



## Use of administrative data in sub-annual statistics – how far can we go?

Paper presented at the OECD Short-term Economic Statistics Expert Group (STESEG),  
Paris, France  
10-11 September 2009

Kathy Connolly

Manager Business Indicators, Statistics New Zealand

[www.stats.govt.nz](http://www.stats.govt.nz)

**Liability statement:** Statistics New Zealand gives no warranty that the information or data supplied in this paper is error free. All care and diligence has been used, however, in processing, analysing and extracting information. Statistics New Zealand will not be liable for any loss or damage suffered by customers consequent upon the use directly, or indirectly, of the information in this paper.

**Reproduction of material:** Any table or other material published in this paper may be reproduced and published without further licence, provided that it does not purport to be published under government authority and that acknowledgement is made of this source.

## ***Summary***

Statistics New Zealand's Economic Statistical Architecture vision has been in existence for a number of years now. Since its inception the strategy has been refined but by and large the core concepts have remained the same - creating a database of business information that covers the whole economy and enables both the production of macro-economic statistics and micro level research. In creating this database, the fundamental premise is that we will move from surveying and supplementing with tax data to using administrative data sources for the majority of businesses with supplemental surveying for complex businesses or variables that aren't covered and can't be modelled well using administrative data alone.

This will result in a reduction in respondent load, gains in efficiency, a substantive increase in the data available to users (for example more services statistics) and an improvement in the quality of existing statistics (for example the production measure of GDP).

We have begun to implement the Statistical Architecture in our Business Statistics with the development of a new flexible IT system and methods for annual tax data and increasing the use of annual business tax returns in our Annual Enterprise Survey. In the second stage we are focussing on sub-annual financial statistics. This work poses many challenges but provides an exciting opportunity to make some quite radical changes in the way that we produce our information.

## ***Background – Statistics New Zealand's Statistical Architecture***

A few years ago, Statistics New Zealand developed a 'Statistical Architecture' for business data. This has become the vision that has shaped the development of our statistical collections. The ultimate goal is a fully-integrated set of statistics that will support a broad and growing range of user needs.

In essence, the Statistical Architecture recognised that we needed to move from thinking like engineers to thinking more like an architect when it came to redesigning our collections. In much of our work, Statistics New Zealand has tended to take an engineering approach, using the best available techniques to build an existing design. An architectural approach goes a step further and thinks about how emerging user needs should be met. This approach uses best practice techniques and materials to produce a design that will deliver all the new features that users require. The purpose of a statistical architecture is to describe an integrated and systematic approach to data collection that will support current and future information needs.

At Statistics NZ we already have a well integrated system of economic statistics. Over the last decade we have developed an efficient and effective system for collecting a broad range of economic information using a combination of administrative data and sample surveys.

The foundation of this collection system is a comprehensive business register called the Business Frame (BF) that provides a single list of businesses and organisations of interest to Statistics NZ. A comprehensive business register has several benefits for the production of economic statistics.

- The Business Frame provides a common reference point for applying standard classifications for all units. This facilitates the integration of statistical outputs by ensuring that classifications are applied consistently across all surveys and statistical outputs.

- The Business Frame links all economic and financial survey data to the tax system, enabling more effective use of tax data to reduce respondent load.
- All administrative datasets are incorporated to our statistics by first being matched to the Business Frame using unique Tax numbers. This eliminates problems with duplications and inconsistent coverage of administrative datasets.
- The populations for all economic surveys are selected from the BF. This ensures coherence of information between different surveys and administrative data sources. Coverage adjustments are unnecessary, because we always know which units are covered by each data source. Where a unit is included in two different data sources, it will be removed from whichever is appropriate to ensure that coverage is coherent.
- Administrative data and survey data can be combined in a statistical output with the Business Frame ensuring coherence between data sources. For example, the frame can be partitioned with tax data being used for one partition and survey data being used for the rest.

A survey-based strategy has served us well, but there is pressure to keep respondent load and collection costs to a minimum. This approach is unable to support some of the emerging needs for statistical analysis - for example services statistics, longitudinal studies of small business, Maori business statistics, and regional data. There is also an expectation that new statistical information will be able to be delivered more quickly.

To respond to longstanding and emerging needs whilst minimising respondent load, Statistics NZ has embarked on a strategy that places a greater use of administrative data, with sample surveys being used to fill the information gaps.

