

Combining Estimates from Multiple Surveys

Michael R. Elliott^{1,2} Trivellore E. Raghunathan^{1,2} Nathaniel Schenker^{3,4}

¹Department of Biostatistics, School of Public Health, University of Michigan

²Survey Methodology Program, Institute for Social Research, University of Michigan

³Joint Program in Survey Methodology, University of Maryland

⁴National Institute of Statistical Sciences

Abstract

Combining estimates from multiple surveys can be very useful, especially when the question of interest cannot be addressed well by a single, existing survey. In this paper, we provide a brief review of methodology for combining estimates, with a focus on dual frame, weighting-based, joint-modeling, missing-data, and small-area methods. Many such methods are useful in situations outside the realm of combining estimates from surveys, such as combining information from surveys with administrative data and combining probability-sample data with non-probability sample, or “big” data. We also provide examples of comparability issues that must be kept in mind when information from different sources is being combined.

Key Words: Administrative data; dual frame; imputation; missing data; non-probability samples; small-area estimation; weighting.

1 Introduction

Increasing demands for information for policy-relevant research cannot always be met with any single survey. Hence, a statistical framework that combines information from multiple sources is urgently needed. The desire to combine data from several surveys can arise for many reasons (Schenker and Raghunathan, 2007) (Lohr and Raghunathan, 2017). In the simplest case, larger sample sizes and thus greater information and precision can be obtained by combining data from two or more independent surveys. More commonly, there may be several surveys that have complementary strengths and weaknesses – for example, one may have a more complete sampling frame, a higher response rate, or smaller measurement error, while the other survey may be larger in sample size, or contain information about small-area identifiers. In other settings different sets of variables may be available on different surveys.

We provide an overview of research in the area of combining data across multiple surveys. A more comprehensive review is given in Lohr and Raghunathan (2017). We consider traditional methods that typically assume the surveys have similar frame and measurement error characteristics, as well as more recent methods that are tailored toward taking advantage of the differing characteristics of the different surveys. We broadly classify this work into weighting-based approaches, joint-modeling approaches, and missing-data and imputation approaches. Because small-area estimation has been a key application of these combining efforts, we consider several examples of this as well.

2 Traditional Dual Frame Estimation and Extensions

Perhaps the earliest version of combining estimates from multiple surveys comes from the use of multiple frames. Harley (1962) provided the earliest formal method for combining data obtained from two sampling frames (Kalton, 2014) A and B under general conditions (no overlap, partial overlap, and subsetting). Assuming that the numbers of the elements in frame A only, frame B only, and the overlap of the two frames are known and given by N_a , N_b , and N_{ab} , and that elements from the overlapping frames could be ascertained after sampling, he proposed an estimator of a population total Y as

$$\hat{Y}(p) = N_a \bar{y}_a^A + N_b \bar{y}_b^B + N_{ab}(p \bar{y}_{ab}^A + (1-p) \bar{y}_{ab}^B) = \hat{Y}_a^A + \hat{Y}_b^B + \hat{Y}_{ab}(p) \quad (1)$$

where $\hat{Y}_a^A = N_a \bar{y}_a^A$, $\hat{Y}_b^B = N_b \bar{y}_b^B$, and $\hat{Y}_{ab}(p) = p \hat{Y}_{ab}^A + (1-p) \hat{Y}_{ab}^B$. Under simple random sampling (SRS) (McLeod, 2014), it can be shown that the approximate variance of (1) is minimized when $p = \frac{f_A}{f_A + f_B}$ with $f_A = n_A/N_A$ for $n_A = n_a + n_{ab}$ and $N_A = N_a + N_{ab}$, and similarly f_B , when n_A and n_B are fixed (i.e., the estimator of the mean in the overlapped portion of the frames is a weighted sum of the means from each frame, weighted proportional to the sampling fractions). However, an optimal allocation for n_A and n_B can be found as a quartic function of the costs, variances, and proportions of the frames that overlap. In the more common setting where N_{ab} is unknown in advance and/or a more complex sample design is used, we can rewrite (1) as $\hat{Y}(p) = \hat{Y}(0) + p \Delta_Y$ for $\Delta_Y = \hat{Y}_{ab}^A - \hat{Y}_{ab}^B$, using design-based estimators for $\hat{Y}_a$, $\hat{Y}_b$, $\hat{Y}_{ab}^A$, and $\hat{Y}_{ab}^B$. Minimizing $\text{var}(\hat{Y}(p))$ yields $\tilde{p} = \text{var}(\hat{Y}(0))/\text{cov}(\hat{Y}(0), \Delta_Y)$, which must be estimated from the data; thus $\tilde{p}$ is typically a non-linear function of each response variable.

