Sample Redesign of Canadian Local Government Surveys

James Ahkong and Martin Renaud
Statistics Canada, R.H. Coats Building, 11th Floor, Ottawa, Ontario, K1A 0T6, Canada
james.ahkong@statcan.gc.ca, martin.renaud@statcan.gc.ca

Abstract

Financial data for Canadian local governments are collected through various quarterly and annual surveys, known collectively as Local Government Surveys. For the last ten years, these surveys have been mailing out questionnaires to the same sample of municipalities. Over the years this has led to an increase in response burden, especially for small municipalities. A redesign has been proposed where annually updated population counts of municipalities would be used to select a new yearly sample as follows: large municipalities with certainty, small municipalities with a probability based on their population size, and the exclusion of the smallest municipalities. This paper will briefly summarize the Local Government program and will mainly focus on the various methodological issues behind the sample redesign of the Local Government Surveys.

Key Words: Lavallée-Hidiroglou algorithm, take-all stratum, take-some stratum, PPS sampling, municipalities

1. Introduction

As part of its public sector statistics program, Statistics Canada is responsible for collecting financial data for local governments. This is achieved through a combination of administrative sources and surveys. Administrative data are collected through a census of all municipalities by the Department of Municipal Affairs (DMA) for most provinces and from the financial statement of municipalities where DMA data are not available. Because these data are not available in a timely manner, a series of surveys, known as the Local Government Surveys (LGS), are also carried out. There are two types of LGS: quarterly and annual. Financial flows, the change of financial data between quarterly periods, are collected through the quarterly surveys. More detailed financial information is collected through the two annual surveys: the Current Revenue and Expenditures Survey and the Capital Expenditures Survey (CES).

All legal municipalities in Canada, with the exception of First Nations type, are in-scope for the surveys. In total there are approximately 3,600 in-scope legal municipalities across Canada. For every legal municipality, the number of residents is known through the use of Census data published by Statistics Canada.

The current sample design for the LGS was created in 1998. A sample of municipalities is selected, and each LGS uses the same sample. Eleven years later in 2009, a sample redesign has been devised that addresses the issues that have arisen over the years in regards to the previous sample design. The next section of this paper will present the methodology behind the 1998 sample design and the issues that have arisen with this design. The methodology behind the

creation of the new sample design for 2009 is presented in Section 3 followed by some results of the new sample that was selected in Section 4. We conclude with a brief overview of additional improvements and future work that is considered to further improve the data quality of the LGS.

2. 1998 Sample Design Methodology and Issues

In 1998 a stratified simple random sample design was constructed for the LGS. Three levels of stratification were used:

- Province
- Municipality Type
- Size Groupings

Municipalities are located in different areas across Canada which are represented by the provinces. Municipalities are defined by type such as city, town or village to name a few. Size groupings were constructed based upon the number of residents in a municipality which was available for every municipality from the 1996 Census of Population. A take-all size grouping, where each municipality is selected for the sample with certainty based upon its size, was created within each province while a variable number of take-some groupings, where municipalities are randomly selected for the sample with a probability less than 1, were constructed depending on the number of municipalities present in a province. Boundaries for the take-all and take-some groupings were determined through the optimization of the stratification with respect to the variance of the mean. Because there were different size groupings and municipality types per province, there were a different number of strata groups formed within each province. In total there were 53 unique strata for this sample design. The sample of approximately 500 municipalities was allocated across the strata using a Bankier *p*-allocation (1988).

Municipalities within a take-some stratum were selected using simple random sampling. Final estimates were calibrated to the total number of residents by province based on the 1996 Census. Provincial estimates were also aggregated to produce estimates at the national level.

In addition to the legal municipalities, there are other entities that are part of the target population. They include 31 regional municipalities and 14 other entities that report expenditures and complement the legal municipalities in our universe. These 45 entities were added to the sample following the selection of the legal municipalities.

