Maximising the use of administrative data in sub-annual business collections

Paper presented at the New Zealand Association of Economists Conference,
at Auckland, New Zealand,
2-4 July 2014

Mathew Page

Senior Analyst, Platform Operations, Statistics New Zealand
Private Bag 4741
Christchurch, New Zealand
mathew.page@stats.govt.nz

www.stats.govt.nz

Nicholas Cox

Statistical Analyst, Platform Operations, Statistics New Zealand
Private Bag 4741
Christchurch, New Zealand
nicholas.cox@stats.govt.nz

www.stats.govt.nz

Liability

The opinions, findings, recommendations, and conclusions expressed in this paper are those of the authors. They do not represent those of Statistics New Zealand, which takes no responsibility for any omissions or errors in the information in this paper.

Citation

Page, M J & Cox, N R (2014, July). *Maximising the use of administrative data in sub-annual business collections.* Paper presented at the New Zealand Association of Economists Conference, Auckland, New Zealand.

Abstract

Statistics New Zealand (Statistics NZ) intends to use administrative data as the primary source of information to produce business statistical outputs. Statistics NZ has undertaken an investigation into developing sub-annual business collections based on Goods and Services Tax (GST) administrative data. GST is a comprehensive 'value added' tax collection undertaken by Inland Revenue. The investigation comprised three parts: establishing the desirable outcomes from using GST in business collections; producing an assessment of the GST data to determine its fitness for use; and outlining the methodological options available for using GST data within business collections. It was found that the underlying structure of the business units and their GST data characteristics were diverse. As each of the methodological options is suitable for different parts of the business population, there is not a ‘one size fits all’ solution for maximising the use of GST in business collections. This is illustrated in some statistical trial outputs from the manufacturing and wholesale trade industries.

Keywords: fitness for use, administrative data, quality, desired outcomes.

Acknowledgements

The authors thank John Stewart, Chen Chen and Vera Costa for their contribution into the development of the ‘assessment model for maximising GST use’ described in this paper. Section 2 is largely based on previous work presented at the International Conference on Establishment Surveys in June 2012.

1 Introduction

1.1 This paper

The purpose of this paper is to present an approach for maximising the use of administrative data for the production of financial statistics within sub-annual business collections. It is expected that the approach outlined here will inform Statistics NZ’s future use of administrative data.

The paper starts by describing an assessment model for using GST data within business collections. It describes how this model looks using the GST ‘sales’ variable in a quarterly sub-annual financial collection. GST is a comprehensive administrative based tax collection undertaken by Inland Revenue. It is a ‘value added’ tax that provides sales and purchases information. The GST ‘sales’ data was chosen for assessment as it is potentially the most comprehensive administrative data available and thus most useful for a sub-annual output. While sub-annual sales data is published in its own right, it is also used by the National Accounts in the production of quarterly Gross Domestic Product (GDP).

The application of the assessment model found that the underlying structure of the business units and their GST data characteristics are diverse. To maximise the use of GST in business collections different methodological solutions are required across the business population. This is illustrated as part of some trial outputs which compare the GST-based sub-annuals estimates with Statistics NZ’s published output from selected quarterly manufacturing and wholesale trade series.

The paper concludes by discussing some issues raised from exploring approaches to maximise the use of GST data. These issues are currently being investigated by Statistics NZ.

1.2 Statistics New Zealand’s use of administrative data in economic statistics

Over the past decade we have seen advances in Statistics NZ's use of administrative data, particularly in the economic area. Following these initial developments, the Statistical Architecture[1] and wider Statistics 2020[2] programmes have been created to provide the strategic foundation upon which Statistics NZ will transform the collection and production of statistics.

[1] McKenzie, R (2008). Statistical architecture. Statistics New Zealand, Christchurch.

[2] Statistics New Zealand Strategic Plan 2010-2020

As part of transforming its collection of statistics, Statistics NZ's vision is to have administrative data as the primary source of data to produce its statistical outputs. Data will be directly collected by Statistics NZ (eg by survey) only when necessary[3]. This would be a significant paradigm shift in the collection and production of business statistics. This need is being driven by increased customer demands for data, to create efficiencies in production of official statistics, and to decrease Government compliance costs upon our business population.

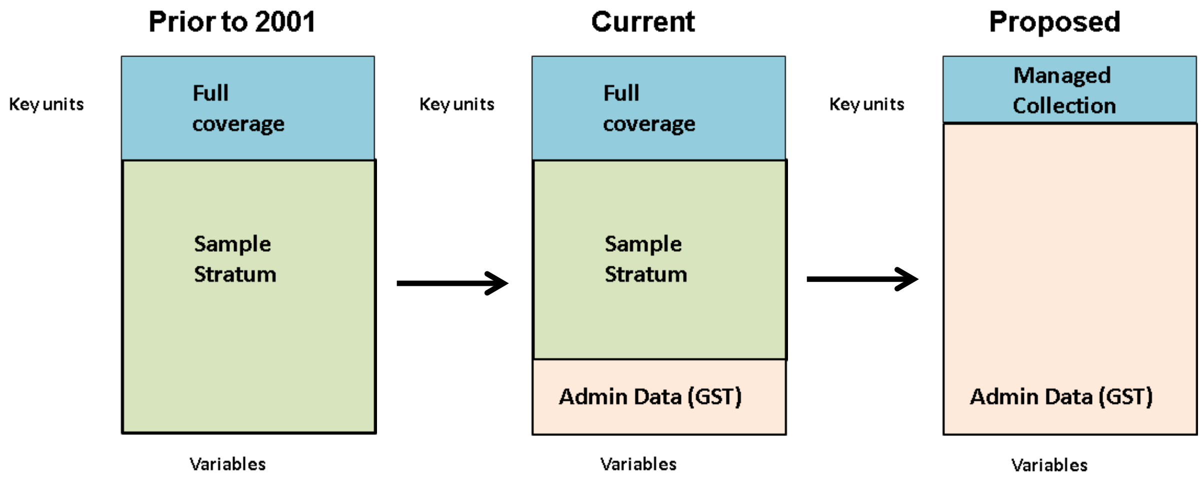
Statistics NZ has been using administrative data like GST in selected business collections for a number of years as a supplement to survey data. As GST data is essentially accounting information similar in nature to that supplied by respondents in Statistics NZ's surveys, it is thought to be suitable for wider use within our collections. Thus we want to assess how Statistics NZ can utilise such data as a first source for sub-annual business collections.

