

## **STATISTICS NEW ZEALAND: RECENT DEVELOPMENTS IN ELECTRONIC DATA COLLECTION**

Nancy McBeth, Stuart Pitts and Steven Johnston<sup>1</sup>

### **ABSTRACT**

Over the past few years Statistics New Zealand (SNZ) has experienced an increase in the volume of business surveys data supplied by email. However up until now SNZ has not had the business processes in place to support electronic collection in a way that meets the needs of both SNZ and data suppliers. To this end, SNZ has invested effort over the last year investigating how best to approach the problems and opportunities presented by electronic data collection. This paper outlines SNZ's plans to move from email supply to a secure data lodgement facility and in the longer term the development of an internet-based data collection system. It also presents a case study of the data currently supplied by email for the Monthly Retail Trade Survey. The case study illustrates some of the benefits of electronic data, but also some of the data quality problems and costs to the organisation, and highlights the need to consider the data collection methodology within the wider context of the total survey cycle.

KEY WORDS:      Email; Retail trade; Data quality.

### **1. INTRODUCTION**

Like other national statistical offices, Statistics New Zealand (SNZ) has been rapidly coming to terms with the implications of increasing volumes of data that is supplied electronically. Up until a few years ago SNZ was not too concerned about the occasional and ad hoc nature of electronic data provision. However, the increasing use of email meant that SNZ found itself in a position where the amount of electronic data became significant, but the business processes were not in place to support the use of the electronic collection medium in a way that best meets the needs of both SNZ and data suppliers. To this end, SNZ has invested a significant proportion of time over the last year investigating how best to approach the problems and opportunities presented by electronic data collection.

### **2. ENVIRONMENT FOR DATA COLLECTION FROM BUSINESSES**

New Zealand is a small economy by world standards, with the total population at the March 2001 Census just under 3.8 million people. Total employment at the same time was approximately 1.8 million, across nearly 300,000 enterprises. Major economic changes in the 1980s and 1990s have changed the nature of the New Zealand economy, from a traditional extensive reliance on agriculture and agricultural products, to an increasingly diversified economy. A key element of these economic changes, was the opening up of the New Zealand economy, not only in terms of markets, but also in terms of ownership. In February 2000, approximately 2.3% of New Zealand enterprises had at least 25% overseas equity. As with many other economies, overseas ownership is not evenly distributed across the economy, as highlighted by the Finance and Insurance industries, where over 12% of enterprises had at least 25 % overseas equity.

---

<sup>1</sup> Nancy McBeth and Stuart Pitts, Survey Design and Development Division, Steven Johnston, Survey Methods Division, Statistics New Zealand, P.O. Box 2922, Wellington, New Zealand

One of the consequences of this feature of globalisation has been that increasingly head offices of large New Zealand corporates are located offshore. As responding to SNZ surveys is more likely to be a head office function, one impact of this for SNZ has been the need to develop a more systematic approach to surveying businesses located offshore.

An important issue for producing reliable and timely economic indicators is the need to balance data demands with respondent load. As with other agencies, SNZ has not been immune from political pressure to reduce respondent load. Increasingly this is taking the form of stressing the potential of reducing respondent load from the use of technology. In a recent review of government compliance costs on business it was recommended that SNZ "...should accelerate work on ensuring secure transfer of data. Improved information flows and consultation using web-based technology should be sought without delay."

The developments outlined in the remainder of this paper should therefore be seen in the context of both increasing pressure to reduce respondent load, as well as the need to operate collection activities in the global economy.

### **3. SNZ DEVELOPMENTS**

#### **3.1 Where we have come from - email**

SNZ operates virtually all of its business surveys on paper, but this has not stopped respondents returning data in a variety of electronic formats - and in some cases this has been actively encouraged. Facsimile has probably been the most common form of electronic return and email is increasingly being used as a standard business tool in much the same way. Since 1997 SNZ has included a generic email address (surveys@stats.govt.nz) on many business survey questionnaires and not surprisingly some respondents started to supply data by sending emails to this address. The most significant survey in terms of the number of email returns has been the Monthly Retail Trade Survey (RTS), described in the case study below.

