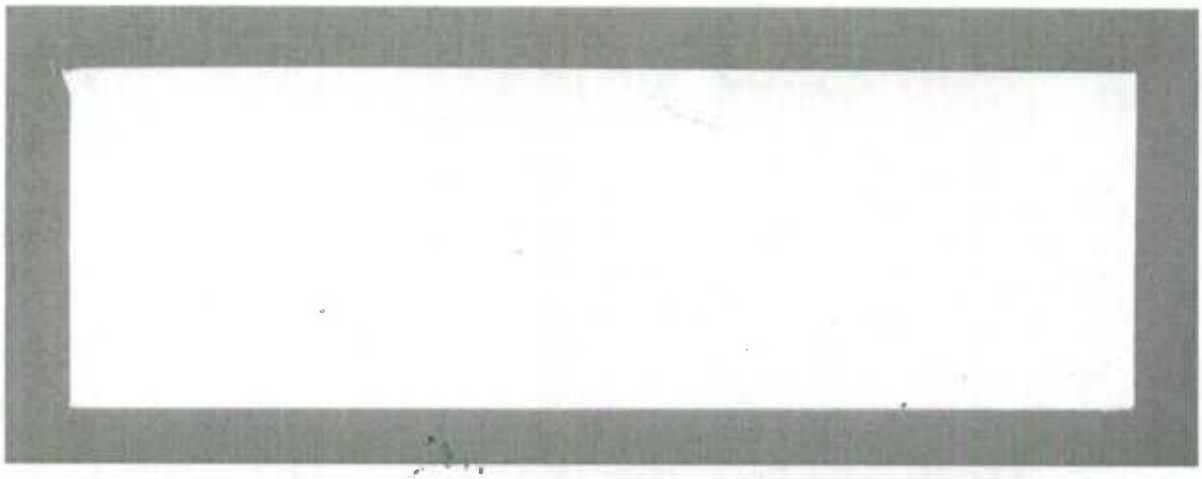
Statistics Canada Statistique Canada



# Methodology Branch

Business Survey Methods Division

# Direction de la méthodologie

Division des méthodes d'enquêtes auprès des entreprises

Canada

# APPLICATIONS OF VARIANCE DUE TO IMPUTATION IN THE SURVEY OF EMPLOYMENT, PAYROLLS AND HOURS

By

Pierre Felx and Eric Rancourt

BSMD-2001-009E

WORKING PAPER
METHODOLOGY BRANCH

# APPLICATIONS OF VARIANCE DUE TO IMPUTATION IN THE SURVEY OF EMPLOYMENT, PAYROLLS AND HOURS

BSMD - 2001 - 009E

Pierre Felx and Eric Rancourt

Business Survey Methods Division
Statistics Canada

September 2001

---

The work presented in this paper is the responsibility of the authors and does not necessarily represent the views or policies of Statistics Canada.

# Applications de la variance due à l'imputation dans l'Enquête sur l'emploi, la rémunération et les heures

Pierre Felx[1] and Eric Rancourt[2]

## RÉSUMÉ

L'imputation est une méthode bien connue de traitement de la non-réponse dans les enquêtes mensuelles auprès des entreprises. Il est aussi bien connu que la formule de variance sous-estime la variance lorsque l'imputation est utilisée. En conséquence, le niveau de variation due à l'imputation devient un calcul nécessaire. Ceci est particulièrement important lorsqu'on requiert des estimations pour des domaines arbitraires. Dans cette situation, plusieurs méthodes ont été proposées pour calculer le niveau de variation due à l'imputation, et ce pour plusieurs méthodes d'imputation. Dans cet article, nous étudions l'approche assistée d'un modèle sous l'échantillonnage aléatoire simple stratifié pour l'imputation par la moyenne, la tendance et le ratio. Les résultats obtenus sont utilisés dans le cas de l'Enquête sur l'emploi, la rémunération et les heures pour mieux estimer la variance totale et pour aider à déterminer s'il y a eu des brisures dans la série d'estimations entre deux phases du remaniement de l'enquête.

**Mots clefs**: Approche assistée d'un modèle; brisures dans la série; imputation simple; phases du remaniement; variance due à l'imputation.

[1] Pierre Felx, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada.
[2] Eric Rancourt, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada.

