A Rough Guide to New Zealand's Longitudinal Business Database (2nd edition)

Richard Fabling and Lynda Sanderson

Motu Working Paper 16-03
Motu Economic and Public Policy Research

February 2016

Author contact details

Richard Fabling
Independent Researcher
richard.fabling@xtra.co.nz

Lynda Sanderson
The Treasury
lynda.sanderson@treasury.govt.nz

Acknowledgements

The authors would like to thank the many staff at Statistics New Zealand and Inland Revenue who fact-checked the metadata, and Dave Maré and Michele Morris for helpful advice and suggestions on the paper as a whole. Richard Fabling gratefully acknowledges funding from Statistics New Zealand.

Motu Economic and Public Policy Research

PO Box 24390
Wellington
New Zealand

Email	info@motu.org.nz
Telephone	+64 4 9394250
Website	www.motu.org.nz

Abstract

New Zealand's Longitudinal Business Database is a rich resource for understanding the behaviour of New Zealand firms. This paper provides an introductory guide to the content and structure of the data aimed at new and prospective users. Where relevant, it references other publications which provide greater detail on particular aspects of the data. It also briefly describes access protocols for researchers, and processes for updating and expanding the database.

JEL codes

C80; D22

Keywords

Firm-level data; New Zealand; Integrated Data Infrastructure (IDI); Longitudinal Business Database (LBD)

Summary Haiku

The LBD shines
But how to harness the light?
Answers lie within

Disclaimer

The results in this paper are not official statistics, they have been created for research purposes from the Integrated Data Infrastructure (IDI) managed by Statistics New Zealand.

The opinions, findings, recommendations and conclusions expressed in this paper are those of the authors not Motu Economic and Public Policy Research, Statistics NZ or the Treasury.

Access to the anonymised data used in this study was provided by Statistics NZ in accordance with security and confidentiality provisions of the Statistics Act 1975. Only people authorised by the Statistics Act 1975 are allowed to see data about a particular person, household, business or organisation and the results in this paper have been confidentialised to protect these groups from identification.

Careful consideration has been given to the privacy, security and confidentiality issues associated with using administrative and survey data in the IDI. Further detail can be found in the privacy impact assessment for the Integrated Data Infrastructure available from www.stats.govt.nz.

The results are based in part on tax data supplied by Inland Revenue to Statistics NZ under the Tax Administration Act 1994. This tax data must be used only for statistical purposes, and no individual information may be published or disclosed in any other form, or provided to Inland Revenue for administrative or regulatory purposes.

Any person who has had access to the unit-record data has certified that they have been shown, have read, and have understood section 81 of the Tax Administration Act 1994, which relates to secrecy. Any discussion of data limitations or weaknesses is in the context of using the IDI for statistical purposes, and is not related to the data's ability to support Inland Revenue's core operational requirements.

Statistics NZ confidentiality protocols were applied to the data sourced from the New Zealand Customs Service, Ministry of Social Development, the Ministry of Business, Innovation and Employment, New Zealand Trade and Enterprise and Te Puni Kokiri. Any discussion of data limitations is not related to the data's ability to support these government agencies' core operational requirements.

Contents

1 Introduction

The Longitudinal Business Database (LBD), maintained by Statistics NZ, is a rich resource for understanding the behaviour and performance of New Zealand firms. The primary selling point of the database is its multi-dimensional nature. By bringing together a range of survey and administrative data sources, the LBD provides over a decade of performance and employment data for the majority of firms operating in New Zealand, together with a detailed view of firm practices across a range of different topics for representative subsets of those firms. In addition, the LBD is linked to the Integrated Data Infrastructure (IDI), which contains individual-level data including information from Inland Revenue (IR), the Ministry of Health, the Ministry of Education, and the Department of Corrections, among others. This additional linking has dramatically expanding the range of research questions answerable using the LBD.

This paper provides an update and extension of the first edition "Rough Guide" (Fabling, 2009). It is intended as an overview of the LBD as a whole, highlighting key elements of the available metadata to supplement collection-specific metadata provided on Statistics NZ's website and in the shared metadata folder available to Datalab users.[1] In addition to describing the structure of the database, this update provides a more detailed description of the content, including recent additions, as an aid to new and prospective users. Section 2 outlines the components of the LBD, providing both a dataset-by-dataset view and spotlights on particular topics of research interest. Where information is already available in the public domain, a brief summary is provided alongside links to the more detailed documents. As the feasibility of many potential research questions is determined by sample size and the joint availability of data across multiple datasets, section 3 describes the pairwise overlap in coverage across selected datasets for different firm-size groups. Section 4 covers access and the process for adding data, and section 5 concludes.

References to the data are based on the December 2014 archive – the most recent complete version of the LBD at the time of writing. A number of changes have been made to the database in the 2015 update, including the removal of aggregated Linked Employer-Employee Data (LEED) variables from the LBD and direct provision (within the IDI) of the individual-level raw tax data previously used to generate those variables. This paper outlines the changes to the LBD which have occurred due to the redevelopment of LEED, but does not otherwise discuss the individual-level data available through the IDI. Fabling & Maré (2015a) provides a detailed discussion of labour measurement using linked LBD/IDI data, and Statistics New Zealand (2013) gives an introduction to the individual-level data in the IDI.

2 Structure of the LBD

The LBD comprises tax- and survey-based financial data, employment data, merchandise and services trade data, sample surveys covering a variety of business practices and outcomes, intellectual property data, and government programme participation lists. These data sources are linked together through the Longitudinal Business Frame (LBF) – a

[1] Metadata on the Web is located at `www.stats.govt.nz` and `http://datainfoplus.stats.govt.nz`; and in the Datalab at *\\wprdfs08\gendata\IBULDD Collaboration\Metadata*.

register of all economically significant businesses in New Zealand. The LBD in turn is linked to the IDI through employer and employee tax identifiers mapped to Statistics NZ firm identifiers (enterprise numbers) and individual unique ids ("*snz_uid*") respectively. The overall structure of the LBD and links to the IDI are depicted in figure 1.

Statistics NZ updates the LBD data annually, and maintains archived versions of each previous iteration of the database.[2] The current, live version of the data is called *ibuldd_clean*, a reference to the original project under which the database was developed – "Improved Business Understanding via Longitudinal Database Development", with earlier archives referred to by the date at which the archive was created. Since becoming available to researchers in April 2007, the database has expanded organically to meet the needs of agencies whose researchers use the data, and now includes a substantial proportion of the business data that Statistics NZ holds and uses to compile official statistics. Table 1 provides a summary of the component datasets of the LBD.

To improve accessibility, each constituent dataset has a "fact table" in which variables are aggregated to a consistent unit of observation and periodicity. Most financial variables are only observed at the enterprise (or tax reporting) level, not at the individual plant, so the enterprise (or "firm") is the common unit of these tables. An annual frequency is imposed on the data by tax-filed financial accounts, working proprietor income declarations, and most sample surveys. Hence, all sub-annual data (goods and services tax, trade, international investment, and employment) are annualised to each firm's financial year, allocated to the 31st March year-end that has the greatest overlap with the financial year.[3] This notional year is dubbed the *dim_year_key* and denoted by YYYY03.

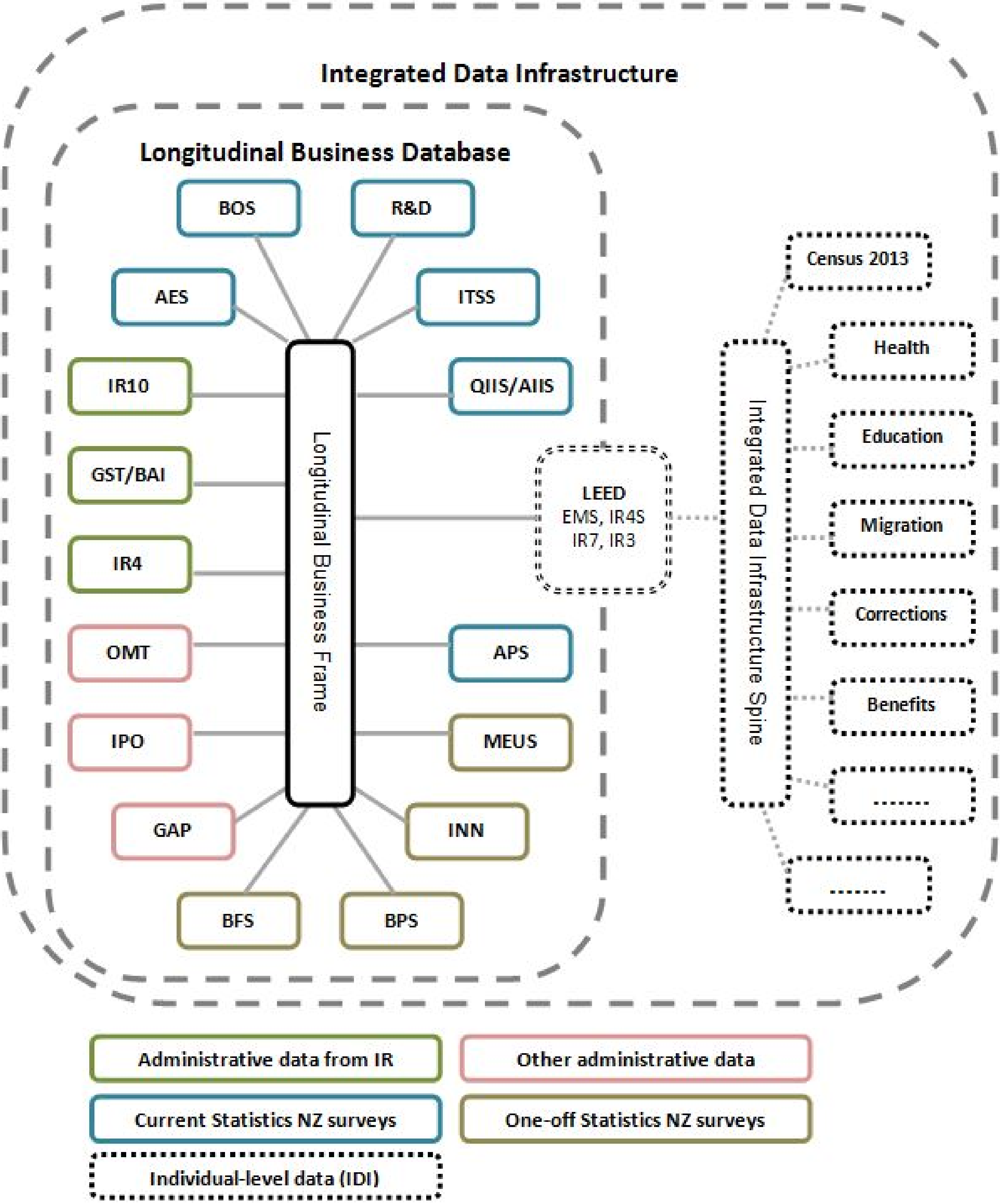
Aside from the construction of fact tables, Statistics NZ applies a light touch to the data – the main technologies added specifically for the benefit of LBD users being editing of tax-filed financial information, and probabilistic matching methods for linking administrative data to enterprise numbers.[4] Where data is transformed, the database usually includes the raw ("load") data so that researchers can choose to undo any processing that has occurred. For example, while Statistics NZ construct an overseas merchandise trade fact table that includes the aggregate annual exports of each firm, the database also includes the underlying daily shipment level data.[5] In addition to the data tables, the LBD also includes a number of reference tables, which provide information about the various classification systems used in the data, and functional tables and procedures to simplify data manipulation.

[2] The update process occurs over a period of time, with component datasets being added to the live version of the database (ibuldd_clean) as they become available for updating. A number of the datasets are subject to historical revisions because of, eg, late filed tax returns, tax id-only firms being subsequently included on the Business Frame (hitting the materiality threshold), changes to link tables or firm-level balance dates. Occasionally data will need to be reloaded into the live version of the database. To ensure they are working with a consistent version of the data over the life of a project, and to enable replication and journal revisions, researchers may prefer to use an archived version unless they need the most recent year of a particular dataset.

[3] Tax filing rules mean that most firms actually have a 31st March balance date. Variations from that rule are generally industry-specific, eg, the public sector operates to an end of June financial year.

[4] Statistics NZ also anonymise the data in accordance with the Statistics Act 1975.

[5] In general, these load tables provide a direct copy of the underlying survey data. However, there are instances where survey responses have been edited prior to inclusion in the LBD, and some cases where not all survey responses are available (eg, the AES survey dataset does not include all the raw responses from the survey form, such as opening book values for fixed assets).


Integrated Data Infrastructure
Longitudinal Business Database
BOS
R&D
AES
ITSS
IR10
QIIS/AIIS
GST/BAI
IR4
OMT
IPO
GAP
BFS
BPS
INN
MEUS
APS
Longitudinal Business Frame
LEED
EMS, IR4S
IR7, IR3
Integrated Data Infrastructure Spine
Census 2013
Health
Education
Migration
Corrections
Benefits
....
....
Administrative data from IR
Other administrative data
Current Statistics NZ surveys
One-off Statistics NZ surveys
Individual-level data (IDI)

IDI tables are provided as an illustration and do not represent the complete content of the IDI.

Alongside the LBD database (*ibuldd_clean*) sits the *ibuldd_research_datalab* database. This is a repository in which researchers can save tables that are required on a semi-permanent basis or need to be accessible by multiple users. The derived productivity dataset (Fabling & Maré, 2015b) is a key example of user-derived data located in *ibuldd_research_datalab.*

Subsequent subsections briefly outline each of the data sources currently linked into the LBD, covering the purpose of the collection, an outline of the content, and the relevant population and sampling methodology. For each data source we also provide a list of "tips and tricks for new users" and, where available, point to methodologies that have been developed by the researcher community to improve the utility of the data.

2.1 Longitudinal Business Frame

Purpose

The LBF is a longitudinal representation of the Business Frame (BF) – the primary sampling frame used by Statistics NZ until April 2014. The LBF unwinds the historical information available in the BF to create a longitudinal (panel) dataset of business (and plant) characteristics. The primary purpose of the Business Frame was to maintain an accurate point-in-time representation of all firms in the economy, from which relevant survey populations could be identified. It was also used directly for the construction of business demography statistics. This frame forms the backbone of the LBD, to which all other firm-level datasets can be linked.

From May 2014, the BF has been replaced by the Business Register (BR). The new BR is designed explicitly to incorporate the longitudinal aspect necessary for the LBD (and other uses). Changes have also been made to data capture and update processes as well as the database management technology. As at the time of writing, Statistics NZ is redeveloping the spine of the LBD to reflect the retirement of the BF and concurrent move to the new BR. From a practical perspective, this move should have little impact on users since the LBF used in the LBD is already being populated from the BR.[6]

Content

The LBF captures basic information on all economically significant firms operating in New Zealand since 1999,[7] including information on location, industry, business type (eg,

[6] At present, the BF is being replicated from the BR as an intermediate step prior to construction of the LBF. The BF will continue to be maintained in this way until all BF users have migrated to using the BR directly.

[7] Enterprises are designated as economically significant if they meet any of the following conditions: (1) >$30,000 annual GST expenses or sales (to avoid existing enterprises repeatedly changing their economic significance indicator, a buffer zone of $25,000 to $35,000 has been established); (2) more than 3 paid employees (BF rolling mean employment variable); (3) in a GST exempt industry, other than residential property leasing and rental (affected units are primarily in the finance and insurance industry, and property operators and developers); (4) part of a BR group; (5a) a new GST registration and has registered for salaries and wages PAYE but has not yet started filing GST returns. These enterprises then have a 12 month window as live units (with economic significance as yes) before the other economic significance criteria are applied; (5b) a new GST registration and part of an IRD GST group return;

limited liability company, partnership), institutional sector (eg, private corporate producer enterprise, central government), and parent-subsidiary relationships. The LBF also includes an annual employment measure, used by Statistics NZ for identifying the appropriate sample for various business surveys (*LBF_rme_nbr*).[8]

In addition to these core variables, the LBF also contains a range of information collected for the purposes of identifying Balance of Payments survey populations. While the specific questions asked have changed over time, these have included indicators of substantial overseas assets, overseas revenue from or expenditure on services, an indicator of foreign subsidiary ownership (*bop_ownership_ind*) and the proportion of foreign ownership (*bop_ownership_rate*).[9]

The LBF identifies a number of different statistical units and structures, including information on the internal structure of firms as well as links between firms through parent-subsidiary relationships. This information is available on a monthly basis through the *load_lbf_fact_business* table.[10] Figure 2 provides a stylised diagram of possible firm structures:

- Enterprises (ENTs) generally correspond to businesses (firms). They are separately tax-reporting legal entities such as companies, incorporated clubs and societies, state owned enterprises and statutory bodies, central and local government bodies, and other entities such as sole proprietors, partnerships, and trusts.
- Geographic Units (GEOs) or "plants" can be a whole ENT or part of an ENT and are normally a physical location from which predominantly one kind of activity takes place on a permanent basis. Location and employment data are available at the GEO level in *load_lbf_fact_business* and *load_lbf_fact_pbn_employee_count* respectively.[11] Statistics NZ also creates Permanent Business Numbers (PBNs), which identify continuing activity at the level of the geographic unit, repairing breaks in GEO numbers.
- Kind of Activity Units (KAUs) are units which engage in one predominant economic activity. KAUs are not restricting to being located in a single geographic area. Thus, a KAU may have several constituent GEOs. KAU-level information is limited and is not reported in the LBF, though KAU-level datasets always identify the relevant ENT. Currently, the Annual Enterprise Survey is the only KAU-level collection held in the LBD.
- Direct majority ownership links between enterprises are identified through the *parent_enterprise_number* and *gte_enterprise_nbr*, where the former refers to the immediate parent company and the latter refers to the Group Top Enterprise – the highest enterprise in the parent-subsidiary chain.

