Detection method for the sources of change in estimates

Serge Godbout
Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6, serge.godbout@statcan.ca

Abstract

In the context of a survey, there is often a need to know and to quantify the main sources of changes observed between two estimates over time. We can then explain these changes or identify the outlier units having substantial impact on the change. Depending on the estimator used, the type of data or the number of units, this exercise can become highly complex.

We will describe a method based on partial derivatives to find the sources of a change. This method can be used in several contexts. Using the estimator formula, we consider the independent variables as functions over time, calculate the partial derivatives of the estimator and approximate the impacts of the units. We can then estimate the impact on the change of each variable and each unit. The main advantage of this simple method is that it can be used to easily and accurately identify which, among all the data, have had a significant impact.

Key words: Measurement of impact, outlier.

1. Introduction

In the context of a survey, there is often a need to know and to quantify the main sources of changes observed between two estimates over time, whether in a repetitive survey or an historic revision. During editing, we can identify outlier units with a significant impact on the estimates, while at the analysis stage, we can use the results to explain these changes. For example, during a panel survey, changes in the lagged variables or in the design weights can change an estimate, sometimes negligibly, but sometimes substantially. Depending on the estimator used, the type of data or the number of entries, it can be a highly complex task to identify the impact of each variable or unit on the final estimate. In the case of manual edits, it can be costly for a survey to verify a large number of units, which often have very little impact on the estimates. In contrast, the accuracy of estimates can suffer from the impact of a small number of overlooked units.

Some of the outlier detection methods often stated include such univariate methods as the sigma-gap, interquartile difference or the Hidiroglou-Berthelot method (1986), or such multivariate methods as the use of graphs or Cook's and Mahalanobis distances (Cook, 1977 ; Franklin, 2000). These methods have the advantage of being quite simple but they cannot directly link an outlier to its impact on estimates. In addition, they cannot be used to explain changes observed in estimates.

We will describe a method based on the definition of partial derivatives to measure the impact of each variable of each unit. We will then use these measures to identify units for verification purposes and to explain the source of a change. Examples using common estimators will illustrate the method. Lastly, we will describe how impact measurement was applied to Statistics Canada's Survey of Employment, Payroll and Hours (SEPH).

1. Measurement of impact

The proposed method involves a transformation built from the estimator where all variables used to produce an estimate are linked to a real number that represents the impact that the change in that variable has on the final estimate. Analysis of the distribution of these impacts allows for identification of the units for verification and identifies the sources of a change observed between two estimates.

Let us consider an estimator f based on the m variables:

$$\mathbf{x} = (x_1, x_2, ..., x_m)$$

We assume that this estimator is a real function of the m real variables, i.e. that $f(\mathbf{x}) : \mathbf{R}^m \rightarrow \mathbf{R}$. For example, in the case of an estimator using a design weight w_k and a variable of interest y_k available for n units, we consider the estimator to be a function of $m = 2n$ real variables.

Let us now assume that this estimator can be derived on an open domain $P \subset \mathbf{R}^m$. Let $\mathbf{x}_0 \in P$ be a point on P. We can say that the plane T tangent to function f at point $\mathbf{x}_0 = (x_1^0, x_2^0, ..., x_m^0)$ is a good approximation of function f around $\mathbf{x}_0$ (Marsden and Tromba, 1988). That is,

$$f(\mathbf{x}) \cong f(\mathbf{x}_0) + \sum_{i=1}^{m} \frac{\partial f(\mathbf{x}_0)}{\partial x_i}\left(x_i - x_i^0\right)$$

Now, let $A=\mathbf{x}_A=\left(x_1^A, x_2^A, ..., x_m^A\right)$ and $B=\mathbf{x}_B=\left(x_1^B, x_2^B, ..., x_m^B\right)$ be two points on P, and let AB be the rectangle defined by the points A and B :