Fuller and Burmeister (1972) considered

$$\hat{Y}_{srs} = (N_A - \hat{N}_{ab,srs}) \bar{y}_a^A + (N_B - \hat{N}_{ab,srs}) \bar{y}_b^B + \hat{N}_{ab,srs} \bar{y}_{ab} \quad (2)$$

where $\bar{y}_{ab} = (n_{ab}^A \bar{y}_{ab}^A + n_{ab}^B \bar{y}_{ab}^B)/(n_{ab}^A + n_{ab}^B)$ and $\hat{N}_{ab,srs}$ is obtained as the smallest root of the quadratic equation $(n_A + n_B)N_{ab}^2 - (n_A n_b + n_b n_a + n_{ab}^A N_A + n_{ab}^B N_B)N_{ab} + (n_{ab}^A + n_{ab}^B)N_A N_B = 0$. Fuller and Burmeister showed that $\hat{N}_{ab,srs}$ is the maximum likelihood estimator of N_{ab} where N_a , N_b , N_{ab}^A , and N_{ab}^B follow a hypergeometric distribution under SRS, and is more efficient than (1) since it takes account of the fact the N_{ab} is estimated from the data. Skinner and Rao (1996) extended (2) by replacing $\hat{N}_{ab,srs}$ with its pseudo-maximum likelihood estimator (PLME) (Binder, 1983) under a general complex sample design, showing this could be obtained by replacing the quantities in the quadratic equation for $\hat{N}_{ab,srs}$ with their design-consistent estimators. The advantage of $\hat{Y}_{srs}$ and its PMLE equivalent is that they are linear in y , allowing a single set of weight adjustments. Lohr and Rao (2000) reviewed these estimators, along with the “single frame” estimators of Bankier (1986) and Skinner (1991) that simply

replace the weights of the overlapping elements with the inverses of the sums of the probabilities of selection in the frames. Lohr and Rao (2000) showed that the single frame estimators are less efficient and developed a simple jackknife weighting estimator that essentially pools the samples into a single “pseudo-frame” and that provides better nominal coverage than the closed form Taylor Series approximation variance estimators for small and medium sample sizes.

Dong et al. (2014) considered a variation on these traditional methods described above that creates synthetic populations from each of the surveys using a nonparametric finite population Bayesian bootstrap, and then obtains an estimate of the population quantity of interest as the weighted sum of the population estimates obtained from the individual surveys, weighted by the precisions of the estimates. This simple estimator is made possible by the fact that the synthetic populations created using Dong et al. are free of design effects, and can be combined in proportion to their precisions (assuming that each of the surveys obtained unbiased estimates). Dong et al. showed that the estimated efficiency of estimates of health insurance coverage in the population obtained by combining three surveys (National Health Interview Survey [NHIS], Medical Expenditure Panel Survey [MEPS] and Behavioral Risk Factor Surveillance System [BRFSS]) (Arnett, 2014) can be increased by several factors over estimators obtained from any one of the single surveys.

3 Weighting-Based Approaches

The traditional approaches reviewed in Section 2 assume that the variables of interest are measured identically in both sampling frames. In practice, some variables may appear in only one of the samples, or the error qualities of the two surveys might differ – one survey might be subject to greater response bias (Frankel, 2014) or measurement error (Stefanski, 2014) than the other survey for any given variable. Most of the remaining methods we discuss will be targeted to these settings, although they are also applicable when the variables measured in two samples are of identical quality.

The first of these approaches (Elliott and Davis, 2005) uses weights, and assumes that the two samples are from surveys a and b , where a is assumed to be of higher quality than b (e.g., higher response rate, reduced mode effects, etc.) and considers adjustments to design weights to reduce bias in b . Elliott and Davis proposed using adjusted weights for b of the form

$$w_i^b = d_i^b \frac{\hat{P}(S_i = a | y_i, \mathbf{x}_i) / \hat{P}(S_i = a)}{\hat{P}(S_i = b | y_i, \mathbf{x}_i) / \hat{P}(S_i = b)} = d_i^b \frac{\hat{P}(S_i = a | y_i, \mathbf{x}_i)}{\hat{P}(S_i = b | y_i, \mathbf{x}_i)} \frac{\hat{P}(S_i = b)}{\hat{P}(S_i = a)} \quad (3)$$

