Implementation of Small Area Estimation Techniques in an Economic Survey: the Experience of the Monthly Survey of Manufacturing in Canada

Sébastien Landry[1]
[1]Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON K1A 0T6

Abstract
The Monthly Survey of Manufacturing conducted by Statistics Canada, which has a sampling design aiming at producing estimates by province, undertook a project to implement small area estimation (SAE) techniques in order to publish estimates of sales for 12 Census Metropolitan Areas without increasing its sample size. Apart from the obvious determination of the proper SAE model to use, challenges in the implementation of the techniques included the handling of the take-none portion of the estimates, the consistency of the SAE estimates with the provincial estimates, the derivation of a quality indicator and the determination of a confidentiality strategy that incorporates survey and modeled estimates. After a brief summary of the survey's current methodology, the solutions to the previously stated challenges will be presented.

Key Words: Small area estimation, consistency between estimates, quality indicators, confidentiality

1. Introduction

Requests have been made to Statistics Canada to produce monthly estimates of sales of goods manufactured for large Census Metropolitan Areas[1] (CMA) by industry groups through its Monthly Survey of Manufacturing (MSM). As the MSM is currently designed to provide estimates at the province level and not the CMA level, the effective sample size for CMA domains may not be large enough to produce reliable estimates. Given this state, two options can be considered to fulfill these requests:

- Redesign the survey so that it ensures a sample at the CMA level large enough to produce reliable estimates for these domains;
- Use small area estimation (SAE) techniques, with the help of auxiliary information, to produce estimates for these domains.

Since the first option would lead to an increase in the survey's sample size and therefore its cost, the second option was selected. However, implementing SAE techniques brings

[1] A Census Metropolitan Area is defined by the formation of one or more adjacent municipalities centred on a population centre (known as the core). A CMA must have a total population of at least 100,000 of which 50,000 or more must live in the core, based on adjusted data from the previous census.

its challenges. This paper will present the challenges faced when implementing SAE techniques in the MSM.

Section 2 will provide an overview of the elements of the MSM that are relevant to this paper. Section 3 will present the selected SAE model and how it was implemented in the MSM. Section 4 will discuss the specific challenges that were faced while implementing SAE techniques in the survey. Section 5 will conclude this paper by presenting results from the implementation of SAE techniques in the survey.

2. Overview of the Monthly Survey of Manufacturing

The Monthly Survey of Manufacturing is a survey conducted monthly over manufacturing businesses. The objective of the survey is to produce estimates of sales of goods manufactured (sales), inventories, unfilled orders, new orders, production and capacity utilization rates for national domains that represent various industry levels (based on the North American Industry Classification System (NAICS)). Estimates of sales and production are also produced for provincial domains.

The sampling frame, built using Statistics Canada's business register, is stratified by industry, province and size (the size variable is the business' annual revenue). Each industry-province cell is divided according to the businesses size, into one take-all stratum, which contains the largest businesses, up to two take-some strata, which contain medium-size businesses, and one take-none stratum, which contains the smallest businesses that globally account for the lowest 10% of the cell's size. The take-all and take-some strata boundaries are derived using the Lavallée-Hidiroglou method (Lavallée and Hidiroglou (1988)).

After the collection and the imputation (if needed) of data, estimates are produced using a ratio estimator. The weighting is done in two steps: first, the design weight is merely the inverse of the selection probability of the business, and second, a weight calibration that is based on an auxiliary variable, in this case the monthly sales derived from Goods and Services Tax (GST) records obtained from the Canada Revenue Agency, which is processed at Statistics Canada to provide a value for all businesses. This ratio estimator ensures that the weighted sum of the auxiliary variable is equal to the calibration group total of that variable. For each industry-province cell, a calibration group is created for the take-all stratum and one is created for the take-some strata. The auxiliary variable total from the take-none stratum is added to the take-some calibration group total so that this portion of the frame is accounted for in the estimates. If no take-some stratum exists for an industry-province cell, the auxiliary variable total from the take-none stratum is added to the take-all calibration group total. Estimates are then calculated from either a weighted sum of a variable or a ratio of two estimated totals.

More information about the MSM can be found in Statistics Canada (2020a).