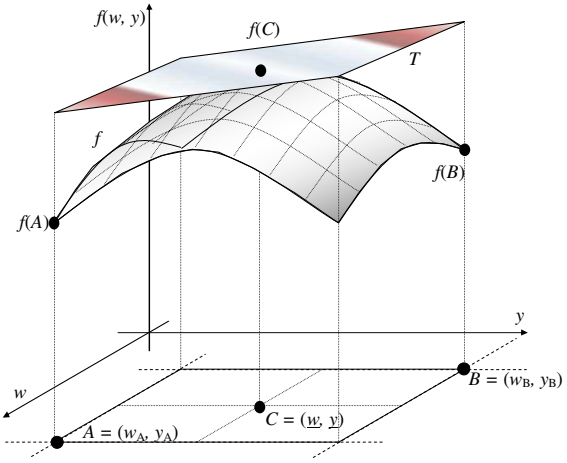
$$AB=\left\{\mathbf{x}=(x_1,...,x_m): x_i^A \le x_i \le x_i^B \ \forall i=1,...,m\right\}$$

If $AB \subset P$, we can conclude that:

$$\Delta f_{AB}=f(B)-f(A) \cong \sum_{i=1}^{m}\frac{\partial f(C)}{\partial x_i}\left(x_i^B-x_i^A\right)$$

for any point $C=\mathbf{x}_C \in AB$. More specifically, we can state that there exists at least one point $C_0 \in AB$ such that the approximation becomes an equality (Rudin, 1976). In contrast, the exact determination of a specific point C_0 is complex, but not essential since by taking C located in the middle of segment $\overline{AB}$, i.e. $C=\frac{1}{2}\left(\mathbf{x}_A+\mathbf{x}_B\right)=\underline{\mathbf{x}}$, we obtain a point that meets the constraint $C \in AB$ and that allows us to calculate the impacts with a sufficient precision. Figure 1 shows the function $f(\mathbf{x})=f(w,y)$ above the rectangle AB and the plane T tangent to the function at point C in the simplified case of a function with only two variables.

Figure 1
Function $f(\mathbf{x})$ and tangent plane



This way, we get an approximation formula of the difference between two estimates Δf_{AB} based on a sum of m impacts caused by the change in each variable used to produce the estimate:

$$\Delta f_{AB}=\Delta f \cong \sum_{i=1}^{m}\frac{\partial f(\underline{\mathbf{x}})}{\partial x_i}\Delta x_i \qquad (1)$$

By analyzing the distribution of all m impacts $\frac{\partial f(\underline{\mathbf{x}})}{\partial x_i}\Delta x_i$ and comparing them to Δf_{AB}, we can identify the influential values to be verified, and thus identify the source of a change observed in an estimate between points A and B.

We can use equation (1) to obtain an approximation of the impacts on the relative change $\left(f(B)-f(A)\right)/f(A)$ in the estimate :

$$\frac{\Delta f_{AB}}{f(A)} \cong \sum_{i=1}^{m}\frac{\partial f(\underline{\mathbf{x}})}{\partial x_i}\frac{\Delta x_i}{f(A)} \qquad (2)$$

This method is very useful in examining the difference between two estimates over time (points A and B) using a group of common units. It can also be used to compare two estimates produced with the same estimator and the same statistical units, but with different values (changes in data processing, revised design weights, etc.). For example, we could consider that the points A and B correspond to data before and after imputation. The impacts measured could be used to accurately identify the sources of changes in the estimates after imputation.

2. Application to different estimators

The advantage of the proposed method is that it provides a general simple algorithm to measure the impact of many types of estimators used in different contexts. To explain how it works, we will show how the method is applied to a few common estimators.

2.1 Horvitz-Thompson estimator of a total and a mean

Let the variable of interest be y. The Horvitz-Thompson estimator of total $\hat{t}_{y\pi}$ is given by:

$$\hat{t}_{y\pi}=\sum_{k=1}^{n} w_k\, y_k$$

where the design weight w_k is the inverse of the probability of selection π_k of unit k (i.e., $w_k=\pi_k^{-1}$). Let us assume that this estimator is repeated over two cycles (A and B) using the same group of n units, namely:

$$\hat{t}_{y\pi}(A)=\sum_{k=1}^{n} w_k^A\, y_k^A$$

$$\hat{t}_{y\pi}(B)=\sum_{k=1}^{n} w_k^B\, y_k^B$$

We are trying to measure the impact of the units on the observed difference:

