Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Combining data from surveys and related sources

by Dexter Cahoy and Joseph Sedransk

Release date: June 30, 2023

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

© His Majesty the King in Right of Canada as represented by the Minister of Industry, 2023

All rights reserved. Use of this publication is governed by the Statistics Canada Open Licence Agreement.

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2023
Vol. 49, No. 1, pp. 69-87
Statistics Canada, Catalogue No. 12-001-X

Combining data from surveys and related sources

Dexter Cahoy and Joseph Sedransk[1]

Abstract

To improve the precision of inferences and reduce costs there is considerable interest in combining data from several sources such as sample surveys and administrative data. Appropriate methodology is required to ensure satisfactory inferences since the target populations and methods for acquiring data may be quite different. To provide improved inferences we use methodology that has a more general structure than the ones in current practice. We start with the case where the analyst has only summary statistics from each of the sources. In our primary method, uncertain pooling, it is assumed that the analyst can regard one source, survey *r*, as the single best choice for inference. This method starts with the data from survey *r* and adds data from those other sources that are shown to form clusters that include survey *r*. We also consider Dirichlet process mixtures, one of the most popular nonparametric Bayesian methods. We use analytical expressions and the results from numerical studies to show properties of the methodology.

Key Words: Administrative data; Bayesian methods; Clustering; Dirichlet process mixture; Pooling data; Survey sampling.

1. Introduction

With substantially reduced response rates and limited budgets there has been an increased emphasis on efficient use of all of the information available to the survey analyst. Specifically, one may be able to improve inferences by using results from several sample surveys and related sources such as administrative records. The methodology that we use to combine information has more structure than the methods currently used in survey sampling, so should lead to better inferences. Starting with the data from the survey that is the best choice for inference, these data are augmented with other, concordant, data. We use analytical expressions and the results from numerical studies to show properties of the methodology.

This research was motivated by a study (Ha and Sedransk, 2019) of health insurance coverage in Florida's counties where the authors noted very different estimates from three surveys. Correspondingly, we could have estimates from a well established probability survey and two non-probability surveys. In either case there is the question about how to make better inferences.

In the sequel we refer to the collection of studies as "surveys", recognizing that these may be probability surveys, non-probability surveys, administrative records and other sources. We consider the case where the analyst has only a point estimate and associated standard error from each survey. This is common, as noted in Section 7 of the review paper, Lohr and Raghunathan (2017). In the motivating example, and in many other cases, there are not any covariates that can be used to improve inferences. Our methodology extends to cases where the inferential objectives and models are more complex.

With survey estimates, $\left\{\hat{Y}_i : i = 1, \ldots, L\right\}$, it is commonly assumed that the $\hat{Y}_i$ are independent

$$\hat{Y}_i \sim N\left(\mu_i, V_i\right) \tag{1.1}$$

1. Dexter Cahoy, University of Houston-Downtown; Joseph Sedransk, University of Maryland. E-mail: jxs123@case.edu.

where the V_i are assumed to be known.

A common prior distribution expressing similarity among $\{\mu_1, \ldots, \mu_L\}$ is

$$\mu_i \,|\, \nu, \delta^2 \sim N(\nu, \delta^2) \tag{1.2}$$

independently for each i, and ν and δ are assigned locally uniform prior distributions.

The resulting posterior expected value of μ_i is a convex combination of the estimate, $\hat{Y}_i$, for that survey, and a weighted average of $\{\hat{Y}_1, \ldots, \hat{Y}_L\}$. The weakness of this approach is that the prior distribution in (1.2) assumes independent sampling of the μ_i from a common distribution. The posterior mean is

$$E\left(\mu_i \,|\, \hat{Y}_1, \ldots, \hat{Y}_L\right) = \lambda_i \hat{Y}_i + (1 - \lambda_i)\, \hat{Y}_w \tag{1.3}$$

where $\lambda_i = \delta^2 / (\delta^2 + V_i)$ and $\hat{Y}_w = \sum_{i=1}^{L} \lambda_i \hat{Y}_i \Big/ \sum_{i=1}^{L} \lambda_i$. This may lead to unsatisfactory inferences when, e.g., $\mu_1, \ldots, \mu_b$ are each close to μ^* while $\mu_{b+1}, \ldots, \mu_L$ are each close to μ^{**} and $\mu^* << \mu^{**}$. Here, estimation of μ_1 would include, perhaps inappropriately, a large contribution from $\hat{Y}_{b+1}, \ldots, \hat{Y}_L$. The difficulty is that the prior distribution is not sufficiently flexible. Here we use more flexible prior distributions that permit the amount and nature of the pooling to be determined by the sample data.

The specification in (1.1) and (1.2) is common in meta-analyses and in situations where inferences for small subpopulations and geographical areas are desired. For example, the U.S. Census Bureau uses such models (augmented by terms to accommodate covariates) to make inference for U.S. county level poverty rates: see example 6.1.2 in Rao and Molina (2015). However, as just noted, the assumptions in (1.1) and (1.2) of full exchangeability may not be appropriate, especially for combining the information from L surveys.

The theory for our principal method, uncertain pooling, was developed by Malec and Sedransk (1992) and Evans and Sedransk (2001) with further work in Evans and Sedransk (1999) and Yan and Sedransk (2011). To our knowledge this methodology has not been used in a survey sampling application: the comprehensive review paper, Lohr and Raghunathan (2017), makes no reference to any technique similar to ours. We also modify (1.1) and (1.2) by using a Dirichlet process mixture (DPM) as, e.g., employed by Polettini (2017) for small area inference.

In this paper we describe the methodology for both uncertain pooling and DPM, and use them to analyze data from Ha and Sedransk (2019). Then we modified these data to exhibit properties of the methods. Finally, there is a simulation study to establish sampling properties. Clearly, the results from this evaluation will apply equally if the three sources were, e.g., a well established probability survey and two non-probability surveys or other choices.

We assume that the sample variances $\{V_1, \ldots, V_L\}$ are specified. In our context none of the alternatives given in the literature for making inferences for the $\{V_1, \ldots, V_L\}$, all based on inferences for small areas, is fully satisfactory. In Section 4 we discuss this challenging problem of making inferences for the sample variances.

Both uncertain pooling and DPM have a more general structure than that in the common specification, (1.1) and (1.2). This should lead to improved inferences. As seen in Section 2 the uncertain pooling model is the natural extension of (1.1) and (1.2); i.e., the model in (1.1) and (1.2) is a special case. Moreover, the output from uncertain pooling includes the posterior probabilities associated with the possible clustering of the L surveys.

Finally, note that we address only one of the many aspects of "combining survey data", well summarized by Lohr and Raghunathan (2017). Their Section 7, "Hierarchical models for combining data sources", gives additional examples where our methodology may be useful.

The methodology that we use is outlined in Section 2, and the results from our numerical studies are summarized in Section 3. A brief summary and discussion are in Section 4.

2. Methodology

As in Section 1 assume that there are L survey estimates, $\hat{Y}_1, \ldots, \hat{Y}_L$, with

$$\hat{Y}_i \overset{\text{ind}}{\sim} N\left(\mu_i, V_i\right) \tag{2.1}$$

where the V_i are assumed to be known.