3. Production of SAE Estimates for CMAs

3.1 Scope of the Project

The 12 CMAs that were selected for this project are Halifax, Quebec City, Montreal, Ottawa-Gatineau, Toronto, Hamilton, Winnipeg, Regina, Saskatoon, Calgary, Edmonton and Vancouver. They are the largest CMAs in their respective provinces, which explains their inclusion in this project.

The estimates for CMA domains are meant to act as a complement to the current survey. This means that provincial estimates produced through the current survey will not be adjusted using SAE techniques.

3.2 Choice of a Small Area Estimation Model

The selected SAE model for this project was the Fay-Herriot area-level model (Fay and Herriot (1979)). This model uses an assumed linear relationship between the domain estimates coming from the survey and auxiliary information (in this project, the auxiliary information (to be defined in section 3.3.2) domain population total) independent from the survey and correlated with the domain estimates coming from the survey to produce estimates that are more reliable than the estimates coming from the survey regular processes.

The Fay-Herriot area-level model is composed of two components:

- a sampling model which shows the relationship between the estimate coming from the survey and the true value:

$$\hat{\theta}_i^{SURVEY} = \theta_i + e_i;$$

- a linking model which shows the relationship between the true value and the auxiliary information:

$$\theta_i = \boldsymbol{z_i}' \boldsymbol{\beta} + v_i;$$

where i is a domain, $\hat{\theta}_i^{SURVEY}$ is the estimate coming from the survey, θ_i is the true value, e_i is the sampling error with $e_i \sim N(0, \sigma_{s.i}^2)$ and $\sigma_{s.i}^2$ is the sampling variance, $\boldsymbol{z_i}'$ is the auxiliary information, v_i is the model error with $v_i \sim N(0, \sigma_m^2)$ and σ_m^2 is the model variance (all these variables are associated to domain i) and $\boldsymbol{\beta}$ is the vector of regression coefficients.

The required inputs for the model, for all the domains for which estimates are required, are the estimates coming from the survey, the variance of the estimates coming from the survey and the auxiliary information. These inputs are used to estimate the linking model regression coefficients ($\widehat{\boldsymbol{\beta}}$) as well as the variance of the model error ($\hat{V}(v_i)$).

After estimating $\widehat{\boldsymbol{\beta}}$, a prediction from the model, or synthetic estimate, for domain i can be derived:

$$\hat{\theta}_i^{SYNTHETIC} = \boldsymbol{z_i}'\widehat{\boldsymbol{\beta}}.$$

The final SAE estimate for domain i is then a linear combination of the estimate coming from the survey and the synthetic estimate:

$$\hat{\theta}_i^{SAE} = \gamma_i \hat{\theta}_i^{SURVEY} + (1 - \gamma_i)\hat{\theta}_i^{SYNTHETIC}$$

with

$$\gamma_i = \frac{\hat{V}(v_i)}{(e_i) + \hat{V}(v_i)}.$$

3.3 Implementation of SAE in the MSM

This section is based on the report by Bocci and Beaumont (2019).

3.3.1 Definition of domains to be used in SAE
A domain is defined as 3-digit NAICS code (except for transportation equipment where the 4-digit NAICS code is used) and CMA. There can be as many as 324 domains (provided that a business exists in each domain).

A domain can be divided in three portions (subdomains): the portion coming from take-all strata (TA subdomain), the portion coming from take-some strata (TS subdomain) and the portion coming from take-none strata (TN subdomain). Since the TA subdomain is self-representative, it was decided to keep its estimate coming from the survey in the final domain estimate. Therefore, only the TS subdomain will be used in the SAE model (the reasons why the TN subdomain is not used in the SAE model will be explained in section 4.1).

3.3.2 Definition of variables to be used in SAE
Although the objective of this project is to produce estimates of totals, the tests conducted to select the SAE model showed that using population means proved to provide with more reliable SAE estimates. Therefore, the input variables are defined as follows.

- Estimate coming from the survey ($\hat{\theta}_i^{SURVEY}$): estimate of total sales from the TS subdomain divided by the TS subdomain population size.
- Variance of estimate coming from the survey ($\hat{V}(\hat{\theta}_i^{SURVEY})$): estimated variance of the estimate of total sales from the TS subdomain, which includes the sampling variance and the variance due to nonresponse, divided by the TS subdomain population size squared.
- Auxiliary information (z_i'): total monthly sales derived from GST records from the TS subdomain divided by the TS subdomain population size[2].

After the model parameters have been estimated, the resulting SAE estimates and mean square error (MSE) are rescaled to obtain estimates of totals.

