Statistics Canada Statistique Canada

Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes-auprès des entreprises

Canada

WORKING PAPER NO. BSMD-2007-009E
METHODOLOGY BRANCH

CAHIER DE TRAVAIL NO. DMEE-2007-009E
DIRECTION DE LA MÉTHODOLOGIE

Methodology for the Processing and Imputation of Corporations Data T2

by

Jessica Andrews (BSMD), Nathalie Hamel (BSMD),
Patrice Martineau (TDD) and Caroline Rondeau (BSMD)

BSMD-2007-009E

METHODOLOGY FOR THE PROCESSING AND IMPUTATION OF CORPORATIONS DATA (T2)

Jessica Andrews, BSMD
Nathalie Hamel, BSMD
Patrice Martineau, TDD
Caroline Rondeau, BSMD

November 2007

Abstract

The Canada Revenue Agency (CRA) shares some administrative data with Statistics Canada including data for incorporated businesses (T2). Those data, which come from the income tax return and the financial statements, are used at Statistics Canada for statistical purposes, such as replacing economic survey data. The T2 data, which come monthly from the CRA, constitute a census of about 1.3 million legal entities.

When they arrive at Statistics Canada, the data are first loaded to the database, processed and then shared with representatives of economic surveys. The processes applied to the data are carried out monthly, with the exception of the enterprise roll-up, which is performed twice a year. Certain partial imputation processes (deterministic imputation) are also carried out on a monthly basis to ensure that there is no missing information (a missing variable in a record) or erroneous information (a variable that fails one or more edits in a record).

Since the survey data are replaced before the whole economic survey universe (known as the survey universe file (SUF)) is received, missing data (records that have not come in when the SUF is completed in September) and erroneous data (records containing variables that failed the edits) are imputed. For example, in September of year *t*, the T2 data for reference year *t-1* are completed to make up the SUF for reference year *t-1*. Historical imputation and nearest-neighbour donor imputation are performed on the data annually. Since the 2004 reference year, estimates of the variance due to historical imputation have also been produced. Estimation of this variance will eventually be added to the annual processes done by Tax data Division (TDD).

This report describes the methodology used to process and impute T2 data and the method for estimating the variance due to historical imputation.

Key words: administrative data, monthly and annual processing of data, monthly and annual imputation of data, estimation of variance due to historical imputation.

Résumé

L'Agence du revenu du Canada (ARC) partage certaines données administratives avec Statistique Canada dont celles sur les sociétés incorporées (T2). Ces renseignements, provenant de la déclaration fiscale et des états financiers, sont utilisés à Statistique Canada à des fins statistiques, notamment pour remplacer les données d'enquêtes économiques. Les données T2, qui proviennent mensuellement de l'ARC, sont des données recensées représentant environ 1,3 million d'entités légales.

Lorsqu'elles arrivent à Statistique Canada, les données T2 reçues sont d'abord chargées sur la base de données, traitées et ensuite partagées avec les représentants des enquêtes économiques. La plupart des traitements sont appliqués mensuellement aux données à l'exception de l'agrégation des entités légales aux entreprises, qui est un traitement appliqué deux fois par année. Certaines procédures d'imputation partielle (imputation déterministe) sont également appliquées mensuellement aux données pour s'assurer de compléter l'information manquante (variable manquante au niveau de l'enregistrement) ou jugée erronée (variable qui échoue certaines règles de contrôle au niveau de l'enregistrement).

Puisque le remplacement des données d'enquêtes a lieu avant la réception de l'univers complet des enquêtes économiques (appelé le fichier de l'univers des enquêtes (FUE)), les données manquantes (les enregistrements qui ne sont pas encore reçus au moment de compléter le FUE, soit en septembre) ou jugées erronées (les enregistrements incluant des variables qui ont échoué les règles de contrôle) sont imputées. Par exemple, en septembre de l'année *t*, les données T2 pour l'année de référence *t-1* sont complétées pour représenter le FUE pour l'année de référence *t-1*. L'imputation historique et l'imputation par donneur selon la méthode du plus proche voisin sont des procédures appliquées annuellement aux données. Depuis l'année de référence 2004, l'estimation de la variance due à l'imputation historique est également estimée. Elle sera éventuellement ajoutée aux procédures annuelles effectuées par la division des données fiscales (DDF).

Ce document présente la méthodologie utilisée pour traiter et imputer les données T2 ainsi que l'estimation de la variance due à l'imputation historique.

Mots clés: données administratives, traitements mensuels et annuels des données, imputations mensuelles et annuelles des données, estimation de la variance due à l'imputation historique.

Table of Contents

1. Introduction

Financial and tax data for corporations (T2) are now playing a prominent role in the production of economic data at Statistics Canada. In recent years, the Agency has expanded its use of such data substantially, primarily to reduce the response burden and costs of economic surveys.

For the majority of economic surveys, the approach involves using T2 data for small, simple units in the take-none stratum,[1] for some selected sample units in the take-some stratum, and for units that did not return the questionnaire.

T2 data are received continuously from the Canada Revenue Agency (CRA). Corporations can choose whatever fiscal year they prefer, and they have up to six months after the end of their fiscal period to file their tax returns.

Each month, Statistics Canada's Tax Data Division (TDD) receives the data processed by the CRA the previous month and puts them through some rigorous processing, including certain control of data quality, before making them available to internal users. Since the survey data are replaced before the whole economic survey universe (known as the survey universe file (SUF)) is received, missing data (records that have not come in when the SUF is completed in September) and erroneous data (records containing variables that failed the edits) are imputed. For example, in September of year *t*, the T2 data for reference year *t-1* are completed to make up the SUF for reference year *t-1*. Historical imputation and nearest-neighbour donor imputation are performed on the data annually.

This report describes the methodology used to process and impute the T2 data and the method of estimating the variance due to historical imputation. Section 2 outlines the procedures for receiving, loading and processing the data. Section 3 describes the imputation procedures, while new developments are presented in section 4. Section 5 contains the conclusion.