A sample of municipalities using this design was first selected in 1998. Every following year since, the same sample of municipalities has been used. The only updates made to the sample on a yearly basis are the result of changes in municipality boundaries, and of amalgamations and de-amalgamations of municipalities. Using the same sample each year created a very high level of response burden on the municipalities. Response burden was a greater issue for the smaller municipalities. In general, larger municipalities have their questionnaires completed by a knowledgeable person from their accounting or finance department. They are also accustomed to being selected for various economic surveys and are willing to participate and provide the required information. Smaller municipalities, however, lack these kinds of resources and the burden of a repeated survey causes a fair share of the municipalities to stop collaborating after a certain period of time.

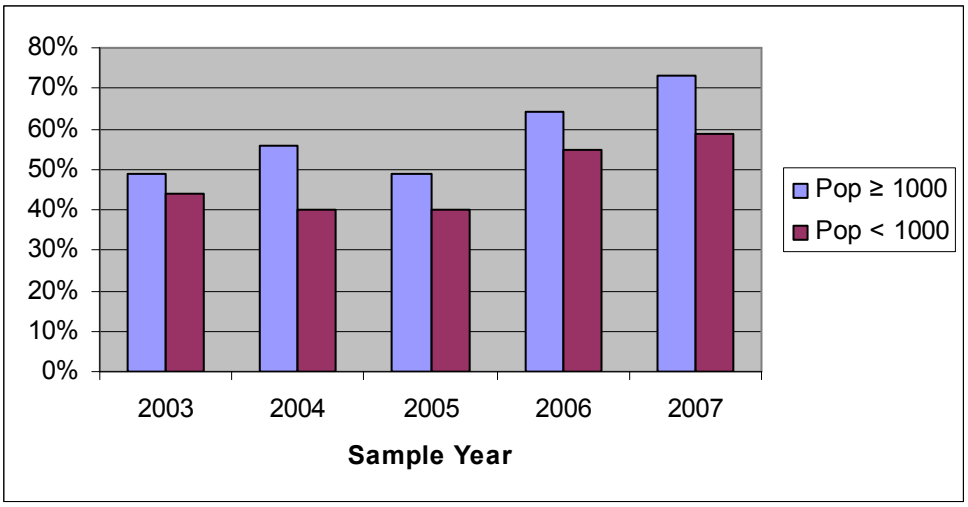

Figure 1: Response Rates of CES

Figure 1 presents the most recent response rates of the CES for municipalities with fewer than 1,000 residents (small) and those for municipalities with at least 1,000 residents (large). The chart shows that the small municipalities consistently respond at a lower rate than the larger municipalities. In 2004 and 2005, their response rate was as low as 40%. It should be noted that the increases in response rates for the past two years are due to an increase in the intensity of the collection process for the survey. Despite the increase in response rates, it should be noted that the smaller municipalities still report at a lower rate than the larger ones.

Having the same sample of municipalities year after year also created problems with coverage. Every municipality created since 1998 had no chance of being selected to the sample. Even though adjustments were made to the sampling weights to account for changes in municipal boundaries, amalgamations and de-amalgamations, the LGS still suffered from under-coverage.

Because of the SRS design used, in giving every unit in a take-some stratum an equal chance of being selected in the sample, a significant number of small municipalities were chosen. The small municipalities were usually small villages and they responded to the questionnaire, as just mentioned, at a lower response rate. Furthermore, these municipalities did not report large amounts of financial data. Another reason why a significant number of small municipalities ended up in the sample was due to the level of detail covered by the stratification. Having municipality type as a stratification variable ensured that a significant portion of villages would end up in the sample. Because the estimates are only produced at a higher level (*i.e.*, by province), perhaps having a level of stratification that detailed was not necessary.

Based on these observations from the 1998 sample design, the time had come to develop a new sample design which would address those issues, namely simplify stratification, select fewer small municipalities, provide better coverage and allow a new sample to be selected on a yearly basis in a simple and efficient manner.

3. New Sample Design for Local Government Surveys

In constructing the new sample design for the LGS, it was decided that the design would be tailored primarily towards the CES, the most complex and highest profile of the LGS. A new sample will be selected every year as opposed to the panel type sample that ended up being

implemented for the past ten years. This will certainly ease the response burden placed on the smaller municipalities that have been contacted repeatedly and provide better coverage of the target population.

