The Survey Statistician, 2025, Vol. 91, 26-33.

New and Emerging Methods

Multiple frame methods for combining data sources

Sharon L. Lohr

Arizona State University, USA
sharon.lohr@asu.edu

Abstract

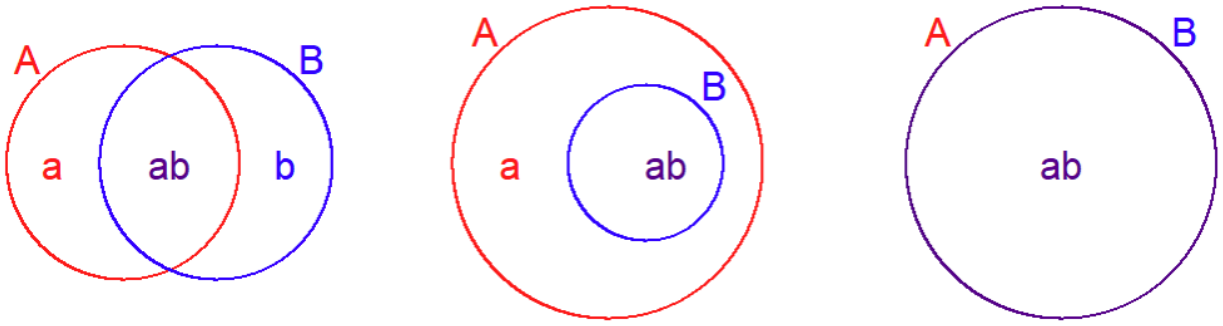
Multiple frame surveys, in which samples are selected independently from overlapping sampling frames, can improve population coverage, increase sample sizes for subpopulations of interest, and reduce costs. I review history and properties of multiple frame surveys when all samples are probability samples. I then discuss extensions of multiple frame theory that can be used with nonprobability samples or when the study variable is measured in only some of the data sources. The assumptions needed for estimators to be unbiased with the claimed variance are strong, and when assumptions are not met a multiple frame approach can increase, rather than decrease, mean squared errors. Finally, I highlight some areas for potential future research.

1 Introduction

With the decreasing response rates of probability samples, many survey organizations have been exploring the use of data from alternative sources to improve, supplement, or replace survey estimates. Sources of statistical information include traditional probability samples; administrative data such as tax or national health service records; commercial data such as credit card transactions or retail sales; location data from cellular telephones; data from satellites and from transportation, environmental, and agricultural sensors; social media and webscraped data; and data from convenience samples. Numerous methods have been developed for combining data from multiple sources; see Lohr and Raghunathan (2017), Yang and Kim (2020), Rao (2021), Kim (2022), National Academies of Sciences, Engineering, and Medicine (2023), and Rao and Lohr (2025) for reviews.

This article reviews combining data from two sources (Lohr, 2021, discussed the general case with more than two data sources) using a multiple frame approach. Let $\mathcal{S}_A$ and $\mathcal{S}_B$ denote the two datasets, and let A and B denote the populations (frames) from which samples $\mathcal{S}_A$ and $\mathcal{S}_B$ are selected. Figure 1 shows frame structures for three dual frame surveys. In Figure 1(a), both frames are incomplete and form three mutually exclusive domains: domain a, which consists of population units

in frame A but not in frame B; domain b, which contains the population units in frame B but not in frame A; and domain ab, which contains the population units in both frames. In Figure 1(b), frame B is a proper subset of frame A so that domain ab coincides with frame B and domain b is empty. In Figure 1(c), the two frames have the same coverage and all units are in overlap domain ab.



(a) (b) (c)

Figure 1: Three frame structures. (a) Frames A and B are both incomplete but overlap. (b) Frame B is a proper subset of frame A. (c) Frames A and B coincide.

Hartley (1962) defined the notation used in Figure 1 and derived point estimators, variances, and optimal allocations when $\mathcal{S}_A$ and $\mathcal{S}_B$ are both probability samples. The motivation for his research came from agriculture, in which area frame A is complete but expensive to sample while frame B, a list of farm operators, is less expensive to sample but incomplete (Figure 1(b)).

Hartley showed that a dual frame approach could yield the same precision as a single frame survey from frame A but with substantial cost savings — because frame B may be less expensive to sample and also because the list frame often contains the largest units (which have the highest variability). Many subsequent developments in multiple frame survey theory have been motivated by agricultural applications (Ferraz et al., 2023). Multiple frame surveys with area and list frames have also long been used to improve coverage and reduce costs for business surveys (Kott and Vogel, 1995).