1.3 Historic use of GST in sub-annual business collections

There are two main variables required as part of the Statistics NZ sub-annual business collections – sales and inventories. These have historically been collected in three separate surveys – the Economic Survey of Manufacturing (ESM), the Wholesale Trade Survey (WTS), and Retail Trade Survey (RTS). While the sales and inventories information is published by Statistics NZ directly from the sub-annual business collections, this data also flows through to the National Accounts and is used in the production of quarterly Gross Domestic Product (GDP).

Prior to 2001 Statistics NZ's sub-annual financial collections used a postal survey design. The design included a full coverage strata for large businesses, where all businesses above a size threshold were surveyed. This is illustrated in Figure 1.

Figure 1
History of sub-annual financial statistical designs



In 2001, a Goods and Services Tax (GST) administrative data component was added to the design for the first time as a supplement to the survey data. This resulted in significant reductions in the postal sample. GST is a comprehensive 'value added' tax collection undertaken by New Zealand's Inland Revenue Department. As GST data is essentially accounting information similar in nature to that supplied by respondents in Statistics NZ's surveys, it is thought to be suitable for use within our collections.

This design is still currently used in statistical production. The GST component has been constrained to contribute no more than 15 percent of sales value for ESM and WTS industries, and 10 percent for RTS industries. While the use of GST data in the current designs are considered fit for purpose, there are concerns that raising the threshold using the current statistical processes will impact on the quality of the outputs. The methods that are used to transform the sales and inventories data for the GST component of the population to make it

[3] Statistics New Zealand Collection Strategy 2011

suitable for use in the sub-annual business outputs have a number of known limitations, including:

- Methods to transform the GST data for businesses that file on a two-monthly and six-monthly basis to a quarterly frequency.
- Apportioning GST data to GST group members to allow homogeneous industrial data to be produced. Some businesses linked by ownership choose to file all their GST through an individual group member (which can include more than one industrial activity).
- Identifying capital transactions and sale of going concern businesses which can be included in the GST data. These transactions are not part of the conceptual measure of sales and purchases required for National Accounting purposes.

Statistics NZ has been working to progress this methodological thinking so the use of GST data can be maximised in the production of sub-annual financial outputs. The proposed design is outlined in Figure 1, where GST data is used wherever possible, supplemented by a managed collection of large and complex businesses where the use of GST data is not suitable. These businesses will be included in a future Statistics NZ quarterly managed collection where we will continue to collect sales, inventories and other key economic variables that are required on a sub-annual basis. The managed collection strategy is fully described in section 4.1.

An important feature of the proposed design is the use of GST allows ‘census-like’ coverage, as opposed to the sample survey approach used in the previous designs. There is also a significant reduction in respondent burden where the number of businesses receiving postal survey forms will decrease. While our current focus is to realise these efficiencies in the existing sub-annual outputs, in the future we intend to apply this design to measure sub-annual financial statistics for industries not currently covered by Statistics NZ. For example, services industry financial indicators.

2 Maximising administrative data use

2.1 The approach

The approach taken has been to develop a process for assessing the production of statistics using administrative data as the primary data source[4]. While the approach could be applied to a wide range of industries and administrative data variables, as a pilot we will concentrate on the use of GST sales data in the quarterly estimation of manufacturing and wholesale trade activity.

The approach used to measure the inventories variable (not available in sub-annual administrative datasets) in the proposed design is assessed separately and is further discussed in section 4.3.

The GST sales assessment gives a clear picture of when, where, how, and why administrative data can and should be used within a statistical output. It is helpful to note that the full use of administrative data may require different approaches for different parts of the business population.

