SSC Annual Meeting, June 2017
Proceedings of the Survey Methods Section

Job Vacancy and Wage Survey: Balancing sampling and operational requirements using the Cube Method

Min Jiang[1]

ABSTRACT

Balanced sampling is a sampling method where the totals estimated with the Horvitz-Thompson estimator are the same or close to the true population totals for a given set of auxiliary variables on the survey frame. If the auxiliary variables are highly correlated to the variables of interest, the variances of the estimators for totals will be small. In Statistics Canada's Job Vacancy and Wage Survey (JVWS), the quarterly sample of business locations has to be split into three monthly subsamples for data collection. In parallel, all locations under the same enterprise must be collected the same month. In this paper, we will show how one can use the balanced sampling method to allocate the sampling units at the enterprise level in a way that keeps the number of employees and the number of locations balanced for each province and industry between months. This allocation strategy was implemented for the JVWS using the Cube Method which was proposed by Deville and Tillé (2004).

KEY WORDS: Balanced sampling; Cube Method; sample allocation

RÉSUMÉ

L'échantillonnage équilibré est une méthode d'échantillonnage selon laquelle les totaux estimés par l'estimateur de Horvitz-Thompson sont les mêmes ou près des vrais totaux de population pour un certain ensemble de variables auxiliaires de la base de sondage. Quand ces variables auxiliaires sont bien corrélées aux variables d'intérêt, la variance des estimateurs pour des totaux est faible. Dans l'Enquête sur les postes vacants et les salaires (EPVS) de Statistique Canada, l'échantillon trimestriel d'emplacements commerciaux doit être réparti en trois sous-échantillons mensuels pour la collecte des données. En parallèle, la collecte des emplacements d'une même entreprise doit se faire le même mois. Dans cet article, nous montrerons comment l'échantillonnage équilibré peut être utilisé pour répartir l'échantillon au niveau de l'entreprise de sorte à balancer le nombre d'employés et le nombre d'emplacements selon les provinces et les secteurs industriels entre les mois. Cette stratégie de répartition a été mise en production dans l'EPVS en utilisant la méthode du Cube proposée par Deville et Tillé (2004).

MOTS CLÉS : Échantillonnage équilibré; Méthode du Cube; répartition de l'échantillon

1. INTRODUCTION

In order to fill gaps in labour statistics data on job vacancies, the Job Vacancy and Wage Survey (JVWS) was launched in 2015 by Statistics Canada. The goal of the job vacancy component of the survey is to produce estimates of the number of job vacancies and vacancy rates by economic region, industry and detailed occupation. The annual wage component was launched in 2016 and the goal of this component is to produce average hourly wage and employment estimates by economic region, industry and detailed occupation. Together, the two components could be used to produce vacancy rates by occupation. Detailed information about JVWS can be found on the Statistics Canada website (http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5217).

The job vacancy component collects data from around 100,000 business locations across Canada that have at least two employees each quarter. Data is collected quarterly for all industries except for federal and provincial public administration, religious organizations and private households. The sample is stratified by economic region, industry and business size. To reduce the data collection burden, approximately 1/3 of the sample is collected each month. For operational and conceptual reasons, the wage component uses the same population and sample as the job vacancy component. However, the wage sample is collected yearly. For each quarter, around 1/4 of the job vacancy sample will be selected in the wage sample.

1 Min Jiang, Methodologist, Business Survey Methods Division, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON, K1A 0T6, Canada, Min.Jiang@canada.ca

In the next section, an overview of data collection for JVWS and challenges will be presented. In the third section, the balanced sampling and Cube Method will be introduced. In the fourth section, an application for sample allocation will be introduced. In the fifth section, a comparison between two different methods will be given. A conclusion will be given in the final section.

