

# SIMULATING SURVEY DESIGN WEIGHTS TO ACCOUNT FOR SAMPLE COORDINATION BETWEEN SURVEYS

Joshua Gutoskie<sup>1</sup>

## ABSTRACT

In an effort to reduce respondent burden, sample coordination often takes place to remove overlapping units between surveys. The removed units are often replaced with other eligible units that were not included in the pre-coordinated sample. Calculating the first-order inclusion probabilities is often quite difficult when trying to account for the removal and replacement of units within a sample. Statistics Canada's Farm Management Survey's unique sample design required several stages of sample coordination not only with other agriculture surveys but also within its own sample, which made calculating the first-order inclusion probabilities challenging. In this paper, we use the Monte Carlo simulation-based approach to estimate first-order inclusion probabilities, as proposed by Thompson and Wu (2008), applying its use to account for the sample coordination done for the Farm Management Survey.

KEY WORDS: Sample coordination; Inclusion probabilities; Monte Carlo simulation

## RÉSUMÉ

Dans le but d'alléger le fardeau du répondant, l'échantillonnage est souvent coordonné pour éliminer le chevauchement des unités entre les questionnaires. Les unités éliminées sont souvent remplacées par d'autres admissibles qui ne sont pas comprises dans l'échantillonnage pré-coordination. Le calcul des probabilités d'inclusion de premier ordre se révèle souvent difficile lorsque nous tentons de prendre en compte l'élimination et le remplacement d'unités dans un échantillonnage. La conception unique de l'Enquête sur la gestion des fermes de Statistique Canada a exigé plusieurs étapes de coordination de l'échantillonnage, non seulement par rapport à d'autres enquêtes portant sur l'agriculture, mais à l'intérieur même de son propre échantillonnage, compliquant ainsi le calcul des probabilités d'inclusion de premier ordre. Pour estimer ces probabilités, nous faisons appel à une méthode de simulation de Monte-Carlo, comme l'ont proposé Thompson et Wu (2008), en appliquant son utilisation à la prise en compte de la coordination de l'échantillonnage pour l'Enquête sur la gestion des fermes.

MOTS CLÉS : Coordination de l'échantillonnage; Probabilités d'inclusion; Simulation de Monte-Carlo

## 1. INTRODUCTION

With any complex survey, producing accurate survey weights are essential when making valid inferences about the population of interest. The first step to create these weights is calculating the first-order inclusion probabilities. These probabilities are often quite simple to calculate. For example, if the sampled units follow a stratified simple random sample without replacement (SSRSWOR) design then the first-order inclusion probabilities are just the ratio of the number of sampled units within a stratum to the total number of units in the population within the same stratum. More formally, let  $N_h, h = 1, 2, \dots, L$  be the population size of stratum  $h$  and  $n_h$  be the sample size of stratum  $h$ . Let  $s$  be the set of sampled units and  $\pi_{hi} = P(i \in s), i = 1, 2, \dots, N_h$  be the first-order inclusion probability of sampling unit  $i$  in stratum  $h$ . The first-order inclusion probability for a SSRSWOR design is given by

$$\pi_{hi} = \frac{n_h}{N_h}. \quad (1)$$

Survey design weights are the inverse of these first-order inclusion probabilities,  $w_{hi} = 1/\pi_{hi}$ . These weights are then further modified to account for other factors, such as non-response within a survey, to produce the final survey weights, which are important for the final analysis.

---

<sup>1</sup> Joshua Gutoskie, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, K1A 0T6, [joshua.gutoskie@canada.ca](mailto:joshua.gutoskie@canada.ca).

Challenges arise when modifications are made following the initial sample selection, whether it is due to operational constraints or other potential circumstances. One such operational constraint is response burden reduction. Statistics Canada has put increased emphasis on maintaining high quality estimates while at the same time, not overly burdening any individual respondent. One of the ways that a survey program can reduce response burden on an individual or business is through negative sample coordination (Royce, 2000), which aims to minimize the overlap of units between samples. Once coordination has taken place between samples, the first-order inclusion probabilities for a SSRSWOR may no longer satisfy (1) and the direct calculation of these probabilities can become quite difficult. In practice, if the amount of coordination is small then this is often ignored and the survey design weights are calculated as if no modification to the original design had been made. However, the larger the modification, the larger the risk that the inferences made about the population may no longer be valid.

In this paper, we use the simulation-based estimation of first-order inclusion probabilities as described by Thompson and Wu (2008) to estimate the survey design weights of Statistics Canada’s Farm Management Survey (FMS). An overview of the FMS, along with a description of its sample design and coordination, will be given in section 2. The simulation method proposed by Thompson and Wu (2008) is described in section 3. The results of the FMS simulation are given in section 4. Finally, concluding remarks will be given in section 5.

