WP.9
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Vienna, Austria, 21–23 April 2008)

Topic (ii): Editing administrative data and combined sources

E&I OF ADMINISTRATIVE DATA USED FOR PRODUCING BUSINESS STATISTICS

Invited Paper

Prepared by Vera Costa, Frances Krsinich and Rudi Van der Mescht,
Statistics New Zealand, New Zealand

I. INTRODUCTION

1. Statistics New Zealand has been using administrative data for supplementing the production of business statistics since 2001. Most of the data is received from other official agencies (Inland Revenue Department and NZ Customs Service) but electronic transaction data from private providers has recently been incorporated into the set of administrative data being used. This paper presents the challenges related with editing and imputation (E&I) methodologies for dealing with private versus government administrative data provided at unit record versus aggregated level.

II. ELECTRONIC CARD TRANSACTION DATA

II.1 BACKGROUND

2. An electronic card transaction occurs when a payment or refund is made between a merchant and customer with the switch company acting as the facilitator for the transaction. Electronic transactions arise from debit, credit and charge card usage. Each of these transactions passes through the switch, working with their respective banks to transfer the money from one account to another. In New Zealand a small number of switches cover all electronic card transactions and we collect data from all of them.

3. The switches keep track of all these transactions and provide aggregated information to Statistics New Zealand within five working days of the end of the reference month. The aggregated information consists of a grand total and totals for each industry subgroup defined by the individual switch.

4. Statistics NZ has been receiving monthly electronic card transaction (ECT) data aggregated by industry since late 2004, with the data starting from January 2000. In July 2005, we began a project to investigate the feasibility of using ECT data as an early estimate of retail trade activity. The main finding of this analysis is that even though ECT actuals almost always move in the same direction as the Statistics NZ Retail Trade Survey (RTS), ECT data does not indicate movements in the RTS to the level of accuracy required by users.

5. Following this work the ECT project team consulted a wide range of technical users as well as representatives of the media to discuss the results of the analysis and ascertain potential uses for the data. Notwithstanding the comparison with the RTS, the consultation showed there is a general desire among

users for ECT data to be published as a statistic in its own right. Technical users see the data as a very timely and useful addition to the range of short term indicators which provide information on consumer spending and economic activity in general. The data are available at least three weeks before other conventionally surveyed data. The media also consider there is general public interest in statistics on debit and credit card usage.

6. Given the user interest in the data the project team recommended that ECT data should be published as stand alone series, based on the current aggregate data received by Statistics NZ. Seasonal adjustment is employed to publish seasonally adjusted and trend series for each of the following series:

- Total Electronic Card Transactions
- Retail Electronic Card Transactions – a subset of total electronic card transactions corresponding to the coverage of the Statistics NZ Retail Trade Survey, which includes the Retail Trade; Accommodation, Cafes & Restaurants; and Personal Services industries as defined by ANZSIC
- Core Retail Electronic Card Transactions – a subset of retail electronic card transactions, excluding the motor vehicle related industries.

7. At the same time the project team has the opportunity to pursue the provision of much more detailed unit record data from the switch houses.

II.2 EDITING AND IMPUTATION

8. ECT aggregated data is sent to Statistics NZ monthly. The fact it is aggregated data has a large impact on the methodologies that can be used for its editing and imputation. The lack of unit record data limits considerably the techniques that can be used for checking its quality. Therefore it is important that mechanisms are put in place to ensure a good relationship with the providers so that Statistics NZ is always informed of any issues that can have an impact on the data.

9. Of special concern is the possibility that an unforeseen operational issue may prevent one of the switches providing complete data for a month. This could happen if a switch was unable to provide results for a certain period of time within the month, despite the fact that electronic card transactions had been made during the period. We consider that it is unlikely that a switch would be unable to provide a total summary, as it is vital for the switch to have a record of all transactions and not having this would be extremely serious for them. On the other hand, the provision of the summary by industry is subject to more risk as these figures are not critical for the switches.