The variance of estimate coming from the survey for a domain can be quite unstable when its sample size is small. The variances of estimates coming from the survey were therefore smoothed using a variance smoothing model. The variance smoothing model is a log-linear regression:

$$\log\left(\hat{V}\left(\hat{\theta}_i^{SURVEY}\right)\right) = x_i'\alpha + \varepsilon_i$$

where $x_i' = (1, \log(z_i), \log(N_i))$ and N_i is the TS subdomain population size. After estimating $\hat{\alpha}$, the smoothed variance can be calculated so that the average of the variances coming from the survey is equal to the average of the smoothed variances:

$$\hat{\tilde{V}}\left(\hat{\theta}_i^{SURVEY}\right) = \exp(x_i'\hat{\alpha})\frac{\sum_i \hat{V}\left(\hat{\theta}_i^{SURVEY}\right)}{\sum_i \exp(x_i'\hat{\alpha})}.$$

[2] No intercept is used in the SAE model in order to avoid the possibility of a negative SAE estimate.

3.3.3 Outlier detection

Outlier detection was conducted on the results coming from the variance smoothing model and the Fay-Herriot area-level model in order to improve the SAE estimates. The outlier strategy is more or less the same for both models: the standardized (or studentized) residuals (r_i) from the models are assumed to follow a N(0,1) distribution, which implies that $r_i^2 \sim \chi_1^2$. The domain with the largest r_i^2 is deemed an outlier if r_i^2>c with c such as P($r_i^2 \leq$c)=0.99. Table 1 describes the specific components of the outlier detection for each model.

Table 1: Components of the Outlier Detection Specific to Each Model

Model	Variance smoothing model	Fay-Herriot area-level model
Calculation of r_i	$\frac{\log\left(\hat{V}\left(\hat{\theta}_i^{SURVEY}\right)\right) - x_i'\alpha}{\sqrt{\hat{V}(\varepsilon_i)}}$	$\frac{\hat{\theta}_i^{SURVEY} - z_i'\hat{\beta}}{\sqrt{\hat{V}(v_i + e_i)}}$
Handling of the outlier domains	$\hat{V}\left(\hat{\theta}_i^{SURVEY}\right)$ is used in the Fay-Herriot model instead of $\hat{\hat{V}}\left(\hat{\theta}_i^{SURVEY}\right)$	$\hat{\theta}_i^{SAE} = \hat{\theta}_i^{SURVEY}$

The outlier detection is an iterative process, i.e. if the domain with the largest r_i^2 is deemed an outlier, the parameters of the model are re-estimated using the remaining domains, test if the domain with the largest r_i^2 is deemed an outlier, and so forth until no more outliers are found.

3.3.4 Final estimate and MSE

The final estimate for a domain will be the sum of the TA subdomain estimate, the TS subdomain estimate and the TN subdomain estimate. As stated earlier, the TA subdomain estimate will be its estimate coming from the survey. The TS subdomain estimate usually will be the SAE estimate. One exception occurs when all the businesses in the TS subdomain were selected in the sample by chance: in this case, the TS subdomain estimate will be the unweighted sum of sales from the subdomain. The TN subdomain estimate will be the synthetic estimate. The following formula summarizes this.

$$\hat{\theta}_i^{SAE} = \hat{\theta}_{i,TA}^{SURVEY} + \hat{\theta}_{i,TS}^{SAE} + \hat{\theta}_{i,TN}^{SYNTHETIC}$$

The MSE of the domain estimate is calculated as the sum of the variance of the TA subdomain estimate, the MSE of the TS subdomain estimate and the MSE of the TN subdomain estimate.

$$\widehat{MSE}(\hat{\theta}_i^{SAE}) = \hat{V}(\hat{\theta}_{i,TA}^{SURVEY}) + \widehat{MSE}(\hat{\theta}_{i,TS}^{SAE}) + \widehat{MSE}(\hat{\theta}_{i,TN}^{SYNTHETIC})$$

4. Challenges Faced While Implementing SAE Techniques

Four additional challenges were encountered during the implementation of SAE techniques in the MSM: calculating the estimation of the TN subdomain, ensuring consistency between the CMA estimates and the provincial estimates, the determination of a quality indicator and building a confidentiality pattern. This section will describe each one and the way they were resolved.

