

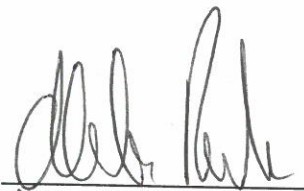
ASML Stratification

Submitted to
UBC Science Co-op Program
The University of British Columbia
Vancouver, BC

By
Yifan Zhao
Student ID: 49510150

The contents of this report is non-confidential and can be released to UBC

Signature of Supervisor:



Date:

Dec 20th / 2018

Signature of Director:



Date:

2019-01-02

Disclaimer: The contents of this report reflect the views of the author and not necessarily the official views or opinions of Statistics Canada.

ASML Stratification

Submitted to
UBC Science Co-op Program
The University of British Columbia
Vancouver, BC

By
Yifan Zhao
Student ID: 49510150

The contents of this report is non-confidential and can be released to UBC

Signature of Supervisor: _____

Date: _____

Signature of Director: _____

Date: _____

Disclaimer: The contents of this report reflect the views of the author and not necessarily the official views or opinions of Statistics Canada.

ASML Stratification

ZHAO, YIFAN

ECONOMIC STATISTICS METHODS DIVISION

DIVISION DES METHODES DE LA STATISTIQUE ECONOMIQUE

Summary

The Annual Survey of Manufacturing and Logging is one of Statistics Canada's annual economic surveys. It covers 253 industries and this large number makes the sample design more complicated than for other economic surveys. For the purpose of stratified sampling, these industries are grouped into industry classes and domains are defined as industry classes crossed by provinces. In the past, industry classes were defined at an aggregate level, which guaranteed a solution to the allocation problem. However, the data were published at a more detailed level, generating highly variable estimates in the published domains. New industry classes are proposed for stratification which align with publication. Our test results show that the new stratification works very well and can be used to generate good quality estimates. This document gives an analysis of the impact of the stratification change on the survey population and the sample allocation, as well as explaining the mechanism of how one change triggers the others.

Table of Contents

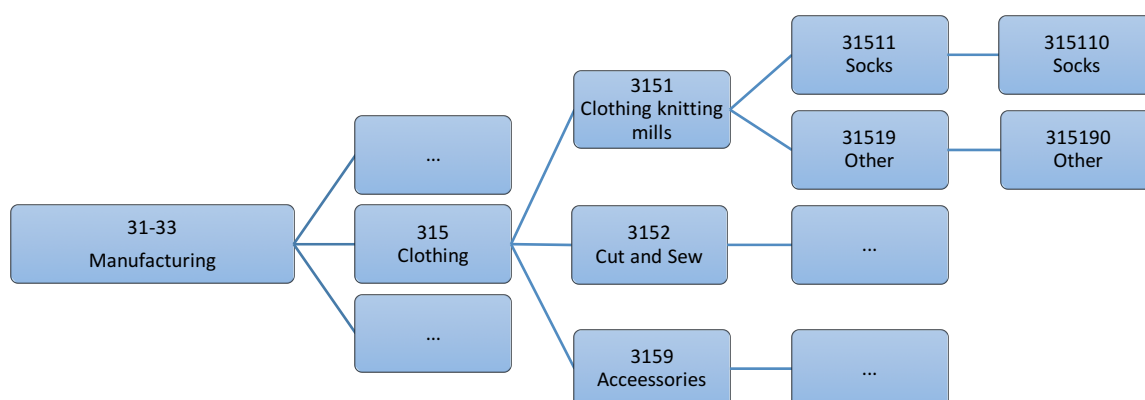
Summary.....	ii
1. Background:.....	1
Survey Frame.....	1
2. Introduction.....	2
Two Stratification Classes.....	2
Sampling Modules.....	2
3. Population Changes	3
Take-None	3
Must-Take	5
Enterprise Classification	6
Summary	6
4. Allocation Changes	7
Size-stratification.....	7
Allocation	7
5. Results.....	9
Coefficient of Variation	9
Results	10
Reasons for CV changes	11
6. Conclusions and Recommendations.....	12
7. References.....	13
8. Appendices	14
Appendix A: Province Code	14
Appendix B: List of Abbreviations	15

