UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Paris, France, 28–30 April 2014)

Topic (v): International collaboration and processing tools

Migration of a large survey onto a micro-economic platform

Prepared by Val Cox, Statistics New Zealand

I.   Introduction

1.   In 2011 Statistics New Zealand began a 10-year transformation programme called Statistics 2020 Te Kāpehu Whetū (Stats 2020). The purpose of Stats 2020 is to address risks, realise opportunities, and create a dynamic, responsive, and sustainable organisation. It is an organisation-wide programme of change to carry out Statistics NZ's Strategic Plan 2010–2020, which focuses on four strategic priorities; leading the Official Statistics System, obtaining more value from official statistics, transforming delivery, and sustainability.

2.   By means of a recent project we will meet at least two of Stats 2020's strategic priorities: transforming how we deliver statistics and getting more value from what we have. First introduced in 2008, a micro-economic platform (MEP) has gradually been transforming the way we process economic surveys. The MEP provides a core platform enabling users to load, process, analyse, and prepare economic data for dissemination. It is designed to maximise the productivity of statistical analysts, providing users with a set of standard IT tools that are flexible enough to give them full control over their business processes. This allows them to design and build their own statistical outputs, empowering users by creating freedom from reliance on IT experts. The basic principle is that administrative data will be used wherever possible, with surveys filling the gaps. The objective is to bring core information about every business in the economy into the Longitudinal Business Database, allowing Statistics NZ to respond quickly to changing needs for economic statistics.

3.   This paper discusses the challenges of building a non-response imputation package for a large survey on the MEP, including the reasons for using both the Banff software for outlier detection and imputation, and the SEVANI (System for Estimation of Variance due to Nonresponse and Imputation) software to estimate sampling errors, taking into account the non-response in the data.

II.   Annual Enterprise Survey

4.   The Annual Enterprise Survey (AES) is the first large, complex survey to be migrated to the MEP at Statistics NZ. The AES is New Zealand's most comprehensive source of financial statistics. It provides statistics on the financial performance and position of almost the entire population of New Zealand businesses, covering most areas of economic activity and capturing approximately 90% of New Zealand's gross domestic product. AES is a complex survey involving four different major data sources; three administrative and one postal survey.

5.   The target population for AES is all economically significant businesses operating within New Zealand; approximately 440,000 units. AES uses a stratified sample design to select the sample from units on the Statistic NZ Business Frame. The Business Frame is a database of all known individual private and public sector business and organisations engaged in producing goods and services in New Zealand that meet significance criteria. Each industry contains between one and four strata, defined by size of turnover and rolling mean employment (RME). Turnover is sourced from the Goods and Services Tax (GST) supplied by Inland Revenue. Each industry has a full-coverage stratum made up of large units with significant economic activity within their industry group. Some units are identified as 'key firms', whose data is essential to the survey. The remaining strata contain a sample of medium-sized units, which are weighted to represent non-sampled units.

6.   The sample design for the upcoming *Annual Enterprise Survey: 2013* includes a postal sample of approximately 18,000 units, slightly higher than the previous year due to a boost to allow for better coverage of Māori businesses. The sample size has been decreasing over the previous few years (e.g. it was 22,500 in 2008). A total of 113 industries are covered in the data and, some industries are different enough to require a separate survey design. For example, some industries' income is not well correlated with GST sales or RME so they need a different strategy. Extreme outliers are generally forced into a full-coverage strata, along with other businesses that we wish to survey with certainty. The main design variables are value-added, total income, and total assets. The target sampling errors for AES are:
*   5% for value-added
*   5% or 10% for total income at the industry level
*   10% for total assets at the sector level.