2.1 Uncertain pooling

The uncertain pooling method is based on Malec and Sedransk (1992) and Evans and Sedransk (2001). They showed that a prior for $\mu = \left(\mu_1, \mu_2, \ldots, \mu_L\right)'$ can be selected to reflect the beliefs that there are subsets of μ such that the μ_i in each subset are "similar", and that there is uncertainty about the composition of such subsets of μ. Let G be the total number of partitions of the set $\mathcal{L} = \{1, \ldots, L\}$, g be a particular partition $\left(g = 1, \ldots, G\right)$, $d\left(g\right)$ be the number of subsets of $\mathcal{L}$ in the g^{th} partition $\left(1 \le d\left(g\right) \le L\right)$, and $S_k\left(g\right)$ be the set of survey labels in subset $k\left(k = 1, \ldots, d\left(g\right)\right)$. For example, for $L = 3$, there are $G = 5$ partitions: $\{g = 1\} \sim \{(123)\}$, $\{g = 2\} \sim \{(13), (2)\}$, $\{g = 3\} \sim \{(12), (3)\}$, $\{g = 4\} \sim \{(23), (1)\}$, $\{g = 5\} \sim \{(1), (2), (3)\}$. Then, $S_1\left(g = 2\right) = \{(13)\}$, $S_2\left(g = 2\right) = \{(2)\}$, $d_1\left(g = 2\right) = 2$ and $d_2\left(g = 2\right) = 1$.

To specify a prior for μ, first condition on g. Malec and Sedransk (1992) and Evans and Sedransk (2001) assume that there is independence between subsets, and within $S_k\left(g\right)$ the μ_i are independent with

$$\mu_i \,|\, \nu_k\left(g\right) \sim N\left(\nu_k\left(g\right), \delta_k^2\left(g\right)\right), \qquad i \in S_k\left(g\right). \tag{2.2}$$

Also, the $\nu_k\left(g\right)$ are mutually independent with

$$\nu_k\left(g\right) \,|\, \theta_k\left(g\right) \sim N\left(\theta_k\left(g\right), \gamma_k^2\left(g\right)\right) \tag{2.3}$$

where the $\theta_k\left(g\right)$ and the $\gamma_k^2\left(g\right)$ are hyperparameters. The definition in (2.3) is the first step in obtaining a reference prior for the $\nu_k\left(g\right)$, i.e., one that is dominated by the likelihood. This will include letting the

$\gamma_k^2(g) \to \infty$, but is considerably more complicated, as described below. The $\delta_k^2(g)$ are also hyperparameters, to be assigned a prior distribution.

The formal definition in (2.3) is included as the first step in obtaining a reference prior for the $\nu_k(g)$, i.e., one that is dominated by the likelihood, as described below in the evaluation of $f(g, \Delta^2 \mid y)$. Conditioning on the $\delta_k^2(g)$ and $\gamma_k^2(g)$ (but suppressing them in our notation), and letting $\gamma_k^2(g) \to \infty$ leads to the following expected results for the posterior moments conditional on the partition g. As discussed below, additional care is needed to obtain the posterior distribution of g.

Defining $y = (\hat{Y}_1, \ldots, \hat{Y}_L)^t$, letting $\Delta^2 = \{\delta_k^2(g) : k = 1, \ldots, d(g); g = 1, \ldots, G\}$ and writing $\hat{\mu}_i = \hat{Y}_i$

$$E(\mu_i \mid y, g, \Delta^2) = \{\lambda_i(g)\}\, \hat{\mu}_i + \{1 - \lambda_i(g)\}\, \hat{\mu}_k(g), \qquad i \in S_k(g) \tag{2.4}$$

and

$$\operatorname{cov}(\mu_i, \mu_j \mid y, g, \Delta^2) = \begin{cases} \delta_k^2(g)\{1-\lambda_i(g)\} + \dfrac{\{1-\lambda_i(g)\}^2 \delta_k^2(g)}{\sum_{i \in S_k(g)} \lambda_i(g)}, & i = j; i, j \in S_k(g)) \\ \dfrac{\{1-\lambda_i(g)\}\{1-\lambda_j(g)\}\delta_k^2(g)}{\sum_{i \in S_k(g)} \lambda_i(g)}, & i \neq j; i, j \in S_k(g) \\ 0, & i \in S_{k_1}(g), j \in S_{k_2}(g), k_1 \neq k_2, \end{cases} \tag{2.5}$$

where

$$\lambda_i(g) = \frac{\delta_k^2(g)}{\delta_k^2(g) + V_i}, \hat{\mu}_k(g) = \frac{\sum_{j \in S_k(g)} \lambda_j(g) \hat{\mu}_j}{\sum_{j \in S_k(g)} \lambda_j(g)}. \tag{2.6}$$

Note that $E(\mu_i \mid y, g, \Delta^2)$ has the familiar form of a weighted average of $\hat{\mu}_i$ and $\hat{\mu}_k(g)$, but, here, $\hat{\mu}_k(g)$ is restricted to the surveys in $S_k(g)$.

Assuming the basic model in (1.1) and (1.2) corresponds, here, to the "pool-all" partition, $\{g = 1\}$, where all of the L surveys comprise a single cluster. Thus, for $\{g = 1\}$ the moments in (2.4), (2.5) and (2.6) are those that would be obtained by an analysis using (1.1) and (1.2). An analysis based on (1.1) and (1.2) is a special case of an analysis based on the uncertain pooling specification.

Inference about μ includes uncertainty about the value of g, i.e.,

$$f(\mu \mid y) = \int f(\mu \mid y, g, \Delta^2)\, f(g, \Delta^2 \mid y)\, dg d\Delta^2 \tag{2.7}$$

where the notation is simplified by using integration rather than summation for g. Using the "most likely" partition g^* (i.e., $p(g^* \mid y) \geq p(g \mid y) : g = 1, \ldots, G$) to make inference would understate the overall precision.

To evaluate (2.7) we need $f(g, \Delta^2 \mid y)$. However, when evaluating $f(g \mid \Delta^2, y)$ one must be careful about specifying the rate at which the $\gamma_k^2(g) \to \infty$: a natural choice leads to an expression for

$f\left(g\middle|\,\Delta^2, y\right)$ that is not invariant to changes in the scale of Y; see Section 4 of Malec and Sedransk (1992). Malec and Sedransk (1992) provided a solution by using an empirical Bayes argument. Here we use a fully Bayesian alternative, described in Section 5 of Evans and Sedransk (2001). It postulates little prior, relative to sample, information about the $\nu_k(g)$, and is invariant to changes in the scale of Y. Let $\nu(g) = \left(\nu_1(g),\ldots,\nu_{d(g)}(g)\right)^t$ and $K\left\{f_1\left(\nu(g)\right), f_2\left(\nu(g)\middle|\, y\right)\right\}$ be the Kullback-Leibler information about $\nu(g)$. With prior $f\left(g, \Delta^2\right) = f(g)f\left(\Delta^2\right)$ and letting the $\gamma_k^2(g) \to \infty$ subject to $K\left\{f_1\left(\nu(g)\right), f_2\left(\nu(g)\middle|\, y\right)\right\} = \text{constant}$,

$$f\left(g, \Delta^2\middle|\, y\right) \propto f\left(\Delta^2\right) f(g) \exp\left\{\frac{-d(g)}{2}\right\} \prod_{k=1}^{d(g)} \prod_{i \in S_k(g)} \left\{1-\lambda_i(g)\right\}^{1/2} \times \exp\left[-\frac{1}{2}\sum_{k=1}^{d(g)} \sum_{i \in S_k(g)} \left\{\frac{\lambda_i(g)}{\delta_k^2(g)}\right\} \left\{\hat{\mu}_i - \hat{\mu}_k(g)\right\}^2\right]. \tag{2.8}$$

The term in the exponent,

$$Q\left\{d(g)\right\} = \sum_{k=1}^{d(g)} \sum_{i \in S_k(g)} \left\{\frac{\lambda_i(g)}{\delta_k^2(g)}\right\} \left\{\hat{\mu}_i - \hat{\mu}_k(g)\right\}^2, \tag{2.9}$$

is likely to decrease as $d(g)$ increases, for example for a new partition of $\bigcup_{k=1}^{d(g)} S_k(g)$ obtained by creating subsets of the existing $\left\{S_k(g)\right\}$. Since $f\left(g, \Delta^2\middle|\, y\right)$ increases as $Q\left\{d(g)\right\}$ decreases, it is helpful to have the second term, $\exp\left\{-d(g)/2\right\}$, that penalizes partitions with larger values of $d(g)$.

