WP. …

ENGLISH ONLY

This paper was originally presented at this conference. It has since been updated to reflect work done during 2005 (updated 10 November 2005).

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ottawa, Canada, 16-18 May 2005)

Topic (i): Editing administrative data and combined data sources

USE AND EDITING OF ADMINISTRATIVE DATA IN THE BUSINESS INDICATORS UNIT, STATISTICS NEW ZEALAND

Invited Paper

Submitted by Statistics New Zealand, New Zealand [1]

I. Introduction

1. The Business Indicators Unit (BIU) at Statistics New Zealand focuses on measuring and reporting on indicators of current economic activity and trade flows for market sectors of the economy. Statistics produced by the BIU are used to underpin the aggregate macro-economic indicators of GDP growth, inflation and current account balance by:

- providing reliable, authoritative and timely short-term economic indicators that are relevant to the emerging needs of identified sectors in a cost-effective way, and that contribute to the understanding of current economic activity
- developing alternative data sources to reduce respondent burden
- developing new economic statistics.

2. Statistics New Zealand uses administrative data across a range of economic data collections in the BIU. The administrative data is used for two purposes:

- as stand-alone data for output, which removes (or reduces) the need to survey and, therefore, also mitigates survey development and collection costs
- to supplement existing surveys, which reduces respondent burden and collection costs.

3. This paper summarises the use of administrative data in the BIU at Statistics New Zealand and then details two BIU collections (Overseas Merchandise Trade and the Business Activity Indicator), including their editing/imputation methods and current developments. See Appendix 1 for a summary about the two collections.

II. Summary of the business indicators unit data collections

1 Prepared by Vera Costa and Blair Cardno.

4. The administrative data used by the BIU comes from several organisations and is used in a variety of ways:

- Inland Revenue (IRD) supplies Goods and Services Tax (GST) data, which are used for the Business Activity Indicator (BAI) series. The BAI data is used to supplement the Wholesale Trade, Retail Trade, and Manufacturing Surveys, removing the need to survey small to medium-sized businesses. Between 75 and 85 percent of 'surveyed' businesses are replaced by BAI data, comprising 15 percent of the total value. The BAI is also modelled at the regional level and output as the Regional Economic Indicator.
- The New Zealand Customs Service (Customs) supplies trade data. Export and import transactions are used to produce Overseas Merchandise Trade statistics. Export, import and excise data are used for alcohol and tobacco statistics. The trade data is also used to derive the Overseas Trade Indexes.
- Local government agencies (Territorial Authorities (TAs)) supply building consent data for output of building consent statistics. The building consents are also used for the sample selection of the Quarterly Building Activity Survey (QBAS).

5. Each of the BIU data collections has one administrative data provider, except for building consents. Building consent data is sent electronically from 74 TAs for output on a monthly basis producing the challenge of obtaining data in consistent file and data formats across the TAs. Data from Customs and IRD flows directly into the processing systems in an agreed electronic format.

6. Fortunately, building consent files total between 6,000 and 7,000 records per month, which is manageable. Having one data source and a consistent format for the other collections is key to allowing efficient processing and development of output. The GST data file is revised monthly and contains approximately 500,000 records. Customs supplies approximately 420,000 records each month, and data arrives daily at Statistics New Zealand.

7. Customs data files have complete information about each entry, including standard classifications. IRD and building consent data requires classifications to be added to the datasets on arrival at Statistics New Zealand. Industry classifications are added automatically to the GST data. The building type classification is added to the building consent data manually for each record due to information about each record being supplied inconsistently.

8. Two different editing strategies are employed. The first involves processing and outputting 'perfect' data. This involves editing every record, including following up every query related to each record, with the aim of providing 100 percent accurate data. No account is taken of user needs, nor the level of accuracy required in the final output, nor how much editing is required to produce these. The trade and building consent collections currently use this strategy and edit at the micro level, with the bulk of the edits applied manually.

9. The second editing strategy can be described as doing what is possible with the resources available, which is not always the most efficient or best practice approach. The BAI currently uses this strategy. More detail about the BAI and Trade editing methods is given later in this paper.

10. The desired editing strategy is to process and output 'fit-for-use' data. This is the preferred option being developed at Statistics New Zealand and will be the likely direction taken by BIU data collection editing in future. In simple terms, a 'fit-for-use' editing strategy requires an understanding of what the data is used for and then specifying the data quality level that is applicable for the collection. The level of editing is intended to meet the specified quality level.

II.1 Investigations into new administrative data developments

11. A project began late in 2003 to investigate the availability and possible uses of electronic transfer data (ETD), (commonly referred to as EFTPOS data) in New Zealand. This is a joint exercise between

Statistics New Zealand and the Reserve Bank of New Zealand. The data is owned by two private companies referred to as switching houses.

12. To date, the project has concentrated on defining the data required from the companies and ensuring data supply. Classification issues are also being resolved so that data can be compared with other published statistics. Early investigations indicate that the data compares favourably with the official retail trade statistics published by Statistics New Zealand. Potential uses of the data are as an early indicator for retail trade statistics and to supplement retail trade sales data. It may also have potential uses in providing information on sales for industries that are outside the coverage of the retail industry, retail sales by region, and expenditure by overseas visitors in New Zealand. Work continues on proving these uses.

