WP. 5
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Neuchâtel, Switzerland, 5–7 October 2009)

Topic (i): Automated editing and imputation and software applications

AUTOMATED EDITING AND IMPUTATION SYSTEM FOR ADMINISTRATIVE FINANCIAL DATA IN NEW ZEALAND

Supporting Paper

Submitted by Statistics New Zealand[1]

I. INTRODUCTION

1. New Zealand produces economic statistics based largely on sample surveys of businesses. Administrative data is used within the survey process to produce sampling frames and for stratification. Increasingly, administrative data is also used to produce estimates for small simple units in an effort to reduce response burden and survey costs. Recent redesigns of business surveys have been undertaken within the context of a new statistical architecture for economic statistics at Statistics NZ. In this context, the word 'architecture' describes an integrated and systematic approach to the collection and organisation of data that will support current and future information needs. This architecture is intended to reduce the volume of information that has to be collected by direct survey from people and businesses, while increasing the range and usefulness of the information produced. The new statistical architecture will be supported by increased use of administrative data.

2. In New Zealand, the tax department (Inland Revenue) collects financial information annually from businesses via the Accounts Information (IR10) return Approximately 500,000 businesses could file an IR10 each year in New Zealand and the majority of them are small to medium sized enterprises. Statistics NZ receives IR10 data monthly from Inland Revenue. On receipt, the data is processed before being made available to users.

3. The IR10 consists of a summary of income and expenses (the profit and loss account) and a summary of assets and liabilities (the balance sheet). IR10 data items are conceptually similar to data items collected in Statistics NZ's main business financial collection, the Annual Enterprise Survey (AES). The AES collects detailed information on financial performance and position by broad industry groups and institutional sector and forms the basis for the calculation of annual benchmarks of current price Gross Domestic Product. This data is currently used in the AES for individuals and partnerships. As a result of the redesign of the AES in 2009, Statistics NZ has expanded the use of IR10 data substantially for small simple businesses.

4. For the use of administrative data in business surveys, six key issues need to be considered: coverage, timeliness, frequency of updating, validity, reliability, and consistency (Hoffmann, 1995). These issues are important for the Annual Enterprise Survey as it is the backbone of Statistics NZ's annual business financial data collection. The particular nature of these quality issues in IR10 data and analytical use to be made of the estimates determine the choice of editing and imputation (E&I) methods.

[1] Prepared by Allyson Seyb, John Stewart, Grace Chiang, Ian Tinkler, Lee Kupferman, Val Cox and Darren Allan of Statistics New Zealand.

Any discussion of data limitations or weaknesses is in the context of using the IR10 return for statistical purposes, and is not related to the ability of the data to support Inland Revenue's core operational requirements.

5. The core of the IR10 editing and imputation system is Banff, an editing and imputation system developed by Statistics Canada. Banff preserves as much of the original data as possible, provides a good range of editing and imputation methods, and provides clear E&I audit trails.

6. This report describes the methodology used to edit and impute IR10 data. Section 2 outlines the editing and imputation strategy adopted. Section 3 describes the editing procedures and imputation procedures. The documentation of the editing and imputation processes is discussed in Section 4 and Section 5 contains the conclusion.

II. EDITING AND IMPUTATION STRATEGY FOR IR10 DATA

7. Statistics NZ's approach to editing and imputation has three objectives (Savage, 2007):

(a) Provide users with fit-for-purpose, plausible data and outputs by the most effective and efficient means.

(b) Ensure all users are well informed about the quality of the data and statistical outputs

(c) Continuously improve our end-to-end business processes and overall data quality.

8. The editing and imputation objectives are supported by seven key principles :

(a) Statistics NZ should maintain, wherever possible, the original data provided by the supplier.

(b) Choose editing and imputation methods that support the two main uses of the data (aggregates and microdata analysis). For example, choose imputation methods that maintain the internal consistency of the IR10 form.

(c) Make the best use of auxiliary and historical data, for example, using Goods and Services Tax data to confirm IR10 data quality.