For our analysis we take $\delta_k^2(g) = \delta^2$ and write $\lambda_i(g) = \delta^2 \big/ \left(\delta^2 + V_i\right)$. Inference for μ is made using (2.7) and (2.8) with

$$\mu\middle|\, y, g, \delta^2 \sim N\left(E\left(\mu\middle|\, y, g, \delta^2\right), V\left(\mu\middle|\, y, g, \delta^2\right)\right) \tag{2.10}$$

where the conditional posterior moments of μ are given in (2.4) and (2.5).

We assume that $f(g)$ is constant, i.e., that all partitions are equally likely, *a priori*, and take the Inverse Beta prior for δ^2, i.e.,

$$f\left(\delta^2\right) \propto 1 \big/ \left(1+\delta^2\right)\sqrt{\delta^2}, \quad 0 < \delta^2 < \infty. \tag{2.11}$$

Inference for μ is made using (2.7). To start, evaluate the right side of (2.8) for

$$\left\{g : g = 1,\ldots,G; \quad R \text{ grid points for } \delta^2\right\}, \tag{2.12}$$

then standardize by dividing the individual terms in the grid by their sum. This provides an approximation for $f\left(g, \delta^2\middle|\, y\right)$. Then select a random sample of size B from the RG normalized values of $f\left(g, \delta^2\middle|\, y\right)$. For each selection, $\left(g_*, \delta_*^2\right)$, sample μ from $f\left(\mu\middle|\, y, g_*, \delta_*^2\right)$. Here, we generated $B =$ 5,000 values of μ. Finally, note that approximations for the marginal posterior distributions, i.e., $f\left(g\middle|\, y\right)$ and $f\left(\delta^2\middle|\, y\right)$, can be obtained directly from the grid approximation of $f\left(g, \delta^2\middle|\, y\right)$.

Assuming that survey r is the single best choice for inference we can consider the posterior distribution corresponding to survey r to be the object of inference. In a common contemporary application there will be data from a well established probability survey (the single best choice) and data from other sources such as non-probability surveys, administrative records, etc. In other settings it is likely that there will be a preference for one of the surveys.

Then, using the posterior expected value for illustration,

$$E\left(\mu_r \mid y\right) = E_{g,\delta^2 \mid y} E\left(\mu_r \mid y, g, \delta^2\right) \tag{2.13}$$

where $E\left(\mu_r \mid y, g, \delta^2\right)$ is defined in (2.4). Thus, inference for μ_r is a function of $\hat{\mu}_r$ together with data from the other $L-1$ studies as determined by the form of (2.4), and, critically, by the likelihood associated with the set of subsets, $S_k(g)$, containing study r. See Evans and Sedransk (2001) for additional details and an application to a notable study of the effect of using aspirin by patients following a myocardial infarction.

The model given by Chakraborty, Datta and Mandal (2014) has a superficial resemblance to the one in (2.1), (2.2), and (2.3). Taking x_i to be the scalar with value 1, the model in (2.1) of Chakraborty et al. (2014) is

$$\hat{Y}_i = \mu_i + e_i, \quad i = 1, \ldots, L \tag{2.14}$$

where $\mu_i = \xi + \left(1-\delta_i\right)\nu_{1i} + \delta_i \nu_{2i} + e_i$ with $e_i, \delta_i, \nu_{1i}, \nu_{2i}$ independent, $p\left(\delta_i = 1 \mid p\right) = 1 - p, \nu_{1i} \sim N\left(0, A_1\right)$ and $\nu_{2i} \sim N\left(0, A_2\right)$. Finally, $e_i \sim N\left(0, V_i\right)$ with V_i known. Thus, unlike the uncertain pooling method, there is only a single focal point, ξ. This permits appropriate treatment of outliers, but does not take advantage of possible clustering of the μ_i. This can also be seen in (2.4) of Chakraborty et al. (2014) where

$$E\left(\mu_i \mid \xi, A_1, A_2, p, y\right) = \hat{Y}_i - \kappa_i\left(\hat{Y}_i - \xi\right) \tag{2.15}$$

and κ_i is a function of $V_i, A_1, A_2, p\left(\delta_i = 0 \mid \xi, A_1, A_2, p, y\right)$ with $y = \left(\hat{Y}_1, \ldots, \hat{Y}_L\right)^t$. Chakraborty et al. (2014) show that if survey i is an outlier, $E\left(\mu_i \mid \xi, A_1, A_2, p, y\right) \approx \hat{Y}_i$, as desired. Now suppose that surveys $\{1, \ldots, b\}$ and $\{b+1, \ldots, L\}$ form two distinct clusters with a very large separation between them. Then inference for μ_1, say, will not, in general, use the data in an appropriate manner. In (2.15) there should be two values of ξ, i.e., corresponding to the two subsets. And appropriate use of information about subsets is the essence of the uncertain pooling method.

2.2 Dirichlet process mixture

An alternative to the uncertain pooling method is to use a Dirichlet process mixture (DPM), one of the most popular nonparametric Bayesian methods. This methodology is presented in detail in Sections 2.1 and 2.2 of Muller, Quintana, Jara and Hanson (2015). For our analyses we have used the R function

DPmeta from the package DPpackage: see Jara, Hanson, Quintana, Muller and Rosner (2011) for details. The model in DPmeta is

$$y_i \mid \theta_i \overset{\text{iid}}{\sim} f_{\theta_i} \tag{2.16}$$

and

$$\theta_i \mid H \overset{\text{iid}}{\sim} H \tag{2.17}$$

with $H \sim \text{DP}(M, H_0)$.

In (2.16) and (2.17) $y_i = \hat{Y}_i, \theta_i = \mu_i, f_{\theta_i}$ is the pdf of a $N(\mu_i, V_i)$ random variable with V_i fixed, and $H_0 = N(\eta, \tau^2)$.

The (independent) hyperparameters are

$$M \mid a_0, b_0 \sim \text{Gamma}(a_0, b_0)$$

$$\eta \mid \eta_b, S_b \sim N(\eta_b, S_b)$$

$$\tau^{-2} \mid \phi_1, \phi_2 \sim \text{Gamma}(\phi_1/2, \phi_2/2). \tag{2.18}$$

Polettini (2017) has proposed using a DPM of this nature for inference about small area parameters. As in Section 2.1 of our paper Polettini (2017) indicates the value of extending the typical random effects model (e.g., the well known Fay-Herriot model) that assumes full exchangeability of the set of small area parameters.