2. Receiving, loading and processing the data

Each month, TDD receives SAS files containing T2 tax data (various CORTAX schedules) and financial information (Income Statement and Balance Sheet, also known as the General Index of Financial Information (GIFI)) from the CRA. From those files, TDD extracts assessment and reassessment data for reference year 2000 and subsequent reference years. The data are then transposed and reformatted for loading into the database.

[1] Units below a fixed threshold.

2.1. Reassessment

Previously, only assessments were input to TDD's T2 database. Following several studies and consultations with the CRA, reassessments were added to the database. This is warranted in the following cases:

- The financial information is different.
- The initial record is considered erroneous (under the CRA's edit rules).
- No initial record was received.

Furthermore, the CRA strongly recommends using reassessments because:

- the initial data may contain errors since they have not been checked;
- if there are revisions, it is because the initial data were wrong.

2.2. Detecting outliers in a record

Big huge numbers (BHNs) are typically generated by field concatenation, which occurs when a missed delimiter causes two values to be captured in one cell. A BHN usually includes a numeric field name, as in the following example: *Total revenue* is captured as 120093681000, but it should actually have been $1,200 for *Total revenue* and $1,000 for *Total expenses*, whose field name is 9368.

Outliers are detected in a record by comparing the financial and tax data for the same record using the following strategy:

1. The largest and second-largest values are found over the entire form for all but certain fields.
2. If the largest value has at least four more digits than the second-largest value, the unit is considered a possible BHN.
3. If the largest value is detected as being less than 10,000,000, it is excluded from being considered a possible BHN.

Records containing at least one outlier are corrected manually later in the process.

2.3. Deterministic edits

The deterministic edits ensure that the results in each section (e.g., assets and components thereof) balance. The edits are performed on two levels.

The first level ensures that the totals in the Balance Sheet and the Income Statement match:

$$\textit{Total assets} = \textit{Total liabilities} + \textit{Total equity} \tag{1}$$

$$\textit{Total profits after adjustments} = \textit{Total income} - \textit{Total expenditures} + \textit{Adjustments}^{2} \tag{2}$$

The second level ensures that the various sections of the Balance Sheet and the Income Statement also balance, when the components are available:

$$\textit{Sum of section components} = \textit{Section total} \tag{3}$$

Records that break one or more edit rules are identified as erroneous and are later corrected manually. Partial totals are calculated only for records that pass the deterministic edits.

2.4. Reviews and manual corrections

Many of the errors identified in records cannot be corrected systematically. In some cases, a variable was simply omitted, or the same digit was entered twice, changing $5,450 into $55,450, for example. The only solution in such cases is to correct the erroneous data manually.

The employees who perform manual corrections use a system developed and implemented by TDD. The system features a number of internal rules that help the correctors ensure that their corrections comply with accounting rules and that the financial statements still balance.

In making their corrections, the correctors use various auxiliary information sources, such as economic survey data, financial and tax data for the current and previous years, and publicly available information (official financial statements and other publications).

2.5. Overlap between fiscal periods

This process identifies records for which the fiscal periods overlap.

If both an imputed missing record and a real record exist for a given nine-character BN (BN-9)[3] for the same reference year, the system will regard them as overlapping even if the fiscal periods do not overlap. In that case, the imputed record will be deleted.

[2] Adjustments reflect taxes and extraordinary items.

[3] The nine characters are the company's unique Business Number.

If the fiscal periods of an imputed received record and a real record for a given nine-character BN (BN-9) for the same reference year have more than a 31-day overlap, the imputed record will be deleted.

If the fiscal periods of two imputed records or two real records for a given 15-character BN (BN-15)[4] for the same reference year have more than a 31-day overlap, the record with the shorter fiscal period will be deleted.

2.6. Correspondence between financial and tax data

The correspondence process involves changing only the fiscal period covered by imputed financial data when real fiscal data exists for that corporation (BN-9).

Example:

Financial Data

BN	Start date	End date	Year	Imputation
XXXXXXXXX	01/31/2002	12/15/2002	2004	Yes

Tax Data

BN	Start date	End date	Year	Imputation
XXXXXXXXX	01/31/2002	12/15/2002	2004	Yes
XXXXXXXXX	01/01/2002	12/31/2002	2004	No

In this case, the fiscal period of the financial data would be changed to match the fiscal period of the tax data, which are real data. The imputed tax data would also be deleted, as shown below.

Financial Data

BN	Start date	End date	Year	Imputation
XXXXXXXXX	01/01/2002	12/31/2002	2004	Yes

[4] The first nine characters make up the BN-9, the 10th and 11th characters are the program identification code (RC for the T2s) and the last four characters are the reference number (a one-up number beginning with 0001 for each account).

Tax Data

BN	Start date	End date	Year	Imputation
XXXXXXXXX	01/01/2002	12/31/2002	2004	No

2.7. Negative values

From an economic standpoint, the various financial fields can be classified on the basis of whether they should take a positive value, a negative value or either one. From an accounting perspective, however, some fields that should theoretically be positive are actually negative.

By means of this process, negative values in fields that should theoretically have a positive value are transferred to an equivalent field where they take a positive value.

For example, if the cash and deposits field, which is part of assets, is negative, the amount will be transferred to bank overdraft (a liabilities field), where it will have a positive value. The cash and deposits field will be reset to zero, and the various totals will be adjusted.

2.8. Consistency between financial and tax data

The consistency edits ensure that the net income from the Income Statement, Balance Sheet and Schedule 1 (Net Income (Loss) for Income Tax Purposes) are the same and give precedence to real data. If any of these sources change, the amount deleted or added in the net income will be shifted to another field. For example, if net profit is changed on the Income Statement, we will adjust an extraordinary item.

In the accounting world, the net profit after taxes and extraordinary items from the Income Statement should be entered in the net income/loss item in the statement of retained earnings, which is part of the Balance Sheet.