1. Background:

The Annual Survey of Manufacturing and Logging (ASML) is one of Statistics Canada's annual economic surveys. In ASML sampling, units are stratified by size within cells defined by crossing industry and geography. For geography, the stratification uses the 13 provinces and territories, notated with a two-digit number (Appendix A). The North American Industry Classification System (NAICS) is used for the industry classification (Statistics Canada 2015). It has a hierarchical structure composed of five levels from NAICS-2 to NAICS-6, as shown in Figure 1 (*Statistics Canada 2017*). Industries under identical sectors and subsectors are grouped together beginning with the same digits to form industry classes. In this report, domains are notated as 'industry class/province' (315110/59).

Figure 1

North American Industry Classification System



There are 253 distinct industries involved in ASML, and these are grouped together to form 88 industry classes. The large number of industries makes the survey design more complicated than other small surveys. In the past, the grouping of industries for stratification was above the level used for estimation, resulting in highly variable estimates in the published domains. To address this, a new grouping of industry classes was proposed for stratification.

This document compares the two stratifications, including all the identified changes and results. Throughout the report, comparisons are made by grouping detailed strata and cells from the new stratification together to match the groups under the old stratification. This is done when comparing counts (population and allocation) as well as CVs.

Survey Frame

Statistics Canada's Business Register (BR) is used as the frame for ASML. It is a data service centre including information about how businesses are organized, the industries and geographic regions they operate in, their revenues, etc. (*Statistics Canada 2015*). Tax data from the Canada Revenue Agency is the main source of the revenue variables in BR. The majority of businesses on the BR are simple enterprises that have only one establishment and can easily be classified to one industry and one geographic region. In contrast, complex enterprises are composed of several establishments classified to multiple industries and/or geographic regions, and usually account for more than 50% of revenue for

each domain (*Statistics Canada 2015*). Sampling for ASML is done at the enterprise level, so if an enterprise is selected, the questionnaire collects data covering all of its establishments within the manufacturing industries. For that reason, enterprises are also known as sampling units (SU) and establishments are known as operating entities (OE). The domain estimates, variances and CVs are all calculated at the establishment level.

2. Introduction

Two Stratification Classes

The list of industry classes that was implemented for reference year (RY) 2017 groups the 253 industry codes into 88 industry classes for the purpose of stratification. For RY2018, we propose to use the published NAICS as a new stratification, which groups the same 253 industry codes into 185 industry classes. Table 1 shows the distribution of NAICS levels for both stratifications.

Table 1

<i>NAICS Level</i>	RY2017	RY2018
<i>NAICS-3</i>	5	0
<i>NAICS-4</i>	63	0
<i>NAICS-5</i>	15	180
<i>NAICS-6</i>	5	5
<i>Industry Classes</i>	88	185
<i>Domain</i>	801	1,509

Generally, moving from RY2017 to RY2018, the stratification is done under a more detailed level in that every industry class is at NAICS-5/6. Combined with provinces, the RY2017 and RY2018 industry groupings generate 801 and 1,509 domains, respectively. In the past, the sample for ASML was selected under a two-phase design, and coordinated with the Statistics Canada's other annual economic surveys. In RY2017, the stratification was chosen to guarantee convergence of allocation under this two-phase design given the budget. For RY2018, the two-phase design has been eliminated and the allocation problem can be solved using the more detailed stratification which aligns with the NAICS levels used for publication.

The proposed RY2018 stratification ensures that the sample is representative at the NAICS-3/4/5 levels and that the sum of the weights is equal to the population total in each published domain. We expect this to decrease the variability of the NAICS-5 estimates compared to the RY2017 stratification which only controlled at the NAICS-3 level.