The uncertain pooling method requires only that one specify a prior distribution for g and δ^2. By contrast DPmeta requires substantial prior input, i.e., values for $a_0, b_0, \eta_b, S_b, \phi_1$ and ϕ_2. Without strong prior information we can't make proper inferences for these quantities with only $L = 3$ surveys. So, we have omitted the specification $M \mid a_0, b_0 \sim \text{Gamma}(a_0, b_0)$ and made inference for a selected set of values of M as suggested by Escobar (1994). Also, we replaced η_b, S_b, ϕ_1 and ϕ_2 with their maximum *a posteriori* probability estimates.

3. Results

Ha and Sedransk (2019) made inference for the proportion of adults without health insurance in each of the 67 Florida counties, and compared these estimates with those from two other sources. Some of the differences were striking, motivating us to consider methodology to make appropriate inferences in such cases. We use these data, and modifications of these data, to show the benefits of using the methodology outlined in Section 2. One source is the Small Area Health Insurance Estimates Program (SAHIE, hereafter survey 1). The SAHIE program uses point estimates from the American Community Survey (ACS) together with administrative data such as Federal income tax returns and Medicaid/Children's

Health Insurance Program (CHIP) participation rates. There is detailed area level modelling. The principal ones are models of ACS estimates of the proportions in income groups and the proportions insured. There are additional models such as ones modelling the number of persons enrolled in Medicaid or CHIP, Supplementary Nutrition Assistance Program (SNAP) participation, and Internal Revenue Service (IRS) tax exemptions. For a full understanding of this program see the twenty-two page technical report, Bauder, Luery and Szelepka (2018). We have added a non-technical summary in the Appendix.

The analyses using both survey 2, denoted by HS, based on Ha and Sedransk (2019), and survey 3 denoted by CDC (Centers for Disease Control and Prevention) use unit level models based on 2010 data from the Behavioral Risk Factor Surveillance System (BRFSS), obtained through telephone interviews. While the sample designs differ somewhat over states, the one in Florida was typical, i.e., a disproportionate stratified sample design. In Florida, the set of telephone numbers were divided into two strata (high and medium density) that were sampled separately. In addition there was a stratification by area codes, i.e., three geographic strata and a fourth stratum consisting of area codes with large estimated Hispanic populations. For additional, general information see http://www.cdc.gov/brfss/annual_data/annual_2010.htm while for technical details see Pierannunzi, Xu, Wallace, Garvin, Greenlund, Bartoli, Ford, Eke and Town (2016) and Ha and Sedransk (2019). Both use, essentially, the same covariates but the modeling in HS is more detailed. Moreover, Pierannunzi et al. (2016) give only point estimates, noting that standard errors were being developed. A further complication is that the CDC analysis is frequentist while the SAHIE and HS analyses are Bayesian. Thus, we have an (empirical) Bayes posterior standard deviation for SAHIE, none for CDC and an estimated SE for HS obtained by taking the 95% credible interval for a county proportion and dividing by 3.92. While the limitations just noted preclude definitive conclusions from these data they illustrate the methodology. Moreover, conditions where standard errors are missing or unreliable are, at least, fairly common for non-probability samples, a focus of this paper.

The first set of analyses is based on the observed data. To show other properties of the methodology a second set of analyses is based on modifications of these data. Finally, to show sampling properties there is a simulation study. Note that each of our analyses is based *only* on data from a single county. Additional research is needed to permit inference using data from all of the sources and counties. See Section 4 for discussion.

3.1 Data-based analyses

Using the uncertain pooling methodology a summary of the results for Dixie county is presented in Table 3.1. These are typical of most of the county-based analyses that we have done. Throughout, SE denotes the sample standard error. There are three panels, corresponding to choices of the CDC SE, taken equal to 0.5, 1.0, 2.0 times the HS SE. For each panel the column headings are the observed proportion, posterior mean of the county proportion, estimated SE of the observed proportion, posterior standard deviation and lower and upper bounds of the 95% credible interval for the county proportion. At the bottom of each panel there are the values of $p(g \mid y)$ with $p(g \mid y) \geq 0.001$ where $\{g = 1\} \sim \{(123)\}$,

$\{g=2\}\sim\{(13),(2)\}$, $\{g=3\}\sim\{(12),(3)\}$, $\{g=4\}\sim\{(2,3),(1)\}$, $\{g=5\}\sim\{(1),(2),(3)\}$, and summaries corresponding to $\{g=1\}$, labelled "pool-all".

We first analyze these data using the uncertain pooling methodology, then compare them with those from DPmeta.

A common way to summarize a set of sample proportions is to assume that the corresponding set of true proportions come from a common source, i.e., $p(g=1)=1$. However, for each of the three cases in Table 3.1, $p(g=1\mid y)\le 0.001$. Thus, there is very little support for pooling all of the data from the three surveys. For further investigation of the effect of assuming a common source, assume $g=1$. Then, as in (1.1) and (1.2),

$$\hat{Y}_i \overset{\text{ind}}{\sim} N(\mu_i, V_i)$$

$$\mu_i \overset{\text{iid}}{\sim} N(\nu, \delta^2), \quad i=1,\ldots,L. \tag{3.1}$$

With a locally uniform prior on ν and the Inverse Beta prior on δ^2 in (2.11)

$$f(\nu\mid y)=\int f(\nu\mid \delta^2, y) f(\delta^2\mid y)\, d\delta^2 \tag{3.2}$$

where the posterior distribution of ν given δ^2 is normal with $E(\nu\mid \delta^2, y)=\frac{\sum_{i=1}^{L}\hat{Y}_i/(V_i+\delta^2)}{\sum_{i=1}^{L}1/(V_i+\delta^2)}$ and $\text{Var}(\nu\mid \delta^2, y)=\left(\sum_{i=1}^{L}(V_i+\delta^2)^{-1}\right)^{-1}$.

For panel 1 in Table 3.1, $E(\nu\mid y)=0.313$, $\text{SD}(\nu\mid y)=0.017$ and the 95% credible interval is (0.290, 0.340). Inferences based on the posterior distribution of ν are not consistent with the notion that any one of the three surveys is the nominal "gold standard". If, for example, survey 1 is taken as the "gold standard", the posterior mean of μ_1, 0.254, is substantially smaller than the posterior mean of ν, 0.313. Moreover, 0.254 is not included in the 95% interval for ν, (0.290, 0.340). The conclusions from panels 2 and 3 are essentially the same. Finally, recall that $p(g=1\mid y)\le 0.001$, indicating very little support for pooling all of the data.

In the following assume, for illustration, that one prefers the HS methodology. Then there may be substantial gains in precision (measured by the posterior standard deviation) by using the uncertain pooling methodology. The gain in precision is measured by comparing the posterior standard deviation from the uncertain pooling methodology with that obtained by using only the data from the specific survey, here HS (survey 2). For the latter and a locally uniform prior for μ_2, the posterior distribution of μ_2 is normal with posterior mean equal to the observed proportion and posterior standard deviation equal to the estimated SE. If we take CDC SE $= k$ (HS SE) for $k=0.5, 1, 2$, then the reductions in the posterior standard deviation for HS (survey 2), and corresponding to $k=0.5, 1, 2$, are 29, 18 and 7%. (For example, from panel 1 of Table 3.1, i.e., $k=0.5$, the percent reduction in the posterior SD for HS is $100\,(0.028-0.020)\,/\,0.028\%=29\%$.) Note that the relatively small SEs for each of the surveys means that the "all singletons" partition, i.e., $\{g=5\}$, has a relatively large posterior probability (about 0.38).

