Exploring Sampling Techniques to Reduce Respondent Burden

Yijun Wei[1], Valbona Bejleri[2]
[1]Yijun Wei, 1200-1398 Independence Ave SW, Washington, DC 20024
[2]Valbona Bejleri, 1200-1398 Independence Ave SW, Washington, DC 20024

The USDA's National Agricultural Statistics Service (NASS) is exploring sampling approaches that incorporate coordination of multiple samples drawn within a year across the population in an effort to control respondent burden. Most of these sampling techniques, including both design-based and model-based approaches, utilize permanent random numbers (PRN) for the purpose of limiting the amount of overlap within a survey or between different surveys to help reduce respondent burden. However, there is little published discussion of the comparative effectiveness of PRNs when used with design-based and model-based approaches or combinations of these two at NASS. A simulation study investigates different sampling strategies (sampling design and estimator) for limiting overlap that utilize design-based and/or model-based inferences. Simulations are based on data from several USDA surveys.

Keywords: Respondent burden, Sampling strategies, Survey coordination

1. Introduction

Response rates among surveys conducted by government agencies have been declining. Reasons include people refusing to respond to surveys (refusals) and failure to make contact with potential respondents. In an ongoing effort to identify ways of improving response rates, the National Agricultural Statistical Services (NASS) of the USDA set up a Response Rate Research team (RRRT) consisting of several subteams. A subteam of the RRRT reviewed sampling techniques currently in use by NASS and by other federal statistical agencies and also surveyed the literature on coordination of sample surveys to reduce sample overlap (unintended repeat selection of the same unit across multiple surveys).

Respondent burden has long posed problems for sample surveys, from an early citation (Chapin, 1920) to a contemporary publication (Fricker *et al.*, 2011). The burden on the respondent can influence how or whether an individual responds to a survey and can impact the quality of the final statistical outputs (Jacqui, 2012). While the definition of respondent burden is multidimensional, the aspect studied here is repeated selection and respondent burden is defined here as the number of times an individual unit is selected to participate in surveys. The focus of this study is on methods for coordination of surveys to control the overlap among samples.

A variety of sampling techniques to handle respondent burden have been developed for use with design-based and with model-based sampling approaches. The most common techniques include those based on permanent random numbers (PRNs), methods using collocated PRNs (CRNs) and those using a coordination function.

PRNs were first introduced by Fan, Muller and Rezucha (1962). PRNs are widely used with different sampling approaches. Jales technique was applied to the Statistics Sweden business registry (Ohlsson, 1992). Bailey and Kott (1997) used PRNs with Multivariate Probability Proportional to Size (MPPS). At the USDA, the Economic Research Service (2016) uses PRNs in conjunction with Sequential Interval Poisson Sampling (SIP).

For simultaneous selection of samples from multiple surveys, Ernst and Casady (2000), and Butani et al. (2000) used (evenly spaced) CRNs to minimize the overlap. A coordination function (Guggemos, Fabien and Sautory, 2012) uses a pre-defined function to transform PRNs without drastically altering the original design to meet a conditional criterion for coordination.

Two simulation studies were designed to evaluate the performance of a coordination function. The first compares its use with different sample designs: Simple random sampling (SRS), PPS and a combination of SRS with PPS. The second study evaluates its contribution when combined with SIP.

This paper is structured as follows: Section 2 discusses PRNs and CRNs as these are used at NASS; Section 3 describes the sampling techniques in two major NASS surveys; and Section 4 gives a detailed discussion of the coordinating function studied here. Section 5 describes the simulation studies and presents results evaluating the performance of the coordination function; summary and discussion appear in Section 6.

2. Technique to reduce respondent burden

2.1 Respondent burden

A large number of surveys are conducted by organizations and government agencies every year. Some businesses or individuals may only receive one survey in a year, while others receive numerous survey requests. From the perspective of the businesses or individuals, these survey requests are burdensome because they incur costs and take up time with no benefit received in return. From the survey organizations' perspective, adequate survey responses are important to the quality of the final statistical outputs. A trade-off therefore exists between the survey demands and the need for quality statistics (Jacqui, 2012). Fricker et al. (2011) concludes that the lower the respondent burden the higher the quality of statistics.

