STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working paper No. 22
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Work Session on Statistical Data Editing
(27 – 29 May 2002, Helsinki, Finland)

Topic (iii): Editing of administrative data

EDIT AND IMPUTATION OF THE GENERAL INDEX OF FINANCIAL INFORMATION

Contributed Paper

Submitted by Statistics Canada[1]

Abstract: Canada Customs and Revenue Agency (CCRA) maintain the General Index of Financial Information database, otherwise known as GIFI database, which is a census of corporate financial statements (T2). This database is shared with Statistics Canada for statistical purposes such as updating or creating sampling frames, adding information to help sample selection and/or imputation, calibrating survey estimations and producing business analysis. The database contains information on businesses such as the financial declaration as well as the income statement and the balance sheet, representing about 700 variables. Developing an edit and imputation system for such a large and diverse database presents a number of significant challenges to achieve data quality. A methodology has been developed over the past few years to deal with these challenges. The current GIFI editing system includes administrative edits, consistency edits, and historical edits. It also includes an imputation process, which is performed using historical imputation, as well as nearest neighbour donor imputation for both partial non-response and total non-response.

KEY WORDS: Editing; Imputation; Partial/Total non-response.

I. INTRODUCTION

1. The General Index of Financial Information database, otherwise known as GIFI database, is a census of corporate financial statements (T2) available from the Canada Customs and Revenue Agency (CCRA). This database is shared with Statistics Canada for statistical purposes such as updating or creating sampling frames, adding information to help sample selection and/or imputation, calibrating survey estimations and producing business analysis.

2. The database contains information on 1.2 million businesses such as the financial declaration (FD) as well as the income statement and the balance sheet, representing about 700 variables, divided into 7 major sections. The sections are non-farm revenue, non-farm expenses, farm revenue, farm expenses, assets, liabilities and shareholder equity. Most companies provide about 30 to 40 items and only 8 fields are actually required by CCRA. These fields are the section totals, and the net income/loss.

3. Developing an edit and imputation system for such a large and diverse database presents a number of significant challenges to achieve data quality. A methodology has been developed over the past few years to deal with these challenges. This methodology is discussed in this paper. First, a pre-fill process using last year data is applied to the database. This step is discussed in section II. The current GIFI editing system includes administrative edits, consistency edits, and historical edits, which are discussed in sections III, IV

[1] Prepared by Nathalie Hamel (nathalie.hamel@statcan.ca) and Richard Belcher (richard.belcher@statcan.ca)

and V respectively. A review and correction facility has been put in place to permit some manual and systematic corrections to the data. This step is discussed in section VI. It also includes an imputation process, which is performed using historical imputation, as well as nearest
neighbour donor imputation for both partial non-response and total non-response. This imputation process is discussed in section 7.

II. GIFI VIEW (PRE-FILL)

4. After reviewing a few options as to how it would be best to proceed for the initialization of the database at time *t*, it was decided to create the GIFI view *t* using all FD at time *t-1* as the base. The GIFI view *t* initially presents the user with the equivalent of a GIFI database at time *t* that would have been historically imputed using all FD at time *t-1*, to which a growth factor of 0% would have been applied.

5. Following the creation of this view is the monthly load, which consists of adding the new FD received at time *t*, to update both the database and the GIFI view *t*. The pre-filled records are replaced with real records at time *t* as they arrive. The users can then access a "live" database at any point. Because the monthly updates may contain data from both the current and previous taxation years, the strategy for the monthly load is designed to cover data for any tax year, which is then loaded at any point during the process. The goal is to make the data load general enough to work for multiple years and for pre and post cut-off records. The intent is to replace existing records with received monthly updates of GIFI and CORTAX data from CCRA. Since the monthly updates contain data covering multiple years, we may receive FDs that are brand new, replacements of FDs received earlier, or replacements of imputed FDs.

6. The monthly update process is divided into the following major stages: Pre-process or Administrative Edits, Consistency Edits and Derivation of Sub-Totals, Historical Edits and Review & Correction. These processes are explained in the following sections.

