Producing labour statistics by detailed geography and occupation: Experiences from the new Canadian Job Vacancy and Wage Survey

Etienne Rassart[1], Zdenek Patak[1]
[1] Senior Methodologists, Business Survey Statistics Division, Statistics Canada,
150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada K1A 0T6
{etienne.rassart, zdenek.patak}@canada.ca

Abstract

In order to fill gaps in labour statistics data on job vacancies, the Canadian Job Vacancy and Wage Survey was launched in July 2014. The goal of the job vacancy component is to produce quarterly estimates of the number of job vacancies by economic region and occupation, while the goal of the annual wage component is to produce average hourly wage and employment estimates by economic region and occupation. Together, the two components could be used to produce vacancy rates by occupation. The components are integrated and use quarterly samples of 100,000 business locations in Canada on a rotating basis, making the survey one of the largest business surveys at Statistics Canada. This paper highlights methodological issues and solutions related to producing statistics at such detailed levels of geography and occupation, while managing respondent burden. One challenge, common to both components, was defining a probabilistic non-response follow-up strategy prioritizing large locations. For the wage component, the main challenge was generating a sample of relevant occupations used to collect detailed wage information from each location. The solutions used sequential Poisson sampling and Pareto sampling.

Key Words: Labour statistics, business survey methodology, respondent burden, job vacancy and wage by occupation, sequential Poisson sampling, Pareto sampling

1. Introduction

In order to fill gaps in labour statistics data on job vacancies, the Job Vacancy and Wage Survey (JVWS) was launched in July 2014 (Statistics Canada, JVWS). The goal of the quarterly job vacancy component of the survey is to produce estimates of the number of job vacancies and vacancy rates by economic region and detailed occupation, while the goal of the annual wage component is to produce average hourly wage and employment estimates, also by economic region and detailed occupation. Together, the two components could be used to produce vacancy rates by occupation. Currently, the ratio of unemployed people from Statistics Canada's Labour Force Survey (LFS) to the number of job vacancies is produced by the agency's Survey of Employment, Payroll and Hours (SEPH). However, because SEPH does not measure occupation, the ratio is only available by industry.

The components of the survey are integrated and use quarterly samples of 100,000 business locations in Canada on a rotating basis, making the survey one of the largest business surveys at Statistics Canada.

In the remainder of this section, an overview of the sampling design, the collection instrument (questionnaire) and the collection process will be presented. In the second section, two methodological challenges encountered during the development of the survey will be introduced. The common thread between the two challenges resides in the similar solutions that they share.

1.1 Overview of the sampling design

Canada is made up of 76 economic regions, sub-provincial partitions which are meant to be the standard geographic units for the analysis of economic activity. The survey uses the North American Industrial Classification System (NAICS) (Statistics Canada, NAICS 2012) to associate industrial groups to business locations, and the National Occupation Classification (NOC) (Statistics Canada, NOC 2011) to associate occupational codes to vacant positions. In the 2011 NOC system, there are 500 detailed occupations.

The JVWS population consists of business locations with employees in all economic regions of Canada and all industrial sectors, including agriculture, fishing, and forestry, with the exception of the federal and provincial administration subsectors. These subsectors will be phased in at a later date. The JVWS population is extracted from Statistics Canada's Business Register.

The population is stratified by industry (NAICS-2 or industrial sector), geography (economic region), and employment (up to four size classes). The fact that certain occupations are prevalent in small industries justifies the inclusion of industry among the stratification variables. Because of the disparity in the size of the Canadian economic regions, a power allocation is used, with the number of employees as the auxiliary variable. This method is known for providing a good balance of quality in the estimates between small and large domains (for example, national vs. provincial). Stratified Bernoulli sampling is used to select the sample. In order to maintain a good balance of quality between trend estimates and cross-sectional estimates, a rotation scheme of 1/8 was implemented. In other words, once selected, a business location will remain in the sample for eight quarters (or two years) before rotating out, except for business locations in take-all strata (certainty units).

1.2 Collection instrument

A vacant position is defined as a position for which a business is actively recruiting externally and planning to fill within the next 30 days. The job vacancy event is a somewhat rare event: based on data from other Statistics Canada surveys, the proportion of businesses with at least one vacant position at any given time ranges from 15% for small businesses to 30% for large businesses.