The corresponding reductions in the posterior standard deviation (uncertain pooling vs. no pooling) for CDC (survey 3) are 7, 18 and 14%.

Table 3.1
Observed proportions, standard errors and posterior summaries from Dixie County, Florida using uncertain pooling

	Survey	ObsProp	PostMean	ObsSE	PostSD	95% Cred Int
CDC SE = 0.5 × HS SE	1	0.254	0.254	0.014	0.014	(0.225, 0.283)
	2	0.361	0.360	0.028	0.020	(0.317, 0.403)
	3	0.359	0.359	0.014	0.013	(0.333, 0.385)
	pool-all		0.313		0.017	(0.290, 0.340)
	$P(g=3\mid y)=0.002;\ P(g=4\mid y)=0.621;\ P(g=5\mid y)=0.377.$					
CDC SE = HS SE	1	0.254	0.254	0.014	0.014	(0.225, 0.283)
	2	0.361	0.360	0.028	0.023	(0.313, 0.406)
	3	0.359	0.359	0.028	0.023	(0.312, 0.404)
	pool-all		0.290		0.011	(0.268, 0.312)
	$P(g=2\mid y)=0.002;\ P(g=3\mid y)=0.002;\ P(g=4\mid y)=0.619;\ P(g=5\mid y)=0.376.$					
CDC SE = 2 × HS SE	1	0.254	0.254	0.014	0.014	(0.226, 0.284)
	2	0.361	0.360	0.028	0.026	(0.307, 0.412)
	3	0.359	0.349	0.056	0.048	(0.256, 0.303)
	pool-all		0.279		0.012	(0.255, 0.305)
	$P(g=1\mid y)=0.001;\ P(g=2\mid y)=0.107;\ P(g=3\mid y)=0.002;\ P(g=4\mid y)=0.554;\ P(g=5\mid y)=0.336.$					

Note: Centers for Disease Control and Prevention (CDC); Standard error (SE); Ha and Sedransk (HS); Standard deviation (SD).

As noted in Section 2, a complete specification of DPmeta requires specifying the values of many hyperparameters, and we have no prior information to make informed choices. So, we have replaced η_b, S_b, ϕ_1 and ϕ_2 with their maximum *a posteriori* probability (MAP) estimates. We have followed Escobar (1994) by considering $M \in \{L^{-1}, L^0, L^1, L^2\} = \{1/3, 1, 3, 9\}$.

From (2.10) in Muller et al. (2015) the prior probability of k clusters is a function of M. Let $p_M = (p_{1M}, p_{2M}, p_{3M})$ where p_{kM} is the *prior* probability of k clusters with precision M. Then p_{kM} can be calculated using the probability associated with any partition, i.e.,

$$\frac{M^{k-1}\prod_{j=1}^{k}\Gamma(L_j)}{(M+1)(M+2)\cdots(M+L-1)} \tag{3.3}$$

where L_j is the number of surveys in cluster j with $\sum_{j=1}^{k} L_j = L$. Then $p_{1/3} = (18/28, 9/28, 1/28)$, $p_1 = (2/6, 3/6, 1/6)$, $p_3 = (2/20, 9/20, 9/20)$ and $p_9 = (2/110, 27/110, 81/110)$. Since $p_{1/3}$ and p_9 are too extreme we have emphasized $M = 1$ and $M = 3$. The results corresponding to $M = 1$ and $M = 3$ are very close, so only the latter are presented in Table 3.2, which has the same format as Table 3.1.

Comparing the results from the uncertain pooling method with those from DPmeta it is apparent that, in general, there is close agreement. For the posterior mean they are similar except in panel 3 where there is greater shrinkage for surveys 2 and 3. The results for the posterior SD are also close except for a larger

value for survey 2 in panel 3. There are only small differences in the intervals except for panel 3 where the DPmeta intervals for surveys 2 and 3 are wider.

Table 3.2
Observed proportions, standard errors and posterior summaries from Dixie County, Florida using DPmeta

	Survey	ObsProp	PostMean	ObsSE	PostSD	95% Cred Interval
CDC SE = 0.5 × HS SE	1	0.254	0.254	0.014	0.014	(0.227, 0.282)
	2	0.361	0.359	0.028	0.013	(0.334, 0.384)
	3	0.359	0.360	0.014	0.012	(0.335, 0.384)
CDC SE = HS SE	1	0.254	0.256	0.014	0.016	(0.227, 0.290)
	2	0.361	0.357	0.028	0.024	(0.291, 0.399)
	3	0.359	0.357	0.028	0.025	(0.290, 0.399)
CDC SE = 2 × HS SE	1	0.254	0.264	0.014	0.018	(0.230, 0.287)
	2	0.361	0.332	0.028	0.044	(0.261, 0.406)
	3	0.359	0.321	0.056	0.048	(0.249, 0.402)

Note: Centers for Disease Control and Prevention (CDC); Standard error (SE); Ha and Sedransk (HS); Standard deviation (SD).

The small value of the SAHIE SE seen in almost all counties limits the scope of our evaluation. Thus, we have used modified data sets based on the original data. Here, as before, we take the CDC SE to be 0.5, 1 and 2 times the HS SE, but also take the SAHIE SE to be 2, 5 and 10 times the HS SE. Table 3.3, with the same format as Table 3.1, shows, for uncertain pooling, the results for Orange county with the CDC SE = 0.5(HS SE). These results are typical of our analyses for a sizeable number of FL counties. Recall, though, that each analysis is based only on the data from the specific county. For panel 1 in Table 3.3, $E\left(\nu \middle| \left\{\hat{Y}_i : i = 1, \ldots, L\right\}\right) = 0.199$, $\mathrm{SD}\left(\nu \middle| \left\{\hat{Y}_i : i = 1, \ldots, L\right\}\right) = 0.008$ and the 95% credible interval is (0.184, 0.215). Inferences based on the posterior distribution of ν are inappropriate if one regards any one of the three surveys as the "gold standard". For example, the posterior mean of μ_1, 0.278, is outside the 95% credible interval for ν. As in Table 3.1, $p\left(g = 1 \middle| y\right) \leq 0.001$, indicating very little support for pooling all of the data.

The percent reductions in the posterior standard deviation of μ_1, i.e., for SAHIE (survey 1), are 11, 26 and 44%, corresponding to the three panels in Table 3.3. As the value of the SAHIE SE is increased, there is, as expected, additional pooling of the SAHIE observed proportion with the CDC observed proportion. Noting that the observed proportion for SAHIE is 0.294, the posterior means for μ_1 decrease from 0.278 (panel 1) to 0.240 (panel 3). One reason for this can be seen by comparing the posterior distributions of g, i.e., $\left\{g, p\left(g \middle| y\right) : g = 1, \ldots, 5\right\}$, given at the bottom of each panel. For example, $p\left(g = 2 \middle| y\right)$ increases from 0.006 to 0.339 while $p\left(g = 5 \middle| y\right)$ decreases from 0.479 to 0.253. These results show that the uncertain pooling methodology is taking proper account of the increased variability associated with the SAHIE estimates, i.e., increasing the likelihood of pooling the data from surveys 1 and 3.