2.2 An assessment model for GST sales data use in sub-annual collections

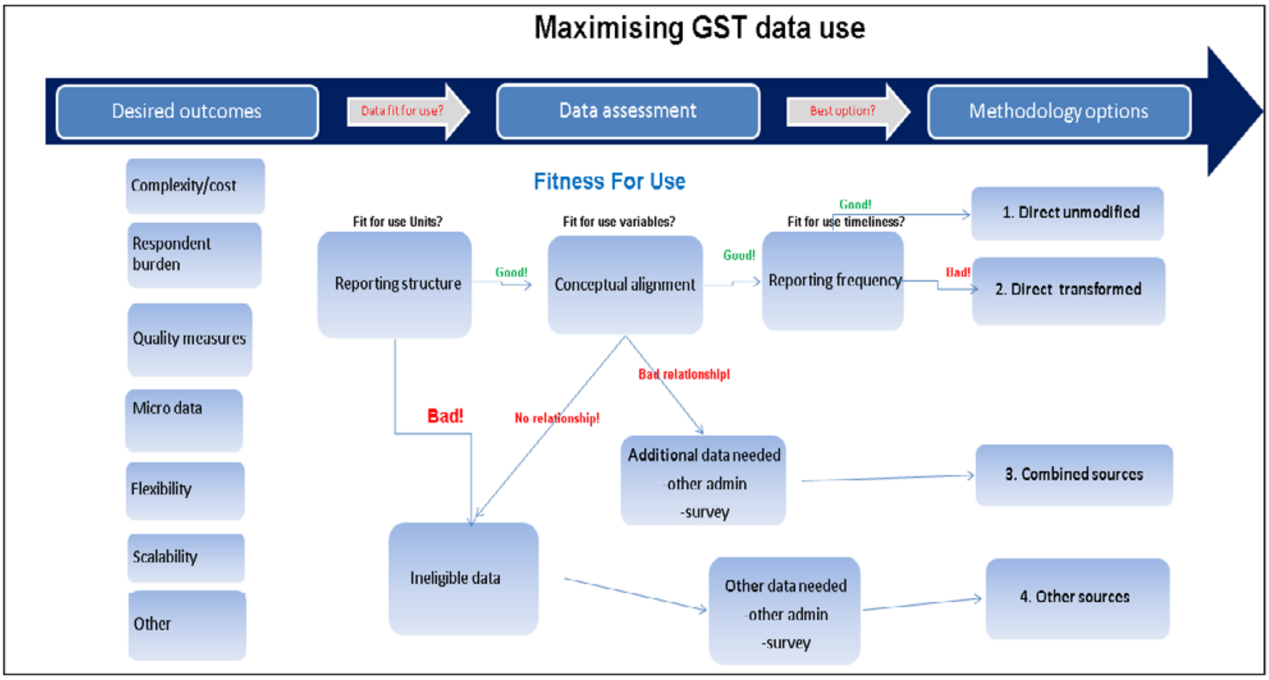
A process for assessing how and where GST data can be used within sub-annual collections using the 'sales' variable is shown in Figure 2.

To maximise the use of GST sales data we established a process of assessment looking at three facets that enable decisions to be made. We assess why the data should be used in a particular way (ie the ‘desired outcomes’ of GST use); what GST data should be used (ie the data’s fitness for use); and how the data should be used (ie the appropriate method). Even though the GST data has not been collected for statistical purposes, it is expected that the majority of the GST data can be used in the production of statistical outputs. This expectation arises as administrative data collected via the tax system is sourced from business accounts and therefore is likely to conform to our needs.

[4] The use of administrative data in the collection design and edit and imputation process are not discussed in this paper.

In Figure 2, the desired outcomes are shown on the left hand side of the diagram. Data is assessed to establish its fitness for use by examining various aspects of the data such as reporting structures and conceptual alignment. These are shown as decision boxes. Depending on the data assessment and its fitness for use, the most appropriate methodology is established.

Figure 2
GST sales assessment model



2.3 The desired outcomes from GST sales use

In reviewing the approaches for the use of GST sales data, it became apparent that there are a number of desirable outcomes that need to be considered. It is recognised that not all desirable outcomes are of equal importance and potential trade-offs between outcomes would need to be considered as part of any use of GST data. Maximising the use of GST data whilst also maximising the outcomes from its use should be our goal.

Key outcomes identified included:

2.3.1 Minimise complexity / cost

The level of complexity associated with each approach needs to be considered. A higher level of complexity will likely result in greater resources and a higher production cost.

2.3.2 Minimise respondent burden

The impact on the level of respondent burden (or compliance cost) needs to be examined. A key aim is to minimise this as much as possible.

2.3.3 Maximise measurable quality

The examination of options for using GST data must maintain the need for suitable quality measures. For instance, traditional survey errors (eg. sample errors) may become less relevant, while the importance of the errors associated with any models used for manipulating the administrative data will increase. Nevertheless, it is necessary to accept that it will be difficult to have measures that encompass all aspects of quality, some error will be difficult to quantify such as a 'non-modelling' error analogous to 'non-sample' error.

2.3.4 Maximise flexibility

A strength of GST data is its level of near ‘census’ coverage. This has the potential to deliver sub-domain estimates at lower levels of detail for the likes of ad-hoc research, and customised data requests. This is generally not possible for sample surveys. The flexibility of each option to enable the ad-hoc production of specific sub domain estimates should be noted.

2.3.5 Maximise scalability

Scalability refers to the ability of an option to be used in the development of new 'green-field' collections. For example, in business surveys this could include the development of collections for industries in the economy which are not covered by existing collections (eg services industries).

2.3.6 Maximise unit record availability

There is an increasing need for micro-data which will underpin future statistical analysis. This supports research and policy's impact evaluation, where the emphasis is on micro-data analysis and the integration of data from different sources. Micro data also makes it easier to meet a range of emerging needs.

2.4 Data assessment

The purpose of the sub-annual financial data series is to: a) provide early identification of turning points in the economy; and b) enable time series analysis. Purpose and fitness for use should guide the use of GST sales data within the collection and shape our methodology.

The main focus for our data assessment is determining whether the data is able to meet our quality criteria. Tax system sourced administrative data like GST is business accounting data that is expected to be fit for use within our collections. However, GST data has not been collected by Statistics NZ and an assessment is required to determine if the unit and sales variable definitions meet our requirements for the data to be fit for use. There are a number of aspects of the data that may affect our ability to use it within a statistical output. Thus the GST data is assessed to determine whether the data can be used and, if so, how it can best be used.

Our process is designed to establish whether the administrative data is fit for use in our statistical outputs in spite of the absence of the statistical controls that would usually ensure this would be the case. To maximise the use of the GST sales data we need to recognise different characteristics within the business population. Different ‘parts’ of the data may be fit for use in different ways. One of the key differences in the data relate to the unit structures and reporting frequency from which the data is obtained. Another is the measure obtained from the data and its relevance to the ‘target concept’. The GST data will have characteristics that need to be assessed to establish whether the data can be confidently used in a statistical output.