Corporations use Schedule 1 to calculate the taxes they owe or the refund to which they are entitled. Schedule 1 captures the differences between the accounting world as reflected in a company's books and the calculations required by the CRA and the *Income Tax Act*. The company then uses the net profit/loss on its books as a starting point for calculating the Net Income (Loss) for Income Tax Purposes. Therefore, the net income (loss) from the Income Statement, from the statement of retained earnings (Balance Sheet) and from Schedule 1 should all be the same. However, a corporation may not have anything to report under Net Income (Loss) for Income Tax Purposes. In that case, the income/loss reported under *Net income (loss) after taxes and extraordinary items* in the

Income Statement will be transferred to *Net income or loss for income tax purposes* in the T2 Corporation Income Tax Return.

2.9. Balancing edits

T2 corporations report their taxable income and calculate their income tax payable on Form 200 (Corporation Income Tax Return). The balancing edit ensures, using the various rates prescribed by law, that all amounts derived from Form 200 are accurate.

For example, if, under CRA rules, line Z must equal 50% of line Y, the edit will check that line Z actually does equal 0.5 * Y.

Alberta, Quebec and Ontario income tax is also calculated by this process, since the tax for those provinces is not shown in CRA schedules. Those provinces have decided to do their own collection of provincial income tax from corporations.

The process checks that taxable income is the same in Schedule 5 (Tax Calculation Supplementary – Corporations) and Form 200. It also verifies that the provincial and territorial totals match. If a corporation operates in just one province or territory, its total taxable income will be copied to the appropriate line in Schedule 5.

2.10. Chart of accounts (COA)

The concepts and definitions used by the CRA and Statistics Canada's economic survey programs are often different. With the increased use of T2 data, it became essential to find a way of achieving consistency across the following data sources: Administrative data, survey data and National Accounts data. The chart of accounts was developed for that purpose. The chart of accounts consists of standardized financial statements, which make it easier to link the three data sources.

Each month, the financial data are aggregated on the basis of the variables in the chart of accounts. This provides users with access to financial data in chart-of-accounts format.

2.11. Enterprise roll-up

Twice a year, the financial and tax data for the various entities in an enterprise are rolled up to produce data for the statistical enterprise. That information is used directly by various T2 data users.

Allocation of statistical enterprise financial data to establishments is carried out by Enterprise Statistics Division (ESD). The information used to make the connections to statistical enterprises comes from Statistics Canada's Business Register (BR).

3. Imputation

T2 data are imputed by various methods and with varying frequency. First, a missing value (either a missing variable or a variable that failed the deterministic edits) at the record level for key variables is imputed by a deterministic method. Missing values are replaced with historical data from the same source or with similar values from other sources. This process is run monthly as the data are received.

The other two types of imputation used on T2 data are trend-adjusted historical imputation and nearest-neighbour donor imputation. These processes are run annually in order to complete the SUF. They impute missing records (records that have not been received when the SUF is completed in September) and erroneous records (records containing variables that failed the deterministic edits). Each of these methods is described in more detail below.

3.1. Deterministic imputation of key variables

Imputation of key variables was introduced to address availability problems with variables that are essential for T2 data users but for which detail is not required by the CRA. Corporations have the option of reporting the variables directly or including them in more generic or rolled-up fields.

Imputation is performed by a deterministic method that involves using historical financial data when available or similar data from other administrative, financial or tax sources to fill in missing values for key financial variables.

To keep the financial statements balanced and avoid affecting the income and expenditure totals, the amount added is subtracted from another financial data field. This reflects the assumption that the missing amount was included in another generic field. The key variables are inventories, amortization, and wages and salaries.

3.1.1 Inventories

When the Inventories variable in the Income Statement is missing, it is imputed using two main sources:

1. Balance Sheet data
2. Historical data

If $CI_{IS} = 0$ *and* $CI_{BS} \neq 0$, take the amount from CI_{BS} and add it to CI_{IS}, provided $CI_{BS} <= A$.

If $OI_{IS} = 0$ and $CI_{-1} \neq 0$, take the amount from CI_{-1} and add it to OI_{IS}, provided $CI_{-1} <= A$.

where

OI_{IS} = opening inventories in the Income Statement
CI_{BS} = closing inventories in the Balance Sheet
CI_{IS} = closing inventories in the Income Statement
CI_{-1} = closing inventories for the preceding Balance Sheet period
A = another financial field from which the amount is taken

The fields from which the imputed amount can be taken are listed below.

- L8320 Material cost and purchase amount
- L8450 Other direct costs
- L8370 Non resource production cost
- L8360 Trades and sub-contracts
- L8400 Resource production costs
- L8435 Unspecified Crown charges

The fields are listed in priority order. If $CI_{-1} <= A$ for the first field, CI_{-1} will be taken entirely from that field. If not, we go on to the next field on the list and repeat the process until we obtain CI_{-1} entirely from one field. If the amount is not available from any of the fields listed, deterministic imputation will not be performed.

3.1.2 Amortization

When the Amortization variable in the Income Statement is missing, it is imputed only from Schedule 1 of the income tax return.

If $Amor_{IS} = 0$ and $Amor_{S1} \neq 0$, take the amount from $Amor_{S1}$ and add it to $Amor_{IS}$, provided $Amor_{S1}$ <= A.

where

$Amor_{S1}$ = amortization in Schedule 1

$Amor_{IS}$ = amortization in the Income Statement

A = another financial field from which the amount is taken

The fields from which the imputed amount can be taken are listed below.

- L8450 Other direct costs (when amortization is added to the cost of goods sold)
- L9270 Other expenses (when amortization is added to non-farming expenses)
- L9284 General and administrative expenses (when amortization is added to non-farming expenses)
- L9760 Unspecified farm machinery expenses (when amortization is added to farming expenses)
- L9790 Unspecified general farm expenses (when amortization is added to farming expenses)

The procedure for extracting the imputation amount is the same as in section 3.1.1.