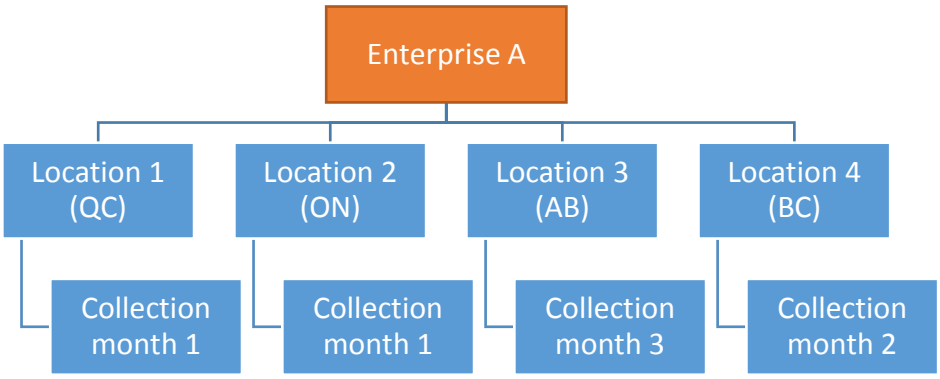
2. DATA COLLECTION FOR JVWS AND CHALLENGES

The job vacancy (JV) data is collected over three months of each quarter to balance the workload for collection. Each unit in the JV sample is assigned a random number in {1, 2, 3} that determines in which month of the quarter it will go to collection. All the units with random number equal to 1 are collected in the 1st month while the sample units with random number equal to 2 are collected in the 2nd month. Finally, all the units with random number equal to 3 are collected in the 3rd month.

Before presenting the challenges for JVWS data collection, some simple definitions will be introduced for the terms location and enterprise. The enterprise is an autonomous unit for which a complete set of financial statements is available. The enterprise, as a statistical unit, is defined as a business unit that directs and controls the allocation of resources relating to its operations, and for which consolidated financial information is maintained. An enterprise is at the top of the operating structure. It can consist of one or more locations. The location, as a statistical unit, is defined as a producing unit at a single geographical location at which or from which economic activity is conducted and for which, at a minimum, employment data are available.

From 2015 quarter 1 to 2016 quarter 4, the collection month was assigned to each sample unit at the location level. This was done randomly and independently. The sampling units are at the location level, and the target respondents are contacted at the location level as well. If the information can be provided by the location, there is no special collection burden. The non-response follow-ups are also done at the location level in this case. However, there are some locations that are unable or unwilling to report for themselves, and the contact person ends up being at the enterprise level. In this case, this could introduce extra response burden for complex enterprises since enterprises with multiple locations might be collected multiple times each quarter. For example, around 3,000 enterprises had been collected for more than one month, and some of them in each month of 2016 quarter 4. The collection team already received some special requests from respondents so that they could report all the units within the same enterprise in the same month. To illustrate the complexity of data collection for complex enterprises, here is a simple example of a complex enterprise that has data being collected each month.

Figure 1: Example of a complex enterprise with data collection in multiple months



In Figure 1, this enterprise has four locations and has to respond three times in one quarter, which is burdensome. It is also burdensome to do non-response follow-up for this enterprise since it will be contacted three times during that quarter if it does not respond.

In order to solve such problems, we propose to assign the collection month at the enterprise level for each sampling unit so that all units within the same enterprise only need to respond once each quarter. At the same time, one also needs to ensure that the workload for collection is approximately equal and that the employment totals of each province and industry are similar for each month in order to minimize the seasonal effect and get correct estimates. The quarterly estimates are technically averages over the three months. If the sample wasn't balanced, we would need to weight the months differently.

For the wage component, we are also attempting to balance the sample in terms of location counts and employment, at the provincial and industrial levels for each quarter of the year. Here is a simple example to illustrate the collection process for both the JV and wage components of JVWS. For example, if a unit has collection month equal to 3 and collection quarter equal to 2, then the unit will go to collection for job vacancy component in the third month of each quarter (March, June, September, December) and go to collection for wage component in the third month of the second quarter (June). In order to avoid that the same enterprise might respond to the wage questionnaire multiple times during different quarters, one also needs to assign the collection quarter at the enterprise level. The same strategy will be applied to assign the collection month for the job vacancy component and the collection quarter for the wage component.