Sampling Modules

Sampling parameters such as the targeted population, stratification methods, Must-Take criteria, sampling method and sample size are defined in the sampling metadata, an Excel file that stores all the information needed to draw the sample. There are 12 steps or sampling modules run to generate

the sample for ASML. Table 2 shows the functions and files created by modules (*Reicker 2017*). The change in stratification will have a direct impact in module 2 to 7, and this document will explain each of these impacts in detail.

Table 2

Sampling Modules	Description
1-2	Population frame is created, Must-Take units are identified
3	Operating Entity Cell is defined, establishments from same industry class and province are grouped together
4	Establishment level Take-None – Royce-Maranda bound is defined, keeping top 90% of the revenue of the cell
5	Enterprise level Take-None – Enterprise is Take-None when each its establishments are Take-None
6	Outlier detection – added to Must-Take population Size-stratification –Take-Some units are stratified by revenue
7	Sample allocation – sample size determined for each stratum
8-10	Sample selection – particular units are selected from the frame
11-12	Output Survey Population Files are created

3. Population Changes

Two tests were run with the identical population using the October 2018 Generic Survey Universe File (G-SUF), which is a snapshot of the Business Register generated on October 1st, 2018. The target population is divided into the Take-None population and the survey population, which is further divided into a Take-Some portion and a Must-Take portion. Two major impacts of changing stratification are the decreased Take-None population and the increased Must-Take population.

Take-None

To reduce the response burden on small units, ASML uses the Royce-Maranda (RM) method to remove small units from the survey population (*Royce and Maranda 1998*). In each cell, the largest establishments, which together make up at least 90% of the revenue, are retained and the remaining establishments are added to the Take-None population (TN). In fact, the RM bounds are selected from a list of fixed numbers ranging from 50,000 to 3,185,000 instead of exactly 90% of revenue in a cell. The threshold is selected in such a way as to retain at least 90% of the revenue in each cell unless the minimum threshold eliminates more than 10% of the cell's revenue (*Gaudet and Stardom 2016*).

Take-None units are not sent for collection but have their data modelled using tax data and added to the final survey estimates prior to publication (*Gaudet and Stardom 2016*). After re-grouping cells from the new stratification to match the RY2017 stratification, we see that for some cells there is an increase in the TN population, but for most there is a decrease, and an overall decrease of 2,531 establishments over the entire survey. These 2,531 establishments are instead part of the survey

population, which increases by 8.9% in terms of establishments (10% in terms of enterprises). Table 3 shows some of the significant changes at NAICS-3 level. Notice that at places where the stratification changed a lot, for example, from 1 industry class to 7 detailed ones, there tends to be more change of Take-None. On the other hand, for places where the stratification has no change, there is no change in TN either.

Table 3

NAICS-3	Number of Industry Classes RY2017	Number of Industry Classes RY2018	Number of TN Units RY2017	Number of TN Units RY2018	Difference in TN Units
113	2	2	22,071	22,071	0
311	9	21	11,282	11,070	-212
313	1	7	1,082	974	-108
315	1	7	8,753	8,380	-373
327	2	11	2,992	2,748	-244
332	8	14	11,748	11,384	-364
336	15	15	3,314	3,314	0
339	2	7	16,777	16,350	-427
...
Total	88	185	126,174	123,643	-2,531

The specific TN boundary may change in a domain, take the domain 332A00/13 as an example. For the RY2017 stratification, the Royce-Maranda boundary is 500,000, with 40 TN units and 7 Must-Take units. Using the RY2018 stratification, with the same domain population, 332A00 is split up into 3 domains: 332910, 332990 and 332210, with RM bounds 3,185,000, 250,000 and 105,000, respectively. We identified a very large establishment in domain 332910, whereas units in the other two domains are very small. This implies that in the RY2017 stratification, the RM bound is skewed up by the large establishment, and many small and medium-sized units are considered TN units.

