Cross-community Comparison and Multi-frame Weighting in REACH U.S.

Peter K. Kwok[*] Hee-Choon Shin[†] Whitney Murphy[*]
Colm O'Muircheartaigh[*] Angela Debello[‡] Kari Carris[‡]

Abstract
REACH U.S. (Racial and Ethnic Approaches to Community Health Across the U.S.) is an umbrella of community-based programs aimed to eliminate health disparities among racial and ethnic groups. Five of the REACH U.S. programs are based in the Greater Los Angeles areas with various overlapping geographies and target populations, and with different combinations of scientific interests in cardiovascular disease, *diabetes mellitus*, adult immunization, and breast and cervical cancer. We will explore in this paper the potential of making cross-community comparisons, and discuss some of the issues. In particular, we evaluate the performance of Lohr's and Rao's pseudo-maximum likelihood estimator and Mecatti's multiplicity estimator.

Key Words: REACH U.S., Address-based sampling, Community-based health program, Multi-frame weighting, Pseudo-maximum likelihood, Multiplicity

1. Introduction

A traditional, multi-frame estimation problem usually begins with a single survey that cannot satisfactorily cover its target population with just one frame. Additional frames are then patched onto the same sampling base to remedy the deficiency in coverage. In other situations, a frame with good coverage exists but is too expensive to be sampled alone. So additional, cheaper frames are supplemented to reduce cost, even though they may induce bias in coverage. At the end, there is still only one target population and only one estimator.

However, there is another type of multi-frame problem which is not primarily motivated by coverage or cost considerations, but by the requirement to compare multiple surveys intrinsically defined by different frames that just so happen to overlap. This time, the surveys can potentially involve multiple target populations and, hence, multiple estimators. But their objectives and designs may be just right for sharing data among them to form a more efficient combined estimate. We will consider such an example in this paper, namely the REACH U.S. Survey. Racial and Ethnic Approaches to Community Health Across the United States (REACH U.S.) is a funding program sponsored by the Centers for Disease Control and Prevention (CDC) to eliminate health disparities among various racial and ethnic groups throughout the United States. Its survey distinguishes from other traditional, multi-frame surveys in two important ways.

First, the REACH U.S. Survey is actually a group of independently owned surveys rather than a single one. The REACH U.S. program itself has a number of participating grantees that target different local populations across the U.S. for different public health services and studies. For each year within a five-year period, surveys are conducted at some of these communities to measure the health behaviors of the local residents. Instead of conducting a single survey at the national level with a unified set of eligibility requirements, the grantees spread the surveys across their communities, keep each of their own

[*] National Opinion Research Center at the University of Chicago, 55 E. Monroe St., Ste. 2000, Chicago, IL 60603

[†] National Opinion Research Center at the University of Chicago, 1155 E. 60th St., Chicago, IL 60637

[‡] National Opinion Research Center at the University of Chicago, 55 E. Monroe St., Ste. 3000, Chicago, IL 60603

requirements, but share similar sampling and questionnaire designs to reduce cost overheads. The questionnaire is consisted of a common module of general health questions, such as diet and frequency of exercise, followed by additional modules that focus on other more specific issues, such as adult immunization. Only the responses of a subset of these modules are of interest to each grantee. Multiple surveys are usually hard to compare if they ask different questions and adopt potentially very different sampling designs. This is not the case for REACH U.S. Therefore, while the REACH U.S. Survey starts with multiple frames and ends with multiple estimates, it creates a rare and unique opportunity to be treated as a multiple-frame, *single*-estimator problem whenever the eligibility requirements match. Since the REACH U.S. Survey has just begun calling at the time of submission of the first draft, we will use only *simulation data* to make our points here. To make our simulations more relevant to real applications, we base our set-up on the five surveys situated in Los Angeles and Orange Counties, California, where the overlapping is the most complex among all REACH U.S. localities.