where $S_i = s$, $s \in \{a, b\}$ is an indicator for survey membership, d_i^s is the design weight associated with element i in survey s , y_i is the outcome of interest, and $\mathbf{x}_i$ is a vector of covariates. (3) can be obtained by pooling the two surveys together with their respective design weights and estimating the conditional odds ratio (first fraction on the right side) by weighted logistic regression on y_i and $\mathbf{x}_i$, and the unconditional odds ratio (second fraction on the right side) from the weighted sample sizes. Elliott and Davis (2005) showed that distributions estimated using the adjusted weights from (3) applied to the data from b will approximately match the distributions estimated from a . They then applied the weight adjustment method to a small-area estimation problem (a common application of combining surveys) where survey b – the BRFSS – contains the small-area identifiers (US counties), and survey a – the NHIS – lacks those identifiers, but, as a face-to-face survey, has a much higher response rate and better frame coverage. This application requires the assumption that $\frac{\hat{P}(S_i=a|y_i, \mathbf{x}_i, G_i=g)}{\hat{P}(S_i=b|y_i, \mathbf{x}_i, G_i=g)} = \frac{\hat{P}(S_i=a|y_i, \mathbf{x}_i, R_i=r)}{\hat{P}(S_i=b|y_i, \mathbf{x}_i, R_i=r)}$ for all small areas g inside “large” area r that is available in both surveys. Variance estimation using the adjusted weights is obtained via a jackknife estimator that incorporates pseudo-estimators obtained by pooling the survey samples, as in Lohr and Rao (2000). Elliott and Davis (2005) provided evidence that the direct small-area estimates of smoking rates and mammogram usage obtained from the BRFSS using the adjusted design weights change in directions consistent with reductions in bias estimated from comparisons with the NHIS. Because the direct weight adjustments can increase variance, Elliott and Davis (2005) also considered a hybrid estimator that estimates the mean square error (MSE) of the unadjusted estimator as $(\hat{\theta}_g^{unadj} - \hat{\theta}_g^{adj})^2 + \hat{V}(\hat{\theta}_g^{unadj})$, estimates the MSE of the adjusted estimator as $\hat{V}(\hat{\theta}_g^{adj})$, and selects the estimator with the smallest estimated MSE. Using the Current Population Survey – Tobacco Use Supplement as a “gold standard” in 165 US counties, Elliott and Davis (2005) showed that the NHIS-adjusted and hybrid estimators generally had improved bias estimates over the unadjusted estimates, and that the hybrid estimator had smaller MSE than either of the others.

4 Joint-Modeling Approaches

Raghunathan et al. (2007) considered an alternative approach to the same application as that of Elliott and Davis (2005), combining BRFSS and NHIS data to produce small-area estimates for county-level risk factor behavior, focusing on a joint model to combined the observed direct small-area estimates. (In contrast to Elliott and Davis (2005), Raghunathan et al. (2007) had access to small-area indicators for the NHIS as well as for the BRFSS.) Raghunathan et al. separated the direct estimates into those obtained from the telephone households (p_{xj}) and those obtained from the non-telephone households in the NHIS (p_{yj}), and those obtained from (telephone) households in the BRFSS (p_{zj}), for small area j . They then assumed a trivariate normal distribution for the joint distribution of these estimates after an arcsine-square root transformation:

$$\begin{pmatrix} x_j = \sin^{-1} \sqrt{p_{xj}} \\ y_j = \sin^{-1} \sqrt{p_{yj}} \\ z_j = \sin^{-1} \sqrt{p_{zj}} \end{pmatrix} \sim \left[\begin{pmatrix} \theta_j \\ \phi_j \\ (1 + \delta_j)\theta_j \end{pmatrix}, (1/4) \begin{pmatrix} \tilde{n}_{xj}^{-1} & \rho(\tilde{n}_{xj}\tilde{n}_{yj})^{-1/2} & 0 \\ & \tilde{n}_{yj}^{-1} & 0 \\ & & \tilde{n}_{zj}^{-1} \end{pmatrix} \right] \quad (4)$$

where $\tilde{n}_{rj} = n_{rj}/d_{rj}$ for design effect d_{rj} in area j , and $r = x, y, z$. The arcsin-square root transformation stabilizes the variance as a function of the sample size, allowing the effective sample sizes $\tilde{n}_{rj}$ to incorporate the complex sample designs of the two surveys. The parameter δ_j is the proportionate bias in the BRFSS, while ρ , the correlation between the NHIS telephone and non-telephone household rates, is assumed constant across the sample areas and is fixed at a value estimated from the national-level data. Raghunathan et al. tied together the county-level means with a prior normal regression on county-level population covariates $\mathbf{U}_j$:

$$\begin{pmatrix} \theta_j \\ \phi_j \\ \delta_j \end{pmatrix} \sim N_3(\beta \mathbf{U}_j, \Sigma) \quad (5)$$

with weak hyperiors placed on β and Σ , and they estimated the parameters of their model via Gibbs sampling (Christen, 2014). The inferential quantities of interest were $\mu_j = M_j \sin^2 \theta_j + (1 - M_j) \sin^2 \phi_j$, where the fraction M_j of households equipped with telephones was obtained from the US Decennial Census. Raghunathan et al. (2007) showed that the combined NHIS/BRFSS estimates of smoking and mammography rates were more stable than the BRFSS-only estimates. Moreover, the combined estimates for smoking and mammography were larger and smaller, respectively, than the corresponding BRFSS-only estimates, consistent with reductions in “middle class” bias due to inclusion of NHIS data with higher response rates and better sampling frame coverage.

Other broadly similar joint-modeling approaches have been considered. Wang et al. (2011) considered the estimation of US corn yield by combining three surveys (Agricultural Yield Survey [AYS], Objective Yield Survey [OYS], and December Agriculture Survey [DAS]). The surveys are conducted with different goals. The AYS is a survey of producers, and the OYS is a field measurement survey; both are conducted monthly during the harvest season (August-December). The DAS is a single survey of producers in December at the end of the harvest season. The DAS is considered to be unbiased, while the AYS tends to be negatively biased and the OYS positively biased. As in Raghunathan et al. (2007), the surveys are considered independent conditional on a survey-specific mean, with a within-year correlation for the AYS and OYS surveys accommodated with an AR(1) model. The vector of months means for the AYS and OYS models, θ_{tk} , $k = \{\text{AYS, OYS}\}$, is linked to the mean of the DAS data, μ_t , by a second AR(1) model, with offset $\mathbf{b}_k$ to accommodate bias (assumed to possibly differ by month, but as a constant level over the years). Finally, the scalar μ_t , the true yield for year t , is modeled as a linear function of year and weather and planting variables, with a compound symmetric error structure across years (this correlation cannot be estimated and thus treated as a sensitivity parameter). A fully Bayesian model is then specified using diffuse hyperpriors, and estimation conducted using Gibbs sampling. Comparing the results of this estimator with a simpler precision-weighted combination of the AYS and OYS estimators (Keller and Olkin, 2002) to predict final December yield from the August to November AYS and OYS data, Wang et al. (2011) found that the three-survey estimate has substantial reductions in RMSE across the six years considered (2004-2009), with greater savings the larger the across-year correlation assumed. Cheng et al. (2015) considered a random error model to combine data from four US Census surveys over a 10-year period to estimate the total number of households in the US. Chen et al. incorporated time-invariant random bias estimates associated with each survey under a sum-to-zero constraint (in the absence of a gold standard), a random-walk error term across years form a truncated (positive) normal prior, and a fixed-effects regression on total population.

5 Missing-Data and Imputation-Based Approaches

In situations where multiple surveys contain different outcomes of interest along with common covariates, combining the datasets for the purposes of imputing the variables not observed in a given survey is a logical approach. Such methods might be particularly valuable when the survey with the missing variable is much larger than the survey with the observed variable, and/or where the available common variables are highly predictive of the outcome. Schenker et al. (2010) considered a setting where health measures such as hypertension, diabetes, and obesity are available only as self-reports in the NHIS, while these same measures are available as both self-reports and direct biometric measures in the National Health and Nutrition Examination Survey (NHANES). Associations between the outcome variable and predictors (including the self-reported status) were estimated via a logistic regression model:

$$\text{logit}(P(Y_i^{NHANES})) = \beta^{NHANES^T} \mathbf{X}_i^{NHANES} \quad (6)$$

and the predicted probabilities for the biometric measures in NHIS were obtained as

$$\frac{\exp(\hat{\beta}^{NHANES^T} \mathbf{X}_i^{NHIS})}{1 + \exp(\hat{\beta}^{NHANES^T} \mathbf{X}_i^{NHIS})} \quad (7)$$

Multiply imputed datasets (Barnard et al., 2014) were then obtained by drawing a Bernoulli variable from (7), with inference carried out using standard multiple-imputation (MI) formulas (Rubin, 1987). (The variance estimated using standard MI is conservative in this setting; the bias in variance estimation decreases as the sample size of the missing-data dataset increases relative to the sample size of the complete-data dataset, and as the predictive power of the common covariates increases.) To relax the assumption that the NHIS dataset is balanced relative to the NHANES, even conditional on X , the logistic regression in (6) was carried out separately conditional on propensity quintiles defined as the estimated propensity for an NHIS case to be in the NHANES sample conditional on X . Schenker et al. (2010) showed considerable bias adjustment (health problems tend to be underreported in self-reports) and dramatic variance reduction, as the NHIS is approximately 17 times larger than the NHANES. They also considered regression estimators obtained using NHIS multiply imputed data, finding little bias reduction and some evidence of slight attenuation, but large variance reductions as well.