#### **3.2 Electronic data collection policy**

Once SNZ began to consider the implementation of an internet based data collection system, it followed that a policy should be developed to enable the consistent application of the internet as a data collection medium. The policy can be very briefly summarised as follows:

Statistics New Zealand will only accept electronic data via the internet that has been both encrypted and authenticated to acceptable industry standards.

The main aim of the policy is to facilitate the adoption of internet based data supply in a way that both meets SNZ's statutory obligations in terms of the Statistics Act and maintains the underpinnings of integrity and trust that allow SNZ to operate as an effective national statistical office.

#### **3.3 Plans for a web based collection facility**

As a first step in complying with the policy SNZ has undertaken to develop a secure electronic lodgement facility, tentatively called the Secure Deposit Box. The Secure Deposit Box approaches the concept of internet data collection at a very basic level and will provide the basis upon which future developments will be made. Respondents will be given the option of providing data in any format (plain text, file attachments) rather than being presented with electronic questionnaires. This complicates full integration with survey processing systems but provides flexibility for respondents. As the Secure Deposit Box will not be able to be used without prior authorisation, a key element of the facility will be a requirement for all users to formalise an electronic collection arrangement with SNZ. SNZ will assume responsibility for other

management aspects of the system, such as password maintenance. The net result of these processes is that SNZ will have tighter control over system usage, which will hopefully facilitate further development and enhancements.

A fully developed web based collection system is likely to be based on improvements/enhancements to the Secure Deposit Box. The first type of enhancement will probably be the integration of electronic survey templates that 'control' the format of returned data. Once the format of data is consistent it is then a fairly straight forward matter of linking with survey processing systems, with electronic data being treated similarly to data derived from imaging, scanning or manual data entry. One of the main concerns SNZ will have to address when moving to this more complex collection model is the potential modal effect associated with different collection methods. As outlined below some preliminary mode effect work has been done in relation to email survey returns that SNZ currently receives. With the introduction of more sophisticated and standardised collection systems, any mode effect should be easier to identify and deal with.

## **4. CASE STUDY: EMAIL RETURNS FOR THE RETAIL TRADE SURVEY**

### **4.1 Background**

The RTS is a monthly collection that collects a single sales figure for each location (called the geographic unit, or GEO) surveyed. Overall monthly retail sales are currently around NZ\$3.6 billion (this goes up to about NZ\$4.5 billion in December). Along with the sales figure, information is also provided on whether the figure includes Goods and Services Tax (GST), as well as the actual period that the return has been completed for. Every third month a closing stock figure is also requested.

The survey is selected at the enterprise level, with each enterprise comprising one or more GEO. All of the retailing GEOs of a selected enterprise are included in the survey. The current sample for the survey consists of 4000 enterprises, resulting in approximately 7900 geographic units. The distribution of geographic units is not homogeneous across enterprises in the sample. Approximately 80% of surveyed enterprises have a single GEO, while one of the larger multi-GEO enterprises has over 200 GEOs. SNZ consistently achieves a response rate for the RTS of between 88 and 90% of GEOs in the sample.

As discussed previously, the surveys@stats.govt.nz email address has been increasingly used as a response medium for the RTS. Email returns were first actively solicited for this survey between April and May 1999 and at about this time a standard response format was developed that most respondents seem to have continued to use. SNZ again solicited email returns in late 1999 as part of an attempt to improve timeliness. The standard collection process is to mail the paper questionnaire, meaning that except in a small number of cases, email is not used as the initial prompt to respond. The same respondents tend to use email to respond each cycle, occasionally a new email respondent will send in a reply, but most tend to be regular email suppliers.

One function of accepting email is that respondents can choose to supply data as either attachments or as plain text. Figure 1 shows the trend in RTS email returns over the months from June 1999 to July 2001.