Second, the REACH U.S. Survey provides enough details on the geographic domains to make it feasible to combine the surveys together. The REACH U.S. Survey adopts an address-based sampling (ABS) design in the sense that residential households are sampled from an address frame before matching to their primary telephone numbers. This method is contrasted to the traditional, random digit dialing (RDD), which samples randomly generated telephone numbers before screening the respondents for their geographic eligibilities. While both approaches involve a step of geographic verification with the respondents, the ABS approach used by the REACH U.S. Survey provides much greater precision and flexibility during the planning and analysis stages. This point is particularly relevant when we evaluate Mecatti's multiplicity estimator later.

The remaining paper will be divided into four sections. The Methodology section briefly introduces the two estimators and our evaluation set-up. Then the Results section presents numerical evidences to highlight the strengths and weaknesses of these estimators. As mentioned before, only simplified simulations will be considered here. But their set-up should be realistic enough to reflect the challenges of actual multi-frame surveys. The Discussions section will review what we learned from those evidences and the challenges of adopting estimators primarily designed for estimation to solve data sharing problems. Finally, we will end the Conclusions section with a few recommendations.

2. Methodology

Let Y_s be the population responses for $s = 1, \ldots, N$, where N is the number of persons in the target population. In the following simulation, we will assume that each simulated frame totally covers its target population. From here on, we will interchange the words "frame" and "population" without any further qualification until we revisit this issue in the Discussions section. In this study, we will simulate a key variable in the common module that measures the respondent's number of days in poor physical health during the past 30 days, that is,

$$Y_s \sim \text{Poisson}(3.4) \quad \text{i.i.d.} \tag{1}$$

Our objective variable is $Y = \sum_{s=1}^{N} Y_s$, which can be interpreted as the community's monthly total of person-days in poor physical health. The mean value is based on real figures interpolated from other studies. The total person-days variable may possibly have implications for public health policy. But for our purpose, both can be treated as arbitrary devices chosen just for convenience while being plausibly close to values of practical interests. The actual REACH U.S. Survey adopts a clustering design which samples the households randomly, and then selects various numbers of eligible household members according to certain rules. Some communities also have stratification and/or oversampling requirements. But for the sake of simulation, we will just simplify the whole scenario to an one-stage, simple random sampling design. In essence, we consider the hypothetical scenario in which every household has exactly one eligible respondent. We will assume no missing data so that any variation in our simulation results is due to sampling error alone.

We will consider two estimators. The first one is Lohr's and Rao's pseudo-maximum likelihood estimator. The use of pseudo-maximum likelihood (PML) in the dual-frame estimator can be traced back to Skinner and Rao (1996); and the concept of dual-frame estimator can be traced back to Hartley (1962, 1974). Hartley's strategy is to partition two sampling frames A and B into domains $a = A \backslash B$, $b = B \backslash A$, and $ab = A \cap B$, and then estimate a key measure Y by

$$\widehat{Y} = \widehat{Y}_a + \widehat{Y}_b + \theta \widehat{Y}_{ab\,\text{est. on}\,A} + (1-\theta)\widehat{Y}_{ab\,\text{est. on}\,B} \ ,$$

where the subscripts indicate the subset condition on which the measure is aggregated. Building on the above expression, Fuller and Burmeister (1972) added a second parameter θ' for an additional adjusted difference

$$\cdots + \theta'(\widehat{N}_{ab\,\text{est. on}\,A} - \widehat{N}_{ab\,\text{est. on}\,B}) \ .$$

Skinner and Rao (1996) first applied the pseudo-maximum likelihood approach to estimators of the form

$$(N_A - \widehat{N}_{ab})\,\widehat{\overline{Y}}_a + (N_B - \widehat{N}_{ab})\,\widehat{\overline{Y}}_b + \widehat{N}_{ab}\,\widehat{\overline{Y}}_{ab} \ .$$

Lohr and Rao (2006) later generalized the above form to any number of frames. We will briefly explain Lohr's and Rao's estimator below, but refer interested readers to their original paper (Lohr & Rao, 2006) for a full exposition.

