# Minimizing Duplication of Samples Drawn from Overlapping Frames of Multiple Surveys

**Peter K. Kwok**[1], **Hee-Choon Shin**[2], **Colm O'Muircheartaigh**[1], **Whitney Murphy**[1], **Angela Debello**[3], **Kari Carris**[3]

National Opinion Research Center at the University of Chicago
[1] 55 E. Monroe St., Ste. 2000, Chicago, IL 60603 kwok-peter@norc.org, colm@norc.uchicago.edu, murphy-whitney@norc.org
[2] 1155 E. 60th St., Chicago, IL 60637 shinh@uchicago.edu
[3] 55 E. Monroe St., Ste. 3000, Chicago, IL 60603 debello-angela@norc.org, carris-kari@norc.org

**Abstract**
REACH U.S. (Racial and Ethnic Approaches to Community Health Across the United States) is a CDC's community-based initiative to eliminate health disparities among various racial and ethnic groups. Five of the 28 participating communities are located within Los Angeles and Orange Counties, California with complicated overlapping among their geographies, and have different scientific interests and eligibility requirements. Although all respondents are asked the same questions, those communities did not intend to share any completed interview at the time of this writing. Therefore, this is a multiple-frame, multiple-survey problem that requires samples to be independently drawn from overlapping areas. We will discuss how the address-based sampling design can meet this demand and what challenges it faces. In particular, we present an estimation algorithm that aims to minimize the impact of sample de-duplication on the independence assumption.

**Keywords.** Address-based sampling, REACH U.S., sample overlap problem, independent sampling, sample de-duplication

## Introduction and Problem Statement

This paper will discuss a new type of optimization problem, called the *Sample De-duplication Problem*, with the objective of searching for an optimal order of drawing samples from overlapping frames in order to address a certain sampling issue that arises from multiple surveys. It is a variation of the traditional sample overlap problem, but is restricted to situations when sample overlap is a result of random sampling and is open to minimization or even elimination. This excludes the cases that are unmeaningful without the sample overlap, such as longitudinal surveys, or the cases due to nonsampling errors, such as record discrepancies or identity frauds. We may not be able to do much if we do not know enough about the frame membership, or if the frames are so small or the domain structure so complex that any benefit of optimization is marginalized. But suppose we get above all these. Then there are situations when it may be beneficial, or even mandatory, to de-duplicate overlapping samples. For instance, there may be a confidentiality agreement to prevent overlapping surveys from sharing data or even coordinating operations. Even if we are allowed to keep a case alive as long as possible until we exclusively assign it to an eligible survey, the combined screener may become too long or too repulsive (if it includes sensitive questions from some survey), and may risk losing respondents. Or the order of screener questions may induce unacceptable bias. Last but not least, it is always easier to plan operations and compute base weights with de-duplicated samples if we can afford it.

Of course, we are *not* saying that every sample overlap should be de-duplicated. After all, it is all about the trade-off. If the samples are de-duplicated before screening, then a case that is ineligible for its assigned sample may not qualify for others, even though it might have been eligible otherwise. When a case is eligible for multiple surveys, the sample drawn first will take it away from others. This may cause a shortage problem for the lower-density samples when the sampling rate differential is too big and the population is too rare. From an analytical point of view, sample overlap can also greatly improve estimation given the advances in dual-frame and multiple-frame approaches. So there are quite a few costs and benefits to weigh in before making a decision. But if you choose to do that, then you are facing a new problem with new challenges. We will use the survey of *Racial and Ethnic Approaches to Community Health Across the United States* (REACH U.S.) to illustrate our point.