# Applications of Variance Due to Imputation in the Survey of Employment, Payrolls and Hours

Pierre Felx[1] and Eric Rancourt[2]

## ABSTRACT

Imputation is a well-known method for dealing with nonresponse in periodic business surveys conducted on a monthly basis. It is also well known that the ordinary variance formula underestimates the variance when imputation is used. As a direct result, the amount of variation due to imputation becomes a necessary calculation. This is especially important when estimates are required for arbitrary domains. Methods have been proposed for calculating the amount of variation due to imputation for arbitrary domains, for many different types of imputation, for a simple random sample without replacement. In this paper, we study the model-assisted approach for stratified simple random sampling when mean, trend and ratio imputation is used. The results obtained are applied to the Canadian Survey of Employment, Payrolls and Hours to better estimate the total variance, and to help determine whether there may be breaks between two redesign phases of the survey.

**Key Words**: Breaks in the series; Model-assisted approach; Redesign phases; Single imputation; Variance due to imputation.

[1] Pierre Felx, Business Survey Methods Division, Statistics Canada.
[2] Eric Rancourt, Household Survey Methods Division, Statistics Canada.

## 1. INTRODUCTION

Since periodic business surveys are used for many purposes including the System of National Accounts, it is important that the estimates obtained be of high quality. It is also of high importance that the users not be misled as to the quality of the data.

Nonresponse is a key factor in assessing the quality of the data for all surveys that experience this problem. In many surveys the approach to deal with nonresponse is to use imputation, while for others, re-weighting may be the method of choice. In some surveys, a mixture of both imputation and re-weighting may be used.

In the Canadian Survey of Employment, Payrolls and Hours (SEPH) the method for dealing with nonresponse is imputation. In determining the overall quality of the data in SEPH it is important not to treat imputed values as respondents since this is known to lead to a significant underestimate of the variance. See Lee, Rancourt and Särndal (1994).

It is now possible to take imputation into account and give a better estimate of the true variance and thus allow the users to have a more precise measure of the quality of the data.

Several other reasons exist for the calculation of the variance due to imputation. In SEPH for instance, it served as a diagnostic to allow for the determination of whether changing the imputation strategy would be more efficient. Also, it helped in determining why there were changes in the data between the old and the new design when SEPH underwent a major redesign. If imputation was ruled out as causing changes in the data then time could be spent in other areas of research.

There are several tools to correctly estimate the precision of estimates when imputation has been used; see Lee, Rancourt and Särndal (2001). The first method developed was multiple imputation by Rubin (1978), but it is by definition not designed for single imputation, which is the case in SEPH. If nonresponse is seen as a second

phase of a two-phase sampling strategy, then the two-phase theory can be applied to obtain the variance due to imputation. This is the variance estimation approach presented in Rao and Sitter (1995). This method is highly sensitive to the assumption that nonresponse is randomly distributed. Resampling methods have also been adapted for variance estimation under imputation. It is the case of the jackknife technique by Rao and Shao (1992), the bootstrap by Shao and Sitter (1996) and the BRR by Rao and Shao (1996). These methods provide an account of the total variance but do not explicitly separate between the sampling variance and the imputation variance, which is one of the goals in SEPH.

The method retained was the model-assisted approach by Särndal (1992). It provides separate estimates of the sampling and imputation variances and is compatible with Statistics Canada's Generalized Estimation System (GES). Further, since the relationships between variables are very strong in SEPH, the model-assisted approach (which does not require that the response mechanism be uniform) is expected to yield good results as shown in Lee, Rancourt and Särndal (1994).

In this paper we present applications of methods to deal with the impact of imputation. As mentioned above, the model-assisted approach is used in all applications. In section 2, we introduce SEPH and its methodology. In section 3, variance due to imputation is presented in more detail. In section 4, applications of variance due to imputation are presented for SEPH along with some results. In section 5, we give a brief discussion of the gains in SEPH from using variance due to imputation. Also in section 5, we give a summary of the results and the work that remains to be completed.

## 2. THE CANADIAN SURVEY OF EMPLOYMENT, PAYROLLS AND HOURS

The Canadian Survey of Employment, Payrolls and Hours (SEPH) was designed to provide estimates of employment, payrolls, working hours, overtime pay and hours, summarized earnings and other related variables. SEPH is a monthly establishment survey and covers all industries in Canada with the exception of agriculture, fishing and trapping, private household services, religious organizations and military services.

Originally SEPH used solely establishment data to provide monthly estimates at the three digit Standard Industrial Classification (SIC) levels for Canada, and the provinces for the variables listed previously (employment, payroll, etc.). In recent years SEPH has undergone a major redesign, involving three phases, to incorporate the use of administrative data for the reduction of costs and the improvement of estimates.

