First, applying measurement error models requires estimates of the C_i. When the X_i's are derived from administrative records, it is not clear how one should define and estimate C_i in the presence of a multitude of possibly unknown sources of errors. Holt (2007) noticed a clear lack of statistical theories for assessing the uncertainty of register-based statistics. Zhang (2012) discussed some possible theories for statistics based on registers, but did not put forward any concrete suggestion for defining the MSE of register-based statistics. Secondly, confidentiality aspects of most administrative data in the U.S. limits the prospect for advancement of research in understanding the theory of statistics derived from administrative data. Finally, though measurement error models assume C_i is known, their optimality properties are not guaranteed when C_i is estimated.

The literature on measurement error can shed some light on what happens when there is random error in covariates derived from adminsitrative records provided that the measurement error model, whether functional or structural, is reasonable for the inherent measurement error in the application in question. Bell et al. (2017) show that the naive model discussed above, which ignores measurement error, can lead to misstated mean squared errors when either the structural or functional measurent error models are true. Exceptions to this arise when the structural measurement error model is true and occur for areas where $C_i = \bar{C}$, where $\bar{C}$ is the mean of the C_is, in which case the predictions are the same for the structural and naive models. It follows that if C_i is constant for all areas then the naive and the structural measurement error predictors give the same results for all areas, but in other cases the results differ. However, the measurement error models discussed above cannot currently be used to correct for measurement error in the covariates drawn from administrative records even when it is reasonable to assume a measurement error model holds, due to the difficulties in determining the C_is.

3.2 Unit-level models

As noted in the last section, uncertainty due to estimation of the sampling variances of direct estimators cannot be incorporated in an area level model without additional information on sample design and within area variation information. In order to capture this additional uncertainty, we need observations within each area. Let n_i be the sample size for the ith area and y_{ij} be the value of the study variable for the jth unit (e.g., person, household, farm, etc.) of the ith small area population ($i = 1, \cdots, m; \quad j = 1, \cdots, n_i$). One possible model is the following:

$$\text{Level 1 (Sampling Distribution): } y_{ij}|\theta_i \overset{ind}{\sim} N(\theta_i, \sigma_e^2), \tag{16}$$

$$\text{Level 2 (Prior Distribution): } \theta_i \overset{ind}{\sim} N(x_i'\beta, \sigma_v^2), \tag{17}$$

$i = 1, \cdots, m; \quad j = 1, \cdots, n_i$, where x_i is a $p \times 1$ vector of area specific known auxiliary variables (some of them could be summaries from administrative records); σ_e^2 and σ_v^2 are within and between area unknown variance components, respectively.

The model described by equations (16)-(17) can be written in the following linear mixed model form:

$$y_{ij} = x_i'\beta + v_i + e_{ij}, \; i = 1, \cdots, m; j = 1, \cdots, n_i, \tag{18}$$

where area specific random effects $\{v_i\}$ and the random error $\{e_{ij}\}$ are independent with $v_i \sim N(0, \sigma_v^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$. We are interested in predicting the mixed effect $\theta_i = x_i'\beta + v_i$. Note that if σ_e^2 is known, we can equivalently use the area level model, discussed in the previous subsection, on the sample mean $\bar{y}_i = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$ with $D_i = \sigma_e^2/n_i$ and $A = \sigma_v^2$.

The two-level model (16)-(17) or the linear mixed model (18) allows for estimation of both the variance components and accounts for uncertainty due to estimation of both the variance components. One advantage of the model is that we do not need an auxiliary variable information at the unit level, which would require complex linking procedures, especially when the auxiliary information comes from administrative records.

The exchangeability assumption in the sampling error model (16), however, is rather strong. It is possible to relax the modeling assumption if the auxiliary variables are available for all sampled units and some summary information about the auxiliary variables is available for the non-sampled units in each small area. Such a situation was considered by Battese et al. (1988).

They proposed an EBLUP method to predict areas under corn and soybean for 12 counties of north-central Iowa using the 1978 June Enumerative Survey (JES) and satellite (LANDSAT) data.

Although satellite data may not be categorized as administrative records, it will be instructive to discuss how Battese et al. (1988) linked LANDSAT data to survey data. The unit for recording the satellite information is a pixel (about 0.45 hectares), a term used for "picture element", and the unit of measurement for the survey data is a farmer. Thus the units of measurements for these two databases are different, which could often be the case when survey data are to be linked with administrative records. Before implementing an EBLUP methodology, a data preparation step was thus needed to define a common unit for which aggregates from both survey and satellite data can be obtained.

Battese et al. (1988) considered segment (about 250 hectares), the primary sampling unit (PSU) of the JES, as the common unit for the two databases. The areas under corn and soybean for each sampled segment were determined by the USDA Statistical Reporting field staff by interviewing farm operators. Using USDA procedures, recordings from LANDSAT during August and June 1978 were used to classify crop cover for all pixels in 12 counties, and this information was used to obtain areas under corn and soybean for all sampled and non-sampled segments in the 12 counties. There can be, however, errors in assigning pixels to the right segment or classifying a pixel to the right crop. There could also be potential errors in compiling data at the segment level from the information obtained from the farmers in JES. Battese et al. (1988) ignored such possible errors in developing their EBLUP method.

Let N_i be the population size and y_{ij} be the value of the study variable for the jth unit of the ith small area population ($i = 1, \cdots, m; \;\; j = 1, \cdots, N_i$). Suppose we are interested in estimating the finite population mean $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$, or, equivalently, the finite population total $N_i \bar{Y}_i$ when N_i is known. In Battese et al. (1988), N_i is the number of segments and $\bar{Y}_i$ is the average hectare of crop per segment for county i, the parameter of interest.

Battese et al. (1988) considered the following linear mixed model, commonly referred to as a nested error regression model:

$$y_{ij} = x_{ij}'\beta + v_i + e_{ij}, \tag{19}$$

$i = 1, \cdots, m; \;\; j = 1, \cdots, N_i$, where x_{ij} is a $p \times 1$ column vector of known

auxiliary variables; $\{v_i\}$ and $\{e_{ij}\}$ are all independent with $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$ and $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$. Thus they assumed the model holds for all the units of the finite population.

We can also write the model as a two-level model:

$$\text{Level 1: } y_{ij}|v_i \overset{ind}{\sim} N(x_{ij}'\beta + v_i,\ \sigma_e^2) \tag{20}$$

$$\text{Level 2: } v_i \overset{iid}{\sim} N(0,\ \sigma_v^2), \tag{21}$$

$i = 1, \cdots, m; j = 1, \cdots, N_i$.