Researchers and practitioners often rely on loose definitions of burden, or continue to employ the interview length as a proxy measure of burden (Sharp et al. 1983). Bradburn's (1978) concept of burden is multidimensional and reflects the influences of interview length, efforts required of respondents, the frequency of interviews, and the amount of stress on respondents. Bradburn suggested several possible factors that could influence respondents' perceptions of the survey task (e.g., interest in survey topic), but only a handful of studies have tried to assess respondents' attitudes and subjective reactions and then examine their impact on burden (e.g., Sharp and Frankel, 1983; Hedlin et al., 2005; Fricker et al., 2011). All but the first of these factors are related to the number of times an individual is selected to participate in a survey. This paper measures this aspect of respondent burden as the number of times an individual is selected to participate in a survey. The focus here is on methods for coordination of surveys and on limiting sample overlap while maintaining statistical properties of the survey design.

2.2 PRNs and CRNs

2.2.1 Equal probability sampling

A simple random sample without replacement of size n from a population of size N can be drawn by associating a random number, uniformly distributed over the interval (0, 1) to each unit in the population. After ordering the random numbers, the first n are selected for the sample.

With PRNs, each unit is associated with the same random number permanently over all surveys. New businesses (births) are assigned new random numbers and closed-down (deaths) are withdrawn from the register. This is especially efficient when multiple samples are considered simultaneously, and when newly rotated-in units and obsolete records are considered (Ohlsson (1992) and Cox at el (1995).

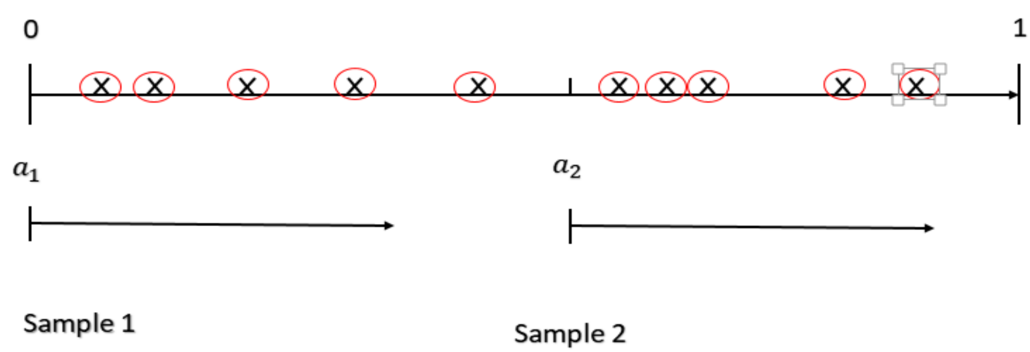

Figure 1: Sample coordination

Figure 1 (adapted from Ohlsson (1992)) shows an illustration of positive and negative coordination. A fixed starting point is chosen for each of two surveys (a_1, a_2). A unit is selected for sampling if its PRN falls into the range of the first *n* smallest PRNs starting from a starting point (a_j). Coordination is negative when the starting points $\{a_j\}$ are distant from each other and positive when they are close together.

It can be difficult to make judicious choices for a_1 and subsequent starting points, especially for small populations, in the presence of large gaps between PRNs or in cases where the PRNs clump. CRNs address this difficulty by transforming the assigned PRNs to equally spaced points. To accomplish this, each PRN is connected to a percentile rank which is left shifted by *ε/sample size* drawn from U(0, 1). Samples are selected from these CRNs as described above for PRNs.

This approach is especially efficient for relatively small populations. Ernst and Casady (2000) show in a simulation that for a sample size greater than 20, the performance of PRNs and of CRNs is similar.