2.4.1 User needs and fitness for use

Considering where, and how, GST sales data can be used within the sub-annual collections requires us to recognise the various user needs and also recognises that not all the needs are of equal priority or needed to the same level of quality. This reinforces that there is no ‘one size fits all’ use of GST data and maximising the benefit of the GST data requires us to assess the data in light of the differing user needs.

2.4.2 GST reporting structures

Administrative data is usually collected for the legal unit and this is the case with GST. Integrating the Statistics NZ business register with the legal units used for GST enables us to identify legal units that are most likely to be statistically fit for use. Such legal units are those which correspond to a statistical unit with activity in a single industry. In addition some more ‘complex’ legal units can be shown to be fit for use for some purposes. However, data obtained from legal units which are ‘complex’ may not be fit for use in industry or regional statistics without some transformation. As the unit structures become more complex, the use of GST sales data becomes more problematic. Thus for these units there is a decrease in quality and potentially an increased need for an alternative to GST data to be used (eg a survey) thus increasing cost.

There may be other reasons why the reported data might be deemed unsuitable. In particular, it is possible that legal or administrative issues might render some population units data unsuitable for statistical use. An example would be where a class or group of units is required or allowed to report on a different basis than others. In the case of GST data the administrative arrangements allow a group of legal units linked by ownership to provide sales data as a total recorded against one unit whilst other units in the group record zero values. This type of group reporting may render the GST data ineligible for use in our output without some further transformation, particularly if the GST group covers multiple industrial activities.

2.4.3 Conceptual alignment

Having established the suitability of the reporting unit for our use we then assess how the GST sales data aligns to the conceptual and definitional requirements of a manufacturing or wholesale trade sub-annual output. The sub-annual collection of 'sales' is defined to be fit for use within the System of National Accounts. The definition of sales for this use is expressed in our existing surveys (the ESM, WTS and Annual Enterprise Survey (AES)). GST is designed as part of a value added tax regime with sales defined to measure the income side of value added. Value added is the concept that the National Accounts is trying to measure using our collections as inputs. It is therefore expected that GST sales data should meet our user need.

As GST data is not collected for statistical purposes there may be circumstances where the sales data does not exactly match the definition needed by users. However, the difference may not be materially significant and the GST data may still meet user needs. As we want to maximise our use of GST data we still accept such data for use. Validation by comparison to values from other sources may well reveal that definitional differences are in practice insignificant for some parts of the population, making direct use of the GST data for these parts of the population viable.

To ensure the alignment and relevance of the GST data to our target concept we first compare the definition of the GST 'sales' data against the user need as represented by existing survey definitions. Then we evaluate the quality of the definition of GST sales by comparison with the values of existing surveys both sub-annual and annual. These surveys have values that are defined and controlled to meet user needs through the questionnaire and survey design process. We examine the relationship between the sets of data based on a pool of common units. Where the evaluation shows that GST data closely matches the survey data then the GST data is deemed to be fit for use directly.

Where this is the case the GST data can be used in one of the 'direct' methodologies that have been developed (see below). Otherwise the data may be ineligible for use or require the use of another methodology (such as being used in combination with other data). Where GST is then used as a first source of data, its quality and relevance over time can be monitored by ongoing comparison of the unit values from an annual survey such as AES.

Our work to date has shown extensive use of GST sales data will be feasible across most manufacturing and wholesale trade industries. However, for a small number of industries we have identified a large level of disparity between annualised GST and AES sales for a significant number of units. For example, in the ‘commission-based wholesaling’ industry businesses can record their sales on either a gross or net (commission) basis. Our target measure of sales is on a net basis.

2.4.4 Reporting frequency

GST data is reported to Inland Revenue on several frequencies depending on turnover size. Higher turnover businesses report monthly, medium turnover businesses two-monthly, and low turnover businesses every six months. This sort of feature of administrative data may not be uncommon and needs to be addressed in our assessment process. Data must be available in time to be used.

For the quarterly manufacturing and wholesale activity statistics considered here, a monthly reporting frequency is ideal. Data for the three months of the quarter can simply be added. However, for the GST reporting frequencies greater than a month obtaining quarterly values requires some manipulation. This issue is addressed through a transformation of the sales data available for the period (eg the first two months of a quarter) using a model that includes the data for the monthly reporters. By doing so we maximise our use of the available GST data whilst also avoiding any recourse to other data. This approach is the 'direct - transformed' method explained below.

2.4.5 Timeliness

There is an underlying assumption that to use administrative data within our statistics outputs, we must have timely data. For use within a first release of an output the GST sales data must be for the current period being measured. Use of GST data solely from previous periods to model or forecast will not be fit for use in outputs which are needed to measure turning points and changes within the economy. However, in order to maximise the use of GST data it may not be possible to wait until all the data needed for the production of an output is available. Where GST data is only partially available for the current period then fit for use data may be able to be obtained through the standard imputation procedures. Where there is a significant proportion of the total value missing due to timeliness issues the problem of potential bias needs to be considered.

This timeliness issue could lead to the development of provisional estimates (eg for the short term economic indicators) which would be revised when all the GST sales data was available.

Assessment of the GST sales data shows that 85 percent to 95 percent of data by value is available within current ESM and WTS publication timeframes (note that the current ESM and WTS survey response rates are 85 percent). This varies across industries and quarters.

2.5 Method options

Our data assessment should enable us to determine what method options can be used for different parts of the business population. Can the sales data be used directly as it is received? Can it be used directly with some internal transformation? Can it be used in conjunction with some other data source? Or can it not be used at all?

2.5.1 Can we use the GST sales data directly within our output?

To use the data directly requires that the data meets some basic statistical standards such as an appropriate statistical unit and the relevance of the GST sales measure to our target concept. If these basic statistical needs are met within fitness for use bounds then the data can be used as a first source for a business collection.

Having established that the GST data is fit for use directly, we have two methodological approaches which enable us to maximise our GST sales data use.