3.1 Correlation of the Auxiliary Variable with the Reported Variable

Every year, Statistics Canada's Geography Division produces a complete list of all legal municipalities in Canada. The latest changes in municipal boundaries are reflected on that list and the number of residents in each municipality is adjusted accordingly. New municipalities that have been created in the last year are added to the list and municipalities that have ceased to exist are removed. This list is used as a frame for the LGS.

However, this frame of Canadian municipalities has a limited number of auxiliary variables available to help develop the new sample design. There are a few geographical and descriptive variables but most importantly, the number of residents in a municipality is available for every single municipality on the frame and is updated annually at Statistics Canada using demographic projections from Census data. Our goal is to use such a valuable piece of auxiliary information in constructing a more effective sample design than the previous one. Before incorporating this variable in the new design, we had to verify that it was sufficiently correlated with our variable of interest, the total capital expenditures. An analysis of the relationship between the number of residents in a municipality and the amount of expenditure reported by a municipality was carried out using administrative data. This analysis was carried out for each province and it was found that the relationship between number of residents and expenditures was quite strong. We explored the relationship in more detail by grouping the municipalities by size (small, medium and large). It was found that the correlation was strong for the larger and smaller municipalities, while the medium size municipalities had a relationship that was weaker but still considered correlated. The results of the relationship showed that the larger the number of residents in a municipality, the greater the amount of expenditures that can be expected from that particular municipality. Based on those results, it was decided to make use of the number of residents per municipality in our new design.

3.2 Exclusion of Municipalities

The distribution of the number of residents in municipalities across Canada and within each province is skewed; there are a few municipalities that have a large number of residents and these municipalities account for a large portion of the total national expenditures. Likewise, there are many smaller municipalities with a small number of residents. As previously mentioned, there is a strong correlation between expenditures and the number of residents in a municipality and we know from administrative data that these smaller municipalities usually have very little expenditures to report. Furthermore, it has historically been more difficult to convince these smaller municipalities to participate in the LGS as they do not see the relevance given the minimal amounts they report. Also, they usually do not have the kind of resources on hand and available to answer the questionnaires. With that in mind, excluding them from the survey was considered a viable option to ensure that they do not end up in the sample. The question was then to determine how many small municipalities should be excluded. To answer that question we decided to apply the exclusions at a provincial level. Instead of excluding a fixed number of municipalities, we chose to use their number of residents and remove the municipalities that totaled the bottom x% of the total number of residents within a province. We considered excluding both the bottom 1% and bottom 2% of municipalities in terms of population size. Table 1 presents the number of exclusions by province for both the 1% and 2% scenarios.

Table 1: Number of Municipalities to Exclude by Province and Territory

Province	Number		
	Total	1% exclusion	2% exclusion
Newfoundland (NL)	282	36	58
Prince Edward Island (PE)	42	5	9
Nova Scotia (NS)	46	8	12
New Brunswick (NB)	104	13	21
Québec (QC)	1159	221	342
Ontario (ON)	415	127	167
Manitoba (MB)	197	27	47
Saskatchewan (SK)	792	156	224
Alberta (AB)	344	136	174
British Columbia (BC)	162	44	61
Yukon (TK)	8	1	1
Northwest Territories (NT)	22	1	2
Nunavut (NU)	25	1	2
Total	**3,598**	**776**	**1,120**

Based on these results, it was decided to go ahead and proceed with excluding those municipalities which represented the bottom 2% of the population in terms of number of residents for each province. This excluded portion of the population is commonly referred to as the take-none stratum. It is interesting to note that these exclusions translate into removing approximately 30% of the total number of municipalities in Canada. Finally, based on administrative data from the previous years, these 1,000 or so municipalities account for less than 0.5% of the total annual expenditures in Canada. As was the case in 1998, the First Nations type of municipalities were excluded from the frame.