4.1 Calculating the Estimation of the TN Subdomain

In the current survey, estimates for the TN subdomain are derived using the sample to reflect the participation of the TN subdomain in the weight calibration. This is achieved through a secondary calibration where the calibration group totals are composed of the auxiliary variable total coming from the take-none strata only. Therefore, an estimate for a CMA TN subdomain can be calculated only if it is represented in the sample. This fact is also true for a CMA TS subdomain. However, some CMA TN subdomains are not represented in the TS subdomain population, so they have no chance of being represented in the sample. Furthermore, the auxiliary variable total coming from the take-none strata only can either be included in the take-all calibration group or the take-some calibration group for calibration, which would prevent consistency in the model inputs if the CMA TN subdomains were to be included in the SAE model.

This resulted in the exclusion of the CMA TN subdomains from the SAE model and the use of the synthetic estimator for the CMA TN subdomain estimates as previously stated. In order to fully remove the TN subdomains from the SAE model, the estimates coming from the survey were recalculated, only to be used for the SAE model, by removing the auxiliary variable total coming from the take-none strata from the calibration group totals.

4.2 Ensuring Consistency between the CMA Estimates and the Provincial Estimates

It is common sense that a provincial estimate must be greater or equal to the sum of the CMA SAE estimates that are in the province. However, since the SAE model has no restrictions in terms of bounds for estimates, it is technically possible that a CMA SAE estimate is greater than the estimate of its province. In such a situation, an adjustment must be made to the CMA SAE estimate to restore the consistency between the CMA SAE estimate and the provincial estimate.

It has been decided that the CMA TA subdomains would not be subject to an adjustment: since the CMA TA subdomain estimates were not calculated using the SAE model, they did not contribute to this issue. Also, the non-CMA TA subdomains (the portion of the province outside of the selected CMAs) should remain unadjusted in the final estimates.

The adjustment consists of calculating the sum of the CMA TS and TN subdomain estimates of a province and, if it's greater than the take-some and take-none portions of the provincial estimate, adjust the CMA TS and TN subdomain estimates so that the sum is equal to the take-some and take-none portions of the provincial estimate. The algebraic expression of the adjustment is shown below (*P* stands for province).

If $\sum_{i \in P} \hat{\theta}_{i,TS}^{SAE} + \hat{\theta}_{i,TN}^{SYNTHETIC} > \hat{\theta}_{P,TS} + \hat{\theta}_{P,TN}$ then multiply $\hat{\theta}_{i,TS}^{SAE}$ and $\hat{\theta}_{i,TN}^{SYNTHETIC}$ by $\frac{\hat{\theta}_{P,TS}+\hat{\theta}_{P,TN}}{\sum_{i \in P} \hat{\theta}_{i,TS}^{SAE}+\hat{\theta}_{i,TN}^{SYNTHETIC}}$.

Ottawa-Gatineau is a CMA that overlaps two provinces (Quebec and Ontario). Since the proportion of the Ottawa-Gatineau SAE estimate allocated to each province is unknown, simply applying the previously shown adjustment is not straightforward as the sum of a province's CMA SAE estimates could legitimately be greater than the estimate of the province. It has been decided that the best way to handle this is to verify the following inequalities.

- The sum of the SAE TS and TN subdomain estimates from the CMAs of a province that are not Ottawa-Gatineau should be less or equal to the take-some and take-none portions of the estimate of the province.
- The sum of all Quebec and Ontario CMA SAE TS and TN subdomain estimates should be less or equal to the sum of the take-some and take-none portions of the estimates of Quebec and Ontario.

To cover the case where more than one of the inequalities would not hold, in which case more than one adjustment factor could be applied, the adjustment factors are calculated by solving a minimization factor.

4.3 Determination of a Quality Indicator

In the MSM, the quality of a domain estimate is summarized through a quality indicator which is a letter grade from A to F, A meaning "excellent" and F meaning "unreliable". This quality indicator is derived from the CV based on the sampling variance of the domain estimate and the response fraction of the domain. However, in the context of small area estimation, the response fraction is a less relevant measure to assess quality.

The quality indicator of a CMA SAE estimate will be derived using $\widehat{MSE}(\hat{\theta}_i^{SAE})$, more specifically the relative root mean square error $\widehat{RRMSE}\left(\hat{\theta}_i^{SAE}\right) = \frac{\sqrt{\widehat{MSE}(\hat{\theta}_i^{SAE})}}{\hat{\theta}_i^{SAE}}$. These measures include the sampling variance, the variance due to nonresponse and the variance due to modelling. Table 2 shows how the quality indicator is defined.