In contrast, domain 321900/10 is a case where there are more TN units under the RY2018. The RM bound is 160,000 and there are 32 TN units. As soon as it is split up into 3 domains, most of the large units are stratified into 2 domains, leaving the small units stratified into the other single domain. As a result, RM bound in those 2 domains with many large units is quite high - 375,000 and 850,000, and units that were above the original RM bound are now under the new RM bound. In this case, the TN population in domain 321900/10 increased.

In total, there are more cases where domains have a decreased TN, which leads to an overall larger survey population in RY2018. With the sample size remaining the same, this means a more variable estimate is expected.

Must-Take

Because of the distribution of revenue in ASML, certain influential units can neither represent other units nor be represented by them (*Gaudet and Stardom 2016*). Such units are stratified as Must-Take (MT) units and are excluded from the simple random sample selection. There are multiple sources of Must-Take, such as subject matter specified (establishments and enterprises identified by the survey manager), complex enterprises (active in more than 5 domains), small cell/stratum units (see below), outliers (in terms of revenue), or other special units (units which cannot report revenue and expenses directly). The changes to the Must-Take stratum have a significant effect on the allocation. In the RY2017 stratification, there are 2,090 MT enterprises, while the RY2018 includes 3,847 MT enterprises, an increase of 84%. Table 4 shows the detailed distribution of increase from each source.

Table 4

Multiple Sources of MT	RY2017	RY2018
Subject Matter Specified	189	189
Complex Enterprises	44	48
Small OE Cells (Establishment level)	249	663
Small SU Cells (Enterprise level)	1,282	2,356
Small Strata	223	472
Outliers	29	36
Special Units	74	83
Total (Enterprises)	2,090	3,847

There is no difference in the list of subject matter specified Must-Take units used in the two tests. Complex Must-Take enterprises are defined as enterprises that contribute to more than 5 domains. Since the RY2018 stratification uses a more detailed level of industry classes, four enterprises become complex Must-Take enterprises under the RY2018 stratification. Cells that contain less than 3 establishments are defined as small OE cells. The corresponding enterprises become Must-Take units (*Gaudet and Stardom 2016*). In this case, there is an increase of 414 enterprises.

The most significant factor is the increase in small SU cells and small strata, although the number of other MT also increases. Small SU cells are cells containing fewer than 10 enterprises above the RM bound and all units are made Must-Take (*Gaudet and Stardom 2016*). In domain 313000/59, there are 28 units above TN and no MT in the RY2017 stratification. In RY2018, with the identical population within domains, the cell split into 7. After the TN is determined, in 5 of the 7 domains, there are only 1 or 2 units (less than 10) in each domain. The 6 units from these 5 domains are made Must-Take. Within the 801 domains, there are 267 whose MT increased, and 25 domains become censuses.

In addition, an enterprise's outlier status can change depending on the stratification. Although the impact to the allocation is not as large as the previous case, it still reveals the primary change:

splitting domains. For example, a large enterprise may not originally be an outlier in a large domain. However, after changing the stratification, the large enterprise may be stratified into a domain where the other units are small. In this case, it becomes an outlier in the new stratification.

Enterprise Classification

The enterprise classification refers to the sampling cell where an enterprise is stratified. When an enterprise has establishments involved in multiple industry classes and provinces, the classification of an enterprise is determined by the maximum SU contribution:

$$\omega_j \frac{x_{ji}}{t_{x_j}}$$

where ω_j is the importance factor of sampling cell j , x_{ji} is the total OE size in sampling cell j of all OEs in sampling unit i , and t_{x_j} is the total OE size of sampling cell j . The importance factor of a cell is related to the proportion of revenue across all industries within a province (*Gaudet and Stardom 2016*), and therefore changes with the change in stratification. However, the main driver of enterprise classification change is the change in t_{x_j} as the cell definitions change. In total, only 71 enterprises changed their classification.

Summary

The primary impact of changing stratification is that the OE cells are changed. As we have discussed above, there is essential change in the survey population, in terms of Take-None, Must-Take and enterprise classification. After the survey population is finalized, size stratification, allocation and sample selection are conducted.