REACH U.S. is a community-based participatory research project funded and organized by the Centers for Diseases Control and Prevention (CDC). Through this project, currently 40 institutions are funded to develop and implement different promotion and research programs to eliminate health disparities among various racial and ethnic groups. Among the grantees, 18 are Centers of Excellence in the Elimination of Health Disparities (CEED), which disseminate data and coordinate public health research efforts; and 22 are Action Communities (AC), which implement and monitor the health programs. Each year within a five-year period, some of them conduct surveys to monitor health behavior in their communities. In the current REACH U.S. Survey, 28 grantees have participated. The questionnaire is consisted of a common module of questions on general health (mostly about diet and exercise), followed by specialized modules on cardiovascular disease (CVD), *diabetes mellitus* (DM), adult/older adult immunization (AI), breast and cervical cancer (BCC), and *hepatitis* B and C, in addition to other screening questions and information. The set of questions (and their options) are almost identical for all respondents with only minor modifications in a handful of exceptional communities, including the UCLA grantee to be discussed below. Each grantee is primarily interested in the responses to only a few modules. But for practical purposes, all modules are asked for all respondents regardless of the scientific objective of the community. On this consideration alone, a universal screener is possible because the surveys are under the same project and ask the same questions. However, the complexity of demographic eligibility poses a problem.

The most complicated locality is consisted of five communities nested within Los Angeles and Orange Counties as summarized in Table 1 and depicted in Figure 1 below. They target different demographic groups among Hispanics/Latinos (HL), African Americans (AA), and various subgroups (*e.g.*, Vietnamese and native Hawaiians) among Asians and Pacific Islanders (A/PI). Each community aims to have 900 completed interviews overall. In addition, certain communities oversample females aged 40 through 64 (F40) or elderly aged 65 and above (65+) to aim for at least 300 completed interviews. All eligible adults are selected from the oversampled group; but only up two are programmatically selected from the non-oversampled group. The geographies of these 5 communities are defined by ZIP code areas within the two counties. All communities above are geographic subsets of the largest one (OCAPI). One community (Biomed) is a subset of another (UCLA) geographically and demographically. But none is a subset of another in terms of geography, demographic group, and scientific objectives all together. To increase accuracy of the screening process, the REACH U.S. Survey rosters all household members to determine demographic eligibility. The risk of losing respondents due to excessively long screening and too many sensitive questions about age, gender, and race/ethnicity became a major concern. Therefore, the REACH U.S. Survey tried the sample de-duplication approach for at least some of its sample releases.

| *Grantee* | *Type* | *Geography in L.A./Orange* | *Demographics (Oversampled)* | *Modules* |
|---|---|---|---|---|
| Community Health Councils, Inc. (*CHC*) | AC | 18 ZIPs | AA | CVD, DM |
| Special Services for Groups (*SSG*) | AC | 32 ZIPs | A/PI (F40) | CVD, DM, AI |
| Los Angeles Biomedical Research Institute at Harbor UCLA Medical Center (*Biomed*) | AC | 20 ZIPs | HL, AA (65+) | AI |
| Regents of the University of California-Los Angeles (*UCLA*) | CEED | 128 ZIPs | AA, HL, A/PI | CVD |
| Orange County Asian and Pacific Islander Community Alliance (*OCAPI*) | CEED | All | A/PI (F40) | BCC |

Table 1: REACH U.S. Communities within Los Angeles and Orange Counties, California

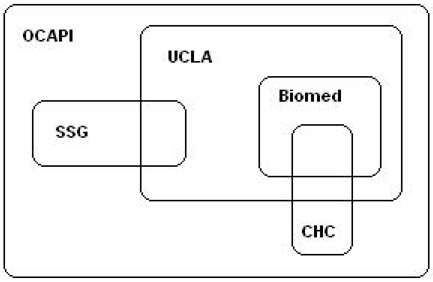

Figure 1: Geographic Relationship of the Five Los Angeles Communities

With such a broad extent of complication, it is a challenge to even just draw the samples. At least it would require a lot of screening for a random-digit dialing (RDD) design in order to maintain a high level of confidence about respondents' eligibilities. Fortunately, the REACH U.S. Survey adopts an address-based sampling (ABS) design by constructing an address frame for each community before matching a primary telephone number to each address. When applicable, the ABS frames are highlighted by list frames that come together with addresses, telephone numbers, and demographic indicators. In such case, the ABS frames are augmented by the list addresses similar to the traditional dual-frame set-up. With or without predictive information from the list frames, the ABS frames across communities can be partitioned by addresses alone. As English *et al.* (2009) have shown, address frames used in the REACH U.S. Survey have good coverage. Thus, we will rely on domain membership determined from the ABS frames to plan for the sampling. In general, ABS gives greater control over the sampling rates, and is better for overlapping surveys of this nature.