Table 2: Derivation of the Quality Indicator for a CMA SAE Estimate

$\widehat{RRMSE}(\hat{\theta}_i^{SAE})$					
[0%-5%[	[5%-10%[	[10%-15%[	[15%-25%[	[25%-35%[	[35% or +
A	B	C	D	E	F

4.4 Building a Confidentiality Pattern

According to Canada's Statistics Act (Department of Justice (2017)), Statistics Canada has the obligation to protect collected data, whether they come from surveys or alternative data sources. One way to accomplish this is to make sure that no published estimate reveals reliable information about an entity (person, business, etc.).

For the scope of this project, confidentiality is achieved by protecting the estimates of domains that are deemed sensitive, i.e. where one business is dominant within its domain, which is done by suppressing (not publishing) estimates of specific domains. The result of the determination of the domains for which their estimate will be suppressed is called a confidentiality pattern. Figure 1 illustrates a classic attack scenario where a suspect tries to derive the value of the dominant business, thus leading to a need for protection.

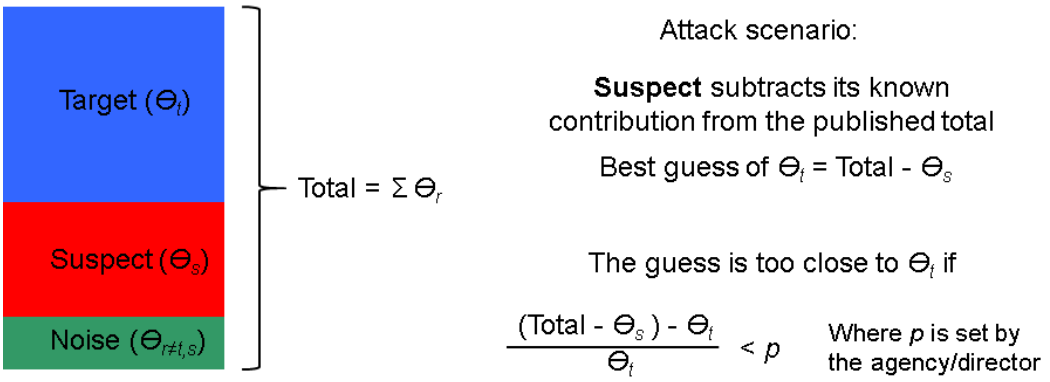

Figure 1: Visualization of a classic attack scenario where a suspect tries to derive the value of target business

The domains that need protection are the CMA domains, the non-CMA domains (although they are not subject to official publication, they can be derived from the published estimates) and the provincial domains. The provincial domains are already subject to a confidentiality pattern and the new confidentiality pattern must take into account this first pattern (i.e. a provincial domain that was published cannot become suppressed and vice versa).

All businesses that participated in the calculation of the estimates (for CMA or provincial domains) must be used for the confidentiality assessment of a domain. Such businesses include those coming from the CMA domains population whether selected or not in the survey sample (since their auxiliary variable was used in the SAE model) and those coming from the sample from the non-CMA domains (that were used to calculate the provincial estimates).

The variable to be used in the confidentiality assessment will be different depending on the sampling status of the businesses. For the businesses in sample, the variable used will be their reported sales from collection, which is the value the respondents will remember providing to Statistics Canada. For the businesses not in the sample, the variable will be a synthetic value derived from the SAE model ($z_j'\hat{\beta}$, where j is a business), which is the best proxy for sales and is fairly close to the auxiliary variable value, given the correlation between sales and the auxiliary variable.

In order to assess confidentiality for a domain, the sum of the variable that needs protection from the businesses participating in the confidentiality assessment in a domain must equal the domain estimate. Since the domain estimate includes the impact of weighting and modelling, this equation will rarely hold as is. Therefore, anonymous records (records that don't need protection in the confidentiality assessment), for which their variable value will be the difference between a subdomain estimate and the sum of the variable from the businesses in that subdomain, are added to the confidentiality assessment population. These anonymous records effectively add to the noise presented in Figure 1.