4. Allocation Changes

Size-stratification

After the Take-None and Must-Take units are defined in each cell, the remaining part is the Take-Some (TS) population, where a probability sample will be selected. Within each cell, units are stratified by size (revenue), which creates more homogeneous classes. Cells containing 10 to 29 TS units will be divided into two size strata, instead of three (all cells with 30 or more TS units). Since our detailed stratification split domains into small sub-domains, many cells have only 2 size strata where the corresponding cell under the RY2017 stratification had 3.

Allocation

For the two stratification tests, we used the same target sample size of 10,078, but we see a difference in the allocation. When looking at the sample sizes by NAICS-3, we find that at places where there is a large increase of industry class, hence a large change in stratification, there tend to be more increase of sample size, especially the MT population. Places with little stratification change have sample sizes that tend to decrease.

Table 5 shows some of the allocation comparisons by NAICS-3.

Table 5

NAICS-3	Number of Industry Classes RY2017	Number of Industry Classes RY2018	Sample size RY2017	Sample size RY2018	TS Difference	MT Difference	Sample size Difference	Percentage Difference in Sample Size
113	2	2	856	577	-279	0	-279	-32.59%
311	9	21	1,029	1,164	-87	222	135	13.12%
313	1	7	72	169	13	84	97	134.72%
314	1	4	114	201	40	47	87	76.32%
315	1	7	247	446	107	92	199	80.57%
323	1	2	440	358	-107	25	-82	-18.64%
327	2	11	328	527	48	151	199	60.67%
332	8	14	1,573	1,314	-396	137	-259	-16.47%
333	7	11	1,020	946	-150	78	-74	-7.25%
337	3	5	707	596	-170	59	-111	-15.70%
...
Total	88	185	10,875	11,177	-1455	1,757	302	2.78%

The difference of 302 units in the final sample size is composed of three sources. First, the sample sizes originally returned by the allocation module are fractions of numbers, which are rounded up. Each stratum contributes on average 0.5 units through rounding, and the published NAICS stratification has 220 more strata (959 compared to 739), so there is an increase in the final sample size. There is also an increase in the number of small-cell units which are selected but not sent to data collection, and these units are ignored during allocation but are included in the final sample size. There is also a difference in the total sample size selected by the allocation module, even though both tests use the same target sample size. Table 6 shows how the differences add up to 302.

Table 6

Source	Not Collected	Allocation	Roundup	Final Sample Size
RY2017	549	9,987.25	338.75	10,875
RY2018	635	10,081.05	460.95	11,177
Difference	86	93.80	122.20	302

5. Results

Coefficient of Variation

After the allocation for each stratum is determined, a simple random sample is conducted. The results from the two stratification tests are compared using the expected coefficient of variation (CV) based on the BR Revenue (available for the full population). The CV is the ratio of the standard error of the estimate to the expected value of the estimate itself, expressed as a percentage. It is a useful measure to assess the size of standard error relative to the estimate of the variable of interest (*Statistics Canada 2003*). In ASML, a quality rating is defined for each estimate based on the value of the CV, as shown in Table 7.

Table 7

CV - Quality Rating	Value
A	0 - 5%
B	5% - 10%
C	10% - 15%
D	15% - 25%
E	25% - 35%
F	> 35%

Results

We evaluate the CV by multiple dimensions – province, industry class, and domain. Figures 2 and 3 shows the distribution of CV by province and NAICS-3 industry class.