3.1.3 Wages and salaries

When the Wages and Salaries variable in the Income Statement is not reported or underestimated, it is imputed from five auxiliary sources:

- CORTAX Schedule 5, Tax Calculation Supplementary - Corporations
- CORTAX Schedule 27, Calculation of Canadian Manufacturing and Processing Profits Deduction
- Historical data for reference year -1
- Historical data for reference year -2
- Data from PD7s, Statement of Account for Current Source Deductions
- Data from T4s, Statement of Remuneration Paid

For various reasons, these sources do not all provide the same figures, and more than one source may be used in processing a given corporation. Consequently, the sources must be placed in priority order, since only one source will be used for imputation. The priority order is as shown above. The sources were prioritized on the basis of five criteria: reliability, relevance, timeliness and periodicity, availability, and control of information.

To avoid outliers and to ensure that the process has some overall effect, only values greater than $1,000 are imputed.

If $WS_{IS} < WS_A$ *and* $(WS_A - WS_{IS}) \geq 1000$, take the amount given by $(WS_A - WS_{IS})$ and add it to WS_{IS}, provided $(WS_A - WS_{IS})$ <= A.

where

WS_A = Wages and salaries from the other source
WS_{IS} = Wages and salaries from the Income Statement
A = another financial field from which the amount is taken

The fields from which the imputed amount can be taken are listed below.

- L8450 Other direct costs
- L9284 General and administrative expenses
- L9270 Other expenses
- L8871 Management and administration fees
- L8400 Resource production costs
- L9110 Sub-contracts
- L8360 Trades and sub-contracts
- L8620 Unspecified employee benefits
- L9790 Unspecified general farm expenses
- L8350 Direct wage benefits cost
- L9808 Farm office expenses
- L8320 Material cost and purchase amount
- L8370 Non resource production cost
- L9273 Selling expenses
- L8810 Office expenses
- L8860 Professional fees
- L8863 Consulting fees
- L8873 Registrar and transfer agent fees

The 18 fields are shown in the order in which they are to be used. They were prioritized on the basis of criteria such as relevance and the probability that they would contain the amount concerned. The procedure for extracting the imputation amount is the same as in section 3.1.1.

Preliminary studies show that in general, as we expected, the amount can be extracted from only one of the 18 fields. As a result, the priority order given above is usually of little consequence.

The Wages and Salaries variable differs slightly from the other two key variables in the way it is processed. This is because all cases in which the variable is underestimated are processed, even if the financial variable is not missing.

3.2. Historical imputation

The basic idea behind historical imputation is to impute a missing record or an erroneous record for the current year with data from the previous year. For financial information, data from the previous year are adjusted using a calculated trend to better reflect the current year data. The trends are calculated within homogeneous imputation classes. No trend is used for tax information.

When historical imputation is required for financial information, both the Income Statement and the Balance Sheet are imputed, except when one of them has been received on time and passed the edits. In that case, only the missing or erroneous form is imputed and the other form is kept as is. The process is referred to as full historical imputation when both forms are imputed and partial historical imputation when only one form is imputed.

The imputation of financial information involves defining the target population, identifying exclusions, calculating and applying the trend and the fiscal period length ratio, and imputing the missing information. The imputation method used for financial information dictates which imputation method will be used for tax information. When financial information is historically imputed for the current year and income tax Form 200 is missing for the current year, Form 200 and Schedules 1 and 5 are historically imputed from the previous year. Each of these steps is detailed below.

The target population consists of current-year SUF records whose data for the previous year are of good quality. Previous year data are considered to be of good quality if they are received data that passed the edits, data that were historically imputed (partially or fully), data that were partially donor imputed, or data that were fully donor imputed in one of the first two imputation runs. Records whose North American Industry Classification System (NAICS) begins with 91 and records whose nine-digit business number (BN) was generated artificially by the BR are excluded.

To calculate trends, we first detect outliers by computing the rate of change in a variable's value from current year to previous year, then using the Hidiroglou-Berthelot (1986) method. After detecting and excluding outliers, we calculate a trend that reflects the growth rate from previous year to current year and apply it to the variables to be imputed. Trends are calculated for imputation classes, which are groups with enough homogenous records to be as representative and robust as possible. Imputation classes are classified by NAICS/province, NAICS/region (Maritimes, Quebec, Ontario, Prairies, and BC/Territories) or

NAICS code. We assume that the records with the same province and six-digit NAICS code are the most homogeneous. An imputation class is then formed by records with the same NAICS-6/province. To avoid an unstable trend resulting from a small imputation class (less than 50 records), if necessary, we relax the imputation class requirement by changing it from same NAICS-6/province to same NAICS-6/region. If the size of the imputation class formed using NAICS-6/region is still less than 50, we further relax the requirement to same NAICS-6, and then to same NAICS-5 if necessary, and then to same NAICS-4 if necessary. We do not relax the requirement any further than same NAICS-3. Cases with missing NAICS-6, missing trends or trends based on less than 10 records at NAICS-3 are imputed with trends based on the entire population.

The variables of interest for historical imputation of financial variables are revenue, expenses, assets and liabilities. A trend is calculated for these four variables and then applied to the data in the appropriate section.

As mentioned previously, a trend is the year-to-year growth rate of a variable within an imputation class. Essentially, we assume that, for a variable of interest *y*, its value for the current year, y^c, and its value for a previous year, y^p, satisfy a regression model

$$\begin{aligned} &y_{ik}^{(c)} = t_k^{(p)} y_{ik}^{(p)} + \varepsilon_{ik} \quad i = 1, \cdots n, \textit{ the index of a record in imputation class } k \\ &\textit{and} \\ &\varepsilon_{ik} \overset{ind}{\approx} N(0, y_{ik}^{(p)} \sigma^2) \end{aligned} \tag{4}$$

Thus, the trend $t_k^{(p)}$ is calculated as

$$\hat{t}_k^{(p)} = \frac{\sum_i y_{ik}^{(c)}}{\sum_i y_{ik}^{(p)}} \tag{5}$$

the total of current year's values divided by the total of the previous year's values for each imputation class. This gives us a growth rate for that variable from year *p* to year *c*. For records needing historical imputation, we can apply the growth rate $t_k^{(p)}$ to the current year data if the financial information for year *p* is used.