Since the five L.A. communities are actually multiple surveys using multiple frames, it seems natural to try to draw the samples independently. To illustrate our main point, we will ignore complex design issues in the actual implementation of the REACH U.S. Survey, and consider only simple random samples without replacement throughout this paper. Suppose we draw a sample $s_A$ of size $n_A$ from a frame $A$ of size $N_A$, and then draw a sample $s_B$ of positive size $n_B$ from a frame $B$ of size $N_B$. We may assume without loss of generality that all sizes under consideration, including that of the frame intersection $ab = A \cap B$, are positive integers. If $s_A$ and $s_B$ are drawn independently, then, for each unit $x \in A \cap B$, we necessarily have

$$P_{\text{ideal}}(x \in s_A \cap s_B) \;=\; P_{\text{ideal}}(x \in s_A) \cdot P_{\text{ideal}}(x \in s_B) \;=\; \frac{n_A}{N_A} \cdot \frac{n_B}{N_B} > 0 \,. \tag{1}$$

However, the five grantees in the Greater L.A. area did not plan to share completed interviews. Since each community counts its own quota, drawing the same unit into multiple samples can benefit at most one community. For that reason, we consider such repeated selection inefficient and undesirable, and will call it a *collision*. To avoid collisions, we have to de-duplicate the samples. It seems convenient to draw samples successively and remove all previously drawn samples from the current one at each step. Under this set-up, $P(x \in s_B \mid x \in s_A) = 0$ by design. Hence, we have instead

$$P_{\text{actual}}(x \in s_A \cap s_B) \;=\; P_{\text{actual}}(x \in s_A) \cdot P_{\text{actual}}(x \in s_B \mid x \in s_A) \;=\; 0 \,. \tag{2}$$

This means that the data separation requirement invalidates the independence assumption. For the rest of this paper, we will describe and assess an algorithm which is not intended to "save" the independence assumption from the contradicting requirement, but rather to "protect" it as much as reasons allow in the hope that the end results will be as close to genuine independent sampling as practical.

### Methodology and Results

Let's start over from the flip side of (1). Re-use the same notations as before. If we are in the ideal situation of independent sampling, then we necessarily have

$$P_{\text{ideal}}(x \notin s_A \cup s_B) \;=\; P_{\text{ideal}}(x \notin s_A) \cdot P_{\text{ideal}}(x \notin s_B) \;=\; \frac{N_A - n_A}{N_A} \cdot \frac{N_B - n_B}{N_B} \,. \tag{3}$$

Suppose $s_A$ is drawn before $s_B$. Denote the size of $s_{ab}^{(A)} = s_A \cap B$ (possibly empty) as

$$n_{ab}^{(A)} = \left| s_A \cap B \right| \,. \tag{4}$$

Here $s_{ab}^{(A)}$ can be interpreted as the *collision of $s_A$ on $B$*, namely the portion de-duplicated from $B$ after $s_A$ is drawn but before $s_B$ is drawn. Its size $n_{ab}^{(A)}$ is interpreted as *collision as an effect on the frame size*. Then, this time we have

$$P_{\text{actual}}(x \notin s_A \cup s_B) \;=\; P_{\text{actual}}(x \notin s_A) \cdot P_{\text{actual}}(x \notin s_B \mid x \notin s_A) \;=\; \frac{N_A - n_A}{N_A} \cdot \frac{N_B - n_{ab}^{(A)} - n_B}{N_B - n_{ab}^{(A)}} \,. \tag{5}$$

Here $n_{ab}^{(A)}$ is a random variable as it potentially changes each time a new $s_A$ is drawn. From this viewpoint, the problem now becomes changing $n_{ab}^{(A)}$ (or, rather, any estimate of its expected value) to minimize the gap between the right-hand sides of (3) and (5). Heuristically, for the case of only two overlapping frames,

$$\frac{N_B - n_B}{N_B} \approx \frac{N_B - n_{ab}^{(A)} - n_B}{N_B - n_{ab}^{(A)}} \iff \frac{n_B}{N_B} \approx \frac{n_B}{N_B - n_{ab}^{(A)}} \iff w_{\text{ideal}}(B) = \frac{N_B}{n_B} \approx \frac{N_B}{n_B} - \frac{n_{ab}^{(A)}}{n_B} = w_{\text{actual}}(B) \iff \frac{n_{ab}^{(A)}}{n_B} \approx 0 \,. \tag{6}$$