Kim and Rao (2012) considered a setting similar to that of Schenker et al. (2010), with a small survey b consisting of complete data on the outcome variable y_i and predictors $\mathbf{x}_i$, and a large survey a consisting of $\mathbf{x}_i$ only, where the publicly-available data will be restricted to dataset a . Kim and Rao (2012) considered a projection estimator

$$\hat{Y} = \sum_{i \in a} w_i \hat{y}_i \quad (8)$$

where $\hat{y}_i = m(\hat{\beta}, \mathbf{x}_i)$. If β is estimated using the design-based estimating equation

$$\sum_{i \in b} w_i (y_i - m(\hat{\beta}, \mathbf{x}_i)) = 0 \quad (9)$$

then (8) yields a consistent estimator of a population total. Using the work of Godambe and Thompson (1986), an equation satisfying (9) can be obtained by assuming $E(y_i | \mathbf{x}_i) = m(\beta, \mathbf{x}_i) = m_i$, $\text{var}(y_i | \mathbf{x}_i) = \sigma^2 s(m_i)$, $\text{cov}(y_i, y_j | \mathbf{x}_i, \mathbf{x}_j) = 0$ for $i \neq j$, and solving for β in

$$\sum_{i \in b} w_i \frac{\partial m_i}{\partial \beta} (y_i - m_i) = 0. \quad (10)$$

In the case where the working model is a generalized linear model with $m_i = g^{-1}(\mathbf{x}_i^T \beta)$, $\frac{\partial m_i}{\partial \beta} = \mathbf{x}_i$, and (9) will hold as long as the first term of $\mathbf{x}_i$ is equal to 1. Thus this is a model-assisted approach, as opposed to a model-based approach (Schenker et al., 2010), and does not require $E(y_i | \mathbf{x}_i) = m_i$, although close approximations will improve the efficiency of the estimator. Kim and Rao (2012) also developed a jackknife variance estimator (Friedl and Stampfer, 2014) that creates K jackknife pseudo-estimates of β from survey b and uses these to generate K jackknife pseudo-estimates of Y , updating the weights in (8) with the appropriate K jackknife replicate weights in survey a . This requires the K pseudo-estimates be created for the jackknife in each of the two datasets.

Chipperfield et al. (2012) also considered a multiple-imputation approach for combining two surveys to estimate a population total, based on a quasi-hot deck method that replaces missing values with means from imputation classes

formed from available common covariates. The total estimator is then obtained as $\hat{Y} = \psi\hat{Y}^I + (1 - \psi)\hat{Y}^A$, where $\hat{Y}^I$ is the estimator obtained from the imputed survey, and $\hat{Y}^A$ from the complete-data survey, and ψ is determined by minimizing the variance of $\hat{Y}$ as in (1):

$$\psi = \frac{\text{var}(\hat{Y}^I) - \text{cov}(\hat{Y}^I, \hat{Y}^A)}{\text{var}(\hat{Y}^I) + \text{var}(\hat{Y}^A) - 2\text{cov}(\hat{Y}^I, \hat{Y}^A)} \quad (11)$$

where $\text{var}(\hat{Y}^I)$ is computed using a variant of standard multiple-imputation approaches and $\text{var}(\hat{Y}^A)$ is the standard design based estimator. (This approach is termed “mass imputation” to distinguish it from standard multiple imputation, since only the mean is imputed within an imputation class, rather than a draw from a predictive distribution.) This approach assumes missingness is random within the imputation cells and that similar measurement error properties hold between the two surveys; Chipperfield et al. provided diagnostics to assess these assumptions.

The problems considered above are similar in structure to problems of bridging across changes in classification systems. In bridging problems, a sample that has been “double-coded” according to old and new classification systems is used to estimate a bridging model, which is then applied to another sample that has been coded using only one of the systems to predict the classifications under the second system. Schenker (2004) discussed applications of bridging to changes in industry and occupation coding for U.S. Census public-use files and changes in reporting of race in Federal data collections.