(d) Automate the editing and imputation process, where possible, ensuring the best possible use of editing resources, for example, minimising clerical intervention. The existing editing approach for IR10 data was mainly manual and targeted the largest units.

(e) Keep an editing and imputation audit trail:

- including unedited and edited data so that sources, types and distributions of errors can be monitored
- including unimputed and imputed data so that the degree, methods and sources of imputation can be monitored
- including producing, monitoring and analysing E&I diagnostics to evaluate and understand the process, including its cost effectiveness and efficiency.

(f) The software application should support continuous improvement, and be flexible and configurable by users to enable future developments.

(g) Wherever possible, new developments should use an editing and imputation tool in the Statistics NZ set of standard tools.

III. DATA EDITING AND IMPUTATION

A. BANFF

9. Statistics NZ has evaluated and selected a number of standard tools to meet the editing and imputation needs in our business surveys. One of these tools is Banff. This project aimed to make the best use of Banff, rather than looking for another off-the-shelf tool or building something in-house. Banff is an editing and imputation system developed by Statistics Canada. Banff is designed to edit and impute continuous numeric data, so is most useful for processing economic or financial data. The system is made up of a collection of specialised SAS procedures, each of which can be used independently, or put together, to satisfy the edit and imputation requirements of a specific collection.

10. Banff preserves as much of the original data as possible, provides a good range of editing and imputation methods and provides clear E&I audit trails. A series of Banff procedures, together with some additional Statistics NZ SAS functionality, provides an end-to-end E&I process that is automatic, consistent, repeatable, and in which the quality is controlled predominantly via parameter settings.

11. The basic functions, or procedures, available in Banff are:

- PROC VERIFYEDITS – allows users to specify and check linear micro edit rules.
- PROC EDITSTATS – applies a group of edits to respondent data, determines the status of each record (pass, fail, miss), and produces tables summarising the status codes.
- PROC OUTLIER – identifies outlying observations using a method described by Hidoroglou and Berthelot (1986).
- PROC ERRORLOC – applies the edit rules to the respondent data, and identifies the minimum (weighted) number of fields that must be changed (imputed) so that the record passes all of the original edit rules, using the Fellegi-Holt method.
- PROC DETERMINISTIC – imputes fields with only one possible value that will allow the record to pass the original edit rules.
- PROC DONORIMPUTATION – imputes records using a nearest neighbour form of donor imputation.
- PROC ESTIMATOR – imputes individual fields using a variety of 20 pre-defined estimators (eg mean, ratio, regression), or user-defined estimators.
- PROC PRORATE – analyses groups of data and adjusts the data one record at a time so that they add to a specified total.
- PROC MASSIMPUTATION – in two-phase sampling, creates a complete rectangular file for a first-phase sample by donor imputing the missing information for the non-sampled second-phase units.

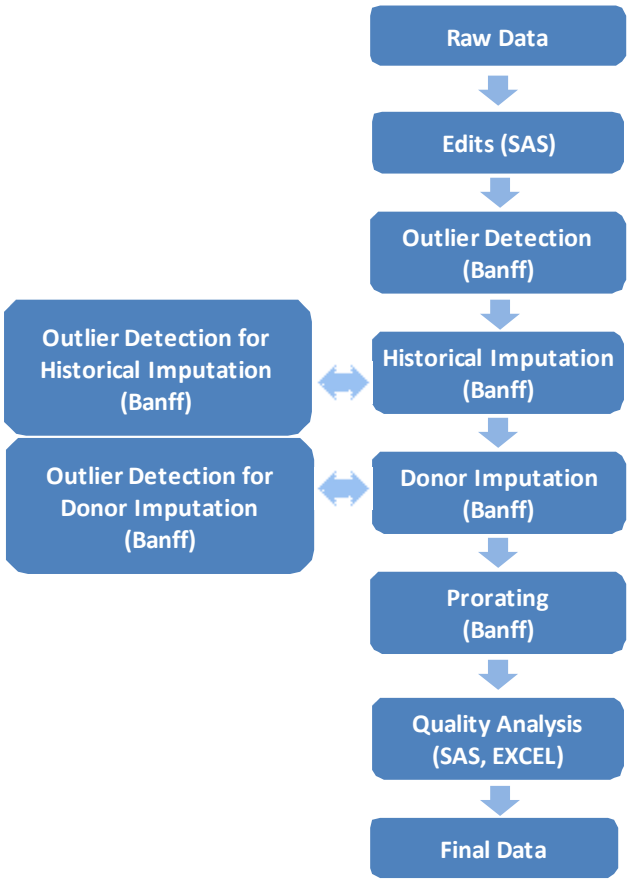
For full descriptions of each procedure, refer to the user guide *Functional Description of the Banff System for Edit and Imputation V1.04* (Statistics Canada, 2005).