$$\Delta f_{AB}=\hat{t}_{y\pi}(B)-\hat{t}_{y\pi}(A)$$

We can consider the estimate $\hat{t}_{y\pi}$ as a function of $m=2n$ variables. That is,

$$\hat{t}_{y\pi}=f\left(w_1, w_2, ..., w_n, y_1, y_2, ..., y_n\right)$$

The $m=2n$ partial derivatives of the estimator are as follows:

$$\frac{\partial \hat{t}_{y\pi}}{\partial w_k} = y_k$$

$$\frac{\partial \hat{t}_{y\pi}}{\partial y_k} = w_k$$

These partial derivatives evaluated at mid point $C = \underline{x}$ become:

$$\frac{\partial \hat{t}_{y\pi}(\underline{\mathbf{x}})}{\partial w_k} = \underline{y_k} = \frac{y_k^A + y_k^b}{2}$$

$$\frac{\partial \hat{t}_{y\pi}(\underline{\mathbf{x}})}{\partial y_k} = \underline{w_k} = \frac{w_k^A + w_k^b}{2}$$

Thus, we get:

$$\begin{aligned}\Delta f_{AB} &\cong \sum_{i=1}^{m} \frac{\partial f(C)}{\partial x_i} \Delta x_i \\ &= \sum_{k=1}^{n} \frac{\partial \hat{t}_{y\pi}(\underline{\mathbf{x}})}{\partial w_k} \Delta w_k + \sum_{k=1}^{n} \frac{\partial \hat{t}_{y\pi}(\underline{\mathbf{x}})}{\partial y_k} \Delta y_k \\ &= \sum_{k=1}^{n} \underline{y_k} \Delta w_k + \sum_{k=1}^{n} \underline{w_k} \Delta y_k\end{aligned}$$

Considering the m=2n elements of this sum, we can measure the impact of the change of each variable of each unit used in the estimate. If a substantial change is observed between estimates at points A and B, it becomes easy to identify the variable and/or the unit responsible for this change. Note that in this specific case, the selection of point C at the middle of segment $\overline{AB}$ transforms the approximation into an exact equality.

A practical way to analyze the impacts is to place the elements of the sum in a two-dimensional table (Table 1). The right-hand column shows the total impact of unit k on the difference between estimates A and B. The bottom row shows the total contribution by variable. By comparing change $\Delta \hat{t}_{y\pi}$ and the distribution of impacts, it is possible to identify a number of influential values, or to identify the source of the change noticed between $f(A)$ and $f(B)$.

This way, we get the impact for each variable related to unit k ($\underline{y_k} \Delta w_k$ and $\underline{w_k} \Delta y_k$), the total impact for each unit k ($\underline{y_k} \Delta w_k + \underline{w_k} \Delta y_k$) and the total impact of each variable common to the units ($\sum_s \underline{y_k} \Delta w_k$ and $\sum_s \underline{w_k} \Delta y_k$). In general, the sum of the impacts is close to the difference observed, but in this example, it is exactly equal to $\Delta \hat{t}_{y\pi}$.

Table 1
Impact of units and variables on the estimate of a total using the Horvitz-Thompson estimator

Unit	Impact of design weight	Impact of variable of interest	Total impact of unit
1	$\underline{y_1} \Delta w_1$	$\underline{w_1} \Delta y_1$	$\underline{y_1} \Delta w_1 + \underline{w_1} \Delta y_1$
2	$\underline{y_2} \Delta w_2$	$\underline{w_2} \Delta y_2$	$\underline{y_2} \Delta w_2 + \underline{w_2} \Delta y_2$
…	…	…	…
n	$\underline{y_n} \Delta w_n$	$\underline{w_n} \Delta y_n$	$\underline{y_n} \Delta w_n + \underline{w_n} \Delta y_n$
Total	$\sum_{k=1}^{n} \underline{y_k} \Delta w_k$	$\sum_{k=1}^{n} \underline{w_k} \Delta y_k$	$\Delta \hat{t}_{y\pi}$

The Horvitz-Thompson estimator $\tilde{y}_{s\pi}$ of the mean is given by:

$$\tilde{y}_{s\pi} = \frac{\sum_{k=1}^{n} w_k y_k}{\sum_{k=1}^{n} w_k} \qquad (3)$$