(6) has a live geographic unit classified to agriculture (typically these units will be registered for GST and/or have paid employees.) (7) IR10 income >$40K.

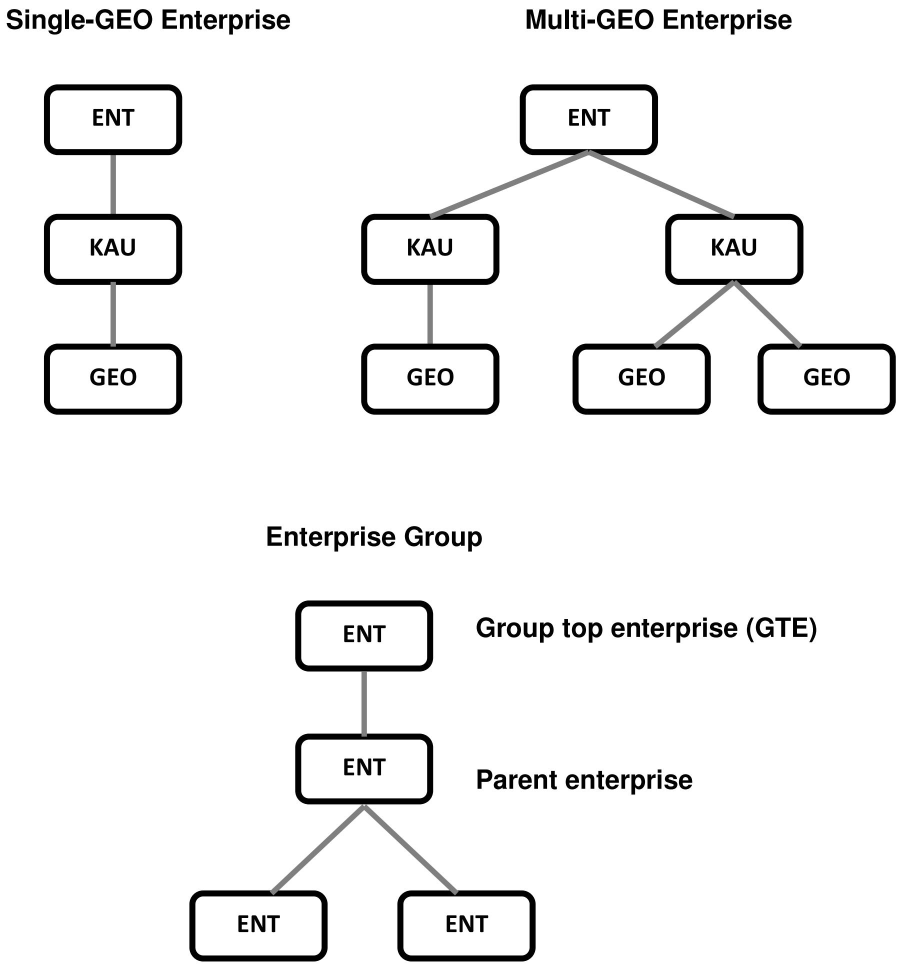
[8] Prior to 2015, the LBF tables in the LBD also included an annual RME count which excluded working proprietors (*leed_rme_at_15th_no_wp_nbr*). This variable has now been discontinued in favour of providing access to the raw employment data through the IDI.

[9] This latter variable has been collected consistently over the period of the LBD. See Sanderson (2013) for more detailed information on foreign ownership indicators over time.

[10] This table records changes of state at the plant level (eg, changes in ownership or industry code) according to the start and end month of each observed state.

[11] Allocation of employees to plants is performed by Statistics NZ (see section 2.5).

Figure 2: Statistical unit structures



The load tables of the LBF provide information on firms' geographic locations at various levels of detail, from Regional Councils to meshblocks (variables such as *geo_meshblock_code* in the *load_lbf_fact_business* table). These can be linked to PBN-level employment counts (*load_lbf_fact_pbn_employee_count*) and geographic information such as x and y coordinates of meshblock centroids (through *ref_meshblock*).[12] Since meshblock boundaries are adjusted annually, a pairwise concordance tables is provided (*ref_meshblock_pair*) to enable users to link geographic information from other sources.

Coverage

LBF data is compiled from a combination of survey responses and administrative data sources. Over the period covered by the LBD, there has been a substantial decline in the use of direct surveying to maintain and update the BF/BR, as part of an effort by Statistics NZ to reduce respondent burden and data collection costs. Surveys have been replaced by increased use of administrative data sources (in particular IR's Client Registration data and information on company registration from the Companies Office) both to populate BF/BR variables at the time of firm birth and to update variables through the life of the firm.

These changes have been implemented through adjustments to the criteria under which firms receive frame/register update surveys. Under BF updating processes since 2004, large and complex (Tier 1) firms were surveyed at birth and received an update survey annually. Medium sized firms (Tier 2) were surveyed at birth and resurveyed at least once every three years.[13][14] Smaller firms (Tier 3) were surveyed at birth only if they registered for both GST and PAYE (ie, they were above the mandatory GST filing threshold or voluntarily registered for GST, and had at least one employee), and were not re-surveyed unless they reached the threshold for Tier 2. Substantial reductions in the annual survey sample over time have been implemented through adjustments to the tier boundaries as well as a range of other sample adjustments.[15]

While BR updating follows a similar system, with resources directed towards maintaining up-to-date information for those firms which make the greatest contribution to national aggregates, the new system makes even greater use of administrative data to populate the information required for new firms, and to update information on incumbents. The BR population is based on the IR Client Register (CR), and covers almost all CR units with observed business tax activity, as well as Clubs and Societies listed as "live" and new CR units which are registered for GST and PAYE but have yet to submit a tax return.[16,17]

The frequency with which firms receive the Business Register Update Survey (BRUS) is

[12] Meshblock coordinates are recorded following the New Zealand Map Grid (NZMG) system.

[13] Update surveys may also have been triggered if the business showed a significant change in tax flows.

[14] Firms in the agriculture and horticulture industries are largely excluded from receiving frame update surveys, as the Agricultural Production Survey (APS) is used to update information on these industries. Similarly, state and integrated schools are excluded as these enterprises are updated using information from the Quarterly Economic Survey (QES).

[15] Sanderson (2013) provides additional information about tier definitions and AFUS sampling for the BF.

[16] Many non-profit organisations are not required to file an annual tax return but will still be recorded as live in the CR. Trusts with no business income are classified to the household sector.

[17] These units are removed from the BR population if they do not show any tax activity within 12 months of registration.

again based on tiers:

- Tier 1 covers "large" or "complex" enterprises, meeting any of the following conditions:
  - more than three geographic units
  - geographic units located in more than one Territorial Authority
  - annual GST sales or purchases ≥$65 million
  - rolling mean employment (RME) >20[18]
  - total income on IR10 ≥$65 million

- Tier 2 covers other enterprise units which are in ANZSIC06 industries deemed more likely to have either BoP-related transactions or to be performing R&D (see table 2), and which meet any of the following conditions:
  - annual GST sales or purchases ≥$1 million
  - rolling mean employment ≥5
  - total income on IR10 ≥$1 million
  - multiple geographic units

- Tier 3 covers other enterprise units which meet the economic significance threshold:
  - annual GST sales or purchases ≥$30,000 [19] or GST sales or purchases ≥$1 if in agriculture
  - total income on IR10 ≥$40,000

- Tier 4 refers to units below the BR economic significance threshold.

In general, tier 1 firms are still surveyed annually and tier 2 firms are surveyed every 3 years. As well as the annual update survey, Tier 1 firms are subject to a range of alternative update methods, including manual updates by Business Register operators on the basis of external information (eg, company reports) and feedback from other Statistics NZ surveys. While tier 2 firms are in principle subject to these same forms of update (in addition to three-yearly surveys), their application is more limited. Tier 3 and 4 firms are maintained purely through automatic updates from administrative data sources.

Monthly birth surveys (MRUS) have also been scaled back substantially in the BR. The number of firms surveyed has dropped from around 1,000 to around 50 per month. At the same time, the paper survey has been replaced by a seven question telephone survey and a four-month delay has been introduced between the appearance of a new firm on IR's CR and the birth survey to enable the collection of relevant administrative data. Table 3 documents the decrease in the number of frame/register update survey responses collected over time.

Updates are applied at the time when new information is made available to Statistics New Zealand, and, where possible, backdated to the point at which the change in status occurred. As many of the sources used to update the LBF are snapshots at a particular time, this "real world" date often reflects the timing of surveys and other data collections,

[18] Sales and employment thresholds incorporate a 10 and 20 percent buffer respectively to avoid firms transitioning into and out of the population due to small fluctuations in size.

[19] Although the definition of economic significance covers all firms that have annual turnover of at least $30,000, there may be some undercoverage of firms with turnover below the mandatory GST filing threshold, which increased from $30,000 to $60,000 between 1999 and 2011.

rather than the actual date of the event.

From a practical perspective, these maintenance rules imply that under both the BF and BR regimes some variables appearing “annually” on the LBF fact table in the LBD are actually observed (updated) less frequently for many firms. At the same time, variables that are not available through administrative sources (eg, Balance of Payments variables such as indicators of offshore financial assets or liabilities) will not be populated for most small and medium sized firms. The accuracy of certain variables may also depend on other characteristics of the firm. For example, enterprise group structures and foreign ownership information for limited liability companies in New Zealand are supported by administrative Companies Office data. In contrast, for other business types, the accuracy of group structures and ownership information relies on survey data and hence is limited for small, non-complex enterprises. Finally, Tier 1 may suffer from undercoverage because some of the identifying characteristics of this subpopulation are (largely) only discovered by surveying the business. In particular, firms that expand from single to multiple locations may not be identified as multi-location unless their expansion also coincides with reaching the sales and/or employment thresholds for Tier 1.

Figures 3-5 provide examples of the changing relative importance of different data sources over time for identifying key events in a firm’s lifecycle. Figures 3 and 4 relate to the entry and exit of firms and plants, respectively, while figure 5 covers changes in firms’ industry classification. Plant entry and exit figures are restricted to events occurring within continuing firms (approx 6 percent of all GEO entry and exit events) to isolate the updating process associated with multi-plant enterprises. In all cases, detailed data sources are aggregated into seven common categories, as shown in table 4, with the four most prevalent data sources for the event in question shown in the figures. Each figure also includes an index of the total number of recorded events (solid black line, right hand axis), with both the shares and the totals expressed as a rolling average over the previous 12 months.

Throughout the period covered by the LBD, enterprise entry and exit is predominantly identified through IR data (figure 3). Over 95 percent of firm entry events are identified through tax registrations, with only the 2002 APS census making a noticeable dent in the share sourced from IR data. In contrast, Statistics New Zealand frame/register update surveys (in the early half of the period) and Companies Office data (in the second half) also play a reasonably significant role for identifying firm exit, jointly accounting for between 15 and 20 percent of updates for most of the period. Measured firm entry and exit have both declined since 2008, though there is a substantial up-tick in the number of firm births since February 2013.

In contrast, the identification of plant entry and exit among continuing firms is much more heavily based on Statistics NZ survey data. These survey sources remain important despite the growing use of IR data over time. AFUS accounts for over 30 percent of observed entry events and forty percent of exits over most of the period, despite the decreasing number of respondents over time, while feedback from other Statistics NZ surveys (primarily the Retail Trade Survey, Quarterly Economic Survey and Group Profile Survey) together accounting for a further 10 to 20 percent in most years. A substantial proportion of plant entry and exit is also derived from other, unspecified data sources. The total number of plant entries and exits peaks somewhat earlier than firm entry and exit, and falls more precipitously after 2006.

The pattern of industry classification updates is dominated by the impact of survey feedback from the APS (figure 5), which account for as much as 65 percent of all changes in APS census years and up to 35 percent in survey years. In order to better identify the shift from update surveys to administrative data over time, panel B of figure 5 excludes APS-sourced changes. While the shift from survey to administrative data over the period is clear, this shift has been far from monotonic. The number of IR-sourced changes spiked in April 2002, and again in February 2004, when the IR10 was added as an additional source for identifying firm births.[20] This latter spike was also accompanied by an unusually high number of updates from other sources. Similar to entry and exit dynamics, the total number of recorded industry changes has fallen substantially since 2004.

It is not clear whether these observed dynamics are driven by real world changes, such as the effects of the Global Financial Crisis or a secular slowdown in the firm turnover rate, or by a reduction in Statistics New Zealand's ability to identify some life cycle events. In particular, while the total number of entry and exit events is likely to have been affected by cyclical and secular variation, it is not clear that a similar argument can be made for the reduction in industry shifts. Moreover, the relative "hit rate" of survey data in terms of the probability that a survey observation will lead to a change in recorded industry is much higher than that observed in the administrative data. This may reflect in part the selected nature of the sample (eg, multi-plant firms and agricultural firms may be more likely to undertake a range of activities at any given point in time, such that an industry change may reflect a relatively minor re-weighting of those activities), but may also reflect a relatively lower propensity to report changes in activity to IR, as industry information is not used to determine firms' tax liability.[21]

Tips and tricks

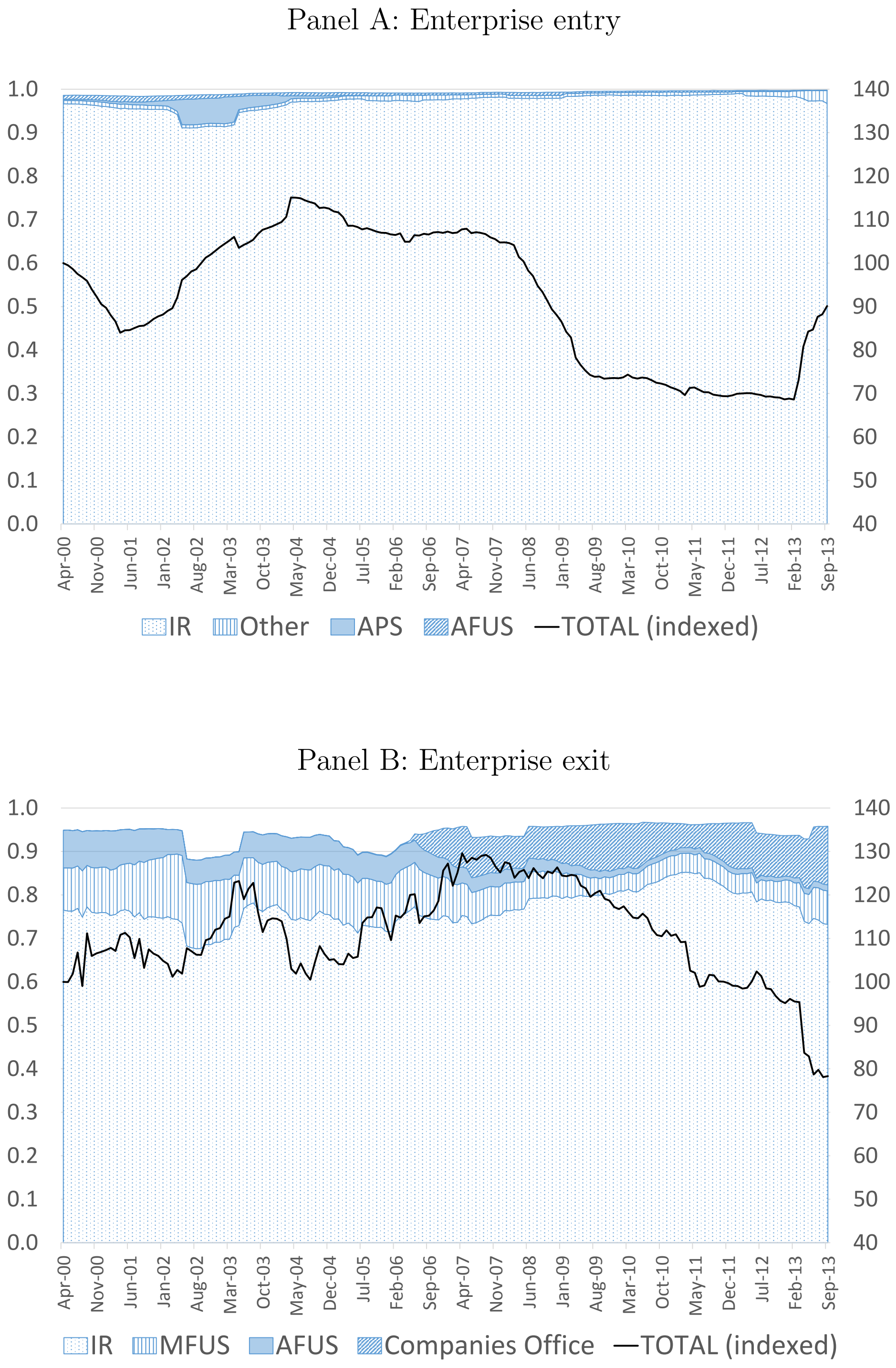
The *fact_LBF_enterprise_year* table is rectangular – that is, every enterprise on the LBF has an entry for every year that the LBF covers. Conversely, firms that have not been allocated an enterprise number by Statistics NZ do not appear on the LBF, even if they appear elsewhere in the data. Consistent with the principle of "keeping everything", administrative records are held on the database where there is filed data that cannot be linked to firm ids. Data that isn't associated with a true enterprise number are identified by creating an "enterprise number" prefix other than EN (eg, IR for tax-only units, CC for Customs clients, FI for firms in the Government Assistance Programme data).[22] These data may not be associated with firms – for example, individuals importing goods for personal use may be present in the Customs data – so that the rate of non-linking to the LBF does not necessarily reflect failure in the matching technology. The economic significance threshold on the LBF is another important reason why IR-ENT links are not present.