10. In any of these cases, erroneous or suspicious data can be identified via the providers telling us there is a problem, or us picking up a problem through our edit checks. Also, it is possible that estimates produced through a time series forecasting method are used for trying to identify potential problems with the data received from the providers (ie treated as expectations). Other possible approaches for identifying problems are comparing the current data with previous data as well as the knowledge about the real world (eg from newspaper clippings).

11. If there is any suspicion about the data quality then the switch will be contacted because, due to the characteristics of the data, this should always be used as the first option to clarify data issues.

12. If the providers are not able to provide complete data for a month, imputation of data for the missing period is needed to allow the production of electronic card transaction estimates for the whole month. Such monthly estimates are important because they can be used as early indicators of the behaviour of the economy. Also, a complete and stable time series is important for users and this can be produced if imputation is used.

13. Performing imputation is not a simple task as it is expected that a number of factors could have an impact on the results. The day (weekend or working day) when the operational issue occurs, the time

of day or night and so on are important effects quite difficult to quantify without lower level data (eg daily and/or merchant data).

14. The use of time series methods for forecasting the total estimate for the month with missing days, as well as the Retail and/or Core Retail estimates, avoids these problems and could be used for coping with the scenario of missing data. It is important to reinforce that such an approach should only be used after all possibilities of obtaining the real data through the providers have been proved unsuccessful, as in this case keeping a good relationship with the providers is considered the preferred option to solve data issues.

15. We have identified two situations that could cause problems for an imputation methodology based on time series methods. The first situation is when a forecasted value is lower than the partial actual result provided by the switch. In this case the value to be imputed would be lower than the original. The second situation is when a big merchant changes switches. In this case there would be a discontinuity in the time series for both providers, which would impact on the quality of time series forecasts. It would be necessary to have mitigation methods to deal with such cases.

16. Other methodology issues to be considered are the quality of the forecast, the potential for bias introduction and the convenience of having thresholds guiding the imputation related actions: for example, if the forecasted value is within +/- x% of the original value, don't adjust; and/or if the percent difference between the estimated and actual value provided in the month is above a certain threshold, don't publish the result. The option of not releasing data for a month needs to be carefully analysed due to the impact that the lack of a month has on the future adjustment of the time series and future imputation / forecasts as well as on the users' confidence in the series.

17. At least one quality indicator related to the amount of the estimate that is imputed is needed in order to inform the data users about the quality of the published data. Also, metadata must be made available regarding the methodology used for performing any editing and imputation. Besides this, a system should be in place to deal with the proposed editing and imputation methodologies. Such system should be automated where possible, and should allow the keeping of an audit trail of all the changes the data went through.

18. Although it would be feasible to develop an imputation system for dealing with problems in the data received from each provider, it was considered that the work involved was too great to justify the benefit, particularly given the low likelihood of such an issue occurring. In the case of a failure from the switches, the mitigation strategy is that Statistics NZ either delays or cancels publication for the month or publishes the data with a caveat about the expected impact. Since the ECT data is being published as an experimental series, this approach is considered appropriate.

19. In fact, technical issues with the source data have already occurred. The release of the ECT series for October and November 2007 was delayed because it was found that an invalid transaction type had been included in the data. Once we received updated data from the supplier the data was revised back to November 2002. The revised series movements were similar to the previously published.

20. Given that a forecast of the upcoming result is not available by switch, expectations and data quality assessment are being done via media monitoring, liaison with the switches and analysis of the data received. There are procedures in place for the checking of the data on receipt in terms of the file formats and the content of the files. Also, the data validation includes the analysis of the received results versus previously observed movements and the observation of zero and partial weights applied in the seasonally adjusted series.