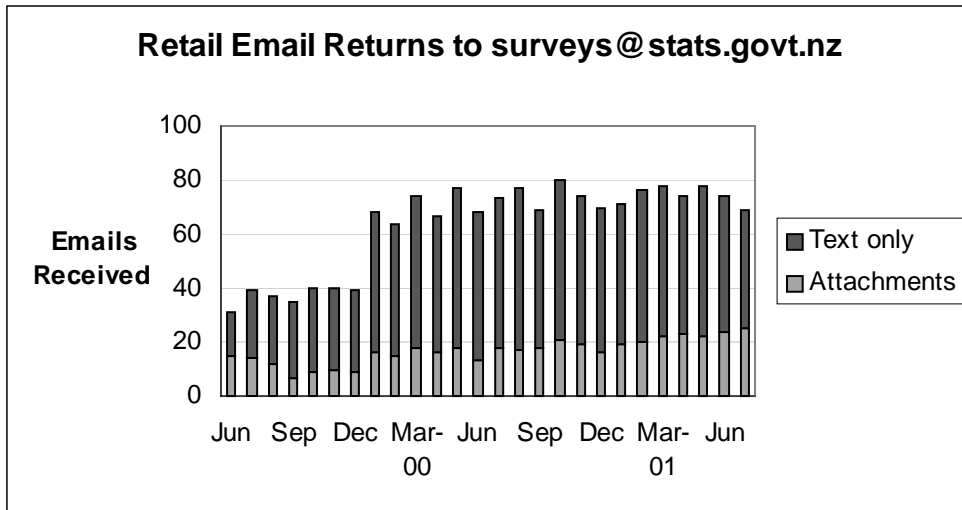


Figure 1

The following observations are based on analysis of four sets of RTS email returns, taking every second month from December 2000. The focus of analysis in this case study is the enterprise, rather than GEO. This is because it is the enterprise rather than the GEO that responds to the survey.

On average, responses from 70 enterprises (equating to approximately 2% of the enterprises in the sample) were received by email. While only a relatively small number of emails were received each month, it should be noted that the enterprises comprised of large numbers of GEOs seem to be more likely to respond by email. Figure 2 shows that just over 60% of the surveyed enterprises comprising more than 50 GEOs responded by email. By comparison, of those enterprises with only one GEO, only 1% of units responded by email.

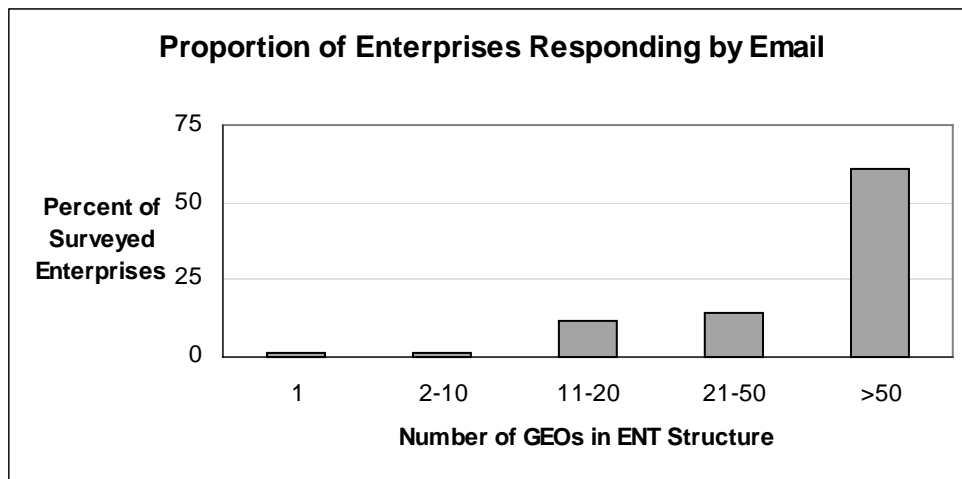


Figure 2

Not only are multi-GEO enterprises more likely to be email respondents, when they do respond by email they are more likely to do so with attachments. Over 80% of multi-GEO enterprises responding by email did so using attached files (mostly Excel spreadsheets), compared with only 7% of single-GEO email respondents. This is most likely related to the way in which large organisations extract survey data from their financial systems. It is easier for a large organisation to set up an automated process to extract

multiple responses all at once and output them to a spreadsheet or other file. The multi-GEO enterprises that returned data as plain text were still quite small, with the largest having only 5 GEOs.