The questionnaire is short and is based on a model developed in the United States by the National Job Vacancy Survey Workgroup and the state of Minnesota. The questionnaire first measures employment and the total number of job vacancies (if none, the questionnaire is finished). Then detailed information is collected for each occupation with vacancies: job title; the number of job vacancies; a short job description (to help coders assign a NOC-4 code to the occupations); number of full-time and part-time job vacancies; number of permanent, temporary, and seasonal job vacancies; offered salary or wage; amount of time the vacant positions have been open; minimum level of education sought; professional accreditation requirements; and methods used to fill the job vacancies.

Electronic questionnaires are the only mode of collection for the survey. A paper questionnaire is made available to respondents who refuse to fill out the electronic questionnaire. Telephone interviews (CATI) are also used for non-response and failed edits follow-up.

1.3 Collection process

For operational and conceptual reasons, the two components of the survey use the same population and sample. In order to balance the measure of job vacancies, the response burden, and the collection operations over the quarter and the year, each sampled unit is assigned to a month of the quarter for the job vacancy component and to a quarter of the year for the wage component. For each month of the quarter, as well as each quarter of the year, an attempt is made to balance the sample in terms of location counts and employment, at the provincial and industrial sector levels. Sampled business locations are asked to provide job vacancy and wage information as of the first day of the month to which they are assigned. Table 1 gives a pictorial representation of the collection process.

Table 1: Collection process for the two components of JVWS

				Q1			Q2			Q3			Q4		
				1	2	3	4	5	6	7	8	9	10	11	12
Quarter assigned	1	Month assigned	1	W											
			2		W										
			3			W									
	2		1				W								
			2					W							
			3						W						
	3		1							W					
			2								W				
			3									W			
	4		1										W		
			2											W	
			3												W

The shaded cells correspond to job vacancy component collection and the cells marked "W" to the wage component collection. In other words, while all locations are collected for the job vacancy component each quarter, only one fourth will be collected for the wage component.

2. Methodological challenges for JVWS

In the course of the development of JVWS, many methodological challenges were encountered. Two were retained for this paper because their solutions share a common thread. Both use fixed sample size variations of Poisson sampling (conditional Poisson sampling) with selection probabilities proportional to size (PPS or πPS). For the survey's

non-response follow-up strategy, sequential Poisson sampling is used, while samples of occupations are selected for the wage component of the survey using Pareto sampling.

2.1 Non-response follow-up

Non-response follow-up (NRFU) by the collection offices for both components of JVWS consists of email reminders, followed by phone calls by interviewers. While email reminders are essentially cost-free, phone calls by interviewers must be prioritized so as to maximize the use of resources.

2.1.1 The challenge for JVWS

In surveys where total (or unit) non-response would be treated through the use of imputation, the normal strategy is to prioritize the larger units. These are surveys for which auxiliary or historical information is available and for which good (hopefully unbiased) imputation models can be constructed. Unfortunately, for JVWS there is very little information on the Business Register that correlates well to job vacancies. Employment correlates best, but the relationship is moderate only. For the time being, the survey also lacks historical information – it is unknown at this point whether the job vacancy counts and the distributions of job vacancies by occupation evolve in time in predictable ways at the business location level or even aggregate levels. As a result, total non-response for the survey is dealt with through non-response adjustments, i.e. the weights of the respondents are adjusted to account for the non-responding units.

A NRFU strategy prioritizing the large business locations only would introduce bias. On the other hand, selecting a subsample of non-respondents for NRFU with equal selection probabilities would be an inefficient use of resources. So a probabilistic way of selecting a sample of non-respondents for NRFU that advantages larger business locations was sought.

2.1.2 Probability proportional to size sampling and sequential Poisson sampling

Sequential Poisson sampling was developed by Ohlsson for the Swedish Consumer Price Index (Ohlsson, 1998). This selection method is meant to provide a fixed sample size modification of Poisson sampling, in particular as a way to draw a probability proportional to size (PPS) sample.

For PPS sampling, each unit has a normalized positive size measure p_i (the notation from Ohlsson's paper is used here). The p_i must sum up to 1, so typically $p_i = X_i/X$ for some positive auxiliary size variable X_i, where X is the sum of the X_i over the population. If a sample of (expected) size n is to be selected, the selection probability of unit i is set to $\pi_i = np_i$. If $np_i > 1$ for some units, they are moved to a take-all stratum, then n and the p_i are recomputed; the process goes on iteratively. A Poisson sample can then be selected by assigning to each unit i a permanent random number u_i and including in the sample all the units for which $u_i < \pi_i = np_i$. Equivalently, the sampled units are the ones with $u_i/p_i = u_iX/X_i < n$.