In the final two phases of the redesign, calculation of employment and payroll is obtained directly from the administrative data while estimation of the other variables is obtained by using regression models on the establishment data and by applying these to the administrative data. Such regression models are used to calculate hours and summarized earnings as well as the other variables. See Hurtubise et al. (2000).

Since the results in this paper are presented only for administrative data, and in that, only the variables employment and payroll, the remaining focus will be on the last two phases of the redesign. More on the redesign can be found in Rancourt and Hidiroglou (1998).

The administrative data itself is received from the Canadian Customs and Revenue Agency (CCRA) on a monthly basis in the form of Payroll Deduction Accounts (PD). There are two types of PD accounts: automatic (PD7A) and twice monthly (PD7TM). PD7As remit once per month to CCRA and whose average monthly remittances in the previous year were less than $15,000. PD7As are selected into the sample using a variation of Bernoulli sampling. Since the sample is selected as Bernoulli trials with probability $\pi_k$ and are post stratified based on size we estimate totals using a post-stratified estimator. PD7TMs are accounts which remit more than once a month and whose average monthly remittances in the previous year were less than $15,000. These units are selected into the sample with probability one.

Nonresponse for the administrative sample approaches the magnitude of 45%. Approximately 25% of this data is considered not to be non-respondent data but rather observations for which the variables employment and payroll are zero. This is determined by using auxiliary information. Imputation of the remaining 20% is

performed using three main methods: trend, ratio and mean imputation. There are several minor methods that are in fact spin-offs of the three main ones.

SEPH uses the variance due to imputation to assess the impact of these different imputation methods and also to give a more realistic measure of data quality.

## 3. VARIANCE DUE TO IMPUTATION

As mentioned in the introduction, the model assisted approach has been chosen to estimate the variance due to imputation in SEPH. The approach allows for the possibility of obtaining separate figures for the variance due to sampling and that due to imputation. This was first laid out in Särndal (1990, 1992) and then in Deville and Särndal (1994).

Let $y_k$ be the variable of interest for unit $k$. The parameter to estimate is

$$Y_U = \sum_U y_k$$

where $U = \{1,...,N\}$ is the population. In the case of complete response, the estimator of $Y_U$, based on $s$ of size $n$, is

$$\hat{Y}_s = \sum_g \sum_p \sum_s w_k y_k,$$

where $g$ is the region, $p$ is the imputation class and $w_k$ is the sampling weight.

To represent the data, a model can be used. In this paper, a simple linear model is adopted. The model is:

$$\xi : y_k = \beta z_k + \varepsilon_k, \text{ where } E_\xi(\varepsilon_k) = 0 \text{ and } E_\xi(\varepsilon_k^2) = \sigma^2 z_k,$$

where $z_k$ is an auxiliary variable available for both respondents and nonrespondents.

When there is nonresponse, the sample is composed of two subsets, $r$ the response set, and $o$ the nonresponse set. The missing values are replaced by imputed values according to one of the methods outlined in Section 2. To denote the data after imputation, $\hat{y}_k$, we use:

$$y_{\bullet k} = \begin{cases} y_k & \text{if } k \in r \\ \hat{y}_k & \text{if } k \in o. \end{cases}$$

For each of the imputation methods ( $I$ ) used in SEPH, we have $\hat{y}_k = \hat{B} z_k$, where

$\hat{B} = \bar{y}_r$, and $z_k = 1$ for Mean Imputation

$\hat{B} = \dfrac{\bar{y}_r}{\bar{z}_r}$, and $z_k = z_k$ for Ratio Imputation

$\hat{B} = \dfrac{\bar{y}_{t,r}}{\bar{y}_{t-1,r}}$, and $z_k = y_{t-1,k}$ for Trend Imputation

$\hat{B} = 1$, and $z_k = z_k$ for Carry over Imputation

$\hat{B} = \dfrac{z_{1,k}}{z_{2,k}}$, and $z_k = 1$ for Single Ratio Imputation.