Figure 2

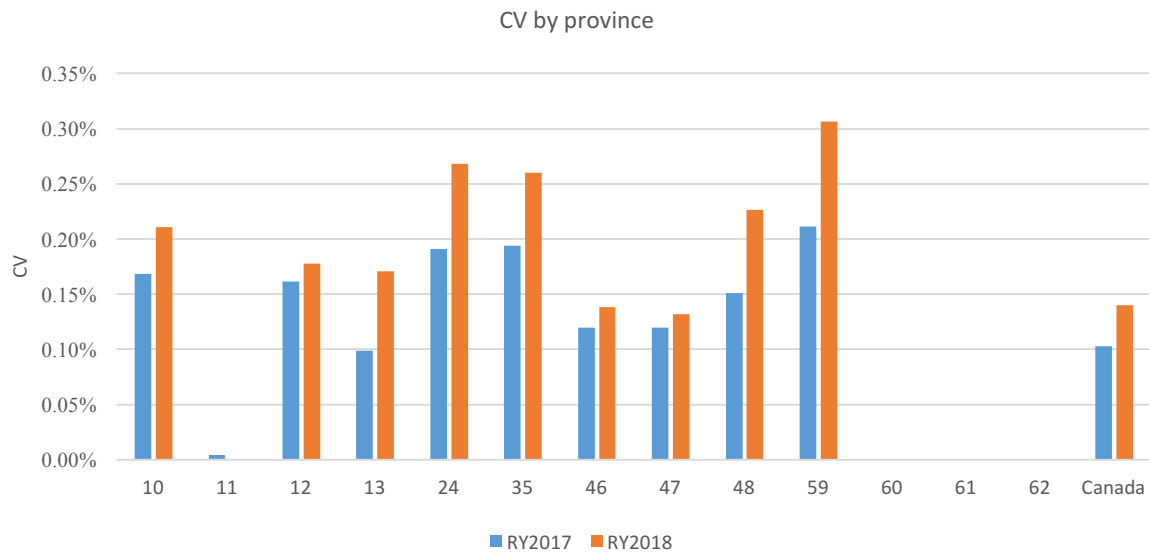
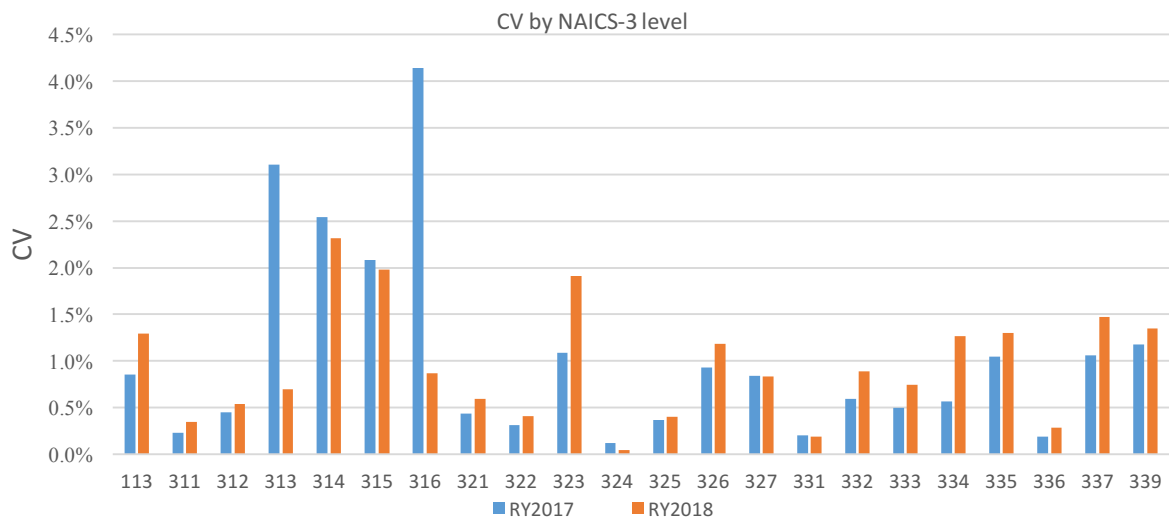