III. Overseas merchandise trade

III.1 Overview

13. The purpose of the Overseas Merchandise Trade (Trade) statistics is to provide statistical information on the export and import of merchandise goods between New Zealand and other countries.

14. Trade data is provided by exporters/importers and their agents to the New Zealand Customs Service. This data is processed and passed on to Statistics New Zealand for compilation into statistics. Considerable reliance is placed on the exporters/importers and their agents submitting correct codes and information, but before the data is released for statistical purposes it is validated and detected errors are corrected.

15. Imports are generally compiled by the date they are cleared for entry by the New Zealand Customs Service. Exports are recorded by their date of export.

16. All overseas merchandise trade data is classified according to the New Zealand Harmonised System (HS) Classification, which is based on the International Harmonised System.

17. The monthly, quarterly and annual outputs include the following data at various levels of detail:

- Overseas merchandise trade (import and exports), actual and trend values
- Exports by destination
- Imports by country of origin
- Exports and imports of main commodities
- Imports by Broad Economic Category group

III.2 Current editing and imputation system

III.2(a) Capture and load

18. Customs uses tariff items and statistical keys and other information contained in the import, export and excise entries to electronically target goods of interest to it and other agencies, and to collect due revenue. Post-border auditing of documents and premises allows Customs to gain a level of assurance over compliance, including accuracy of Trade reporting and return of revenue. Data from such entries and declarations is also analysed for strategic purposes, and extracted for release to other agencies, industry and the community.

19. Customs also undertakes agreed data checks and corrections before sending Statistics New Zealand daily text files. For example, it conducts reality checks on values and corrects missing entry identifier information. Statistics New Zealand also has an employee working in the Customs audit team one day each week to deal with queries.

20. Each text file is supplied according to agreed specifications for loading into the Trade system. The file contains new import and export entries, amendments, deletions, restored entries, and new client details. It also contains excise data that is forwarded to another area of Statistics New Zealand for output.

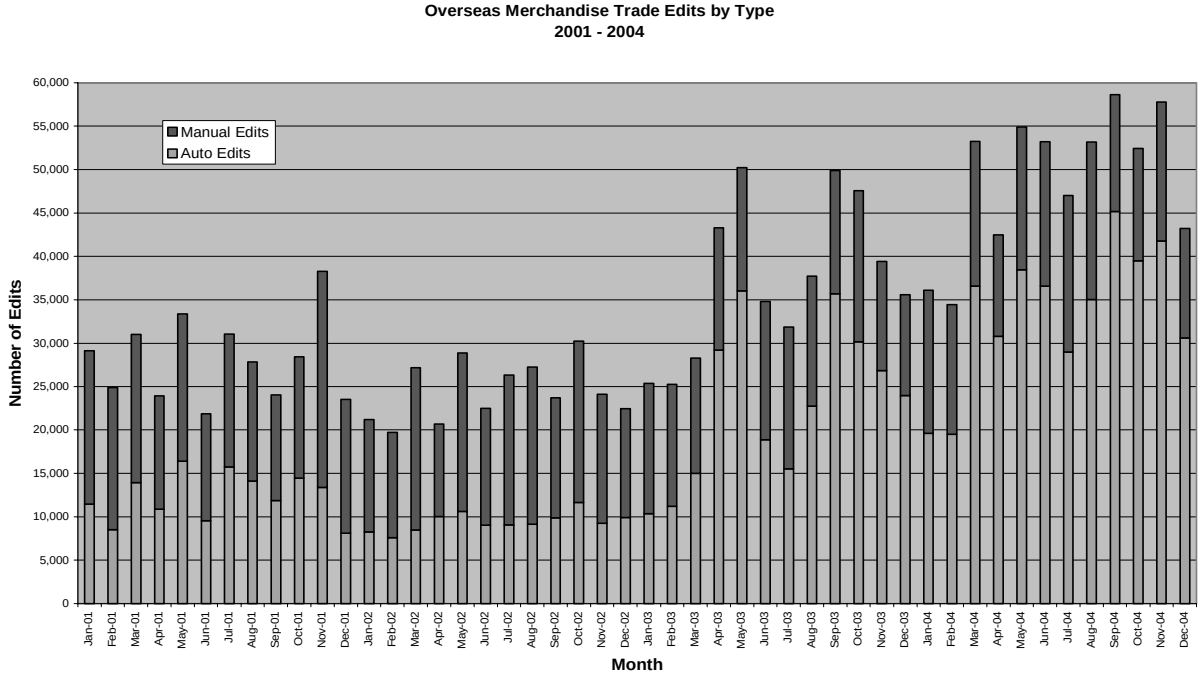
21. The daily text file is run through a capture/edit program, which checks the file for the correct formats and does an inter-record validation (checking for duplicates etc) on each entry. Any formatting (fatal) errors are repaired before the file is loaded into the edit database. For example, if a tariff number has fewer than an expected number of characters, then the file will not load until the number is corrected.