If the frames contain sufficient information to allow each individual record from frame A to be linked with a record from the same entity in frame B — and to determine when a record in one frame has no counterpart in the other — then record linkage can be used to merge the data sources. If the frames are linked before sampling, the resulting single frame concatenates the records from frames A and B while eliminating duplicates. If records in $\mathcal{S}_A$ and $\mathcal{S}_B$ are linked to frames B and A, respectively, after the samples have been collected, weights for sampled units in domain ab in the concatenated sample are modified so that the population members represented by those units are not counted twice.

Multiple frame methods can combine information from $\mathcal{S}_A$ and $\mathcal{S}_B$, however, even when the identifying information is not sufficient to enable linkage. One need know only whether each record in $\mathcal{S}_A$ and $\mathcal{S}_B$ could also have been sampled from the other frame, not *which* record in the other frame belongs to the same entity as the sampled record. The records in the two samples can be from different types of entities — for example, records in $\mathcal{S}_A$ might be for households while those in $\mathcal{S}_B$ are for individual persons — as long as domain membership can be determined for each unit in $\mathcal{S}_A$ and $\mathcal{S}_B$. Multiple frame methods can also be used when individual records are not available. An organization may be unwilling to disclose individual records but might be willing to share domain summary statistics.

This article reviews multiple frame methods for combining data from different sources with the goal of estimating the population total $Y = \sum_{i=1}^{N} y_i$ from a population of size N. Section 2 reviews classical multiple frame surveys along with typical assumptions made for inference, and summarizes advantages of using multiple sources. Section 3 looks at recent research on extensions of multiple frame methods to other situations. Section 4 examines some pitfalls of using multiple frame methods when

assumptions are not met. Finally, Section 5 discusses potential research that could extend these methods.

2 Dual frame assumptions and estimation

For the general dual frame survey in Figure 1(a), the population total can be written as $Y = Y_a + Y_{ab} + Y_b$, where $Y_a = \sum_{i \in A}(1-\delta_i)y_i$, $Y_b = \sum_{i \in B}(1-\delta_i)y_i$, $Y_{ab} = \sum_{i \in A}\delta_i y_i = \sum_{i \in B}\delta_i y_i$, and $\delta_i = 1$ if unit i is in the overlap domain ab and 0 otherwise. If w_i^A and w_i^B are final weights for the two samples $\mathcal{S}_A$ and $\mathcal{S}_B$, the domain totals can be estimated as $\hat{Y}_a^A = \sum_{i \in \mathcal{S}_A} w_i^A(1-\delta_i)y_i$, $\hat{Y}_b^B = \sum_{i \in \mathcal{S}_B} w_i^B(1-\delta_i)y_i$, $\hat{Y}_{ab}^A = \sum_{i \in \mathcal{S}_A} w_i^A \delta_i y_i$, and $\hat{Y}_{ab}^B = \sum_{i \in \mathcal{S}_B} w_i^B \delta_i y_i$. There are two estimators for the overlap domain total Y_{ab}, and much of the multiple frame literature deals with how to combine the two estimates of Y_{ab} to ensure that sampled units appearing in both frames are not "overcounted" when estimating Y. Lohr (2021) summarized methods that have been proposed for addressing multiplicity with samples from two or more frames.

A simple multiplicity-adjusted estimator is of the form

$$\hat{Y} = \hat{Y}_a^A + \alpha\hat{Y}_{ab}^A + (1-\alpha)\hat{Y}_{ab}^B + \hat{Y}_b^B, \tag{1}$$

where α is typically between 0 and 1. This estimator applies to all three situations in Figure 1, with domain b empty for Figure 1(b) and domains a and b empty for Figure 1(c). If α equals 1 (or 0), $\hat{Y}$ is a *screening* estimator: the overlap domain ab is estimated using only observations from frame A (or B) and any observations in ab that are sampled from the other frame are discarded.