From here on, dimension index i always ranges from 1 through d; and j, from 1 through Q. Suppose we have Q frames of sizes $\mathbf{N}^{(Q)} = \left[N_j^{(Q)}\right]_{Q\times 1}$. Their union can be partitioned into d domains (*i.e.*, nonempty, maximal subsets for each combination of frames) of sizes $\mathbf{N}^{(d)} = \left[N_i^{(d)}\right]_{1\times d}$. Let $\mathbf{N} = [N_{i,j}]_{d\times Q}$ denote the sizes of domain i in frame j. In the simulations, $\mathbf{N}$ will be pre-assigned two sets of values. Given $\mathbf{N}$ we independently draw simple random samples of sizes $\mathbf{n}^{(Q)} = \left[n_j^{(Q)}\right]_{Q\times 1}$ with sampling fractions $\mathbf{f} = [f_j]_{Q\times 1} = \left[n_j^{(q)}/N_j^{(q)}\right]_{Q\times 1}$. That is, we randomly draw $n_j^{(Q)}$ out of $N_j^{(Q)}$ cases from frame j. Let $\mathbf{n} = [n_{i,j}]_{d\times Q}$, $\mathbf{y} = [y_{i,j}]_{d\times Q}$, and $\overline{\mathbf{y}} = \left[\overline{y}_{i,j}\right]_{d\times Q} = [y_{i,j}/n_{i,j}]_{d\times Q}$ denote the sample sizes, totals, and means for each combination of domain and frame, respectively. Let $\mathbf{Y} = [Y_{i,j}]_{d\times Q}$ be the population totals by domains and frames, and let $\overline{\mathbf{Y}}^{(d)} = \left[\overline{Y}_i^{(d)}\right]_{d\times 1} = \left[\sum_{j=1}^{Q} Y_{i,j}/N_i^{(d)}\right]_{d\times 1}$ be the population means by domains. The fundamental strategy of pseudo-maximum likelihood estimator is to re-write the true population total $Y = \sum_{s=1}^{N} Y_s$ as $Y = \mathbf{N}^{(d)}\overline{\mathbf{Y}}^{(d)}$, and then estimate by

$$\widehat{Y} = \widehat{\mathbf{N}}^{(d)}\widehat{\overline{\mathbf{Y}}}^{(d)}. \tag{2}$$

For sufficiently large samples, we have approximately $\overline{y}_{i,j} \mid n_{i,j} \sim N\left(\ \overline{Y}_i^{(d)},\ \sigma_i^2/n_{i,j}\ \right)$. Furthermore, $n_{i,j}$ roughly follows a multinomial distribution with sample size $n_j^{(Q)}$ and success probability $N_{i,j}/N_j^{(Q)}$. The joint likelihood $\mathcal{L}\Big(\mathbf{N}^{(d)}, \overline{\mathbf{Y}}^{(d)}\Big)$ can then be approximated by the marginal product $\mathcal{L}\Big(\mathbf{N}^{(d)}\Big)\mathcal{L}\Big(\overline{\mathbf{Y}}^{(d)}\Big)$. The term *pseudo*-maximum likelihood signals that maximization is taken over this approximated function rather than the true likelihood. Maximizing $\mathcal{L}\Big(\overline{\mathbf{Y}}^{(d)}\Big)$ leads to a self-weighted estimate in closed form:

$$\widehat{\overline{\mathbf{Y}}}^{(d)} = \left[\sum_{j=1}^{Q} \frac{y_{i,j}}{f_j}\right]_{d\times 1} . \tag{3}$$