III. ADMINISTRATIVE EDITS

7. The pre-process was re-designed entirely last year as CCRA began sending us the data in a totally different format. Since then, CCRA has provided Statistics Canada with the CORTAX files (which include the GIFI files and three other forms such as schedules 1, 5 and 200) on a monthly basis. The monthly updates include every modification that was performed during the previous month at CCRA. After a first evaluation of the impact of the subsequent modifications on the GIFI data, it was decided that only the initial modifications would be loaded, for each Tax Filer identified using the Business Number (BN), for each specific fiscal period.

8. The first edit applied is the big huge number detector (BHND). It is applied on the income statement and balance sheet amount fields separately. The purpose of this edit is to find data entry errors by comparing the largest value reported by a particular business (in any field) to the second largest. If the difference between these two values is too large then the largest value is considered an error. FDs which are found to have unusually large amounts in either the income statement or the balance sheet are set aside for manual intervention.

9. The documents "GIFI TY2000 – Edit & Imputation Strategy & Documentation" and "Load and edit process description" describe the process in more details.

IV. CONSISTENCY EDITS

10. The GIFI format consists of 700 variables, divided into 7 major sections. For each section, we want to ensure that the total reported is equal to the sum of the corresponding detail items reported. The editing

strategy trusts the totals first and the items second. The edits are applied first to ensure that the financial statements (balance sheet or income statement) balance.

11. The following equations are applied:

- Total Assets = Total Liabilities + Total Shareholder Equity
- Total Revenue – Total Expense +/- Adjustments = Net Income after Adjustments for the income statement

- Sum of components = Section Total

12. If the section total is not equal to the sum of the components for that section, then we prorate the difference to the components within that section. If the proration factor (ratio) is negative, it would make the section "balance", but would reverse the sign of every component. In that case, the proration is not completed and the record is flagged for manual review and loaded onto a separate database, called the Error Table, to await review.

13. Qualified FDs are the FDs that pass the edits and do not require imputation. All qualified FDs are added to the database on a monthly basis and also automatically replace pre-filled FDs in the GIFI view on the Production Table. FDs that failed the edits are placed on the Error Table and are made available through the Review and Corrections Facility. Uncorrected records will become recipients for the imputation process. Sub-Totals are derived on a monthly basis for all qualified FDs placed on the TMS_PROD schema.

14. The document "GIFI TY2000 – Edit & Imputation Strategy & Documentation" describes the process in more details.

V. HISTORICAL EDITS

15. The historical edits are performed in two different ways. First of all, the outlier records are detected using the Hidiroglou and Berthelot (1986) method. The most important outliers are then manually reviewed and corrected if necessary. Secondly, some analytical tables are produced for analysis purposes, as a complement to the outlier detection process.

16. The outliers are detected by comparing year *t-1* and *t* at the micro level. As a start, we perform the comparison from year to year using the Annualized Derived Total Revenue variable from the Income Statement and the Total Assets variable from the Balance Sheet, for every qualified FD available in both years. Qualified FDs available in both years must have the following characteristics: must be non-imputed in time *t-1*; must have a positive, non-zero value in each year; must have a fiscal period of 31 days or longer, since the others will have odd properties, particularly when the revenue is annualized. Records with the largest differences from one year to the next are identified. What is considered to be a "large" difference depends on the size of the corporation; for smaller corporations, a larger percentage variation is accepted. All outliers are stored on a particular table, with some basic information, from which lists are produced on a monthly basis for review and correction if necessary. The corrections to these FDs are made directly on the TMS_PROD database, using the Review and Corrections Facility described later.

17. Since April 2001, the GIFI database has been updated on a monthly basis. In order to verify the quality and consistency of the GIFI database at a macro level, a series of analytical tables are produced and reviewed each month. These analytical tables help to identify inconsistencies in the data at an early stage, and provide Subject Matter areas with some useful information about the quality and consistency of the GIFI database, as it is growing.