The estimator in (1) can be calibrated to information that may be known about the frame population sizes N_A and N_B; the domain population sizes $N_a = \sum_{i \in A}(1-\delta_i)$, $N_{ab} = \sum_{i \in A}\delta_i = \sum_{i \in B}\delta_i$, and $N_b = \sum_{i \in B}(1-\delta_i)$; and other auxiliary variables. Hartley (1962) derived properties of an estimator that poststratifies to N_a, N_{ab}, and N_b:

$$\hat{Y}_{ps} = N_a\hat{\bar{Y}}_a^A + N_{ab}\alpha\hat{\bar{Y}}_{ab}^A + N_{ab}(1-\alpha)\hat{\bar{Y}}_{ab}^B + N_b\hat{\bar{Y}}_{ab}^B, \tag{2}$$

where $\hat{\bar{Y}}_q^F = \hat{Y}_q^F/\hat{N}_q^F$ for $q \in \{a, ab, b\}$ and $F \in \{A, B\}$. Ranalli et al. (2016) presented a thorough treatment of calibration for multiple frame surveys.

The statistical properties of the estimators in (1) and (2) depend on the characteristics of the target population, frames, and samples. The following assumptions affect these properties.

1. The union of frames A and B is the target population.
2. The samples $\mathcal{S}_A$ (of size n_A) and $\mathcal{S}_B$ (of size n_B) are selected independently.
3. The domain membership δ_i is known for each unit i in $\mathcal{S}_A$ and $\mathcal{S}_B$. This assumption is less restrictive than assumptions needed for data linkage, where one needs to be able to match each entity from $\mathcal{S}_A$ with one from frame B (or determine that no such match exists).
4. The domain estimators from $\mathcal{S}_A$ are unbiased, so that $E[\hat{Y}_a^A] = Y_a$ and $E[\hat{Y}_{ab}^A] = Y_{ab}$. This assumption will be met if $\mathcal{S}_A$ is a census or a probability sample with full response (or if the weighting adjustments fully compensate for selection bias), and if there is no measurement error for the study variable y for units in $\mathcal{S}_A$.
5. The domain estimators from $\mathcal{S}_B$ are unbiased so that $E[\hat{Y}_b^B] = Y_b$ and $E[\hat{Y}_{ab}^B] = Y_{ab}$.

If assumptions 1 and 2 are met, the mean squared error (MSE) of $\hat{Y}$ in (1) is

$$\begin{aligned}\text{MSE}(\hat{Y}) = \text{MSE}\left[\hat{Y}_a^A + \alpha\hat{Y}_{ab}^A\right] + \text{MSE}\left[(1-\alpha)\hat{Y}_{ab}^B + \hat{Y}_b^B\right] \\ + 2\,\text{Bias}\left[\hat{Y}_a^A + \alpha\hat{Y}_{ab}^A\right]\text{Bias}\left[(1-\alpha)\hat{Y}_{ab}^B + \hat{Y}_b^B\right].\end{aligned} \tag{3}$$

If the samples satisfy assumptions 1–5, $\hat{Y}$ is an unbiased estimator of Y and the bias terms in (3) are zero. Under these conditions, a multiple frame approach can yield several advantages relative to a single frame survey for estimating population characteristics.

The first advantage is the ability to improve population coverage when both frames are incomplete, as in Figure 1(a). The multiple frame survey has higher population coverage than a single-frame survey from either of the incomplete frames. When cellular telephones began replacing landlines, many telephone survey organizations started taking independent samples from cellular and landline telephone frames to obtain better coverage of the population. The multiple frame approach can also be used to estimate the population size (if it is unknown) by setting $y_i = 1$ for all units i in (1).

The second advantage relates to cost savings/efficiency gains, particularly for the situation in Figure 1(b) where frame A is complete. If data collection is cheaper for frame B than frame A, the multiple frame survey can reduce costs by relying on the less expensive information from B for observations in domain ab. Hartley (1962) calculated the optimal values of n_A, n_B, and α under simple random sampling and demonstrated that a dual frame sample can achieve large reductions in variance relative to a single frame survey from A having the same total cost. The largest gains in efficiency occur when N_{ab}/N is large (so that frame B has high coverage) or the cost to sample a unit in B is much less than the cost to sample a unit in A. If domain membership can be identified for frame A units before sampling, then a screening design in which all units in $\mathcal{S}_A$ have $\delta_i = 0$ is optimal. A screening design may also be desirable if domain membership can be determined inexpensively relative to measurement of y: a first-phase sample from A determines domain membership, and only units in a are sampled in the second phase.

A multiple frame survey design can increase the sample size for members of rare or hard-to-find populations. For example, statisticians are sparse in a complete frame A of the general population, but are highly concentrated in the membership list of the International Statistical Institute. The list frame B gives an inexpensive way of obtaining a large sample of statisticians. To obtain coverage of statisticians who are not members of ISI, one could supplement with a probability sample from frame A, although $\mathcal{S}_A$ would likely contain few (if any) statisticians and the increased coverage from such an approach might not be worth the effort. A more feasible option for improving coverage would be to take additional samples from other organizations that statisticians might join.