[20] At the same time, the update procedures were improved to better identify firms which restarted business after having been ceased on the BF.

[21] An alternative source for industry information would be the Accident Compensation Corporation (ACC), as industry classification affects the rate at which ACC levies are set, increasing the incentive to notify ACC of changes which result in a reduction of levies and raising the probability that sanctions will be imposed for failure to notify. ACC data is not currently used by Statistics NZ for BR updates.

[22] IR and Customs client codes are anonymised as part of the process that creates these identifiers.

Figure 3: LBF updating: Firm entry and exit

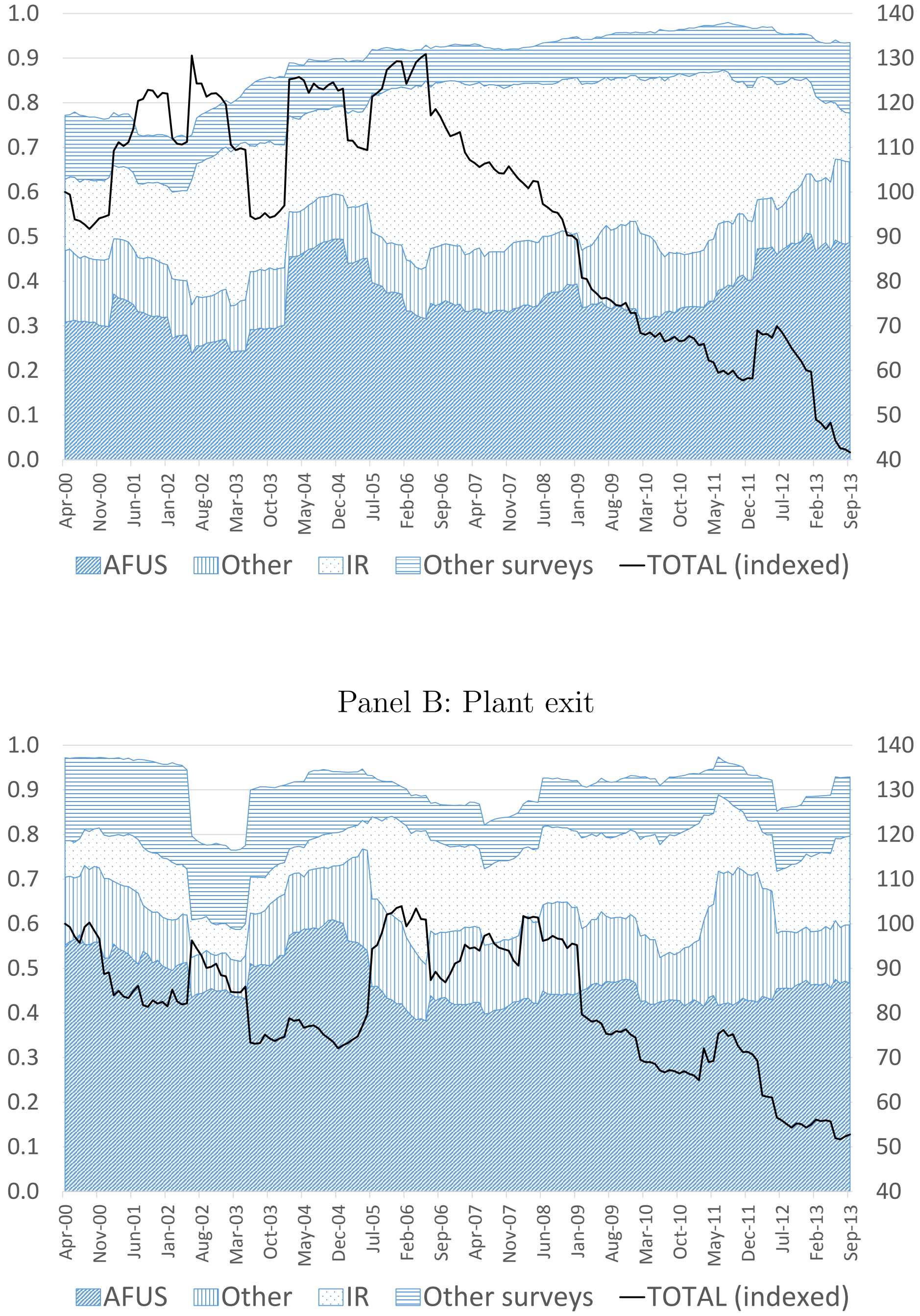


Panel A: Enterprise entry

Panel B: Enterprise exit

Left axis: 12 month rolling average of the share of each data source in firm entry and exit events recorded on the LBF. Right axis: Index of total number of recorded events.

Figure 4: LBF updating: Plant entry and exit in continuing firms



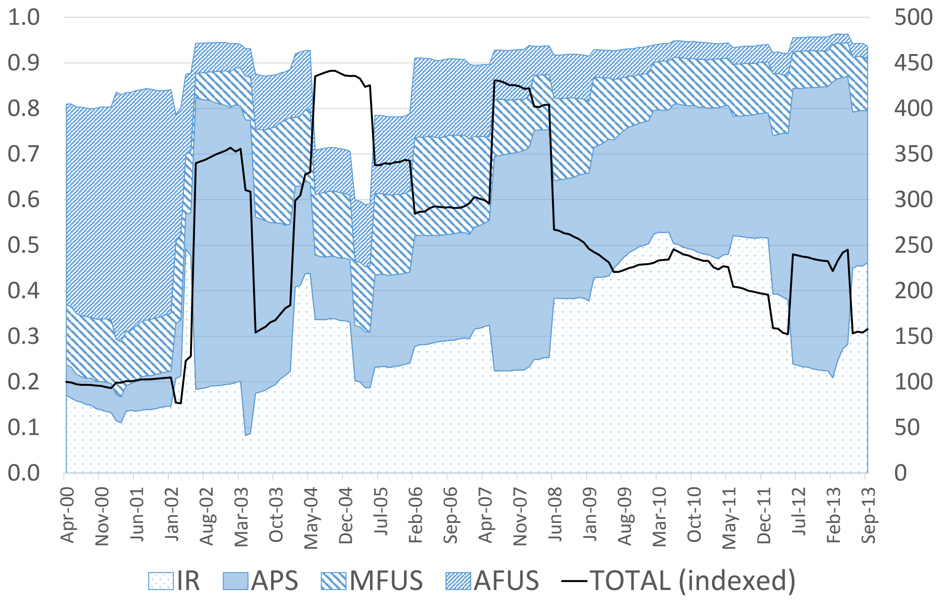
Panel A: Plant entry

Panel B: Plant exit

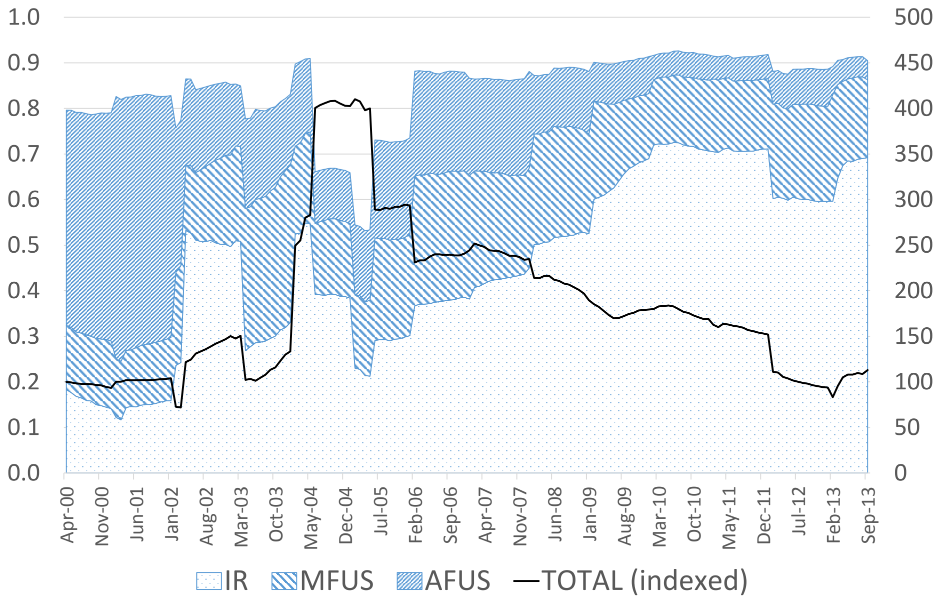
Left axis: 12 month rolling average of the share of each data source in plant entry and exit events recorded on the LBF. Right axis: Index of total number of recorded events. Excludes plant entry and exit associated with firm birth and deaths.

Figure 5: LBF updating: ANZSIC changes

Panel A: All sources



Panel B: Excluding APS-derived changes



Left axis: 12 month rolling average of the share of each data source in industry changes recorded on the LBF. Right axis: Index of total number of recorded changes. Where ANZSIC96 and ANZSIC06 classifications change simultaneously, each change is given a weight of 0.5.

Firms that are classified as "live" can be identified in the LBF as having *life_cycle_code* set to *birt* (birthed) or *reac* (reactivated).[23] We use alternative definitions of active businesses based on observed employment and tax activity because the LBF status does not always align with observed activities. This incongruence affects around 10 percent of private-for-profit firms in each year (table 5). Related to this, firm birth dates are recorded in the LBF, but are not always consistent with the observed data from other sources. One response to this issue is to calculate firm age based either on first observed employment or on recorded birth date, whichever is the earlier. Exclusively using employment data may not be an adequate approach because these data only start in April 1999, and a substantial proportion of currently active businesses began operation before that date.

Because the Business Frame tracks legal entities, firm id continuity can be broken by events that do not imply the exit of a firm. For example, if the partners in a firm decide to reestablish their existing business as a limited liability company, that firm may be issued a new firm id despite continuity in location, economic activity and ownership. Statistics NZ puts effort into repairing plant-level ids, making use of continuity of location and employees as measured in the Linked Employer-Employee Data (LEED) to generate Permanent Business Numbers (PBNs). Fabling (2011) provides a method for constructing "permanent" enterprise numbers (PENTs) exploiting the continuity of PBNs. Except where otherwise specified, firm counts presented in this paper are based on PENTs.

The coverage of the LBD extends beyond those entities which would generally be classed as "businesses". For many research purposes it is sensible to restrict attention to private-for-profit firms. We exclude firms with *business_type96_code* above 6,[24] those with *inst_sector96_code* other than 1111,1121,1311,1321,2211,2221,2291,2311,2411 (in short, private producer enterprises, central and local government enterprises, and private financial institutions) and those with ANZSIC96 codes beginning M (Government administration and defence) and Q97 (Private households employing staff).[25]

LBF load tables begin a new record (spell) whenever any of the firm's attributes change. Simple rules are used to aggregate LBF variables to the annual, enterprise level frequency required for the LBF fact table. As not all LBF variables are easily aggregable (eg, industry classification), users who are interested in whether multiple states occur during a year are advised to use the load tables to construct relevant indicators.

During the period covered by the LBD, Statistics NZ shifted from the ANZSIC96 to the ANZSIC06 industry classification. From 1999 to 2012, most firms were dual coded, but firms which had ceased prior to the change may be missing the later code, while those birthed onto the BF after dual coding ceased may not have an ANZSIC96 code. As part of the BR redevelopment, ANZSIC06 codes have been retrospectively applied to all enterprises, KAUs and GEOs that were live at any time since April 1999. However, there are some examples in which these historical changes made on the BR are not accurately captured by the Business Frame Replicator, leaving some firms with missing ANZSIC06 codes on the LBF. Since the switch to the BR in 2012, existing ANZSIC96 codes have not generally been updated to reflect changes in industry. Prior to February 2013, on average 92 percent of ANZSIC changes were applied either to both classifications or solely to the

[23] Other possible states are *bbir* (before birth), *ceas* (ceased) and *inva* (invalid).

[24] An exception is often made for government owned trading entities (*business_type96_code*=7, *inst_sector96_code*=1311) and local authority trading entities (*business_type96_code*=9, *inst_sector96_code*=1321)

[25] Under ANSZIC06, the equivalent industry criteria are codes beginning with O and S96, respectively.

ANZSIC96 code. From that date onwards, only 35 percent of ANZSIC06 updates were also accompanied by an ANZSIC96 change. From 2012 on, researchers therefore need to use the ANZSIC06 classification, or impute ANZSIC96 code changes from ANZSIC06 code changes using an appropriate concordances.[26]

The IDI includes a subset of the firm-level tables from the LBF (identified by the *br_clean* prefix). The IDI and LBD versions of the LBF are not generally identical, with the IDI being updated more regularly than the LBD, and missing some variables that are available in the corresponding LBD tables.

The country identified in firms' contact addresses is listed in the LBF. This is the address supplied to Statistics NZ for the purpose of survey mailouts, and is not necessarily related to the location of the firm's head office, or that firm's parent or group-top enterprise. For example, it may reflect the location of an accounting division of the firm or a regional office of a multi-national company.

Balance of Payments (BoP) indicator questions included in the AFUS survey were substantively revised in 2008. One notable change made at this time was the consolidation of a range of services trade indicators into a single indicator for commercial services imports and exports. At the same time, new indicators were introduced to capture activities such as contract manufacturing and merchanting (the purchase and on-selling of goods outside New Zealand), and payments made or income received for the use of intellectual property.

2.2 Annual Enterprise Survey (AES)

Purpose

The term Annual Enterprise Survey (AES) is used by Statistics NZ to refer to two distinct things: the KAU-level postal Annual Enterprise Survey (yielding approximately 16,000 enterprise-level observations per year) which collects information on firms' financial performance and financial position;[27] and a dataset compiled by Statistics NZ from responses to that survey alongside information from a number of administrative sources.[28] The main objective of the AES is to provide annual data for financial performance and financial position by broad industry groups and to support the construction of GDP statistics.

[26] Concordances between ANZSIC96 and ANZSIC06 are available but not all codes have a one-to-one correspondence across the two versions (`http://www.stats.govt.nz/methods/classifications-and-standards/classification-related-stats-standards/industrial-classification.aspx`). One possible response to this issue, which we have used for contributions to OECD research where consistency is required both across countries and over time, is to randomly allocate firms which are not coded under the relevant ANZSIC vintage an industry code based on the observed frequency of the potential codes among those firms which are dual coded. Alternatively, the Fabling-Maré productivity dataset infers a permanent industry based on most likely concordance, again based on observed frequencies in the data.

[27] The AES postal sample has fallen from over 20,000 in 1999, with the most substantial reductions occurring from 2009 onwards.

[28] The source of data in the AES table is identified by the variable *data_srce_code*. The two main sources are the AES postal survey (*postout*) and the IR10 (*tax*). The IR10 is included independently in the LBD. Other source codes are *government* and *superannuation*.

Administrative and survey data sources undergo processing by Statistics NZ for the purposes of constructing National Accounts, such that it is not generally possible to link the transformed AES dataset back to the individual question responses. While this processing is designed to maximize comparability of the administrative data and the postal survey for National Accounts purposes, it is not clear that the tax component of the data is suitable for longitudinal firm-level research. As such, much of the research using the LBD to date has used AES postal responses in combination with the raw IR10 data from *fact_i10_enterprise_year*, enabling more control over the transformations applied to the latter.[29] The remainder of this section is therefore focused on the AES postal collection.

Content

AES contains information on firms' financial position and performance. Financial performance data include various categories of income and expenditure; dividend payments; information about fixed and intangible assets, including closing book values,[30] additions, disposals, depreciation or amortisation, revaluations, and the value of capital work undertaken by employees. Financial position information include current and non-current assets and liabilities, and equity and shareholder funds. From this data, Statistics NZ derive key National Accounts variables, such as gross output, value added, intermediate consumption, and gross fixed capital formation, components of which are included in the data at the unit record level.