18. The documents "GIFI 2000 – Historical edits (Outlier detection)" and "GIFI Analytical Tables" describe the process in more details.

VI. REVIEW AND CORRECTION FACILITY

19. The Review facility is intended to provide a simple tool with which to view micro-data. The use of the tool includes providing a way to find the data of interest, and display them in a logical fashion that reflects the organization of the data being displayed. This facility is also intended to show related information that would be commonly requested for a particular declaration e.g. North American Industry Classification System (NAICS) code for the related business. The history of the data (all versions of a declaration throughout all processing stages) is also displayable. In particular, this facility is made available to provide tax data users a facility to review and correct the records identified as outliers as well as to allow the review and correction of certain FD received, that have failed the edits. Only changes that respect the consistency edits are allowed. The document "Review and Correction" describes the process in more details.

VII. IMPUTATION PROCESS

20. Prior to releasing the Final GIFI database, the imputation process takes place. Two sets of records need imputation. The first set consists of the FDs that were received and failed the edits, and the second set are the records for which no FD has been received at the time of imputation. The imputation process is based on the population of the Statistics Canada's Unified Enterprise Survey which is derived from Statistics Canada's Business Register (BR). As mentioned earlier, the population of GIFI is approximately 1.2 million businesses, and information from about 900,000 of these is received each year from CCRA. There are about 150,000 GIFI records received each year that are out of scope (outside the population) and some that fail the edits. As a result, about 400,000 to 500,000 records will need full imputation.

21. Before 2000, the imputation was done using only a nearest neighbour donor imputation technique. Matching variables were sourced from the Administrative file from CCRA (ADMIN), historical GIFI information from CCRA, or the BR from Statistics Canada. For 2000 and beyond, the ADMIN file is no longer available as an independent source of information as it was in the past. Using the BR as a source has created some challenges. As a result, we decided that it would be a great improvement to develop a method for better exploiting historical GIFI data to help in treating total non-response. Since 2000, the imputation process has included a historical imputation technique. Now, the imputation is divided into two processes, the historical imputation and the donor imputation.

22. The documents "GIFI 2000 Historical Imputation Specifications", "GIFI RY2000 – Pre-GEIS Process Specifications", "GIFI RY2000 – GEIS Process Specifications", and "GIFI 2000 – Post-GEIS process specifications" describe the process in more details. Note that the imputation of other forms from CORTAX is not considered in that paper.

A. Historical Imputation

23. In order to create the best imputed records possible for GIFI, we use a historical imputation method that takes a non-respondent's own data from the previous year (if available), and adjusts the data according to trends measured for the responding records in the same imputation class or data group. Data groups for historical imputation are defined 'on the fly'. We try to use the most detailed NAICS code level possible, but want to avoid calculating trends that are unstable due to a small number of records in a group. The NAICS code is composed of 6 digits, which represent the levels. We start at the NAICS6 level, and move up to a less-detailed NAICS level only when the group does not contain enough units. In this way, we end up with some groups based on NAICS6 groupings, and others based on NAICS3, 4, or 5. A minimum group size of 50 records is considered sufficient for calculating stable trends.

24. Trends are calculated for each data group by taking the sum of the chosen variable in that group for the current year, and dividing by the sum of the chosen variable for the same group for the previous year. This gives us a growth factor for that variable. Then, for records needing historical imputation, we can apply the growth factor to the chosen variable and all other variables in the section. The natural divisions in the GIFI format are revenue, expenses, assets, and liabilities, but analysis has shown that the revenue and expenses sections can be combined without seriously affecting the results. These two variables are highly correlated. Creating three separate trends means a little extra work to keep the imputed records balanced. A balancing strategy has been created to avoid any unbalancing problems.

25. Outliers are removed from the trend calculation, using the same method used for historical edits (Hidiroglou and Berthelot method (1986)). In historical edits, outliers are flagged for manual review. In the historical imputation process, however, outliers are automatically excluded from the trend calculations.