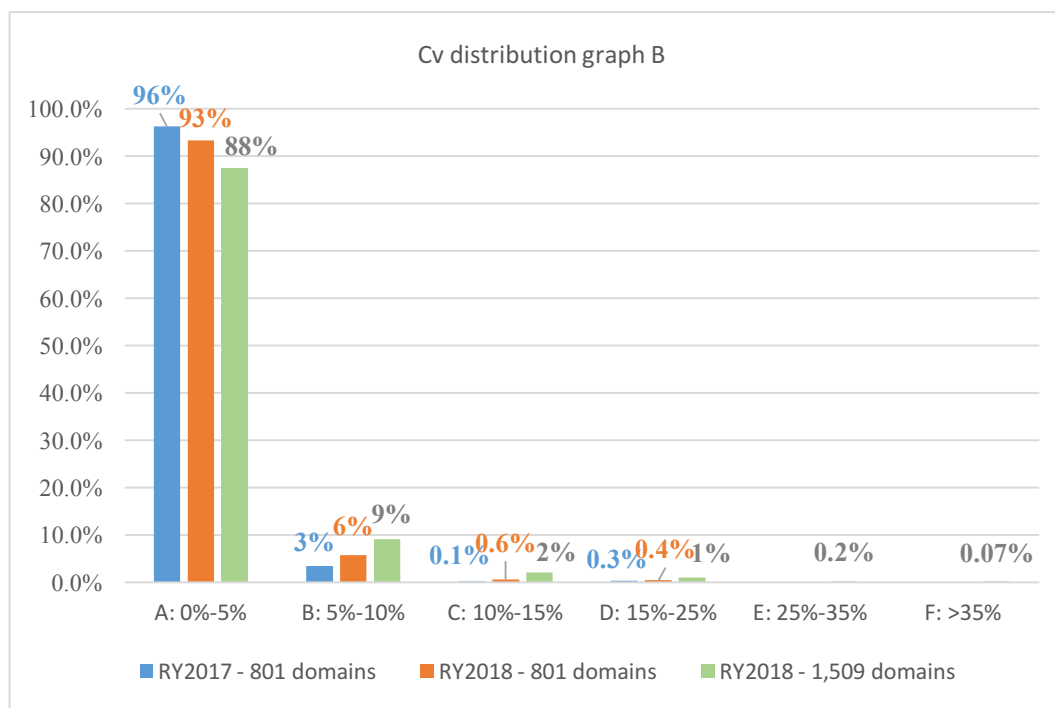
Figure 3



As shown in figures 2 and 3, we can see a slight increase at both provincial and industry levels. However, the increases of expected CV are all within 1% at the industry level, and within 0.1% at the provincial level, which are not significant increases. Domains that have high CVs before tend to have relatively high CVs using the new stratification as well. PEI is an exception and it is entirely selected under the new stratification. And NAICS 313 and 316 have a huge drop of CV, where there is a big change of stratification and increased sample size. Figure 4 shows a general distribution of the CV quality for the two stratification methods by domain. Note that the percentage of Y-axis refers to the

proportion of total cells. The comparison of CV is done across the 801 domains used for the RY2017 stratification (Figure 4 and Table 8) and the 1,509 domains for the RY2018 stratification (Figure 4). To compare at the RY2017 stratification level, the variances of the contributing RY2018 strata are added.

Figure 4



The RY2018 stratification revealed CVs with quality ratings of E and F which were concealed using the RY2017 stratification. In other words, the new stratification method gives us the opportunity of identifying and controlling domains that could be problematic.

Reasons for CV changes

Table 8 shows the distribution of difference of CV in the 801 domains generated by the RY2017 stratification.

Table 8

Difference of CV $CV_{RY2018} - CV_{RY2017}$	Count
10% ~ 15%	1
5% ~ 10%	10
0% ~ 5%	163
0%	487
-5% ~ 0%	132
Less than -5%	8
Total	801

There are various reasons why the CV may increase in a given domain. First, the number of establishments in the survey population above the TN boundary has increased by 8.9%. Since the target sample size is constant, we expect a larger CV overall. For example, TN in domain 337900/24 dropped from 78 to 73 and lead to a 5.3% increase of CV. Second, based on the fact that some domains have a large increase in MT, and that the sample size for both tests is the same, there must be other domains where the sample size decreases. Domains where the stratification changes very little lose the most sample. Specifically, in the 11 domains whose CV increased by more than 5%, 5 of them are the result of less sample selection. For domain 113311/46, although the sample size did not change, the allocation between small and medium strata changed. This occurs when the solver reaches the numerical tolerance and stops before finding the global optimum. The proposed solution is to decrease the numerical tolerance of the optimization algorithm.