In the presence of imputation, we are interested in assessing the error

$$\hat{Y}_{\bullet s} - Y_U = \left(\hat{Y}_s - Y_U\right) + \left(\hat{Y}_{\bullet s} - \hat{Y}_s\right),$$

that is, the sampling error and the imputation error where

$$\hat{Y}_{\bullet s} = \sum_g \sum_p \sum_s w_k y_{\bullet k} .$$

If the bias is null, then to estimate the variance, it is only necessary to evaluate the model expectation of

$$\left(\hat{Y}_{\bullet s} - Y_U\right)^2 = \left(\hat{Y}_s - Y_U\right)^2 + \left(\hat{Y}_{\bullet s} - \hat{Y}_s\right)^2 + 2\left(\hat{Y}_s - Y_U\right)\left(\hat{Y}_{\bullet s} - \hat{Y}_s\right)$$

which then corresponds to

$$V_{\mathrm{TOT}} = V_{\mathrm{SAM}} + V_{\mathrm{IMP}} + 2V_{\mathrm{MIX}}.$$

For a particular domain of interest $d$,

$$V_{\mathrm{TOT}}(d) = V_{\mathrm{SAM}}(d) + V_{\mathrm{IMP}}(d) + 2V_{\mathrm{MIX}}(d).$$

For estimating $V_{\mathrm{SAM}}(d)$, we use $\hat{V}_{\mathrm{ORD}}(d) + \hat{V}_{\mathrm{DIF}}(d)$, where

$$\hat{V}_{ORD}(d) = \sum_g \left(1 - \frac{n_g}{N_g}\right) \sum_p \frac{N_{pg}^2}{n_{pg}} \left(\frac{1}{n_{pg} - 1}\right) \sum_s \left\{ y_{\bullet k}(d) - \left( \frac{\sum y_{\bullet k}(d)}{n_{pg}} \right) \right\}^2$$

with

$$y_{\bullet k}(d) = \begin{cases} y_{\bullet k} & \text{if } \ k \in d \\ 0 & \text{otherwise.} \end{cases}$$

See Särndal, Swensson and Wretman (1992) for more details on domain estimation. Then the model assisted approach provides a method for estimating each component by using model $\xi$.

Letting the superscript $I$ denote the imputation method used, for ratio, trend and mean imputation, the formulae are:

$$\hat{V}_{DIF}^{I}(d) = \sum_g \left(1 - \frac{n_g}{N_g}\right) \sum_p \frac{N_{pg}^2}{n_{pg}^2} \sum_{k \in o^I} z_{kpg}^I(d) \hat{\sigma}_{pg}^{I\,2}$$

$$\hat{V}_{IMP}^{I}(d) = \sum_g \sum_p \frac{N_{pg}^2}{n_{pg}^2} \sum_{k \in o^I} z_{kpg}^I(d) \left( \frac{\sum_{k \in o^I} z_{kpg}^I(d)}{\sum_{k \in r} z_{kpg}^I} + 1 \right) \hat{\sigma}_{pg}^{I\,2}$$

$$\hat{V}^{I}_{MIX}(d) = \sum_{g}\left(1-\frac{n_g}{N_g}\right)\sum_{p}\left(\frac{N^2_{pg}}{n^2_{pg}}\left\{\sum_{k\in o^I} z^I_{kpg}(d)\left(\frac{\sum_{k\in r} z^I_{kpg}(d)}{\sum_{k\in r} z^I_{kpg}}+1\right)\hat{\sigma}^{I\ 2}_{pg}\right\}\right)$$

with

$$\hat{\sigma}^{I\ 2}_{pg} = \frac{\sum_{k\in r} e^{I\,2}_k}{\sum_{k\in r} z^I_k};\quad e^I_k = y_k - \hat{\beta}^I_r z^I_k.$$

$o^I$ is the set of non-respondents imputed using method $I$. Note that the response set $r$ does not depend on $I$.

For carry over and single ratio imputation, the formulae are:

$$\hat{V}^I_{DIF}(d) = \sum_g\left(1-\frac{n_g}{N_g}\right)\sum_p \frac{N^2_{pg}}{n^2_{pg}}\sum_{k\in o^I} z^I_{kpg}(d)\hat{\sigma}^{I\ 2}_{pg}$$

$$\hat{V}^I_{IMP}(d) = \sum_g\sum_p \frac{N^2_{pg}}{n^2_{pg}}\left\{\sum_{k\in o^I} z^I_{kpg}(d)\hat{\sigma}^{I\ 2}_{pg}\right\}$$

$$\hat{V}^I_{MIX}(d) = \sum_g\left(1-\frac{n_g}{N_g}\right)\sum_p \frac{N^2_{pg}}{n^2_{pg}}\sum_{k\in o^I} z^I_{kpg}(d)\hat{\sigma}^{I\ 2}_{pg}$$

with

$$\hat{\sigma}^{I\ 2}_{pg} = \frac{\sum_{k\in r} e^{I\,2}_k}{\sum_{k\in r} z^I_k};\quad e^I_k = y_k - \hat{\beta}^I_r z^I_k.$$

For more than one imputation method we sum across the $I$,

$$\hat{V}_{DIF}(d) = \sum_I \hat{V}^I_{DIF}(d)$$

and similarly for $\hat{V}_{IMP}(d)$ and $\hat{V}_{MIX}(d)$.