3. BALANCED SAMPLING AND THE CUBE METHOD

A sample is said to be balanced with respect to the auxiliary variable $\boldsymbol{x} = (x_1, \dots, x_Q)$ if the following equation holds:

$$\widehat{\boldsymbol{x}}_{\boldsymbol{\pi}} = \sum_{i \in s} \frac{\boldsymbol{x}_i}{\pi_i} = \sum_{i \in U} \boldsymbol{x}_i, \tag{1}$$

where π_i is the selection probability of population unit *i*; *s* is the selected sample; *U* is the population. The variables $x_1, \dots, x_Q$ are called balancing variables, and they are usually frame variables such as employment or population counts. $\widehat{\boldsymbol{x}}_{\boldsymbol{\pi}}$ is the Horvitz-Thompson estimator of $\boldsymbol{x}$. Equation (1) is a form of calibration at the sampling stage, which means that the Horvitz-Thompson estimate of $\boldsymbol{x}$ will always be equal to its population total for all balanced samples *s*. If the variable of interest *y* is strongly correlated with $\boldsymbol{x}$, then final estimates of *y* will be very efficient if the sample *s* is balanced.

There are many methods that can yield a balanced sample, such as ordered systematic sampling, stratified sampling (see Kott 1986), complete enumeration and rejective method. However, these methods all suffer from different kinds of constraints. For example, the inclusion probability for the rejective method is difficult to calculate (see Fuller 2009). The complete enumeration method cannot be applied for large size populations. The Cube Method which was proposed by Deville and Tillé (2004) leads to the selection of approximately balanced samples such that the inclusion probabilities $\boldsymbol{\pi} = (\pi_1, \dots, \pi_N)^T$ are exactly satisfied. A sample can be represented by a vector of sample selection indicators $\boldsymbol{I} = (I_1, \dots, I_N)^T$, where $I_i = 1$ if $I_i \in s$ and 0 otherwise. Let matrix $\boldsymbol{A}$ be defined as

$$\boldsymbol{A} = \left(\frac{\boldsymbol{x_1}}{\pi_1}, \dots, \frac{\boldsymbol{x_N}}{\pi_N}\right).$$

Then the balancing constraint $\widehat{\boldsymbol{X}}_{\boldsymbol{\pi}} = \boldsymbol{X}$ can be re-written as

$$\boldsymbol{A}(\boldsymbol{I} - \boldsymbol{\pi}) = 0, \tag{2}$$

since

$$\widehat{\boldsymbol{X}}_{\boldsymbol{\pi}} = \sum_{i \in s} \frac{\boldsymbol{x}_i}{\pi_i} = \sum_{i \in U} \frac{\boldsymbol{x}_i}{\pi_i} I_i = \boldsymbol{AI} \tag{3}$$