B. Donor Imputation

26. The donor imputation is done using the Generalized Edit & Imputation System (GEIS) developed by Statistics Canada. Some manipulations of the data are needed to allow us to use GEIS. Therefore, this donor imputation is divided into three steps: Pre-GEIS, GEIS and Post-GEIS.

27. The Pre-GEIS step creates input files for GEIS. It identifies the donor and recipient records, and populates matching fields using GIFI, ADMIN or the BR. While using the GIFI file would be our first choice, the information is available for donors only. When using the ADMIN file, the information is available for past years only, so it requires trending to approximate current year information. This trended information is our first choice for populating recipients' matching fields. When using our second option, the BR, it is impossible to estimate the legal entity information when many legal entities are linked to the same statistical enterprise. For these cases, a cold deck technique is considered. This information is used to populate recipients' matching fields. The matching fields are region, revenue, expenses, assets, fiscal period end date and fiscal period length.

28. The GEIS step finds a donor for each recipient. It uses data groups based on NAICS. It can perform several runs with different rules. First runs are very strict, for instance, the donor must have the same 6-digit NAICS code as the recipient, must be in the same geographical region, have a similar fiscal period length and end date, and the revenue and assets must be in the same range (e.g. +/- 33%). Then the rules are relaxed through each run until the final one where the donor must simply have the same 4-digit NAICS code as the recipient. A Partial imputation is first performed on recipients needing imputation only for the Cost of Sales section. This is the only partial imputation performed on GIFI data. All records flagged for partial imputation of any other sections will be fully imputed (the real data available is not use in that case). After the partial imputation is completed, the full imputation follows and we allow the partially imputed recipients to become donors.

29. The Post-GEIS step passes the donated values to recipients. The partially imputed records are processed first to complete the information before processing the fully imputed records. This is done outside of GEIS due to large number of fields in GIFI. The recipients are then prorated according to the revenue and assets fields for the Income Statement and Balance Sheet respectively. Finally, the whole record is re-balanced as needed.

VIII. CONCLUSIONS

30. The edit and imputation of such a large and diverse database presents a number of significant challenges to achieve good data quality. A methodology is in place to try to deal with these challenges. The current GIFI editing system includes administrative edits, consistency edits, and historical edits. It also includes an imputation process, which is performed using historical imputation, as well as nearest neighbour

donor imputation for both partial non-response and total non-response. Due to the high number of variables and records available on that database, it has been important to improve the methodology over the past years and still in the future. Our future plan is to improve the partial imputation and also edits to try to maximize using the real data as much as possible and minimize the use of full imputation. Since new files are received from CCRA, the potential of using this information to improve our imputation will also be evaluated.

References

René Beaudoin, GIFI TY2000 – Edit & Imputation Strategy & Documentation, Statistics Canada internal document, October 1, 2001.

René Beaudoin, Load and edit process description, Statistics Canada internal document, April 9, 2001.

Richard Belcher, GIFI 2000 – Historical edits (Outlier detection), Statistics Canada internal document, April 23, 2001.

M.A. Hidiroglou and J.-M. Berthelot, Statistical Editing and Imputing for Periodic Business Surveys, Survey Methodology, June 1986, Vol. 12, No. 1, pp. 73-83, Statistics Canada.

René Beaudoin, GIFI Analytical Tables, Statistics Canada internal document, May 25, 2001.

René Beaudoin, Review and Correction, Statistics Canada internal document, June 21, 2001.

Richard Belcher, GIFI 2000 Historical Imputation Specifications, Statistics Canada internal document, July 17, 2001.

Richard Belcher, Caroline Rondeau and Eric Zentner, GIFI RY2000 – Pre-GEIS Process Specifications, Statistics Canada internal document, October 26, 2001.

Richard Belcher and Caroline Rondeau, GIFI RY2000 – GEIS Process Specifications, Statistics Canada internal document, September 28, 2001.

Richard Belcher, GIFI 2000 – Post-GEIS process specifications, Statistics Canada internal document, November 19, 2001.