The third approximation in (6) implies that it suffices to minimize the deviation of sampling weights in the actual design from the ideal scenario. Such deviation is interpreted as *collision as an effect on the weight*. From here on, we will be casual with the use of language, and no longer distinguish collision as a set or as an effect on frame size or weight. The fourth and last approximation in (6) tells us to minimize the *total collisions* (or its estimated expectation).

In the REACH U.S. Survey under consideration, we have $k = 5$ overlapping frames. Without loss of generality, we consider a specific ordering of frame indices from $A$ to $E$, denoted by $A \succ B \succ C \succ D \succ E$ (read "A, then B,..., then E"). Re-partition the 5 frames into mutually exclusive and exhaustive domains. A domain is a maximal intersection of frames with identical membership combination, and is represented as an ordered subset of the frame index set $\{a,b,c,d,e\}$. Group the domains into $\mathscr{D}_A$ through $\mathscr{D}_E$ such that each element in $\mathscr{D}_A$ begins with the letter $a$ to indicate the first drawn frame $A$, so on and forth. Then we can define the collisions similar to (4) to obtain a new objective function

$$\begin{aligned} f(A \succ \cdots \succ E) = {} & \frac{n_{ab}^{(A)}}{n_B} + \frac{n_{ac}^{(A)} + n_{abc}^{(A)} + n_{abc}^{(B)} + n_{bc}^{(B)}}{n_C} + \\ & \frac{n_{ad}^{(A)} + n_{abd}^{(A)} + n_{acd}^{(A)} + n_{abcd}^{(A)} + n_{bd}^{(B)} + n_{abd}^{(B)} + n_{bcd}^{(B)} + n_{abcd}^{(B)} + n_{cd}^{(C)} + n_{acd}^{(C)} + n_{bcd}^{(C)} + n_{abcd}^{(C)}}{n_D} + \frac{n_{ae}^{(A)} + \cdots}{n_E} \end{aligned} \tag{7}$$

subject to

$$\begin{aligned} & n_B \geq n_{ab}^{(A)} \\ & n_C \geq n_{ac}^{(A)} + n_{abc}^{(A)} + n_{abc}^{(B)} + n_{bc}^{(B)} \\ & \quad \cdots \end{aligned}$$

In the general case of $k$ frames $F_1, \ldots, F_k$, (7) can be rewritten as

$$f(A \succ \cdots \succ E) = \underbrace{\sum_{\alpha=F_1}^{F_k}}_{\text{frames}} \underbrace{\sum_{\delta \in \mathscr{D}_\alpha}}_{\text{domains}} \underbrace{\sum_{\alpha \neq \beta \in \delta} \frac{n_\delta^{(\alpha)}}{n_\beta}}_{\text{collisions}} \tag{8}$$

subject to

$$n_\beta \geq \sum_{\alpha=A}^{D} \sum_{\beta \in \delta \in \mathscr{D}_\alpha} n_\delta^{(\alpha)} \quad \text{for all } \beta = B, \ldots, E \ .$$

It suffices to just estimate $n_\delta^{(\alpha)}$ by the expected proportion of $s_A$ within $\delta$, that is,

$$\widehat{n}_\delta^{(\alpha)} \ = \ \widehat{E}\left(n_\delta^{(\alpha)}\right) \ = \ n_\alpha \times \frac{N_\delta}{N_\alpha} \ . \tag{9}$$

The above expression is a little complicated. It is often more convenient to maximize the weights instead of minimizing the collisions. The new, alternative objective function becomes

$$W = w_1 + \cdots + w_k = \frac{N_{1*}}{n_1} + \cdots + \frac{N_{k*}}{n_k} \ , \tag{10}$$

where $N_{i*}$ is the frame size actually available for drawing sample $s_i$ for $i = 1, \ldots, k$ subject to $N_{i*} \leq n_i$. Since the number of frames under consideration is only $k = 5$, we can afford the time and resources to inspect each of the $k! = 5! = 120$ orderings of sample drawings to select the one that minimizes the objective function $f$.