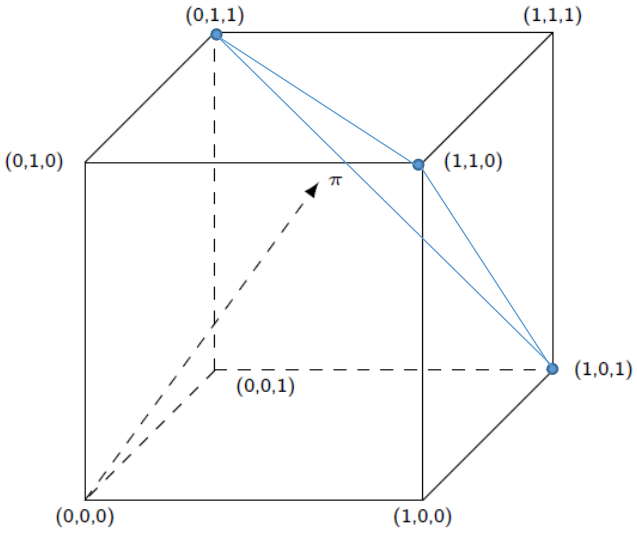
and

$$\boldsymbol{X} = \sum_{i \in U} \boldsymbol{x}_i = \sum_{i \in U} \frac{\boldsymbol{x}_i}{\pi_i} \pi_i = \boldsymbol{A\pi}. \tag{4}$$

The Cube Method tries to find a sample $\boldsymbol{I} = \left(I_1, \dots, I_N\right)^T$ such that $\boldsymbol{E}(\boldsymbol{I}) = \boldsymbol{\pi}$ which means that inclusion probabilities are exactly satisfied while at the same time also satisfying the balancing constraint defined by equation (2).

In the following example, there are eight possible samples in total for the population size of 3. A sample can be viewed as a vertex of a cube in Figure 2 which comes from Haziza and Bocci (2014). If one is interested in drawing a sample with sample size equal to exactly 2, the Cube Method consists of a stepwise transformation of the selection probability vector into one of the three vectors corresponding to the vertices of the cubes lying on the plane $I_1 + I_2 + I_3 = 2$, in such a way that $E(I_i) = \pi_i$ at each step.

Figure 2: Possible samples in a population of size N=3 with sample size n=2



The Cube Method consists of two phases: the flight phase and the landing phase. The idea of the flight phase is to start from the selection probability vector $\boldsymbol{\pi} = (\pi_1, \dots, \pi_N)^T$. After T steps, it will be transformed into the sample vector $\boldsymbol{I} = (I_1, \dots, I_N)^T$ while at the same time satisfying the balance constraints. The flight step is a random walk in the null space $N(\boldsymbol{A})$ of matrix A. Here is detailed description of the flight phase algorithm from Haziza and Bocci (2014).

Cube Method-flight phase

- Start with $\boldsymbol{\pi}(\boldsymbol{0}) = \boldsymbol{\pi}$;
- In step $t,\ t = 2, \dots, T,$
  - Generate a vector $\boldsymbol{\mu}(t) \in N(\boldsymbol{A})$ such that $\boldsymbol{A\mu}(t) = \boldsymbol{0}$
  - Generate

$$\boldsymbol{\pi}(t) = \begin{cases} \boldsymbol{\pi}(t-1) + k_1\boldsymbol{\mu}(t) & with\ prob\ \ k_2/(k_1+k_2) \\ \boldsymbol{\pi}(t-1) - k_2\boldsymbol{\mu}(t) & with\ prob\ \ k_1/(k_1+k_2) \end{cases}$$

    where k_1 and k_2 are positive constants
- After T steps, the vector $\boldsymbol{\pi}$ will be transformed into a sample vector $\boldsymbol{I}$.

At each step of the algorithm, we have

$$\begin{aligned} E[\boldsymbol{\pi}(t)|\boldsymbol{\pi}(t-1)] &= [\boldsymbol{\pi}(t-1) + k_1\boldsymbol{\mu}(t)] \times \frac{k_2}{k_1+k_2} + [\boldsymbol{\pi}(t-1) - k_2\boldsymbol{\mu}(t)] \times \frac{k_1}{k_1+k_2} \\ &= \boldsymbol{\pi}(t-1) \end{aligned}$$

Applying this equation recursively, we have

$$E[\boldsymbol{\pi}(t)] = E[E[\boldsymbol{\pi}(t)|\boldsymbol{\pi}(t-1)]\,] = E[\boldsymbol{\pi}(t-1)] = \cdots = E[\boldsymbol{\pi}(0)] = \boldsymbol{\pi}$$