While the actual number of email returns does not seem significant, the dollar value represented by those returns accounts for around 9% of the total sales figure published from the survey. Looking at the breakdown of total sales by type of retailer, there are wide variations in the contribution of data supplied by email. As the analysis in Figure 3 shows, the contribution of data supplied by email ranges from Department Stores, where over 60% of data is supplied by email; to Liquor, Footwear and Furniture and Floor Covering storetypes, where no data is supplied by email.

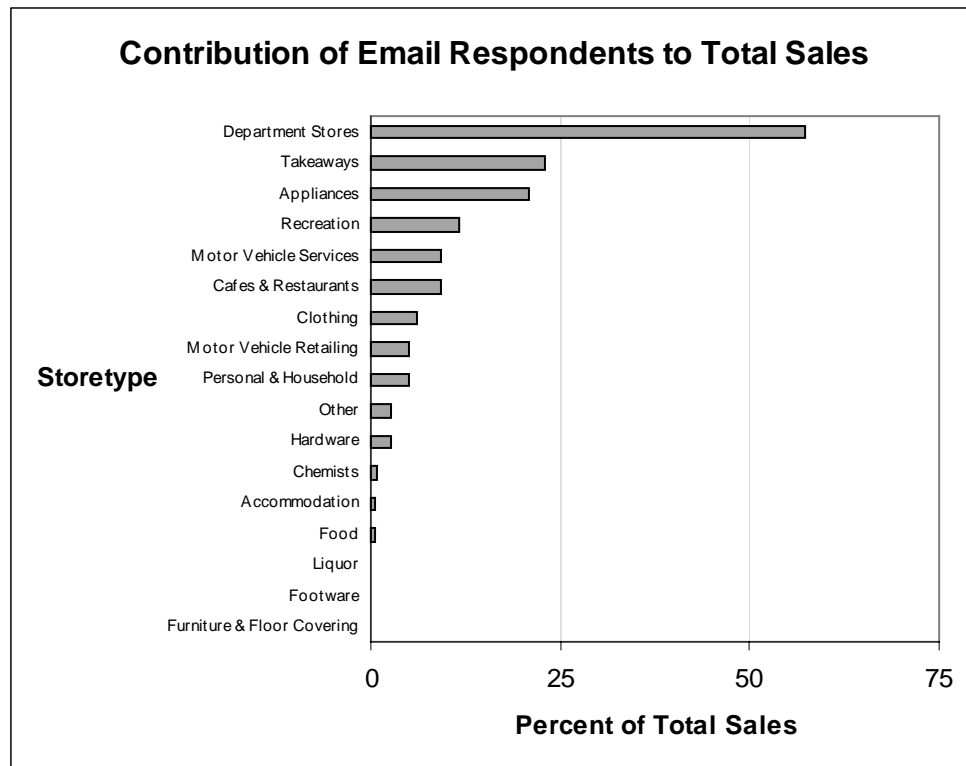


Figure 3

\*Contribution is an average of the four months analysed

## 4.2 Timeliness

In terms of timeliness, it is interesting to observe the pattern of email responses across the collection period. It has often been assumed that the use of email is an important response tool at the critical end of the collection cycle. However this assumption does not seem to be borne out by the actual results. This can be seen by averaging the cumulative response rate across 3 months (December excluded because of public holidays) and comparing email to the overall response rate, as shown in Figure 4.