In all cases, only non-respondents imputed by each method are used, all others are set to zero.

## 4. APPLICATIONS IN SEPH

As SEPH has recently undergone the final of three major phases of a redesign, it has become very important to monitor the overall data quality. Not only is it considered important to monitor the quality of the current data but also to assess how the quality of the current data compares to that in previous phases and that of possible future phases (or occasions).

There were many changes in each phase of the redesign so it has become important to isolate the causes of changes and look at them individually. In this paper, concern is placed in the methodological changes in imputation and areas that have an impact on imputation.

The question of how to assess the impact of changing imputation strategies was raised. Evaluation of the variance due to imputation was determined as being a good method in determining the impact of imputation. The model-assisted approach by Särndal (1992) allows for the calculation of variance due to imputation separately from the variance due to sampling.

There were three main areas where the calculation of the variance due to imputation was helpful for analysis. Firstly, it was useful in determining the possible existence of breaks in a data series. Secondly, it allowed for the determination of whether a change in imputation could or would be an improvement. Finally, the calculation of the variance due to imputation allows SEPH to give users a more realistic estimate of data quality. In this study we concentrate on the first and second issues, as the third is a direct result of the first two.

The redesign of SEPH started in 1995. Phase II had been in place from May 1996 until April 1998 when the change to phase III occurred. Because phase II had been in place for an extended period of time it was deemed more important to concentrate on the changes between phase II and phase III since phase II levels had been accepted as being of good quality.

The major change between phase II and phase III with regards to imputation was the introduction of forced records. Forced records are observations that are forced into the sample and thus carry a weight of one. Several of these forced records were selected randomly in phase II and contributed highly to the variance due to imputation. The effect of changing these records from being randomly selected to being forced into the sample should have a positive effect on the variance due to imputation, that is, lowering the variance and thus providing higher quality data.

Previously in determining the possible existence of a break in the data series the sampling variance was calculated for each of the consecutive months and a confidence interval was calculated. The covariance was assumed to be negligible. Introducing the variance due to imputation allowed the calculation of a new confidence interval, one including the variance due to imputation.

Differences in the data series could be explained now being due to sampling and/or imputation. Previously only sampling could be used to eliminate data differences as being a break.

To determine whether a break exists, confidence intervals were calculated in the last month of phase II and compared with those calculated in the first month of phase III. In fact, what is of interest is to determine whether $\hat{Y}_{III} - \hat{Y}_{II} = 0$, where $\hat{Y}_{III}$ is the phase III estimate of employment (or payroll) and $\hat{Y}_{II}$ is the phase II estimate. The variance of the difference is

$$V\left(\hat{Y}_{III} - \hat{Y}_{II}\right) = V\left(\hat{Y}_{III}\right) - V\left(\hat{Y}_{II}\right) - 2Cov\left(\hat{Y}_{III}, \hat{Y}_{II}\right),$$

and as seen in section 3,

$$V\left(\hat{Y}_{III}\right)=V_{SAM}\left(\hat{Y}_{III}\right)+V_{DIF}\left(\hat{Y}_{II}\right)+2V_{MIX}\left(\hat{Y}_{II}\right)$$

and

$$V\left(\hat{Y}_{III}\right)=V_{SAM}\left(\hat{Y}_{III}\right)+V_{DIF}\left(\hat{Y}_{II}\right)+2V_{MIX}\left(\hat{Y}_{II}\right)$$

As $Cov\left(\hat{Y}_{III},\hat{Y}_{II}\right)$ is likely to be greater than zero, it is also likely that

$$V\left(\hat{Y}_{III}\right)+V\left(\hat{Y}_{II}\right)>V\left(\hat{Y}_{III}-\hat{Y}_{II}\right).$$

Since we wanted to be conservative on the number of declared breaks and since sampling and imputation are but two sources of variability, we simply compound intervals built around $\hat{Y}_{III}$ and $\hat{Y}_{II}$ (thereby implicitly assuming that $Cov\left(\hat{Y}_{III},\hat{Y}_{II}\right)=0$). See Schenker and Gentleman, (2001) for a detailed discussion.