So, for each step of the algorithm, the inclusion probabilities $\boldsymbol{\pi}$ are exactly satisfied. If a solution exists after T steps which means that $\boldsymbol{\pi}(T)$ is successfully transformed to a sample vector $\boldsymbol{I}$, the balancing constraint defined by (2) should also be automatically satisfied. This is because:

$$\begin{aligned} \boldsymbol{A\pi}(t) &= \begin{cases} \boldsymbol{A\pi}(t-1) + k_1\boldsymbol{A\mu}(t) & with\ prob\ \ k_2/(k_1+k_2) \\ \boldsymbol{A\pi}(t-1) - k_2\boldsymbol{A\mu}(t) & with\ prob\ \ k_1/(k_1+k_2) \end{cases} \\ &= \boldsymbol{A}\,\boldsymbol{\pi}(t-1) \end{aligned}$$

since $\boldsymbol{A\mu}(t) = \boldsymbol{0}$ for any t. Thus,

$$\widehat{\boldsymbol{X}}_{\boldsymbol{\pi}} = \boldsymbol{AI} = \boldsymbol{A}\,\boldsymbol{\pi}(T) = \boldsymbol{A}\,\boldsymbol{\pi}(T-1) = \boldsymbol{A}\,\boldsymbol{\pi}(0) = \boldsymbol{A}\,\boldsymbol{\pi} = \boldsymbol{X}$$

However, in most cases, an exact balanced sample does not exist. The goal of the landing phase is to end the selection such that the inclusion probabilities are exactly satisfied while the balancing constraints are approximately satisfied. One option is to relax the balancing constraints by dropping them one by one.

4. APPLICATION FOR SAMPLE ALLOCATION

In this section, the application of the balanced sampling for JVWS sample allocation by using the Cube Method will be shown. For the job vacancy component, allocating the sample units into three months for data collection is equivalent to selecting three mutually exclusive subsamples with inclusion probability equal to 1/3 for each month. For each enterprise in the sample, the total weighted number of locations and weighted employment for each province and for each industry are calculated. These are the auxiliary variable $\boldsymbol{x}$ in the description of balanced sampling above. Each enterprise has selection probability equal to 1/3. The selected sample is balanced on the number of locations and employment by province and industry. The first selected sample will be collected in the first month of each quarter. One half of the remaining sample will be selected and allocated to the second collection month. The rest of the sample will be assigned to the third collection month.

Subject matter specialists also provided a list with a few groups of enterprises that must be kept together for collection. For these units, we need to randomly assign the collection month and quarter at the group level by using Bernoulli sampling. Then, all enterprises under the same group will get the same collection month and quarter. After that, one needs to adjust the population total and exclude these units before applying the Cube Method to balance the allocation for the rest of the sample units.