III. LONGITUDINAL BUSINESS DATABASE

III.1. BACKGROUND

21. There is a growing demand for different aspects of official statistics. Policy makers and several interest groups need official statistics that allow them to look at issues in different and dynamic ways (e.g. longitudinal, linking information across time / sectors). The Longitudinal Business Database aims at generating information on microeconomic performance of New Zealand businesses that previously was not possible to produce.

22. The Longitudinal Business Database integrates, using a deterministic link, tax and business survey data with the Statistics NZ's Longitudinal Business Frame. This frame is a by-product of our business sampling frame and contains longitudinal information (e.g. industry, ownership type and sector) on a wide population of firms.

23. The full-coverage tax data incorporated into the Longitudinal Business Database includes

- sales and purchases from Goods and Services Tax (GST) returns, annualised from a monthly, bi-monthly or six-monthly basis
- financial performance and position variables from "IR10" tax forms
- salaries and wages, and employee counts from monthly Pay-As-You-earn (PAYE) returns.

24. A team started work on the project in January 2006 and the first nine months involved understanding the various data sources and reaching agreement on data integration structure and rules. Then the data was checked to ensure the integration process hadn't impacted on the data itself.

25. To enable consistency of aggregate outputs missing values are imputed but, as any one imputation methodology can't fulfil all requirements for analysis, all imputed records are clearly flagged so that users of the microdata can choose to remove the imputed values and adjust for missing data as appropriate to their specific needs.

26. The strategy being used for the imputation methodology development encompasses an iterative approach consisting initially of a broad brush implementation, followed by gradual improvement in successive iterations. The idea behind such approach is that the production of a usable database early on enables input from researchers and such feedback can be incorporated in future developments.

27. On the other hand, there is a risk of having no incentive to do further improvements, especially if researchers seem to be successfully using the database already built. Also, users of the data must be aware that there may be future revisions of the data and this can cause changes to inferences made using such provisional data.

28. By April 2007, the data was considered suitable for use by researchers and four papers on firm dynamics, market structure and performance were presented at the New Zealand Association of Economists' Conference in June. Nevertheless, these pieces of research done through Statistics NZ Datalab have not made direct use of the imputed data and the work for improving the imputation methodology is still progressing. In some cases the imputed data has been used for checking the robustness of the estimates.

29. Currently the database contains data from 2000 to 2005 and the many related variables across different data sources give us much auxiliary information that can be incorporated into the methods used to impute missing data. There are around 800,000 enterprises in the database.

30. Beyond the desire to meet demand for new official statistics, Statistics NZ hopes that the permanent construction of a Longitudinal Business Database will:

- reduce the response burden on firms by shifting from surveyed to administrative sources for quantitative data
- contribute to an improved statistical architecture for sampling our surveys
- enable a wide range of future longitudinal research to be conducted on a stable, high quality micro dataset.

III.2. EDITING AND IMPUTATION

31. Due to non-response, and the editing of poor-quality responses, approximately 6 percent of the sales and purchases data (based on the goods and services tax, hence referred to as GST data), and 30 percent of the financial performance and position data (based on data from the IR10 tax forms, hence referred to as IR10 data) is missing.

32. Most of the work done for the creation of the Longitudinal Business Database targeted the development of an imputation methodology with almost no attention being paid to its prior editing. Therefore, the tax financial data (IR10 data) only went through the tax automatic edits already in place for data used for the production of the Annual Enterprise Survey (AES).

33. Most of the edits used for the tax data being used in AES involves additivity checks which are automatically corrected if the responses are out within a given threshold. When an edit is failed by more than the given threshold, and the unit has goods and services tax sales of more than $1m, then the unit is flagged as error and is edited by an AES team member.

34. There is a large number of tax units included in AES and individually these units have a small contribution to the overall survey results. Therefore resources are concentrated on the more significant units and all other respondents who fail one or more edits by more than the given threshold have their data removed and a new response imputed.

35. No manual resolution of edit failures has been performed specifically for the Longitudinal Business Database project.

36. A range of imputation methods was considered in the first year of the project and, of those, the imputation methods tested were mean imputation, multiple imputation and donor imputation.