For Income Statement imputation where current year data were received and failed the edits, the information from the previous year is copied to the record for the current year. Where multiple previous year declarations are available for this BN, we impute using the previous year record with the longest and most recent fiscal period. The trend $t_k^{(p)}$ is applied to the imputed data. We also calculate and apply a fiscal period length ratio *R* based on the current year *c* and the previous year *p*, as follows:

$$R = \frac{length\ of\ fiscal\ period_c}{length\ of\ fiscal\ period_p} \tag{6}$$

For Balance Sheet imputation where current year data were received and failed the edits, the information from the previous year is copied to the record for the current year. Where multiple previous year declarations are available for this BN, we impute using the most recent record. The trend $t_k^{(p)}$ is applied to the imputed data.

For Income Statement and Balance Sheet imputation where no data were received for the current year, all the information from the previous year is copied to the record for the current year. The trend $t_k^{(p)}$ is applied to the imputed data.

For the imputation of tax information, only dollar value fields are copied from the previous year to the current year record for Form 200 and Schedules 1 and 5. Tax information is imputed after financial information.

3.3. Donor imputation

Donor imputation is based on the nearest-neighbour method. For financial information, partial or full imputation may be used for erroneous records, while only full imputation is performed for missing records. Partial imputation is used for erroneous records that fail the deterministic edits for the *Cost of sales* section only (note that this imputation will be discontinued in the 2006 reference year because of the low imputation rate). Other erroneous records are fully imputed. For tax information, only full imputation is used for missing records. In this case, there are no erroneous records because deterministic edits are not performed for tax data. Donor imputation consists of three processes: pre-imputation, imputation and post-imputation. The three processes are described in detail below.

3.3.1 Pre-imputation

The pre-imputation process for financial information identifies missing and erroneous records for which no historical imputation will be performed (since historical imputation follows donor imputation). The process involves the following steps: defining the target population and exclusions, identifying recipients and donors, setting start and end dates, and generating the matching variables (to find the pool of donors) and other fields needed for imputation.

The target population is the current year SUF. Records with good previous year information for historical imputation, a many-to-one link to an enterprise, a generated BN or a NAICS code beginning with 91 are excluded. The start and end dates are determined using the information in the financial statements or the tax information. If more than one tax information record for the previous year is available for the unit, the information from the record with the highest account number and the longest and most recent fiscal period is used.

For records requiring partial imputation of financial information, matching variables are derived using current-year financial information. Only records that have a non-zero total for cost of sales and do not require any kind of imputation are used as donors. The matching variables derived are revenue, expenses, total cost of sales and the net income sign.

For records requiring full imputation of financial information, matching variables are derived using financial information from previous years, administrative files for 1999 or 1998, or the BR, in that order. Note that beginning in the 2006 reference year, matching variables are derived using financial information from the two previous years (reference year -2 and reference year -3) or the BR only. When financial information or administrative files are used, the matching variables derived are fiscal period end date, fiscal period, revenue, assets, revenue-to-expenses ratio and net income sign. The revenue, expenses and assets variables are multiplied by a trend factor to control for changes from previous years to the current year. The revenue and expenses variables are multiplied by a revenue trend, and the assets variable is multiplied by an assets trend. The methodology used to calculate these trends is the same as the one presented in the historical imputation section.

When the BR is used, the same matching variables are derived. However, since the BR does not have an expenses variable, we use a linear regression model to derive expenses from revenue. The estimated beta, $\hat{\beta}$, is the coefficient in the regression model that predicts expenses based on revenue.

$\hat{\beta}$ is obtained using current-year financial information and is calculated for the NAICS-4 imputation class. When the size of the imputation class is less than 15, the resulting $\hat{\beta}$ is out of range (0.5, 1.5) or the $\hat{\beta}$ results from a regression model with relatively weak linear relationship ($R^2 <.95$), then $\hat{\beta}$ is based on the entire population. In the calculation of $\hat{\beta}$, outliers based on the revenue-to-expenses ratio for the current year are excluded using the Hidiroglou-Berthelot method.

This $\hat{\beta}$ is then applied to the revenue in the BR to derive the expenses variable:

$$y_{ik} = \hat{\beta}_k x_{ik} \qquad i = 1, \cdots n, \textit{ the index of a record in imputation class } k \qquad (7)$$

For records requiring full imputation of tax information, the donor is the same as the one used for financial information. Otherwise, the matching variables are derived using the current year of the financial information. The matching variables derived are absolute revenue, net income and assets, ratio of absolute revenue to absolute net income, ratio of absolute assets to absolute net income and signs of revenue, assets and net income.

3.3.2 Imputation

The imputation process for financial variables takes the information created by the pre-imputation process and finds donors for all recipients needing imputation. The generalized system used to find a donor for each recipient is BANFF. The BANFF system transforms the value of the matching variables as follows:

- It sorts the valid values from donors and recipients in ascending order.
- It assigns a rank to each value. If the values are the same, the same rank (i.e., the average of the two ranks) is assigned.
- It divides each rank by the total number of valid values plus one.

For each recipient, we choose the donor that minimizes the expression:

$$MAX \{| t_1(r) - t_1 (d) |, | t_2(r) - t_2 (d) |, \ldots,| t_n(r) - t_n (d) |\} \qquad (8)$$

Where $t_1(d), t_2(d), \ldots$ *and* $t_n(d)$ are the transformed donor's values for the first, second, ... and nth matching variables, and $t_1(r), t_2(r),\ldots$ and $t_n(r)$ are the transformed recipient's values for the first, second,... and nth matching variables.

The donor must also pass the post-imputation edits to be accepted. If the closest donor does not pass the edits, the next-closest donor is considered, and so on until we find an acceptable donor. Note that there is a maximum number of donors that will be checked in each run for each recipient.