3.3 Stratification

The 1998 design had a level of stratification that was considered too detailed. Besides stratifying by province, the sample design also stratified by size of municipalities (with a varying number of size groups from province to province) and by municipality type. Since no estimates were produced by municipality type and this stratification variable did not contribute to improvements in the efficiency of the design, it was decided to remove it and use only province and size in the new stratification. For the size variable, size groupings using the number of residents in a municipality were constructed. The idea was to create a take-all size grouping and at least one take-some grouping within each province. One of our goals was to include the larger municipalities of each province in the sample with certainty. The challenges associated with the creation of the size groupings were two-fold. First, how does one determine the cut-offs used to separate the groupings and second, how many take-some groups should be created? These challenges are addressed in the next section.

3.4 Sample Allocation

Going into the sample redesign, it was known that based upon the budget, the target sample size for the LGS would once again be approximately 500 municipalities. We therefore had to find a way to separate the take-all and take-some strata, determine how many take-some strata were needed, and allocate the sample while respecting the total sample size restriction. To address

these challenges, the auxiliary information of the number of residents in a municipality based on the 2006 Census was put to use through the Lavallée-Hidiroglou (1988) algorithm. This method calculates a cut-off value to separate the take-all and take-some strata while minimizing the sample size required to achieve a certain target level of precision. For this method to be effective, the auxiliary variable must be correlated to the variable of interest which is the case for the LGS. The application of the Lavallée-Hidiroglou method was tested by using different combinations of the target level of precision across the provinces, the maximum sampling weight, and the percentage of total population to exclude. Target coefficients of variation (CVs) of 5% and 10% were used in conjunction with maximum sampling weights of 10 and 20, and an exclusion of the bottom 1% and 2% of the total provincial population for the take-none stratum. Because of the highly skewed distribution of the population, it was difficult to create more than one take-some size grouping for every province. Therefore, we decided to focus on creating only a single take-some stratum within each province.

Table 2: National Sample and Size Strata Counts Excluding Bottom 2% of Population

Target CV	Maximum Weight	Sample Size	Number of TA Units	Number of TS Units
10%	20	283	150	133
10%	10	383	150	233
5%	20	406	256	150
5%	10	484	256	228

Table 3: National Sample and Size Strata Counts Excluding Bottom 1% of Population

Target CV	Maximum Weight	Sample Size	Number of TA Units	Number of TS Units
10%	20	312	163	149
10%	10	430	163	267
5%	20	443	274	169
5%	10	536	274	262

Table 4: National Sample and Size Strata Counts With No Exclusions

Target CV	Maximum Weight	Sample Size	Number of TA Units	Number of TS Units
10%	20	372	180	192
10%	10	524	180	344
5%	20	536	328	208
5%	10	662	328	334

Looking at the results of Tables 2 to 4, it was decided to adopt the following parameters which provided a total sample size that was close to the target of 500: a target CV of 5%, a maximum sample weight of 10, and the exclusion of the bottom 2% of the population for a total sample size of 484 municipalities.

3.5 Sample Selection

The Lavallée-Hidiroglou algorithm has been developed on the assumption that the sample selected from the take-some strata would be a simple random sample. However, in the case of the LGS, we could observe large fluctuations in the number of residents for the municipalities in the take-some stratum in most provinces. Given that situation, our objective was to give larger

municipalities a greater chance of being selected to the sample. Selecting the municipalities using probability proportional to size sampling (with the number of residents in a municipality as the size variable) was an option to be considered. Another option would be to use simple random sampling as was done in 1998 with the difference being that the very small municipalities would not appear in the sample since we had already decided to exclude them from the design. Any use of simple random sampling would be accompanied by the use of a ratio estimator in order to calibrate to the number of residents in the take-some stratum.

3.5.1 Simulation of Methods for Sample Selection of Take-Some Strata

To decide which sampling strategy should be adopted for the take-some strata, we decided to carry out simulations to compare the different options. The sample selection methods used in the simulation were simple random sampling (SRS), probability proportional to size sampling (PPS) and systematic probability proportional to size sampling (PPS Sys). In the case of PPS Sys, prior to selection of units the sample was ordered by size. For SRS, even though early indications pointed out that one take-some stratum per province was preferable we also considered the two take-some strata scenario in our simulations. Similar to the sample design of 1998, a ratio estimator using the number of residents as the auxiliary variable was used for the SRS methods. For each sample selection method, 1,000 samples were generated. Each of the methods was compared using the CV and relative absolute bias of its estimator. For the simulation, our population of Canadian municipalities was created using capital expenditures administrative data from the year 2005.