2.2.2 Unequal probability sampling

Poisson sampling is a PPS sampling scheme that is usually carried out in the context of PRNs. To select a sample of size *n*, where unit *i* is included with probability proportional to the size p_i, $\sum_{i=1}^{N} p_i = 1$, assign unit *i* a random numberPRN_i in (0,1) for all units *i*=1, . . ., *N*. Determine a starting point a_j for the *j-th* survey. To select the sample for survey j, for each unit *i* apply a selection rule: select unit *i* for survey *j* if the interval $(a_j, a_j + n * p_i)$ contains $\mathrm{PRN_i}$ otherwise not. This method, also known as Bernoulli sampling (Sarndal et al., 1992), yields a fixed sampling fraction but not a fixed sample size. Poisson sampling gives better positive and negative coordination than the equal probability approach and provides simplicity in variance estimation and in the rotation of sample units (Ohlsson 1992).

3. Sampling Techniques used at National Agricultural Statistics Service

At USDA, NASS is exploring sampling approaches that allow for coordination of samples drawn for multiple surveys within a year. In designing surveys, NASS is concerned with reducing burden but also with deriving efficient estimators, as well as feasibility of implementation. NASS currently uses three main sampling procedures: Multivariate Probability Proportional to Size (MPPS), Sequential Interval Poisson (SIP) sampling, and stratified simple random sampling.

3.1 MPPS

NASS conducts the Crop APS quarterly in March, June, September, and December in order to set crop estimates at the State and National level. Two supplemental surveys are conducted separately from the Crop APS, i.e. Agricultural Yield Row Crops (AYR) and Agricultural Yield Small Grains (AYS). Therefore, the samples are selected sequentially using Poisson sampling: AYR sample first, then the AYS sample is selected from the small-grain population excluding any units selected by AYR sample. Finally the Crop APS sample is then selected excluding all units selected by AYR sample or AYS sample (Bailey at el, 1997).

Since the NASS Crop APS gathers data on many different crops, sample selection utilizes MPPS, with multiple auxiliary variables to define the selection probabilities for these surveys for sample allocation. The inclusion probability is defined as the minimum of 1 and the maximum of selection probabilities for each of M crops (commodities of interest), i.e., $\pi_i = min\{1, max\{p_i^{(m)}, m = 1 \ldots M\}\}$, where $p_i^{(m)}$ is the crop (commodity) m selection probability for unit i. Then, Poisson sampling is used to select the sample.

3.2 SIP

The Agricultural Resource Management Survey (ARMS) is the main source of information on production practices, financial condition and resource use by America's farm business and also on the economic well-being of America's farm households (*ARMS Farm Financial and Crop Production Practices, n.d.*). ARMS seeks to limit response burden both within the current year (with the Crop APS) and relative to the preceding year's ARMS.

The sample design for ARMS uses Sequential Interval Poisson sampling (SIP). In a Poisson or Bernoulli sample design, for each unit i, PRN_i is drawn from U(0, 1) and a selection probability p_i is assigned (unequal probabilities in the case of ARMS). Then each unit in the population is subjected to a Bernoulli trial with the probability of inclusion for unit i proportional to p_i. The mechanism to control the overlap between the ARMS and related surveys is by either deletion of units already sampled or by alteration (decreasing) p_i.

4. Coordination function

The coordination function introduced first in Guggemos and Sautory (2012) is a measurable function g of the PRNs that preserves the uniform, U(0,1), probabilities for assigning of PRNs but at the same time, decreases the probability of repeated selection for different surveys.

In PPS sampling, the coordination function is defined as

$$g_{i,j}(w_i) = \begin{cases} w_i + a_j & if \ \ w_i + a_j \le 1 \\ w_i + a_j - 1 & if \ w_i + a_j > 1 \end{cases},$$

where i indexes unit, j indexes sample, w_i is the random number drawn for unit i and a_j is a constant for the j-th sample. Cumulative burden function $\Gamma_{i,j}(w_i) = \sum_{u \le j} \gamma_{i,u} I(w_i)$. $\gamma_{i,u}$ is usually set as 1 (and is omitted hereafter); u is an integer between 0 and j.