Coverage

The AES population includes all active businesses. Of these, a stratified random sample is selected to receive the annual postal survey. Stratification is based on industry, type of legal entity and firm size. Over time, there has been a significant decrease in the sample size for AES, with greater use being made of administrative data sources (table 6). Firms which make a substantial contribution to National Accounts aggregates continue to be selected with a probability close to one. In 1999, 8.1 percent of AES dataset observations were based on postal responses, compared to 5.3 percent in 2012.

The AES questionnaire was redesigned in 1999 and 2009. Prior to 2009, only around 60 percent of AES survey respondents were asked to provide financial position data (table 6). Firms without financial position data appear to be a random sample of the survey population, and are recorded in the LBD as having zeros for the financial position variables.

The AES sample is broadly stable over consecutive years, in order to provide time-series consistency for the aggregate estimates. The sample is then reselected periodically, most recently in 2009. The sample was boosted in 2006 to support the transfer of industrial classification from ANZSIC96 to ANZSIC06.

[29] Fabling & Maré (2015b) use AES postal data with IR10s to construct consistent measures of multi-factor productivity, and provide detailed information about the two datasets.

[30] Although both opening and closing book values are collected, only the closing values are available in the LBD.

While the survey sample is drawn at the enterprise level, separate survey forms are sent to each KAU within the enterprise.[31] This enables the collection of industry-specific data in the case where enterprises undertake a range of different activities. There are 26 different AES forms, differentiated by industry. This allows for the collection of particular industry-relevant variables, enabling industry-specific treatment of the data (eg, stock of aircraft and helicopters is required only for the Services to Agriculture and Transport and Storage industries). Table 7 summarises the fixed asset breakdown by form type.[32]

Tips and tricks

Due to the way Statistics NZ samples business surveys, there is a high probability that firms that meet the population criteria for multiple surveys will be surveyed for both. AES is the exception, and is sampled in a way designed to minimise the overlap with other business surveys, at least for firms in less than full coverage strata (usually smaller firms).[33]

The Agriculture and Horticulture industries are excluded from the AES postal sample, as information for these industries is collected via the Agricultural Production Survey (APS). Services to agriculture are included in the AES postal sample (form type AC).

AES values are recorded in thousands of dollars, so need to be multiplied by 1000 to be comparable with other LBD data. As with most LBD data, values are recorded exclusive of GST.

Around 25 percent of AES postal responses are imputed. This is identified by the variable *unit_status_code*, which can take values: C (clean); S (suppressed warnings); E (error); M (imputed due to data quality); R (imputed due to non-response); and W (imputed financial position data). Status codes C and S indicate (non-imputed) data of usable quality. Imputation flags are missing for the 1999/00 year and earlier. Table 6 identifies the number of firms in each year with non-imputed AES data, distinguishing according to whether financial position data is available.

Because there are multiple form types, variables must be aggregated to get to consistent fixed asset definitions. In particular, due to AES survey processing, the "plant, machinery and equipment" variable available in the AES dataset includes the "All other fixed assets" category.

Asset-specific tax depreciation rates in New Zealand are intended to align with economic depreciation rates. This makes it more defensible to use book values to estimate the value of capital services. See Fabling & Maré (2015b) for a comprehensive discussion of the calculation of capital services measures from the LBD. As the data underlying AES and IR10s are prepared following accounting standards and tax rules, any changes in

[31] AES load tables are provided at the KAU level in *load_AES_data1* and *load_AES_data2* and can be aggregated to the enterprise level using *load_AES_header*.

[32] There is also a form type (AL) which is used to identify transactions between different parts of the same business. The content of this form type is not comparable to the other types.

[33] Statistics NZ allocates a permanent random number (between zero and one) to each firm when it first appears on the Business Frame. Random sampling for most surveys is then conducted by starting at a specified point on the number line and picking up firms to the right until the appropriate sample fraction is reached. For AES, the sample is drawn from the top end of the number line, working to the left. In the absence of entry/exit or redesign of the stratification, the same firms would be drawn each year.

these guidelines will affect the reported data. For example, the removal of loadings from depreciation allowances which took place in May 2010 appears to have affected reported book values (see Fabling & Maré, 2015b).

Industry codes reported in the AES fact table should not be used. These are a simple aggregation of KAU-level industry codes used for sampling, and do not necessarily accord with firm-level industry codes in the LBF since the latter capture predominant activity for multi-KAU enterprises.

AES load tables hold multiple extracts of the data. Use the version with the highest *extract_ver_nbr* in each year.

2.3 Tax-filed accounts information (IR10)

Purpose

The IR10 is an abbreviated set of financial accounts composed of a profit and loss statement and a balance sheet. The information is collected by IR for analytical purposes (such as tax compliance risk analysis, policy and strategic research)[34] and provides an alternative to filing a full set of financial accounts with IR as part of the firm's annual tax return.

Content

The IR10 form consists of a front page, which includes a breakdown of income (sales, interest and dividends received etc) and expenditure (including items such as depreciation, and salaries and wages), and a back page, which summarises the firm's balance sheet. Balance sheet items include fixed assets (broken down into vehicles; plant and machinery; furniture and fittings; land and buildings; and other), liabilities (current and term), and shareholders' funds.

In the 2012/13 tax year, the IR10 *Accounts Information* form was replaced by the substantively revised and renamed IR10 *Financial statements summary.* This will begin to affect the LBD when the 201303 *dim_year_key* data is included in the database. Changes to the form were made in order to better align the IR10 with taxpayers' financial statements, to improve the clarity of response categories, and to meet the evolving information needs of IR.[35]

From a longitudinal research perspective, the IR10 form changes are significant. While some response items can be tracked across the two versions, many of the items used to generate key variables have been changed and new variables have been created. Alongside changes to specific items, the redeveloped IR10 no longer asks firms to indicate whether their accounts are GST-exclusive, or whether they are reporting for a 12 month period.[36] It is unclear at present what effect these changes will have on research methods, since the new data is not presently available in the LBD.

[34] Source: http://www.ird.govt.nz/technical-tax/general-articles/ga-revised-ir10-summary.html

[35] ibid.

[36] In the new Guide, firms are instructed that "The IR 10 should record GST-exclusive figures, unless the financial statements are prepared on a GST-inclusive basis".

Coverage

When submitting their annual tax returns, firms are required to provide accounts information to IR. This can take one of several forms, depending on the business type. The presence of multiple filing options for compliance means that IR10 coverage is less than complete. In particular, larger firms are more likely to submit their financial accounts, while smaller firms are more likely to submit an IR10, biasing IR10 coverage towards smaller firms. Around 68 percent of employing firms with ten or fewer full-time equivalent employees (FTE) returned a useable IR10 form in 2012, compared to 37 percent among firms with more than 50 FTE (table 12).

Tips and tricks

Statistics NZ carry out consistency checks separately for the front and back pages of the IR10, following the same rules used in AES processing of IR10 data. The outcome of these checks is reflected in the variables *fp_edit_status_code* and *bp_edit_status_code*, which can take values: P – pass; F – fail; and Z – zero-punched (data present but all values are zero). These consistency checks make limited edits to the data, including adjusting values in "other" categories to ensure components sum to reported totals, and transforming the data to be GST-exclusive using simple assumptions about the GST status of income and expenditure categories.[37]

IR load tables (including GST and IR4s as well as IR10s) are referenced to the confidentialised unique IR firm identifier *ird_uid.* This can be linked to a Statistics NZ enterprise number within the LBD using the *load_lbf_enterprise_ird_link* table. The confidentialisation of IR numbers happens independently in the LBD and IDI, so that *ird_uid*s for the same IR number are unlikely to be the same across the two databases. Users must access the IDI table *security.ibuldd_to_xref_ileed_ird_uid* for a concordance between the two if linking on IR numbers. Most users will probably prefer to use enterprise number (which is consistent) for linking.

Research and development expenditure is commonly under-reported in the IR10 because the form requires salaries and wages to be reported separately (see the box on page 42 for further details).

2.4 GST and the Business Activity Indicator (GST, BAI)

Purpose

Goods and Services Tax (GST) returns are collected by IR as a core part of the administration of the tax system. These are processed by Statistics NZ to produce the Business Activity Indicator, which provides monthly aggregate and industry level statistics on GST sales and purchases.

[37] As already noted, it is not clear what the changes to the IR10 form mean for these edits. Raw data is included in year-by-year load tables (*load_i10_return_YY*).

Content

GST data provide information on sales and purchases and are collected on a monthly, bi-monthly or six-monthly basis by IR, depending on the size of the firm. Statistics NZ manipulate this raw data to create the Business Activity Indicator (BAI) dataset (also included in the LBD). The primary manipulations applied to generate the BAI data are to temporally apportion the GST data down to a monthly frequency, apportion returns across GST group members, remove "capital spikes," and apply limited imputation in cases where a single return appears to be missing.

Coverage

As GST in New Zealand has very broad coverage and non-compliance is low, these data are often treated as full coverage, with the following exceptions: GST sales data are a bad indicator of revenues for the financial services and residential rental accommodation industries, and for suppliers of fine metals, as the main products of these industries are GST-exempt; and exported goods are zero-rated, though the value of these sales are captured in GST filings.[38]

Tips and tricks

GST/BAI sales and purchases data are GST-inclusive whereas most other financial data, including survey responses, are on a GST-exclusive basis. Adjustments to GST sales data to be GST-exclusive should make use of zero-rated sales data, noting that GST total sales *includes* the zero-rated sales amount.[39]

The GST rate rose from 12.5 percent to 15 percent on 1 October 2010. A composite rate of 13.75 percent is often applied when using annual data for the year ending March 2011.

Firms may file GST returns monthly, bi-monthly, or six-monthly, identified by *gst_freq_code*.[40] BAI apportionment uses industry (ANZSIC96) seasonal patterns to apportion values for bi-monthly filers and divides evenly by six for six-monthly filers. Firms are advised by IR to align one of these filings with their balance date (ie, the end of the firm's financial year).

Some researchers have used zero-rated GST sales as an indicator of exporting. Some income sources unrelated to exports are also zero-rated for GST purposes implying a trade-off between accuracy and the comprehensive nature of a GST-based indicator of exporting. Alternative sources of trade data are summarised on page 41.

[38] Detail of GST-exempt and zero-rated supplies can be found at `http://www.ird.govt.nz/gst/additional-calcs/calc-spec-supplies/calc-exempt/`.

[39] If X is *gst_total_sales_amt*, Z is *gst_zero_rated_sales_amt*, and t is the GST rate (as a decimal) then total tales excluding GST is $\frac{1}{1+t}(X-Z)+Z$. For the period prior to October 2010, $t = 0.125$ and the formula simplifies to $(8X+Z)/9$.

[40] *gst_freq_code* also identifies which months the firm files in. For example, code SA refers to six-monthly filers that file in January (and hence also in July), while TA refers to two-monthly filers that file in January (and also in March, etc).

A similar trade off exists for the measurement of value-added. In the GST data, sales and purchases of capital equipment are observed in the year the transaction occurs, and would be attributed to sales and intermediate consumption, respectively, in that year. Thus, an expanding firm may look less productive due to capital investment, while a failing firm could look more productive due to asset sell-downs. AES and IR10 data provide a more appropriate measure of intermediate consumption which excludes capital purchases and identifies depreciation expenses.

2.5 Linked Employer-Employee Data (LEED) and Individual Tax Returns (IR3, IR7, IR4S)

Purpose

Detailed employment data in the LBD is sourced from IR data, collected as part of the tax administration system. Until 2015, these were processed by Statistics NZ and provided to researchers in the LBD as aggregated firm-level employment data, generated through the LEED processing system. From 2015 onwards, this aggregation has been discontinued, with researchers instead provided with access to the underlying individual level data – the Employer Monthly Schedule (EMS), Company Shareholder Details (IR4S), Income Tax Return for Partnerships (IR7/IR7P), and Individual Tax Return (IR3).

Content

Fabling & Maré (2015a) provides a detailed description of these data and a methodology for using the annual and monthly tax information together to generate monthly estimates of headcount and full-time equivalent employment, and annual estimates of working proprietor labour input. These data are available in *ibuldd_research_datalab* as tables *pent_year_L_IDI_YYYYMMDD* and *pent_pbn_month_L_IDI_YYYYMMDD*, where *YYYYMMDD* refers to the IDI instance the tables are based on.

Coverage

The EMS is a mandatory monthly reporting requirement for all employing firms. We assume that the absence of employment data implies zero employees on the grounds that personal income tax non-compliance is likely to be negligible in the population of firms that meet the mandatory GST filing threshold.

Annual income tax returns are mandatory for the relevant types of businesses and individuals. IR3s are required of individuals who earned income other than salary, wages, interest, dividends and/or taxable Māori authority distributions. The IR4 (and IR4S) is required from all active New Zealand resident companies. The IR7P is required from all partnerships.

Tips and tricks

Currently, researchers who wish to access the individual-level data in conjunction with the LBD firm data must do so through the IDI server (*snz-idiresearch-prd-sql\ileed*), linking back to the LBD server (*wprdsql31\ibuldd*) via a link server.[41]

Plant-level employment counts and employee information in the LBD rely on the allocation of individuals to physical locations (PBNs) by Statistics NZ. Multi-location firms provide Statistics NZ with employee counts for each of their locations as part of the BF/BR update process. This information is collected annually for known multi-location firms (see section 2.1). In order to link individual employees to work locations, Statistics NZ makes use of contact address information provided to IR, linking individual workers to the closest location to their address, subject to matching the LBF plant-level employment counts. Once an individual has been allocated to a work location, they remain associated with that location. When Statistics NZ receives updated information about plant-level employment, individuals can be reallocated across plants in order to satisfy the updated employment counts.

The accuracy of the allocation relies on worker addresses held by IR being up-to-date *and* identifying the individual's residential address.[42] Ongoing development of the IDI is expected to improve the quality of address information (including historical addresses) by allowing Statistics NZ to make use of address data from a range of sources. However, at present, it is likely that the individual IR address information is inaccurate for a significant proportion of employees, since most taxpayers no longer receive mailed correspondence from IR.

At the same time, ongoing reductions in the sampling coverage of the BR update survey imply that the ability to identify multiple location firms, and the frequency at which plant-level employment counts are observed, will continue to decrease over time (see section 2.1).[43]

2.6 Company Income Tax Return (IR4)

Purpose

The IR4 is collected by IR as part of the administration of the corporate tax system.

[41] For example, to access LBD data from within the IDI an SQL query would take the form of *select top 10 * from [IBULDD_PROD_CLEAN].ibuldd_clean.dbo.fact_lbf_enterprise_year*, where *[IBULDD_PROD_CLEAN]* is the link server. Link servers operate in one direction only and there is not currently a link server set up in the other direction.

[42] Contact addresses may not accurately reflect residential addresses if, eg, workers choose to receive IR correspondence through an accountant or at an alternate address. Accurate allocation also relies on geographic proximity being a sufficient indicator of employment location, which may be problematic if a firm has multiple plants within commuting distance.

[43] At the same time, the increasing availability of individual contact addresses due to the integration of additional individual-level data sources may enable new methods for identifying multi-location plants to be developed, using the existence of clusters of employees to indicate the potential location of new plants.

Content

The IR4 form is a declaration of taxable income for companies and, as such, include values of overseas income, interest and dividends, income from "business or rental activities," and the value of net loss carry forwards, net loss payments and subvention payments.[44] The IR4 also includes a number of disclosure questions, which provide indicators of non-resident ownership or control, payments made to non-residents, receipt of foreign-sourced dividends, and the value of any shares repurchased, redeemed or cancelled during the year.

Coverage

IR4 filing is an annual mandatory requirement for "all active New Zealand resident companies" including body corporate and trusts, with the exception of look-through companies.

The coverage rate of the IR4 appears to be high for firms in a target population derived from Statistics NZ data. IR4 returns are available for around 90 percent of active, private-for-profit registered limited liability companies (LLCs); 75 percent of co-operative companies; and 65 percent of branches of companies incorporated overseas. LLCs in turn account for 58 percent of employing firms (66 percent of employment) in 2012.[45]

Tips and tricks

The IR4 includes an annual indicator of foreign ownership or control (*ir4_nrcontr_ind*), as used by Maré et al. (2014). This question is used by IR to identify situations where companies may be able to transfer profits internationally through related-party transactions such as transfer pricing. The definition of non-resident control is quite broad, including both New Zealand subsidiaries owned by foreign companies and also New Zealand companies owned or controlled by non-resident natural persons.

The IR4 also includes a question on overseas income received (*ir4_osinc*) and overseas tax paid (*ir4_ostaxpd*), but does not identify the source of this income (eg, overseas sales, earnings from assets etc).

Along with the main firm-level IR4 data, shareholder remuneration (the final page of the IR4 form) is available through the IDI – *ir_clean.ird_attachments_ir4s*, showing non-PAYE salary payments made to working proprietors of the business.

[44] Net loss carry forwards allow continuing firms to offset net losses from a previous year against their current year's tax liability. Subvention payments and net loss payments allow for companies under common ownership to offset the losses of one company against the tax liability of the other company.