7.   The individual data sources feeding into AES provide all the information needed for the specific businesses included in each source. The data sources and their approximate percentages of the population covered are:
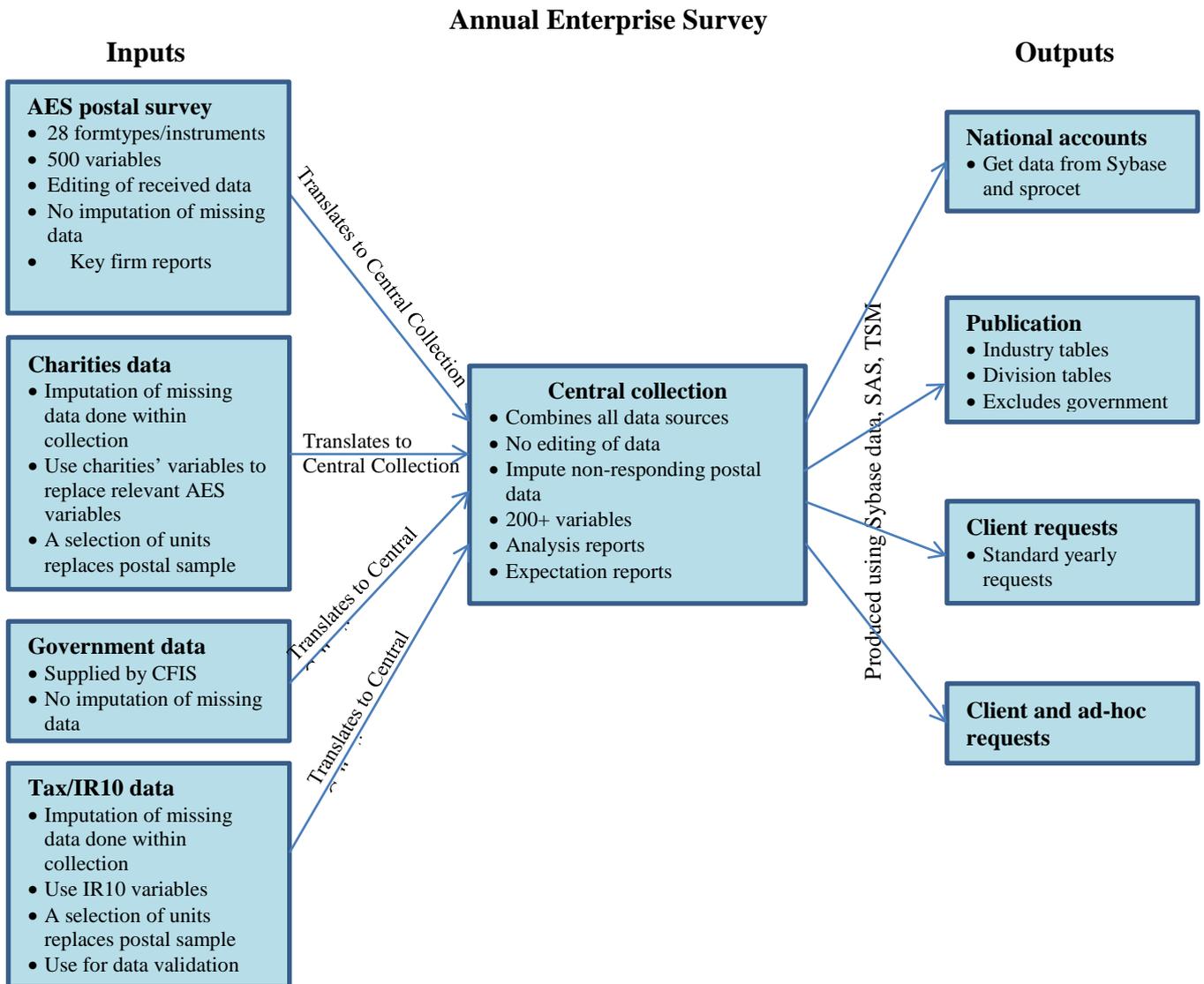
(a) IR10 units, 70%. The IR10 tax data was one of the first data collections to be fully migrated to the MEP. It is a financial statement summary collected by Inland Revenue for the purpose of determining the amount of income tax a business is required to pay. On receipt from Inland Revenue the data undergoes editing and non-response imputation within its own MEP collection, and no further processing occurs when it is used in AES. AES uses data from a selection of units in this collection. The initial introduction of IR10 data substituted tax data for small, insignificant business units in order to reduce the postal sample. Further replacement took place with the introduction of income thresholds and the replacement of whole strata for selected industries. The goal is to eventually use IR10 tax data as the main source with survey data only being used to supplement this where necessary.