In stratified SRS, before defining coordination function $g_{i,j}(w_i)$, $b_{i,j}(x)$ is defined as

$$b_{i,j}(x) = 1 - \int_{u=0}^{x} \frac{1}{B(n,q)} x^{n-1}(1-x)^{q-1} du$$

where $q = N - n$, $B(n,q) = (n-1)!\,(q-1)!/(N-1)!$, and N is the population size.

To calculate $g_{i,j}(w)$, [0, 1] is divided into L equal intervals (L is usually a large enough integer often greater than 50). A piecewise function $\widetilde{b_{i,J}}$ defined on these intervals takes the value $b_{i,j}$ at

the endpoints of the interval. $g_{i,j}(w) = \widetilde{b_{i,j}}(w_l) + (w - w_l)$ where w_l represents the closest interval endpoint for w's interval.
Cumulative burden function $\Gamma_{i,j}(w) = \sum_{u \le j} \gamma_{i,u} b_{i,j}(x)$. $\gamma_{i,u}$ is usually set as 1 and u is an integer between 0 and j.

Although the coordination functions in PPS sampling and SRS sampling are different, the sample selection procedure follows the same steps:
Select sample S_1:

- Set $\delta_{i,0}(w) = 0,\ \forall\, i,\ \forall w\ \in [0,1] \rightarrow\ g_{i,1}(w) = w, \forall\, i,\ \forall w\ \in [0,1]$
- $I_{i,1}(w) = I_{[0,\pi_{i,1}]}(w)$
- Cumulative burden function: $\Gamma_{i,j}(w) = \sum_{u \le j} \gamma_{i,u} I_{i,u}(w)$

Select sample S_n:

- Use cumulative burden function $\Gamma_{i,n-1}(w)$ as a criteria C to build a coordination function $g_{i,n}$ for selection of S_n
- $K \in S_n \leftrightarrow g_{i,n}(w_i) \in [0, \pi_{i,n}]$

The coordination function is evaluated using simulated data as well as data from NASS agricultural surveys. In next section, the coordination function is evaluated using simulated data as well as data from the NASS surveys and results from two simulation studies are discussed.

5. Two Simulations Experiments and Results

Two simulation studies illustrate how the coordination function introduced by Guggemos and Sautory (2012) works. The first study compares the contribution of the coordination function with results for designs without it. The second study evaluates performance of the coordination function with SIP used alone.

5.1 Simulation of Performance based on Survey Design

Four simulations make up the first study, with a population of 100 units and 10 samples ("surveys") of 25 units each. Units' sizes were assigned independent random numbers from U(0,1).Then four scenarios were simulated each with a thousand runs: 1) 10 samples selected using SRS, 2) 10 samples selected with PPS, 3) 5 samples selected using SRS plus 5 samples selected using PPS sampling, and 4) 10 samples drawn using the coordination function in Section 4. CRNs were used to avoid clumping that PRNs might have produced.

5.1 First simulation study

Number of appearance	Coordination function(%)	SRS(%)	PPS (%)	SRS and PPS(%)
1	10.83	20.064	20.779	21.540
2	22.574	29.705	21.544	27.974
3	68.892	26.419	21.275	24.593
4	0.6671	15.615	17.081	15.458
5	0.0740	6.098	11.396	7.222
6	0.0039	1.713	5.375	2.483
7	0.0001	0.330	1.939	0.627
8	0	0.005	0.529	0.092
9	0	0.001	0.075	0.011
10	0	0	0.007	0

Table 1. Results of simulating 10 samples of size 25 from a population of size 100

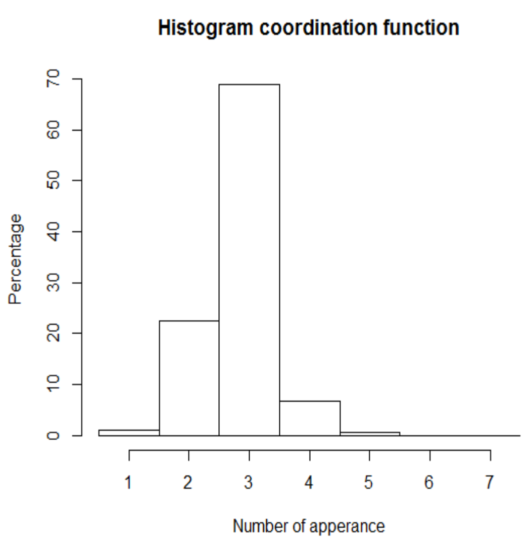

Figure 2: Histogram of number of appearances over all samples when coordination is used