Table 5: CVs and Relative Absolute Biases of Estimators by Province and Territory

	SRS-1		SRS-2		PPS		PPS Sys	
Province	**CV**	**RB**	**CV**	**RB**	**CV**	**RB**	**CV**	**RB**
NL	3.35%	2.57%	3.21%	2.48%	3.23%	2.52%	4.13%	2.83%
PE	2.16%	1.92%	1.38%	1.18%	1.81%	1.56%	1.10%	0.91%
NS	5.38%	4.30%	5.34%	4.33%	4.97%	3.83%	4.44%	3.66%
NB	9.22%	7.39%	9.21%	7.33%	10.50%	7.39%	9.06%	6.06%
QC	4.01%	3.23%	4.32%	3.42%	3.92%	3.12%	4.85%	3.65%
ON	4.50%	3.65%	4.36%	3.48%	3.50%	2.78%	3.60%	2.96%
MB	5.67%	4.70%	5.14%	4.21%	3.99%	3.21%	2.49%	1.97%
SK	1.98%	1.54%	1.93%	1.53%	1.22%	0.97%	0.79%	0.64%
AB	5.16%	3.78%	5.41%	3.90%	3.99%	3.22%	3.81%	2.90%
BC	4.59%	3.55%	4.35%	3.33%	3.81%	3.01%	3.86%	3.24%
YK	2.04%	1.64%	1.93%	1.67%	1.87%	1.50%	1.87%	1.68%
NT	10.17%	8.54%	5.77%	5.45%	6.78%	6.09%	5.57%	5.52%
NU	7.13%	5.68%	3.63%	2.96%	7.50%	6.14%	7.25%	5.98%

Table 6: Ranks of Estimators by Province and Territory

	Coefficient of Variation				Relative Absolute Bias			
Province	**SRS-1**	**SRS-2**	**PPS**	**PPS-Sys**	**SRS-1**	**SRS-2**	**PPS**	**PPS-Sys**
NL	3	1	2	4	3	1	2	4
PE	4	2	3	1	4	2	3	1
NS	4	3	2	1	3	4	2	1
NB	3	2	4	1	3.5	2	3.5	1
QC	2	3	1	4	2	3	1	4
ON	4	3	1	2	4	3	1	2
MB	4	3	2	1	4	3	2	1
SK	4	3	2	1	4	3	2	1
AB	3	4	2	1	3	4	2	1
BC	4	3	1	2	4	3	1	2
YK	4	3	1.5	1.5	2	3	1	4
NT	4	2	3	1	4	1	3	2
NU	2	1	4	3	2	1	4	3
Total	**45**	**33**	**28.5**	**23.5**	**42.5**	**33**	**27.5**	**27**