Also, the derivation of $Cov\left(\hat{Y}_{III},\hat{Y}_{II}\right)$ is complex since it involves several terms. It is also likely to be greater than 0. This is consistent with the positive correlation between phase III estimates ($\hat{Y}_{III}$) and phase II estimates ($\hat{Y}_{II}$). In the case of SEPH the correlation coefficient on SIC3, province estimates for employment between $\hat{Y}_{II}$ and $\hat{Y}_{III}$ is 0.83.

To determine if a break exists we proceeded as follows.

If the confidence intervals fail to overlap then this is considered as a break. Phase II and phase III data are compared at the SIC3, province level to determine if a break exists. There is also the chance that a break may exist if both confidence intervals overlap but neither confidence interval covers the others midpoint. There are three levels of severity for a break.

I. The confidence intervals do not overlap at any point. This is the most severe and is considered a definite break.

II. The confidence intervals overlap but neither midpoint is covered by the other confidence interval. This is considered as having a high chance of being a break.

III. One confidence interval covers the others midpoint but both midpoints are not covered. This is considered as having a low chance of being a break.

If both confidence intervals cover the others midpoint then this is not considered a break.

Note that, as pointed out in Schenker and Gentleman (2001), it would not be desirable to use only levels II and III or level III alone. One should note that although there is the existence of a break it does not mean that there is no reason for this. It could be due to a significant change in the data that is completely justified. There may have been a large increase in employment or payroll between the two months due to a large number of "real" job gains or loses in society.

In Table 1 and 2, we see the distribution of breaks for both employment and payroll as they occurred prior to the addition of the variance due to imputation.

**Table 1:** Employment

| **Possibility of a Break** | **Frequency (#)** | **Percentage (%)** |
|---|---|---|
| Definite | 844 | 29.2 |
| High | 159 | 5.5 |
| Low | 579 | 20.0 |
| None | 1309 | 45.3 |
| TOTAL | 2891 | 100.0 |

**Table 2:** Payroll

| **Possibility of a Break** | **Frequency (#)** | **Percentage (%)** |
|---|---|---|
| Definite | 1027 | 35.5 |
| High | 159 | 5.5 |
| Low | 611 | 21.1 |
| None | 1094 | 37.8 |
| TOTAL | 2891 | 100.0 |

After including the variance due to imputation into the confidence intervals we find the following breakdown.

**Table 3:** Employment

| Possibility of a Break | Frequency (#) | Percentage (%) |
|---|---|---|
| Definite | 611 | 21.1 |
| High | 172 | 5.9 |
| Low | 645 | 22.3 |
| None | 1463 | 50.6 |
| TOTAL | 2891 | 100.0 |

**Table 4:** Payroll

| Possibility of a Break | Frequency (#) | Percentage (%) |
|---|---|---|
| Definite | 728 | 25.2 |
| High | 174 | 6.0 |
| Low | 677 | 23.4 |
| None | 1312 | 45.4 |
| TOTAL | 2891 | 100.0 |

In employment we see there is a decrease of over 200 SIC3, province combinations being considered as a definite break. In payroll the number is even larger with just under 300 SIC3, province combinations changing from a definite break to something other than a definite break.

Considering the variance due to imputation allows for the possibility of more in-depth study of the remaining definite breaks since there are a lot fewer to contend with. Overall this leads to better quality data and a better impression of the true change in quality of the data.

Recently, SEPH has changed from the SIC categories to the new NAICS codes. This change in industry class structure effects the classes and inherently, how imputation is performed. Also, the use of an auxiliary variable containing monthly remittances will be used to improve the current methodology.

Before these changes were adopted it was important to assess whether the data resulting from the changes was going to improve data quality. Again, it was felt that the calculation of variance due to imputation could be a great help in the assessment of

change in data quality and the determination of whether the use of remittances and new imputation classes would provide better quality data.

Table 5 shows the variance due to imputation at the province level for employment for PD7As prior to the introduction of the new imputation methods for remittances and the new NAICS classification. Table 6 shows the variance due to imputation after the introduction of the new imputation methods for remittances and the new NAICS classification.