2.5.1.1 Direct unmodified approach

The direct unmodified approach uses the GST sales data as it is received. The data becomes the measurement for the units in the target population.

This option can be used when the administrative data reporting structures are suitable, the desired statistical outputs are closely aligned with the concept measured, and the reporting frequency enables timely use within the collection. The approach allows the production of aggregate estimates and provides additional micro unit level data for analysis and other purposes.

Evaluation of the quality of any estimates produced through this approach is very straightforward as actual data is used and therefore there is no uncertainty arising from the use of modelling and/or sampling.

The outcomes of using this approach are:

- Is a simple low cost option - using inexpensive GST data;
- Has low respondent burden - reduced burden due to GST data use;
- Can easily measure quality since actual data is used;
- Is flexible. Where used, this approach is essentially a unit census and thus supports sub domain estimates and changes in output needs;
- Is easily extended to green fields collections;

- Supports micro data analysis.

In terms of GST sales data use for sub-annual collections, this method is applicable to the monthly filers. Since the values for each unit for each month can be directly added, any level of aggregation can be produced which makes this approach very flexible and able to be used to produce many different outputs. It is, however, limited to a subset of the available GST data ie monthly filers.

2.5.1.2 Direct transformed approach

The direct transformed approach uses one part of the GST sales data to ‘complete’ another part of the data. A deficiency in some GST data can potentially be overcome by modelling its relationship to some other GST data that is itself fit for use. This allows us to maximise the use of available data.

This option can be used when the administrative data reporting structures are suitable, and the desired statistical outputs are closely aligned with the concept measured, but some data transformation is needed to use it. For example, data manipulation may be needed to deal with some GST sales data characteristics such as the units' reporting frequency. In this case the data is used 'directly' in that it is not reliant on any other data source (eg a sample survey for modelling conceptual differences). The approach allows the production of aggregate estimates and may provide additional micro unit level data for analysis and other purposes.

The evaluation of the quality of any aggregate estimates produced through this approach must be based on the model(s) used for producing such estimates.

In terms of GST sales data use this method is applicable to the two-monthly filers, since in this case some data manipulation is needed to cope with the fact that it can't be added directly to produce a quarterly result. Modelling in conjunction with the monthly filers can transform the two-monthly data to quarterly estimates. No other data sources are needed if the concepts measured align with the estimates to be produced.

2.5.2 Can the GST sales data be used in conjunction with other data?

If the data is not fit for use directly due to issues with the conceptual alignment of the GST sales measure and the target concept it may still be able to be used if combined with another data source that can be used to align the GST data.

2.5.2.1 Combined sources approach

The combined sources approach combines the GST sales data with data from other sources (eg. non-administrative data) which is known to meet our fitness for use requirements. The approach can be used when there is some type of relationship between the administrative data and the statistical outputs, despite their concepts not being closely aligned. In this case unmodified GST data is used in combination with other data to model the estimates to be produced.

The evaluation of the quality of any aggregate estimates produced through this approach must be based on the model(s) used for producing such estimates.

This approach can reduce the respondent load for the survey and is considered an improvement over sample surveying without any use of auxiliary data. However, it does not permit the full realisation of benefits of only using GST sales data.

For example, this method is applicable to small and medium sized GST groups, where Employer Monthly Schedule (EMS)[5] data can be used to derive ratios to apportion the GST sales data to the individual group members.

[5] Employer Monthly Schedule data is collected on a monthly frequency by Inland Revenue for the purpose of administering New Zealand’s taxation system. It contains details of employee numbers, earnings, tax type, and tax deducted.

2.5.3 Is the GST sales data unfit for statistical use?

Whilst we endeavour to maximise our use of the GST sales data, in some cases the data will be unusable. As noted, to be fit for use requires that the data meets some basic statistical standards such as an appropriate statistical unit and some relevance of the GST sales measure to our target concept. Where the data does not meet these basic standards then it will be deemed as ineligible for use.

2.5.3.1 Other sources approach

The other sources approach is the use of non-administrative data sources. This may include the use of a ‘managed collection’ of large and complex enterprises where the use of GST sales data is not suitable (as depicted in figure 1).

It may also include the use of a sample survey more widely across an industry, if for example there are large conceptual differences between the GST sales data and the target concept. A sample survey is an established method that will not be discussed here in depth.

If ‘other sources’ are used, the approach has the strengths of being well understood, of measurable quality, and based around consistent user defined concepts. However, it is costly, and imposes high respondent burden.

2.6 Application of the assessment model

The data assessment undertaken ensures that any GST sales data use will meet our quality needs. From this assessment the most appropriate method can be selected that ensures we make maximum use of the available GST sales data. The methods in turn are assessed against the desired outcomes to identify the potential benefits that arise from the use of the GST data.

When reviewing the options, there are some general points that need to be recognised. Options should not necessarily be considered mutually exclusive. That is, it is not desirable to look for a ‘one size fits all’ option. So the methodological approaches presented should be considered as 'building blocks' that could be combined in different ways to produce the desired outputs.

It would appear that the two direct options (unmodified and transformed) are to be preferred when the GST sales data permits. These two options have the potential to meet all our desired outcomes. The combined sources and other sources options whilst not meeting all our desired outcomes will still be useful in many situations.

This GST sales assessment model has been used to produce an experimental sales series of Christchurch retail trade and hospitality activity, following the Canterbury earthquakes of 2010 and 2011[6]. Since this initial release, we have been focusing on applying the GST sales assessment model to sub-industries within the current ESM and WTS collections.

3 Trial sales outputs

The application of the GST sales assessment model has allowed us to generate some trial sales series for analysis, using the proposed design as illustrated in figure 1. The assessment model was applied to historical sub-annual financial and GST sales data to allow us to compare the trial series with Statistics NZ published results. The trial series need to be analysed with caution as they are based on experimental methodologies that have not been through the same rigorous statistical processes as our official outputs.