[45] The proportion of LLCs in the active firm population has increased substantially over the period covered by the LBD, from 35 percent in 2000. This has primarily been a move towards incorporation among smaller firms, with almost no change in the proportion of employment covered by LLCs.

2.7 Government Assistance Programme lists (GAP)

Purpose

The New Zealand government provides a range of assistance for firms, with the intention of enhancing economic growth and international competitiveness. These programmes provide support for activities such as research and development, training, investment and market development. In many cases the support provided is financial, but can also involve in-kind support such as advice and training. The GAP dataset is constructed from data generated as part of the administration of these programmes.

The GAP data included in the LBD was provided by three main sources – the Ministry of Business Innovation and Employment (MBIE) and its predecessor agencies;[46] New Zealand Trade and Enterprise (NZTE); and the New Zealand Venture Investment Fund (NZVIF). A small number of programmes administered by Te Puni Kokiri and the Ministry of Social Development are also included, though these data have not been updated since 2007.

Content

The data available and the structure of the collection reflect the needs of the administering agencies, and differ across providers and assistance programmes. In general, the data includes the period in which firms received assistance, the specific assistance programme in which they participated and, where appropriate and available, the value of assistance awarded and the amount disbursed.

Coverage

GAP datasets are, in principle, full coverage for the assistance schemes for which data has been supplied. As the data is probabilistically matched on name and address information, some observations are not linked to businesses on the LBF. Overall, 91 percent of businesses that received assistance between 2009 and 2013 could be matched to the BF (Statistics New Zealand, 2015). For 2001-2008, overall match rates were lower, but this was mainly due to one particular scheme (NZTE's Enterprise Training Programme) with particularly low match rates (de Beer et al., 2010). Match rates tend to be lower for schemes which target very small firms since they may not meet the threshold for appearing on the BF, or may change their business name or address prior to becoming economically significant. Programme information from non-matched businesses is included in the fact tables, identified by "enterprise numbers" beginning with FI.

Tips and tricks

To date, linking of GAP data to the LBD has been undertaken on an ad hoc basis, in collaboration with the Ministry of Business, Innovation and Employment. The most

[46] MBIE data is restricted to funding associated with business R&D and includes funding from Callaghan Innovation.

recent data link (up to 2013) was completed in 2014, and is documented in Statistics New Zealand (2015). The initial linking is documented in de Beer et al. (2010).

Firms are instructed to include the value of grants and subsidies received (eg, through government assistance programmes) in their GST sales, as these grants are liable for GST. An exception is made for grants intended for international overseas development, where GST is payable only on the portion of the grant used for administration and capability building in New Zealand. As such, caution should be taken using the GST sales data as a metric in evaluation of GAP outcomes.

2.8 Merchandise trade data (OMT)

Purpose

Merchandise trade data is collected from importers and exporters for the administration of the Customs and Excise Act 1996. The New Zealand Customs Service collects information on all shipments of goods above a minimal value threshold coming into or leaving New Zealand. The information is used as part of Customs' security screening and to calculate taxes payable on imported goods.

Content

Merchandise trade data contains daily shipment-level information for over two decades (1988-2014) covering (10-digit Harmonised System) goods, countries of origin and destination, values, quantities, weights, currency of trade, port of entry/exit and mode of transportation.

Coverage

Customs declarations are legally required for all shipments of goods valued over NZD1000. In line with the principle of keeping everything, merchandise trade data which cannot be matched to a firm (for example, Customs declarations of private individuals) is retained, using dummy "enterprise numbers" starting with EC.

Tips and tricks

On average, from 2000 onward, over 99 percent of export value and 98 percent of import value is linked to firms on the LBF. However, the linking rate has fallen over time (from an average of 0.997 between 200003-200203 to 0.986 for the period 201103-201303 for exports, and 0.993 and 0.989 for imports for the same periods). This may relate in part to increasing trade by private individuals but may also reflect some deterioration in the match rate.

Trading firms are allocated a Customs Client Code when they first file export or import data with the New Zealand Customs Service. Customs Client Codes are linked to the LBF initially using tax ids, then probabilistic matching on names and addresses, and finally manually to link any remaining unmatched large-value Customs clients. The quality of the match deteriorates (at least in terms of the share of aggregate trade linked) prior to 1996, and some currency variables are only available on a comprehensive basis subsequent to the introduction of mandatory electronic filing (April 2004).

The existence of enterprise groups (ie, parent-subsidiary relationships) complicates the analysis of overseas merchandise trade data, as exported goods may pass (physically or administratively) between members of a business group (eg, from the manufacturing enterprise to the head office or a wholesale enterprise) before crossing the border. In addition, as Customs Client Codes are linked to an enterprise number at a single point in time, restructurings which shift activity between group members can mean that active Customs Clients continue to be associated with apparently inactive (non-employing) firms. As restructurings occured among some of New Zealand's largest exporters over the 2000s, a failure to identify and reallocate this trade can substantively affect reported aggregates for any analysis which links trade to firm characteristics. Group aggregation has sometimes been used to alleviate this problem.[47] Fabling & Sanderson (2010) present a methodology for reallocating trade back to the producing manufacturer.

Unlike most data sources, disaggregated merchandise trade data is held in fact (eg, *fact_cus_export_header, fact_cus_export_line*), rather than load, tables, which simply pull all the relevant load data into more easily usable tables.

Where goods are invoiced in currencies other than the NZD, values are provided as recorded by the exporter both in the currency of invoice (*val_in_curr_code*) and translated into NZD at the exchange rate provided by the exporter (*local_curr_val_nbr*), as well as the value used by Statistics NZ (*free_on_board_val_nbr* = *val_in_curr_code* × a common monthly exchange rate used by Statistics NZ). The exchange rate provided by the exporter is also recorded (*xchng_rate_nbr*), as well as an indicator of whether the transaction was explicitly hedged against exchange rate risk (*export_exchange_rate_code* = C for transactions with forward cover or a fixed exchange rate, F for floating exchange rates, and N for NZD).

In general, quantities are provided both by statistical unit of measure and apportioned gross weight. The unit of measure is specific to the good (harmonised system code) in question. For example, meat products are generally assigned the unit KGM (kilograms - industry standard), whereas beer is assigned the unit LAL (litres of alcohol - for the collection of duty). Some HS codes do not have an associated unit, reflecting a category for which a relevant unit has not been identified (eg, goods described as "not elsewhere classified" commonly do not have a unit of measure).[48] Apportioned gross weight takes the weight of the shipment as a whole and apportions it to the component parts, so is unlikely to provide an accurate volume measure for shipments containing a mixture of goods.

Between 1988 and 2015, there have been 49 minor and 4 major revisions to the Harmonised System classifications. Fabling & Sanderson (2013) use Fortran code developed and documented by Abowd et al. (2002) alongside detailed HS concordances provided by

[47] While group allocation is a feasible solution for cross-sectional comparisons, it is not appropriate for many potential longitudinal analyses, as longitudinal group aggregation (aggregating together all firms which ever share group membership) results in the majority of exports being associated with a small number of "super" groups.

[48] A full list of statistical unit codes is available through `www.customs.govt.nz`.

Statistics NZ to identify the minimal HS grouping required to concord codes over time. This concordance is stored on *ibuldd_research_datalab* as table *hs2012_concordance*, and relates each HS10 code which was in use between 1996 and 2012 to an indicative code from the 2012 revision.[49]

2.9 Intellectual Property Office data (IPO)

Purpose

The IPO dataset is based on administrative records from the Intellectual Property Office of New Zealand (IPONZ), the government agency responsible for granting and registration of intellectual property rights. Registration gives firms and individuals legal protection for their intellectual creations, including inventions, designs, symbols, names and images, enabling them to earn recognition or financial benefit from what they invent or create.

Content

The IPO data includes patents, trademarks, designs and plant variety rights filed with IPONZ. The data distinguishes between filed applications and registrations. Patents data is divided into four technology categories: biotech, chemical, electrical, and mechanical.

Coverage

In the past, IPO data has been linked to the LBF on an ad hoc basis. The most recent update was in 2010. The data has been probabilistically matched to enterprise names and addresses in the Business Frame, then aggregated to an annual frequency.[50]

As the IPONZ data includes applications filed by overseas businesses and by individuals, only around 60 percent of applications were matched to an enterprise number. At the time of writing, Statistics NZ is investigating updating the IPONZ data, and using a different probabilistic matching method (but not attempting to link the IPONZ data to individual applicants within firms, which may be feasible). It is unclear what the structure of the resulting tables will look like on the LBD.

2.10 Business Operations Survey (BOS)

Purpose

The Business Operations Survey (BOS) is the only survey element within the LBD which is explicitly designed to support longitudinal firm-level research, as well as producing a range of cross-sectional aggregates. BOS also collects key firm-level information required

[49] The variable *hs2012_indicative* is intended as a semi-informative label for the grouped codes. The HS2012 code chosen is the highest HS2012 code within the group. Seven codes ceased to exist over this period and are not included in the table.

[50] Load tables are not available for confidentiality reasons.

for international comparisons – particularly, the biennial collection of innovation and business ICT use statistics.

Content

BOS in an omnibus collection of survey data related to firm behaviour and performance. The survey has a modular design with Module A collecting annual financial and employment data, and qualitative information on firm performance.[51] Module B alternates between collecting innovation statistics (in odd years, using the revised Oslo Manual classification of four types: product, process, marketing and organisational design) and information and communication technology use (in even years, following OECD guidelines). Finally, Module C content is open to competitive bidding between government agencies. Table 8 identifies module topics and sponsoring agencies to date.

The general business practices module and innovation module have predecessor surveys also included in the LBD – the Business Practices Survey (BPS 2001) and the Innovation Survey (INN 2003) respectively.

Coverage

The population for BOS is all private-for-profit firms with a rolling mean employment of at least six (roughly 35,000 firms). From that population between five and seven thousand useable responses are collected (depending on the year), with the realised response rate always being at least eighty percent.[52]

BOS responses come from a mail-out to a random sample of the population, stratified on industry and firm size, and enhanced with a longitudinal top-up sample. The initial top-up ensured that all respondents in 2005 were resurveyed annually until 2011, even if they would not otherwise have been included in the later random survey samples. The panel was reset in 2012, and a new longitudinal top-up was started for the years from 2013 onwards. The sample also underwent a major redesign in 2007 to account for the shift from ANZSIC96 to ANZSIC06. The survey was dual sampled under both the old and the new system in that year, resulting in a one-off increase in the number of respondents to approximately 7,200.[53] At the same time, the scope of BOS was expanded to include several previously excluded industries, primarily in personal services.

Table 9 reports the overlap across survey years for BOS and its predecessors. Of the 2,754 respondent firms for the Business Practices Survey 2001, 37.1 percent also responded to the Innovation Survey in 2003, 52.5 percent to BOS 2005, and so on. The overlap declines in 2012, due to the panel refresh. Overlap in coverage is also much weaker for the three non-BOS surveys, which had a different population and sampling strategy and did not

[51] The quantitative data collected has largely been dropped since 2009 on the grounds that most of this data is now available from other LBD sources. Outside of the research sphere, the associated reduction in BOS respondent load is probably the greatest achievement of the LBD.

[52] Response is mandatory under the provisions of the Statistics Act for all the surveys in the LBD.

[53] Statistics NZ created two versions of the 2007 BOS data – one using ANZSIC96 sampling weights, strata code and imputation, and the other using ANZSIC06 weights, strata and imputation. The LBD holds the ANZSIC96 version.

include a panel element in sampling.[54] Looking across all nine years of BOS (2005-2013), of the 13,857 firms which submit a useable BOS response, 1,437 are observed in every year, and a further 2,304 are observed in at least 7 years, with each firm providing an average of four responses. Conversely, over a quarter of the firms which ever appear, appear only once (table 10).

Tips and tricks

Statistics NZ uses a range of response codes (*response_code*) to identify the usability of survey responses in BOS: R – responded (response used in official statistics); F – useable responses not included in official statistics (eg, firm was included in panel top-up); C – ceased (firm is no longer operating); Q – firm did not answer sufficient questions to be considered a useable response (60% threshold, after accounting for survey routing); L – response was received after the survey was closed; U – unusable (eg, all "don't know"). In 2005, F indicates firms which were oversampled due to a mail-out error in which subsidiary firms were initially excluded from the population. In 2006, the panel top-up were included in the official statistics, and so were recorded as response code R, and with final weight of one. In 2007, F includes both panel top-up and firms which were sampled under ANZSIC06 but would not have been sampled under ANZSIC96.

The last digit of the stratum code give a consistent indicator of whether firms are in the panel top-up: 1-4 are firm size groups ($[6, 20)$, $[20, 30)$, $[30, 50)$, $>= 50$ respectively); 5 is state-owned enterprises (SOEs) and local area trading authorities (LATEs); and 6 are panel top-up observations.[55] The first two characters of the stratum code refer to the industry stratification, which is largely two-digit ANZSIC.

Survey respondents are allocated a weight at the time of selection into the survey (*selection_weight*), which is updated to account for non-response and ceased firms to give the final weighting used in published outputs (*final_weight*).

BOS variables are identified in the LBD according to their survey line code (eg, *A2301*), corresponding exactly to the relevant survey questionnaire. In addition to the data itself, each variable also has edit and imputation flags (eg, *A2301_e, A2301_i*), which identify whether the item in question has been edited by Statistics NZ (eg, in response to internal inconsistencies) or whether the item has been imputed based on the response of a "nearest neighbour" firm in the same stratum (ie, donor imputation at the question level). Edits can involve Statistics NZ staff contacting the respondent to clarify a particular response, as well as the application of simple edit rules,[56] and we take these to be an accurate reflection of the true response.

With the exception of the removal of financial questions from Module A in 2009, Module A and B questions have been highly consistent over the life of the survey, though question numbers and hence variable names have changed. As question numbers and content differ across years, BOS data is included in the LBD with separate fact tables for each year.

[54] As BOS is approximately full coverage for large firms, the overlap across time tends to be composed of larger firms (not shown).

[55] SOEs and LATEs were not included in the population in 2005 and, since it was the first year of the survey, only stratum numbers 1-4 were used.

[56] For example, if a respondent checks both "yes" and "no" in a routing question, but continues to answer the related questions, the initial routing question may be edited to match the subsequent responses.

Firms are asked to provide any financial information (financial questions prior to 2009 and R&D spend in all surveys) using GST exclusive values if possible, and are asked to indicate whether they have used GST inclusive or exclusive values. Variable names preceded by "D" or "LD" are "derived," which involves adjusting to make numbers GST exclusive, and pro-rating values for firms whose responses relate to less than a full year. Pro-rating is used in official statistics, but relies on strong assumptions which are unlikely to be appropriate for research use. Pro-rating stopped in 2009, and the part-year operation question is no longer asked in the survey.

In 2007, there was a substantial change in the routing of the Innovation module. In the 2005 version of the survey, only firms which indicated they had innovated in the past two years were asked to identify activities they had undertaken to support innovation. In the 2007 and later versions, all firms are asked to identify such activities, and firms which indicate that the activities were in support of innovation are also routed so that they answer the rest of the module content.

Statistics NZ's headline figures on firm innovation follow OECD guidelines by including firms which have abandoned one or more activities intended to result in an innovation as "innovating firms." Aggregate innovation statistics derived from the micro data will not replicate official statistics unless the same inclusion is made.

2.11 Research and Development Survey (R&D)

Purpose

The Research and Development (R&D) Survey is a joint survey between the Ministry of Business, Innovation and Employment (MBIE) and Statistics NZ. It is the main source of aggregate information on R&D expenditure in New Zealand, including information from universities and government departments as well as firms.

Content

The R&D Survey collects information on current and capital expenditure and employment in R&D as well as high-level information on funding sources and the purpose(s) of the research.

Coverage

The survey is conducted every even year and excludes industries that are deemed to be non-R&D performing (eg, wholesale and retail trade). Around 2,000 to 3,000 useable responses are collected each year. As the core purpose of the survey is the production of aggregate information, the sample is focused on known R&D performers, including recipients of government funding, though the practical implementation of this focus has varied over time. R&D activity in large firms is identified through the AFUS/BRUS indicators, with a threshold of $5000, as well as a range of other sources.

Tips and tricks

The R&D survey asks respondents to identify total R&D expenditure, then later asks them to allocate this expenditure to a range of different expense categories, identifying the shares associated with different funding sources as well as the shares associated with different purposes. In some cases, firms are given the option of whether to provide dollar values or to allocate percentage shares of the total. Statistics NZ then uses the percentage information to derive dollar values, which are identified in the dataset by the prefix "D" or "LD." These derived variables also adjust responses to be GST-exclusive and, where respondents indicate that they are reporting for a part-year, pro-rate the responses. As with BOS, pro-rating probably doesn't make sense for micro analysis.

Edits and imputation flags are provided for each response item, identified by the suffixes *_e* and *_i* respectively.

From 2006 onwards, response codes are used to identify unusable surveys, as discussed in section 2.10 with respect to the BOS survey. However, in the R&D survey (and other Statistics NZ surveys), F refers to firms which failed to return a response. An additional code used is K, which indicates that key questions have not been answered. Earlier years used alternative indicators (*response_flag* in 2002, *non_response_reason_code* in 2004).