22. A system of auto-edits introduced into the Trade system in 1999 allows automatic changes to be made to simple errors where a change can be made without intervention or an editor having to use discretion. Up to three criteria are used for each edit, and if all are met for an individual entry the field is automatically corrected (deterministically imputed with known corrections). For example, exporting lobsters (HS code 030622) from New Zealand is not possible, but exporting rock lobsters (HS code 030621) is possible because they live in our waters. Therefore, if an entry shows an export of lobsters (HS code 030622) from New Zealand, it is automatically changed to rock lobster (HS code 303621). After auto-editing, the entry stays in the edit database and remains subject to all other error checks.

23. Before establishing a new auto-edit, care is taken to ensure the change is likely to be accurate and will not affect other data. Existing auto-edits are reviewed periodically to ensure effectiveness. However, the impact on data quality due to the introduction of auto-edits has not been quantified.

24. The introduction of the auto-edit system has been successful in reducing the number of manual edits and has consequently saved resources and time. In 2004, manual edits made up on average 31 percent of all edits per month, compared with 100 percent prior to the introduction of the auto-edits in 1999. Edit team resources were struggling to cope with increasing volumes prior to 2000. Since the introduction of the auto-edits, no overtime has been worked and more time is available for quality analysis of the resulting statistics.

Figure 1: Overseas Merchandise Trade Edits by Type



III.2(b) Manual editing

25. The loaded entries (including those that have undergone auto-edit) are next checked against the edit system. Entries that fail are flagged as 'errors', which are definitely incorrect and must be fixed, and 'warnings', which require further investigation but not necessarily any changes. There are more than 100 different errors and warnings in the system. An error typically involves an impossible situation. For example, an error notification appears if the processing port is not valid when checked against standard port codes. A warning signals a suspicious situation that is possible and therefore requires reality checking. For example, a 'warning' notification appears if a value is outside an expected range for an item. The editor follows predetermined instructions for investigating and correcting the errors/warnings. The edits have developed and evolved over time and decisions on different issues are based on historical experiences.

26. The unit value ranges are expected prices per unit. These are updated using finalised data on a quarterly basis to ensure the unit value range moves with the market. The process to update the ranges is simple, but the results require time-consuming manual validation. Analysis is undertaken at the most detailed HS level (10) of the 300 largest and 300 smallest differences between the old and new ranges to investigate the impact of outliers or small numbers of entries used in the range update.

27. Individual entries with errors/warnings are edited manually. Every entry that passes the errors/warning traps successfully remains in the edit database for macro checks.

28. The editors can run queries on 'warnings' based on historical data. They can compare previous data for the same entry type over specifiable times and can also search client details to reference previous client entries on the system.

29. Entries in the edit environment can be altered at any time. Revisions to the output database are very rare and are major changes. There has been only one since 1996.

III.2(c) Pre-output assessment (macro checks)

30. Before the data is 'handed-over' from the edit environment to the output environment, pre-hand-over checks are done on special entries (eg million dollar entries). Pre-set reports and checks on various fields (eg tariffs, ports, values) are undertaken and any issues are resolved by manual queries to the system or even back to Customs or the brokers.

31. At the end of the month, before hand-over to the output database, macro totals are produced to gauge significant movements. Individual transactions that contribute to large aggregate changes can be identified by drilling down into the data by HS code. The values at the micro and macro levels are analysed against percentage and absolute value changes to determine if the value is higher or lower than the expected range. The ranges are arbitrary and used as a guide. If a value is outside the range, the value is investigated with other information such as industry knowledge and explained. The purpose of these checks is to pick up incorrect data that may have been missed by the micro editing. It also enables validation of 'output' data in the edit environment through analysis, and provides key explanations for any movements in the Trade series.

32. Quality/query work to follow up significant entries from editing actions happens throughout the editing process.

33. Trade data is considered final after four months and changes are made to the output database.

III.2(d) Handover to output

34. A hand-over program updates the output database. The update covers four months – three provisional months and the month to be finalised.

35. During the hand-over process, the gross weight for each entry is apportioned across the different items in the entry. The system uses a combination of net kilogram quantities (where available) and gross weight factors to apportion the total gross weight to all lines in the entry. This provides a standard measure for all commodities. The gross weight factors used in the apportionment calculations are expected ranges using a standard base for all commodities and are expressed in kilograms per dollar. The factors are updated for price changes resulting from external influences (eg exchange rate changes) and, like the unit value ranges, require manual validation.

III.2(e) Summary of edits

36. Average number of errors/warnings per month in 2004 = 57,854
Average number of manual corrections per month in 2004 = 15,167
Average number of automated adjustments per month in 2004 = 33,548
(Note: Approximately three-quarters of automated adjustments are due to an exchange rate indicator compatibility issue.)

III.3 Future developments

37. Initial research carried out as part of the Overseas Trade Review in 2004 indicated that there was potential to make resource savings in the Overseas Trade area. Editing currently accounts for 41 percent of the Trade section's baseline budget. There are two main projects under way to improve the editing and imputation processes currently used in Trade to move towards a fitness-for-use editing approach.

III.3(a) Threshold editing

38. The key objective of introducing threshold editing to the Trade data is to meet an immediate need to reduce the amount of editing action (because of a reduction in staffing).