We used a trial managed collection to create the new series which is further discussed in section 4.1. As businesses in the trial managed collection are large and complex, the majority were already included in our existing sub-annual survey designs. So we had a valuable supply of historic survey data available to use as part of our simulations. The references to the ‘other sources’ methodology option in this section in essence refers to the data from the trial managed collection.

The ‘petroleum and coal product manufacturing’, ‘fruit, oil, cereal, and other food manufacturing’, and ‘machinery and equipment wholesaling’ industries have been selected across the ESM and

[6] Christchurch retail trade indicator (information releases), Statistics New Zealand, Wellington, 2014.

WTS outputs for comparison and are illustrated in figure 3. The graphs compare the two sales series from the June 2010 quarter to the December 2013 quarter.

The graphs show that for all industries in general the pattern of change over time is similar for the two series - the lines on each graph move up and down almost in sync with each other. The results for these three industries are typical of most industries across the manufacturing and wholesaling industries – that is while the comparisons look encouraging there are some examples where the series are moving in different directions and/or there is a temporary or consistent level difference between the two series. Our investigations to date have identified a number of common themes which help explain the differences in the two series:

- The published series are often based on a postal sample which are designed to give statistics to a certain level of accuracy. The trial series are based on the entire business population with no sampling used. This sometimes results in a consistent level difference between the two series if there is some degree of statistical sample bias in the published estimates.
- The trial series includes capital transactions and sale of going concern businesses which can be included in the source GST sales data. These transactions are not included in the published series.
- The managed collection in the trial series uses a ‘snapshot’ of enterprises as at the June 2011 quarter which is applied across the entire timeseries. While our analysis has found the group of large and complex businesses in the NZ economy tends to remain relatively stable over time, there are sometimes changes to these businesses in the real world (eg business restructures) which can result in disparities between the two series.

Table 1 shows the percentage of value by methodology option from applying the GST sales assessment model. The three industries presented use the methodology options to varying degrees.

The ‘petroleum and coal product manufacturing’ industry is dominated by a small number of large and complex businesses where the use of GST sales data not suitable, so most of its value can be attributed to ‘other sources’ (or the trial managed collection).

The ‘fruit, oil, cereal, and other food manufacturing’ industry makes more use of GST sales, although over half the value can still be attributed to ‘other sources’.

The ‘machinery and equipment wholesaling’ industry predominantly uses GST sales data, where two-thirds of value can be attributed to administrative data sources.

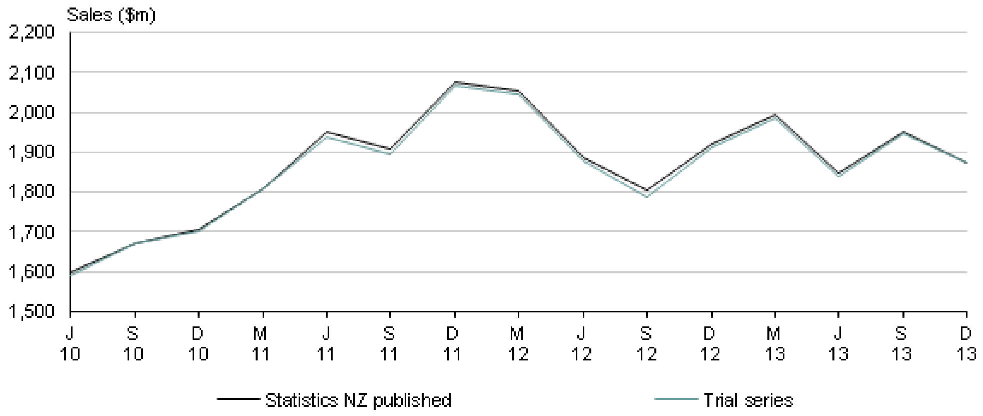
As described in section 1.3, the current designs constrain the use of GST data to no more than 15 percent of value, so the trial outputs show much more extensive use of GST data. But these industries show there is no ‘one size fits all’ option and different methodological solutions are required across the ESM and WTS industries.

Table 1
Percentage of value by methodology option

Method	Petroleum and coal product manufacturing	Fruit, oil, cereal and other food manufacturing	Machinery and equipment wholesaling
Direct unmodified	0	18	27
Direct transformed	1	24	39
Combined sources	0	3	1
Other sources	99	55	34

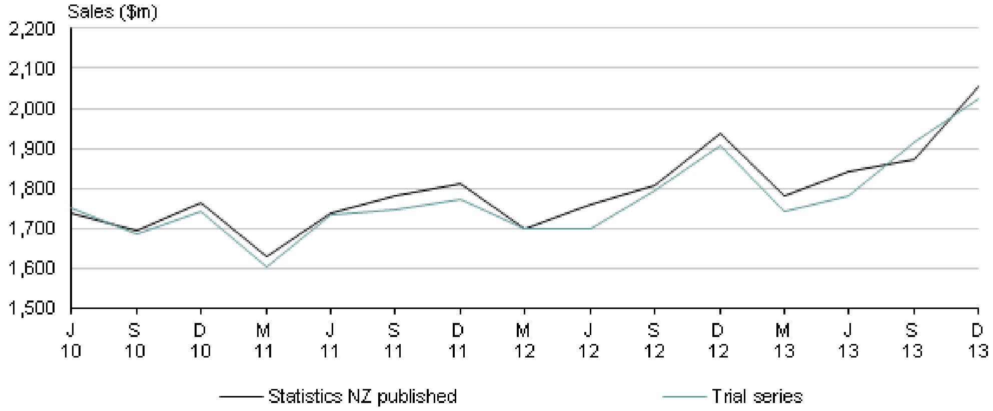
Figure 3
Trial sales outputs - June 2010 quarter to December 2013 quarter (unadjusted)