Response items are also allocated an "imputation status" flag (*_is*), which identifies whether the observation is valid for use as a donor (V = a valid entry for use in imputing other units, U = unlinked, excluded from being used to impute other units). In the case of imputed items, the imputation status provides information on the imputation method used (M = current mean method, D = donor method, P = pro-rate method). If a survey form was identified as unusable (*response_code* = F, Q or C), no imputation is performed.

BOS and the R&D survey both collect R&D information, but the content and coverage differs (see page 42).

2.12 Agricultural Production Survey/Census (APS)

Purpose

The APS is conducted by Statistics NZ to provide information about agricultural, horticultural and forestry activity. APS responses are used in the derivation of GDP measures for agriculture, horticulture and forestry, and these industries are excluded from the AES survey population.

Content

The APS collects information on the inputs and outputs of agricultural production, including the total area of land by land use, application of various types of fertilizer, livestock numbers, and production of various crops.

Coverage

The population for the APS is all businesses involved in agriculture (livestock and arable farming), horticulture or forestry production. Around 20,000 enterprises return useable responses in most years, with a full census held every five years resulting in around 45,000 useable responses. Until 2012, horticulture firms were surveyed every second (even) year.[57]

The survey unit is the "farm" – one or more blocks of land managed as a single operation engaged in agricultural activity. In early years, the APS was surveyed at the "sub-KAU" level – a contiguous block of land (one or more GEOs) operating in the same industry ("kind of activity"). APS was the only survey using sub-KAUs and has recently shifted to surveying at the GEO level. Where a farm is operated by someone other than the owner (eg, a sharemilker) the questionnaire is intended to be completed by the owner, with input from the operator as required.

Tips and tricks

The APS survey form has undergone multiple revisions over the period from 2002-2013. Documentation of the line items collected in each year has been prepared by Richard Fabling and is available in the shared metadata folder in the Datalab (Fabling, 2015).[58] The APS fact table compiles all the consistently measured variables over time. Other variables are available in survey year-specific load tables.

The LBD-derived flag *ibd_imputed* indicates an imputed observation. Similarly, *ibd_exited* indicates farms deemed to have left the population.

2.13 International Trade in Services and Royalties Survey/Census (ITSS)

Purpose

The purpose of the ITSS is to collect information on international trade in selected services and royalties. This information is used in compiling Balance of Payments (BoP) statistics, which are a record of New Zealand's international economic transactions, and to inform New Zealand's trade negotiations.

[57] From 2012 onwards, one observation of horticulture firms has been dropped from the five-year survey cycle, resulting in the following pattern of surveying: 2012 – Full Census; 2013 – Survey, excluding horticulture; 2014 – Census of all horticulture units, survey of other units; 2015 – Survey, excluding horticulture; 2016 – Survey, excluding horticulture; 2017 – Full Census (start of new survey cycle).

[58] Potential researchers who are not in the Datalab can email the author directly for a copy.

Content

The ITSS provides information on the value of imports and exports of commercial services, and payments made or received abroad for royalties.[59] The ITSS tables included in the LBD are compiled from the quarterly ITSS, which captures the bulk of the value of international services trade, and an ad hoc census, run in 1999, 2005 and 2011, which informs the non-sampled estimate of commercial services and royalties trade for BoP statistics. Both sources provide information on partner country and service type, with the census including additional questions on mode of supply.[60] As at 2015, the census component of the ITSS has been discontinued.

Coverage

As the focus of the ITSS collection is the production of aggregate values for National Accounts purposes, the sampling methodology is focused on high-value traders. The quarterly survey is designed to capture 95 percent of all commercial services and royalties transactions within the BoP framework. The remaining 5 percent is estimated based on the census.

In the case of enterprise groups linked by parent-subsidiary relationships, the survey is intended to be completed by the New Zealand head office (the top link in the ownership chain located in New Zealand). However, in some cases business units within the group complete the survey separately. In either case, the response is linked to the enterprise which provided the data to Statistics NZ.

Tips and tricks

The compilation of BoP statistics follows international guidelines produced by the International Monetary Fund in the form of the Balance of Payments and International Investment Position Manual (BPM). To date, the BoP data included in the LBD has been based on the BPM5 manual, issued in 1993. The 2015 update of the LBD is the first to incorporate changes to the data collection and processing methodologies to reflect the revised BPM6 guidelines, published in 2009.

The ITSS census in June 2011 was the first survey form to reflect the new guidelines, which were adopted in the quarterly surveys from September 2011 onwards. The first published data using the new BPM6 definitions was released in June 2014. As the underlying data collected in the ITSS is sufficiently detailed to support both sets of guidelines, Statistics NZ is able to backcast the Quarterly ITSS data under the new definitions using the unit record data collected under the BPM5 methodology. The 2015 update to the database strips out data from the censuses, as this collection has been discontinued and Statistics

[59] The ITSS data is supplemented with a range of other data sources to give the full services trade estimates published by Statistics NZ.

[60] Mode of supply options include cross-border supply (where the work is performed in NZ and delivered to a customer overseas), presence of natural persons (where the work is performed by a New Zealand employee working in a foreign country), and consumption abroad (where the customer travels to New Zealand to take delivery of the service).

NZ does not have a complete concordance between the BPM5 and BPM6 definitions for use with the census.[61]

The sampling structure of the BoP surveys makes it difficult to compare goods and services exporters. Goods trade is based on comprehensive merchandise trade data collected at the firm level (though with allocation issues as detailed in section 2.8), while services data is collected at the group level and relies on a secondary source (AFUS/BRUS) with a much higher threshold for identification of activity.

2.14 Quarterly and Annual International Investment Surveys (QIIS/AIIS)

Purpose

The QIIS and AIIS surveys are used to estimate the International Investment Position (IIP) and Balance of Payments (BoP). The IIP is New Zealand's balance sheet with the rest of the world, and includes measures of New Zealand's overseas debt. The BoP measure New Zealand's business transactions with the rest of the world. QIIS and AIIS are used in estimates of investment income earned and paid (the current account),[62] and transactions in the economy's international equity assets and liabilities, and international borrowing and lending (the financial account).

Content

QIIS and AIIS are the primary source of data for BoP statistics on New Zealand firms' international investment transactions and positions (balance sheet). Information on assets and liabilities (values of assets and liabilities held, and international borrowing and lending) is captured in both the annual and quarterly surveys, while the QIIS also provides details on the composition of changes in investment values (financial accounts, market price revaluations, exchange rate adjustments and other changes) and flows of investment income and expenditure (profits, interest, dividends).

The QIIS collects data on the breakdown of foreign investment in New Zealand and New Zealand's investment abroad by country, type of instrument, maturity, currency, and by type of asset or liability. The detailed quarterly data is held in *load_bop_comp_answer_QIIS*. LBD fact tables aggregate the quarterly data to an annual frequency – selecting the closing quarter for stock values and summing up reported flow data.[63] The load tables provide country-level data[64] as well as including QIIS and AIIS data based on a range of different classification methods.

[61] The historical census data will still be available (under BPM5) through archived versions of the LBD.

[62] The other major component of the current account is trade in goods and services, captured by the ITSS, merchandise trade data, and a range of other sector- or product-specific collections.

[63] Stocks are NZIIP5CLB categories starting with an A; flows are NZIIP5CLB categories beginning with a C.

[64] Country names are available by linking the load tables to *load_bop_category* on *country_code=cat_code* and *classfn_nbr=2368.*

Coverage

The population of interest for the QIIS and AIIS surveys is those New Zealand businesses located in New Zealand that: are wholly or partly foreign owned, including New Zealand branches of overseas businesses; have ownership interests in businesses located overseas, including branches; and/or have financial asset and liability positions with overseas residents. These businesses are primarily identified from the BF/BR, based on BoP indicator questions from AFUS/BRUS. Additional firms may also be identified using administrative data or through other information sources.

The sampling strategy for QIIS is similar to that for the ITSS. The population covers approximately 4,000 firms with known international investment links, including banks, corporate enterprises, the Reserve Bank and the Treasury. Of these, approximately 500 firms are selected quarterly, including all New Zealand located banks and all other businesses with significant international balance sheets. A response rate of 100 percent of "key" and 80 percent of "non-key" enterprises is required, with around 600 to 700 firms providing at least one useable response each year. This sample captures approximately 95 percent of the total value of New Zealand's international financial asset and liabilities. The remaining 5 percent is estimated through the annual AIIS, which is sent to a stratified sample of other firms that have large balance sheets. Every three years the AIIS is run as a census of all the firms which are known to have international assets or liabilities but which have not received the QIIS. In non-census years, non-sample estimates are generated by Statistics NZ and attached to a "dummy company."[65]

Like the ITSS, the QIIS and AIIS are sampled at the group level, rather than the individual enterprise. The survey goes to the New Zealand group top enterprise – the highest enterprise within the group which has a geographic unit located in New Zealand – and that firm is expected to respond on behalf of all its subsidiaries, including those located offshore. This survey method provides the extensive detail required to produce the IIP and BoP aggregates efficiently, reducing both processing costs and respondent burden, as well as eliminating the potential for double-counting of assets and liabilities by multiple members of an enterprise group.

Tips and tricks

The sampling methodology for AIIS/QIIS creates challenges for micro-economic research. First, any analysis using linked data from other sources must take account of the grouping structures in the data, including the potential for restructuring or mergers and acquisitions to affect the reporting unit and/or the coverage of the response. Second, as the sampling method is focused on firms which are known to have substantial international assets or liabilities, it cannot be easily used to examine transitions into and out of international investment activity. Finally, due to the small number of firms with international investments, the level of detail which can be reported is limited by confidentiality considerations.

Fact table information is aggregated to the reporting enterprise level, with partner-country level information available in the load tables. Load tables are also used to pro-

[65] This dummy company is excluded from the fact table data, and the associated enterprise number does not appear on the LBF.

vide reference information (eg, in *load_bop_answer_source_type*, and *load_bop_category*). Descriptions of the asset and liability types captured in the fact tables can be found in *ref_bop_comp_description*.

Load table variables identify the source of AIIS/QIIS data in *ans_srce_type_code*. This indicates whether a given response is an actual survey response (SUR), a simple formula-based derivation based on actual survey responses (FOR), an estimate provided by the respondent (EST - eg for new companies where financial data is not yet available), a modeled or imputed item (MOD, IMP), a weighted survey estimate (WES), an estimate (BPE) or adjustment (BPA) made by analysts in the BoP team based on other information such as financial accounts, or from ad hoc BoP surveys (BPH). Formula-based items are calculated from both imputed and actual responses, with the formula response code superceding the other two codes. Fact tables identify analyst estimates (BPE) and imputed items (IMP) as being imputed.

In fact tables, quarterly reference periods are assigned to financial years to maximise the overlap of the flows data. This is only inexact for firms with rare balance dates, as the most common balance dates (March, June) align with QIIS reference period ends.

For each form of investment, fact tables separately identify assets and liabilities. Although in principle a "negative asset" is a liability, the method of capturing assets and liabilities in the survey allows for both negative assets and positive liabilities to exist. These occur for a number of reasons: (1) if accumulated losses are greater than total equity, this will show up as a negative asset value; (2) equity capital is distinguished from other capital. The latter includes permanent loans from parents and subsidiaries. If a firm has loans from its parents but also lends to its overseas subsidiaries, this can show up as a negative liability; (3) firms can hold both reserve assets and reserve liabilities which, when summed, can result in negative values. Negative assets and positive liabilities are reasonably common in the unit record data, and even occur in the aggregate published data.

Load tables contain data based on multiple different classification methods (*method_classfn*). These are often duplicates or derivations of each other, which enable BoP aggregates to be calculated easily according to different reporting systems. The fact tables restrict attention to a single classification system - NZIIP5CLB (New Zealand International Investment Position Manual 5 Closing Balances).[66]

Output version numbers identify where the data for a certain BoP output (a quarter by classification method) has been revised. For example, *output_ver_nbr*=1 indicates that the data is from the initial published version, while larger numbers indicate later revisions. *Answer_ver_nbr* has the same purpose but indicates the number of edits to the data, rather than the number of published outputs.

2.15 Other sample surveys (BPS, INN, BFS, MEUS)

The Business Practices Survey (BPS) 2001 was a one-off, economy-wide survey of firms' business practices, including the use of information technology, adoption of specific management practices, and innovation outcomes. It has been superceded by the Business

[66] From the 2015 update onwards, this will be replaced by the updated Manual 6.

Operations Survey and the four-yearly Business Practices Module C (2005, 2009, 2013).

The Innovation Survey (INN) 2003 was conducted to provide information on the characteristics of innovation activity of private sector New Zealand businesses, including levels of firm innovation, collaboration across firms and between firms and institutions for the purposes of innovation, factors which affect firms' ability to innovate, and the outcomes of innovation. The survey was designed to collect innovation data in a form that met OECD guidelines described in the Oslo manual. A two-way stratified sample (on industry and firm size) was used, restricting to firms with 10 or more employees. The Innovation Survey has been superceded by the biennial BOS innovation module B.

The Business Finance Survey (BFS) 2004 was a one-off survey of micro to medium-sized businesses (1-500 employees), including a mix of quantitative questions about balance sheet structure and qualitative information about the availability of finance. It was developed by the Ministry of Economic Development and Statistics NZ, and surveyed selected industries which were considered to be relevant for policy purposes and for which it was deemed that there was not currently a suitable source of information on finance. Since 2005, qualitative questions on the success of firms seeking debt and equity financing have been included in the annual BOS Module A. An expanded set of questions on finance issues is included in the 2010 and 2014 BOS, based largely on questions originally asked in the BFS.

The Manufacturing Energy Use Survey (MEUS) 2006 is a precursor to the New Zealand Energy Use Survey (NZEUS). MEUS contains energy use broken down by type (eg, oil, electricity, etc). NZEUS has rolling industry coverage, with the primary sector, services sector, and industrial and trade sector surveyed over a three year period.[67] Only the 2006 survey has been linked to the LBD, and the data remains in a load table (*load_eus_enterprise_2006_mj*).

2.16 Reference tables and programmability

Finally, Statistics NZ have added a range of tables to the database to improve the useability of the data sources above. These include reference tables which detail classification systems used in the data tables (eg, *ref_anzsic06* and *ref_business_type96*), geographic information (eg, *ref_meshblock*,and *ref_meshblock_pair*) and survey-specific reference material (eg, *ref_bop_comp_description*, *ref_energy_conversion* and *ref_domain_value*).

Three tables have been included to simplify the use of month and date variables:

1. *dim_bal_date_year* shows, for every combination of *balance_month_nbr* and *dim_year_key*, which calendar months are included in that financial year
2. *dim_mar_year* provides start and end information for financial years ending in March
3. *dim_month* provides alternative formatting options for every month, including a *month_seq_nbr* which starts in January 1988 and increases by one for every month

Within the database, under *Programmability* \ *Functions* \ *Scalar-valued functions*, there are also a set of programming functions which can be used to manipulate the dates

[67] While MEUS was surveyed at the enterprise level, NZEUS is surveyed at the KAU level. If a firm has KAUs in multiple sectors, those KAUs will be surveyed in different years.

recorded in the YYYYMM format used throughout the LBD, and to convert numbers to string format. Particularly useful are *fn_month_get_next*(month,offset) which counts forward or backward a specified number of months, and *fn_month_span*(start month,end month) which returns the number of months between two specified dates (inclusive of the end months).

Summary: Basic firm information

In many cases, the LBD contains multiple different data sources which capture similar information. The table below provides a brief indication of where to look for various sources of basic firm data. Where there is a clearly preferred data source (eg, the EMS for employment data) other data sources are not mentioned. However, where there are multiple sources which provide useful information (eg, foreign ownership variables) each of these sources is mentioned. Data on international trade, innovation and R&D are excluded from this table as they are covered in more detail on pages 41 and 42.

Data type	Source	Notes
Industry	LBF	*load_lbf_fact_enterprise* and *load_lbf_fact_business* includes ENT- and PBN-level industry codes, respectively, and record the timing of changes in industry. The enterprise table also includes business type and sector
Location	LBF	*load_lbf_fact_business* provides meshblock-level location information for PBNs
Birth date	LBF	Recorded birth date not always consistent with observed activity (see section 2.1)
Employment, wages	Employer Monthly Survey	PENT-level measures of headcount and FTE employment, and working proprietor input are available in *ibuldd_research_datalab* tables *pent_year_L_IDI_YYYYMMDD* and *pent_pbn_month_L_IDI_YYYYMMDD*. The construction of these tables is described in detail in Fabling & Maré (2015a)
Sales and purchases	GST,BAI,AES,IR10	
Gross output, intermediate consumption, value-added	AES,IR10	Key variables used to measure productivity are available in *ibuldd_research_datalab* tables *pent_prod_IDI_YYYYMMDD*, constructed as described in Fabling & Maré (2015b).
Profit	AES,IR10,IR4	
Capital stock	AES,IR10	Closing book values by asset type, including intangible assets
Capital investment	AES	Includes data on additions, disposals, revaluations and depreciation
Foreign ownership	LBF,IR4,BOS,AIIS/QIIS	Definitions and coverage differ across sources, see Sanderson (2013) for further details
Ownership of a foreign subsidiary	LBF,BOS,AIIS/QIIS	

Summary: Trade and international engagement

The LBD provides a number of indicators of firms' international activities, with varying degree of coverage, detail and definitions. This section outlines the main sources of international trade data – BOS, OMT, ITSS, LBF, and GST/BAI. There are also four potential sources for International Investment data – the LBF, BOS, AIIS/QIIS and IR4. These are discussed in detail in Sanderson (2013).