6 Small-Area Estimation

Small-area estimation (Rao, 2014) is a key application for combining data from multiple surveys, as can already be seen from some of the examples above. Sanchez et al. (2008) considered a variant of Raghunathan et al. (2007), where there is a main survey with outcome Y_{ji} for observation $i = 1, \dots, n_j$ in small area $j = 1, \dots, J$ that follows a generalized linear model of the form $g(\mu_{ji}) = \beta_0 + \beta_1\theta_j + \beta_2^T \mathbf{X}_{ji}$, where $E(Y_{ji}) = \mu_{ji}$, $g(\cdot)$ is a known link function, $\mathbf{X}_{ji}$ is a set of subject-level covariates, and θ_j is the population mean of a relevant neighborhood measure estimated from an ancillary survey using observed $U_{jk} \mid \theta_j \sim N(\theta_j, \sigma^2)$. This base model can be extended by considering interactions between θ_j and $\mathbf{X}_{ji}$ in the outcome model, and by regressing θ_j on known population-level characteristics $\mathbf{Z}_j$ of the small area: $\theta_j \sim N(\gamma^T \mathbf{Z}_j, \tau^2)$. The key to improving small-area estimation using this model is the strength of the association between Y_{ji} and θ_j measured by β_1 . Estimation can proceed by a joint likelihood approach, or a two-stage method that first obtains empirical Bayes estimates of θ_j using the ancillary survey only and plugs in the resulting values into the generalized linear model for the main outcome data. Sanchez et al. (2008) applied this model to the estimation of hypertension in 436 neighborhoods (US Census tracts) using the Multi-Ethnic Study of Atherosclerosis (MESA) for the main outcome survey and an ancillary survey measuring healthy food availability.

Many of the previous methods assume that the survey data for analysis is available at the individual level. In some small-area estimation settings this is not the case. Thus Manzi et al. (2011) considered combining smoking prevalence estimates for $i = 1, \dots, I$ local areas (LAs) in eastern England from $j = 1, \dots, J$ data sources that provide point estimates y_{ij} and confidence intervals and allow computation of variances σ_{ij} treated as known. Manzi et al. proposed the model

$$\begin{aligned} y_{ij} \mid \delta_{ij} &\sim N(\theta_i + \delta_{ij}, \sigma_{ij}^2) \\ \delta_{ij} &\sim N(\mu_j, \tau_j^2) \end{aligned} \quad (12)$$

This assumes a common average bias μ_j for each survey across the LAs, so that θ_i , the mean of the smoking rate in the i th LA, is the parameter of main inference. To maintain identifiability, the mean smoking rate across all of the LAs, $I^{-1} \sum_{i=1}^I \theta_i$, is treated as known and is obtained from yet another source (the UK General Household Survey). Flat independent priors are placed on μ_j and τ_j , and estimation proceeds via Gibbs sampling.

Maples (2017) considered a model related to that used Kim and Rao (2012), but focused toward small-area estimation. Maples replaced the design-based estimating equation (9) from the small sample b with a mixed-effect generalized linear model for exponential distribution f :

$$y_{ij} \mid z_i \sim f(y_{ij}; \theta_{ij}, \phi) \quad (13)$$

$$g(\theta_{ij}) = \beta^T \mathbf{x}_{ij}^T + \sigma z_i$$

where $j = 1, \dots, J$ indexes the small areas, $i = 1, \dots, n_j$ indexes the i th observation in the j th small area, $E(y_{ij} | \mathbf{x}_{ij}) = \theta_{ij}$, $\text{var}(y_i | \mathbf{x}_i) = \phi s(\theta_{ij})$, $z_i \sim N(0, 1)$, and g and s are known functions. The parameter β in (13) is estimated via pseudo-maximum likelihood to account for unequal sampling probabilities in b . The small-area estimator of a population total is then obtained from the large sample survey a using a design-consistent estimator $\hat{Y}_i = \sum_{j \in A_i} w_{ij} E_{z|y}[g^{-1}(\beta^T \mathbf{x}_{ij}^T + \sigma z_i)]$, where $E_{z|y}[g^{-1}(\beta^T \mathbf{x}_{ij}^T + \sigma z_i)] = \frac{\int_{-\infty}^{\infty} [g^{-1}(\beta^T \mathbf{x}_{ij}^T + \sigma z_i)] f(y_{ij}|z_i) f(z_i) dz_i}{\int_{-\infty}^{\infty} f(y_{ij}|z_i) f(z_i) dz_i}$ can be obtained using the PMLEs for β and σ from survey b combined with either analytic or numeric integration. Maples (2017) also developed a sandwich-type design-consistent estimator for the variance and MSE of $\hat{Y}_i$, and considered an application to the estimation of disability levels in US states using the Survey of Income and Program Participation (SIPP) as the high-quality small sample b (approximately 70,000 persons aged 15+ across the 50 US states) and the American Community Survey (ACS) as the large sample a (approximately 3.7 million persons aged 15+ across the 50 US states). Common covariates between the surveys included age, sex, race/ethnicity, and the number of mobility limitations reported.