Petroleum and coal product manufacturing



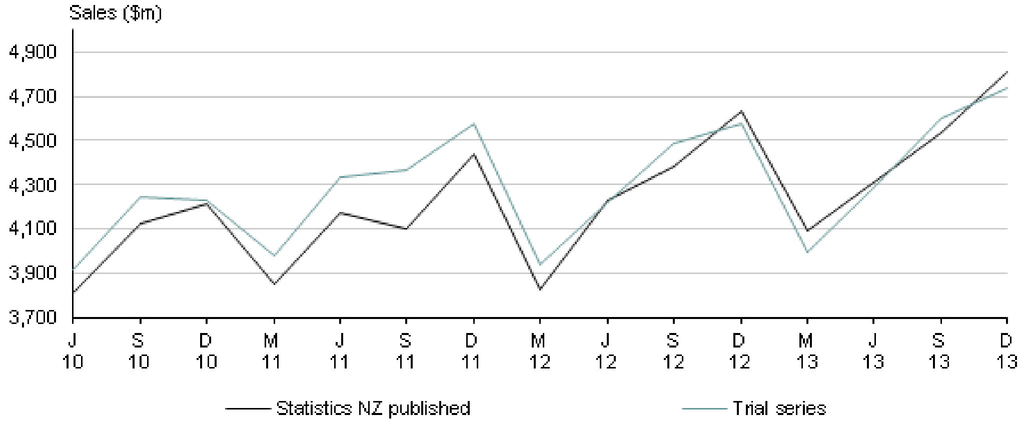
Source: Statistics New Zealand

Fruit, oil, cereal, and other food manufacturing



Source: Statistics New Zealand

Machinery and equipment wholesaling



Source: Statistics New Zealand

4 Current work

Statistics NZ is currently investigating three issues which need to be resolved before the proposed approach is used in the production of official statistics. This includes the development of:

- a managed collection strategy where the use of administrative data is not suitable
- an editing strategy for dealing with outliers (eg inclusion of capital) in the GST data
- methods for measuring inventories, a variable not available in administrative datasets on a sub-annual basis.

4.1 Managed collection strategy

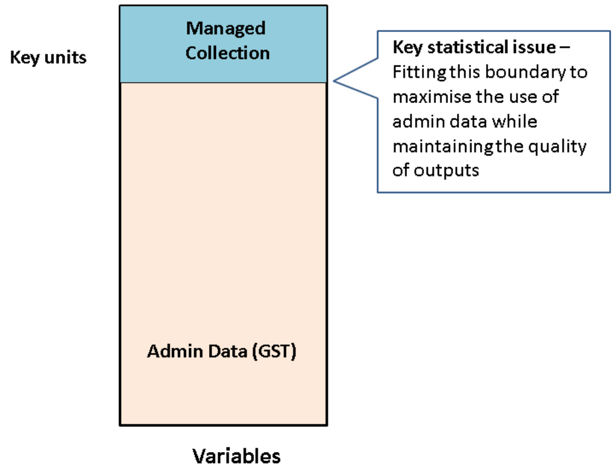
The use of GST sales data for large and complex businesses is at times not suitable. For example as described in section 2, administrative data for businesses with complex reporting structures can be deemed to be unsuitable for our statistical use. These businesses will be included in a Statistics NZ quarterly managed collection where we will continue to collect sales, inventories and other key economic variables that are required on a sub-annual basis.

Very large businesses with simple reporting structures may also be included in the managed collection so we can retain a degree of 'statistical control' over their reporting. This is so we can continue to contact these businesses with queries about their data, collect any additional variables that are not available from the administrative data on a sub-annual basis, and minimise any risk associated with using their GST data directly. The risks from using GST directly for large businesses include the inclusion of items outside the true conceptual measures for National Accounting purposes, or any GST processing lags that do not meet our publication timeliness requirements. Statistical adjustments of the GST sales data to deal with both of these facets for large businesses may have a significant material impact on our outputs.

The key statistical issue for defining a managed collection strategy is determining the boundary to maximise the use of administrative data while continuing to maintain the quality of our outputs. We have been producing a number of simulated series using different business rules, and comparing the results against currently published output. The business rules used to produce a managed collection for creation of the trial outputs are described below. Further work is required before we can confirm these rules will be used in the future to create the proposed new series.

- A $100 million significance rule - if an enterprise, or group of enterprises linked by ownership, have an annual GST turnover of more than $100 million.
- A 3 percent industry dominance rule - if an enterprise makes more than a 3 percent contribution to annual total income for an industry.
- All enterprises that have a significant level of activity across more than one industry.

Figure 4
Managed collection strategy


Key units
Managed Collection
Admin Data (GST)
Variables
Key statistical issue – Fitting this boundary to maximise the use of admin data while maintaining the quality of outputs

We also are developing statistical processes for maintaining the managed collection over time. This is to ensure the managed collection continues to reflect economic reality on a quarterly basis, and takes account of any business restructures, mergers, split-offs etc, as well as significant growth or reduction in the size of large businesses. Our analysis to date has found that the group of large and complex businesses in the NZ economy tends to remain relatively stable over time. As such, we used a 'snapshot' of enterprises as at the June 2011 quarter to form our managed collection for the trial outputs.

Work is continuing to develop a robust managed collection strategy to meet these longer term goals.

4.2 GST editing strategy

GST administrative data is not collected for the purpose of producing financial or economic statistics. There may be circumstances where the data does not exactly match the definitions needed by users. For instance, there are two known transactions that can be included in the GST data that are not part of the conceptual measure of sales required for National Accounting purposes – sales of capital items and of going concern businesses.