(b) Postal survey, 4%. AES covers a total of 113 different industries. Many of these industries are very different from each other and require different questions to be asked in order to procure the necessary data. As a result of this a total of 28 different questionnaires or form types are used, which collect approximately 500 different variables. Some editing occurs as the data is received, but no non-response imputation takes place at this stage. As mentioned above, there has been a strategy over several years to reduce the size of the postal sample. Units whose data is available from The Charities Commission, will be removed from the postal sample and units that meet certain criteria, such as industry and income, will have their IR10 data used as a replacement, where possible.

(c) Government data, 1%. This is a full-coverage data source with central government financial information coming from The Treasury's Crown Financial Information System. It covers all New Zealand businesses that are owned by the government, including those in education and health, and other government departments. Some changes are made to the data after comparison with annual reports.

(d) Charities data, 0.5%. The Charities Commission supplies this data for not-for-profit businesses that meet the criteria of having charitable status in New Zealand. We have not yet migrated the data into its own collection on the MEP, but load it directly from Excel after editing. AES uses data from a selection of units in this collection to replace whole units that would otherwise have been in the postal sample. At present any non-response in the charities data undergoes imputation with the postal collection, but this will change on migration to the MEP.

The chart below illustrates the structure of the AES before migration to the MEP.

Annual Enterprise Survey

Inputs

Outputs

AES postal survey
- 28 formtypes/instruments
- 500 variables
- Editing of received data
- No imputation of missing data
- Key firm reports

Translates to Central Collection

Charities data
- Imputation of missing data done within collection
- Use charities' variables to replace relevant AES variables
- A selection of units replaces postal sample

Translates to Central Collection

Government data
- Supplied by CFIS
- No imputation of missing data

Translates to Central

Tax/IR10 data
- Imputation of missing data done within collection
- Use IR10 variables
- A selection of units replaces postal sample
- Use for data validation

Translates to Central

Central collection
- Combines all data sources
- No editing of data
- Impute non-responding postal data
- 200+ variables
- Analysis reports
- Expectation reports

Produced using Sybase data, SAS, TSM

National accounts
- Get data from Sybase and sprocet

Publication
- Industry tables
- Division tables
- Excludes government

Client requests
- Standard yearly requests

Client and ad-hoc requests

8.   Over the past few years we have carried out a strategy to increase the amount of administrative data used in order to decrease the size of the postal sample. If the data for a business is available from an administrative source then it will be used where possible. This has seen both the introduction of a new data source (charities data from The Charities Commission, which has been included since 2011) and the increase in the use of the IR10 financial statements summary data from Inland Revenue (IR10 tax data). The main assumption of the migration was that the survey methodology would be replicated where possible and that analysis of the various components, such as population and sample selection, and editing and imputation (E&I) would be based on a comparison with the existing, published survey results.

9.   The data from the various input data sources contains some variables that have slightly different definitions from each other so it is necessary to transform these into a standard set of variables. The transformations take place when the data from the input sources is 'translated' into the central collection. In many cases there is a direct relationship between two variables, others are summed to derive the central variables. The central collection is a complete dataset of all the new variables (approximately 200) for each AES unit. Non-response imputation is now carried out using the data from the central collection.

10.   The outputs of AES are:
(a)   Annual release. An annual release is published with industry and industry division tables. AES output variables include income, expenditure, profit, purchases of fixed assets, and equity. From this data, we can derive economic ratios such as the return on assets and profit margin on sales.

(b)   National accounts. The national accounts division is one of the main users of the AES data, which captures approximately 90% of New Zealand's gross domestic product. The data is pulled directly from sybase tables by their processing system and includes value-added, gross output, and gross fixed-capital formation.

(c)   Client requests. There are standard annual and ad hoc requests for the AES data.

III. Editing and imputation strategy for AES data on the MEP

11.   We developed the editing and imputation strategy for migrating AES data onto the MEP using the Methodological Standard for Editing and Imputation. This standard provides a comprehensive source of principles, requirements, and guidelines to ensure effective and efficient editing and imputation is applied to Statistics NZ data collections.