7 Discussion

As the above article demonstrates, the issues around combining data from multiple surveys are often idiosyncratic to the specific surveys to be combined and the scientific question of interest. While some of the early papers developed general methodologies, they are mainly useful only in somewhat narrow settings where assumptions of unbiased estimation and overlapped sampling frames are reasonable.

There are a number of closely related topics to this issue of combining data from multiple surveys. *Statistical matching* considers settings where the variables of interest X and Y are never observed together in any of the surveys, but there are common covariates Z so that the conditional distributions of $X | Z$ and $Y | Z$ are estimable (Rodgers, 1984) (Moriarity and Scheuren, 2001) (Raessler, 2002). *Data harmonization* considers the problem of combining data from multiple datasets where the characteristics of the questions might differ to small or large degrees across surveys, bringing portability into question (Fortier et al., 2017). Although we have emphasized combining information from multiple surveys, joint-modeling and imputation approaches can also be used to *combine information from surveys and administrative data sources* (Citro, 2014). For an example of a project combining survey and non-survey sources, see He et al. (2014).

Though combining data from multiple sources may have considerable benefits, one also should be careful about the comparability of the candidate surveys. Here we provide five such comparability issues that should be considered when developing methods for combining information (see Schenker et al. (2002) and Schenker and Raghunathan (2007) for more examples):

1. Differences in the type of respondents and source of responses. For example, one may be a face-to-face interview of respondents reporting on health conditions and the other may be a physician reporting about the patients based on medical records.
2. Differing modes of data collection: Mail, Telephone, face-to-face or a mix.
3. Survey context: One survey may be conducted by a well known Federal Agency and another may be a reputed institution, but not that well known. Response error properties might differ in the two surveys.
4. Differences in the survey design. For example, the NHIS is a face-to-face survey and the NHANES involves a face-to-face survey as well as measurement and laboratory components. Respondent recalling abilities may differ under these two survey-design settings.
5. Differences in the question wording or the placement of the questions with the same wording may provide different stimuli to respondents and, hence, different error properties. This may be even more important when information is combined from a survey (where every respondent receives the same stimuli) and an administrative data source (absence of or unknown nature of stimuli).

With these challenges, *combining data from a mix of probability and non-probability sources* provides exciting opportunities for the increasing world of “big data,” where large quantities of poor or unknown quality data in terms of representativeness and measurement error can be improved with the use of high quality probability sample data; see Elliott and Valliant (2017) for a recent review.

8 Related Articles in Wiley StatsRef

Braverman, Amy J. (2014). Data Fusion. *Wiley StatsRef: Statistics Reference Online*. Wiley.

Dalla Valle, L. (2014). Data Integration. *Wiley StatsRef: Statistics Reference Online*. Wiley.

Rodgers, W.L. (2014). Statistical Matching. *Wiley StatsRef: Statistics Reference Online*. Wiley.

References

Arnett, R. (2014). Health services data sources in the us. In *Wiley StatsRef: Statistics Reference Online*. Wiley.

Bankier, M. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81:1074–1079.

Barnard, J., Rubin, D. B., and Schenker, N. (2014). Multiple imputation methods. In *Wiley StatsRef: Statistics Reference Online*. Wiley.

Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51:279–292.

Cheng, Y., Chakraborty, A., and Datta, G. (2015). Hierarchical bayesian methods for combining surveys. In *Proceedings of the Survey Research Methods Section*, pages 4099–4111. American Statistical Association.

Chipperfield, J., Chessman, J., and R., L. (2012). Combining household surveys using mass imputation to estimate population totals. *Australian and New Zealand Journal of Statistics*, 54:223–238.

Christen, J. (2014). Gibbs sampling. In *Wiley StatsRef: Statistics Reference Online*. Wiley.

Citro, C. F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, 40:137–161.

Dong, Q., Elliott, M., and Raghunathan, T. (2014). Combining information from multiple complex surveys. *Survey Methodology*, 40:347–354.

Elliott, M. and Davis, W. (2005). Obtaining cancer risk factor prevalence estimates in small areas: combining data from two surveys. *Journal of the Royal Statistical Society*, C54:595–609.

Elliott, M. and Valliant, R. (2017). Inference for non-probability samples. *Statistical Science*, 32:249–264.