However, maximizing $\mathcal{L}\Big(\mathbf{N}^{(d)}\Big)$ involves solving for a linearization $\tilde{\mathbf{N}}$ of the optimal estimate $\widehat{\mathbf{N}}^{(d)}$ of $\mathbf{N}^{(d)}$ through the following iterative matrix equation:

$$\begin{pmatrix} (\mathbf{I}-\mathbf{M}\mathbf{M}^{+})(\operatorname{diag}\tilde{\mathbf{N}}_k)^{-1}(\operatorname{diag}\mathbf{M}\mathbf{f}) \\ \mathbf{M}' \end{pmatrix} \tilde{\mathbf{N}}_{k+1} = \begin{pmatrix} (\mathbf{I}-\mathbf{M}\mathbf{M}^{+})(\operatorname{diag}\tilde{\mathbf{N}}_k)^{-1}(\operatorname{diag}\widehat{\mathbf{H}}\mathbf{f}) \\ \mathbf{N}^{(Q)} \end{pmatrix} , \tag{4}$$

where k is the iteration index at most 100; $\mathbf{M} = [\delta_{i,j}]_{d\times Q}$ is the domain-in-frame indicator matrix, such that $\delta_{i,j} = 1$ or 0 according to whether domain i is in frame j or not; $\mathbf{M}^{+}$, the Moore-Penrose inverse of $\mathbf{M}$; and $\widehat{\mathbf{H}} = [f_i n_{i,j}]_{d\times Q}$. The initial estimate $\tilde{\mathbf{N}}_0$ is chosen to be $[\max_j n_{i,j}]_{d\times 1}$. If case s is in domain i, then its pseudo-maximum likelihood weight is

$$\tilde{w_s} = \frac{\tilde{N}_i}{n_i^{(d)}} \tag{5}$$

on all frames. The estimated total is then

$$\tilde{Y} = \sum_{j=1}^{Q} \sum_{s\in\text{sample } j} \tilde{w}_s y_{s,i} . \tag{6}$$

The second estimator to be evaluated is Mecatti's multiplicity estimator (Mecatti, 2005, 2007). In addition to the previous conventions, we need a variable $\mathbf{m} = [m_1, \ldots, m_N]$ to indicate the number of frames to which each case belong. The multiplicity weights are then defined by

$$w_s^{(j)} = \frac{\delta_j(s)}{f_j m_s} , \tag{7}$$

where $\delta_i(s) = 1$ or 0 according to whether case s is in frame j or not. Unlike their pseudo-maximum likelihood counterparts, the multiplicity weights for the same case are different across frames. Under our simple random sampling design, the variance of the estimated population total can be estimated by

$$\widehat{\operatorname{Var}}(\widehat{Y}) = \sum_{j=1}^{Q} \left(\frac{1}{f_j} - 1\right) \frac{1}{f_j(N_j^{(Q)} - 1)} \left[N_j^{(Q)} \sum_{s\in\text{sample}j} \frac{y_s^2}{m_s^2} - \frac{1}{f_j} \left(\sum_{s\in\text{sample}j} \frac{y_s}{m_s} \right)^2 \right] . \tag{8}$$

3. Results

We chose to simulate for $Q = 5$ frames and, among their combinations, $d = 8$ domains. The domain sizes are set in two ways to generate simulation frames Data Set 1 and Data Set 2. For reasons to be explained shortly, we draw 1000 rounds of samples from Data Set 1, but only one round from Data Set 2. In each round, a simple random sample of size between 901 to 905 is drawn from each of the 5 frames. Note that the sampling fractions are small (and all less than 1%) under this set-up.

Table 1: Simulated Results (First Round) of Data Set 1 (Converged)

Community					**True Values**		**Unrestricted PMLE**	
A	B	C	D	E	$N^{(d)}$	$\%N^{(d)}$	$\tilde{N}$	$\%\tilde{N}$
0	0	0	0	1	1,000,000	44.8	1,000,115	44.8
0	0	0	1	1	1,000,000	44.8	999,817	44.8
0	0	1	1	1	10,000	0.4	10,068	0.5
0	1	0	0	1	10,000	0.4	10,170	0.5
0	1	0	1	1	100,000	4.5	99,830	4.5
1	0	0	0	1	1,000	0.0	715	0.0
1	0	0	1	1	10,000	0.4	10,353	0.5
1	0	1	1	1	100,000	4.5	99,932	4.5

Data Set 1 (Table 1) simulated $N = 2,231,000$ cases, and converged within 6 iterations in all rounds when the tolerance at iteration k is $\sqrt{\frac{1}{d}\sum_{i=1}^{d}\left[(\tilde{N}_i)_k - (\tilde{N}_i)_{k-1}\right]^2} < 1$. The first 5 columns of the table correspond to the domain-in-frame indicator matrix $\mathbf{M}$. The estimated domain sizes are quite close to the true values. Table 3 below is an example of one of the rounds. Comparing the pseudo-maximum likelihood estimate $7,621,918$ to the true population total $7,585,350$, the relative difference is only 0.48%.