3 Extensions of multiple frame survey theory

The theory in Section 2 was developed under the supposition that $\mathcal{S}_A$ and $\mathcal{S}_B$ are both full-response probability samples designed to measure y. In most dual frame agricultural surveys or cellular/landline telephone surveys, for example, the same questionnaire is administered to both samples. The two surveys are designed to be as similar as possible, although in practice domain misclassification, mode effects, or different response rates in the two samples may affect assumptions 3–5.

The remainder of this article assumes that frame A equals the target population, so that the frame structure is described by Figure 1(b) or (c). Lohr (2014) and Kim and Tam (2021) noted that Figure 1(b) includes the special case in which $\mathcal{S}_B$ is a census of its frame. In this situation, the units in $\mathcal{S}_B$ represent themselves alone so that selection bias for $\mathcal{S}_B$ is not an issue. The census of frame

B can come from a set of governmental or commercial records, sensor or satellite data, or even a convenience sample. If $\mathcal{S}_B$ is a set of administrative records, or another large dataset that is already being collected for a nonstatistical purpose, cost savings can be substantial. Using $\mathcal{S}_B$ can reduce respondent burden and provide more granular data for frame B than can be obtained from a typically much smaller probability sample. When assumption 5 is met, $\hat{Y}_{ab}^B = Y_{ab}$ with no sampling error and $\mathcal{S}_A$ is needed only for estimation of Y_a. For many applications, however, the frame B data collection is not designed in concert with $\mathcal{S}_A$ and there are many potential reasons why assumption 5 might not be satisfied in practice (see Section 4).

In some situations corresponding to Figure 1(b), the study variable y is measured in only one of the samples, but both samples measure the same set of auxiliary variables $\mathbf{x}$. Kim (2022) and Wu (2022, 2023) reviewed model-based data integration methods in which y is measured only in a nonprobability sample $\mathcal{S}_B$. One approach is to consider frame B to equal $\mathcal{S}_B$ (so that the population units in domain ab are the units in $\mathcal{S}_B$), develop a regression model on $\mathcal{S}_B$ predicting y from $\mathbf{x}$, apply the model to obtain predicted values $\hat{y}_i$ for units in $\mathcal{S}_A$, and estimate the population total by $\sum_{i\in\mathcal{S}_A}(1-\delta_i)w_i^A\hat{y}_i + \sum_{i\in\mathcal{S}_B} y_i$. Strong assumptions are needed for this estimator and its variance estimator to be approximately unbiased: 1 to 3, 5, and that $\mathbf{x}$ has no measurement error (if unit i is in both frames, $\mathbf{x}_i$ will have the same value if measured in $\mathcal{S}_A$ as it will if measured in $\mathcal{S}_B$). Additionally, it must be assumed that the model developed on domain ab provides unbiased predictions for the units in domain a with accurate measures of the imputation error.

An alternative approach is to assume that $\mathcal{S}_B$ can be generalized to the population in frame A by estimating the inclusion probabilities $\psi_i^B = P(i \in \mathcal{S}_B|\mathbf{x}_i, y_i)$ (using information in $\mathcal{S}_A$) and estimating Y by $\sum_{i\in\mathcal{S}_B} y_i/\hat{\psi}_i^B$. This approach also requires strong additional assumptions — that all inclusion probabilities are positive and that participation in $\mathcal{S}_B$ is conditionally independent of y given $\mathbf{x}$ (see, for example, Savitsky et al., 2023). Liu et al. (2024) explored nonignorable participation mechanisms for $\mathcal{S}_B$.

4 What can go wrong with a multiple frame approach?

A multiple frame survey is more complicated than using data from a single source, and strong assumptions are needed for estimators to be approximately unbiased with the claimed variance. If the assumptions are not met, bias from domain misclassification or nonsampling errors can cause estimates to have higher MSE than single frame estimates from a high-quality probability sample.