**Table 5:** PD7A Employment (000s)

| PROV | $V_{IMP}$ | $V_{DIF}$ | $V_{MIX}$ |
|---|---|---|---|
| 10 | 222 | 152 | 178 |
| 11 | 87 | 60 | 66 |
| 12 | 376 | 270 | 347 |
| 13 | 289 | 209 | 264 |
| 24 | 3,433 | 2,604 | 4,622 |
| 35 | 9,129 | 6,507 | 12,012 |
| 46 | 219 | 143 | 207 |
| 47 | 292 | 187 | 267 |
| 48 | 1,354 | 910 | 1,700 |
| 59 | 2,669 | 1,988 | 3,721 |
| 60 | 7 | 0 | 0 |
| 61 | 7 | 0 | 0 |

**Table 6:** PD7A Employment (000s)

| PROV | $V_{IMP}$ | $V_{DIF}$ | $V_{MIX}$ |
|---|---|---|---|
| 10 | 142 | 56 | 111 |
| 11 | 71 | 41 | 81 |
| 12 | 266 | 134 | 269 |
| 13 | 162 | 90 | 180 |
| 24 | 3,077 | 2,215 | 4,429 |
| 35 | 4,631 | 3,044 | 6,088 |
| 46 | 131 | 75 | 149 |
| 47 | 132 | 69 | 137 |
| 48 | 1,205 | 626 | 1,252 |
| 59 | 1,523 | 2,001 | 2,002 |
| 60 | 2 | 0 | 0 |
| 61 | 3 | 0 | 0 |

These tables show that there is a dramatic reduction in variance due to imputation proving that the use of another auxiliary variable, remittances, is highly useful in SEPH. More results can be found in Appendix A for variance due to imputation for payroll of PD7As and employment and payroll of PD7TMs.

## 5. CONCLUSION

There are several advantages to using the total variance including that due to imputation versus using only the ordinary variance. Some of these include a more detailed breakdown of variance and more precise estimates of the true variance leading to more precise knowledge about the reliability of the data.

In SEPH it has proven to be a useful tool to assess the magnitude of a change in imputation methods and also to verify the data when changes in imputation have been performed.

Because of the dramatic reductions in variance due to imputation from using the auxiliary variable remittances in imputation is has been decided to incorporate these changes. The introduction of remittances into SEPH improves micro data and also global estimates. With these changes the consistency and reliability in SEPH data are increased and hopefully lead to better acceptance by the users.

Performing data analysis using variance due to imputation allowed for more efficient use of man power to verify "breaks" in the data. It not only eliminates causes of "breaks" but also helps in pointing out areas where "breaks" may occur.

As a result of the complexity of the imputation in SEPH, several assumptions were made to reach the conclusions mentioned in the paper. One major assumption was that imputation classes were assumed to have been at the same level without a hierarchy of class collapsing. A middle level was chosen to minimize the bias caused by this assumption.

Based on the experience of the SEPH survey, it is recommended that surveys start incorporating variance due to imputation as part of their variance estimation process and as part of their assessment of methodological changes in imputation.

## 6. REFERENCES


Deville, J.-C. and Särndal, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thomson estimator. *Journal of Official Statistics*, 10, 381-394.

Hurtubise, D., Morin, Y., Lavallée, P. and Hidiroglou, M. (2000). Variance estimation for synthetic estimators in the context of an establishment survey. *Proceeding of the Second International Conference on Establishment Surveys*.

Lee, H., Rancourt, E. and Särndal, C.-E. (1994). Experiment with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.

Lee, H., Rancourt, E. and Särndal, C.-E. (2000). Variance estimation from survey data under single value imputation. *Working paper* HSMD-2000-006E, Statistics Canada.

Lee, H., Rancourt, E. and Särndal, C.-E. (2001). Variance estimation from survey data under single value imputation. In *Survey Nonresponse,* Groves, R., Dillman, D., Eltinge, J. and Little, R. eds. J. Wiley, to appear.

Rao, J.N.K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot-deck imputation. *Biometrika*, 79, 811-822.

Rao, J.N.K. and Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

Rancourt, E. and Hidiroglou, M.A. (1998). Use of administrative records in the Canadian Survey of Employment, Payrolls and Hours. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 39-47.

Rubin, D.B. (1978). Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 20-34.

Särndal, C.-E. (1990). Methods for estimating the precision of survey estimates when imputation has been used. *Proceedings of Statistics Canada Symposium 90: Measurement and improvement of data quality*, 337-347.

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241-252.

Schenker, N. and Gentleman, J.F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. The American Statistician, 55.3, 182-186.

Shao, J., Chen, Y. and Chen, Y. (1998). Balanced repeated replication for stratified multistage survey data under imputation. *Journal of the American Statistical Association*, 93, 819-831.

Shao, J. and Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.