Once the list of domains that need protection and are subject to the primary suppression has been established, the estimate of additional domains needs to be suppressed to effectively protect the sensitive domains (this is known as secondary suppression). The domains that are subject to the secondary suppression are selected to minimize the sum of

the suppressed domains' variable while ensuring the protection of the sensitive domains. The non-CMA domains, that are not officially published, will be prioritized for secondary suppression.

5. Results and Conclusion

The SAE estimates have been produced for the specified domains for all the reference periods since January 2009 and the production of the CMA estimates is now integrated in the survey's regular production. The CMA estimates can be viewed in Statistics Canada (2020b).

Various diagnostics are calculated when building the SAE model to assess its validity. Two of these diagnostics are presented below for a specific reference period. Figure 2 compares the TS subdomains estimate coming from the survey with their SAE estimates. It is expected that the SAE estimate stays relatively close to the estimate coming from the survey especially for large domains. Figure 3 compares the TS subdomains CV coming from the survey with their SAE RRMSE. It is expected that the SAE RRMSE is lower than the CV coming from the survey. As expected, SAE estimates outperform in terms of the ratio of RRMSE and CV the estimates coming from the survey. Notice that, the estimate coming from the survey CVs are the one obtained after applying the log-linear regression to smooth the variance.

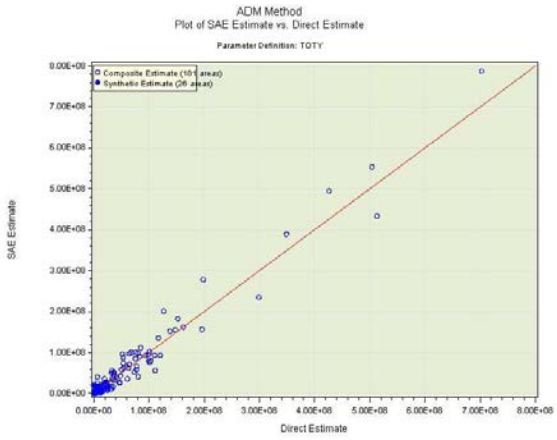

Figure 2: Comparison between the estimate coming from the survey and the SAE estimate for the CMA TS subdomains for the reference period May 2020

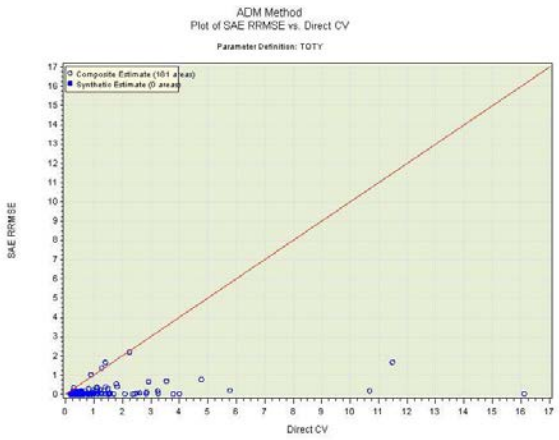

Figure 3: Comparison between the CV coming from the survey and the SAE RRMSE for the CMA TS subdomains for the reference period May 2020

The confidentiality pattern that has been established allows for around 60% of the CMA domains, representing around 85% of the total sales, to be published every month.

In terms of ongoing and future project, seasonal adjustment is currently being performed on the CMA estimates and will be published in a near future, then additional CMAs may be added to the scope of this project.

Acknowledgements

The author would like to thank the numerous people who worked on this project for their contribution. The author would also like to thank the reviewers for their useful comments.

References

Bocci, C., and Beaumont, J.-F. (2019), Small area estimation methodology applied to the Monthly Survey of Manufacturing- UPDATE, internal document.

Department of Justice (2017), Statistics Act, https://laws-lois.justice.gc.ca/eng/acts/s-19/fulltext.html.

Fay, R.E., and Herriot, R.A. (1979), Estimation of income from small places: An application of James-Stein procedures to census data. Journal of the American Statistical Association, 74, 269-277.

Lavallée, P., and Hidiroglou, M.A. (1988), On the stratification of skewed population, Survey Methodology, 14, 33-43.

Statistics Canada (2020a), Monthly Survey of Manufactures (MSM), https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=2101.

Statistics Canada (2020b), Table 16-10-0011-01 Manufacturing sales, by industry for 12 Selected Census Metropolitan Areas (x 1,000), https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1610001101.