37. Mean imputation was tested because it is a very common and simple method, used often for business surveys. A fair amount of work was put into the testing of multiple imputation but software limitations meant that testing had to be put on hold. Version 9 SAS is required, which was not available on the Statistics NZ SAS server at the time, and the dataset was too large for multiple imputation to run on a stand-alone PC. Donor imputation was seen to be a promising method, on the basis that it will preserve distributions in the data, which is a key consideration in enabling unit record analysis, one of the key objectives of the Longitudinal Business Database.

38. An evaluation was undertaken to compare the results of applying both mean and donor imputation. This was on a dataset with simulated 'missingness', to enable differences between real and imputed values to be calculated and compared for the two methods. In general, donor imputation performed well compared to mean imputation, but some results were affected by a small percentage of outlying values.

39. After consideration of the properties of both mean and donor imputation, and based on the evaluation, the decision was made to implement donor imputation on the Year 1 database, to enable the development of new official statistics, and to provide a complete dataset for use by researchers in our Datalab. Flags were included to indicate imputed fields so that researchers have the option of excluding the imputed values from their analysis where appropriate.

40. The donor imputation used to impute missing values for the Year 1 dataset is achieved using *proc donorimputation*, a SAS procedure available in Statistics Canada's *Banff* system. The donor imputation procedure uses a nearest neighbour approach to find, for each record requiring imputation (recipient), the valid record (donor) that is most similar to it and that will allow the imputed recipient record to pass user-specified edits. The only edit specified was a non-negativity constraint. Donors must be complete records – that is, they can't have any missing fields.

41. Initially, imputation was attempted for all the variables, across all years, at once. There are 58 IR10 tax variables and four GST variables across 6 years (2000 to 2005) so this meant imputing approximately 400 variables at once. We weren't able to get an imputation run to complete with this number of variables. Imputation didn't finish running in a feasible time frame.

42. Eventually it was decided to cut down the number of variables instead, to just those required for the upcoming official statistics development work (ten IR10 tax variables and three GST variables). This meant there was a reduction in the number of variables imputed for, from 62 to 13. Across all years this means imputing for approximately 80 variables instead of 400 variables.

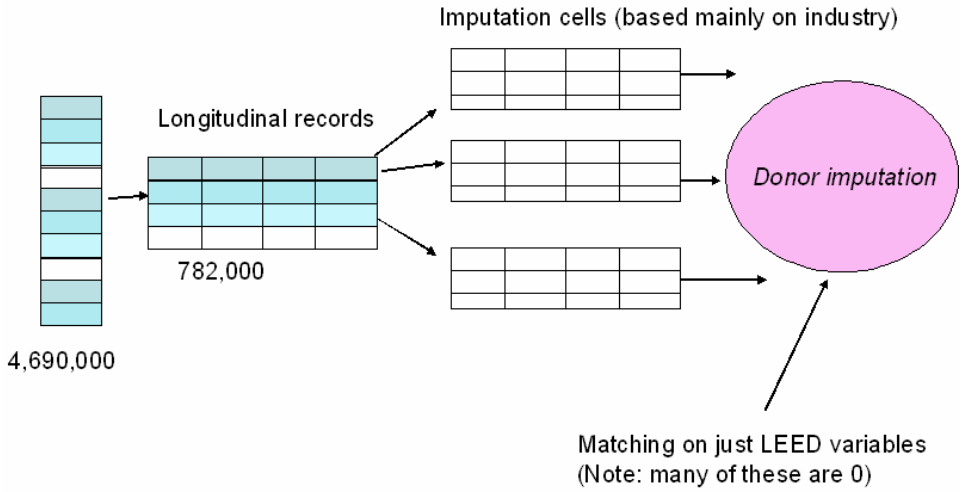
43. The donor imputation across the longitudinal record used the Linked Employer-Employee Database (LEED) variables (employee count and salaries_and_wages) to determine the nearest neighbour. One donor was used for each recipient across the full longitudinal record.