The table below summarises the coverage of the various sources of international trade data. The annual BOS Module A asks three questions about firms' export activities (percentage of sales from exports, percentage of sales from tourism, and whether the firm entered any new export markets). In addition, Module C of the 2007, 2011 and 2015 BOS surveys ask a wide range of questions on firms' international engagement, organised into sections on overseas revenue, overseas production and overseas purchasing (2007 and 2011) or overseas sales of goods and services, workforce and/or offices overseas, use of goods and services sourced overseas and assistance with international engagement (2015). The export questions in Module A combine exports of goods and services, while those in Module C ask firms to identify the sources of their overseas revenue (eg, manufactured goods, services) before asking a range of additional questions which refer to all sources of overseas income. The 2010 Module C on Pricing also includes some export-specific questions, related to firms' export pricing strategies. Two precursors to BOS – the Business Practices Survey 2001 and the Innovation Survey 2003 – also include questions on export sales as a proportion of revenue.

Overseas Merchandise Trade data provide comprehensive, detailed information about goods imports and exports, but does not cover services trade.

The ITSS provides detailed information about services imports and exports, but only for a selected sample of firms which are known to have significant services trade.

Zero-rated GST sales from the GST/BAI data has sometimes been used as an indicator of exporting. These data are comprehensive, in that in principle they capture all firms and all types of export sales, but the indicator is noisy as there are other reasons why income may be zero-rated for GST purposes.

Finally, the LBF includes a range of indicators of international trade, collected in order to establish whether a firm is in scope for the main BoP surveys. This includes questions on merchanting and contract manufacturing, as well as indicators of whether the firm had transactions worth over $20K for commercial services activity or intellectual property with any overseas firm (with purchases and sales combined as a single variable). Specific questions asked have varied over time. As with all LBF data, these indicators are collected annually only for a subset of the population, with historical values remaining in place until new information becomes available.

		BOS A	BOS C	OMT	ITSS	GST/BAI	LBF	BPS	INN
Coverage	Selected				✓		✓		
	Representative	✓	✓					✓	✓
	Comprehensive			✓		✓			
Type of trade	Goods		✓	✓					
	Services		✓		✓		(d)		
	Mixed	✓	✓			✓		✓	✓
Periodicity	Ad hoc/one-off		(a)					✓	✓
	Annual	✓					(c)		
	Sub-annual			✓	✓	✓			
Reporting unit	Enterprise	✓	✓	(b)		✓	✓	✓	✓
	Group				✓				

(a) BOS International Engagement modules have been run in 2007, 2011 and 2015. (b) For OMT data, the reporting firm within an enterprise group may not be the producer. (c) LBF data is recorded as annual but is updated annually only for large or complex firms. (d) Trade variables in the LBF indicate that the firm has international trade but do not distinguish exports from imports.

Summary: R&D and innovation

A range of datasets provide information on firms' R&D expenditures and innovation outcomes. The main sources of R&D and innovation data are outlined briefly below. In addition, as the standard definition of innovation includes process, marketing and organisational innovations, measures of innovation can also be constructed using panel data from surveys and administrative data. For example, Fabling & Grimes (2014) identify changes in HR practices using the 2001 BPS and 2005 BOS surveys (an organisational innovation) while Fabling & Sanderson (2010) identify entry into new markets and the introduction of new export products using OMT data (potentially a marketing/product innovation).

Detailed data on R&D expenditures are available from the R&D survey, including a breakdown of the expenditure by funding source and purpose. These data are collected primarily from known R&D performers. Less detailed, but more representative (of the population as a whole), data are available from the BOS, in which firms are asked about their R&D expenditure and the proportion which was carried out in-house. Both the Business Practices Survey 2001 and the Innovation Survey 2003 also have indicators of innovation expenditures, with the BPS asking a wider question about expenditure on innovation while the Innovation survey includes the more standard questions on R&D expenditure as well as a question on expenditure to develop new or significantly improved products, processes or services. IR10 data has greater coverage, but tends to be selectively completed by smaller and medium sized firms and does not accurately capture total R&D spend.

The expenditure numbers do not, in general, match up across sources, reflecting differences in the calculation of R&D expenditure. While the BOS and R&D survey both measure R&D expenditure, IR10s exclude wages and salaries for R&D staff, which the R&D survey shows account for around 70 percent of all R&D expenditure. There are also differences across surveys in the instructions provided for what should be included in certain types of R&D expenditure. For example, BOS specifies that remuneration of subcontractors who work in-house should be counted as in-house R&D, while the Innovation survey only includes expenditure on R&D carried out "within your business for developing new or significantly improved products, processes or services". GAP provides information about the value of a range of government R&D grants, but not the total value of R&D performed by the firm.

Data on innovation outcomes are available from the BOS, the Innovation Survey, and Intellectual Property Office information. BOS Module A provides a single indicator of innovation – whether the firm developed or introduced any new or significantly improved products, operational processes, organisational or managerial processes or marketing methods (one-year). The biennial innovation module expands on this measure, separating out different types of innovation (two-year) and including information on the degree of novelty of product innovations as well as the activities and cooperative arrangements undertaken to support innovation, sources of ideas and information, and reasons for innovation.

Counts of patents and trademarks applied for and granted are available in the IPO dataset, with the BOS innovation module and its predecessor surveys also including an indicator question for various methods of intellectual property protection.

		BOS A	BOS B	R&D	IR10	IPO	INN	BPS	GAP
Coverage	Selected			✓	(a)				
	Representative	✓	✓				✓	✓	
	Comprehensive					✓			✓
Type of Activity	Input measures	✓	✓	✓	(b)		✓	✓	✓
	Output measures	✓	✓				✓	✓	
	IP protection		✓			✓	✓	✓	
Periodicity	Ad hoc/one-off					(c)	✓	✓	(c)
	Biennial		✓	✓					
	Annual	✓			✓				

(a) Although IR10s are not selected on the basis of expected R&D or innovation activity, they are more often filed by small and medium sized firms, whose R&D propensity tends to be low. (b) IR10s include R&D as an expense item, but total wages and salaries (a key component of R&D spending) are itemised separately. (c) While IPO and GAP data are collected continuously, linking to the LBD has been done on an ad hoc basis.

3 Population coverage and overlap

A key benefit of the LBD is the ability to use a range of data from different sources simultaneously. However, the feasibility of many possible research questions depends on sample size and the degree of overlap between the different sources. This section briefly outlines the coverage of each of the individual datasets, then examines the joint coverage across pairs of datasets.

Table 11 reports the base population – active, private-for-profit businesses – in each year. The total number varies from 440,000 to 512,000, of which between 310,000 and 350,000 have some labour input (employees, working proprietors, or both). The LBF contains, in addition, between 50,000 and 75,000 active enterprises which are not classed as private-for-profit (eg, government agencies, universities) and a further 7-8 hundred thousand firms per year which are classed as inactive, as they have no observed labour input or GST sales or purchases. This is an artefact of the rectangular nature of the table (ie, each firm is represented in each year), coupled with substantial firm turnover (at least for small firms).

Table 12 identifies the proportion of the population from table 11 covered by each data source, by firm size. Figures for the 2012 year are reported for all available data sources. Where there have been substantial changes in coverage over time (eg, due to changes in the population (IR4) or a rolling sample/census approach (AIIS, APS), additional years are presented for comparison. IR4 and APS data are presented both as a share of the full population and as a share of the target group (companies, and agricultural firms, respectively). Group level collections (ITSS, QIIS, AIIS) report coverage of filing units (PENTs), not coverage allowing for group structures (which would yield a higher coverage rate, particularly for employment-based measures). Variation in coverage across datasets and firm sizes reflects both real factors (eg, import and export propensity is higher among larger firms) and sampling (eg, BOS is restricted to firms with 6 or more rolling mean employment). In addition to the data sources within the LBD, table 12 also reports coverage rates for productivity data (PROD), constructed as described in Fabling & Maré (2015b) and available as table *pent_prod_IDI_YYYYMMDD* in *ibuldd_research_datalab*.

Table 12 shows the difference in coverage rates by firm size between the two main sources of performance data – AES and IR10. While coverage increases strongly with firm size for AES reflecting higher sampling rates for large firms, the coverage of IR10s is relatively stable across small and medium sized firms then drops off for larger firms, which are more likely to supply IR with a copy of their accounts rather than submitting an IR10 form. The productivity data uses a combination of AES and IR10 data, but has lower coverage because firms are excluded from the productivity data if they do not have positive values for the key productivity components (gross output, intermediate consumption, labour and capital services) or are in industries which are not part of the "measured sector."[68]

IR4 coverage is consistently high (at least 75 percent) among the target population, with small and medium-sized companies having the highest coverage rates. While coverage rates as a share of the target population have been fairly stable over time (not shown), the

[68] Significant industry restrictions include education, health services, community services, property development and real estate, and central and local government administration. The government sector is also excluded by the private-for-profit criteria.

share of the total population filing IR4s has increased substantially since 2000 reflecting higher rates of incorporation, particularly among small firms.

Tables 13-16 look at the share of firms for which data is available from each specified data source, conditional on being available in another specified data source. Table 13 pools all firms together, including those with no employees, while tables 14-16 report the same statistics for small, medium and large employment firms. Data are presented for the 2012 year, for active, private-for-profit permanent enterprises. Each row reports the total number of PENTs observed in that dataset-year, then identifies the proportion of that number that are covered by each of the relevant datasets referenced in the column headers. For example, of the 987 firms with GAP data in 2012, 30 percent also have AES data and 54 percent have IR10 data (table 13, row 5).

These tables show the strong overlap between the main survey sources which support National Accounts and BoP statistics – AES, AIIS, QIIS, ITSS – reflecting a sampling strategy which focuses on firms which contribute most to relevant aggregates, and the tendency for international trade and investment to be concentrated among larger firms. As BAI data are almost full coverage for active firms (partly as a reflection of the definition of active, which relies on GST sales and purchases as well as labour input), all other data sources show very strong overlap with BAI (BAI column), while the BAI row largely reflects the overall coverage of the other datasets (as shown in table 12). Overlap between APS and AES is very low, reflecting the exclusion of the main primary sector industries from AES. Most cases of overlap between these two sources relate to enterprises which have both agricultural and non-agricultural production units in the relevant year. Most data sources exhibit reasonable overlap with the productivity dataset (generally over 60 percent among employing firms). While this provides opportunities to compare performance across a wide range of firm characteristics, interpretation still relies on an understanding of the sampling and coverage of each of the component datasets.

4 Protocols

As can be seen from the breadth of agencies contributing unit-record data, the New Zealand legislative environment presents few obstacles to supplying data to Statistics NZ, provided there is an end goal of producing official statistics (including research). This section covers the legal basis for accessing the LBD, and the practical processes of applying for research access and for adding new data sources to the database.

4.1 Legal framework

Access to the LBD (and other data) is granted at the discretion of the Government Statistician in accordance with the Statistics Act 1975. Section 37 of the Act conveys the authority to grant access (under controlled conditions) for research purposes. In 2012, the Act was amended to enable access for researchers outside government departments.

> **Section 37C Disclosure of individual schedules for bona fide research or statistical purposes**
> (1) . . . the Statistician may disclose individual schedules to any person if – (a) the information contained in the schedules is to be used by that person solely for bona fide research or statistical purposes in relation to a matter of public interest; and (b) the Statistician is satisfied that the person has the necessary research experience, knowledge, and skills to access and use the information contained in the schedules . . .

Because the LBD includes filed tax returns, the privacy and confidentiality provisions of the Tax Administration Act 1994 also apply to researchers using the data, and IR is consulted regarding who accesses the database. Researchers sign declarations to the effect that they have read and understood the relevant provisions of both acts. Access to IR business data and, therefore, the LBD more generally, is currently restricted to researchers working for, or on behalf of, the New Zealand Government.

4.2 Access

The starting point for accessing the LBD is a Datalab application outlining the data required, the research question, and the reasons why there is no alternative to using microdata. If approved, researchers access an anonymised version of the LBD at one of Statistics NZ's three offices (Wellington, Auckland or Christchurch). Organisations can also apply for remote access, enabling researchers to access the data through a remote desktop from their home institution. In order to be approved for remote access, the organisation must show evidence that they can provide a secure environment for data access. This includes both the physical environment and the IT systems, as well as the standard requirements for appropriate use required of microdata researchers. Remote access has been established for government agencies, universities, and private research institutes.

Results of research are subject to confidentiality provisions designed to protect individual businesses from identification.[69] The methods used to confidentialise outputs are blunt tools, since they must be effective across an unknowable range of situations. Researchers apply these confidentiality rules to their own results and then must demonstrate to Statistics NZ checkers that the rules have been correctly applied. The turn-around time for checking outputs ("Phase 1" checking) is a maximum of three working days, but is often less in practice.[70]

After results have been written up into a paper or other format intended for public release (eg, presentations, reports to Ministers), Statistics NZ require the completed work to be returned to them for a final check ("Phase 2" checking) to confirm that the researcher has not inadvertently breached any of the rules through, eg, their description of their method or interpretation of their results, or through combining outputs in a way that allows confidential information to be inferred. The phase 2 checking process can take up

[69] Code written by researchers within Statistics NZ's offices must also be submitted for checking if it is to be taken outside the secure environment.

[70] Researchers with significant experience using and clearing Statistics NZ microdata may be appointed an Accredited Researcher by the Government Statistician. Accredited Researchers are permitted to self-release Phase 1 outputs for specified datasets, including the LBD (after applying the relevant confidentiality provisions). However, IR data is not subject to this expedited process and must still be submitted for checking by an employee of Statistics NZ.

to ten working days. Papers must display a disclaimer explaining how the principles of the legislation have been met and disclaiming Statistics NZ from the analysis (as in page iii of this paper).

With the support of domestic government agencies and Statistics NZ it has been possible for overseas-based researchers to work with the data though unit record data is accessible only within New Zealand.[71]

Statistics NZ charges for access to the data on a cost recovery basis which may include a fixed fee for any data preparation or linking, and for confidentiality checking. At the time of writing, the only fees relevant for use of the LBD are a $500 application fee (payable only for successful applications), a charge of $115 per hour for the creation of non-standard datasets (eg, additional data linking), and a fee of $115 per hour for confidentiality checking (beyond the first 15 hours for each project).

4.3 Adding data to the LBD

The primary way to add new firm-level data is via a Datalab application explaining the research need for such data to be linked to the LBD.[72] Examples of data that have been added through this process include the Government Assistance Programme (GAP) data and the trade in services (ITSS) data. Where data integrated in this manner has not been subject to an official first release, Statistics NZ may choose to publish an official statistic prior to allowing research use. Applicants are required to cover the cost of new data integration.

The other mechanism for adding new data to the LBD is through commissioned BOS content (section 2.10 above), since the BOS dataset is updated annually within the database. The point of difference for BOS is the ability to design content in the knowledge that additional longitudinal data can be accessed via the LBD (particularly quantitative performance data and other BOS survey years).

Once an agency has successfully applied for additional data to be added to the LBD it becomes accessible to all research users.[73] That is, researchers all have access to the same dynamic database.

Researchers can also apply for access to linked individual and firm-level data through the IDI. Unlike the LBD, access to the IDI is restricted by data provider, and outside a few core tables provided in conjunction with the LBD (which include the Business Register,

[71] There are currently no “general use” business Confidentialised Unit-Record Files (CURFs) in New Zealand. While there are some off-site anonymised and confidentialised datasets, notably unlinked BOS survey data, these are held under tight security for the sole use of the funding government departments. Unconfidentialised unit record data only leaves Statistics NZ in limited circumstances: for BOS, when explicit consent for a follow-up study has been given by the respondent; and where the survey has been run jointly with another agency (for example, the R&D Survey is jointly run by Statistics NZ and the Ministry of Business, Innovation and Employment). That data remains subject to the provisions of the Statistics Act and Statistics NZ’s confidentiality rules.

[72] This section covers the integration of additional microdata. There are no issues with researchers bringing in macro data sources (eg, deflators or trade partner country characteristics) or publicly available regional information (eg, weather data or Census aggregates) and linking that data directly themselves.

[73] Subject to the constraint that the work using the data fits with the research question outlined in the researcher’s own Datalab application.

IR data and a table of basic personal characteristics), researchers must provide a rationale for access to each “schema” that they require (eg, health data, migration data). Access to additional schema must, generally, be approved by the data provider as well as by Statistics NZ.

5 Conclusions

While the LBD presents a huge opportunity for researchers looking to understand the function and performance of firms in the New Zealand economy, these data are not without their challenges. In particular, both the content and the coverage reflect the fact that the database is constructed from data collected by several different agencies for a range of different purposes. This guide has set out some of the major characteristics of the data which need to be accommodated when using the LBD for research purposes.