## 2. FARM MANAGEMENT SURVEY

### 2.1 Survey Overview

The Farm Management Survey (FMS) is conducted by Statistics Canada every five years, following the Canadian Census of Agriculture. The data gathered from the FMS is used to help measure management practices within the Canadian agriculture industry, to address federal and provincial policy needs, and to support the development of effective agricultural programs pertaining to sustainable development. In past occasions of the FMS (known as the Farm Environmental Management Survey prior to the 2018 occasion), eligible farms were classified as either crop farms, livestock farms or both and, if selected, the farm operator would receive either a crop or livestock questionnaire. For the 2018 occasion, the survey further divided the population into seven agricultural subsectors: dairy, beef, pig, and poultry farms for the livestock sector and field crop, forage crop, and vegetable, fruit, nut, and berry farms (referred to as horticulture farms in this paper) for the crop sector. Due to the more detailed scope of the survey for the 2018 occasion, the sample design and coordination strategy needed to be redesigned.

### 2.2 Sample Design and Coordination

The FMS can be seen as seven different surveys, one for each subsector of interest. Each subsector’s sample of farms is selected independently from a frame of farms undertaking this activity using a SSRSWOR design, stratified by sub-provincial region and by a subsector-specific size measure. For example, the size measure used for the field crop subsector was the farm’s field crop acreage obtained from the 2016 Census of Agriculture, while the size measure used for the dairy subsector was the farm’s number of dairy cattle. The total sample size for the FMS is 18000 farms.

Since farms can produce more than one agricultural product, farms can be included in the sample frame of more than one subsector. These overlapping farms can be initially selected in the sample of more than one subsector. Table 1 below shows the proportion of farms by the number of subsectors that they belong to. This is broken down for all units in the FMS population, as well as the farms that were selected in the initial sample.

**Table 1 – Proportion of farms by number of subsectors**

Number of subsectors	Proportion of farms on frame (%)	Proportion of farms initially sampled (%)
1	65.7%	93.5%
2	25.6%	6.1%
3	8.3%	0.4%
4+	0.4%	< 0.1%

There is a large amount of overlap between the seven subsectors. 34.3% of farms are included in more than one subsector's frame. At first glance, it would seem that there would be significant overlap once the initial sample is drawn. However, only 6.5% of the farms selected in the initial sample were chosen in multiple subsectors.

The FMS is a lengthy survey so in order to reduce response burden, only one questionnaire would be sent to a farm. If a farm was chosen in more than one subsector's sample then it would be randomly kept in only one sample. Once the overlap between the subsectors was removed, replacement farms were randomly selected from the affected strata to return the sample to its initial size.

On top of the coordination between the subsectors within FMS, sample coordination was conducted between the FMS and Statistics Canada's Farm Financial Survey. The Farm Financial Survey is a burdensome survey, like the FMS, with both surveys being collected around the same time of the year. Because of this, we wanted to reduce the number of farms that would be contacted for both surveys. To initially minimize the overlap between the two surveys, the microstrata method (Rivière, 2001) was used. Microstrata are created by overlapping the two sample designs and minimizing the overlap in the selected samples between the two surveys within the microstrata. This method is unbiased for both surveys. However, having a sufficient number of units within a microstratum to meet the sample size needs of both surveys was not always possible, resulting in some overlap remaining between the surveys. In order to further reduce the remaining overlap between these surveys, additional coordination steps were needed. However, these steps were potentially biased. Unlike the coordination done between the subsectors, the overlap between FMS and the Farm Financial Survey could not be completely mitigated. In the end, 9.9% of the farms in the final FMS sample were chosen as replacements for the overlapping farms. Table 2 below shows the population size, final sample size, as well as the percentage of replacements needed following all coordination steps for all subsectors.

**Table 2 – FMS population and sample counts, and replacement percentages by subsector**

Subsector	Population	Sample	Replacement (%)
Dairy	7913	1658	6.6%
Beef	24346	3042	10.4%
Poultry	2315	1146	11.0%
Pigs	2137	823	10.1%
Field Crops	44119	4864	11.3%
Forage Crops	43759	4236	11.0%
Horticulture	4972	2231	6.1%

### 3. SIMULATION METHOD

Thompson and Wu (2008) describe the use of a Monte Carlo simulation-based method to estimate the first-order inclusion probabilities for surveys that require sampled units to be replaced. In their example, they look at the International Tobacco Control Policy Evaluation Survey of China (ITC China Survey). The survey uses a multi-stage unequal probability sample design to select smoking and non-smoking adults from clusters of districts and blocks within seven cities. Several of these clusters refused to participate in the survey. Replacement clusters were randomly selected from the remaining clusters that were not included in the initial sample. In this example, a modification to the original sample design was needed because of the refusing units, which is out of the control of the survey methodologist. For the FMS, the modification to the original design was imposed due to the constraint of the coordination.