B. IR10 EDITING AND IMPUTATION SYSTEM

12. The IR10 E&I system has been developed as a coherent process that edits and imputes the raw IR10 data and then provides analytical views to assess the quality of the edited and imputed data. This system is designed to be run end-to-end automatically with the output of one procedure feeding smoothly into the following procedure.

13. The IR10 E&I system consists of an initial set of user-defined editing and imputation procedures, a series of Banff procedures, and an Excel-based analysis tool. Process flags at each step indicate which value has been manipulated by which method. See the flow diagram on the following page for the order of the processes and further discussion.

FLOW OF E&I PROCESSES



C. EDITS

14. Administrative data quality is dependent on the statistical suitability of the return that is completed and the way it has been interpreted by respondents. In the case of the IR10 return, there are particular issues with both the return layout and respondent interpretation that affect the data quality. Edits to detect logical errors arising from these issues have to be corrected, irrespective of size, to have consistent data. Therefore, initial data corrections are necessary before running Banff procedures.

15. The IR10 return is set of financial accounts that logically groups transactions into nine 'balance' groups. For example, the trading account consists of: Sales - (Opening Stock - Purchases + Closing Stock) = Gross Profit. Therefore, the IR10 return has internal relationships that enable logical errors to be identified and corrections to be made. Moreover, there are common response errors arising largely from misinterpretation of the return. These can be analysed and rules for correcting common problems formulated.

16. Some examples are:

- The form asks the respondent to transfer the Gross Profit value from their accounts. Many respondents don't have a 'Gross Profit' in their accounts, as it is only applicable to businesses that sell goods. In these cases, Gross Profit will be missing (or zero) whereas it should be equal to Sales.
- The form asks respondents to add up Gross Profit and other income to obtain Total Income. This calculation assumes that the Sales value has been included in Gross Profit. If as noted above, Gross Profit is missing, some businesses have filed correctly by placing their Sales in Other Income in their calculation, whereas some report the wrong Total Income.

17. Such logical edits dramatically reduce the number of errors present in the data, which shows the importance of taking into consideration the practical aspects of the data collection when processing administrative data.

18. Some of the logical errors could have been corrected using Proc Deterministic in Banff. However, many of the errors required complex conditional edit rules to identify and resolve them. As a result, we developed our own edit functionality in SAS to carry these out.

D. OUTLIER DETECTION

19. Following the correction of logical errors, the data is processed to identify values requiring imputation. This identification is done by a three-step process combining Banff Proc Outlier, a size threshold, and confirmation against auxiliary data. Edits are mainly ratios such as a year-on-year value movements. Range edits were originally proposed but later considered redundant.

20. Firstly, the data is processed by the Banff procedure Proc Outlier. Very suspicious values, or outliers, are identified as potentially requiring imputation. We use two types of edit to identify suspicious values. These edits are year-on-year comparison for the same record and current-period comparison to other data within the record. The year-on-year comparison will identify any unusual movement of variables at the unit record level. We expect that the year-on-year comparison will detect the majority of errors appearing in the IR10 data. Where there is a strong relationship between variables within a record, this can be used to detect errors in the current period. For example, the relationship between sales and purchases of a business can be used to check for potential outliers. Values that are very different from their comparable values can be considered to be in error. Such values are flagged for imputation. Some other suspicious values, which are not sufficiently extreme to be considered an error, are flagged for exclusion from use in the imputation methods. Outlier detection is carried out for all numeric variables on the IR10 return.