The sample selection and weighting are determined by SU (enterprise) cell, while the calculation of CV is done by OE (establishment) cell, and this can have an impact on the CV. In some cases, a large increase in the CV of a certain domain is because of a change in the sampling fraction in another domain. For domain 316000/48, there is a large establishment whose enterprise is stratified outside of 316000/48. In the RY2017 stratification, it has weight of 1, while under the RY2018 stratification, its weight increased to 1.17. This leads to the CV increasing from 14% to 20%.

6. Conclusions and Recommendations

The proposed stratification controls the sample at the NAICS-5 level and does so while maintaining low expected CVs for the NAICS-3 level domain estimates. These very positive results were presented to the survey managers and this new stratification based on the published NAICS will be implemented for RY2018. Many small units are selected at the expense of larger, more influential units, which contributes to a slight increase in the overall CV, from 0.1% to 0.14%. In addition, units are moved from industry classes with small changes between the two stratifications to industry classes with large changes, which increases the expected CV in some domains. The biggest benefit comes from aligning the stratification with publication. The overall survey population changes from Take-None, Must-Take and enterprise classification fluctuations have been analyzed and understood, as well as the changes to the size-stratification. After studying all the impacts from changing stratification, we conclude that the primary impact is that most OE cells/domains split into detailed ones, which leads to an increase in the survey population and a large increase in the Must-Take population.

The proposed level of stratification for RY2018 is very detailed and this increases the size of the Must Take population by adding small and medium-sized units in small cells. To minimize the response burden on these units, and to allow more of the sample to be allocated where it will have the biggest impact on the overall CV, we are looking into defining a new stratification by combining certain small domains. These refinements of the stratification will allow Statistics Canada to produce high quality estimates to Canadians and the decision makers who depend on the data from the Annual Survey of Manufacturing and Logging.

7. References

Gaudet, J., Stardom, J. (2016) An Overview of the Integrated Business Statistics Program Sampling Methodology, Economic Statistics Methods Division, Statistics Canada.

Reicker, A. (2017) Files Created by the Sampling Modules, Internal Statistics Canada document.

Royce, D., Maranda, F. (1998) Task Group on Data Acquisition for Enterprises, Internal Statistics Canada document.

Statistics Canada (2003) Survey methods and practices. National Library of Canada Cataloguing in Publication Data.

Statistics Canada (2015). Integrated Business Statistics Program Overview, Internal Statistics Canada document.

Statistics Canada (2017). Definitions, data sources and methods. Retrieved from <http://www.statcan.gc.ca/eng/concepts/index>

8. Appendices

Appendix A: Province Code

Code	Abbreviation	Province and territory
10	NL	Newfoundland and Labrador
11	PEI	Prince Edward Island
12	NS	Nova Scotia
13	NB	New Brunswick
24	QC	Quebec
35	ON	Ontario
46	MB	Manitoba
47	SK	Saskatchewan
48	AB	Alberta
59	BC	British Columbia
60	YT	Yukon
61	NT	Northwest Territories
62	NU	Nunavut

Appendix B: List of Abbreviations

Acronym	Explanation
ASML	Annual Survey of Manufacturing and Logging
BR	Business Register
CV	Coefficient of Variation
MT	Must-Take
NAICS	North American Industry Classification System
OE	Operating Entity (Establishment)
RM	Royce Maranda
RY	Reference Year
SU	Sampling Unit (Enterprise)
TN	Take-None
TS	Take-Some