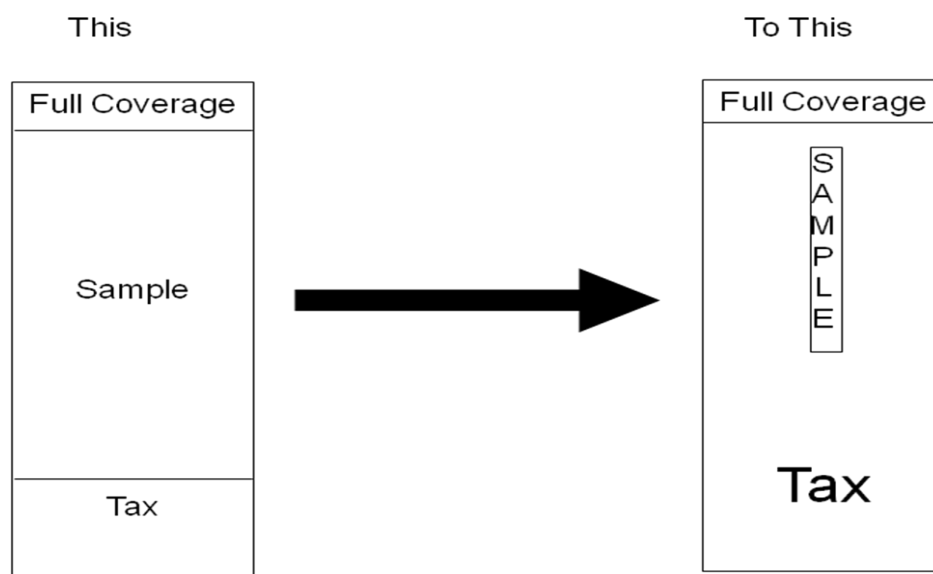
The basis of the future statistical system will be an integrated set of core information about businesses. New and existing collections will add information to this core infrastructure in a way that increases the power of statistical analysis. An optimal mix of survey and administrative data should reduce the volume of information that needs to be collected from businesses, while increasing the range and usefulness of the information produced. Statistical outputs will not be based on a single survey, but will be compiled by combining data from several different sources.

The statistical architecture is shaped by the following general design rules.

- i. Information can only be collected if a clear user need has been established.
- ii. Information should only be collected once.
- iii. Administrative data will eventually be used as the primary source of data.
- iv. Surveys will only be used to fill the gaps that cannot be met from administrative sources.
- v. Survey and administrative data will be integrated using a comprehensive business register.
- vi. Information should only be collected from units that can readily provide reliable and meaningful estimates.
- vii. Large complex business units will be closely managed to facilitate the collection of all the data that is needed from them.
- viii. Information quality will continue to be fit for purpose.

- ix. Reliance on administrative data will increase in incremental steps, beginning with the parts of the population for which tax data is robust and then expanding into more difficult areas as data issues are resolved.

Using administrative data first with surveys filling the gaps is a reversal of the current strategy of using surveys for medium and large businesses information with administrative data being used for small businesses where the contribution is less significant.



### ***Current use of administrative data***

We use a range of administrative data at Statistics NZ. In some areas we make full use of data collected by other organisations, examples include;

- Building consents – supplied by local authorities, edited and published monthly as an indicator of future building activity and also used as a frame for a Building Activity Survey
- Merchandise Trade – supplied by Customs, edited and released monthly
- Electronic Card Transactions – supplied by private organisations processing the data. This data is not edited, but compiled and briefly analysed before being published 7 working days after the reference month.
- a longitudinal series of payroll data linked to a longitudinal employer series from the Statistics NZ Business Frame to produce Linked Employer-Employee Data (LEED) which measures labour market dynamics at various levels
- a Longitudinal Business Database (LBD) links business-related data from both administrative and sample survey data to form a micro-data research tool which is used by many policy agencies

In other instances we make only partial use of the administrative data available to us. For example, use of tax data supplied by the Inland Revenue department at this stage is in most cases limited to replacing data for small businesses. It is here that we see a lot of potential in increasing the use of tax data and implementing the Statistical Architecture.

In New Zealand Goods and Services Tax (GST) is a value added tax levied on most goods and services. GST is currently used in the following collections:

- Monthly and quarterly Retail Trade (sales monthly, inventories quarterly) for small businesses for up to 10% of the total value
- Quarterly Manufacturing (purchases, sales, inventories, salaries & wages) for small businesses for up to 15% of the total value
- Quarterly Wholesale Trade (sales and inventories) for small businesses for up to 15% of the total value

Other uses of tax data include;

- use of an annual tax return data to supplement direct surveying in the Annual Enterprise Survey
- updating our Business Frame – tax data is used to identify new businesses, and the deaths of existing businesses and to update classifications for businesses below the predetermined size threshold (largely small and mediums size businesses out of 450,000 businesses 90% are updated using information from tax data)
- a Business Activity Indicator (BAI) - this is a quarterly indicator of economic activity based on GST data

### ***Focus on existing use of GST data in the sub-annuals***

A decade of using tax data to supplement surveying in our sub-annuals has given us confidence that we want to use it far more than we currently do. Experience has given us good insight into the strengths and weaknesses of the data and we see opportunities to improve the quality of it so that we can maximise its use.