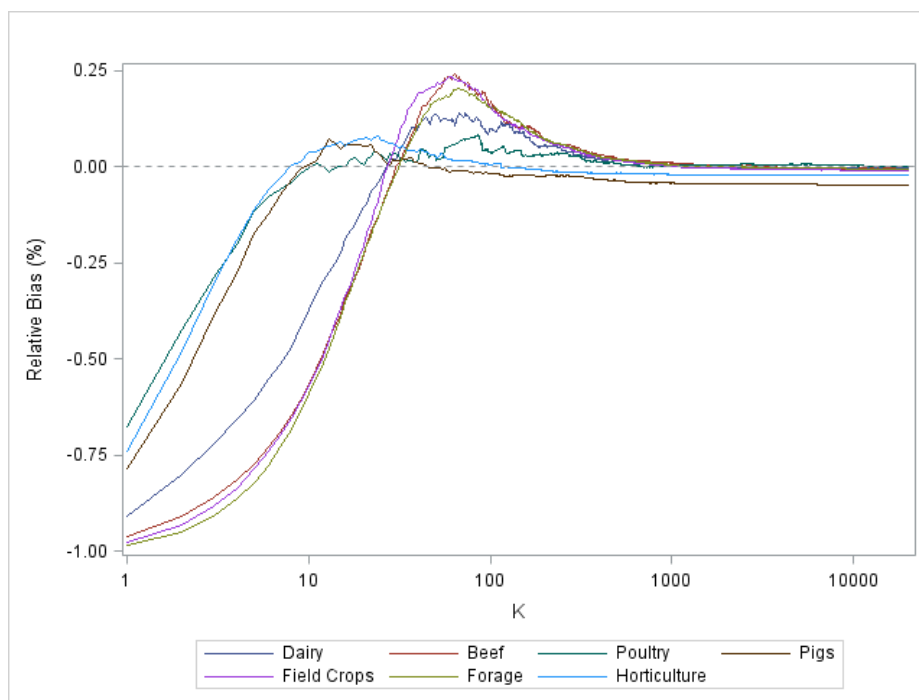
To estimate the first-order inclusion probabilities for the FMS using a simulation-based method, we assume a completely specified probability sample design, including any potential sample coordination, denoted by  $p$ . Select  $K$  independent samples, where each sample follows  $p$ . Let  $M_i$  be the number of samples for which unit  $i$  was included in the final sample. The first-order inclusion probability  $\pi_i = P(i \in s)$  for unit  $i$  can be estimated by  $\pi_i^* = M_i/K$ . The design weight of unit  $i$  is estimated by

$$w_i^* = \frac{K}{M_i} \tag{2}$$

There are some drawbacks to using simulation methods, of which an important one is computational resources. In their paper, Thompson and Wu (2008) show that, theoretically, one would need  $K = 10^8$  independent samples for the Horvitz-Thompson estimator using simulated probabilities to perform relatively on par with the Horvitz-Thompson estimator using the true probabilities. Though this is an upper bound, in the situation where the difference between the population size and the sample size is equal to 100 (as is the case with the ITC China Survey), computing  $10^8$  samples can take a large amount of computing time. As shown earlier, the difference between the FMS population size and sample size was much larger than 100, thus more than  $10^8$  samples would have been needed to achieve this. Computing  $10^8$  or even  $10^6$  is not always practical so a compromise must be made between the number of samples needed for the simulation and the potential bias that this would introduce.

Taking into account this compromise between  $K$  and the potential bias that this could introduce, snapshots were taken at various points of the FMS simulation to assess the convergence of the simulated design weights. The relative bias between the population size  $N$  and the Horvitz-Thompson estimator for  $N$  using the simulated weights,  $\tilde{N} = \sum_{i=1}^n w_i^*$  was used to evaluate if the simulated weights had converged. The relative bias is given by  $(\tilde{N} - N)/N$ . Figure 3 below shows how the relative biases of the population counts behave as  $K$  increases for all subsectors.

**Figure 3 – Relative bias of population counts by subsector**



Interestingly, there looks to be two different paths taken. The subsectors with larger populations, beef, field crops, forage crops, all begin with the lowest relative biases then peak to approximately 25% before converging to approximately 0%. The smaller subsectors, pigs and horticulture, both converge to a relative bias below 0%. One possible reason for this is due to the fact that there were not always replacement units available within these subsectors during the coordination steps. For all subsectors, the relative bias of the population counts converged by the time that  $K = 20000$ .