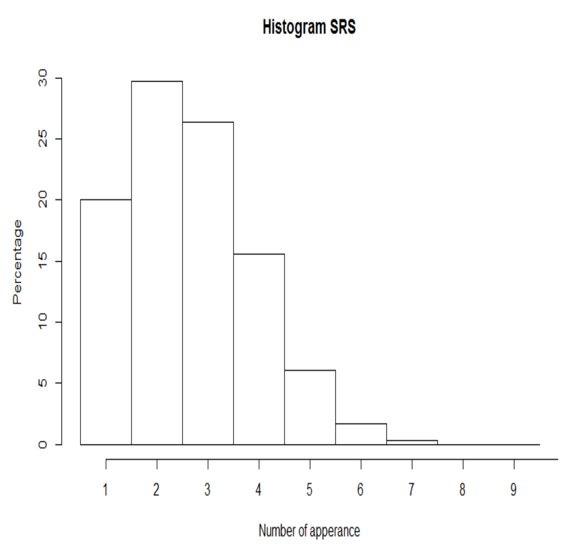

Figure 3: Histogram of number of appearances over all samples when SRS is used

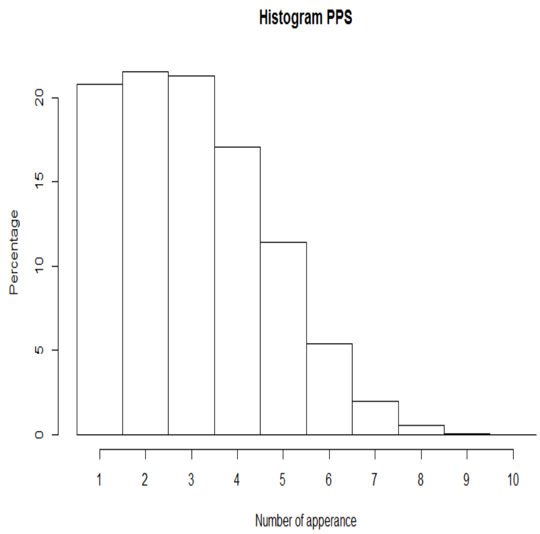

Figure 4: Histogram of number of appearances over all samples when PPS is used

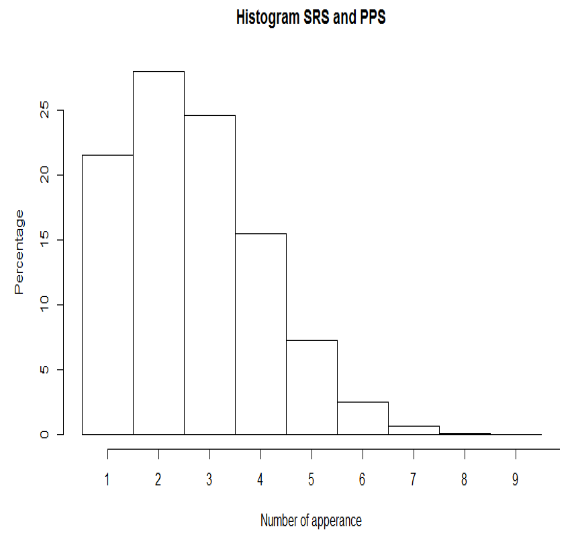

Figure 5: Histogram of number of appearances over all samples when PPS and SRS is used

In Table 1, “Number of appearance” indicates the average number of samples that a unit appears among 10 samples over 1000 runs, “coordination function” indicates using the coordination function, “SRS” implies using SRS, “PPS” implies using PPS, and “SRS and PPS” implies 5 samples selected by SRS and 5 samples selected by PPS. The percentage of number of appearance in Table 1 are visualized in figures 2, 3, 4, and 5 for coordination function, SRS, PPS, and a combination of SRS and PPS separately.