A model-dependent estimator of $\bar{Y}_i$ can be written as

$$\hat{\bar{Y}}_i = f_i \bar{y}_i + (1 - f_i)\hat{\bar{Y}}_{ir}, \tag{22}$$

where $f_i = n_i/N_i$, the sampling fraction, and $\hat{\bar{Y}}_{ir}$ is a model-dependent predictor of $\bar{Y}_{ir} = (N_i - n_i)^{-1} \sum_{j \notin s_i} y_{ij}$, with s_i being the sample for area i $(i = 1, \cdots, m)$. For the Bayes or Best Predictor of $\bar{Y}_{ir}$, we have

$$\hat{\bar{Y}}_{ir} = \bar{x}_{ir}'\beta + \tilde{v}_i, \tag{23}$$

where $\bar{x}_{ir} = (N_i - n_i)^{-1} \sum_{j \notin s_i} x_{ij}$, and $\tilde{v}_i \equiv \tilde{v}_i(\beta, \lambda) = [1 - B_i(\lambda)](\bar{y}_i - \bar{x}_i'\beta)$, with $B_i \equiv B_i(\lambda) = \lambda/(\lambda + n_i)$, and $\lambda = \sigma_e^2/\sigma_v^2$.

In an EB setting, one would estimate the hyperparameters using a classical method. For example, one can estimate β by the weighted least squares estimator with estimated variance components and REML to estimate the variance components. One can then use a resampling method (e.g., Jiang et al. 2002) or Taylor series (e.g., Datta and Lahiri 1999, Das et al. 2004) method to estimate the MSE. Confidence intervals can be obtained using the parametric bootstrap method of Chatterjee, Lahiri and Li (2008).

In a HB setting, one would put a prior on the hyperparameters. Typically, enough data will be available to estimate β and σ_e^2 so that one can use any reasonable noninformative prior distribution. For example, one can assume that apriori β and σ_e are independent and β and σ_e have improper uniform priors in the p-dimensional Euclidean space and positive part of the real line, respectively. The prior on σ_v is less clear cut. See Gelman (2006). One suggestion is to put an improper uniform prior on σ_v. MCMC can be applied to carry out the fully Bayesian data analysis for a variety of inferential problems. See Molina and Rao (2015) for various extensions of the nested

error model proposed by Battese et al. (1988) and the different estimation methods.

Heteroscedasticity in the distribution of e_{ij} is common in real applications. One solution proposed in the literature is to assume nested error regression model (19) on some suitable transformations (e.g., logarithm) of the data of study and auxiliary variables. As the heterogeneity problem persists even after such transformations, Molina et al. (2014) suggested to replace σ_e^2 by $k_{ij}\sigma_e^2$ in the nested error regression model, where k_{ij} is known. In many real applications, however, it is not easy to choose k_{ij}. In a poverty mapping application, Elbers et al. (2003) discussed this problem of heteroscedasticity at length and suggested certain complex modeling on the sampling variances. The nested error regression modeling assumption on the transformed study variable and/or the estimation of complex non-linear parameters such as the poverty gap and poverty severity (see Foster et al. 1984) require availability of auxiliary variables not only for the sampled units but also for all units in the population, which limits the use of such models in many applications.

In an effort to provide a solution to the heteroscedasticity problem that does not require transformations, Bellow and Lahiri (2012) considered a nested error regression model (19) with σ_e^2 replaced by $x_{ij}^{\delta}\sigma_e^2$, where x_{ij} is a size variable available in a list frame, and δ is a parameter to be estimated from the data.

Another potential solution that avoids transformation was suggested by Gershunskaya and Lahiri (2018). The predictor of $\bar{Y}_{ir}$ is derived from a model (denoted N2) that is obtained from the nested error regression model (19) with the distribution of e_{ij} following a mixture of two normal distributions with zero mean but different variances:

$$e_{ij}|z_{ij} \overset{ind}{\sim} (1 - z_{ij})N(0, \sigma_1^2) + z_{ij}N(0, \sigma_2^2), \tag{24}$$

where the mixture part indicators z_{ij} are independently identically distributed Bernoulli random binomial variables with a common success probability π (probability of belonging to part 2). Gershunskaya and Lahiri (2018) obtained the following empirical best predictor (EBP) of $\bar{Y}_{ir}$:

$$\hat{\bar{Y}}_{ir}^{N2} = \bar{x}_{ir}'\hat{\beta}^{N2} + \hat{v}_i^{N2}, \tag{25}$$

where

$$\hat{\beta}^{N2} = \left(\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}x_{ij}x_{ij}'\right)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}x_{ij}(y_{ij}-\hat{v}_i^{N2}) \quad (26)$$

$$\hat{v}_i^{N2} = \frac{\sigma_v^2}{D_i^{N2}+\sigma_v^2}(\hat{\bar{y}}_i^{N2}-\hat{\bar{x}}_i^{N2\prime}\hat{\beta}^{N2}) \quad (27)$$

with

$$w_{ij} = \hat{\sigma}_1^{-2}(1-\hat{z}_{ij})+\hat{\sigma}_2^{-2}\hat{z}_{ij},\ \hat{z}_{ij}=E(z_{ij}|y_{ij},x_{ij},\hat{\phi})$$
$$D_i^{N2} = \left(\sum_{j=1}^{n_i} w_{ij}\right)^{-1},$$

$$\hat{\bar{y}}_i^{N2}=\left(\sum_{j=1}^{n_i} w_{ij}\right)^{-1}\sum_{j=1}^{n_i} w_{ij}y_{ij},\ \hat{\bar{x}}_i^{N2}=\left(\sum_{j=1}^{n_i} w_{ij}\right)^{-1}\sum_{j=1}^{n_i} w_{ij}x_{ij},$$

and the hyperparameters $\phi=(\beta,\sigma_1^2,\sigma_2^2,\sigma_v^2,\pi)$ are estimated using EM algorithm.

Note that $\hat{\bar{y}}_i^{N2}$ accounts for outliers. While $\hat{\bar{y}}_i^{N2}$ resembles a "direct estimator," unlike direct estimators, it depends on units from other areas through the estimates of variances and the probabilities of belonging to part 2 of the mixture. Each observation has its own conditional probability $P(z_{ij}=1|y_{ij},x_{ij},\hat{\phi})=E(z_{ij}=1|y_{ij},x_{ij},\hat{\phi})$ of belonging to part 2 of the mixture so that the observations in the sample can be ranked according to these probabilities. The estimate of $\hat{\beta}^{N2}$ (thus, the synthetic part of the estimator) is outlier robust because the outlying observations would be classified with a higher probability to the higher variance part of the mixture; hence, they would be "down-weighted" according to the formula for $\hat{\beta}^{N2}$.