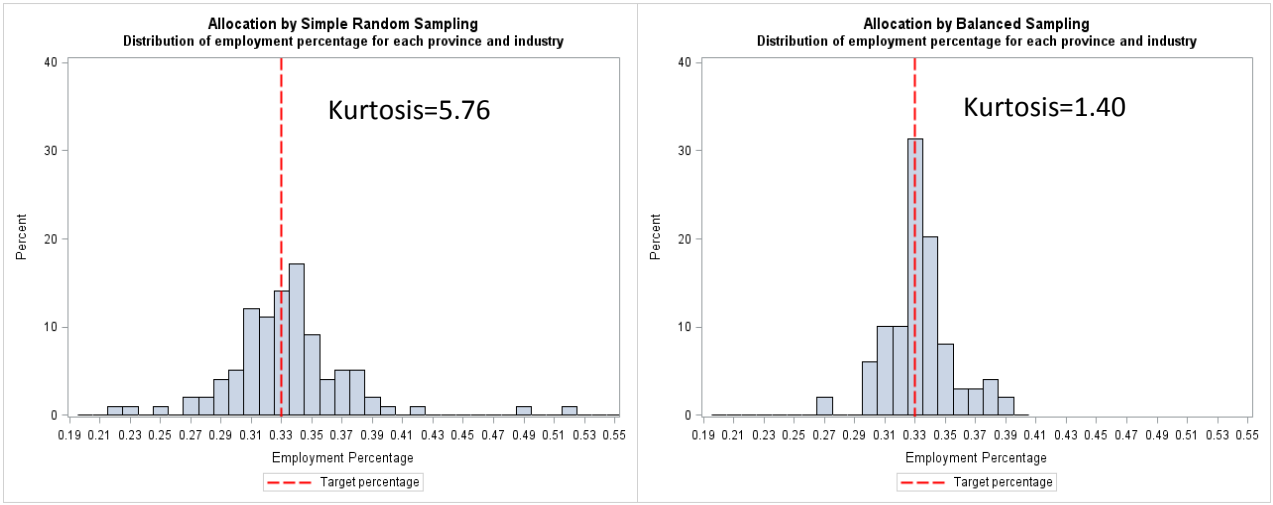
For the wage component, the allocation strategy is similar. Allocating the sample units into four quarters for data collection is equivalent to selecting four mutually exclusive subsamples with inclusion probabilities equal to 1/4 for each month. The wage sample allocation is done independently of the job vacancy allocation. Each enterprise now has a selection probability equal to 1/4. The selected sample is also balanced on the number of locations and employment by province and industry. The first selected sample will be collected in the first quarter of each year. 1/3 of the remaining sample will be selected and allocated to the second collection quarter. 1/2 of the remaining sample will be selected and allocated to the third collection quarter. Finally, the rest of the sample will be assigned to the fourth collection quarter. All the collection month and quarter variables for each enterprise selected in the sample are saved in a permanent table. This will ensure that the enterprises selected in the next cycle will always have the same collection month and quarter as in the previous cycle. This is not a theoretical necessity, but a practical one to avoid confusing respondents.

Starting in 2017 quarter 1, in order to maintain a good balance of quality between trend estimates and cross-sectional estimates, a rotation scheme of 1/8 was chosen. Therefore once selected, a business location will then remain in the sample for eight quarters (or two years) before rotating out, except for business locations in take-all strata. One also needs to take the sample rotation into consideration when allocating the new units in the sample. New locations belonging to an enterprise that had locations sampled in a previous cycle are assigned that enterprise's collection month and quarter. For locations belonging to an enterprise that never had locations in sample before, a new collection month and quarter will be assigned independently by using the same strategy described in the previous paragraphs. This means that the allocation of collection month and quarter for the new enterprises does not depend on the collection month and quarter assigned to the enterprises that were previously in sample.