Many attempts are made to find a donor for each recipient, but the majority of recipients find their donor on the first or the second run. In the first run of each module (partial and full imputations of financial information and full imputation of tax information), we use very restrictive imputation classes, based on NAICS-6 code, region, etc. As we progress through the runs, we expand the donor pool for each recipient by combining certain imputation classes, for example, based on NAICS-4 code and same net income sign.

3.3.3 Post-imputation

The post-imputation process takes the information created by the imputation process and imputes values for the recipients from the donor information.

Records that have undergone partial imputation of financial information are allowed to be donors for the full imputation of financial information. Therefore, partial imputation must be completed before full imputation. This will ensure that all full imputation donors have a complete direct expenses section before donating their values. In partial imputation, the values of the fields in the direct expenses section of the donor's Income Statement are copied to the corresponding fields in the recipient's Income Statement, but the total is not overwritten. The values are then prorated to make their sum equal to the recipient's total. The prorating value is defined as follows:

$$\textit{Prorating value} = \begin{cases} 1, & \textit{if } y \textit{ is null or prorating value} \leq 0 \\ \dfrac{x}{y}, & \textit{otherwise} \end{cases} \qquad (9)$$

where *x* represents the recipient's reported total and *y* represents the sum of donated detail items.

For full imputation of the Income Statement, the values of the fields in the donor's Income Statement are copied to the corresponding fields in the recipient's record. The values are then prorated to improve the fit for revenue. The prorating value is the same as above, where *x* represents the recipient's revenue and *y* represents the donor's revenue.

For full imputation of the Balance Sheet, the values of the fields in the donor's Balance Sheet are copied to the corresponding fields in the recipient's record, except fields from the equity section (for more details on the equity section, see the BSMD and TDD system specifications). The imputed values are then prorated to improve the fit for assets. The prorating value is the same as above, where *x* represents the recipient's assets and *y* represents the donor's assets.

For full imputation of tax information, dollar value fields in the donor's Schedule 1 are copied to the corresponding fields in the recipient's record, except the description of other additions and other deductions. Form 200 is imputed in the same way. The values are not prorated.

3.4. Generic-to-detail allocation (GDA)

Corporations can report their data in generic field, in detail or both. For the CRA, only eight fields are compulsory. The other fields are not checked. Statistics Canada, on the other hand, needs the details, since T2 data are used to replace economic survey data. To meet users' requirements, the financial data are put through a generic-to-detail allocation process. The process is carried out monthly for incoming data and annually for the SUF data.

Two methods have been used for the generic-to-detail allocation: the old methodology and the new methodology. The new methodology is used for reference year 2006 and beyond, but only for the Income Statement.
Both methods involve imputing a distribution based on details reported by corporations. This distribution is then used to impute missing details for which generic data are present, within an imputation class for a given block. The block is defined as follows for *Office expenses*, for example:

Definition of the *Office expenses* block based on the generic field and its details

Block: **Office expenses**	**Field number**	**Field description**
Generic field	8810	Office expenses
Details	8811	Office stationery and supplies
	8812	Office utilities
	8813	Data processing

The difference between the two methods has to do with the way the imputation classes are defined. In the old methodology, classes are defined by industry and revenue size, whereas in the new methodology, they are defined by a hierarchical classification and a discriminant analysis using statistically significant reported fields (details, generics, sum of details, etc.). Studies have shown that imputation classes are more homogeneous with the new methodology and that the micro-level results are more accurate. Thus, the new methodology meets users' needs more effectively. The two methodologies are described in more detail below. You can also refer to Andrews, Brisebois and Hamel (2007).

3.4.1 Old methodology

First, we make sure that the sum of the partial totals (subtotals) is within $1,000 of the grand total, and that the sum of the generic fields and their details matches the partial total within $1,000. When the totals do not match, the GDA ratios are used to correct the partial totals in the former case and the generics and details in the latter case.

Then we start processing records through the GDA. Certain blocks and details are excluded from the GDA process: blocks that do not have exhaustive details assigned to a single chart-of-accounts variable, the *Other Revenue* block when it is less than zero, and details previously assigned through wage and salary edit or amortization edit. Note also that for the Balance Sheet, certain amortization blocks use the ratios for the corresponding blocks when necessary to avoid creating imbalances in the financial information.

Then we assign all businesses to imputation classes. The imputation classes are based on the first two digits of the NAICS code (which provides 25 different industries), and on business size. Business size has three revenue levels: large (over \$25 million), small (under \$5 million) and mid-size (between \$5 million and \$25 million). Within each block, all imputation classes must contain at least 25 businesses that report only detail amounts; otherwise, classes are combined in accordance with the following rules:

1. Industry by size group (small-medium, medium-large or small-medium-large);
2. Group of industries (financial or non financial) by original sizes;
3. Group of industries by size group;
4. All industries together by original size;
5. All industries by all sizes.

For each imputation class, ratio distributions are estimated by

$$R_j = \frac{\sum_i Y_{ij}}{\sum_i \sum_j Y_{ij}} \tag{10}$$

where Y_{ij} is the non-zero value reported for detail j (which is reported by at least 10% of businesses, or 5% for the *General farm expenses* block) by business i, where all such businesses belong to the imputation class and report only detail amounts. If any of the estimated ratios are negative for details in which negative values are not permitted, the ratios are reassigned such that

$$R_j = \frac{1}{\sum_j |R_j|} |R_j| \tag{11}$$

Once all ratios have been estimated, imputation is performed. Imputation is carried out for all businesses reporting a generic amount.