To select a sequential Poisson sample of size n with probability proportional to size, each unit i is assigned a permanent random number u_i. The population units are ordered by increasing u_i/p_i and the first n units in this ordering form the sample. Ohlsson showed that the estimators associated with Poisson and sequential Poisson sampling are asymptotically normally distributed, asymptotically unbiased, and equally efficient. Note that since X is constant, ordering by increasing $u_i/p_i = u_iX/X_i$ is equivalent to ordering by increasing u_i/X_i.

2.1.3 The advantages of sequential Poisson sampling for JVWS

This last observation is the important one for the purposes of the JVWS NRFU because it means that the size n of the NRFU subsample does not need to be specified in advance. The ordering can be used to assign each unit in the whole sample a rank for potential NRFU. The NRFU subsample size is determined a posteriori as the stopping rank. The responding units are simply skipped over during NRFU activities.

Because sequential Poisson sampling approximately yields PPS selection probabilities, the NRFU strategy favors the locations with the largest number of employees and therefore, because the number of employees is positively correlated with the number of vacant positions, the locations most likely to contribute significantly to the job vacancy totals.

In practice, to make sure that the largest units in the sample benefit from NRFU, units above a certain threshold on the number of employees (currently 500) are all assigned to a take-all stratum for NRFU. Of about 33,000 units in each monthly sample for the job vacancy component, these priority cases (certainty units) only represent 800 units.

2.2 Selection of occupations for the wage component

Because job vacancies are somewhat rare events, business locations are asked to provide an exhaustive list of their job vacancies. For the wage component, small business locations (currently 10 employees or less) are asked to provide employment and wage information for all their occupations. However, even mid-size business locations may have a wide spectrum of occupations, which makes the wage component more burdensome. As a result, larger business locations (more than 10 employees) are prompted for detailed wage information for only a subset of their occupations.

In practice, each sampled business location receives a list of 10 occupations, for which they are asked to provide detailed wage information. This number of occupations was assessed through questionnaire testing by Statistics Canada's Questionnaire Design Resource Center. If the business location has more than 10 employees, they are asked to provide wage information for only those occupations given on the list. Business locations with 10 or fewer employees are asked to provide wage data for any occupation that they may have that is not part of the list (i.e. they are asked to provide a census of occupations). By using a list of 10 occupations even for the smaller business locations, the amount of manual coding of occupations is reduced. Questionnaire testing also showed that respondents prefer lists to providing job titles and descriptions, so part of their work is simplified.

2.2.1 Which occupations to select?

The question becomes how to select the 10 occupations to be presented on the list. These lists or samples of occupations form the second stage of the survey design for the wage component, the first stage being the selection of the business locations (shared with the job vacancy component). Ideally, the lists should be tailor-made for each business location, so that the respondents recognize as many occupations as possible. However, there are no sources of administrative data available to Statistics Canada that provides an exhaustive list of occupations for each location. Since the business locations' industries are available on the Business Register, the best proxy comes in the form of lists of occupations by industry, obtained from Statistics Canada's Labour Force Survey (Statistics Canada, LFS).

Respondents to the LFS are asked to provide both a description of their occupation and, if employed, of their employer's industry. From the LFS data, a distribution of occupations at the 4-digit NOC level can be obtained for each 4-digit NAICS industry. JVWS uses the last 90 months of LFS data to derive these distributions. Occupations not found in a specific industry over this period are considered non-existent in that industry.

In order to obtain unbiased estimates, each occupation should have a positive selection probability. Selecting 10 occupations with equal probabilities would not be efficient since some occupations are more prevalent than others within an industry. This suggests selecting occupations for a location in a certain industry with probabilities proportional to the prevalence rates of the occupations in that industry. Using straight prevalence rates has disadvantages, though: very prevalent occupations (e.g. cashiers in the retail industry) will appear on the lists too often, at the expense of less prevalent occupations. To mitigate this, the square root of the prevalence rates are used as the size measure. This is akin to the choice of a power allocation for the first stage of the survey design, to yield a good balance of quality between national and regional job vacancy and employment estimates. (With the square root transformation, the occupations remain in the same order of prevalence rates, but the distribution of the transformed prevalence rates falls less sharply. As a result, rare occupations have less extreme selection probabilities and weights.)