44. The LEED variables are available for all enterprises, but a high proportion (around 60 percent) is zero. For the enterprises with zero employee count and salaries_and_wages, this meant that the imputation was effectively a hot-deck one. That is, recipients were randomly assigned a donor from within their imputation cell.

45. Imputation cells were defined by industry, and where this was not sufficient to get a small enough cell, we subsetted further by whether or not the enterprise had zero-valued LEED data, and then, for those that were still too large, by an average rank of the imputable variables. The requirement that cells were not too large was due to the processing time, which was initially very long given our misunderstanding of how best to specify the donor procedure.

46. Figure 1 shows the overall approach to imputation used in the first year of the project.

Figure 1
Year 1 Approach to Imputation



47. By the end of the Year 1 imputation work we undertook a literature review in order to identify any imputation methods that we hadn't already considered. The general conclusion was that the approach being used was sensible, and that we should also take advantage of existing data to impute from where possible, for example, by interpolation and historical imputation. We took this into account when determining the enhancements for the second version of the imputation.

48. A known limitation of the year 1 imputation was that donors must have been live in the population for the whole reference period (2000 to 2005). This means that the imputed values were biased towards the characteristics of 'long stayers'. Also, the pool of potential donors was smaller than it would otherwise be.

49. To overcome this problem, an important modification to the way donor imputation was applied has been made in the project's second year: the data should be subsetted by the (approximate) 'lifespan' of the enterprises before imputing, to allow recipient enterprises who were not 'alive' over the whole reference period to seek donors with a similar life span.

50. Also, instead of using only the LEED variables for performing the matching, the year 2 imputation also uses some IR10 and GST variables to do so. Nevertheless, by incorporating more matching variables there is potential for quality issues of the matching variable to impact on the 'noisiness' of the match. So, further research is needed to determine the optimal set of variables to include in the matching.

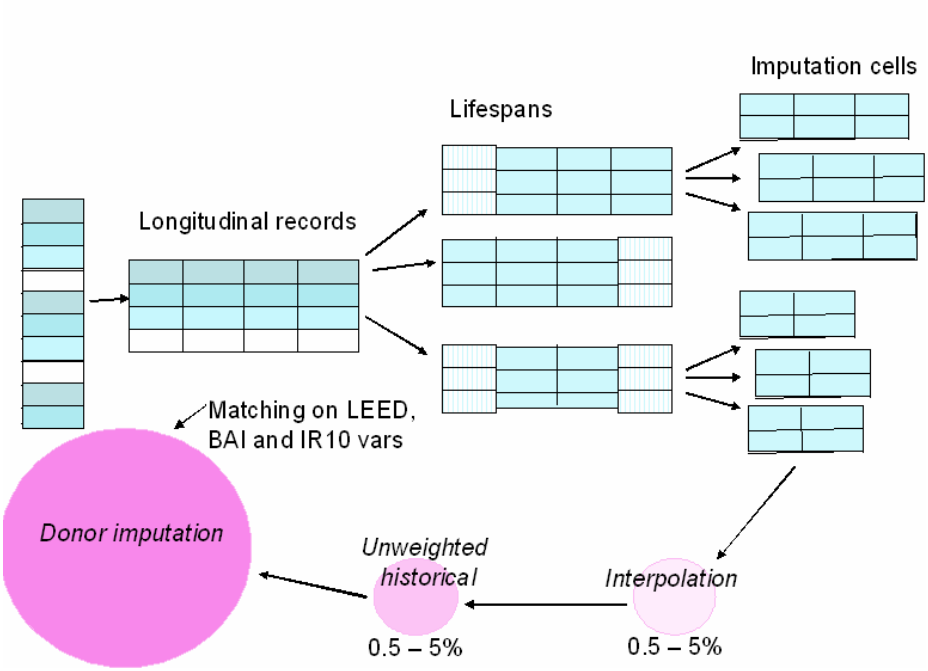
51. Other changes in the imputation methodology were the use of interpolation and the carry forward of a previous value (called unweighted historical). Both approaches use more information from each business for which imputation is needed and are therefore considered more appropriate than using donor imputation.