12.   Statistics NZ's approach to editing and imputation has four objectives (Methodological Standard for Editing and Imputation):
(a) To ensure collections in Statistics NZ have shared objectives for editing and imputation of fit-for-purpose, plausible data, and continuous improvement of processes and data quality.
(b) To ensure that all subject matter areas have a common understanding of the overall principles of quality, efficiency, and practicality.
(c) To guide our development of editing and imputation processes and methods, and inform future decisions about editing and imputation solutions and tools to be implemented across all Statistics NZ collections.
(d) To help establish a common, informed basis from which we can proceed in processing our data collections.

13.   The editing and imputation objectives are supported by eight 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 AES form.
(c) Make the best use of auxiliary and historical data, for example, using IR10 data to confirm AES 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 AES 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
  • unimputed and imputed data so that the degree, methods, and sources of imputation can be monitored

- producing, monitoring, and analysing editing and imputation (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.

(h) Editing and imputation are specialised processes that require skilled, trained staff to perform them effectively and efficiently.

IV. Data editing and imputation

A.   BANFF

14.   Banff is Statistics NZ's standard tool to meet the editing and imputation needs in our business surveys. Banff is an editing and imputation system developed by Statistics Canada and 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. Banff has been used successfully for a number of collections that have already been migrated to the MEP.

15.   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 where the quality is controlled predominantly via parameter settings.

B.   AES editing and imputation system

16.   The AES 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, to keep an E&I audit trail and provide E&I quality indicators.

17.   The AES E&I system has been developed as a coherent process that edits the raw AES data, imputes for non-response, and then provides analytical views to assess the quality of the edited and imputed data. This system is designed to automatically run end-to-end with the output of one procedure feeding smoothly into the following procedure. The end-to-end system has been divided into three separate stages due to the nature and complexity of the data:

- Firstly, the postal edits are run on each of the 28 different forms. This stage finds and fixes as many errors as possible to ensure the forms add up correctly.
- Secondly, the different data sources are translated into the central collection, so that all businesses have variables with consistent meanings.
- Thirdly, the non-response imputation is run to fill all the gaps in the central collection.

C.   AES postal editing

18.   The goal for building the postal edits in the MEP was to automate the process as much as possible so that minimal manual intervention is undertaken. In the old processing system the majority of edits were reviewed manually by analysts, who then left comments on the changes made. The postal edits are required to carry out the following functions:

- Time and GST adjustments – where units report for less than a year or in GST-inclusive amounts, the values are reapportioned

- Consistency checking – where a box has been ticked and a corresponding amount is required, checks are made that the results are consistent

6

- Deriving new values or variables – some edits are set up to derive new values (e.g. converting percentages into values, or to rescale some variables)

- Additivity edits – edits are built to check that values throughout the form add up to the total given by the respondent

- Checking for scanning errors (e.g. a 2 incorrectly read in as a 7 or a negative sign being missed)

- Looking for magnitude errors by comparing the unit's results to previous years (as a result of the edits flagged for the unit).

19. The AES formtypes are sets of financial accounts that logically group transactions into 'balance' groups. For example, the trading account consists of:

$$\text{sales - (opening stock - purchases + closing stock) = gross profit}$$

Therefore, the AES formtypes have internal relationships that enable logical errors to be identified and corrections to be made. Also, there are common response errors arising largely from misinterpretation of the return. These were analysed and rules for correcting common problems formulated.

20. Some edits required very simple fixes, for example, removing GST from a GST-inclusive value or deriving a new variable. In these cases we defined simple SAS-based rules to identify the variable whose value needs to be changed and carry out the change in a single step.

21. Other edits were much more complex and required detailed conditions to be met in order to identify a variable whose value is erroneous. An example of this is where a negative sign has not been read when the questionnaire has been scanned and a variable (such as Owner's Equity) is reported as being positive when it should be negative. In this case the criteria to be met are that the absolute value of the sum and/or difference of the equity components equals two times the absolute value of the equity total, and the equity total is greater than zero. The fix is simple; the equity total is multiplied by negative one. Although this edit is reasonably complex it can still be dealt with in a single step using SAS.

22. 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.

D. AES outlier detection

23. The postal data is processed by the Banff procedure Proc Outlier, following the correction of logical errors,. Very suspicious values, or outliers, are identified as potentially requiring exclusion from the calculation of imputed values for non-respondents. This is known as unlinking, which involves the removal of unusual and significantly differing units, as well as those that cannot be used in the calculation of imputed values for non-respondents. Four main types of imputation are used in AES; ratio imputation with GST, ratio imputation with RME historical imputation, and mean imputation. Similarly, Proc Outlier is used with three different algorithms to identify the outliers for the different imputation methods; the ratio algorithm with GST and RME as auxiliary variables, the historic algorithm with the unit's previous value as an auxiliary variable, and the current algorithm.