The $m = 2n$ partial derivatives of the estimator are:

$$\frac{\partial \tilde{y}_{s\pi}}{\partial w_k} = \left(y_k - \tilde{y}_{s\pi}\right) \left(\sum_{k=1}^{n} w_k \right)^{-1}$$

$$\frac{\partial \tilde{y}_{s\pi}}{\partial y_k} = w_k \left(\sum_{k=1}^{n} w_k \right)^{-1}$$

This gives us:

$$\begin{aligned}\Delta f_{AB} &\cong \sum_{i=1}^{m} \frac{\partial f(C)}{\partial x_i} \Delta x_i \\ &= \sum_{k=1}^{n} \frac{\partial f(\underline{\mathbf{x}})}{\partial w_k} \Delta w_k + \sum_{k=1}^{n} \frac{\partial f(\underline{\mathbf{x}})}{\partial y_k} \Delta y_k \\ &= \frac{\sum_{k=1}^{n} \left(\underline{y_k} - \underline{\tilde{y}_{s\pi}}\right) \Delta w_k + \sum_{k=1}^{n} \underline{w_k} \Delta y_k}{\sum_{k=1}^{n} \underline{w_k}} \qquad (4)\end{aligned}$$

We will illustrate this last result using an example based on a fictitious 4-unit sample, presented in Table 2.

Table 2
Example of the calculation of impacts

Unit k	Design weight w_k		Variable of interest		Impact		
	A	B	A	B	w	y	Total
1	2	1	50	70	-1.2	2.7	1.5
2	2	4	20	10	-5.7	-2.7	-8.4
3	3	6	70	60	5.1	-4.1	1.0
4	4	0	50	50	-1.3	0.0	-1.3
Total					-3.2	-4.1	-7.3

From the sample in Table 2, we can estimate the mean of variable of interest y using the Horvitz-Thompson estimator given by formula (3):

$$\tilde{y}_{s\pi}(A) = 50$$

$$\tilde{y}_{s\pi}(B) = 42.73$$

The difference observed in the estimate of the mean is: 42.73 – 50 = -7.27. The last three columns of Table 2 give impact values measured using formula (4). For instance:

$$IMPACT_{w,1} = \left(\underline{y_1} - \underline{\tilde{y}_{s\pi}}\right)\Delta w_1 \Big/ \sum_{k=1}^{n} \underline{w_k}$$

$$= (60 - 46.4)(1-2)/11 = -1.2$$

Note that in this example, ignoring the rounding effect, the sum of the impacts exactly equals the difference between the two estimates.

By analyzing the impacts, we are able to observe that unit k=2 is primarily responsible for the change in the estimated mean, due mainly to the change in its design weight.

2.2 Classic ratio estimator of a total

Let the variable of interest be y and the auxiliary variable be x. The classic ratio estimator (Särndal, Swensson and Wretman, 1992) of total $\hat{t}_{yra}$ is given by:

$$\hat{t}_{yra} = t_{xU} \frac{\sum_{k=1}^{n} y_k}{\sum_{k=1}^{n} x_k}$$

In this example, only the total $t_{xU} = \sum_{k=1}^{N} x_k$ is known. Estimate $\hat{t}_{yra}$ is a function of $m = 2n + 1$ variables, specifically:

$$\hat{t}_{yra} = f\left(t_{xU}, x_1, x_2, ..., x_n, y_1, y_2, ..., y_n\right)$$

The partial derivatives are:

$$\frac{\partial \hat{t}_{yra}}{\partial t_{xU}} = \sum_{k=1}^{n} y_k \left(\sum_{k=1}^{n} x_k\right)^{-1}$$

$$\frac{\partial \hat{t}_{yra}}{\partial x_k} = -\hat{t}_{yra} \left(\sum_{k=1}^{n} x_k\right)^{-1}$$

$$\frac{\partial \hat{t}_{yra}}{\partial y_k} = t_{xU} \left(\sum_{k=1}^{n} x_k\right)^{-1}$$

The impact formula of the classic ratio estimator of total $\hat{t}_{yra}$ is given by:

$$\Delta \hat{t}_{yra} \cong \frac{\Delta t_{xU} \sum_{k=1}^{n} \underline{y_k} + \sum_{k=1}^{n} \Delta y_k \, \underline{t_{xU}} - \sum_{k=1}^{n} \Delta x_k \, \underline{\hat{t}_{yra}}}{\sum_{k=1}^{n} \underline{x_k}}$$