52. For missing values where a previous and a subsequent value exist, we interpolate to impute a value. That is, the imputed value is the average of the values existing in the previous and subsequent years. So, this interpolation method makes the most use of the information available for the unit.

53. Where it is not possible to use interpolation but a previous value exists, we carry forward the previous value. At this stage, we are not incorporating the trend in each variable within imputation cells (called weighted historical imputation). The reason for this is that there was no time for developing any unlinking procedures for the monitoring so unweighted historical imputation was felt to be more robust.

54. Figure 2 shows the approach to imputation currently being used.

Figure 2
Year 2 Approach to Imputation



55. Approximately 30 percent of the enterprises have missing data for any given 'block' of variables (IR10 and/or GST). It is possible to use interpolation as the imputation method for around 2 percent of the enterprises, 5 percent can be historically imputed and the about 20 percent remaining must be dealt with through donor imputation.

56. Although progress has been made so far, it is recognised that the imputation methodology still requires more development. The iterative approach we have taken to developing the imputation methodology means that the imputation is adequate, and the 'year 2' approach is better than the 'year 1' approach, but it is still not at the stage where it would be appropriate to put it straight into production.

IV. FUTURE WORK

57. Due to the nature of the Electronic Card Transaction data and its implications for editing and imputation, Statistics NZ has not developed an imputation system for dealing with problems in the data received from the providers. Suggestions from other statistical agencies about how E&I related issues are dealt with when aggregated data is received from the providers of administrative data would be greatly appreciated.

58. Regarding the Longitudinal Business Database, we think that some thought should be given to other methods. An obvious one is the Little and Su method (Little & Su, 1989) that is being used for the Household, Income and Labour Dynamics in Australia (HILDA). The decision to use donor imputation was made under tight time constraints with a choice between donor or mean imputation. Our literature search and further thinking has confirmed that this is an adequate method, but more work is required to confirm that this is the most appropriate method for the Longitudinal Business Database.

59. If the current donor-imputation approach is retained, there is further work required such as:

- Defining imputation cells - this could be honed further by basing on the number of donors rather than the entire number in the cell (ie including those requiring imputation).
- Determining whether the matching variables being used are the optimum ones.

- Determining whether categorical variables can be more directly incorporated into the match than through the definition of imputation cells.
- Making more use of past and future values in the longitudinal record. That is, generalising the historical and interpolation approach used in the year 2 imputation, using backcasting, and from further away than just adjoining periods.
- Consider using donors to impute movements rather than levels, particularly for those that are outlying with respect to size (employment) variables.
- Consider treating large non-respondents separately. Our testing of imputation results showed that the very outlyingly large enterprises are often non-respondents (with respect to IR10 data, at least). As these are likely to be 'key firms' for the AES survey, and therefore likely to have AES data, then imputing IR10 data from AES data for these is a possibility.

Consider a moving longitudinal window, as too long a longitudinal record may become too 'noisy' We would have to research what would be the optimum window length and how we link together across the edge of the window. For undertaking these researches we probably need a longer time series than we currently have now. Also, we need some thought/education about the fact that the data will change each year.

60. The recommendation at this stage is that imputation for all Longitudinal Business Database periods be rerun each year, rather than just the most recent year. This means that past years' data will change, and so there needs to be management of users' expectations about this. But this issue should be investigated as part of the development of a production system, with pros and cons weighed up explicitly.

V. REFERENCES

Little, R.J.A. and Su, H.L. (1989): Item Non-Response in Panel Surveys. In: Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M.P. (Eds.): *Panel Surveys* (pp. 400–425). New York: John Wiley.