5. COMPARISON AND RESULTS

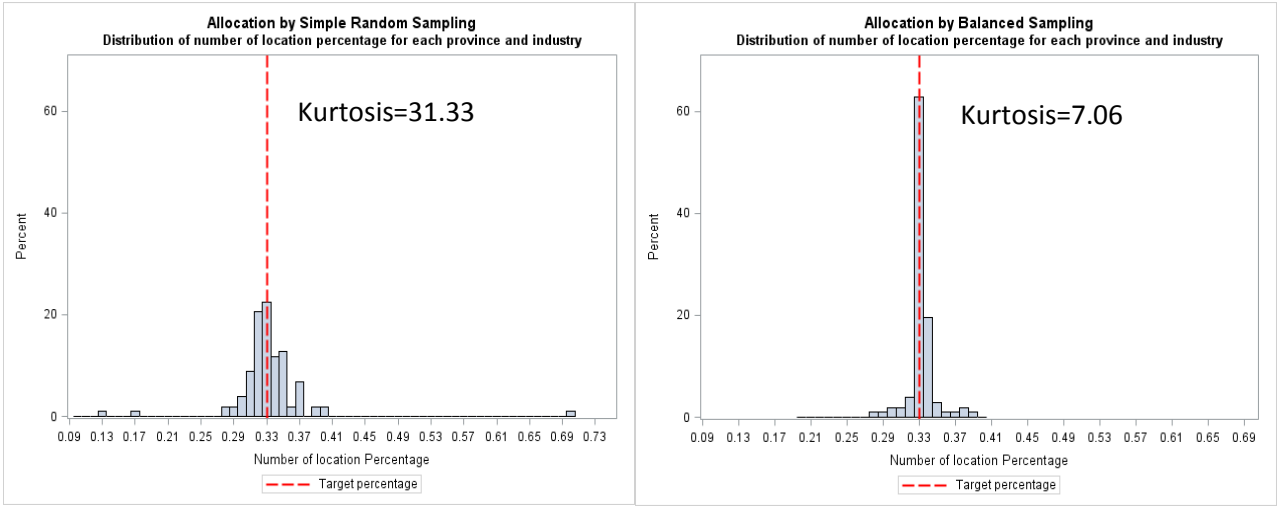
In this section, we compare the allocation results between simple random sampling (SRS) and the proposed balanced sampling by Cube Method for the job vacancy sample from 2017 quarter 1. In Figure 3, the employment percentage variable was defined as the total weighted employment for each month over the total employment in the population. The data points in Figure 3 are the employment percentages for each province and industry. Ideally, if the sample was well balanced for employment, all the employment percentages should be close to 1/3 for each province and industry, which is the red line in the histograms. The left panel of Figure 3 is the histogram of employment percentage obtained by using the SRS method. The right panel of Figure 3 is the histogram of employment percentage obtained by using the Cube Method. By comparing these two histograms in Figure 3, it is clear that the allocation by Cube Method outperformed the simple random sampling method since the employment percentage is more concentrated around the red line.

Figure 3: Percentage of employment per month for 2017 quarter 1



Similarly, the number of location percentage was defined as total weighted number of locations for each month over the total number of locations in the population. The data points in Figure 4 is the percentage of number of locations for each province and industry. Thus, if the sample was well balanced, all the percentages should be close to 1/3 for each province and industry, which is the red line in the histograms. The left panel of Figure 4 is the histogram of the percentage obtained by using the SRS method while the right panel of Figure 4 is the histogram of the percentage obtained by using the Cube Method. Similar conclusions can be made by comparing these two histograms in Figure 4. The allocation by Cube Method performed better than the simple random sampling method again.

Figure 4: Percentage of number of locations per month for 2017 quarter 1



6. CONCLUSION

By using the balanced sampling method, we are able to achieve many goals at the same time: all locations under the same enterprise are collected in the same month for the job vacancy component and in the same quarter for the wage component; the number of locations is balanced for each province and industry between the collection month and quarter; the number of employees is balanced for each province and industry between the collection month and quarter. In addition, allocation using balanced sampling performs better for most provinces or industries in terms of achieving an equal representation of the sample for each collection month and quarter.

ACKNOWLEDGEMENTS

The author would like to thank Danielle Lebrasseur, Etienne Rassart, Bertrand Ouellet-Léveillé and Leon Jang for their insightful comments. He would also like to thank Marc St-Denis and Wesley Yung for their support in making this project possible.

REFERENCES

Deville, J.-C. and Tillé Y. (2004). “Efficient balanced sampling: The cube method”. *Biometrika,* **91**,*4,* 893-912.

Fuller¸W.A. (2009). “Some design properties of a rejective sampling procedure”. *Biometrika,* **96,** 933-944.

Haziza, D. and Bocci, C. (2014). *Calibration and Balanced Sampling in Survey*. Statistics Canada internal training material.

Kott, P. S. (1986). "Some asymptotic results for the systematic and stratified sampling of a finite population." *Biometrika* 73, 2, 485-491.

Statistics Canada, Job Vacancy and Wage Survey.
http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5217