E. AES non-response imputation

24. As stated above, ratio, historical, and mean imputation are used in AES to impute for missing records by means of the Banff procedure Proc Estimator. These methods align with the imputation strategy of preserving the relationships between data items on the AES questionnaire. Imputation is carried out in imputation cells – groups of records of a similar size and industry. For each of the AES central variables a preferred order of imputation has been decided, based on the correlation with and availability of various auxiliary variables. If the preferred method fails there is usually a second, third, or even fourth method available, with mean imputation as the final option.

25.   Ratio imputation uses the Banff algorithm 'CURRATIO' with either GST or RME as the auxiliary value. Variables that use ratio imputation with GST as their first preference have a good correlation between their key value, total income, and GST sales. In this case the missing value is replaced by the GST sales value, but adjusted by multiplying with the ratio of the mean of the values of the responding units in the imputation cell to the mean of the GST sales values for the same units. Only units with non-zero GST sales values are eligible for ratio imputation with GST. In a similar way, ratio imputation with RME is used as a first preference for variables that are highly correlated with RME and can only be used if the RME value for that business is non-zero.

26.   Historical imputation uses the Banff algorithm 'DIFTREND' which replaces the missing value by the value declared at the previous collection period, but modified by multiplying by a forward movement factor. This is calculated from a group of similar units in the imputation cell by dividing the mean of the current values by the mean of the historic values. For variables that are not highly correlated with GST sales or RME, the next preferred imputation method is usually historical imputation, but only units with previous data are eligible for historical imputation.

27.   Mean imputation uses the Banff algorithm 'CURMEAN' and replaces the missing value by the mean of all values of the variable in the imputation cell. As the quality of mean-imputed data is usually inferior to other methods, this is the least preferred imputation method and is only used as a last resort.

V. Challenges and solutions for E&I

A.   Postal editing

28.   The AES is a very large, complex survey, with 28 different formtypes, 113 industries, and approximately 180 variables for 18,000 postal units to be edited and imputed. The sheer volume of data was a challenge due to the time the system took to run, although this did improve significantly along the way. To assist with building the editing and imputation processes on the MEP we used a 'thin slice' approach, where we initially restricted the dataset to one formtype and one industry to prove that all the different stages of the E&I were working. Once this was done we then expanded the dataset to include more industries, and eventually replicated the E&I for all 28 formtypes.

29.   The AES form contains several groups of variables that add or subtract to a subtotal or total. If the data provided by the respondent does not agree with the total given it is not always clear which variable is erroneous – there may be several variables that could be changed to make the equation correct. For this situation we decided to use Banff's Proc Errorloc (Error Localisation), which identifies the variable(s) to be changed in each erroneous record so that it can be made to pass the edit. Banff uses the strategy of minimising the number of variables to be changed in order for the record to pass the edit. This has the disadvantage that Errorloc will always choose to change one variable by a large amount, rather than two fields by smaller amounts. In the example given above;

$$\text{sales - (opening stock - purchases + closing stock) = gross profit}$$

we might prefer a small change to each of the components rather than a large change to one. In order to mitigate this risk, we can extend the rule of minimum change to include weights for the variables. If one of several variables could be changed to pass the edit, Errorloc will choose one randomly, but by adding weights we can have more control over the choice. We used our subject matter knowledge to assist us in the weighting. For example, some variables (especially totals) are more reliable than others – the respondent pays income tax based on gross profit so we assume that this will be correct. By weighting this variable considerably lower than the others, we ensure that it will never be changed. Similarly, we know that businesses do not often carry out complete stock assessments and so the stock-level variables may not be so reliable. We can give these variables higher weights. If more than one variable is a candidate for change, Errorloc will choose the one that has the highest weight.

30. The error localisation procedure identifies the best variable to change based on our weighting criteria, but does not make any changes. These are done by running Proc Deterministic immediately following Proc Errorloc. The variable identified in the first step is changed so that the record passes the edit.