2.2.2 Pareto sampling

Like sequential Poisson sampling, Pareto sampling is a fixed sample size variation on probability proportional to size (PPS) sampling. It was introduced by Rosén in 1997 (Rosén, 1997a). The two methods are similar: while for sequential Poisson sampling the n units with the smallest values of u_i/X_i (or u_i/p_i) form the sample, for Pareto sample it is the n units with the smallest values of

$$\frac{u_i/(1-u_i)}{\pi_i/(1-\pi_i)}$$

where the π_i are the target PPS selection probabilities as before. Rosén showed that Pareto sampling is the most efficient design among many similar order sampling designs (Rosén, 1997a, 1997b, 2000a, 2000b) (Bondesson et al., 2006) (Aires, 2000).

2.2.3 The advantages of Pareto sampling for JVWS

In the same way that sequential Poisson sampling yields a probabilistic way of organizing the non-response follow-up strategy in a manner that prioritizes the larger business locations, Pareto sampling – or PPS sampling in general – gives more prevalent occupations in an industry a higher probability of being selected in the lists of occupations for business locations in that industry. Pareto sampling has the additional advantage of producing lists of the same length, an advantage both in terms of variance and in terms of electronic questionnaire design.

By using the square root of the prevalence rates by industry instead of the prevalence rates themselves, the most prevalent occupations are not oversampled and the number of less prevalent occupations for which estimates of reasonable quality can be produced is increased.

3. Conclusion

Poisson sampling has many advantages due to the fact that the sample units are selected independently. The estimation of the variance of linear estimators is simplified since

knowledge of the distribution of the joint selection probabilities is not required. The additional variance introduced by the variability of the sample size can be lessened through the use of calibration or by conditioning on a fixed sample size. This latter approach, conditional Poisson sampling, provides an interesting way of producing fixed size samples with selection probabilities very close to the PPS selection probabilities. The designs used to meet the two challenges described above are both examples of conditional Poisson sampling with probabilities proportional to measure of size – business size for the non-response follow-up strategy and prevalence rate by industry for the selection of occupations.

Because the sample size is known in advance for the selection of occupations – each list contains 10 occupations – the use of the more efficient Pareto sampling for the second stage of the JVWS sampling design was implemented. Pareto sampling has stronger theoretical foundations than sequential Poisson sampling, for which mostly only simulation results exist. Preliminary results show that the selection method for occupations yields occupation prevalence rates coherent with LFS for common occupations (at the 4-digit NOC level), as well as a reasonable spectrum of occupations by economic region (for reasonably sized economic regions).

For the survey's non-response follow-up strategy, on the other hand, sequential Poisson sampling proved extremely helpful as the fact that the sample size need not be known in advance permitted a ranking of the sample units for non-response follow-up. The method is currently being used in the field to prioritize cases.

Acknowledgements

We would like to thank Pierre Lavallée, Jean-François Beaumont, and J.N.K. Rao for bringing sequential Poisson sampling and Pareto sampling to our attention. We would also like to thank Statistics Canada's Advisory Committee on Statistical Methods for helpful recommendations on both components of the survey.

References

Aires, N. 2000. Comparisons Between Conditional Poisson Sampling and Pareto Sampling Designs. *ICES II.*

Bondesson, L., I. Traat and A. Lundqvist. 2006. Pareto Sampling versus Sampford and Conditional Poisson Sampling. *Scandinavian Journal of Statistics, 33*(4). 699-720.

Ohlsson, E. 1998. Sequential Poisson Sampling. *Journal of Official Statistics, 14*(2). 149-162.

Rosén, B. 1997a. Asymptotic theory for order sampling. *J. Statist. Plann. Inference, 62*. 135-158.

Rosén, B. 1997b. On sampling with probability proportional to size. *J. Statist. Plann. Inference, 62*. 159-191.

Rosén, B. 2000a. On inclusion probabilities for order πps sampling. *J. Statist. Plann. Inference, 90*. 117-143.

Rosén, B. 2000b. A user's guide to Pareto πps sampling. *ICES II.*

Statistics Canada. *North American Industry Classification System (NAICS) 2012*. Retrieved from http://www.statcan.gc.ca/eng/subjects/standard/naics/2012/introduction

Statistics Canada. *National Occupation Classification (NOC) 2011*. Retrieved from http://www.statcan.gc.ca/eng/subjects/standard/noc/2011/introduction

Statistics Canada. *Job Vacancy and Wage Survey (JVWS)*. Retrieved from http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5217

Statistics Canada. *Labour Force Survey (LFS)*. Retrieved from http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3701