Gershunskaya and Lahiri (2018) proposed the following overall bias-corrected EBP of $\bar{Y}_{ir}$:

$$\hat{Y}_{ir}^{N2+OBC}=\hat{Y}_{ir}^{N2}+n^{-1}s^R\sum_{i=1}^{m}\sum_{j=1}^{n_i}\psi_b(e_{ij}^{N2}/s^R),$$

where $e_{ij}^{N2}=y_{ij}-x_{ij}'\hat{\beta}^{N2}-\hat{v}_i^{N2}$, s^R is a robust measure of scale for the set of residuals $\{e_{ij}^{N2},\ i=1,\cdots,m;j=1,\cdots,n_i\}$, and $n=\sum_{i=1}n_i$, the overall

sample size. They considered $s^R = \text{med}|e_{ij}^{N2} - \text{med}(e_{ij}^{N2})|/0.6745$, and ψ_b is a bounded Huber's function with the tuning parameter of $b = 5$.

One potential problem in using unit level covariates from administrative records is that covariate information could be missing for a number of units. In such a case, one may consider a solution proposed by Bellow and Lahiri (2012) although they encountered the problem in a covariate from a list sampling frame. Following Bellow and Lahiri (2012), one can divide the administrative records into two groups – group 1 where unit level covariates are not missing and group 2 where unit level covariates are missing. One can then apply an EBP or Bayesian method using the nested error model (19) with the unit level covariate for group 1 and apply an alternative method for group 2. For instance, one can use EBP or Bayesian approaches without the covariates that are missing units or a simpler method such as a synthetic method used by Bellow and Lahiri (2012). Estimates from these two groups can then be combined in an appropriate way. More research is needed to address the issue of missing unit level covariates.

4 An Application

We have mentioned the U.S. Census Bureau's SAIPE program throughout the chapter to illustrate concepts related to data preparation and modeling. In fact, SAIPE is a good example of a successful implementation of a small area program by a government agency. In this section, we use past data similar to those used by SAIPE to show readers how one might analyze survey data using covariates from administrative records to produce small area statistics.

From Section 2, recall SAIPE produces poverty statistics for various age groups at different levels of geographic aggregation in the U.S. – at the state level, county level, and at the school district level. The estimates for related[2] school-aged (aged 5-17) children in poverty at the school district level are used for the allocation of funds by the U.S. Department of Education – over $16 billion in the fiscal year 2014. The primary data source for SAIPE's area level models are estimates from the American Community Survey (ACS), which samples approximately 3.5 million addresses per year. SAIPE uses data from administrative records as a source of auxiliary information for all

[2] "Related" here refers to children in families.

age groups and levels of geographic aggregation. The main sources of covariates from administrative records are selected tax records obtained from an interagency agreement with the IRS and data from the Supplemental Nutritional Assistance Program (SNAP). For estimates for the over 65 population SAIPE uses data from the Supplemental Security Income (SSI) program instead of SNAP. Some of the issues associated with using these administrative records as covariates were discussed in Section 2. SAIPE also uses estimates from the 2000 Census long-form as covariates. The census long-form, a survey that used to be part of the decennial census data collection, was discontinued after 2000 and was replaced by the ACS. Research has been conducted to explore replacing the 2000 Census long-form estimates by more current estimates based on 5 years of ACS data collection (Huang and Bell, 2012, Franco and Bell, 2015).

Much has been written about SAIPE over the years. See, for instance, the recent book chapter by Bell et al. (2016), or the many other publications available at the SAIPE website: `https://www.census.gov/programs-surveys/saipe/library.html`. Here, we focus on how to analyze data for school-aged children in poverty at the state level using data from the Current Population Survey (CPS) and associated administrative records tabulations for the year 1993. The CPS, sponsored jointly by the U.S. Census Bureau and the Bureau of Labor statistics, is primarily designed to produce monthly estimates related to labor force participation and employment, and is used to produce national unemployment rates, among other statistics. It has a multistage probability sample design. For more information about the CPS, see `http://www.census.gov/programs-surveys/cps.html`. As a historical note, SAIPE used data from the CPS from its inception until the year 2004. Starting in 2005, SAIPE began using data from the ACS due to its larger sample size – the ACS sampled approximately 3 million addresses per year at the time, whereas CPS sampled approximately 100,000 addresses in 2005. In the year 1993, CPS sampled about 60,000 addresses.

The data set we use here is from Bell and Franco (2017), available at `https://www.census.gov/srd/csrmreports/byyear.html`. The compressed set of files included there contains the text files cps93p.txt and CEN89RES.txt, among others. The colums corresponding to the ages 5-17 in these files contain the variables we use here, at the state level, for the year 1993. All but the last variable listed below are in the cps93p.txt file. The last variable is in the CEN89RES.txt file.

- **cps93** – The direct CPS estimated poverty rates for related children ages 5-17.

- **irspr93** – The pseudo-poverty rates tabulated from IRS tax data. These are defined as the number of child tax exemptions for poor households divided by the total number of child tax exemptions.

- **irsnf93** – The tax non-filer rates tabulated from IRS tax data, defined as the difference between the estimated population and number of tax exemptions under age 65, divided by the estimated population under age 65.

- **fs93** – The Food Stamp participation proportions. As pointed out in Section 2, the Food Stamp Program changed its name to the Supplemental Nutritional Assistance Program (SNAP) in 2008 . This variable is the average monthly number of individuals receiving food stamps over a 12-month period, as a percentage of the population.

- **smpsize** – The CPS sample size (number of interviewed households).

- **fnlse** – The GVF estimates of sampling standard errors from the CPS. These are computed using the GVF developed by Otto and Bell (1995), using an iterative procedure that alternates between estimation of model parameters via maximum likelihood and estimation of the sampling standard errors.

- **cen89rsd** – The residuals obtained by fitting a Fay-Herriot model to the estimates of children in poverty from the 1990 census, with analogous covariates to those used here but for the year 1989. These are found in the "Age 5-17" column of the file CEN89RES.

In some years, the census residuals were replaced in the SAIPE production model by the census estimates of children in poverty. See Bell et al. (2016) for more details.

We conduct analysis in the spirit of Bell et al. (2007), which performs analysis of ACS data and related covariates to produce county-level poverty estimates. We use some of the same model selection and diagnostic tools as in this technical report as an illustration of how such an analysis might be done in practice. The data set we use in this chapter is for research purposes only and may differ sightly from that used in actual SAIPE production.

We fit a Fay-Herriot model, given by equations (1) and (2). Here, y_i is given by **cps93** for each state, x_i is given by an intercept term, and **irspr93**, **irsnf93**, **fs93**, **cenrsd**, and D_i is given by **fnlse**^{2}, again, for each state. The analysis here is done with the "sae" package in R (Molina and Marhueda, 2015), and we invite the reader to replicate the analysis as an exercise.