21. Proc Outlier enables us to identify suspicious values. However, many of the edits used to identify these values as suspicious are unreliable when applied to small values. Proc Outlier includes an option to incorporate the size of the value as a factor in the outlier determination. This means that small deviations in a large value can be deemed more suspicious than similar deviations in a small value. With large administrative datasets like IR10, it is difficult to avoid flagging many small units even with the highest sensitivity to size incorporated in Proc Outlier. Also, unfortunately, too high a sensitivity to size results in even modest changes in large values being erroneously identified as suspicious. The resolution for the IR10 E&I was to incorporate via SAS a threshold size at which suspicious values would be accepted and not flagged as such. While all records pass through the Proc Outlier procedure the resultant outlier identification was then subject to reversal where appropriate. Thresholds can be varied by variable, by industry and for different types of businesses using a user-configurable interface.

22. Within Statistics NZ, we have robust auxiliary data available that can be used to further improve the identification of suspicious values. In particular, and relevant to the IR10 collection, we have comprehensive administrative measures of business sales and purchases, and a measure of salary and wage payments. These particular auxiliary data enable us to confirm whether the corresponding IR10 value is in error or not. As the variables sales, purchases and salaries, and wages are key data within the IR10, this provides a significant quality enhancement to our outlier detection process overall. After testing, it was found that Proc Outlier in combination with auxiliary data comparison was roughly twice as good at correctly identifying errors as either method on its own. This combination of edits, using Proc Outlier, a size threshold and confirmation against auxiliary data makes best possible use of methods and data available. We have implemented this method into a single, seamless process that can be tailored to various data, or subsets of data, using parameters entered in a user interface.

23. As noted, outlier detection is also conducted before historical and donor imputation. Outlier detection before historical imputation checks for unusual movements between years in a data item. Units identified as outliers are excluded from the calculation of the forward movement factor in historical imputation. Outlier detection before donor imputation checks for unusual current values in a data item. Units identified as outliers are excluded from the donor pool.

E. HISTORICAL IMPUTATION

24. Historical and donor imputation are used in the IR10 system. They impute missing records and erroneous records. These methods align with the imputation strategy of preserving the relationships between data items on the IR10 return. Imputation methods considered in the preliminary design were historical, donor, regression, and mean imputation, and each of the methods was carefully assessed. Both methods were discarded because neither preserves the relationships between data items on the IR10 return: mean imputation can distort the data distribution and result in unusual movements within unit records, and only three data items of the 59 on the IR10 return have potential auxiliary data that could be used in regression imputation.

25. Historical imputation means the missing value is replaced by the value declared at the previous collection, but modified according to a trend calculated from a group of similar units. Imputation is carried out in imputation cells – groups of records of a similar size and industry. Only units with previous data are eligible for historical imputation. This is the first imputation method applied after editing.

26. The Banff function DIFTREND is used for imputing IR10 data items. The equation of this function is:

$$\hat{y}_{IC} = \frac{\bar{y}_C}{\bar{y}_H} y_{IH}$$

This means the value from the previous collection for the same unit, with a trend adjustment calculated from the ratio of reported values for the variable, is imputed. The numerator and denominator are based on the same units. For example, if a unit provides the sales figure in the current period, but not in the previous period, its current sales figure will be excluded from the ratio calculation in the numerator.

27. Imputation is performed using all eligible records. The number of eligible records is a dimension of the quality of the imputation. Imputation cells, original values, denominator, numerator, and trend adjustment factors are also preserved for quality checking purposes.

F. DONOR IMPUTATION

28. Units without historical information available are imputed by donor imputation. Donor imputation obtains the replacement value from a similar record in the current file that has passed the edits. Banff uses a nearest neighbour approach to find, for each record requiring imputation, the valid record that is most similar to it. Donor imputation takes all fields requiring imputation from the same donor and this approach retains the relationships between the imputed variables.