**APPENDIX A:**

PD7As variance due to imputation for payroll:

**Table A1: SIC (000,000,000s)**

| PROV | $V_{IMP}$ | $V_{DIF}$ | $V_{MIX}$ |
|---|---|---|---|
| 10 | 1,030 | 695 | 827 |
| 11 | 577 | 382 | 415 |
| 12 | 1,782 | 1,206 | 1,667 |
| 13 | 1,653 | 1,136 | 1,463 |
| 24 | 16,178 | 11,492 | 21,077 |
| 35 | 63,453 | 34,257 | 66,784 |
| 46 | 1,135 | 709 | 1,053 |
| 47 | 1,194 | 764 | 1,143 |
| 48 | 8,911 | 5,766 | 11,341 |
| 59 | 15,154 | 11,694 | 22,749 |
| 60 | 24 | 0 | 0 |
| 61 | 38 | 0 | 0 |

**Table A2: NAICS (000,000,000s)**

| PROV | $V_{IMP}$ | $V_{DIF}$ | $V_{MIX}$ |
|---|---|---|---|
| 10 | 419 | 143 | 286 |
| 11 | 109 | 42 | 85 |
| 12 | 1,270 | 611 | 1,222 |
| 13 | 2,719 | 552 | 1,103 |
| 24 | 10,533 | 7,609 | 15,218 |
| 35 | 26,802 | 14,132 | 28,265 |
| 46 | 994 | 431 | 862 |
| 47 | 556 | 287 | 573 |
| 48 | 4,883 | 2,351 | 4,703 |
| 59 | 8,207 | 4,790 | 9,581 |
| 60 | 10 | 0 | 0 |
| 61 | 58,638 | 0 | 0 |

**APPENDIX A (Continued):**

PD7TMs variance due to imputation for employment:

**Table A3: SIC (000s)**

| PROV | $V_{IMP}$ | $V_{DIF}$ | $V_{MIX}$ |
|---|---|---|---|
| 10 | 1,524 | 0 | 0 |
| 11 | 632 | 0 | 0 |
| 12 | 473 | 0 | 0 |
| 13 | 201 | 0 | 0 |
| 24 | 13,219 | 0 | 0 |
| 35 | 90,804 | 0 | 0 |
| 46 | 1,557 | 0 | 0 |
| 47 | 3,634 | 0 | 0 |
| 48 | 3,217 | 0 | 0 |
| 59 | 6,229 | 0 | 0 |
| 60 | 5 | 0 | 0 |
| 61 | 8 | 0 | 0 |

**Table A4: NAICS (000s)**

| PROV | $V_{IMP}$ | $V_{DIF}$ | $V_{MIX}$ |
|---|---|---|---|
| 10 | 80 | 0 | 0 |
| 11 | 21 | 0 | 0 |
| 12 | 116 | 0 | 0 |
| 13 | 23 | 0 | 0 |
| 24 | 8,643 | 0 | 0 |
| 35 | 26,551 | 0 | 0 |
| 46 | 304 | 0 | 0 |
| 47 | 872 | 0 | 0 |
| 48 | 828 | 0 | 0 |
| 59 | 1,639 | 0 | 0 |
| 60 | 0.04 | 0 | 0 |
| 61 | 0.56 | 0 | 0 |

**APPENDIX A (Continued):**

PD7TMs variance due to imputation for payroll:

**Table A5: SIC (000,000,000s)**

| PROV | $V_{IMP}$ | $V_{DIF}$ | $V_{MIX}$ |
|---|---|---|---|
| 10 | 19,441 | 0 | 0 |
| 11 | 716 | 0 | 0 |
| 12 | 22,313 | 0 | 0 |
| 13 | 4,793 | 0 | 0 |
| 24 | 352,610 | 0 | 0 |
| 35 | 715,933 | 0 | 0 |
| 46 | 16,476 | 0 | 0 |
| 47 | 45,975 | 0 | 0 |
| 48 | 134,075 | 0 | 0 |
| 59 | 50,430 | 0 | 0 |
| 60 | 28 | 0 | 0 |
| 61 | 273 | 0 | 0 |

**Table A6: NAICS (000,000,000s)**

| PROV | $V_{IMP}$ | $V_{DIF}$ | $V_{MIX}$ |
|---|---|---|---|
| 10 | 12,090 | 0 | 0 |
| 11 | 23 | 0 | 0 |
| 12 | 26,625 | 0 | 0 |
| 13 | 293 | 0 | 0 |
| 24 | 226,984 | 0 | 0 |
| 35 | 983,175 | 0 | 0 |
| 46 | 10,025 | 0 | 0 |
| 47 | 12,054 | 0 | 0 |
| 48 | 85,754 | 0 | 0 |
| 59 | 45,754 | 0 | 0 |
| 60 | 0.25 | 0 | 0 |
| 61 | 333 | 0 | 0 |