We first explore the relationships between the covariates and the response. Figure 1 plots these against each other. All covariates appear to have a positive correlation with CPS direct estimates of poverty, though these relationship appears to be stronger for the Food Stamp participation rate and the IRS pseudo poverty rate. Both of those covariates are from administrative records. Note that the true relationship between the covariates and the true poverty rate is masked by the sampling variance of the CPS estimates.

We then fit the Fay-Herriot model to the CPS poverty rates using all combinations of the four covariates described above. The 15 models considered are listed in Table 2.

The models are fit using using restricted maximum likelihood estimation to estimate the model variance. Table 1 shows results for the regression coefficients. Note that all regression coefficients are significant at the 0.05 significance level, and that, as suggested by Figure 1, all coefficients are positive. Table 2 shows the estimated model variances and AIC's for all models. The lowest AIC and the lowest model variance corresponds to the model including all covariates. This is not surprising since all the t-statistics are significant and since the model with all covariates is the one that was used for production. The largest estimated model variance is for the model with the Census residuals as the only covariate. However, the census residual covariate is intended to be used in conjunction with the other covariates. Since a model similar to that used for the year 1993 was used to produce them (using census poverty estimates and covariates corresponding to the year 1989), the residuals reflect variation in the model not explained by the administrative record covariates. This is only relevant when such covariates are in the model. The next highest model variance and AIC is for the model with only the IRS non-filer rate. A high model variance implies less shrinkage towards the synthetic estimator, which contains the information from the auxiliary variables, and thus a greater weight on the direct survey estimate.

In practice, one may also consider other model forms and transformations of the data, but we do not pursue them here. We now look at some model diagnostics for the model with all four covariates (M1234), which had the lowest AIC and model variance. In particular, we study the standardized

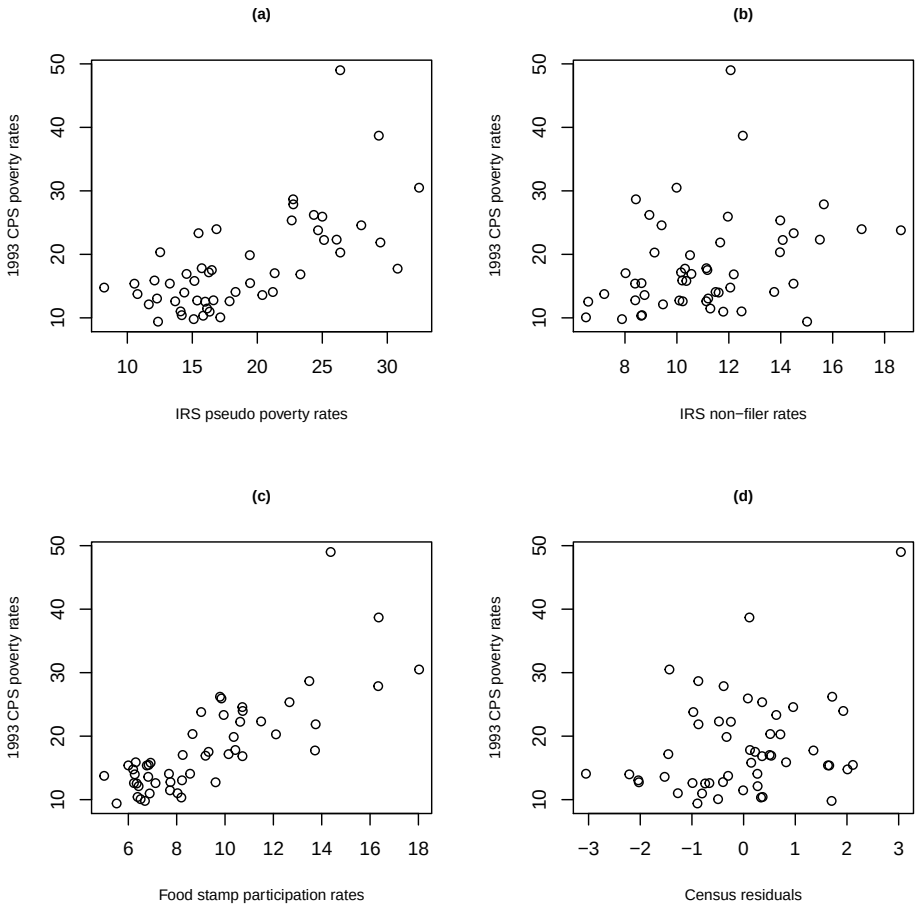

Figure 1: CPS poverty rates for school-aged children plotted against (a) IRS pseudo-poverty rates, (b) IRS non-filer rates (c) Food Stamp (SNAP) participation rates, (d) Census 1990 residuals

residuals, defined here as:

$$r_i = \frac{(y_i - x_i'\hat{\beta})}{\sqrt{var(y_i - x_i'\hat{\beta})}}.$$

When the parameters are known, $var(y_i - x_i'\beta) = D_i + A$, so we use $D_i + \hat{A}$ as a somewhat naive approximation to $var(y_i - x_i'\hat{\beta})$. More accurate vari-

Table 1: Regression prediction results for model with four regressors

Variable	Coefficient	S.E.	t	$Pr > \|t\|$
Intercept	−3.477	2.224	−1.564	0.118
IRS pseudo-poverty rate	0.267	0.125	2.144	0.032
IRS non-filer rate	0.509	0.156	3.261	0.001
Food stamp participation rate	1.185	0.268	4.429	<0.001
Census residuals	1.261	0.413	3.050	0.002

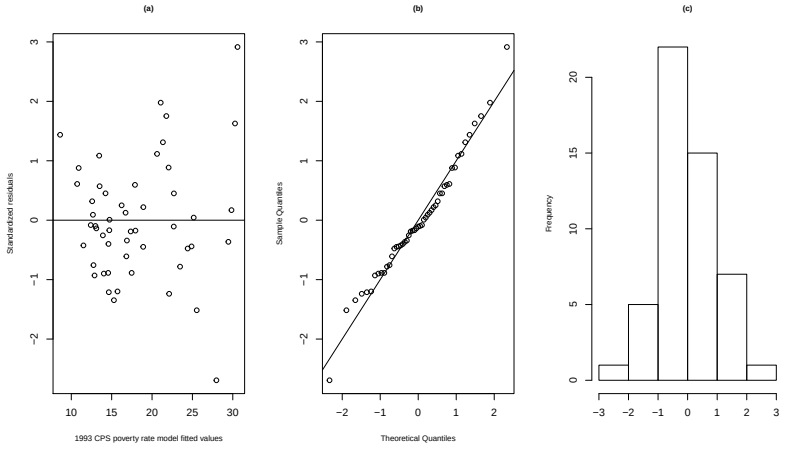

Figure 2: Model diagnostic plots: (a) Standardized residuals plotted against model predictions (b) Quantile to quantile plot of standardized residuals (c) Histogram of standardized residuals

ance estimates can also be calculated under this model, but for simplicity we omit these calculations here. Figure 2(a) shows these standardized residuals plotted against the model fitted values $x_i'\hat{\beta}$. All values are between -3 and 3, indicating an absence of extreme outliers, and there does not appear to be a systematic difference between the standardized residuals and the fitted values. Under the model, these residuals should be normal and independent of the fitted values, and Figure 2(a) does not appear to contradict this. Figures 2 (b)-(c), a quantile to quantile plot with a normal distribution and a histogram of the residuals, also do not suggest severe deviations from the normality assumption.