Finally, we note without showing computational details that the optimal result is selected (with arbitrary tie breaker) to be

$$(A \succ B \succ C \succ D \succ E) \ = \ (\text{UCLA} \succ \text{OCAPI} \succ \text{SSG} \succ \text{Biomed} \succ \text{CHC}) \ . \tag{11}$$

Note that the samples with lower sampling rates tend to go to the front of the ordering because their collisions on subsequent frames tend to have a lesser impact, while the ones with higher rates tend to go to the back. Also, the SSG frame turns out to be geographically disjoint from the Biomed and CHC frames. Since disjoint frames have zero collision on one another, we can swap the order of last two groups in (11) to obtain another minimum ordering, which is also closest to being independent. Reversing either minimum ordering maximizes the expected total collisions, and would deviate the most from being independent.

## Discussions and Conclusions

We have considered a new type of problem called the Sample De-duplication Problem, and have presented a heuristic algorithm that optimizes a sample drawing order to be as close to independent sampling as reasonable. However, our proposal has three obvious shortcomings that need to be acknowledged and addressed.

First, while the above algorithm can be easily generalized to any number of frames, the practice of searching through all $k!$ possible frame orderings is computationally feasible for only small values, say, $k \leq 10$. It is unproductive or even infeasible for $k \geq 15$ at our current state of technology. For the purpose of REACH U.S., $k = 5$ is about the maximum number of localities it is going to have within a close vicinity. While the algorithm is suitable for our purposes, we acknowledge that it is not suitable for applications with a lot of frames (*e.g.*, longitudinal frames with frequent edits and updates). However, we argue that every algorithm has its limitation, and that the best we can do is to keep a small problem to an easy solution that it is worth.

Second, the algorithm above does not guarantee a sufficient sample size after augmenting from previously drawn samples. However, such potential issue can be resolved by drawing, say, 2 to 3 times the originally planned sample size on each frame. In our case, we still had enough left after sample de-duplication. If the domain sizes are not large enough to accommodate the sample inflation, then we can either reduce the inflation factor, or divide the inflated samples into smaller replicates and then release just the right amount to match the originally planned sample sizes. This method may not be feasible if some domains are smaller than the projected sample size after inflation and augmentation. In that case, some smaller domains may need to merge before re-running through the algorithm.

Third and lastly, the above algorithm is applicable only when the samples are all available so that they can be drawn in any of the $k!$ orders. This is a mild assumption when the overlapping frames are all released at once, usually before the start of interviewing operations. However, if the overlapping frames have to be released in part or in the orders dictated by any release schedule such that some sample has to go in a particular order, then this algorithm cannot be used as is.

Traditional sample overlap problems, as Ernst (1999) has comprehensively surveyed, exclusively deal with frequency count optimization. The motivation is usually related to alleviation of respondent burden (for minimizing overlaps) or interviewing cost saving (for maximizing overlaps). Their solutions are also sensitive to computational limitations and sample size uncertainty just like our algorithm. However, in this paper we have considered the problem from a different angle. The object of optimization now becomes the sum of weights (or collisions) under the effect of successive sampling and augmentation. As address-based sampling is gaining popularity, and as surveys tend to cover more but smaller areas with complicated overlapping, this type of sampling issues may arise more frequently in the future. Hopefully this can serve as a first step to raise people's awareness and ultimately lead to better solutions.

## References

Ned English, Colm O'Muircheartaigh, and Stephanie Eckman (2009). "Coverage Rates and Coverage Bias in Housing Unit Frames," presented at the *Joint Statistical Meeting (Section on Survey Methods and Research)*, Washington, D.C., August 5, 2009. (Proceedings forthcoming in 2010.)

Lawrence R. Ernst (1999). "The Maximization and Minimization of Sample Overlap Problems: A Half Century of Results," *Bulletin of the International Statistical Institute, Proceedings Tome LVII*, Book 2, 293–296.