Statistics NZ is developing editing methodologies to attempt to identify these transactions. This is to ensure the GST data being used for businesses outside the managed collection is fit for purpose in the production of sub-annual financial outputs.

The proposed methods reflect internationally recognised practice, being based on Statistics Canada's Banff editing and imputation processes[7]. They apply the statistical process control method, which means for a given business in the GST data collection, all values of one variable are expected to fall within a certain range; the range being a function of the values previously observed. Any values that fall outside the range are then identified as being potentially anomalous.

Our analysis to date (which has mainly focused on the manufacturing and wholesale trade industries) has shown only a relatively small number of enterprises are identified as potentially anomalous on a quarterly basis. Therefore, in a future statistical production process we anticipate analyst input will be obtained before an adjustment is made. The factors that will be considered before any adjustment is made include:

- How large is the anomalous value and what material impact will the adjustment have on aggregate industry level outputs?
- What GST item is the anomalous value sourced from? For example, a sale of a going concern business is 'zero-rated' for GST purposes and recorded separately. Capital items are more difficult to identify as they are recorded with general sales transactions.
- What is the behaviour of other administrative variables? For example, is a large increase in GST sales matched by a similar rise in GST purchases, or employee numbers in Inland Revenue's EMS?

4.3 Measurement of inventories

Quarterly inventory measures are required by National Accounts in the production of GDP. In the current sub-annual design inventories are primarily measured from data collected in the postal survey. A small proportion of inventories are modelled for the administrative data component of the population (refer to figure 1).

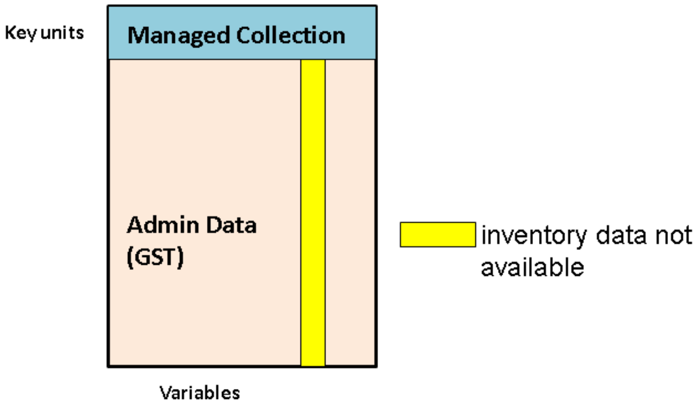
In the proposed design, we will continue to collect inventories data for businesses in the managed collection. However, there will be no quarterly inventories data available for businesses outside the managed collection (illustrated in figure 5). Statistics NZ has therefore been investigating methods for measuring quarterly inventories in the proposed design for the administrative data component of the population (which will be considerably larger than in the current design).

[7] Banff Version 1.04 (2005, June). Generalized System Methods Section & Business Survey Methods Division, Statistics Canada, Montreal.

The work to date has found different methods will be required across the manufacturing and wholesale trade sub-industries. The methods include:

- Using the managed collection inventories data to calculate statistical estimates directly. The managed collection enterprises are at times responsible for most (if not all) of industry inventory estimates.
- Modelling quarterly inventory estimates from administrative data sources such as GST sales and purchases, and annual inventories in Inland Revenue's IR10 return.
- Using a targeted postal collection (in addition to the managed collection) to estimate inventories. This is similar to the current methods.

Figure 5
Measurement of inventories



5 Summary

This paper presented an approach for maximising the use of administrative data in the production of statistics within sub-annual business collections. We describe how an assessment model can be used to determine what method options are suitable for maximising GST sales data use in different parts of the business population. The assessment model ensures the application of GST sales data meets user needs through a fitness for use assessment, and describes how the method options meet desirable outcomes from using GST in business collections.

The assessment model is applied to produce trial sales series for selected manufacturing and wholesale trade industries. These series are compared against existing Statistics NZ published output. While the comparison looks encouraging, there is further work required by Statistics NZ before the proposed approach can be used to produce official statistics. This includes the development of a managed collection and GST editing strategy, and methods to measure inventories on a quarterly basis which is not available in administrative datasets.

References

Brisebois, F, Girard, J, & Xie, H (2007, June). Challenges surrounding the use of tax data in the development of quarterly indicators for various services industries. Statistical Society of Canada Annual Meeting, St John’s.

Brodeur, M, & Ravindra, D (2010, December). Statistics Canada’s new use of administrative data for survey replacement. Conference on Administrative Simplification in Official Statistics, Ghent.

McKenzie, R (2008). Statistical architecture. Statistics New Zealand, Christchurch (internal paper).

McKenzie, R (2009). Managing the quality of administrative data in the production of economic statistics. International Statistical Institute, Durban.

Orchard, C, Moore, K, & Langford, A (2010). Practices for using VAT turnover data within the UK to produce estimates of growth and monthly turnover. Office for National Statistics, Newport.

Statistics New Zealand (2010). Statistics New Zealand Strategic Plan 2010-20. Wellington.

Statistics New Zealand (2011). Statistics New Zealand Collection Strategy 2011. Wellington.

Statistics New Zealand (2014). Christchurch retail trade indicator (information releases). Wellington.

Stewart, J, Costa, V, Page, M, & Chen, C (2011). Towards an architecture for sub-annual business collections. Statistics New Zealand, Christchurch (internal paper).

Stewart, J, Costa, V, Page, M, & Chen, C (2012, June). Maximising the use of administrative data in sub-annual business collections. International Conference on Establishment Surveys, Montreal.

Yung, W, & Lys, P (2008). Use of administrative data in Statistics Canada’s business surveys – the way forward. International Association for Official Statistics Conference, Shanghai.