39. The aim is to have the threshold editing methodology in the production environment by 30 March 2005. This involves producing a methodology that is easily introduced and has minimal impact on current editing procedures and the quality of key outputs. All other existing editing systems will remain in place when the threshold editing method is introduced.

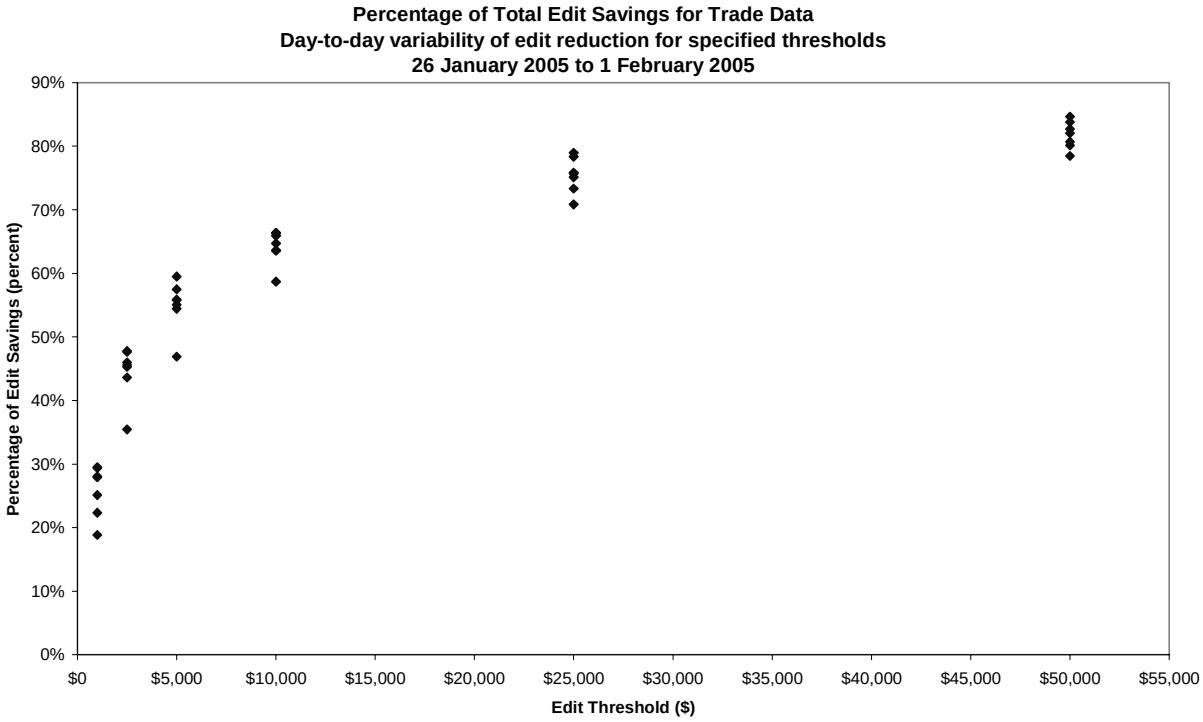
40. It is proposed that 'small valued' entries for imports and exports with non-missing dollar values below a specifiable threshold not be brought up for manual editor action. The threshold relates only to 'warning' edits, which require editor action. All formatting errors continue to be actioned as before.

41. In early testing, the effect of edit changes for levels HS2, HS4, and HS10[2] and country estimates were analysed for April and May 2004. The analysis indicated that the top 10 percent of changes by raw value were the only ones that made a significant change to the final estimates. Almost no change at high HS levels was perceptible from the changes made to the lower 50 percent by raw value.

42. The threshold has yet to be determined, but its introduction is expected to reduce the editing effort by a significant amount, as Figure 2 shows.

2 The Harmonised System used for Trade classification has several levels of detail. HS2 is the basic level, HS4 is the more detailed level below HS2, and HS10 is the most detailed level.

Figure 2: Percentage of Total Edit Savings for Overseas Merchandise Trade Data



43. The advantages of the threshold approach are that it can be implemented relatively easily and quickly, and the editing team can continue its current procedures without needing to train in new procedures.

44. The key disadvantage is that setting the threshold predetermines the quality of the output data. However, the impact on quality has been estimated in advance by analysis of back data.

45. Using this approach, regular changes to the threshold value are possible. Historical data will be analysed and continual review will indicate when (or if) threshold changes are necessary. The history of any threshold changes, with the date and identity of the change agent, will be retained for auditing purposes.

46. The small valued entries will bypass the edit system altogether and move unchanged to the output database. High risk entries will continue to come up for editing action regardless of their value if there is large quantity or large gross weights.

47. Since the writing of this paper the threshold approach has been introduced. It was decided initially to be very risk averse and introduce a threshold of $1000. This reduced the amount of editing by 25 percent. This threshold was re-evaluated two months after its introduction and it was decided at that time that hadn't quite achieved our target of reduction of workload, so after more analysis, it was decided to raise the threshold to $2500. The threshold was then raised to $5000 three months after the initial introduction of threshold editing. This has meant that the editing effort from editing 12 percent of entries has been reduced to 3 percent, without any reduction in the quality of published statistics.