We now explore the differences between the direct estimators and the

Table 2: Model Comparison

Model	Regressors	Model Variance	AIC
M1	IRS pseudo poverty rate	12.803	316.638
M2	IRS non-filer rate	25.307	341.525
M3	Food Stamp participation rate	6.049	294.007
M4	Census residuals	30.631	345.437
M12	IRS pseudo poverty rate, IRS non-filer rate	7.972	308.810
M13	IRS pseudo poverty rate, Food Stamp participation rate	6.051	295.332
M14	IRS pseudo poverty rate, Census residuals	9.741	311.074
M23	IRS non-filer rate, Food Stamp participation rate	3.449	290.033
M24	IRS non-filer rate, Census residuals	24.310	339.345
M34	Food Stamp participation rate, Census residuals	5.468	289.568
M123	IRS pseudo poverty rate, IRS non-filer rate, Food Stamp participation rate	3.061	290.188
M124	IRS pseudo poverty rate, IRS non-filer rate, Census residuals	4.418	300.373
M134	IRS pseudo poverty rate, Food Stamp participation rate, Census residuals	4.85	289.749
M234	IRS non-filer rate, Food Stamp participation rate, Census residuals	3.257	285.264
M1234	IRS pseudo poverty rate, IRS non-filer rate, Food Stamp participation rate Census residuals	1.703	282.601

model predictions. Figure 3 plots these against each other, along with the

$y = x$ line. Note the difference between the domain and the range – the direct estimators have more extreme values. This is because shrinkage causes smoothing; for instance, the two highest values for the direct estimates correspond to smaller model predictions.

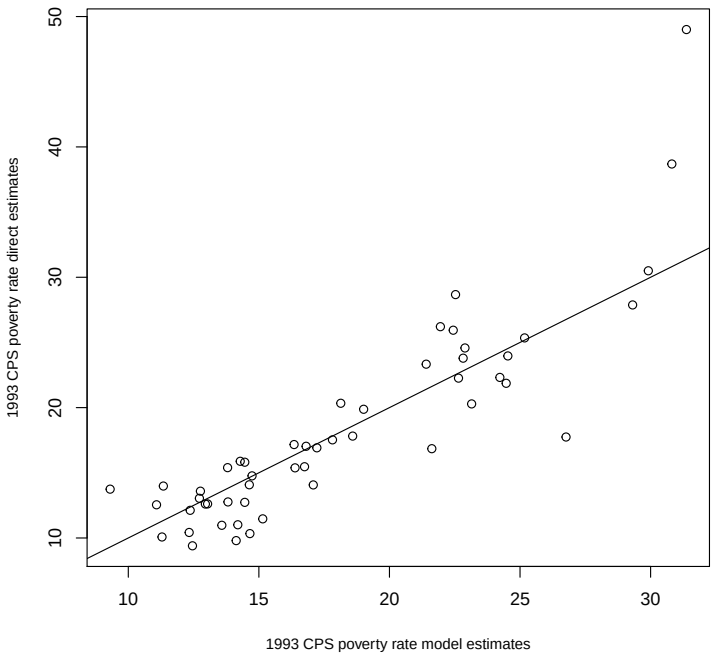

Figure 3: Model predictors vs. direct estimates for 1993 school-aged children in poverty based on CPS data, and $y = x$ line

Figures 4 (a) and (b) displays the ratios of standard errors and coefficients of variation of the modeled estimates over the direct estimates, respectively. Note large reductions in both standard errors and coefficients of variations can be achieved by modeling in this application. In fact, in all but one state, the standard errors are lower for the model estimates than for the direct estimates, with a median decrease of 57%, and range of decrease of -1% to 67%.

Of course, measurement error in the covariates, if present, would not be captured or reflected in the results under this model. Nonetheless, this appli-

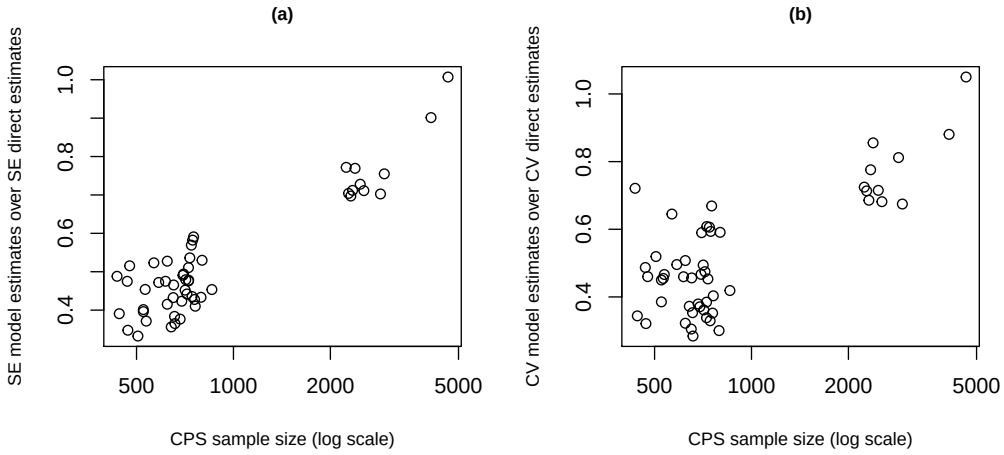

Figure 4: Ratios of standard errors (a) and of coefficients of variation (b) of modeled estimates over direct CPS estimates

cation illustrates that impressive improvement in inference can be achieved through using administrative records in small area estimation models.

5 Concluding Remarks

We critically reviewed different statistical models and methods that can be used to improve small area estimation by utilizing information extracted from administrative records. While extensive research has been done in developing small area methodology that combines survey data with aggregate-level statistics derived from administrative records, more research is still needed to incorporate linkage errors that arise from probabilistically linking records from different databases. Some early work in this area can be found in Han and Lahiri (2018) and Han (2018). Sustained collaboration between researchers in government agencies, industry, and academia may significantly advance progress in this challenging research area.

We also need to emphasize that protecting data confidentiality associated with administrative data is vital. One way to maintain confidentiality and yet have usable auxiliary data is to develop more synthetic data that match the key properties of the real data but are not themselves confidential or

sensitive.