Comparing the results from the uncertain pooling method in Table 3.3 with those from DPmeta in Table 3.4 it is apparent that there are greater differences than those seen in Tables 3.1 and 3.2. For the posterior means it is notable that for survey 1 the posterior mean from DPmeta is somewhat smaller than that from uncertain pooling. This reflects greater pooling of the data from survey 1 with that from survey

3. For surveys 2 and 3 the two sets of posterior means are similar. The most notable difference is for the posterior SDs where, for survey 1 (panels 2 and 3) the values are very much smaller from DPmeta than from uncertain pooling. For the other seven cases, the two sets of posterior SDs are similar. Correspondingly, for survey 1 the intervals from DPmeta are much shorter than those from uncertain pooling while those for surveys 2 and 3 are only a little wider. From Table 3.4 and for survey 1 the percent reductions in the posterior SD (relative to the ObsSEs) are (42%, 55%, 75%). There are increases, though, for survey 2 in panels 2 and 3.

Table 3.3
Observed proportions, standard deviations and posterior summaries from Orange County, Florida, where CDC SE = 0.5 × HS SE using uncertain pooling

	Survey	ObsProp	PostMean	ObsSE	PostSD	95% Cred Int
SAHIE SE = 2 × HS SE	1	0.294	0.278	0.036	0.032	(0.227, 0.352)
	2	0.257	0.261	0.018	0.017	(0.226, 0.294)
	3	0.179	0.179	0.009	0.009	(0.162, 0.197)
	pool-all		0.199		0.008	(0.184, 0.215)
	$P(g=2\mid y)=$ 0.006; $P(g=3\mid y)=$ 0.514; $P(g=5\mid y)=$ 0.479.					
SAHIE SE = 5 × HS SE	1	0.294	0.251	0.089	0.066	(0.162, 0.417)
	2	0.257	0.258	0.018	0.018	(0.223, 0.293)
	3	0.179	0.179	0.009	0.009	(0.162, 0.197)
	pool-all		0.195		0.008	(0.180, 0.211)
	$P(g=2\mid y)=$ 0.224; $P(g=3\mid y)=$ 0.468; $P(g=5\mid y)=$ 0.308.					
SAHIE SE = 10 × HS SE	1	0.294	0.240	0.179	0.101	(0.059, 0.520)
	2	0.257	0.257	0.018	0.018	(0.222, 0.292)
	3	0.179	0.179	0.009	0.009	(0.162, 0.197)
	pool-all		0.195		0.008	(0.179, 0.211)
	$P(g=2\mid y)=$ 0.339; $P(g=3\mid y)=$ 0.408; $P(g=5\mid y)=$ 0.253.					

Note: Small Area Health Insurance Estimates (SAHIE); Centers for Disease Control and Prevention (CDC); Standard error (SE); Ha and Sedransk (HS); Standard deviation (SD).

Table 3.4
Observed proportions, standard errors and posterior summaries from Orange County, Florida using DPmeta

	Survey	ObsProp	PostMean	ObsSE	PostSD	95% Cred Interval
SAHIE SE = 2 × HS SE	1	0.294	0.262	0.036	0.021	(0.202, 0.297)
	2	0.257	0.263	0.018	0.018	(0.222, 0.296)
	3	0.179	0.180	0.009	0.009	(0.162, 0.199)
SAHIE SE = 5 × HS SE	1	0.294	0.226	0.089	0.040	(0.168, 0.290)
	2	0.257	0.246	0.018	0.023	(0.186, 0.291)
	3	0.179	0.182	0.009	0.011	(0.163, 0.205)
SAHIE SE = 10 × HS SE	1	0.294	0.217	0.179	0.044	(0.166, 0.288)
	2	0.257	0.243	0.018	0.031	(0.185, 0.290)
	3	0.179	0.183	0.009	0.011	(0.162, 0.205)

Note: Small Area Health Insurance Estimates (SAHIE); Standard error (SE); Ha and Sedransk (HS); Standard deviation (SD).

3.2 Results from simulation study

To evaluate properties such as bias and coverage of the credible interval we have carried out a simulation study based on several modifications of the Orange county data. Specifically, we generate $\{\hat{Y}_i : i = 1, 2, 3\}$ from

$$\hat{Y}_1 \sim N(\psi_1, V_1)$$

$$\hat{Y}_2 \sim N(\psi_1, V_2)$$

$$\hat{Y}_3 \sim N(\psi_2 + \Delta, V_2) \tag{3.4}$$

where ψ_1 and ψ_2 are from the third panel of Table 3.3; ψ_1 is the average of the observed proportions from surveys 1 and 2 while ψ_2 is the observed proportion from survey 3 (CDC). Also, we took V_1 to be much larger than V_2. These choices were made to represent a common situation where survey 1 is a probability sample, with relatively large sample variance while surveys 2 and 3 are non-probability samples with much smaller sample variances. Finally, $\Delta \in \{0, 4(0.0193) = 0.0772, 8(0.0193) = 0.1544\}$.

Table 3.5 gives the values of ψ_1, ψ_2, V_1 and V_2 in the footnote. There are three rows, corresponding to $\Delta = 0, 0.0772, 0.1544$. In each row there are the medians over 500 replications of $\{p(g \mid y) : g = 1, \ldots, 5\}, \{E(\mu_i \mid y) : i = 1, 2, 3\}$ and $\{\text{SD}(\mu_i \mid y) : i = 1, 2, 3\}$, together with the estimated coverages.

Table 3.5
Simulation results from 500 replications of (3.4)

Inc Size	$P(g \mid \text{data})$					Coverage			PostMean			PostSD		
	g : 1	2	3	4	5	i : 1	2	3	i : 1	2	3	i : 1	2	3
$\Delta = 0$	0	0.148	0.462	0	0.332	0.973	0.958	0.960	0.262	0.275	0.179	0.050	0.006	0.006
$\Delta = 0.0772$	0.032	0.292	0.303	0.033	0.249	0.984	0.941	0.939	0.269	0.275	0.257	0.039	0.006	0.006
$\Delta = 0.1544$	0	0.280	0.401	0	0.300	0.971	0.958	0.952	0.292	0.275	0.333	0.044	0.006	0.006

$\psi_1 = 0.276$, $\psi_2 = 0.179$, $V_1 = 0.06^2$, $V_2 = 0.006^2$.

Note: Standard deviation (SD).

The principal findings are: (a) the medians of the posterior means are close to the values used to generate the data, i.e., ψ_1 and ψ_2, (b) the coverages are close to the nominal 95%, and (c) there are significant reductions in the posterior standard deviation for survey 1 (SAHIE), i.e., 16.7%, 35.0% and 26.7% corresponding to $\Delta = 0, 0.0772$ and 0.1544. There are no reductions in the posterior standard deviations for surveys 2 and 3.

For $\Delta = 0.1544$ note that $p(g = 2 \mid y) = 0.280$ while $p(g = 4 \mid y) = 0$. That is, we pool data from surveys 1(SAHIE) and 3(CDC), $\{g = 2\}$, because of the relatively large SE for survey 1(SAHIE). However, we do not pool data from surveys 2(HS) and 3(CDC), $\{g = 4\}$, because of the relatively small SEs for survey 2(HS) and survey 3 (CDC). Of course, we pool data from surveys 1 and 2, $\{g = 3\}$, because they have the same mean, ψ_1.

4. Discussion and summary

With reduced response rates and diminished resources there is considerable interest in combining data from several sources such as sample surveys and administrative data. Currently there is special interest

when the sources include non-probability surveys. Appropriate methodology is required to ensure satisfactory inferences since the target populations and data acquisition methods may be quite different.