Potential concerns include:

- Undercoverage. Assumption 1 posits that the union of the frames contains all population units, but some population units might be missing from all of the frames. Undercoverage can be remedied by including a complete frame such as an area frame (if one is available). If sampling from a complete frame is prohibitively expensive, however, it may be better to simply define the target population to be the union of the frames and limit inference to that population.
- Domain misclassification. The estimator in (1) downweights units in domain ab by α for units in $\mathcal{S}_A$ and $(1-\alpha)$ for units in $\mathcal{S}_B$ because population units in ab can potentially be selected twice. An incorrect multiplicity factor may be applied to observations classified in the wrong domain, potentially causing bias.
- Measurement error. The study variable y may be measured differently in $\mathcal{S}_A$ than in $\mathcal{S}_B$. This is of particular concern when dual frame methods merge samples with different data collection protocols — for example, when $\mathcal{S}_A$ is a designed probability sample and $\mathcal{S}_B$ is from administrative data. Measurement error in $\mathbf{x}$ may affect the properties of calibrated estimators or of the imputations described in Section 3.

- Missing data or duplicate responses. One or both samples may have nonresponse that biases estimates. A convenience sample $\mathcal{S}_B$ may contain multiple responses from the same person, and these must be deduplicated before combining with $\mathcal{S}_A$.

Suppose that assumptions 1 to 4 are met for the situation in Figure 1 (b) but estimates from $\mathcal{S}_B$ are biased. Following Elliott and Haviland (2007), suppose that the variance within each domain is σ^2 and that Bias$(\hat{\bar{Y}}_{ab}^B)/\sigma = E$. If $\mathcal{S}_A$ is a simple random sample, $\mathcal{S}_B$ is a census of its population (frame B), and $p_B = N_{ab}/N = N_B/N$, the MSE of the poststratified dual frame estimator in (2) is larger than MSE$(\hat{Y}_a^A + \hat{Y}_{ab}^A)$ if $E^2 > (1+\alpha)/[(1-\alpha)n_A p_B]$. Just a small amount of bias from $\mathcal{S}_B$ can cause the dual frame estimator to have higher MSE than an estimator based on $\mathcal{S}_A$ alone.

Ang et al. (2024) compared the bias and MSE of various estimators when assumptions 1 to 4 are met and $\mathcal{S}_B$ is a census of its frame. They found that the screening dual frame estimator $\hat{Y}_a^A + \hat{Y}_{ab}^B$ is unbiased with low MSE when $\mathcal{S}_B$ has no measurement error, but has large bias and MSE when y is measured with error in $\mathcal{S}_B$.

5 Discussion

De Waal et al. (2020) pointed out that many statistics currently calculated from probability samples could potentially be produced more quickly and inexpensively by taking advantage of datasets already collected for other purposes. Multiple frame methods can merge data sources to create population estimates and obtain more granular information about subpopulations. If frame B has a high concentration of individuals from a subpopulation of interest and A is a complete population frame, then supplementing $\mathcal{S}_A$ with a sample from B can augment the sample size for the subpopulation.

The same mechanism that augments the sample size, however, also leads to higher precision for the mean of domain ab than the mean of domain a. When $\mathcal{S}_B$ is a census of its population to be supplemented by a screened probability sample from frame A, then each observation in $\mathcal{S}_B$ has weight 1 while observations in $\mathcal{S}_A$ may have high weights. There is much more information about subpopulations in B than in a, and research is needed on how to address this information inequality.

The assumptions needed for multiple frame methods to produce unbiased estimates with accurate measures of uncertainty are strong. The discussions in Sections 3 and 4 assume the existence of a gold standard data source $\mathcal{S}_A$ that produces unbiased estimates for the target population. In the current data world, however, such gold standard sources are becoming rarer and most data sources can be expected to have undercoverage or bias. If the bias terms in (3) have the same sign, using both samples can reinforce the bias.

Multiple frame methods can be used to study relative bias among different data sources. De Waal et al. (2020) described methods that can be used to detect and model measurement error when records can be linked across sources. If records cannot be linked, bias can be studied by comparing population and subpopulation estimators in the overlap domain. Since these in theory have the same expected value, significant discrepancies indicate a bias problem in one or both samples that can be investigated further. Such investigations can identify problems and improve measurement in both data sources, although they will not necessarily detect biases that have the same direction.

Dual frame estimates for the situations in Figure 1 (a) and (b) depend on correctly determining the domain for each observation. Lohr (2011) and Lin et al. (2019) modified estimators to deal with domain misclassification, but these modifications depend on having independent estimates of misclassification probabilities. A better solution is to design the study to reduce potential misclassification. Research is needed on variables that might be included in $\mathcal{S}_A$ and $\mathcal{S}_B$ to reduce domain misclassification, and on estimators that are robust to misclassification. What information is needed to be

able to determine δ_i from each source? If domain membership is uncertain or imputed as in Kim and Tam (2021), how does that uncertainty propagate to the estimates? How does domain misclassification differ across subpopulations, since some subpopulations may have lower quality information for domain identification?