There are three possible ways to impute GDA ratios. In the first case, only the generic amount is reported and the ratios are used exactly as they are. In the second case, a generic amount is reported along with a detail amount for all details with non-zero ratios. In this case, too, the ratios are applied to the generic. In the third case, a generic amount is reported along with amounts for some, but not all, details with non-zero ratios. In this case, the ratios are changed so that none of the generic is assigned to details that have been reported. Thus, the following ratios are applied to business i:

$$R_j = \begin{cases} \dfrac{R_j}{\sum\limits_{j\in(j,Y_{ij}=0)} R_j}, & if \quad Y_{ij}=0 \\ 0, \; if \quad Y_{ij}\neq 0 \end{cases} \tag{12}$$

Certain special edits have been added for industries such as *Real estate*. These are performed as a last step in order to ensure that values are reported in the expected details and not in details that do not apply to the industry.

3.4.2 New methodology

The new GDA methodology (based on Huang and Ladiray (2005)) is similar to the old one, except the choice of imputation classes differs. In the new methodology, businesses are assigned to imputation classes using models with a variety of variables where the groups correspond to the distribution of block totals over detail variables. Different blocks use different cluster definitions, but in most cases, dominant detail clusters (also called attractor clusters) are used. Business *i* belongs to cluster *j* of attractor *x*%, where *x>50*, if

$$\frac{Y_{ij}}{\sum\limits_{j=1}^{J} Y_{ij}} \geq \frac{x}{100} \tag{13}$$

where Y_{ij} is the total value reported by business *i* in detail *j*. If this statement is not true for any detail *j*, then the business is assigned to cluster $J+1$, where a block contains J details in total. When attractor clusters have not been used, clusters are based instead on common reporting patterns or abbreviated attractor clusters (where rare details do not have their own imputation class but are assigned to the general catch-all class).

Once imputation classes have been defined for businesses reporting details only, businesses for which imputation is needed (e.g., some amount is reported in the generic field) are divided into imputation classes using discriminant models. These discriminant models can be parametric or non-parametric. In the first case, a model based on the values of reported variables assigns each business to an imputation class. In the second case, the imputation class is based on the 15 closest neighbours in terms of the explanatory variables.

Once imputation classes have been generated, ratios of all details are estimated for businesses with known distributions as in the previous section. This is done without any restrictions on the proportion of businesses that provided responses for each detail. Generic amounts are then assigned to details using the ratio distributions and the current assignment rules. In the case of an imputation class with less than 25 businesses that reported details only, the group is absorbed into the general $J+1$ group. The exception to this is the *Other revenue* block, to which special rules apply.

3.5. Estimation of the variance due to historical imputation

The variance due to historical imputation is estimated using the EVITA program (Hurtubise, 2006). We assume that the historical trends estimated for individual variables are the same as those estimated for the section totals for expenses, revenue, assets or liabilities. This assumption is essential if we want to use the EVITA formulas; it is consistent with the assumptions underlying historical imputation of tax information (see section 3.2).

Contributors to the historical trend by EVITA are subject to the same restrictions as are used for the estimation of historical trends in the imputation process. The only difference is that for all businesses, the trend used is that of the full NAICS code and province for 2004 and subsequent years, and the full NAICS code for other years (instead of using NAICS codes with fewer digits for small imputation classes). This is due to EVITA's requirement that imputation classes not overlap. Changing the definition of imputation classes had little impact on the results, and it was found that most imputations were performed at these higher levels.

In the following equations, we use:

d: domain of interest
p: imputation class
g: region or stratum
o: set of non-respondents
r: set of respondents
m: current year
$m\text{-}1$: previous year

Similarly:

y_m : the variable of interest in the current year
y_{m-1} : the variable of interest in the previous year
w_f : weight used in sampling (equal to 1 since T2 data are the result of a census)
w_S : weight used to estimate σ^2, set to 1 for all data contributing to estimation of the trend since it is a census (0 for non-contributors)
w_β : weight used to estimate β, set equal to w_f

In the following,

$$N_{pg} = \sum_{i \in g} \sum_{i \in p} w_f , N_g = \sum_{i \in g} w_f \qquad (14)$$

and

$$\hat{y}_{pgk,m} = y_{pgk,m-1} \frac{\sum_{k \in r_{pg}} w_{\beta,pgk} y_{pgk,m}}{\sum_{k \in r_{pg}} w_{\beta,pgk} y_{pgk,m-1}} \qquad (15)$$

Finally, the formula for the variance is given by

$$\hat{V}_{IMPt}(d) = \sum_{g \in d} \sum_{p \in d} \left(\frac{N_{pg}}{n_{pg}} \right)^2 \left(\sum_{k \in o_{pg}} y_{pgk,m-1}(d) \right) \left(\frac{\sum_{k \in o_{pg}} y_{pgk,m-1}(d)}{\sum_{k \in r_{pg}} y_{pgk,m-1}} + 1 \right) \hat{\sigma}^2_{pg} , \qquad (16)$$

where

$$\hat{\sigma}^2_{pg} = \left(\frac{\sum_{k \in r_{pg}} w_{S,pgk} \hat{e}^2_{pgk}}{\sum_{k \in r_{pg}} w_{S,pgk} y_{pgk,m-1}} \right), \qquad \hat{e}^2_{pgk} = \left(y_{pgk,m} - \hat{y}_{pgk,m} \right)^2 . \qquad (17)$$

3.6. Identification of inactive units, misclassified units and double counts

Identification of inactive units, misclassified units, and double counts is very important, to prevent overestimation of T2 data. In order to limit overestimation, TDD introduced many processes to identify problematic units.

For most of these units, no imputation is performed. However, units identified as being inactive by using the Goods and services tax (GST) and PD7 data, are imputed. For these units, data are historically imputed, but Income Statement and tax data linked to activities are set to zero.

Inactive and dead corporations

The following corporations are considered to be inactive or dead:

- Corporations that previously reported (Form 200) they were going to be merged or dissolved;

- Corporations identified as inactive on the BR and who have not reported T2 data for the previous year;

- Corporations with a bankruptcy date prior to the beginning of the reference year.

- Corporations identified on the BR as dead or merged.

- Corporations with GST sales of zero.