Our existing use of GST data in sub-annual Retail, Manufacturing and Wholesale collections makes use of the Business Activity Indicator (BAI) which was a series first released in October 1998. The original idea was that a quarterly indicator of activity across all the industries in the economy could be produced from the GST data. There were a number of challenges that had to be overcome:

- differences between the tax unit and the statistical unit; in this case where there was not a one to one match, tax data was apportioned using sales from annual tax returns and employee count
- standardising the reference period to monthly; Enterprises submit GST returns either monthly, two-monthly or six-monthly. The two-monthly and six-monthly data are apportioned either using seasonal factors from their industry or by simply dividing them equally
- variable coverage not meeting statistical needs: specifically a) inclusion of capital purchases and sales; in theory these should be removed however only large values are removed, this is done manually; b) no inventories were collected

- missing returns; GST data is filed monthly, two-monthly or six-monthly, depending on level of turnover. For two and six-monthly filers that have not filed in the most recent months and for non-response, an estimate is made using either historical or mean imputation.

The usefulness of the BAI as an indicator of economic activity was significantly reduced by its lack of timeliness and being a poor predictor of GDP in some cases. The BAI did, however, prove to be a useful resource for researchers and very importantly as a replacement for survey data for small businesses in sub-annual financial surveys. Statistics NZ's current use of the BAI in our sub-annual surveys is to fully replace responses for small businesses – up to a maximum of 15% by value of the industry.

Timeliness is probably the largest single issue for use of GST data. Tax filers are given until the end of the month following the reference period to file their return and we receive it around 8 weeks after the reference period. This means that for the quarterly manufacturing and wholesale surveys there are only 2 months of 'real' data. For the monthly Retail Trade series all the data is modelled (based on historical returns). The series are not revised when actual data is received as this would have very little impact on the results. In addition, variables such as inventories are modelled using postal responses.

The BAI has served very well as a survey replacement for small businesses data. It was developed as an experimental series over a decade ago with a small budget, a tight timeframe and a different purpose in mind. Despite this, many of the methods used were 'best-practice' at the time and it is still a good collection vehicle for small businesses saving them the cost of having to complete questionnaires. Now we want to significantly increase our use of GST data for our sub-annual collections and this requires a fresh look at the data and a change of approach.

### ***Increasing the use of administrative data- work to date***

To date, a number of projects have progressed the implementation of the statistical architecture. These include the development of a linked employer-employee database, a research longitudinal business database and most recently a redesign of the Annual Enterprise Survey (AES).

The AES redesign developed an end-to-end edit and imputation system to process large scale administrative data volumes. The IT solution is scalable, uses standard tools (BANFF and SAS), is flexible for user-needs and facilitates easy data investigations and what-if analysis. The initial AES sample redesign has made conservative use of the clean tax data, but investigations to date indicate that much more use can be made. Over the next few years we intend to significantly increase our use of tax data in AES.

The IT infrastructure and methods developed for AES will now be used to manage GST data and in replacing the current BAI.

### ***Using more GST data***

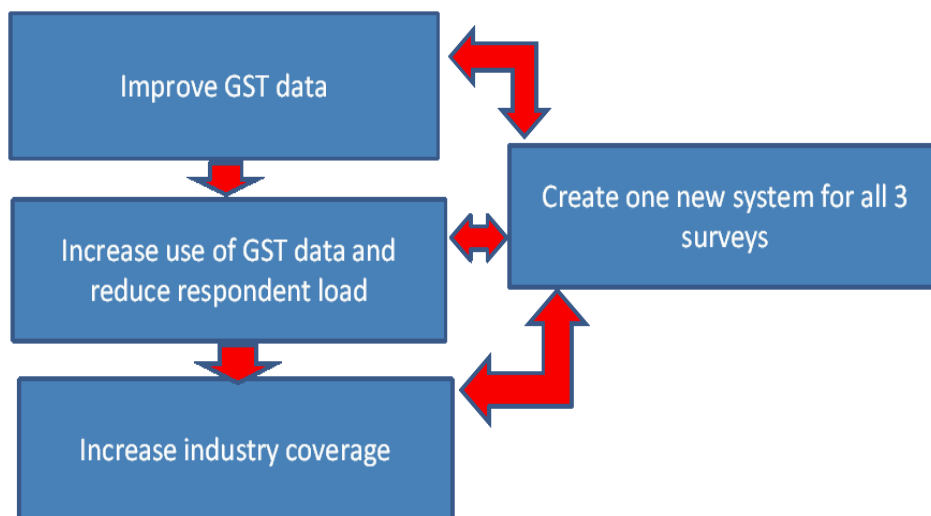
With success in implementing the statistical architecture in our annual business statistics, the logical next step was the sub-annuals. The motivators for change in the sub-annual financial statistics - currently comprising the Retail, Wholesale and Manufacturing surveys include:

- Lowering respondent load: businesses and government want to spend less time providing us with data and form-filling. We have full access to GST data and yet we are only using a fraction of what we could so we have a responsibility to our respondents to expand our use as effectively as we can while still producing fit-for-purpose statistics.