29. In the IR10 system, nearest neighbours are identified based on auxiliary information such as industry and size variables from Statistics NZ's Business Frame and information on salaries and wages from our Linked Employer-Employee Data. This auxiliary data is comprehensive and of good quality providing a high degree of confidence in the quality of the IR10 donor imputation. Moreover, unlike survey data where donors, even 'nearest neighbours', may be quite different to the non-respondent despite our best efforts, the volume of administrative data provides a larger and more comprehensive set of records from which to choose a donor.

G. PRORATE

30. After imputation, units may need to be re-balanced. The purpose of the prorate procedure is to ensure that the sum of parts adds up to the desired total within each balance group in each record. An assessment of the relative quality of totals against components led to a two-pronged approach to re-balancing. Where the total is known to be of lower quality than the components, the components are accepted and the total altered; where the total is known to be of higher quality than the components, Banff Prorate is used to adjust the components. In Prorate, the components are adjusted relative to their size to achieve the correct total. The relative quality of a total and its components is largely determined by the E&I processes that each has been subjected to. For example, the Sales and Purchases data items are edited, so we expect their quality to be high. Gross Profit on the other hand is almost impossible to

edit because of its nature. So, we would consider the components of Gross Profit to be more reliable than Gross Profit itself.

H. QUALITY ANALYSIS

31. It is important that the quality of the E&I process is understood for two reasons. Firstly, the E&I process is seen as evolving – it is subject to continuous improvement. The collection of administrative data is typically outside the control of the statistical agency, and any changes in the underlying data need to be identified and the E&I process changed accordingly. Secondly, the automatic nature of the E&I process needs active assessment to avoid a 'black box' mindset that may lead to quality problems. Quality analysis takes two forms, namely, quality measures and output analysis.

32. The aim of quality indicators is to inform users about the data quality before and after the E&I process, and about the performance of the E&I process. Quality indicators can be calculated on any procedure used in E&I and can be calculated overall or by variable or record. The indicators may be used as standard reporting devices that are routinely calculated, and for optimising the performance of the E&I process (Luzi, Di Zio, Guarnera, Manzari, De Waal, Pannekoek, et al, 2007).

33. The indicators and measurements in the IR10 system are based on the set of indicators described in the EDIMBUS manual and include indicators of resources (eg elapsed time for the E&I process), indicators that can be used to monitor E&I methods (eg failure rate for an edit), and indicators that can be used in documenting E&I processes (eg imputation rates).

34. Output analysis includes assessment of the impact of the E&I process and an assessment and resolution of data issues. Quality analysis enables us to alter the process. The flexibility and parameter driven setup means changes can be made quickly and easily and the new configuration re-run to assess the impact of the changes.

IV. NEXT STEPS

35. Our next steps are to focus on improving the quality of the IR10 editing and imputation system, both from a methodological perspective and from the perspective of improving the way the system operates.

36. In addition, we plan to reuse the new system with another set of administrative data, the Goods and Services Tax (GST) return data. This may require the inclusion of additional E&I methods, primarily Banff procedures.

37. Once the system has matured, we plan to investigate the feasibility of using the system for survey data. This feasibility study will require the inclusion of additional functionality (for example, to carry out sample selection), with the ultimate aim of having a single configuration that carries out the entire survey production process from collection right through to dissemination.

V. REFERENCES

Hoffmann, E (1995). We must use administrative data for official statistics – but how should we use them?, *Statistical Journal of the United Nations ECE 12,* 41–48

Luzi, O, Di Zio, M, Guarnera, U, Manzari, A, De Waal, T, Pannekoek, J (et al) (2007). *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. ISTAT, CBS, SFSO, EUROSTAT. Available from: http://edimbus.istat.it/EDIMBUS1/document/RPM_EDIMBUS/RPM_EDIMBUS.pdf

Savage, T (2007). Statistics NZ Editing and Imputation Principles and Strategy V1.0. Unpublished technical report. Wellington: Statistics New Zealand.

Statistics Canada (2005). *Functional Description of the Banff System for Edit and Imputation V1.04*. Ontario: Author.