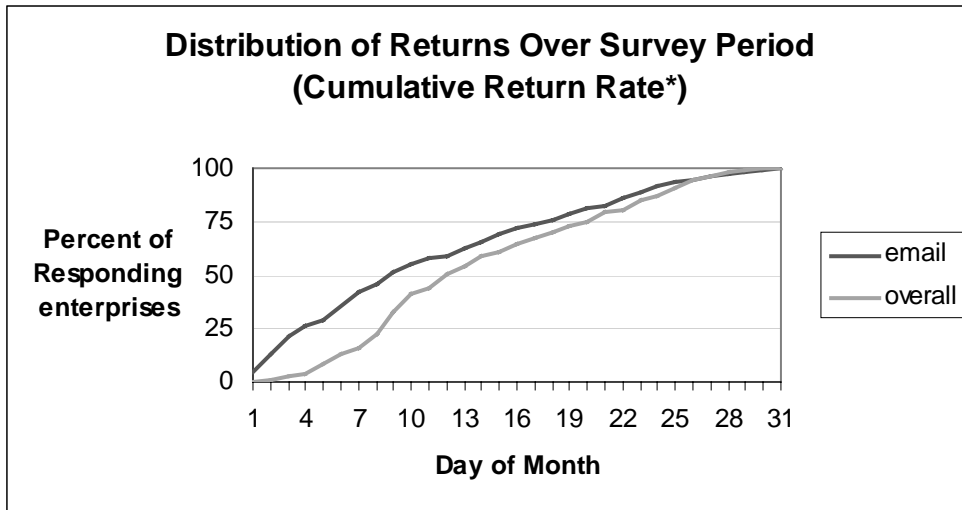


Figure 4

\*Since the number of email non-respondents is not known the return rate is based on the proportion of all returns received

There are two things that should be kept in mind when looking at this graph. Firstly, the post-out for any given month takes place towards the end of that month and before any respondents actually have the full period of data to return. This means the earliest possible return will be received after trading on the last day of the month surveyed. Secondly, the overall mark-in process is currently done as questionnaires are processed rather than when they were received, but the email receipt dates used are based on the date when they were actually arrived. In most instances questionnaires will be processed on the day they were received so any distortion should be minimised.

Having said this, it is still possible to see a clear difference in response timeliness. For example, 50% of email had been returned by day 9 compared to the overall 50% return rate not being achieved until day 12. Figure 5 shows response timeliness in another way and reinforces the fact that email 'peaks' earlier in the response cycle. This can perhaps be attributed to the relative transit time of the response (that is, it only takes a few minutes for an email transfer of data versus a few days for mail). As noted above email does not seem to be significantly more important at the tail end of the response period.

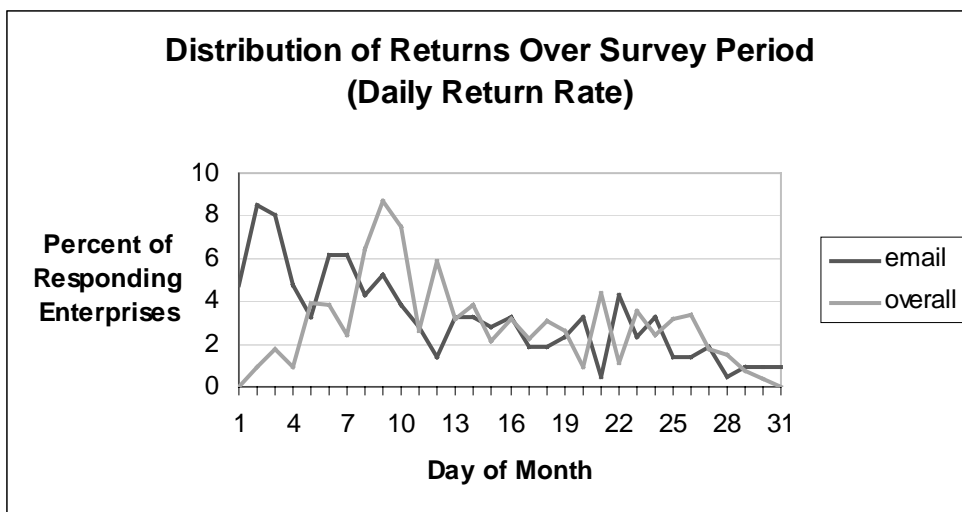


Figure 5

### 4.3 Data quality

In terms of the quality of the email returns, it has not been possible to properly compare the raw data received through email with that received by post as only email data has been examined in detail. However this does not preclude the identification of four main issues that seem to be of concern:

- absence of GST indicators
- absence of stock figures
- absence of data for some GEOs
- absence of respondent identifiers.