Assuming that the variable x_k is known for all units of the population, it is possible to measure the impact of x_k for each unit of the population, rather limiting ourselves to the impact of total t_{xU}. This way, we can rewrite the classic ratio estimator of the total as:

$$\hat{t}_{yra2} = \sum_{k=1}^{N} x_k \frac{\sum_{k=1}^{N} \delta_k y_k}{\sum_{k=1}^{N} \delta_k x_k} \quad \text{where } \delta_k = \begin{cases} 1 & \text{If } k \in s \\ 0 & \text{Otherwise} \end{cases}$$

The estimator $\hat{t}_{yra}$ now becomes a function of $m = 3N$ variables, specifically:

$$\hat{t}_{yra} = f\left(x_1, ..., x_N, \delta_1, ..., \delta_N, y_1, ..., y_N\right)$$

The partial derivatives are:

$$\frac{\partial \hat{t}_{yra2}}{\partial x_k} = \left(\hat{t}_y - \delta_k \hat{t}_{yra2}\right)\left(\sum_{k=1}^{N} \delta_k x_k\right)^{-1}$$

$$\frac{\partial \hat{t}_{yra2}}{\partial \delta_k} = \left(y_k t_{xU} - x_k \hat{t}_{yra2}\right)\left(\sum_{k=1}^{N} \delta_k x_k\right)^{-1}$$

$$\frac{\partial \hat{t}_{yra2}}{\partial y_k} = \left(\delta_k t_{xU}\right)\left(\sum_{k=1}^{N} \delta_k x_k\right)^{-1}$$

The impact formula of the classic ratio estimator of total $\hat{t}_{yra2}$ is then given by:

$$\Delta\hat{t}_{yra2} \cong \left(\sum_{k=1}^{N} \underline{\delta_k x_k}\right)^{-1} \left(\sum_{k=1}^{N} \Delta x_k \left(\underline{\hat{t}_y} - \underline{\delta_k \hat{t}_{yra2}}\right) + \sum_{k=1}^{N} \Delta y_k \, \underline{\delta_k t_{xU}} + \sum_{k=1}^{N} \Delta\delta_k \left(\underline{y_k t_{xU}} - \underline{x_k \hat{t}_{yra2}}\right)\right)$$

In this case, we can measure the impact of a change Δx_k in the values of the auxiliary variables for all units of the population, a change Δy_k in the values of variables of interest for the sampled units, and the inclusion of a unit in the sample (represented by $\Delta\delta_k$).

2.3 Selection of explanatory variables

The selection of explanatory variables is the choice of the analyst, depending on the complexity of the estimator and the desired level of details of the impacts.

For example, let us consider the case of a weighted estimator of the total of the variable y where the estimation weight w'_k consists of a design weight w_k, an adjustment for non-response a_k and a calibration factor g_k ($w'_k = w_k a_k g_k$). The impacts can then be measured on w'_k, or on w_k, a_k and g_k. Similarly, if the variable y itself is obtained by the product or the ratio of two lagged variables ($y_k = y_{1k}\, y_{2k}$ or $y_k = y_{1k} / y_{2k}$), it is possible to obtain the impact on y_k, or on y_{1k} and y_{2k}.

Even though it is natural to consider the units as microdata, the method can be used to measure the impacts by cluster, stratum, modeling group, domain, etc. In the case of a repeated survey with independent samples, the method can be used to measure the impact of microdata, even though the distribution of the impacts would be more dispersed. In this case, if we consider microdata analysis unnecessary, it is possible to simply measure the impacts at the stratum, domain or modeling group level. For example, in the case of the Horvitz-Thompson estimator of a total, we could use the following forms to measure respectively the impacts at the level of the n sampled units, the H strata or the D domains:

$$\hat{t}_{\pi y} = \sum_{k=1}^{n} w_k y_k = \sum_{h=1}^{H} \hat{t}_{yh} = \sum_{d=1}^{D} \hat{t}_{yd}$$

3. Impact measurement applied to the Survey of Employment, Payrolls and Hours

Statistics Canada's Survey of Employment, Payrolls and Hours (SEPH) is a monthly survey using two sources of data: a census of administrative data, and a sample survey of establishments called the Business Payrolls Survey (BPS). The purpose of SEPH is to produce estimates of levels and trends in employment, earnings, hours, and other related variables, by province and by industry (Godbout, Grondin and Lavallée, 2005).