However, not every set-up can run so smoothly. Data Set 2 (Table 2) simulated $N = 4,365,000$ cases, and failed to converge within 100 iterations even when the tolerance is relaxed to values as large as d. A more serious concern is that the PML algorithm, when run without any restriction, led to a negative value in some domain estimate of $\tilde{N}^{(d)}$. This result is probably the price paid for linearizing the optimal estimate in order to open up the possibility of solving the iterative matrix equation. For simplicity, we considered only one alternative to try to remedy the above issue, that is, we imposed a lower bound on the estimates so that all elements of $\tilde{N}^{(d)*}$ are non-negative. More precisely, we reset all negative domain sizes during the iterations to a uniform random number between 2 and 10 to try to avoid being trapped at a suboptimal point. However, with all the efforts mentioned above, the algorithm still did not converge. Surprisingly, in spite of the divergence, the lower-bound-imposed pseudo-maximum likelihood estimate has value $15,402,326$ and is still within 3% of the true total $14,844,712$. This suggests that, while the domain size

Table 2: Simulated Results of Data Set 2 (Diverged whether Unrestrictd or with Lower Bound Imposed)

Community					True Values		Unrestricted PMLE		LB-Imposed PMLE	
A	B	C	D	E	$N^{(d)}$	$\%N^{(d)}$	$\tilde{N}$	$\%\tilde{N}$	$\tilde{N}^*$	$\%\tilde{N}^*$
0	0	0	0	1	2,500,000	57.3	1,600,417	36.7	1,835,115	38.3
0	0	0	1	1	1,000,000	22.9	1,252,083	28.7	984,154	20.5
0	0	1	1	1	50,000	1.1	697,500	16.0	730,731	15.2
0	1	0	0	1	50,000	1.1	449,167	10.3	50,673	1.1
0	1	0	1	1	500,000	11.5	100,833	2.3	499,327	10.4
1	0	0	0	1	5,000	0.1	505,417	11.6	669,212	14.0
1	0	0	1	1	10,000	0.2	157,083	3.6	26,519	0.6
1	0	1	1	1	250,000	5.7	$-397,500$	-9.1	6	0.0

estimates may be biased, the bias is linear in a sense that overestimation in one domain will likely be balanced out by underestimation in another domain. This feature can be a strength of this method.

Data Set 2 experienced difficulties for other choices of samples. Thus, the two estimators were only stress-tested against Data Set 1. More precisely, we drew 1000 rounds of samples, and calculated the pseudo-maximum likelihood and multiplicity estimates for each round's total sample. We then derived the relative error $(\widehat{Y} - Y)/Y$. Figure 1 shows that the two estimators have very similar distributions (left) and are highly correlated (right). The relative error of the PML estimator ranges from -0.041 to 0.048; while that of the multiplicity estimator ranges from -0.055 to 0.050. Also, he multiplicity estimator has smaller absolute relative error only 41.3% of the time. Thus, the pseudo-maximum likelihood estimator performs just slightly better in terms of being consistently close to the true value in this particular setup. But for all practical purposes, the accuracy advantage is not obvious.