The first issue is the inclusion of an indicator for GST. As with any data of this type, it is important to understand whether data from an individual respondent is inclusive or exclusive of GST. As can be seen in Figure 6, around 24% of email returns had no indication of GST and this trend remained consistent over the four months examined; whereas a preliminary examination of paper questionnaires returned suggests that around 4% had failed to indicate whether or not GST was included in the sales figure provided. A modal effect does seem to be present. While it is necessary to do further investigation, it is possible that this is related to the separation of the questionnaire from the response.

It should be noted that where no GST indication is given, the GST status of the data can usually be assumed from how the individual unit has responded previously, as business do not tend to change their GST reporting status. While this may ease our concerns about missing GST indicators somewhat, there is still a potential problem if a respondent does change their behaviour (for example, if the person responding for that business changes).

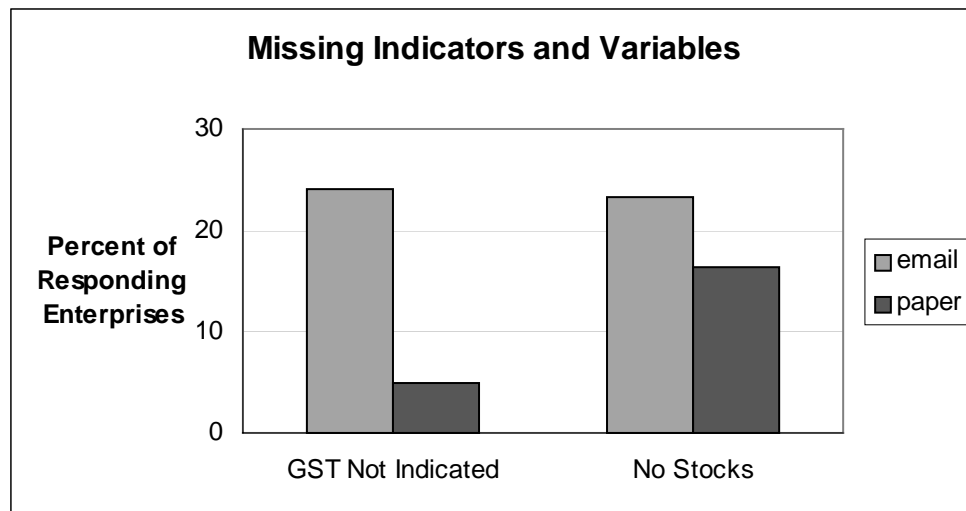


Figure 6

Analysis of the stock question asked every third month is less conclusive. The email returns that were examined included two of these stock months, December 2000 and June 2001. Figure 6 shows that of the enterprises responding by email, 23% provided sales but not stocks for at least one GEO (note that a stock figure of \$0 or any other indication that no stocks were carried counts as a stock figure provided). In comparison, the proportion for enterprises returning paper questionnaires in the same months was around 16%. While it is possible that the results from just these two months are not truly representative, there is some evidence that email respondents are more likely to not provide a stock figure. Further analysis may show more clearly whether there is a mode effect here (for example, analysis in relation to the relative importance of the missing stock figures).

Anecdotal evidence from survey processing staff suggests that stock months are problematic and it is not uncommon for respondents to have difficulty supplying stock figures. It is also apparent from both paper questionnaires and email responses that a significant number of stock figures seem to be estimates, since they are given in round figures (for example, to the nearest thousand) whereas sales seem to be more exact.