The administrative source consists of payroll deduction data provided by the Canada Revenue Agency (CRA). Because these data cannot be linked to a specific province or industry, they must be aggregated at the business level and then disaggregated at the level of establishments, which are linked to a single province or industry. For each establishment k of universe U of size $N \cong 900{,}000$, we get the number of employees E_k and the average monthly earnings per employee x_k.

The survey portion consists of a stratified sample s of about 11,000 establishments, selected from a list frame, which is the Statistics Canada's Business Register (BR). These establishments can be linked to the administrative source. Among the variables collected for each unit $k \in s$, we find the average weekly earnings per employee y_k. The design weight w_k of the unit k in stratum h is $w_k = 1/p_k$, where p_k is the probability of selection of unit k.

The estimator of average weekly earnings for domain d included in modeling group g (or corresponding to modeling group g) is a simple projection estimator based on a linear regression model given by:

$$\tilde{y}_{proj} = \frac{\sum_U \delta_k E_k \mathbf{x}'_k}{\sum_U \delta_k E_k} \hat{\boldsymbol{\beta}}_g = \overline{\mathbf{x}}'_d \hat{\boldsymbol{\beta}}_g$$

where:

$$\delta_k = \begin{cases} 1 & \text{if } k \in d \\ 0 & \text{if } k \notin d \end{cases}$$

$$\mathbf{x}_k = (1, x_k)'$$

$$\overline{\mathbf{x}}_d = \sum_U \delta_k E_k \mathbf{x}_k \Big/ \sum_U \delta_k E_k$$

$$\hat{\boldsymbol{\beta}}_g = (\hat{\beta}_{0,g}, \hat{\beta}_{1,g})'$$

$$= \left(\sum_{s_g} w_k E_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \sum_{s_g} w_k E_k \mathbf{x}_k y_k$$

$$= \hat{\mathbf{T}}^{-1} \hat{\mathbf{t}}$$

The estimator $\tilde{y}_{proj}$ is a function of $m = 3N + n$ dimensions:

$$\tilde{y}_{proj} = f\left(E_1,...E_N, x_1,...x_N, w_1,..., w_N, y_1,...y_n\right)$$

The partial derivatives are:

$$\frac{\partial \tilde{y}_{proj}}{\partial E_k} = \frac{\delta_k\left(\mathbf{x}'_k - \bar{\mathbf{x}}'_d\right)}{\sum_U \delta_k E_k}\hat{\boldsymbol{\beta}}_g + \bar{\mathbf{x}}'_d\hat{\mathbf{T}}^{-1}w_k\mathbf{x}_k\left(y_k - \mathbf{x}'_k\hat{\boldsymbol{\beta}}_g\right)$$

$$\frac{\partial \tilde{y}_{proj}}{\partial x_k} = \frac{\delta_k E_k \mathbf{D}'_{x_k}}{\sum_U \delta_k E_k}\hat{\boldsymbol{\beta}}_g + \bar{\mathbf{x}}'_d\hat{\mathbf{T}}^{-1}w_k E_k\left(\mathbf{D}_{x_k}y_k - (\mathbf{D}_{x_k}\mathbf{x}'_k + \mathbf{x}_k\mathbf{D}'_{x_k})\hat{\boldsymbol{\beta}}_g\right)$$

$$\frac{\partial \tilde{y}_{proj}}{\partial w_k} = \bar{\mathbf{x}}'_d\hat{\mathbf{T}}^{-1}E_k\mathbf{x}_k\left(y_k - \mathbf{x}'_k\hat{\boldsymbol{\beta}}_g\right)$$

$$\frac{\partial \tilde{y}_{proj}}{\partial y_k} = \bar{\mathbf{x}}'_d\hat{\mathbf{T}}^{-1}w_k E_k\mathbf{x}_k$$

where $\mathbf{D}_{x_k} = \frac{\partial \mathbf{x}_k}{\partial x_k} = (0,1)'$.