Since 10 samples are selected at 25% sampling rate in each simulation and each simulation is repeated 1000 times, each unit is expected to be selected 2.5 times in average. Table 1 shows that a unit has been selected for 2.6, 2.6, 3, and 2.6 in average for four sampling techniques. However, Table 1 shows that, when the coordination function is used most of the population units appear in 3 samples or less (more than 99%) compare to few (less than 0.1%) that appear in 5 or more samples . The percentage of population units appearing in 3 samples or less decreases when other sampling approaches are used to: 77% for SRS, 63.5% for PPS and 74% for the combination of SRS and PPS sampling. On the other side, the percentage of population units appearing in five samples or more increases to 8% for SRS, 19.5% for PPS and 15.5% for the combination of SRS and PPS sampling.

This is an illustration of the effectiveness of the coordination function in reducing respondent burden when applied to and compared with SRS, PPS, and a combination of SRS and PPS.

5.2 Coordination function on Agricultural Yield Small Grains, Agricultural Yield Row Crops, and Crop APS

In the second study, two scenarios were simulated based on SIP sampling and on the coordination function. Data from NASS agricultural surveys formed the population (N=11652 units). From this population from which to select three samples were selected of 2325, 1154 and 736, i.e. at sampling rates of 20% for AYS, 10% for AYR, and 7% for AS. As above, 1000 repetitions were run of each scenario.

Number of appearance	Coordination function(%)	SIP(%)
1	79	76.3
2	21	22.8
3	0	0.9

Table 2. Coordination function applied in Agricultural Yield Small Grains, Agricultural Yield Row Crops, and Crop APS of NASS

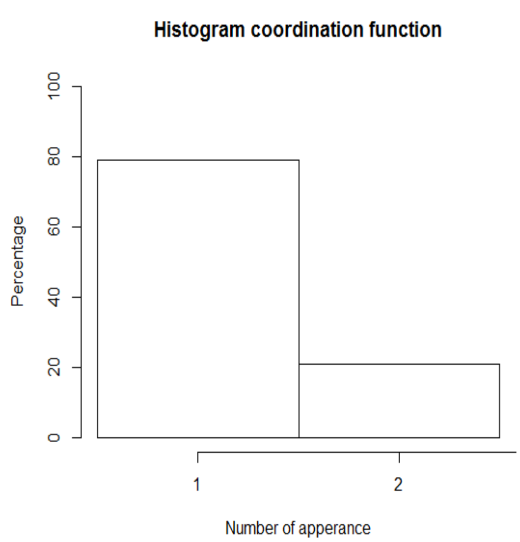

Figure 6: Histogram of number of appearances

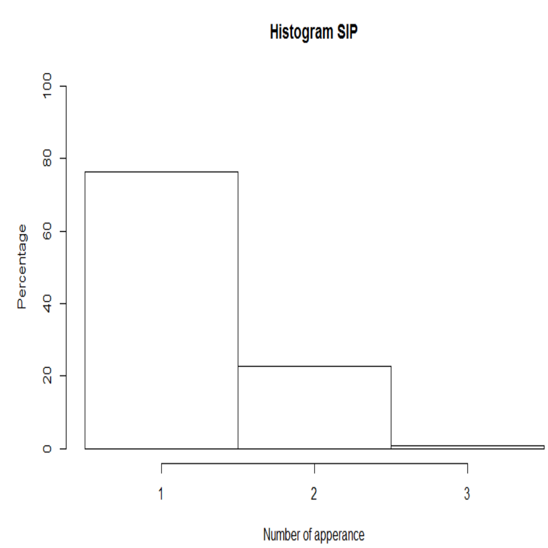

Figure 7: Histogram of number of appearances

over all samples when coordination is used in Agricultural survey data

over all samples when SIP is used in Agricultural survey data

In Table 2, “Number of appearances” indicates the number of samples in which a unit appears, while “coordination function” indicates use of the coordination function, “SIP” implies using SIP sampling.