III.3(b) Selective editing project

48. The external environment in which merchandise trade statistics are collected has changed in response to new international security or anti-terrorism measures. This has improved the integrity of export entries, which are the primary source for trade statistics. To minimise the likelihood of Customs intervention in export transactions, New Zealand trading partners and on-the-wharves trading partners,

New Zealand exporters are becoming 'Secure Export Partners'. To become a secure export partner, exporters must have a security plan approved by Customs, and security seals on consignments, and be subject to auditing.

49. Statistics New Zealand is in the process of moving to a more efficient business model. As part of its Business model Transformation Strategy (BmTS), project business processes are planned to be generalised, parameterised modules that will support the various statistical outputs.

50. Against this background of improved data collection and with the key outcome for Statistics New Zealand to deliver value, it was considered timely to introduce selective editing as a means of introducing more targeted editing effort by concentrating on the editing that will make a difference to the published totals and to develop a generalised selective editing methodology to fit into the BmTS.

51. The Selective Editing Project will develop a generalised selective editing methodology to fit into BmTS. Trade data will be used as the initial case study due to its potential to provide large resource savings (as identified in the Overseas Trade Review 2004), before the methodology is tested in other areas.

52. Since the writing of this paper a selective editing methodology has been finalised and is awaiting implementation into the overseas trade system. This is expected to happen in early 2006 and will replace the threshold approach.

IV. Business Activity Indicators

IV.1 Overview

53. The BAI uses GST data from Inland Revenue (IRD) that is matched to the Statistics New Zealand Business Frame (BF) to provide monthly indicators of business activity in New Zealand by industry, two months after the reference month.

54. Goods and Services Tax is a tax on the consumption of most goods and services in New Zealand. Because it is charged on most taxable activities, the data includes almost all New Zealand businesses.

55. Statistics New Zealand has had access to IRD tax records since the introduction of the IRD's computer system in 1995 and is currently using GST records from the IRD to update and maintain the frame of New Zealand businesses (BF).

56. In summary, the BAI data processing involves:

- initial data editing at IRD to look for invalid IRD numbers
- data extraction on the first Sunday of each month. The IRD extracts from its client registration system all active taxpayers that are not classified as individual salary and wage earners. The file sent to Statistics New Zealand has about 500,000 records
- data matching of GST units by their IRD identification number to the BF so that industry and employment variables can be added
- data manipulation, to deal with two issues:
    - Businesses may choose to form a group for GST purposes. Where a group return has been filed which covers a number of enterprises, GST values are apportioned by size to the enterprises included in the group
    - Businesses can submit GST returns either monthly, two monthly or six monthly, depending on their annual turnover. To standardise the reference period to monthly, the GST value in a return is apportioned through modelling over the reference period of the unit
- data editing, which is done manually to correct values identified as driving the movement of some industries between the previous and the current month.

IV.2 Outputs

57. The BAI was initially released as an experimental series. This meant the series was subject to revision while the methodology underwent further development. Seasonally adjusted and trend series were produced for GST sales and GST purchases by 17 standard industrial classifications.

58. The Regional Economic Indicators (REI) series is derived from the BAI dataset. This series is also experimental and was developed with the objective of estimating economic activity at a local level. The estimates are model-based and include GST sale and purchase data for the 16 New Zealand regions by 17 standard industrial classifications.

59. In June 2001, the BAI series became part of the production system of the Quarterly Manufacturing Survey (QMS). BAI data is used in the QMS to model quarterly manufacturing data for small businesses in an effort to ease their respondent burden. Later, similar methodologies were introduced in redesigns of the Wholesale Quarterly Survey (WQS) and Retail Trade Survey (RTS).

60. In addition, annual GST sales are used as an indicator of size in some surveys (eg Annual Enterprise Survey, Quarterly Manufacturing Survey) in addition to the employment variable.

61. At the beginning of 2002, the business activity (GST) scheduled publications were temporarily suspended pending a review of the changes necessary to elevate the series from experimental to the status of an official statistic. This led to a system redevelopment in 2004 which included a revision of the whole processing system as well as the development of a completely different editing methodology.

62. This redevelopment is currently being implemented.

IV.3 Current imputation system

63. Depending on turnover, businesses file GST returns six monthly, two monthly or every month. Therefore, enterprises with no GST in the current month include:

- one, two and six-monthly filers who should have filed a return but did not (non-respondents)
- two and six-monthly filers who filed a return but need their data apportioned back over the reference period (temporal apportionment)
- two and six-monthly filers who are not expected to file a return and need their data forecasted for the next reference period (forward estimation).

64. Imputation for non-respondents is carried out at imputation cell level. Such cells are defined by a combination of industrial classification and filing frequency. Mean imputation is used to impute values for all non-responding BAI births (in the current month). Historical imputation is used to impute values for all enterprises that existed previously and did not respond in the current month. If there is no historical factor for a cell or historical data for the non-responding unit, then mean imputation is used.

65. Two and six-monthly returns are apportioned over appropriate previous months. The temporal apportionment uses seasonal factors in the case of two-monthly units and simple division for six-monthly units.