  To be identified as inactive, they also must be part of an industry where at least 50 % of the corporations report GST data (this does not include corporations of exempted industries), have no PD7 data for the same reference year, and have been historically imputed the previous year (this means that the corporation has produced T2 data in the past, and has been inactive for at least a year).

Misclassified corporations

The following corporations are considered as misclassified:

- Units coded with NAICS 91 Public Administration. Most of them are not T2 corporations, which explains why TDD receives few declarations for that NAICS.

- Units without a T2 account, according to CRA files.

Double counts

The following cases are considered as double counts:

- Generated BNs. These BNs are used to identify units without an official BN. Often, the results for these units are reported under a different BN.

- Units where many BNs are linked to the same enterprise, because the results are usually reported under a different BN.

4. Developments

4.1. Estimation of the variance due to donor imputation

Currently, we cannot estimate the variance due to donor imputation using EVITA. This is because the EVITA method is based on the assumption that the donor is chosen in a way that relies on only one auxiliary variable. For T2 data, several different variables contribute to the choice of the donor (see section 3.3), and it was found that models depending on only one variable could not be created so as to produce results that were sufficiently similar.

We will have the capability to estimate this variance in the near future. A new method, in which dependence on more than one auxiliary variable is allowed, is being developed for unincorporated business (T1) data.

4.2. Outlier detection

Previously, outlier detection by the Hidiroglou-Berthelot (1986) method was performed on financial data when they were loaded. Records identified as potential outliers were then checked manually and corrected if necessary.

In processing the tax data for the 2004 reference year, TDD decided to discontinue this process because it was detecting about 10,000 potential outliers each month. Because of the high volume of potential outliers, it was impossible to check each one before the next set of data was loaded. What's more, few of the values identified were actually outliers. A number of edit rules had been added over the years to correct some systematic errors during data processing.

In the future, this method will be revised again in an effort to determine how effective it might be in meeting data requirements. We will also assess other outlier detection methods, such as those used for unincorporated corporations (T1) data, goods and services tax (GST) data and Unified Enterprise Survey (UES) data, ratios between two variables (e.g., income and expenses), and methods available in the BANFF system.

5. Conclusion

T2 data are used at Statistics Canada to replace economic survey data. Provided by the CRA, they consist of administrative data on corporations. Since the CRA requires only eight fields – the section totals and net profit/loss – Statistics Canada has to carry out data processing and imputation processes to ensure that the data are complete and of good quality and meet users' needs. To satisfy these requirements, T2 data processing and imputation procedures were developed and implemented by TDD in association with BSMD and in consultation with users. TDD is also responsible for maintaining the data and sharing them with Statistics Canada users.

The data are received from the CRA on a monthly basis. They are loaded into the TDD database, and some processes are performed each month on the incoming data so that users can access them as soon as possible. Some processes are carried out annually – in September of the year following the T2 reference year – to ensure that the SUF is complete.

Year after year, the various processes are reviewed and improved as required. There are regular consultations between TDD, BSMD and users. Some developments were described earlier in this report.

Acknowledgements

We would like to thank revisers François Brisebois, Wisner Jocelyn, Denise Kelly, Richard Laroche and Mathieu Thomassin for their valuable advice.

Bibliography

Andrews, J., Brisebois, F. and Hamel, N. (2007). Methodology of Allocating Generic Field to its Details. ICESIII, The Third International Conference on Establishment Surveys, June 18 to 21, 2007, Montreal, Canada. American Statistical Association, Alexandria (Virginia).

Hamel, N. (2005). "Traitement des données fiscales pour le remplacement des données d'enquêtes". "Quatrième colloque francophone sur les sondages" Acts.

Hamel, N. and Martineau, P. (2007). Assessment of the quality of data for corporations (T2) produced by Tax Data Division – 2004 data. Working paper, Statistics Canada, BSMD-2007-005F/E.

Hidiroglou, M.A. and Berthelot, J.-M. (1986). Statistical Editing and Imputation for Periodic Business Surveys. Survey Methodology, 12, 73-83.

Huang, R. and Ladiray, D. (2005). Imputing Distribution in Administrative Tax Data. Proceedings of Statistics Canada's Symposium 2005 - Catalogue no. 11-522-XIE.

Hurtubise, D. (2006). User Guide for … Evita, Version1 or Estimation of the Variance due to Imputation for Treatment and Analysis. BSMD internal document.

Mair, D. (2005). Survey with T2 Users. TDD internal document.

Mair, D. (2005). Review of tax data usage. TDD internal document.

Martineau, P. (2003). "Portrait des sociétés ayant fourni une déclaration de revenus (T2) mais n'apparaissant pas sur le Post-SUF 2000". TDD internal document.

Martineau, P. (2004). "Traitement des valeurs negatives". TDD internal document.

Martineau, P. (2004). "Traitement de la variable salaire". TDD internal document.

Rondeau, C. (2005). "Évaluation de la qualité de l'imputation". BSMD internal document.

Rondeau, C. (2005). "Comparaison des systèmes d'imputation". BSMD internal document.

System specifications produced by TDD and BSMD:

- T2 financial and tax information data load
- CORTAX BHND specification
- GIFI BHND specification
- Derivation of subtotals – GIFI
- Deterministic edits – GIFI
- Specifications of correction by allocation
- Specifications of overlapping records
- Correspondence between GIFI and CORTAX
- "Valeurs négatives – Sommaire"
- "Salaires et traitements – Sommaire"
- Inventory edits
- Depreciation edits
- GDA documentation
- TDD COA vs GIFI
- Enterprise Roll-up
- GIFI and CORTAX RYyyyy Pre-imputation process specifications
- GIFI and CORTAX RYyyyy BANFF process specifications
- GIFI RYyyyy Post-imputation process specifications
- CORTAX RYyyyy Post-imputation process specifications
- GIFI and CORTAX RYyyyy Historical imputation process specifications
- GIFI RYyyyy Trend and Beta Calculation for Historical and Donor Imputations Specifications
- GIFI and CORTAX RYyyyy cases inactive specifications.