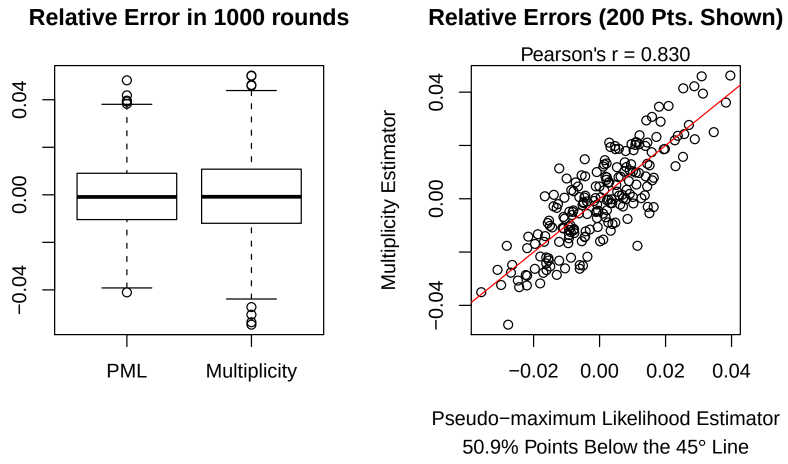

Figure 1: Comparison of Two Estimators: (left) both pseudo-maximum likelihood and multiplicity estimators have similar distributions; (right) both estimators follow the same trend.

Figure 2 shows that 94.9% of the multiplicity estimates fall within 1.96 times the standard error under a bell-shaped curve (left), and that the Shapiro-Wilk test of the critical values has p-value 0.883 (right). Both lend support to its normality and ultimately to the applicability of the asymptotic variance.

4. Discussions

Our simulation frames are set up to perfectly cover the hypothetical target population. Of course, actual frames are almost never perfect in that way. We can imagine Data Set 1 to be a frame that undercovers a hypothetical target population defined by Data Set 2. Then the deficiency in the sampling frames may provide an illusion of computational stability (or lack thereof) while the use of the full target population could have led to a different conclusion. This problem is especially severe in the presence of any small domain that spans across many frames.

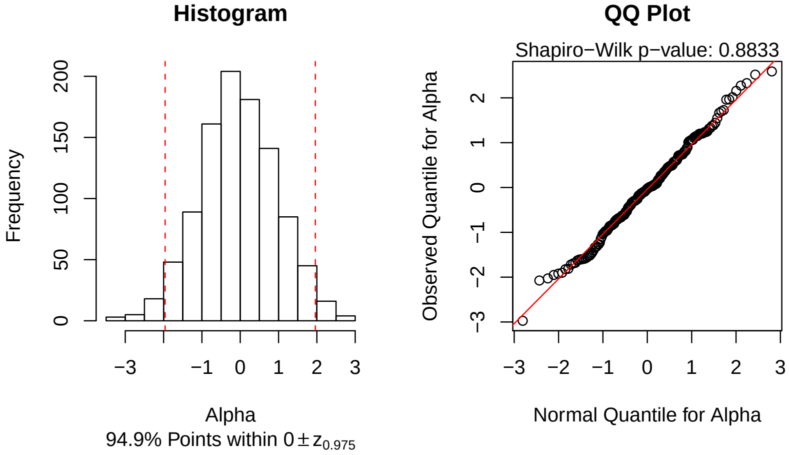

Figure 2: Multiplicity Estimator Alpha: (left) the distribution of α values for the multiplicity estimator has the normal curve shape; (right) the QQ plot also suggests close fit to the normal theory.

Lohr and Rao explored the domain collapsing technique to avoid unstable estimates from domains where no or very few sample units are found. This could potentially alleviate the divergence problem. However, there are two limitations to this approach in the context of our data sharing problem. First, different orders of frame merging would result in different sets of weights (Lohr & Rao, 2006, p. 1023). Second, their simulations were based on randomly generated frames. While this tends to cover a variety of domain combinations and sizes, real applications usually involve only a few domains– all with fairly large sizes. Take Data Set 2 for example. Although its domain sizes are artificially chosen, its domain combinations and sizes closely resemble the actual ones in the REACH U.S. Survey. It turns out the pseudo-maximum likelihood method requires more involved computational considerations as convergence is not guaranteed for unrestricted maximization. While we do not rule out the existence of more sophisticated (and, hence, more complicated) methods that may remedy the negative estimate problem, we did show that any other solutions, if they exist at all, are probably not straightforward. Although the PML method does not require knowing the true domain sizes (or even frame sizes), this is not a distinct advantage in our case as the frames are already believed to closely resemble the true population distributions.