- Improving quality: We know that our users want better sub-annual measures of the economy - this includes wider coverage of the economy, particularly services, and more recently this means more timely data. These changes should improve efficiency of the process and provide the opportunity to expand coverage at a later date.
- Moving off legacy systems: Our existing systems are aging and siloed with little support available for the software they are written in.

So how do we achieve it? There are four key steps:

- Improve the quality of the GST data for use to replace surveying: There are known weaknesses with the current methods used in the BAI which need addressing as we look at using GST in a different way
- Increase the use of GST data and reduce respondent load: extending the use of GST past the current low levels to deliver savings in respondent load and collection costs
- Create a new system that can be used for all surveys: Having a single system that replaces existing siloed systems will lower the current risk to Statistics New Zealand. The system will be easier to maintain, reduce training costs, and an expanded pool of expertise with more people using the same system
- Increase the coverage of the services industries: we currently only produce manufacturing, retail and wholesale series sub-annually, there is a high demand for better coverage of the services industries



### ***Stage one of the Sub-annuals redesign***

Following on from the successful annuals project where a new system was designed; bringing GST data into the new system is a logical first step. From a development perspective, the new system provides flexibility which is very useful for trialling different options for edit, imputation and modelling methods to solve the GST challenges noted above.

Although this system currently only allows for the edit and imputation and modelling of tax data and not for any survey data editing or other survey requirements, this is still a good starting point in terms of replacing our existing aged systems.

Once the GST data is in the new system we will be working through our options for edits, imputation and modelling to improve the quality of the GST data available for use in the collections and for other uses, including the Business Frame and longitudinal business database. In the past we have taken an all or nothing approach to using GST data, attempting to meld it to our statistical view of the world. For many large or complex businesses this required applying methods to apportion data across a number of statistical units, the result was not ideal.

What we propose now is to look at the GST data to see where there is a good match between the tax unit and the statistical unit. We know that we can do better at removing the capital component of sales and purchases with a better outlier methodology. We are still at the thinking stage in terms of methods for producing industry based estimates. It's likely that we will produce unit record data to be used as this retains flexibility and will allow incremental increases in the use of administrative data over time – an approach that is more palatable for users. We will also look at the option of modelling to produce industry estimates.

Ideally, we want to be able to have a number of versions of GST data for example– 'raw', removal of capital expenditure, standardising reporting period, so that we can pick and choose which version we want for which particular use.

In this stage of the project we need to confront the challenges that we are aware of and those that will invariably arise when we do more investigation. We are confident that there will be a significant part of the GST data set that will be fit for use. Our challenge is how to maximise its potential.

### ***Future stages***

Once we have good quality GST data, we can then progress work on increasing the use of it in our collections. This will require investigation of the modelling methods to account for timing issues and non-response.

Timeliness of GST is an issue. For the foreseeable future we are unlikely to get GST data earlier than we currently do. Currently we use 2 months of actual responses and forecast the last month in our quarterly estimates. Because the level of GST use is low (currently a maximum of 15% by value of GST data to replace surveying small businesses), this is not a problem. Before we can significantly extend the use of GST data we need a method of forecasting the latest month that doesn't result in major quality concerns. The existing method may be good enough but this needs to be tested. There are also options including publishing a provisional series which is revised when more data is received.

A particular challenge is creating a quality measure for administrative data, as current measures (eg sample error) are no longer appropriate. We need to develop new methods that reflect quality for the new collection approach.

The final phase is expanding coverage. Today, more than ever, data on the rest of economy - particularly services - is sorely missed. If we can use GST extensively and have an easily extensible system, then expanding coverage is made far simpler and can be accomplished without a significant increase in respondent load.

This project has a number of challenges. However we are confident that we can significantly increase our use of administrative data in our sub-annual surveys, delivering respite for respondents, the possibility of lower collection costs and the potential to be able to expand the industry coverage with significantly less surveying necessary than if we were to use our existing methods. This project will also provide a better understanding of the quality of the series and ensure fitness for purpose.