Table 2 shows that, when the coordination function is used most of the population units appear in 3 samples or less (more than 99%) compare to few (less than 0.1%) that appear in 5 or more samples . The percentage of population units appearing in 3 samples or less decreases when other sampling approaches are used: to 77% for SRS, to 63.5% for PPS and to 74% for the combination of SRS and PPS sampling. On the other hand, the percentage of population units appearing in five samples or more increases to 8% for SRS, to 19.5% for PPS and to 15.5% for the combination of SRS and PPS sampling.

Using coordination function leads to 21% of population units being selected twice, while using SIP leads to 22.8% of population units being selected twice and 0.9% population units being selected 3 times. The major reason leading to more than 20% of population units being selected is the large inclusion probability of some units. This shows that when a coordination function is applied in AYS, AYR, and AS of NASS, i.e. and, as in this simulation the sampling rate is small, it marginally could improve the sample selection.

The difference between performances of the coordination function in the two simulation studies is due to the fact that the NASS data used here involves small sampling rates while the coordination function performs better with large sampling rates.

6. Summary and discussion

In this paper, the history and definition of respondent burden is discussed. Several sampling techniques are reviewed, including MPPS and SIP procedures used at USDA NASS. Also, the coordination function is tested on simulated data and data from three NASS agricultural surveys in two separate studies. In the first study, coordination function was compared with SRS, PPS, and a combination of SRS and PPS on simulated data, and the result showed that coordination function approach outperforms all three other approaches. In the second study, coordination function was found leading to marginal reduction in respondent burden compared to SIP. Therefore, as sampling rate increases, respondent burden increases with multiple samples and coordination function is more effective at reducing respondent burden as sampling rate increases. The next step of the work will be exploring model-assisted approaches and testing them using simulation studies.

Reference:

Bailey, Jeffrey T., and Phillip S. Kott. (1997). "An application of multiple list frame sampling for multipurpose surveys." In ASA Proceedings of the Section on Survey Research Methods (ASA Proceedings of the Section on Survey Research Methods) p. 496-500.

Berzofsky, ME, Brandon Welch, RL Williams, and Paul Biemer. (2008). Using a model-aided sampling paradigm instead of a traditional sampling paradigm in a nationally representative establishment survey. Research Triangle Park, NC: RTI Press.

Berzofsky, Marcus, McRitchie Susan, and Brendle Mark. (2012). Model-aided sampling: An empirical review. (2012). In proceedings of the Fourth International Conference on Establishment Surveys, 2012.

Bradburn, N. (1978). Respondent Burden. Proceedings of the Survey Research Methods Section of the American Statistical Association, 1978: 35-40.

Butani, Shail, Kenneth W. Robertson, and Kirk Mueller. (2000). Assigning Permanent Random Numbers to the Bureau of Labor Statistics Longitudinal (Universe) Data Base. Proceedings of the Section Research Methods.

Causey, B. D., L. H. Cox, and L. R. Ernst. (1985). Applications of transportation theory to statistical problems. Journal of the American Statistical Association, 80(392), 903-909.

Chapin, Francis Stuart. Field work and social research. Century Company, 1920.

Brenda G. Cox, David A. Binder, B. Nanjamma Chinnappa, Anders Christianson, Michael J. Colledge, Phillip S. Kott. (1995). Business Survey Methods: 153-169. Wiley.

Deville, J.-C. (1991). A theory of quota surveys. Survey Methodology, 17, 163-181.

Dodge, Y. A. (1981). Mathematical programming in statistics. John Wiley & Sons.

Ernst, Lawrence R. (1986). "Maximizing the overlap between surveys when information is incomplete." European Journal of Operational Research 27(2), p. 192-200.

Ernst, Lawrence R. (1995). Maximizing and minimizing overlap of ultimate sampling units. . Proceedings of the Section on Survey Research Methods, American Statistical Association.

Ernst, Lawrence R. (Sep 1998). "Maximizing and Minimizing Overlap When Selecting a Large Number of Units per Stratum Simultaneously for Two Designs." Journal of Official Statistics 14(3), p.297.