31. The 'thin slice' work verified that postal editing in the MEP could increase the level of automation done for correcting failed units through the editing run. The automation has increased for both finding and fixing errors. The main way we have increased this automation has been by using the Banff Error localisation and Deterministic imputation procedures. The use of the error localisation procedure is the most significant methodological change and is the first successful implementation of the procedure at Statistics NZ. All edit steps have been set up to only apply to responding units.

B.   Outlier detection

32. There were significant differences in the strategies used to find outliers, between Banff and the old system used for processing AES. The main difference was that in the old system outliers were looked for in only three key variables, but when an outlier was found for a unit all the variables for that unit were unlinked. Banff deals with variables individually and there is no functionality to replicate the old system. To evaluate which variables should be run through Proc Outlier we performed two runs of the data through the entire E&I process and compared the results with those from the old system. The first run had only the three key variables set as outliers and the second had all the variables that were included in each imputation step. Some variables gave closer results between the old AES and Banff imputations when they did not have outliers detected and removed from the imputation calculations. The reason for this is probably that the nature of the variable does not lend itself to having outliers detected very well. For example, variables with many zeroes are known to be difficult to satisfactorily deal with any outlier process that compares values to their own distribution. Other variables, however, gave closer results when outliers were removed. Consequently, we did not look for outliers in every variable, just the core three plus those variables that benefit from having them removed.

C.   Non-response imputation

33. As mentioned above, the complexity of the AES survey data proved to be very challenging. The ideal situation would have been to set up the imputation parameters for each variable separately, but with over 200 variables in the central collection, and with several industries behaving differently from others, and up to 4 potential imputation methods, this would have been very long and time-consuming. Instead the variables were grouped by preferred imputation method as well as several other differences that proved to be significant:
  - Variables that are allowed to be negative (e.g. equity and gains and losses on fixed assets) were problematic when used with DIFTREND due to large fluctuations in the calculation of the forward moving factor. A variable that changes sign from one year to the next can result in a large negative forward-movement factor and strange imputation results. We changed the algorithm to 'HISTREG', which does a linear regression on the historic values of the variable.
  - Some industries' variables have different characteristics from those of other industries and are exceptions to the preferred normal imputation method. These needed to be treated separately. Some examples are fishing, employment services, and financial asset investors.
  - Statistics NZ has an implicit standard that at least 60% of units in an imputation cell should be valid responses (linked). Some variables gave better results if this linked criterion was reduced to 50%.

34. We divided the imputation process into four main sections based on the four methods mentioned in section 22. The first stage consisted of all the variables for which each of the methods was their first preference, followed by stage 2 for the second preference and so on. Each step in the stages divided the variables up into groups that had the same valued parameters and used the same imputation method. Each step (configuration) contains the list of variables, values of parameters, and any criteria to distinguish the units that it applies to. The first stage alone has 35 different configurations. Following stages have fewer steps as some variables only have two or three possible imputation methods.

35.   After analysing the results of running the data through the four stages of imputation, it was clear that there were some variables that had failed to impute. In some cases the reason was that the imputation cells were too small and failed to meet the criteria of a minimum count and linked percentage of valid respondents. Following the rules for merging imputation cells in AES we set up a cell allocation table at the beginning of the E&I run to combine two or occasionally three imputation cells into one merged cell. The four imputation stages then became eight as each had to be repeated with the merged cell as the 'by' variable of the imputation. If a variable failed to impute by imputation cell in stage 1, it would pass through stage 2 and try again by merged cell. If it failed again it would then try again with its second preferred imputation method and so on.

36.   Even with creating merged imputation cells a few variables failed to impute after eight stages. To fix this situation we added a ninth 'catch-all' stage at the end to carry out mean imputation by industry. Even though a very coarse solution, it was felt to be an improvement on the alternative of no imputation at all.

37.   There were many more challenges, not all of which had a solution. For example:
   • It took several hours to run the large dataset through the E&I, then be written back to the main data storage area of the MEP in order to view the data in a cube, so analysis of improvements to the E&I was quite slow.
   • While we were trying to produce results as close as possible to the published data this wasn't always possible to achieve so the difficult thing to decide was when to stop trying. What was the 'right' answer?