The third quality issue is another type of item non-response; multi-GEO enterprises that provide data for some but not all of their GEOs. Figure 7 shows the proportion of responding multi-GEO enterprises where there is at least one GEO for which no data has been provided. What is most noticeable here is the considerable variation in this data quality measure for the email respondents across the four months, ranging from 28% to 5%. This is perhaps attributable in part to the small number of multi-GEO enterprises responding by email. However there is clearly some evidence here that email respondents are more likely to 'miss' GEOs. It is possible that this also is related to the separation of the response from the questionnaire, as the locations that an enterprise is requested to provide data for are pre-printed on the questionnaire.

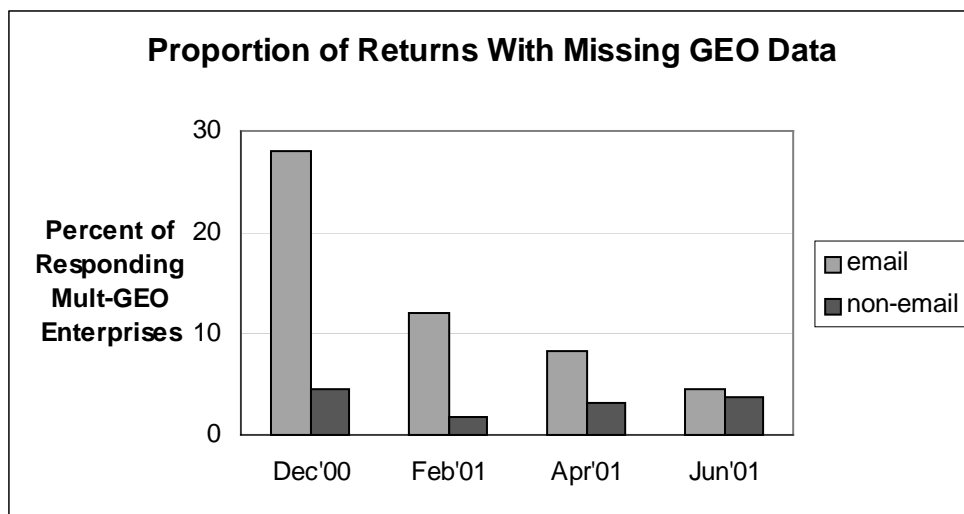


Figure 7

A further quality issue is the absence of identifiers on some email returns; a significant proportion of email returns come in without an indication of which statistical unit the data should be associated with. Almost 14% of email returns were lacking any identifiers, and a further 6% had an enterprise identifier but no identifiers at the GEO level. For single unit structures this can usually be resolved by matching a name with units on SNZ's Business Frame, but this can be a laborious task if more complex units are involved. In some instances the identifiers provided (either enterprise or GEO) had actually been mistyped. Again this problem is most likely related to the separation of questionnaire from answers, as all paper questionnaires sent out by SNZ (including those sent out as reminders) are labelled or printed with identifiers.

#### 4.4 Process integration

To date SNZ survey systems have generally been designed around the receipt and capture of paper documents. The RTS is no different. Email data has proven to be difficult to integrate into the existing system. For example, all the emails that contain only text (as compared to those with files attached) have to be examined by SNZ staff and the data transferred manually into the processing system. This manual process is helped in a small way by the consistency of format in the emails themselves; most respondents seem to have been provided with a format to use and have done so consistently over a number of survey cycles.

Emails with attachments require even more manual intervention than those with plain text. Recent attempts to automate the manual processes, via automatic uploading of data in comma-separated-variable format to the processing system, revealed the variations in format. While not all attachments are Excel spreadsheets, even when only the Excel attachments are considered, virtually every attachment was different in terms of layout and data format. Some of the differences include:

- positioning within the spreadsheet
- spacing in relation to rows
- inclusion of location identifiers
- figures given in \$000s
- figures given as negative values.

In order to make use of the automatic uploading system, each attachment must be reworked so that it appears in a format consistent to every other. In some cases the amount of work required to do this is quite considerable. In one extreme example, figures are provided for close to 600 different 'locations' that do not correspond to SNZ statistical units.