In contrast, Mecatti's method is more straightforward. Since only the multiplicity is required, this method is less sensitive to domain misclassification. However, this advantage seems to be less relevant to ABS because 1) multiplicity in ABS usually cannot be determined without making some assumptions about the domain membership; and 2) the risk of domain misclassification is small when the ABS frames have good coverage, as in the case of REACH U.S. All in all, the multiplicity method still appears to be a good tradeoff in handling the stability problem that the pseudo-maximum likelihood method is less able to avoid.

Between the two methods, Mecatti's estimator is easier to calculate. It is also conceptually simpler, as it just averages the self-weighted domain size estimates. However, since the multiplicity weights change across frames, they can be cumbersome to report. For traditional, multiple-frame estimation problems, this is not a concern because they usually consider only two or three frames. But, if the REACH U.S. model becomes a viable option for data sharing among smaller surveys, then perhaps more than 5 surveys will be packed into the same area. In that case, management and reporting issues will need to be addressed. Aside from such potential inconvenience, the computational simplicity and stability of Mecatti's estimator make it very competitive for data sharing problems that combine many relatively small frames together.

In the simulations considered so far, we have always assumed that the five simulation frames (and, hence, the corresponding surveys) are comparable. We justified this by appealing to a question in the common module. Although the geographic overlaps appear to be large, the total overlaps required to make meaningful comparisons are small, because the communities usually have different demographic targets and scientific aims. In the REACH U.S. Survey, it turns out that only one larger community intersects two other smaller ones geographically, demographically, and teleologically. The two smaller ones are geographically disjoint. Therefore, if the larger community turns out to have run an effective intervention, then the other two smaller communities would appear to be more successful. This implies that, if we want to fully resolve the overlapping issue, then it is not enough to just combine surveys together.

5. Conclusions

Based on the limited simulation evidences that compared against only one alternative (*i.e.*, pseudo-maximum likelihood weighting), Mecatti's multiplicity weighting is computationally more efficient and more stable for data sharing. However, even when additional information appears to be available, extra care must be taken to ensure the cross-survey comparison makes sense in the first place. And even when comparison is warranted, estimators primarily designed for coverage enhancement and cost reduction may not be good fits for multi-survey projects such as REACH U.S. It is because, while those traditional multi-frame estimators are good at handling uncertain domain membership and unknown frame sizes, they tend to anticipate large overlaps among frames, a condition often unfulfilled or even unwanted in our data sharing problems. Further research is needed to explore the right type of estimator to meet the new demands.

REFERENCES

Fuller, W. A., and Burmeister, L. F. (1972), "Estimators for Samples Selected From Two Overlapping Frames," in *Proceedings of the Social Statistics Section*, American Statistical Association, 245–249.

Hartley, H. O. (1962), "Multiple Frame Surveys," in *Proceedings of the Social Statistics Section*, American Statistical Association, 203–206.

——— (1974), "Multiple Frame Methodology and Selected Applications," *Sankhyā: the Indian Journal of Statistics*, Series C, **36**, 99–118.

Lohr, S., and Rao, J. N. K. (2006), "Estimation in Multiple-Frame Surveys," *Journal of the American Statistical Association* **101**:475, 1019–1030.

Mecatti, F. (2005), "Single Frame Estimation in Multiple Frame Survey," *Proceedings of Statistics Canada Symposium 2005*: Methodological Challenges for Future Information Needs, Ottawa, Ontario, October 27, 2005.

——— (2007), "A Single Frame Multiplicity Estimation for Multiple Frame Survey," *Survey Methodology* **33**:2, 151–158.

Skinner, C. J., and Rao, J. N. K. (1996), "Estimation in Dual-Frame Surveys," *Journal of the American Statistical Association* **91**:433, 349–356.