#### 4. SIMULATION RESULTS

To assess how well the simulation method estimated the first-order inclusion probabilities, we look at the performance of the Horvitz-Thompson estimator for a population total. Recall that the Horvitz-Thompson estimator for a population total  $T = \sum_{i=1}^N y_i$  for a variable of interest  $y$  is given by  $\hat{T} = \sum_{i=1}^n w_i y_i$ . To test the performance, we look at the relative bias between the population total  $T$ , and the Horvitz-Thompson estimator using the estimated design weights,  $\tilde{T} = \sum_{i \in S} w_i^* y_i$ . The relative bias is given by  $(\tilde{T} - T)/T$ . In this study, a subsector-specific response variable  $y \geq 0$  was simulated for all units in each subsector using the model  $y_i = \beta_0 + \beta_i x_i + \varepsilon_i, i = 1, 2, \dots, N$ , where  $x_i$  is the size variable that was used for stratification for the subsector (field crop acreage, heads of dairy cattle, etc.) and  $\varepsilon_i$  is independently and identically

distributed following a normal distribution with mean 0 and population variance  $\sigma^2$ . We also wanted to see how the relative bias changed for differing values of  $\sigma^2$ . Three different population variances were chosen such that the population correlation coefficients  $\rho_{xy}$  between  $y_i$  and  $x_i$  were 0.1, 0.5, and 0.9 respectively. Table 4 shows the relative biases of the three different population totals for the seven subsectors.

**Table 4 – Relative bias of Horvitz-Thompson estimator using simulated weights**

Subsector	Relative Bias (%)		
	$\rho_{xy} = 0.1$	$\rho_{xy} = 0.5$	$\rho_{xy} = 0.9$
Dairy	0.1%	-0.2%	-0.2%
Beef	0.3%	-0.3%	-0.5%
Poultry	0.5%	0.4%	0.2%
Pigs	-5.7%	-4.6%	-4.6%
Field Crops	-0.6%	-0.9%	-0.8%
Forage Crops	-0.6%	-0.6%	-0.5%
Horticulture	-2.5%	-2.1%	-2.1%

For most of the subsectors, the relative bias is below 1% with the exception of the pigs and horticulture subsectors. As pointed out by Reicker and Mohl (2018), the coordination strategy that was used for the FMS was not unbiased. A larger bias occurs in populations where a large farm is overlapping but there are no replacement farms available in the stratum. The relative biases in Table 4 are very similar to the relative biases seen in the population counts. To reduce this bias, a calibration estimator controlling for strata counts, is proposed. Table 5 shows the relative bias of the same three different population totals when applying the calibration estimator.

**Table 5 – Relative bias of calibration estimator using simulated weights**

Subsector	Relative Bias (%)		
	$\rho_{xy} = 0.1$	$\rho_{xy} = 0.5$	$\rho_{xy} = 0.9$
Dairy	0.3%	< 0.1%	< 0.1%
Beef	0.7%	0.1%	< 0.1%
Poultry	-0.1%	< 0.1%	< 0.1%
Pigs	-0.8%	0.2%	0.1%
Field Crops	0.2%	-0.1%	< 0.1%
Forage Crops	-0.1%	-0.1%	< 0.1%
Horticulture	-0.4%	< 0.0%	-0.1%

Calibrating to the strata counts does reduce the relative bias for the majority of the estimates, resulting in relative biases below 1% for all subsectors and populations. It is important to note that the particular variables of interest for the FMS may have a significant less bias given that management practices are expected to be more uniform than the size of the farm.

## 5. CONCLUSION

Due to operational constraints, such as sample coordination, it is not always easy to directly calculate first-order inclusion probabilities that are necessary when analyzing complex survey data. In the situation where these probabilities cannot be accurately calculated, they can be estimated through a Monte Carlo simulation. This was successfully done in a production environment at Statistics Canada for the Farm Management Survey. The bias that was introduced due to the simulation was negligible on the estimates, compared to the bias that was introduced due to the coordination. This bias is reduced when using a calibration estimator, which controls for the strata counts. The simulation method can be extended to any sample design, as long as all information of the design is available.

## ACKNOWLEDGEMENTS

The author would like to thank Chris Mohl, John Marshall, and Wesley Yung for reviewing this paper and for their insightful comments.

## DISCLAIMER

The content of this article represents the position of the author and may not necessarily represent that of Statistics Canada.

## REFERENCES

- Reicker, A., and Mohl, C. (2018). "Farm management survey 2016 sample plan". *Statistics Canada internal document*.
- Rivière, P. (2001). "Coordinating samples using the microstrata methodology". *Proceedings of Statistics Canada Symposium 2001*.
- Royce, D. (2000). "Issues in coordinated sampling at Statistics Canada". *Proceedings of the Second International Conference on Establishment Surveys, American Statistical Association*.
- Thompson, M.E., and Wu, C. (2008). "Simulation-based randomized systematic PPS sampling under substitution of units". *Survey Methodology*, **34** (1): 3-10. Statistics Canada Catalogue no. 12-001-X.