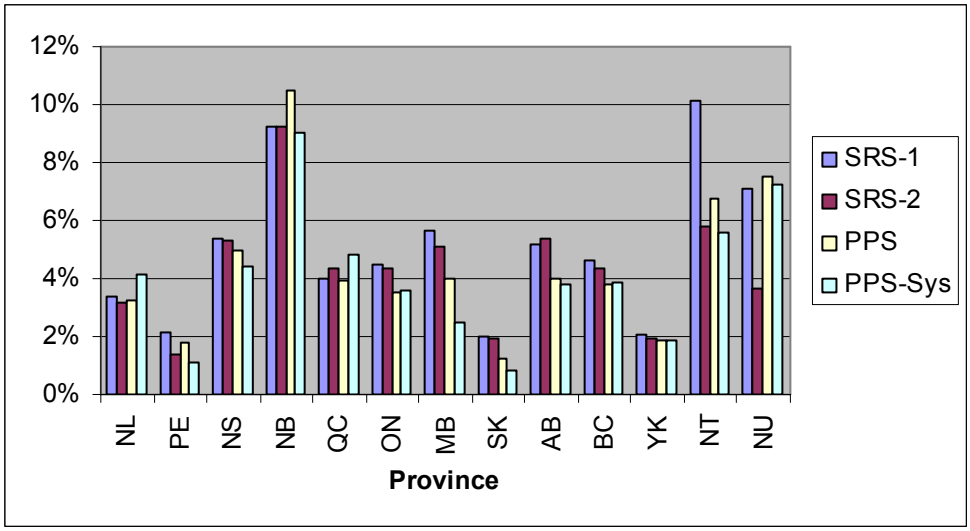

Figure 2: CV of Estimators by Province and Territory

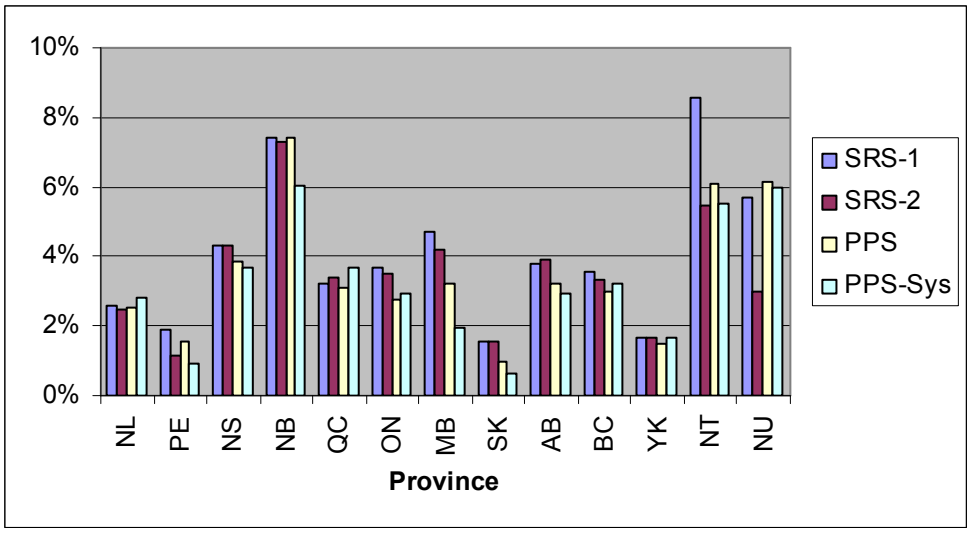

Figure 3: Absolute Relative Bias of Estimators by Province and Territory

The results of the simulation are presented in Tables 5 and 6, and Figures 2 and 3. The results show that going from one take-some stratum to two take-some strata did not significantly improve the CV or the bias of the estimate with the exception of the Northwest Territories and Nunavut. Comparing SRS to PPS, it is shown that both methods lead to similar CVs and biases, however, the PPS methods have lower CVs and biases. The PPS systematic method generally produced the best results and therefore, we decided to use this method to select municipalities from the take-some strata.

4. 2009 Sample

For the first time, a sample for the LGS was selected with the new design in early 2009. The distribution of the sample is presented in Table 7. Including the 45 regional municipalities and other entities, 529 municipalities were selected, an increase of around 30 municipalities compared to the old sample. In total, 256 municipalities were selected from the take-all strata and 228 municipalities were selected from take-some strata across Canada.

Table 7: Sample Distribution by Province and Territory for the 2009 Design

Province	Take-All	Take-Some	Regional Municipalities & Other Entities	Total Sample Size
NF	33	19	1	53
PE	6	4	1	11
NS	5	6	1	12
NB	19	9	1	29
QC	60	71	2	133
ON	36	21	30	87
MB	16	13	1	30
SK	28	54	1	83
AB	15	16	1	32
BC	24	8	4	36
YK	1	2	1	4
NT	3	2	1	6
NU	10	3	0	13
CANADA	**256**	**228**	**45**	**529**