66. For the last five months of the series, some of the filing frequency types will not have responded because their next reference period end date is still in the future. Forward estimates for these periods are imputed using either historical or mean imputation.

IV.4 Current editing system

67. The current system has a facility for allowing the value for a unit to be edited. After all the apportionments and imputations are done, the units are grouped by industrial classification. The ones driving industry movement between the previous and current month are manually checked and corrected, if necessary.

68. For example, units may have capital spikes as sales and purchases of capital items (such as buildings or machinery) are included in the GST data provided by IRD. Capital items with the potential to breach the confidentiality rules are identified using macro edits at industry level and removed from the series. However, only capital transactions large enough to stand out at industry level are removed.

IV.5 Current issues

69. Although there are methodological issues related to imputation (forward estimation and non-response), group apportionment, temporal apportionment and producing data that can be reliably published at the regional level, the main concerns are about the stability of the whole system and the editing approach used.

70. The existing BAI editing system involves manually identifying and removing large capital transactions from the series, as well as correcting group structural issues. The BAI system is not user-friendly so only a limited amount of quality adjustments can be made each period prior to hand-over to the sub-annual surveys, in which edit checks identify any unusual movements in the data.

71. Many problems occur each period because most of the BAI analysis is aimed at fixing large companies' data and group apportionment problems, whereas the sub-annuals use data for the small enterprises, which are not targeted in the BAI editing. The problems identified during monitoring within each sub-annual processing system often take time to correct. This impacts on the amount of time available to analyse survey data prior to publication.

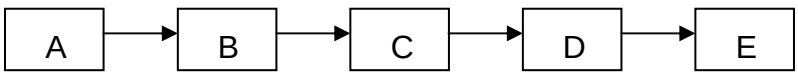
IV.6 Proposed changes

72. Significant system changes are being made during the redevelopment project. These changes have been concentrated around two aspects:

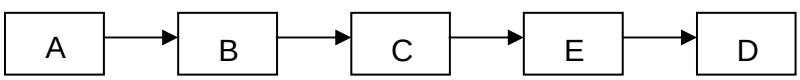
- improving the stability of the processing system to minimise potential hand-over delays to the sub-annual surveys as well as improving its user-friendliness
- developing an automated editing system, including clear principles on how capital transactions should be dealt with and moving the editing stage to earlier in the process.

73. The following diagrams show the different processing order:

Current system:



Future system:



In both cases, the modules represent:

A – Initial data editing at IRD
B – Data extraction
C – Data matching

D – Data manipulation
E – Data editing (the methodologies are different in the current and future systems).

74. The editing system being developed comprises:

- outliers detection – the values identified as outliers at this stage are sent to the editing phase
- outliers editing – the outliers are prioritised based on their contribution to a specific group. The values with a contribution above a fixed threshold are manually edited and the ones below it are automatically edited.

75. It is expected that most of the gains will be made in improving the existing system and business processes. However, there are a number of other potential benefits. The availability of GST-based sales and expenditure data on a monthly basis provides Statistics New Zealand with the opportunity to improve data quality in many new and existing quarterly surveys and to potentially lower its data collection costs. In addition, compliance costs on small businesses can be reduced by increasing the thresholds used to define the strata modelled based on tax data in current sub-annual surveys. This fits with one of Statistics New Zealand's key goals of reducing the overall burden of survey activity.

76. If the quality of the data is improved, Statistics New Zealand could eventually introduce it as a validation tool for different subject matter areas or for preparing external analytical papers.

77. A preliminary investigation of the expected benefits of the new editing system indicated that significant savings could be achieved. Currently, around 200 units are manually edited to exclude some identified capital spikes from the BAI series. As this targets only the largest businesses, the sub-annual surveys that use BAI data for small units in their tax strata need to revisit the editing process during their monitoring stage. This is another time-consuming process that is expected to be reduced after the new editing system is implemented.

78. In fact, a rough investigation of the editing work undertaken for the Retail Trade Survey showed that almost 90 percent of the 96 units in the tax strata that were manually inspected in November 2004 would have been automatically handled if the new editing system had been in place. Considering that this is a monthly survey and that the two other sub-annual surveys use BAI data to obtain the data for their tax strata, the new system can lead to significant changes in monitoring time.

IV.7 New editing methodology

79. The new editing methodology proposed for the BAI takes into account the nature of the data: all data to be edited is numeric, it is available at the unit record level, and a long time series exists for most units. As the BAI is a census of all businesses registered for GST, there are no weights involved in the data use, so this aspect does not need to be taken into account in the editing methodology.

80. The new methodology edits the data provided by Inland Revenue, prior to any adjustments being made. This avoids the problems encountered previously, when the data was edited at the end of the process; there is now no need to work backwards through any adjustments made (eg group apportionment).

81. The selective editing approach used now is more efficient and effective and better targets the effort spent editing data. It also has the flexibility to change the amount of resources employed to edit data in any given month. Introducing automatic editing of smaller units is expected to significantly reduce the time spent monitoring tax data that feeds into the three sub-annual surveys and to improve the REI estimates.

IV.7(a) Outliers detection

82. Outliers detection is based on the Statistical Process Control idea which expects, for a system under control, that all values of one variable will be within a certain range, which is a function of the

values previously observed. This has the advantage of taking into account the ‘natural’ variability of the data.