Section 3 discusses applying multiple frame methods to datasets collected for nonstatistical purposes. More research is needed on designing an integrated data system that uses alternative data sources but is flexible enough to handle changes in measurement or availability of a source. A screening design for $\mathcal{S}_A$ may be optimal when $\mathcal{S}_B$ has no sampling or measurement error, but then $\mathcal{S}_A$ has no observations in ab for assessing discrepancies between the data sources (Holmberg et al., 2024). Exploration of multipurpose multiple frame designs — that can detect or be robust to violations of assumptions, provide accurate and cost-effective estimates of population characteristics, and have flexibility for changes in data sources — could yield many benefits for future use of multiple data sources.

References

Ang, L., R. Clark, B. Loong, and A. Holmberg (2024). An empirical comparison of methods to produce business statistics using non-probability data. `https://arxiv.org/abs/2405.14208`, accessed November 4, 2024.

De Waal, T., A. van Delden, and S. Scholtus (2020). Multi-source statistics: Basic situations and methods. *International Statistical Review 88*(1), 203–228.

Elliott, M. N. and A. Haviland (2007). Use of a web-based convenience sample to supplement a probability sample. *Survey Methodology 33*(2), 211–215.

Ferraz, C., F. Mecatti, and J. Torres (2023). Dual frame design in agricultural surveys: Reviewing roots and methodological perspectives. *Statistical Methods & Applications 32*(2), 593–617.

Hartley, H. O. (1962). Multiple frame surveys. In *Proceedings of the Social Statistics Section*, pp. 203–206. Alexandria, VA: American Statistical Association.

Holmberg, A., L. Ang, R. Clark, and B. Loong (2024). Propensity score estimation and optimal sampling design when integrating probability samples with non-probability data. Presented at Statistics Canada International Methodology Symposium, November 1, 2024.

Kim, J. K. (2022). A gentle introduction to data integration in survey sampling. *The Survey Statistician 85*, 19–29.

Kim, J. K. and S.-M. Tam (2021). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review 89*(2), 382–401.

Kott, P. S. and F. A. Vogel (1995). Multiple-frame business surveys. In B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, and P. S. Kott (Eds.), *Business Survey Methods*, pp. 185–203. New York: Wiley.

Lin, D., Z. Liu, and L. Stokes (2019). A method to correct for frame membership error in dual frame estimators. *Survey Methodology 45*(3), 543–565.

Liu, Y., M. Yuan, P. Li, and C. Wu (2024). Statistical inference with nonignorable non-probability survey samples. `https://arxiv.org/abs/2410.02920v1`, accessed November 15, 2024.

Lohr, S. L. (2011). Alternative survey sample designs: Sampling with multiple overlapping frames. *Survey Methodology 37*(2), 197–213.

Lohr, S. L. (2014). When should a multiple frame survey be used? *The Survey Statistician 69*, 17–21.

Lohr, S. L. (2021). Multiple-frame surveys for a multiple-data-source world. *Survey Methodology 47*(2), 229–263.

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

National Academies of Sciences, Engineering, and Medicine (2023). *Toward a 21st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources*. Washington, DC: The National Academies Press.

Ranalli, M. G., A. Arcos, M. d. M. Rueda, and A. Teodoro (2016). Calibration estimation in dual-frame surveys. *Statistical Methods & Applications 25*, 321–349.

Rao, J. N. K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhyā, Series B 83*(1), 242–272.

Rao, J. N. K. and S. L. Lohr (2025). Trends and directions in sample survey theory and methods. *Survey Methodology*, in press.

Savitsky, T. D., M. R. Williams, J. Gershunskaya, and V. Beresovsky (2023). Methods for combining probability and nonprobability samples under unknown overlaps. *Statistics in Transition 24*(5), 1–34.

Wu, C. (2022). Statistical inference with non-probability survey samples (with discussion). *Survey Methodology 48*(2), 283–373.

Wu, C. (2023). Calibration techniques for model-based prediction and doubly robust estimation. *The Survey Statistician 88*, 86–93.

Yang, S. and J. K. Kim (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science 3*, 625–650.