There are many situations where it may be beneficial to combine such data, as shown in the review paper by Lohr and Raghunathan (2017). Here, we have investigated the case where the analyst has only summary statistics from each of the sources, and where one can think of one source, r, as the single best source for inference. While it is often beneficial to use the data from related sources to improve inferences from r, it is essential that the data that are combined be concordant with the data from r. The methodology in this paper can also be used in settings where the data are not limited to summary statistics and inferential objectives and models are more complex. As seen in this paper, failure to consider biases due to pooling "unlike" data may lead to poor inference. Using analytical expressions and examples we have shown that both the uncertain pooling and DPM methods provide appropriate inferences. However, our analyses based on uncertain pooling are fully Bayes while those from DPmeta are empirical Bayes – due to the need to specify values for many hyperparameters. Moreover, the uncertain pooling method provides additional information in the form of the posterior probabilities for the partitions, g.

The methods can be implemented. For DPmeta there is an R package, DPpackage (Jara et al., 2011) while an R package is being developed for the uncertain pooling method. When completed it will be submitted to The Comprehensive R Archive Network. It contains functions that allow Bayesian analyses of the type described in this paper (a) with user-supplied point estimates and associated variances, or (b) with binomial data, cases (y) and total counts (n). Case (b) provides an analysis based on the logit transformation of the sample proportion. We have implemented the latter when there are eleven surveys.

Making inference for the sample variances, $V_1, \ldots, V_L$ is a very challenging problem. Polettini (2017) provides an extensive discussion of methods that have been proposed. Of particular interest are solutions proposed by You and Chapman (2006), Sugasawa, Tamae and Kubokawa (2017) and Polettini (2017). However, these solutions are posed in the context of small area inference, not when the objective is combining data from surveys and related sources.

While the discussion below is in the context of extending the DPM method (Section 2.2), the ideas are also relevant for the uncertain pooling method (Section 2.1). Polettini (2017) augments the DPM model in Section 2.2 with

$$\delta_i S_i^2 \overset{\text{ind}}{\sim} V_i \chi^2_{\delta_i}, \quad i = 1, \ldots, L \tag{4.1}$$

and

$$V_i^{-1} \overset{\text{iid}}{\sim} \text{Gamma}(a_1, b_1) \tag{4.2}$$

where S_i^2 is the sampling variance and δ_i is a measure of the degrees of freedom.

As noted by Polettini (2017) the assumption of the χ^2 distribution in (4.1) is questionable, surely so when there is a complex survey design. With only a single sample one cannot verify the *sampling distribution* of S_i^2, a point also made by Polettini (2017) on page 731. Moreover, in survey sampling the

form of S_i^2 is likely to be a complex function of the values of the variable of interest, Y, and survey weights. Thus, it is unlikely that the distribution of the observed Y's can be used to infer a reasonable approximation for the distribution of S_i^2.

The assumption of constant population parameters in (4.2) is problematic for our case, i.e., combining data. We expect considerable differences among the surveys, e.g., for a collection of probability and non-probability samples. You and Chapman (2006) generalize by replacing (4.2) with

$$V_i \overset{\text{ind}}{\sim} \text{Inverse Gamma}\left(a_i, b_i\right). \tag{4.3}$$

This requires values for $\left(a_i, b_i\right)$, a difficult choice without prior information. Moreover, Gelman (2006) shows that selecting both a_i and b_i very small, a natural choice (and one made by You and Chapman (2006)), may lead to poor inferences. Sugasawa et al. (2017) provide an alternative to You and Chapman (2006) by assuming

$$V_i \overset{\text{ind}}{\sim} \text{Inverse Gamma}\left(a_i, b_i\gamma\right), \tag{4.4}$$

with a prior on γ, but this, too, requires specifying values for a_i and b_i. Clearly, making better inference for the sample variances is an important topic for future research.

There has been an increased interest in making inference for small subpopulations, i.e., "small area" inference, when there are several data sources; see, e.g., Manzi, Spiegelhalter, Turner, Flowers and Thompson (2011) and Nandram, Berg and Barboza (2014). While further research is needed to extend the uncertain pooling methodology to this case the approach is clear. Let j denote a small area, e.g., a US county, and i denote a data source where $j = 1, \ldots, J$ and $i = 1, \ldots, L$. As above g denotes a generic partition with generic subset $S_k(g)$ for $k = 1, \ldots, d(g)$. Define $\mathcal{G} = \left\{(ij): j = 1, \ldots, J; i = 1, \ldots, L\right\}$. Then for fixed g, $S_k(g)$ is a subset of $\mathcal{G}$ with $S_k(g) \bigcap S_m(g) = \varnothing$ for $k \neq m$ and $\bigcup_{k=1}^{d(g)} S_k(g) = \mathcal{G}$. For example, let $J = 2$ and $L = 3$. Then each partition is a collection of the disjoint subsets of $\mathcal{G} = \left\{(11), (12), (21), (22), (31), (32)\right\}$ whose union is $\mathcal{G}$. By analogy with the discussion in Section 2, there would be a single best source for each small area, identified as $\left(j, i(j)\right)$ for some i in small area j.

Then the following model, analogous to the one in Section 2, is

$$\hat{Y}_{ij} \overset{\text{ind}}{\sim} N\left(\mu_{ij}, V_{ij}\right). \tag{4.5}$$

By analogy with (2.2)

$$\mu_{ij} \overset{\text{ind}}{\sim} N\left(\nu_k(g), \delta_k^2(g)\right), \qquad ij \in S_k(g). \tag{4.6}$$

If the same (limit) assumptions are made about the $\gamma_k^2(g)$ and $\delta_k^2(g) = \delta^2$, the expressions for posterior inference for the μ_{ij} will be the same as in Section 2. However, the assumption of constant δ^2 may not be reasonable. Since there will be a very large number of partitions computation will be challenging, especially since it is expected that many $p\left(g \mid y\right)$ will be very small.

The premise of our work is that one should include the possibility that the parameters associated with different surveys may not be exchangeable. (With a probability sample and several non-probability

samples this may be especially important.) Similarly, it is natural to generalize so that the parameters associated with the small areas are not assumed to be exchangeable. However, if exchangeability across both surveys and small areas can be assumed, the model in Section 2.1 of Kim, Park and Kim (2015) (possibly modified to accommodate Bayesian inference) should be easier to implement.

Although the very large number of partitions of $\mathcal{G}$ may pose an obstacle to implementation, one may be able to apply DPmeta when there are data from a set of small areas and several data sources. One problem is the specification in DPmeta of a common distribution for the μ_{ij}, i.e., over small areas and surveys, which is unlikely to be appropriate. The possibilities include an ANOVA model (Section 4.4.2) or nested model (Section 7.3.1) of Muller et al. (2015), although the ANOVA model has no interaction terms and our model is a cross-classified one.

Future research should include making inference for the sample variances, as noted above. Also, we need improved methodology to handle the extension to small area inference when there are data from several surveys. In some cases one may be able to simplify the model for the μ_{ij}. Using a grid-based method for sampling g and δ^2 is difficult to implement when G is extremely large. So, using a standard MCMC approach, possibly with an informative prior on g, may be a better way to make inference. For example, see Dahl, Day and Tsai (2017).