83. In Statistics New Zealand’s case, the GST values for each business are treated as a separate time series and the values from previous periods can be used to create the range within which a current value is expected to be. Therefore, the current value is compared with a range of values instead of a single expected value. The use of ranges of historical values avoids the need to adjust for any seasonal variation in the current period and simplifies the editing process.

84. Nevertheless, in order to create the expected range of values it is necessary to access the back series for each unit. Depending on the way the data is stored, this can be a challenging and time-consuming exercise, which would need to be repeated every period. At Statistics New Zealand, all data is stored in one table, having the enterprise IRD code and date (month/year) as unique identifiers. The table has around 74 million records at present, with about 500,000 added each month. This gives an idea of the amount of work that could be required to retrieve the back data.

85. Besides this, in order to have the same number of periods of back data for each unit it would be necessary to go back though different periods of time depending on the filing frequency of the unit (one, two or six monthly), the number of periods with missing data, the number of invalid data, and so on. This adds another complication to the retrieval of the back data.

86. To avoid these problems, a simpler process was put in place to summarise the five years of back data needed for creating the range of expected values for each unit. It is based on the fact that the mean of the previous values can be considered the best estimator of the location of such data, and the standard deviation represents the same regarding their variability.

87. It also takes advantage of the fact that, when a new value is added to the time series, there is a simple relationship between the new mean, the old mean, the number of observations in the time series and the new value. It can be represented as:

$$\bar{x}_t = \frac{x_t + (t-1) * \bar{x}_{t-1}}{t}$$

where:
x_t = current value
$\bar{x}_t$ = mean value up to the current month
$\bar{x}_{t-1}$ = mean value up to the previous month
t = total number of months up to the current month.

88. This expression shows that it is very easy to update the mean value for a unit once a mean for the back data is created for it.

89. A similar approach can be used to update the unit's coefficient of variation, although the expression involving the new value, the old coefficient of variation, the old mean and the number of observations in the time series is much more complex.

90. As the coefficient of variation is a measure of variation independent of the unit size, we used it to indicate the data variability. This has the advantages of being potentially more meaningful to monitor and to choose general thresholds for.

91. The current value for a unit is considered an outlier if it is outside the range with the following boundaries:

$$LowerLimit = \bar{x} * (1 - k * cv)$$

$$UpperLimit = \bar{x} * (1 + k * cv)$$

where:

$\bar{x}$ = mean value

cv = coefficient of variation

k = threshold (to be determined).

92. Although it is desirable to exclude the true outliers, it is also desirable not to disturb the original data too much. The analysis based on back data indicated that a *k* value equal to 3.3 was the threshold to be used. For normally distributed data, it is expected that the interval created using such threshold will contain 99.90 percent of the data.

93. This approach leads to an outlier detection methodology that is based on the creation of a single additional table for use in the editing system. This table contains some reference data for each unit as well as some data that will be updated whenever new data is reported. It has about 800,000 records because there is a single record for each IRD unit. Such table includes indicators to keep track of the edits made either manually or automatically, as explained later.

IV.7(b) Outliers editing

94. As described in the previous section, the outliers detection is done at the micro level, ie the time series for each business is treated separately. This can be done automatically, provided a threshold is fixed for setting the boundaries of the range of expected values.

95. On the other hand, depending on the threshold selected for outliers identification, the number of units requiring editing can be quite large and it would not be feasible to manually examine all of them. Therefore, further consideration is needed about the importance of the contribution of the unit to a certain group to decide whether it should be automatically or manually edited.

96. The comparison of the mean GST value for a unit that has a new GST value identified as an outlier with the mean total GST value for the group to which the unit belongs gives an indication of the importance of the contribution of the unit to the group's total. If this contribution is above a fixed threshold, the unit is selected for manual inspection to determine an edit value. Otherwise, the edit is done automatically, according to the editing rules developed. The main editing rule is to use the previous year's mean value for the unit.

97. One problem at this stage is that unit record GST values can be used for very different purposes; the ideal grouping for determining the importance of a unit is not the same regarding the estimation cells in the Retail Trade Survey and the region-by-division industry breakdown used in the Regional Economic Indicator, for example. Also, it is desirable that the groups not be too large (risking not being representative), nor too small (risking the creation of updating problems due to lack of data). Having many groups is undesirable for monitoring purposes too.

98. All these considerations led to the choice of the four-digit ANZSIC industry code as the grouping for determining the importance of the unit and whether any potential editing for it should be manual or automatic. Such choice took into account the input from the Statistics New Zealand output division involved with the BAI redevelopment.

99. As can be seen, the approach used for assigning manual or automatic editing to a detected outlier is based on the previously observed contribution of the unit to the group's total. The expected change to the group's total due to the editing process does not play a role in such decision.

100. The fixed cut-off for manual/automatic editing for each group is determined by the expected manageable number of manual edits. No reporting error distributions have been analysed to select such thresholds.

101. At present, the BAI is a key data source for three sub-annual surveys, as well as being used for the production of the Regional Economic Indicators. Therefore, any impact on data delivery and quality as a result of the redeveloped system must be well managed.