6 Exercises

Data for the following exercises are from Bell and Franco (2017) at `https://www.census.gov/srd/csrmreports/byyear.html`. See Section 4 for more information on these data.

Exercise 1. Use the files cps93p.txt and CEN89RES.txt, as described in Section 4, to perform the following analysis:

(a) Fit the full Fay-Herriot model to the data, using all of the available covariates, and using REML to estimate the model parameters. Provide summary results for the estimated model parameters. Determine whether the estimated parameters associated with the different administrative data sources are significant. These results should match those of Section 4.

(b) Compute estimated shrinkage coefficients $B_i = \frac{D_i}{D_i + \hat{A}}$, for all the states.

Exercise 2. Use the data in *cps97p.txt* and *CEN89RES.txt* for ages 5-17 to answer the following questions:

(a) Fit the full Fay-Herriot model to the data, using all of the available covariates, and using REML to estimate the model parameters. The description of the variables for the 1993 and the 1997 datasets is the same, and is given in Section 4. Provide summary results for the estimated model parameters. Determine whether the estimated parameters associated with the different administrative data sources are significant.

(b) What implications does the estimate of the model variance $\hat{A}$ have on the weights placed on the direct estimates? What are the shrinkage coefficients for each of the states?

Exercise 3. Replicate the analysis in *Exercise 1* and *Exercise 2* using a hierarchical Bayesian approach. See Section 3.1. for choices of prior distributions to β and A and for choices of model fit and estimation. How do the model parameter estimates compare? In each case, provide the shrinkage

coefficients.

Exercise 4. Lahiri and Pramanik (2011) computed estimates for the shrinkage coefficient, B_i, based on the full Fay-Herriot models fitted to the 1993 and the 1997 data, using the following three methods:

- Exact Bayesian posterior distribution of B_i,
- The estimation method (ADM) mentioned in Section 3.1,
- The first-order Laplace approximation (Kass and Steffey, 1989).

Table 3 displays the exact posterior means and variances of $\hat{B}_i$, for the four chosen states, California (CA), North Carolina (NC), Indiana (IN) and Mississippi (MS), representing both small (i.e., large D_i) and large (i.e., small D_i) states.

Table 3: Estimates of the shrinkage coefficients based on Fay-Herriot models to SAIPE state-level data; see Lahiri and Pramanik (2011) for details.

Year		Posterior mean			Posterior variance		
	State	Exact	ADM	Laplace	Exact	ADM	Laplace
1993	CA	0.47	0.37	0.56	0.038	0.023	0.093
	NC	0.62	0.55	0.72	0.030	0.025	0.061
	IN	0.80	0.77	0.87	0.014	0.014	0.019
	MS	0.81	0.79	0.89	0.012	0.012	0.015
1997	CA	0.68	0.60	1.00	0.037	0.041	0.987
	NC	0.84	0.81	1.00	0.014	0.018	0.120
	IN	0.87	0.85	1.00	0.010	0.013	0.071
	MS	0.92	0.91	1.00	0.005	0.005	0.021

Use the data from *cps93p.txt*, *cps97p.txt*, and *CEN89RES.txt* for ages 5-17 to answer the following questions:

(a) Compare your results in *Exercise 1 (c)*, *Exercise 2 (d)*, for California (CA), North Carolina (NC), Indiana (IN) and Mississippi (MS), with the posterior means of B_i in column *Laplace*, in Table 3.

(b) Using the posterior means of B_i in column *ADM*, in Table 3, compute the estimated model (random effects) variance under the Li-Lahiri method.

(c) Give a Taylor approximation expression to the variance of the shrinkage coefficient for the general Fay-Herriot model.

(d) Compute estimates of shrinkage coefficients and their estimated Taylor approximated variance using the expression in (c), evaluated at the REML point estimates for the model (random effects) variances constructed in Exercises 1 and 2, and at the ADM point estimate for the model (random effects) variances constructed in (b). How do the estimated Taylor approximated variances compare to the estimated variances in Table 3?

Acknowledgements

The authors thank the editors for a few constructive suggestions that led to improvement of an earlier version of the chapter. The authors also thank William Bell, Jerry Maples, and David Powers for their very useful comments during the Census Internal Review. The research of the third author was supported in part by the National Science Foundation Grant Number SES-1534413.

References

[1] Arima, S., Bell, W. R., Datta, G. S., Franco, C., and Liseo, B. (2018). Multivariate Fay-Herriot Bayesian estimation of small area means under functional measurement error model. *Journal of the Royal Statistical Sociery–Series A*,180, 4, 1191-1209.

[2] Arora, V. and Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems, Statistica Sinica, 7, 1053-1063.

[3] Battese, G. E., Harter, R. M. and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association,* 83, 28-36.

[4] Bauder, M., Luery, D., and Szelepka S. (2018). Small Area Estimation of Health Insurance Coverage in 2010-2016. U. S. Census Bureau. Available online at `https://www2.census.gov/programs-surveys/sahie/technical-documentation/methodology/2008-2016-methods/sahie-tech-2010-to-2016.pdf`[accessed October 26 2018]

[5] Bell, W. R., and Franco, C. (2017). Small Area Estimation – State Poverty Rate Model Research Data Files. Available at `https://www.census.gov/srd/csrmreports/byyear.html` [accessed October 22, 2018]

[6] Bell, W. R., Chung, H. C., Datta, G. S, Franco, C. (2018). Measurement Error in Small Area Estimation: Functional Versus Structural Versus Naive Models. To appear in *Survey Methodology.*

[7] Bell, W. R, Basel W. W., Cruse C. S., Dalzell L., Maples J. J., O'Hara, B. J, Powers D. S. (2007). Use of ACS Data to Produce SAIPE Model-Based Estimates of Poverty for Counties. Available at `https://www.census.gov/content/dam/Census/library/working-papers/2007/demo/bellreport.pdf`.

[8] Bell, W. R., Basel W. W., Maples, J. J. (2016). An Overview of the U.S. Census Bureau's Small Area Income and Poverty Estimates Program. In M. Pratesi (Ed.) *Analysis of Poverty Data by Small Area Estimation* (pp. 349-377). West Sussex: Wiley & Sons, Inc.

[9] Bellow, M. and Lahiri, P. (2012). Evaluation of Methods for County Level Estimation of Crop Harvested Area that Employ Mixed Models. *Proceedings of the ICES-IV.* http://www.amstat.org/meetings/ices/2012/papers/302087.pdf

[10] Berger J.O. (1985). Statistical Decision Theory and Bayesian Analysis. Springer-Verlag.