Other approaches could also be explored. For example, Park, Kim and Stukel (2017) suggest a different approach for combining data from two surveys. Here, there are covariates, X, observed in each survey while Y_1, the study variable of interest, is observed only in survey 1 and Y_2 is observed only in survey 2. Inference for the population mean of Y_1 is desired, given data from both surveys. The densities that they use are $f_1\left(Y_1 \mid X, \theta_1\right)$, $f_2\left(Y_2 \mid X, Y_1, \theta_2\right)$ and, for identifiability, it is assumed that $f_2\left(Y_2 \mid X, Y_1\right) = f_2\left(Y_2 \mid Y_1\right)$. For a Bayesian analysis an extension to more than two surveys would be needed, together with specification of appropriate prior distributions for the parameters. It does not seem to be straightforward to model the distribution of Y_2 given Y_1.

Acknowledgements

The authors are grateful to the reviewers whose extensive comments have resulted in a paper that has greater focus and increased breadth. They also appreciate research allocation grants from ACCESS's Pittsburgh Supercomputing Center.

Appendix

Small Area Health Insurance Estimates (SAHIE) Program

The following summary paraphrases relevant parts of US Census Bureau (2021). To avoid distortion of the authors' meaning we have retained the first-person text.

The SAHIE program produces model-based estimates of health insurance coverage for demographic groups within counties and states. We publish county estimates by sex, age and income. The income

groups are defined by the income-to-poverty ratio (IPR) – the ratio of family income to the appropriate federal poverty level.

For estimation, SAHIE uses models that combine survey data from the American Community Survey (ACS) with administrative records data and Census data. The models are "area-level" models because we use survey estimates and administrative data at certain levels of aggregation, rather than individual survey and administrative records. Our modeling approach is similar to that of common models developed for small area estimation, but with additional complexities.

The published estimates are based on aggregates of modeled demographic groups. For counties, we model at a base level defined by age, sex and income groups.

We use estimates from the Census Bureau's Population Estimates Program for the population in groups defined for county by age and sex. We treat these populations as known. Within each of these groups, the number with health insurance coverage in any of the income categories is given by that population multiplied by two unknown proportions to be estimated: the proportion in the income category and the proportion insured within that income category. The models have two largely distinct parts – an "income part" and an "insurance part" – that correspond to these proportions. We use survey estimates of the proportions in the income groups and of the proportions insured within those groups. We assume these survey estimates are unbiased and follow known distributions. We also assume functional forms for the variances of the survey estimates that involve parameters that are estimated. We treat supplemental variables that predict one or both of unknown income and insurance proportions in one of two ways:

Some of these variables are used as fixed predictors in a regression model. There is a regression component in both the income and insurance parts of the model. In each case, a transformation of the proportion is predicted by a linear combination of fixed predictors. Some of these predictors are categorical variables that define the demographic groups we model. Others are continuous. The continuous fixed predictors include variables regarding employment, educational attainment, and demographic population.

We also utilize random continuous predictors, which include data from 5-year ACS, Internal Revenue Service, Supplemental Nutrition Assistance Program, and Medicaid/Children's Health Insurance Program. These are not fixed predictors in the model. Instead, we treat them as random, in a way similar to survey estimates, but not as unbiased estimators of the numbers. Instead, we assume that their expectations are linear functions of the number in an income group or the number insured within an income group. We typically assume they are normally distributed with variances that depend on unknown parameters.

We formulate the model in a Bayesian framework and report the posterior means as the point estimates. We use the posterior means and variances together with a normal approximation to calculate symmetric 90-percent confidence intervals, and report their half-widths as the margins of error.

We control the estimates to be consistent with specified national totals.

References

Bauder, M., Luery, D. and Szelepka, S. (2018). *Small Area Estimation of Health Insurance Coverage in 2010-2016*. Technical Report, U.S. Census Bureau.

Chakraborty, A., Datta, G.S. and Mandal, A. (2016). A two-component normal mixture alternative to the Fay-Herriot model. *Statistics in Transition new series and Survey Methodology*, 17, 1, 67-90, https://doi.org/10.21307/stattrans-2016-006.

Dahl, D., Day, R. and Tsai, J. (2017). Random partition distribution indexed by pairwise information. *Journal of the American Statistical Association*, 112, 721-732.

Escobar, M. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89, 268-277.

Evans, R., and Sedransk, J. (1999). Methodology for pooling subpopulation regressions when there is uncertainty about which subpopulations are similar. *Statistica Sinica*, 9, 345-359.

Evans, R., and Sedransk, J. (2001). Combining data from experiments that may be similar. *Biometrika*, 88(3), 643-656.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515-533.

Ha, N., and Sedransk, J. (2019). Assessing health insurance coverage in Florida using the Behavioral Risk Factor Surveillance System. *Statistics in Medicine*, 38(13), 2332-2352, https://doi.org/10.1002/sim.8108.

Jara, A., Hanson, T., Quintana, F., Müller, P. and Rosner, G. (2011). DPpackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software*, 40(5), 1-30, https://doi.org/10.18637/jss.v040.i05.

Kim, J.-K., Park, S. and Kim, S.-Y. (2015). Small area estimation combining information from several sources. *Survey Methodology*, 41, 1, 21-36. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2015001/article/14150-eng.pdf.

Lohr, S., and Raghunathan, T. (2017). Combining survey data with other data sources. *Statistical Science*, 32(2), 293-312.

Malec, D., and Sedransk, J. (1992). Bayesian methodology for combining the results from different experiments when the specifications for pooling are uncertain. *Biometrika*, 79(3), 593-601.

Manzi, G., Spiegelhalter, D.J., Turner, R.M., Flowers, J. and Thompson, S.G. (2011). Modelling bias in combining small area prevalence estimates from multiple surveys. *Journal of the Royal Statistical Society, A*, 174(1), 31-50, https://doi.org/10.1111/j.1467-985X.2010.00648.x.

Muller, P., Quintana, F., Jara, A. and Hanson, T. (2015). *Bayesian Nonparametric Data Analysis*. Springer.

Nandram, B., Berg, E. and Barboza, W. (2014). A hierarchical Bayesian model for forecasting state-level corn yield. *Environmental and Ecological Statistics*, 21, 507-530, https://doi.org/10.1007/s10651-013-0266-z.

Park, S., Kim, J. and Stukel, D. (2017). A measurement error model approach to survey data integration: Combining information from two surveys. *Metron*, 75, 345-357.

Pierannunzi, C., Xu, F., Wallace, R., Garvin, W., Greenlund, K., Bartoli, W., Ford, D., Eke, P. and Town, G. (2016). A methodological approach to small area estimation for the Behavioral Risk Factor Surveillance System. *Preventing Chronic Disease*, 2016 Jul. 14, 13, http://dx.doi.org/10.5888/pcd13.150480.

Polettini, S. (2017).A generalised semiparametric Bayesian Fay-Herriot model for small area estimation shrinking both means and variances. *Bayesian Analysis*, 2017, 12, 729-752.

Rao, J., and Molina, I. (2015). *Small Area Estimation*. 2nd Edition. New York: John Wiley & Sons, Inc.

Sugasawa, S., Tamae, H. and Kubokawa, T. (2017). Bayesian estimators for small area models shrinking both means and variances. *Scandinavian Journal of Statistics*, 44, 150-167.

US Census Bureau (2021). SAHIE 2008 – 2015 demographic and income model methodology: Summary for counties and for states. https://www.Census.gov/Small Area Health Insurance Estimates (SAHIE) Program/Technical Documentation/Methodology/Demographic and Income Model Methodology (2008-2015).

Yan, G., and Sedransk, J. (2011). Improved inference for a linear mixed-effects model when the subpopulations are clustered. *Journal of Statistical Planning and Inference*, 141, 3489-3497.

You, Y., and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32, 1, 97-103. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006001/article/9263-eng.pdf.