102. The analysis that led to the development of the new editing system indicates that it is not expected to have a large impact on the estimates from the sub-annual surveys. This is due to the current approach of further editing of BAI-sourced data within the survey when such data produces unusual movements between periods. On the other hand, it is expected that the new editing methodology will substantially reduce the monitoring time spent in these collections.

103. Until now, no assessment has been made of the impact of the editing methodology on the Regional Economic Indicators. At the moment, quality indicators prevent the data for some combinations of region and industry being released if it is considered not up to the required standards. After the new editing system is in place, more data is expected to meet the standards and be able to be released.

104. Due to the importance of the collections involved, it is planned to have both the old and new editing systems running in parallel for around six months, because the benefits of a long dual run are considered to outweigh the costs involved.

V. Conclusions

105. The two collections focused on in this paper are at different stages in their development of new editing methods. The key similarities between them are the aim to have more automated editing systems and the proposed introduction of a selective editing strategy. This editing strategy will enable the collections to move to a ‘fit-for-use’ data approach.

106. A ‘fit-for-use’ rather than a ‘perfect’ data editing strategy as used in Overseas Merchandise Trade or ad hoc manual editing as done in the Business Activity Indicators series should result in greater resource efficiencies for processing the data.

107. Developing a better understanding of the uses of the data and the needs of the data users, both within and outside Statistics New Zealand, is necessary to develop a ‘fit-for-use’ editing strategy. Gaining a better understanding of the data uses will allow informed decisions to be made on how best to apply selective editing methods.

108. The introduction of a generic selective editing strategy is one component of Statistics New Zealand’s move towards a more efficient business model. The challenges in developing a generic selective editing strategy to apply to both collections have yet to be identified. However, aspects of the collections need to be noted as part of the development. The table in Appendix 1 summarises some of these aspects, and a brief discussion of the most important ones is present below.

109. These collections are compiled from administrative data that is collected for different purposes. The GST data used for the BAI series is collected as administrative tax data without consideration for statistical uses and little editing is done by the agency which provides the data. On the other hand, the Customs data used for Overseas Merchandise Trade has the advantage of being collected with statistical use in mind. Also, Statistics New Zealand has an agreement with Customs to ensure the data arrives in the best form possible for statistical output.

110. Although the amount of data provided for such collections each month is about the same (500,000 records), the Customs data is provided every day, whereas the IRD data is received once a month.

111. In addition, the data used for the BAI series is numeric and advantage can be taken of the fact that there is a time series for each unit. This makes it easier to determine expected values that can be used in the editing process. The Overseas Merchandise Trade data is a mixture of numeric and categorical variables and, even for the numeric variables, it is difficult to use a time series approach because of the large number of different items and businesses involved.

112. The GST tax data requires much manipulation and modelling to achieve useful results, unlike Customs data.

113. A key aspect of developing a selective editing strategy is defining the groups to be used to determine the importance of an edit. This is an important difference between the two collections in this paper. The businesses for which tax data is provided can be grouped in several different ways according to the output to be produced. For example, businesses can be combined by geographic area to produce regional outputs, or by industrial classification to produce data for the Retail Trade Survey. The existence of several possibilities for arranging the data makes the choice of the grouping to be used for selective editing more difficult. As the outputs from the Overseas Merchandise Trade are not so diversified, the definition of groups for the selective editing process is easier.

114. Administrative data is crucial in producing statistical outputs in the Business Indicators Unit. While surveys remain the initial collection tool for some data collections, administrative data is continually being investigated for use as stand-alone data, and to supplement surveys in order to reduce respondent burden and survey costs.

115. Additionally, the amount of data available at different government agencies offers potential for the production of a new range of statistical information. However, the amount of data makes it mandatory to have automatic systems in place for data validation.

116. These considerations highlight the importance of the editing systems if good quality data is to be produced.

Appendix 1: Summary table

	Trade Data	**BAI**
Source	Customs.	Inland Revenue Goods and Services Tax data.
State when received	Basic validation.	Raw data as entered by the IRD.
Number of Units	Volumes per month: • Imports 300,000. • Exports 110,000.	Volume per month: • 500,000 records.
Uses	• Customised series output. • Input into quarterly GDP. • Input into balance of payments.	• Customised series output. • Supplementary data in business surveys. • Regional series.
Level of outputs	• Imports and exports at various detailed levels.	• GST sales, purchases and net sales, by 78 GST industry divisions.
Current types of editing	• Micro edits. • Macro and micro checks. • Automated and manual.	• Macro warning. • Micro edits. • Manual.
Current imputation	Automatic edits for known corrections. Gross weight apportionment for some entries.	• Non-response imputation. • Temporal apportionment. • Forward estimation. • Group apportionment.
Key issues	• 41 percent of the trade budget goes on editing. • Reduce editing effort and retain 'fitness for use'. • Move from 'perfect data' to 'fit for use' data.	• Imputation method. • Editing system. • Reliable regional data.
Future developments	• Threshold editing to reduce editing effort. • Selective editing (long term editing solution).	• Automated editing system. • Stability of the system. • Apportion GST data to regional level.