VI. Sampling errors

38.   A sampling error is the measure of the variability that occurs by chance because a sample is surveyed rather than the whole population. As well as sampling errors, which are quantifiable, there are also non-sampling errors which are not. Non-sampling errors can have a variety of causes in a survey, such as unintentional mistakes made by respondents when completing questionnaires, and processing errors. Another cause of non-sampling variance in the estimates of a survey is non-response imputation. Studies have shown that for surveys where variance due to imputation is not taken into account the sampling errors produced are generally understated.

39.   Most surveys in Statistics NZ either publish sampling errors with their releases or have them available on request. Each survey has a SAS program that is run once the estimates are finalised, and these programs are customised to the specific survey. One of the means by which the Stats 2020 strategic priority of transforming how statistics are delivered will be met is by standardisation. Creating a generic SAS code for estimating sampling errors inside the MEP was an important goal for our project.

40.   In order to build a means of estimating sampling errors within the MEP that was as generic as possible we had two options:

   (a) Adapt the existing sampling error code for use within the platform, making it generic enough to read in the number and names of required variables from the configuration for each survey.

   (b) Use a new tool that takes into account the variance due to imputation, adapt it for use on the platform and make it generic enough to read in the number and names of variables required from the configuration for each survey.

A.   SEVANI

41.   SEVANI (the System for Estimation of Variance due to Non-response and Imputation) is a statistical tool developed by Statistics Canada, which has already been evaluated at Statistics NZ. It is used in a survey context, when domain totals or means are to be estimated, and is particularly useful for estimating variances due to non-response and imputation of missing values.

42.   As SEVANI operates in a SAS environment through the use of macros, it was not too difficult to write a SAS script so that SEVANI can be run within the MEP. The more difficult part was to make it generic enough to run on multiple surveys. The main challenge was that SEVANI will only produce output for one variable at a time, but we wanted to automate the process and name all the required variables in the configuration. This was achieved by creating a macro for the variable names, running the code for each one and renaming the output datasets so they would not be overwritten.

43.   In order to estimate variance due to non-response, SEVANI requires information about the method of imputation used and the unlinking information for each imputation method. The MEP produces all this information as the data is processed; the method of imputation is a status attached to the imputed value and the unlinking information is the status from the Banff Outlier procedure. Code was added to the SAS script to bring in the extra datasets needed to merge the additional variables with the original dataset. SEVANI also needs to set up a lot of parameters relating to the imputation methods and names of variables, such as weight, selection cells, imputation cells, and auxiliary variables. This proved relatively easy to do through a configuration in the MEP, with SAS code added to the script to read these variables in to the SEVANI macro call.

44.   Although the estimation of sampling errors using SEVANI can be processed within the MEP, the output cannot. This is due to the nature of the MEP being unit based, whereas SEVANI produces output at aggregate levels. The easiest way to get the output is to copy the SAS code out of the MEP and run it again in SAS.

VII. Next steps

45.   The most critical next step is to educate the users of the new AES system on the MEP. One of the underlying aims for the project was to minimise the amount of manual intervention by automating processes as much as possible. There will be a period of adjustment to this new way of thinking for members of the AES subject matter area, as well as a rearrangement of the structure of the team and the way they work. This will take some getting used to. However, the advantages for them are a faster, more streamlined, and transparent system.

46.   As the AES team get used to working in the new environment there will be the potential to make improvements to the editing and imputation system. The configurations are transparent in that all the details, parameters, and equations relating to the E&I are set out plainly, with nothing hard-coded in SAS script behind the scenes. This allows the AES team to make changes to a postal edit (for example) if they find that the current setup is not working as they would expect. These changes can be made without any IT involvement.

47.   As mentioned earlier in the paper, the Charities Commission data has not yet been migrated onto the MEP. A future task, therefore, will be to create a new the MEP collection for Charities data with its own editing and imputation configurations.

VIII. References

Seyb, A, et. al. (2012). *On tap: Developments in statistical data editing at Statistics New Zealand*. Paper presented at the 2012 UNECE Work Session on Statistical Data, 24–26 September 2014, Oslo, Norway.

Statistics New Zealand (2010). *Methodological standard for editing and imputation*. Internal report, Wellington: Felipa Zabala.