[11] Brackstone, G. J. (1987). Small Area Data: Policy Issues and Technical Challenges, in R. Platek, J.N.K. Rao, C.-E. Sarndall, and M.P. Singh (Eds.), *Small Area Statistics*, New York: John Wiley & Sons, Inc., pp. 3-20.

[12] Carter, G. M. and Rolph, J. F. (1974). Empirical Bayes methods applied to estimating fire alarm probabilities. *J. Amer. Statist. Assoc.* **69** 880-885.

[13] Casas Cordero Valenciaa, C., Encina, J., Lahiri, P. (2016). Poverty Mapping in Chilean Comunas. In M. Pratesi (Ed.) *Analysis of Poverty Data by Small Area Estimation* (pp. 379-403). West Sussex: Wiley & Sons, Inc.

[14] Chatterjee, S., Lahiri, P. and Li, H. (2008). On small area prediction interval problems, *Annals of Statistics*, **36**, 1221-1245.

[15] Cox, D. R. (1976). Prediction intervals and empirical Bayes confidence intervals. in: J.Gani(Ed.), *Perspectives in Probability and Statistics, Papers in Honor of M.S. Bartlett, Academic Press.* 47-55.

[16] Cruse C. S. and Powers D. S. (2006). Estimating School District Poverty with Free and Reduced-Price Lunch Data. Available online at https://www.census.gov/content/dam/Census/library/working-papers/2006/demo/crusepowers2006asa.pdf.

[17] Datta, G. (2009). Model-Based Approach to Small Area Estimation. In C. R. Rao (Ed.) *Handbook of Statistics Sample Surveys: Inference and Analysis* Volume 29, Part B, Pages i-xxiii, 3-642

[18] Datta, G. S., Fay, R. E., and Ghosh, M. (1991). Hierarchical and Empirical Bayes Multivariate Analysis in Small Area Estimation, in *Proceeding of the US Census Bureau 1991 Annual Research Conference*, U.S. Census Bureau, Washington, DC, pp. 63-79.

[19] Datta, G.S. and Lahiri, P. (1999). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems, *Statistica Sinica,* **10**, 613-627.

[20] Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the U.S. states, *Journal of the American Statistical Association*, **94**, 1074-1082.

[21] Elbers, C., J. O. Lanjouw & P. Lanjouw (2003). Micro-level estimation of poverty and inequality. *Econometrica,* 71, 355-364.

[22] Efron, B. and Morris, C. N. (1973). Stein's Estimation Rule and Its Competitors – An Empirical Bayes Approach. *J. Amer. Statist. Assoc.* **68** 117-130.

[23] Efron, B. and Morris, C. N. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70** 311-319.

[24] Erciulescu A.L., Cruze N., Nandram B. (2018) "Model-Based County-Level Crop Estimates Incorporating Auxiliary Sources of Information. *Journal of the Royal Statistical Society, Series A*. DOI 10.1111/rssa.12390.

[25] Esteban, M. D, Morales, D., Perez, A. (2016). Area-level Spatio-Temporal Small Area Estimation Models. In M. Pratesi (Ed.) *Analysis of Poverty Data by Small Area Estimation* (pp. 349-377). West Sussex: Wiley & Sons, Inc.

[26] Fay R.E. and Herriot R. (1979). Estimation of income from small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association* 74, 269-277.

[27] Fay, R.E. (1987). Application of Multivariate Regression to Small Domain Estimation, in R. Platek, J.N.K. Rao, C.-E. Sarndall, and M.P. Singh (Eds.), *Small Area Statistics*, New York: John Wiley & Sons, Inc., 91-102.

[28] Foster, J., Greer, J., Thorbecke, E. (1984). A class of decomposable poverty measures, *Econometrica,* 52, 761-766.

[29] Franco, C. and Bell, W. R. (2015). Borrowing information over time in binomial/logit normal models for small area estimation. Joint Special Issue of *Statistics in Transition* and *Survey Methodology*. 16, 4, 563-584.

[30] Fuller, W. A. and Goyeneche, J. J. (1998), Estimation of the state variance component, *Unpublished manuscript*.

[31] Ganesh, N. and Lahiri, P. (2008). A new class of average moment matching prior, *Biometrika,* **95**, 514-520.

[32] Gershunskaya, J. and Lahiri, P. (2018), Robust Empirical Best Small Area Finite Population Mean Estimation Using a Mixture Model, Calcutta Statistical Association Bulletin, 69 (2) 183 4, DOI: 10.1177/0008068317722297.

[33] Ghosh, M., Nangia, N., Kim, D. H. (1996). Estimation of Median Income of Four-Person Families: A Bayesian Time Series Approach. *Journal of the American Statistical Association*, 91, 1423-1431.

[34] Ghosh, M and Rao, J.N.K. (1994) Small Area Estimation: An Appraisal (with Discussion), *Statistical Science*, 9, 55-93.

[35] Ha, N. (2013). Hierarchical Bayesian Estimation of Small Area Means Using Complex Survey Data, PhD Dissertation, University of Maryland, College Park, USA.

[36] Ha, N. S., Lahiri, P. and Parsons, V. (2014). Methods and results for small area estimation using smoking data from the 2008 National Health Interview Survey, *Statistics in Medicine*. **33**. 22.

[37] Hall, P and Maiti, T. (2006). On parametric bootstrap methods for small-area prediction. *J. Roy. Statist. Soc. Ser. B.* **68** 221-238.

[38] Han, Y. (2018), Statistical Inference Using Data From Multiple Files Combined Through Record Linkage, PhD. Dissertation, University of Maryland, College Park.

[39] Han, Y. and Lahiri, P. (2018), Statistical Analysis with Linked Data, to appear in *International Statistical Review*.

[40] Hawala, S. and Lahiri, P. (2018) Variance Modeling for Domains, to appear in *Journal of the Indian Society of Agricultural Statistics*..

[41] Henderson, C.R. (1953) Estimation of Variance and Covariance Components, *Biometrics*, 9, 226-252.

[42] Holt, T. (2007). The Official Statistics Olympic Challenge: Wider, Deeper, Quicker, Better, Cheaper. (With discussions). *The American Statistician*, vol. 61, pp. 1- 15.

[43] Huang, E. T., Bell, W. R. (2012). An Empirical Study on Using Previous American Community Survey Data Versus Census 2000 Data in SAIPE Models for Poverty Estimates. Research Report Number RRS2012–4, Center for Statistical Research and Methodology, U.S. Census Bureau, `http://www.census.gov/srd/papers/pdf/rrs2012_04.pdf`.

[44] Jiang, J., Nguyen, T. and Rao, J. S. (2011). Best predictive small area estimation, J. Amer. Statist. Assoc. 106, 732-745.

[45] Jiang, J., Lahiri, P. and Wan, S. (2002). A unified jackknife theory for empirical best prediction with M-estimation, *Annals of Statistics,* **30**, 1782-1810.