Fortier, I., Raina, P., Van den Heuvel, E., Griffith, L., Craig, C., Saiba, M., Doiron, D., Stolk, R., Knoppers, B., Ferretti, V., Granda, P., and Burton, P. (2017). Malestrom research guidelines for rigorous retrospective data harmonization. *International Journal of Epidemiology*, 46:103–115.

Frankel, M. (2014). Response bias. In *Wiley StatsRef: Statistics Reference Online*. Wiley.

Friedl, H. and Stampfer, E. (2014). Jackknife resampling. In *Wiley StatsRef: Statistics Reference Online*. Wiley.

Fuller, W. and Burmeister, L. (1972). Estimators for samples selected from two overlapping frames. In *Proceedings of the Social Statistics Section*, pages 245–249. American Statistical Association.

Godambe, V. and Thompson, M. (1986). Parameters of superpopulation and survey population: their relationship and estimation. *International Statistical Review*, 54:127–138.

- Harley, H. (1962). Multiple frame surveys. In *Proceedings of the Social Statistics Section*, pages 203–206. American Statistical Association.
- He, Y., Landrum, M. B., and Zaslavsky, A. M. (2014). Combining information from two data sources with misreporting and incompleteness to assess hospice use among cancer patients: a multiple imputation approach. *Statistics in Medicine*, 33:3710–3724.
- Kalton, G. (2014). Survey sampling. In *Wiley StatsRef: Statistics Reference Online*. Wiley.
- Keller, T. and Olkin, I. (2002). Combining correlated unbiased estimators of the mean of a normal distribution. Technical report, National Agricultural Service.
- Kim, J. and Rao, J. (2012). Combining data from two independent surveys: a model-assisted approach. *Biometrika*, 99:85–100.
- Lohr, S. and Raghunathan, T. E. (2017). Combining survey data with other data sources. *Statistical Science*, 32:293–312.
- Lohr, S. and Rao, J. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, 95:271–280.
- Manzi, G., Spiegelhalter, D., Turner, R., Flowers, J., and Thompson, S. (2011). Modelling bias in combining small area prevalence estimates from multiple surveys. *Journal of the Royal Statistical Society, A174*:31–50.
- Maples, J. (2017). Improving small area estimate of disability: combining the american communities survey with the survey of income and program participation. *Journal of the Royal Statistical Society, Series A*, In Press.
- McLeod, A. I. (2014). Simple random sampling. In *Wiley StatsRef: Statistics Reference Online*. Wiley.
- Moriarty, C. and Scheuren, F. (2001). Statistical matching: a paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics*, 17:407–422.
- Raessler, S. (2002). *Statistical Matching*. Springer, New York.
- Raghunathan, T., Xie, D., Schenker, N., Parsons, V., Davis, W., Dodd, K., and Feuer, D. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *Journal of the American Statistical Association*, 102:474–486.
- Rao, J. N. K. (2014). Small-area estimation. In *Wiley StatsRef: Statistics Reference Online*. Wiley.
- Rodgers, W. (1984). An evaluation of statistical matching. *Journal of Business and Economic Statistics*, 2:91–102.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.
- Sanchez, B., Raghunathan, T., Diez Roux, A., and Lee, O. (2008). Combining data from primary and ancillary surveys to assess the association between neighborhood-level characteristics and health outcomes: the multi-ethnic study of atherosclerosis. *Statistics in Medicine*, 27:5745–5763.
- Schenker, N. (2004). Bridging across changes in classification systems. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*, pages 117–128. Wiley.
- Schenker, N., Gentleman, J. F., Rose, D., Hing, E., and Shimizu, I. M. (2002). Combining estimates from complementary surveys: a case study using prevalence estimates from national health surveys of households and nursing homes. *Public Health Reports*, 117:393–407.
- Schenker, N. and Raghunathan, T. (2007). Combining information from multiple surveys to enhance estimation of measures of health. *Statistics in Medicine*, 26:1802–1811.
- Schenker, N., Raghunathan, T., and Bondarenko, I. (2010). Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey. *Statistics in Medicine*, 29:533–545.

- Skinner, C. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86:779–784.
- Skinner, C. and Rao, J. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91:348–356.
- Stefanski, L. (2014). Measurement error. In *Wiley StatsRef: Statistics Reference Online*. Wiley.
- Wang, J., Holan, S., Nandram, B., Barboza, W., Toto, C., and Anderson, E. (2011). A bayesian approach to estimating agricultural yield based on multiple repeated surveys. *Journal of the Agricultural, Biological, and Environmental Statistics*, 17:84–106.