We then define a score for each unit of the population by taking the absolute value of the impact, weighted by a factor determined by validation rules:

$$SCORE_k = Fact_k\left|\frac{\partial \tilde{y}_{proj}(\underline{\mathbf{x}})}{\partial E_k}\frac{\Delta E_k}{\tilde{y}_{proj}(A)} + \frac{\partial \tilde{y}_{proj}(\underline{\mathbf{x}})}{\partial x_k}\frac{\Delta x_k}{\tilde{y}_{proj}(A)} + \frac{\partial \tilde{y}_{proj}(\underline{\mathbf{x}})}{\partial y_k}\frac{\Delta y_k}{\tilde{y}_{proj}(A)} + \frac{\partial \tilde{y}_{proj}(\underline{\mathbf{x}})}{\partial w_k}\frac{\Delta w_k}{\tilde{y}_{proj}(A)}\right|$$

where $Fact_k = \begin{cases} 4 & \text{if at least 1 rule fails} \\ 1 & \text{Otherwise} \end{cases}$

The next step is to identify the units to be verified by retaining the largest scores by domain, along with all scores exceeding a defined threshold. We use the impacts by variable to identify the elements we want to focus on.

Lastly, for those domains where we observe a change outside the defined interval, we compare this change to the impacts measured from the microdata to verify the main responsible units, correct any errors, and identify the source of these major changes.

Note that we did not include the inclusion variable δ_k among the variables of the function $f(\mathbf{x})$ because we preferred to calculate $\bar{\mathbf{x}}_d$ using a modified version of variable E_k: $E'_k = E_k$ if $k \in U_d$ and $E'_k = 0$ otherwise. However, it would have been possible to include the δ_k and to measure an impact on the estimate arising from this inclusion variable.

4. Conclusion

We have presented a general method for obtaining a formula that can be used to measure the impact of all variables involved when a change is observed in an estimate. This method can be applied to several types of estimators in different contexts. The selection of explanatory variables and of the level of the units is the choice of the analyst.

Once the impacts are determined, it is possible to apply various common univariate methods (sigma gap, interquartile differences, etc.) to identify the influencing units. Similarly, when analyzing the estimates, the impacts measured can be used to explain the source of the observed changes.

It should however be noted that the formula remains an approximation and that the sum of the impacts can be different from the observed change, even if, in some cases (as with the Horvitz-Thompson estimators of the total and the mean), the selection of mid point $C = \underline{\mathbf{x}}$ gives a sum of the impacts exactly equal to the observed change. The estimator used to produce the two estimates must be the same. In addition, the function of the estimator must be differentiable, which excludes some estimates such as percentiles.

5. Acknowledgements

I would like to thank the editors and the team of SEPH methodologists (Yanick Beaucage, Lisa Corscadden, Anne-Marie Houle, Édith Hovington, Yves Morin, Sharon Wirth).

6. Bibliography

Cook R.D. (1977). Detection of Influential Observation in Linear Regression, Technometrics, Vol. 19, No. 1.

Franklin S. et al. (2000). Robust Multivariate Outlier Detection Using Mahalanobis Distance and Modified Stahel-Donoho Estimators, the Second International Conference on Establishment Surveys, pp. 697-706, Ottawa, Canada : Statistics Canada.

Godbout S., Grondin C. and Lavallée P. (2005). Current Methodology of the Survey of

Employment, Payrolls and Hours, Ottawa, Canada : Statistics Canada.

Hidiroglou M.A. and Berthelot J.-M. (1986). Statistical Editing and Imputation for Periodic Business Surveys, Survey Methodology, Vol. 12, No. 1, pp. 73-83, Ottawa, Canada : Statistics Canada

Marsden J.E. and Tromba A.J. (1988). Vector Calculus, Third Edition, W.H. Freeman and Company, New York, 655 pages

Rudin W. (1976). Principles of Mathematical Analysis, McGraw-Hill, New York, 342 pages

Särndal C.-E., Swensson B. and Wretman J. (1992). Model Assisted Survey Sampling, Springler-Verlag, New York, 694 pages