[46] Kim J.K., Wang Z., Zhu Z., Cruze N. (2018), Combining Survey and Non-Survey Data for Improved Sub-Area Prediction Using a Multi-Level Model, *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 23, **2**, pp. 175-189.

[47] Lahiri, P. and Li, H. (2009). Generalized maximum likelihood method in linear mixed models with an application in smallarea estimation. In Proceedings of the Federal Committee on Statistical Methodology Research Conference. Available at `http://www.fcsm.gov/events/papers2009.html`.

[48] Lahiri, P. and Pramanik, S. (2011), Discussion of "Estimation of random effects via adjustment for density maximization," by C. Morris and R. Tang, Statistical Science, http://dx.doi.org/10.1214/10- STS349, 271-298.

[49] Lahiri, P. and Suntornchost, J. (2015) Variable Selection for a Regression model when dependent variable is subject to measurement errors, Sankhya, Series B. DOI 10.1007/s13571-015-0096-0

[50] Li, H. and Lahiri, P. (2010). Adjusted maximum method for solving small area estimation problems, *Journal of Multivariate Analysis*, **101**, 882-892, doi: 10.1016/j.jmva.2009.10.009.

[51] Liu, B. (2009) Hierarchical Bayes Estimation and Empirical Best Prediction of Small Area proportions, PhD. Dissertation, University of Maryland, College Park.

[52] Liu, B., Lahiri, P. and Kalton, G. (2014). Hierarchical Bayes Modeling of Survey-Weighted Small Area Proportions. *Survey Methodology*. 40. 1-13.

[53] Luery, D. M. Small Area Income and Poverty Estimates Program. (2010) U.S. Census Bureau. Availlable online at `https://www.census.gov/library/working-papers/2010/demo/luery-01.html`

[54] Maples J.J. (2008) Calculating coeffcient of variation for the minimum change school district poverty estimates and the assessment of the impact of nongeocoded tax returns. Research Report RRS2008/10, Center for Statistical Research and Methodology, U.S. Census Bureau, available online at `https://www.census.gov/srd/papers/pdf/rrs2008-10.pdf` [accessed July 15-16] [accessed July-10-16]

[55] Maples J.J. and Bell W.R. (2007). Small area estimation of school district child population and poverty: Studying the use of IRS income tax data. Research Report RRS2007/11, Center for Statistical Research and Methodology, U.S. Census Bureau, available at online at `https://www.census.gov/srd/papers/pdf/rrs2007-11.pdf`. [accessed July-15-16]

[56] Molina, I. and Marhuenda, Y. (2015) sae: An R Package for Small Area Estimation.

[57] Morris, C. and Tang, R. (2011). Estimating random effects via adjustment for density maximization. *The R Journal*. 7, 1, pp 81-82. *Statistical Science*, 26: 271-287.

[58] National Research Council (2000a). Small-Area Estimates of School-Age Children in Poverty: Evaluation of Current Methodology (eds Citro C.F. and Kalton G.), Panel on Estimates of Poverty for Small Geographic Areas, Committee on National Statistics, Washington, DC: National Academy Press.

[59] National Research Council (2000b). Small Area Income and Poverty Estimates: Priorities for 2000 and Beyond (eds Citro C.F. and Kalton G.), Panel on Estimates of Poverty for Small Geographic Areas, Committee on National Statistics, Washington, DC: National Academy Press.

[60] Otto M.C. and Bell W.R. (1995). Sampling error modelling of poverty and income statistics for states. *Proceedings of the American Statistical Association*, Social Statistics Section, 160165. Avaiable online at `https://www.census.gov/library/working-papers/1995/demo/otto-01.html` [accessed October 25 2018]

[61] Pfeffermann, D. (2013). New Important Developments in Small Area Estimation. *Statistical Science*, 28, 1, 40-68

[62] Powers, D. Bowers L., Basel, W. and Szelepka, S. (2016). Medicaid and CHIP Data Methodology for SAHIE Models. Proceedings of the 2015 Federal Committee on Statistical Methodology (FCSM) Research Conference. Available online at `http://www.census.gov/content/dam/Census/library/working-papers/2016/demo/powers-bowers-basel-szelepka-fcsm.pdf` [accessed July 15 2016]

[63] Prasad, N. G. N. and Rao, J. N. K. (1990). The estimation of the mean squared error of small area estimators. *J. Amer. Statist. Assoc.* **85** 163-171.

[64] Vogt, M. (2010). Bayesian Spatial Modeling: Propriety and Applications to Small Area Estimation with Focus on the German Census 2011, PhD Dissertaion, Univrsity of Trier.

[65] Rao J.N.K. (2003). *Small Area Estimation.* Hoboken: John Wiley & Sons, Inc.

[66] Rao J.N.K. and Molina, I. (2015). *Small Area Estimation (2nd. ed.).* Hoboken: John Wiley & Sons, Inc.

[67] Rao, J.N.K. and Yu, M. (1994). Small Area Estimation by Combining Time Series and Cross-Sectional Data, *Canadian Journal of Statistics*, 22, 511-528.

[68] Torabi, M. and Rao, J. N. K. (2014), On small area estimation under a subarea level model, *Journal of Multivariate Analysis*, 127, 36-55.

[69] Winkler, William E. Matching and Record Linkage. U.S. Bureau of the Census. Available online at `http://www.census.gov/srd/papers/pdf/rr93-8.pdf` [accessed July 15-16]

[70] Wolter, K.M. (2007) *Introdcution to Variance Estimation*, (2nd Edition), New York: Springer-Verlag.

[71] Ybarra, L.M.R. and Lohr, S.L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, **95**, 919-931.

[72] Yoshimori, M. and Lahiri, P. (2014a), A new adjusted maximum likelihood method for the Fay-Herriot small area model, *Journal of Multivariate Analysis,* 124, 281-294, `http://dx.doi.org/10.1016/j.jmva.2013.10.012`

[73] Yoshimori, M. and Lahiri, P. (2014b). A second-order efficient empirical Bayes confidence interval, *The Annals of Statistics*, 42, No. 4, 1233-1261 DOI: 10.1214/14-AOS1219.

[74] Zanutto, E. and Zaslavsky, A. (2002). Using Administrative Data to Improve Small Area Estimation: An Example from the U.S. Decinnial Census, *Journal of Official Statistics*, 18, 559-576.

[75] Zhang, L.C. and and Giusti, C. (2016). Small Area Methods and Administrative Data Integration. In M. Pratesi (Ed.) *Analysis of Poverty Data by Small Area Estimation* (pp. 379-403). West Sussex: Wiley & Sons, Inc.

[76] Zhang L-C (2012) Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, **66**, 41-63.