#### **4.5 Lessons from the case study**

The case study highlights some interesting experiences in the use of email as a means of data collection. While the analysis in this paper does not suggest that the use of email contributes to any systematic biases in the data, the case study highlights some areas for further investigation. To ensure that SNZ remains responsive to the needs of respondents in supplying data, SNZ has been prepared to offer a range of data collection media. The case study illustrates the need to examine whether any of the data quality issues noted here are also evident in those other media. Work is currently underway to conduct a similar analysis of respondents who provide data by fax and telephone. This will help SNZ and data users understand the implications of the collection methods on the data.

The extra analysis will also be helpful as a context when addressing the data quality issues that seem to be related to the means by which data has been provided, namely the problems with missing GST indicators, missing GEOs and missing survey identifiers. It will be important to understand whether these problems are more related to the size of the respondent, rather than the data supply method. Anecdotal evidence prior to the uptake of email suggests that some respondents were still creating spreadsheets and supplying SNZ with printed or faxed copies. If the problems are related to the separation of response from questionnaire, then they cannot be considered as solely related to the means by which data is returned.

The RTS is about to undergo major changes in terms of developing a new survey processing system and relocating most of the processing and output work to another centre (the inward receipt of all types of data will still be co-ordinated from the current location). As noted previously, SNZ is also in the process of developing a web based data collection system that will be used instead of email. The changes to the RTS in general, combined with a new collection system, is an opportunity to take steps to resolve some of the problems. A much closer relationship with respondents who supply data electronically is one of the intended outcomes of moving collection to the web based system. By establishing and maintaining these relationships it should be possible to exercise more control over the type and format of data that is currently being supplied. An associated outcome will be a requirement to provide at least one high-level survey identifier with any given submission. This will not necessarily help with problems related to multi returns, but it will at least ensure that every electronic response comes from an identifiable source.

Another means of achieving better data quality could be the creation of Excel templates or specifications that can be given to large multi respondents, as well as or instead of paper questionnaires. If data can be supplied in a consistent format by different respondents, then the amount of extra work currently done to prepare spreadsheet returns for automatic uploading to survey systems could be greatly reduced. Individualised templates may also help respondents associate the data they supply with survey identifiers and provide a consistent prompt for GST indicators.

At the smaller end of the scale, where respondents tend to supply data as plain text, a type of e-form could be incorporated into the web based collection system. This would help greatly in ensuring that all survey information required is provided as part of the submission.

The pressure from respondents to provide data in a format that best suits them is not only found in the RTS. In an agency such as SNZ, where the responsibility for respondent management has been centralised, this pressure has been managed by remaining flexible to the needs of respondents. The case study however highlights the importance of ensuring that those responsible for data collection receive regular systematic feedback on the impact of different data collection methods on the quality of the final data.

## **5. CONCLUSIONS**

The case study of the experiences of using email for the collection of data for the Monthly Retail Trade Survey highlights some of the benefits and data quality issues that can arise from data collection in an electronic environment.

The benefits include the early provision of data and a more even distribution of returns across the response period. However the case study also highlights some important data quality problems, which may not always exist in a pen and paper environment. These include absence of identifiers, as well as absence of key explanatory variables. While in part these quality problems can be seen as caused by the separation of questions from the answers in the final return, the problems also highlight the need for more investigations into respondent behaviour.

The case study also highlights the need to consider the impact of changes in collection methodology within the wider survey cycle. The experience of the RTS is that the use of email for collection has not resulted in the cost reduction that might be expected. The costs of the survey, not only in terms of resolving the quality problems discussed above, but also in transforming the email data into formats that can mirror the existing processing environments, are still higher than they should be.

The environment for business survey data collection in New Zealand is considerably more difficult than it was in the past. Pressure exists to ensure that the respondent load on businesses, especially small and medium businesses is minimised. Technology is increasingly seen as providing some of the solutions. However, as the case study reminds us, solutions adopted in the collection phase must be integrated into the total survey cycle in order to meet the need for quality, timely data in a cost effective manner.