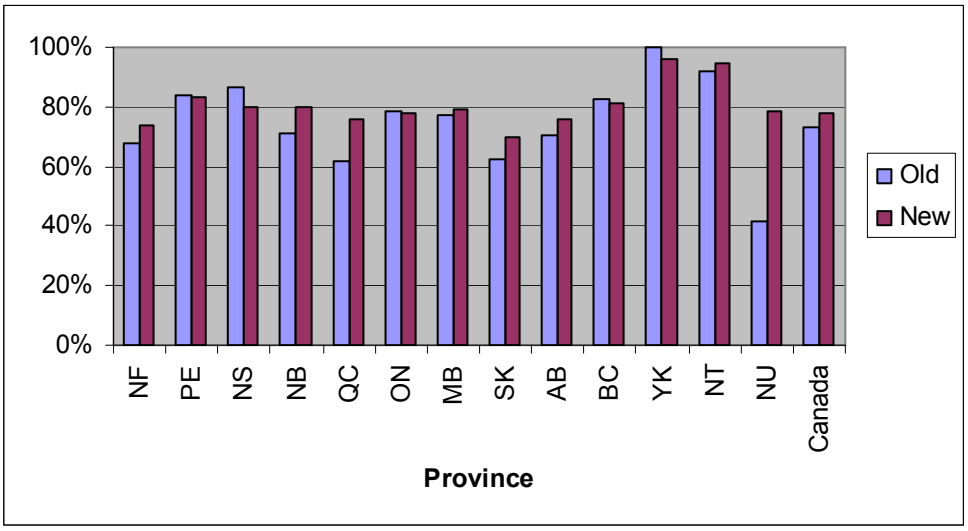

Figure 4: Coverage of Sample by Number of Residents – 1998 Design vs. 2009 Design

Looking at the coverage of the sample in terms of the number of residents in each municipality, Figure 4 shows that the coverage is more evenly distributed amongst provinces. With the new design, for each province the total number of residents of the municipalities in the sample is at least 73% of the total number of residents in the province. With the old design this percentage was as low as 42% in Nunavut.

At the time of writing this article, data collection for the new design was almost complete, though it was still too early to verify if the new design achieved its purpose of producing better quality data while reducing the response burden on smaller municipalities. A thorough assessment of the successes and shortcomings of the new design will take a couple of years to complete while future iterations of the LGS provide more evidence.

5. Conclusion and Future Work

After ten years of using the same sample for the LGS, it was more than time to proceed with a redesign. Compared to the previous design, the new one is simpler. It has a less detailed level of stratification and a more uniform way of establishing size groupings across the provinces. Given that a new sample will be selected every year, the new design will also greatly contribute to reducing response burden for the smaller municipalities. Furthermore, it will provide a more timely and improved coverage of the target population.

Given the decision to have only one take-some stratum per province, in the future we will consider the possibility of using the Hidiroglou (1986) method (the predecessor to the Lavallée-Hidiroglou algorithm) to determine the cut-off points between the take-some and take-all strata. We will also try to develop an estimation strategy for the bottom 2% of the population that is now excluded from the survey even though we know that the under-coverage resulting from those exclusions is minimal. More work will also be invested into refreshing the imputation strategy which has also been left mostly unchanged for the past ten years. Finally, future studies will assess the possibility of replacing survey data with administrative data given that the timing would be right. Supplementing survey data with administrative data, perhaps as part of the imputation or estimation strategy, is also a direction that will be considered.

Acknowledgements

The authors would like to thank Joseph Duggan, Marie-Claude Duval and Claude Girard for their valuable comments in improving the quality of this article.

References

Bankier, M. (1988). Power Allocations: Determining Sample Sizes for Subnational Areas. *The American Statistician*, August 1988, vol. 42, no. 3, pp. 174-177.

Lavallée, P. and Hidiroglou, M.A. (1988). On the Stratification of Skewed Populations. *Survey Methodology*, June 1988, vol. 14, no. 2, pp. 33-43.

Hidiroglou, M.A. (1986). The Construction of a Self-Representing Stratum of Large Units in Survey Design. *The American Statistician*, vol. 40, pp. 27-31.