Ernst, Lawrence R, and Robert J. Casady. (2000). "Permanent and collocated random number sampling and the coverage of births and deaths." Journal of Official Statistics 16.3: 211.

Ernst, Lawrence R. (2001)." Retrospective Assignment of Permanent Random Numbers for OHLSSON's Exponential Sampling Overlap Maximization Procedure for Designs with more than One Sample Unit Per Stratum." Proceedings of the Annual Meeting of the American Statistical Association.

Fan, C. T., Mervin E. Muller, and Ivan Rezucha. (1962). "Development of sampling plans by using sequential (item by item) selection techniques and digital computers." Journal of the American Statistical Association 57.298: 387-402.

Fricker, S., J. Gonzalez, and L. Tan. (2011). Are You Burdened? Let's Find Out. Paper Presented at the Annual Conference of the American Association for Public Opinion Research, Phoenix, AZ.

Guggemos, Fabien, and Olivier Sautory. (2012). "Sampling coordination of business surveys conducted by Insee." In proceedings of the Fourth International Conference of Establishment Surveys.

Hedlin, Dan, Trine Dale, Gustav Haraldsen, and Jaqui Jones. (2005). Developing Methods for Assessing Perceived Response Burden. Research report, Stockholm: Statistics Sweden, Oslo: Statistics Norway, and London: Office for National Statistics.

Jacqui Jones (2012, June 11-14). Response Burden: Introductory Overview Lecture. Retrieved from https://ww2.amstat.org/meetings/ices/2012/papers/302289.pdf

King, B. (1985). Surveys combining probability and quota methods of sampling. Journal of the American Statistical Association, 80(392), 890-896.

Keyfitz, Nathan. (1951). Sampling with probabilities proportional to size: Adjustment for changes in the probabilities. Journal of the American Statistical Association. Vol 46. 253

Nedyalkova, Desislava, and Yves Tillé. 2008. "Optimal sampling and estimation strategies under the linear model." Biometrika 95.3: 521-537.

Ohlsson, Esbjörn. (1990). Sequential poisson sampling from a business register and its application to the Swedish consumer price index. Statistiska centralbyrån.

Ohlsson, Esbjorn. (1992). SAMU, The system for Co-ordination of Samples from the Business Register at Statistics Sweden. Statistics Sweden. R&D Report. Statistics Sweden. Research-Methods-Development.

Ohlsson, E. (1998). Sequential poisson sampling. Journal of official Statistics, 14(2), 149.

Petroni, Rita, Richard Sigman, Diane Willimack, Steve Cohen, and Clyde Tucker. (2004). Response rates and nonresponse in establishment surveys–BLS and Census Bureau. Federal Economic Statistics Advisory Committee, 1, 50.

Raj, D. (1957). On estimating parametric functions in stratified sampling designs. Sankhyā: The Indian Journal of Statistics (1933-1960), 17(4), 361-366.

Rosén, B. (1997). On sampling with probability proportional to size. Journal of Statistical Planning and Inference, 62(2), 159-191.

Royall, R. M. (1992). Robustness and optimal design under prediction models for finite populations. Survey methodology, 18(2), 179-185.

Saavedra, P. (1995). Fixed sample size PPS approximations with a permanent random number. 1995 Joint Statistical Meetings American Statistical Association. Orlando, Florida.

Särdnal, Carl-Erik, Bengt Swensson, and Jan Wretman. (1992). Model assisted survey sampling.

Sharp, Laure M, and Joanne Frankel. (1983). Respondent Burden: A Test of Some Common Assumptions. Public Opinion Quarterly, 47, 36-53.

Smith, T. M. F. (1983). On the validity of inference from non-random samples. Journal of the Royal Statistical Society: Series A., Pt. 4, 146, 394-403

USDA Economic Research Service. (October 12, 2016). ARMS Farm Financial and Crop Production Practices. http://www.ers.usda.gov/data-products/arms-farm-financial-and-crop-production-practices/documentation.aspx.

Stephenson, C. B. (1979). Probability sampling with quotas: An experiment. *Public Opinion Quarterly*, *43*(4), 477-496.