In addition to identifying dataset-specific idiosyncracies, we have also looked across the database as a whole, with the view to providing guidance for new users on where to look for specific types of data. In particular, we have focused on the types of variables that have attracted substantial research attention over the past 8 years of the LBD – basic firm characteristics, innovation and international engagement. This focus reflects use of the data to date, but does not imply that these are the areas where most use can be made of the data in future. As researchers from a wider range of backgrounds come to make use of the data, knowledge and understanding of both the data itself and the firms which produce it are expected to widen.

The discussion in this paper reflects the LBD in its current state. As changes are made to the collection and processing of the underlying data, as part of the ongoing operation of the collection agencies, new challenges and opportunities will emerge. In addition to annual updates, there is scope to expand the number of components in the database. Statistics NZ hold a number of unlinked business datasets that could form the basis of interesting research projects. While not an exhaustive list, these data include the Quarterly Employment Survey (providing a survey measure of hours paid); input and output producer price data; and a number of industry-specific surveys (eg, Retail and Wholesale Trade Surveys, and the Accommodation Survey). Currently, the integration of any of these data would require government agencies to satisfy the Datalab process outlined above and to fund the integration work.

References

Abowd, J., Creecy, R., & Kramarz, F. (2002). Computing person and firm effects using linked longitudinal employer-employee data. LEHD Program Technical Papers TP-2002-06, US Census Bureau.

de Beer, Y., Greet, P., & Morris, M. (2010). Business participation in government assistance programmes. Technical report, Statistics New Zealand, Wellington.

Fabling, R. (2009). A rough guide to New Zealand's Longitudinal Business Database. Global COE Hi-Stat Discussion Papers No. 103, Institute of Economic Research, Hitotsubashi University.

Fabling, R. (2011). Keeping it together: Tracking firms in New Zealand's Longitudinal Business Database. Working Paper 11-01, Motu Economic and Public Policy Research.

Fabling, R. (2015). APS variable concordance (extended, 2002-2014). Mimeo, Statistics New Zealand.

Fabling, R. & Grimes, A. (2014). The "suite" smell of success: Complementary personnel practices and firm performance. *ILR Review*, *67*(4), 1095–1126.

Fabling, R. & Maré, D. (2015a). Addressing the absence of hours information in linked employer-employee data. Working Paper 15-17, Motu Economic and Public Policy Research.

Fabling, R. & Maré, D. (2015b). Production function estimation using New Zealand's Longitudinal Business Database. Working Paper 15-15, Motu Economic and Public Policy Research.

Fabling, R. & Sanderson, L. (2010). Entrepreneurship and aggregate merchandise trade growth in New Zealand. *Journal of International Entrepreneurship*, *8*(2), 182–199.

Fabling, R. & Sanderson, L. (2013). Exporting and firm performance: Market entry, investment and expansion. *Journal of International Economics*, *89*(2), 422–431.

Maré, D., Sanderson, L., & Fabling, R. (2014). Earnings and employment in foreign-owned firms. Working paper 14-16, New Zealand Treasury.

Sanderson, L. (2013). Sources of international investment data in the Longitudinal Business Database. Working Papers 2013/31, New Zealand Treasury.

Statistics New Zealand (2013). Introduction to the Integrated Data Infrastructure 2013. Technical report, Statistics New Zealand: Wellington.

Statistics New Zealand (2015). Business participation in government assistance programmes: 2013 update. Technical report, Statistics New Zealand: Wellington.

Tables

Table 1: Datasets integrated into the Longitudinal Business Database - December 2014 archive

Acronym	**Component**	**Frequency**	**Availability**	**Source**	**Linking**	**Key population restrictions**	**Updated (to LBD)**
LBF	Longitudinal Business Frame	Monthly	Apr99-Sep13	SNZ		Non-employing firms below GST filing threshold	Annually
AES	Annual Enterprise Survey	Financial year	1999-2012	SNZ		Stratified random sample; IR10s used for small firms	Annually
IR10	Tax-filed financial accounts (IR10)	Financial year	1999-2012	IR	Tax id		Annually
BAI	Business Activity Indicator	Monthly	Apr92-Sep13	IR	Tax id	Mandatory filing threshold	Annually
GST	Goods and Services Tax (GST 101 form)	1, 2 or 6-monthly	Apr92-Dec07*	IR	Tax id	Mandatory filing threshold	Annually
LEED	Linked Employer-Employee Data	Monthly	Apr99-Mar14	IR	Tax id	See Fabling and Maré (2015)	
IR4	Company tax return (IR4)	Financial year	1999-2012	IR	Tax id	Registered companies only	Annually
GAP	Government Assistance Programmes	Various	2000-2013	Various	Matching		Ad-hoc
OMT	Overseas Merchandise Trade	Daily	Oct88-Sep14	Customs	Tax id; matching	Mandatory filing threshold ($1K); match quality deteriorates pre-1996	Annually
IPO	Intellectual Property Office	Annual	1996-2010	IPONZ	Matching		Ad-hoc
BOS	Business Operations Survey	Financial year	2005-2013	SNZ		Stratified random sample; 6+ RME	Annually
R&D	Research and Development Survey	Financial year	1996-2012(biennially)	SNZ		Stratified random sample; excludes ”non-R&D” industries	Biennially
APS	Agricultural Production Survey	Financial year	2002-2013	SNZ		Census or stratified random sample; some non-census years exclude horticulture and related questions	Annually
ITSS	International Trade in Services Survey	Quarterly	96Q2-13Q3	SNZ		Approximate census satisfying BOP materiality thresholds	Annually
AIIS	Annual International Investment Survey	Financial year	2001-2012	SNZ		Stratified random sample; firms w/ known international investors/investments; Census in 2003,06,09,13	Annually
QIIS	Quarterly International Investment Survey	Quarterly	00Q2-13Q3	SNZ		Stratified random sample; firms w/ known international investors/investments	Annually
BPS	Business Practices Survey	Financial year	2001	SNZ		Stratified random sample; 6+FTE	One-off
INN	Innovation Survey	Financial year	2003	SNZ		Stratified random sample; 10+FTE	One-off
BFS	Business Finance Survey	Financial year	2004	SNZ		Stratified random sample; <100FTE	One-off
MEUS	Manufacturing Energy Use Survey	Financial year	2006	SNZ		Stratified random sample; manufacturing; 10+RME	One-off

Availability of financial year data relates to “notional” March 31st balance dates (eg, the 1999 year is the financial year ending 31st March 1999 or the financial year that has the greatest overlap with that period if the firm’s balance date is not March 31st). FTE=Full-Time Equivalents; RME=Rolling Mean Employment. * Raw GST data has not been updated since the original archived version of the LBD, but is being refreshed in the current update and will be available over the same time period as BAI data in future.

Table 2: Industry restrictions for BRUS Tier 2 sampling

A0201	C170	G425	J562	L661	M700
A0202	*C2499*	G431	J580	L664	N722
A0203	D261	I481	J592	*L6712*	N729
A0510	D264	I490	K622	M691	O751
B070	F349	I510	K624	M692	P810
C113	F360	I521	K632	M693	R911
C119	F372	I529	K641	M694	R920
C121	F373	J551	K642	M696	*S9551*

Firms are selected for 3-yearly BRUS (tier 2) only if they are in one of the industries listed above. Restrictions are generally applied at the three-digit industry level. Items in italic indicate restrictions applied at a more detailed level of aggregation.

Table 3: Estimate of usable Business Frame/Register Update surveys by year

Year	AFUS	MFUS	BRUS	MRUS
2002	91,466	*	N/A	N/A
2003	83,523	*	N/A	N/A
2004	67,975	*	N/A	N/A
2005	67,083	*	N/A	N/A
2006	59,068	*	N/A	N/A
2007	52,655	*	N/A	N/A
2008	44,133	*	N/A	N/A
2009	32,175	10,956	N/A	N/A
2010	25,320	9,918	N/A	N/A
2011	24,915	10,512	N/A	N/A
2012	25,677	10,329	N/A	N/A
2013	21,792	10,587	N/A	N/A
2014	12,138	8,979	N/A	N/A
2015	N/A	N/A	12,543	564

Underlying firm counts random rounded base three in accordance with Statistics NZ confidentiality procedures. Survey response counts estimated at the enterprise level, and relate to a March year end. Counts are approximate for the AFUS because we estimate usable returned responses based on post-out size prior to October 2008, multiplying by a factor of 0.85 (estimated from response rates for later years). This affects the 2002-2009 years (data for the 2000 and 2001 years not available). Counts are approximate for MFUS in the 2009 year because we have coverage of usable responses from October 2008 (and no post-out counts for earlier periods, represented by * in the table), so we multiply by a factor of two to adjust for undercoverage in the 2009 year. Finally, the MRUS survey is sent out four months after the birth of a firm to allow relevant administrative data to be collected first. As a consequence, the raw count of responses received is probably only a half the number expected in a full year of the survey (allowing for a delay in response), so we double our raw count to compensate.

Table 4: Proportion of LBF updates by data source, May 1999-September 2013

	Enterprise		Plant		Industry
	Entry	Exit	Entry	Exit	change
AFUS	0.006	0.043	0.355	0.471	0.113
MFUS	0.005	0.084	0.077	0.017	0.135
APS	0.008	0.021	0.045	0.051	0.324
Other surveys	0.003	0.012	0.110	0.137	0.013
IR	0.968	0.767	0.282	0.150	0.309
Companies Office	0.000	0.048	0.001	0.020	0.001
Other	0.010	0.025	0.130	0.155	0.105
Total	859,404	728,310	47,631	40,341	304,881

Industry change counts include both ANZSIC96 and ANZSIC06 changes. Where there is a simultaneous change in ANZSIC96 and ANZSIC06, each change is given a weight of 0.5. Plant entry and exit figures are restricted to events occurring within continuing firms. Reference period refers to "real world" dates, rather than dates when changes were implemented on the BF.

Table 5: Comparing definitions of activity – "Live" vs "Active" enterprises

Year ended March	Live on LBF	Of which, active	Active	Of which, live
2000	428,259	0.922	440,307	0.896
2001	430,674	0.931	451,281	0.889
2002	431,202	0.931	452,505	0.887
2003	443,298	0.929	461,919	0.892
2004	455,010	0.934	476,505	0.892
2005	473,712	0.929	487,674	0.902
2006	488,691	0.924	496,899	0.909
2007	496,512	0.926	506,724	0.908
2008	501,786	0.928	512,346	0.909
2009	497,013	0.927	507,234	0.908
2010	488,298	0.927	497,943	0.909
2011	480,870	0.933	492,954	0.910
2012	477,153	0.930	486,762	0.912
2013	468,948	0.935	480,471	0.913

Private-for-profit enterprises only. Firm counts random rounded base three in accordance with Statistics NZ confidentiality procedures. "Live" defined as having *life_cycle_code = 'birt'* (birthed) or *'reac'* (reactivated) on the LBF. "Active" defined as having positive labour input (employees or working proprietors) and/or positive GST sales or purchases.

Table 6: Coverage of AES postal component by enterprise

Year	No. of postal observations	Of which, non-imputed	Of which, has financial position
1999	20,454	*	*
2000	20,256	*	*
2001	20,193	15,597	7,740
2002	20,313	15,171	7,875
2003	20,340	15,246	8,109
2004	21,825	16,488	9,303
2005	22,203	16,674	9,780
2006	24,102	17,853	10,164
2007	21,897	16,593	10,320
2008	22,353	16,815	10,704
2009	20,340	15,723	15,723
2010	18,228	13,722	13,722
2011	18,495	13,713	13,713
2012	17,499	13,431	13,431

Firm counts random rounded base three in accordance with Statistics NZ confidentiality procedures. *Imputation flags not available prior to 2001.

Table 7: Fixed asset schedule by form type

IR10		LAND&BUILDINGS					F&F	PLANT&MACH + OTH				VEHICLES				
Intermediate		LAND		BUILDINGS				PME/OTH		COMP						
	AES	land	li	rb	nrb	oc	fandf	pme***	lift	hard	soft**	mv	ships	plane	bus	
Services to agriculture	AC	y			y*		y	y		y	y	y		y		
Construction	BC	y		y	y		y	y		y	y	y				
Building Trades	BT	y			y*		y	y		y	y	y				
Education	ED	y			y*	y	y	y		y	y	y				
Electricity Gas Supply	EL	y			y*	y	y	y		y	y	y				
Fishing	FH	y			y*		y	y		y	y	y	y			
Forestry and Logging	FL	y	y		y*	y		y		y	y	y				
Financial Services	FS	y		y	y		y	y		y	y	y				
Gas Supply	GA	y			y*	y		y		y	y	y				99/00 form
General Insurance	GI	y		y	y		y	y		y	y	y				
Hospitals and Nursing Homes	HP	y		y	y		y	y		y	y	y				
Life and Medical Insurance	LI	y		y	y		y	y		y	y	y				
Medical Services	MD	y			y*		y	y		y	y	y				
Mining and Quarrying	MQ	y			y*	y	y	y		y	y	y				
Manufacturing and Wholesale	MW	y			y*		y	y	y	y	y	y				
Education (incl. PNP)	NE	y			y*	y	y	y		y	y	y				
Hospitals and Nursing Homes (incl. PNP)	NH	y		y	y		y	y		y	y	y				
Medical Services (incl. PNP)	NM	y			y*		y	y		y	y	y				
Services (incl. PNP)	NS	y			y*		y	y		y	y	y				
Oil Companies	OC	y			y*	y	y	y		y	y	y				
Producer Boards	PB	y			y*		y	y		y	y	y				01/02 form
Racing Clubs	RC	y			y*	y	y	y		y	y	y				
Real Estate	RE	y		y	y		y	y		y	y	y				
Retail Trade and Services	RS	y			y*		y	y		y	y	y				01/02 form
Retail Trade	RT	y			y*		y	y		y	y	y				
Services	SV	y			y*		y	y		y	y	y				
TAB/NZ Racing Board	TA	y			y*		y	y		y	y	y				
Transport and Storage	TP	y			y*	y	y	y		y	y	y	y	y	y	

Asset types based on 2008/09 form (or last available for defunct form types)
Survey forms prior to 1998/99 year (96/97 & 97/98) collected different categories so are excluded from analysis

*Buildings assumed "non-residential"
**Respondents are instructed to report hard+soft under hard if they cannot report separately
***The AES pme variable on the LBD includes the "All other fixed assets" category

"Common" form type includes: land; building (or nrb+rb); fandf (ex FL/GA); pme; hard; soft; mv
Form-specific items that affect consistency of variables across form types
(eg, pme excludes lifting equipment in Manu, but not in any other industry)

var	Description#
land	Land
li	Land improvements
rb	Residential buildings
nrb	Non-residential buildings
oc	Other construction
fandf	Furniture and fittings
pme	Other plant, machinery & equipment + All other fixed assets
lift	Lifting and handling equipment
hard	Computer hardware
soft	Computer software
mv	Motor vehicles and other transport equipment
ships	Ships
plane	Aircraft and helicopters
bus	Buses and coaches

#"Includes and excludes" differ by form type

Table 8: Business Operations Survey Contestable Modules, 2005-2016

Year	Topic	Sponsoring agency
2005	Business Practices	MED
2006	Employment Practices	MED/DoL
2007	International Engagement	MED/TSY/NZTE/MFAT
2008	Business Strategy & Skills	MED/DoL/MoRST/TSY
2009	Business Practices	MED
2010	Price & Wage Setting	RBNZ/VUW/UoT
	Financing	MED
2011	International Engagement	MBIE
2012	Regulation	MfE/TSY/MBIE
2013	Business Practices	MBIE
	Skill Needs & Recruitment	MBIE
2014	Skills Acquisition	MBIE
	Business Finance	MBIE
2015	International Engagement	TSY/MBIE/MFAT
2016	Skills Acquisition	MBIE
	Regulation	MBIE

Sponsoring agencies: The New Zealand Treasury (TSY); Ministry of Economic Development (MED); Department of Labour (DoL); New Zealand Trade and Enterprise (NZTE); Ministry of Foreign Affairs and Trade (MFAT; Ministry of Research, Science and Technology (MoRST); Reserve Bank of New Zealand (RBNZ); Victoria University of Wellington (VUW); University of Tasmania (UoT); Ministry for the Environment (MfE). MED, DoL, and MoRST are now part of the Ministry of Business, Innovation and Employment (MBIE).

Table 9: Overlap between BOS (by year) and predecessor surveys

		BPS01	INN03	BFS04	BOS05	BOS06	BOS07	BOS08	BOS09	BOS10	BOS11	BOS12	BOS13
BPS01	2,754	1.000	0.371	0.141	0.525	0.481	0.442	0.422	0.402	0.400	0.368	0.339	0.330
INN03	2,988	0.342	1.000	0.163	0.682	0.615	0.588	0.567	0.545	0.523	0.494	0.438	0.429
BFS04	4,560	0.085	0.107	1.000	0.276	0.214	0.209	0.191	0.189	0.179	0.170	0.143	0.140
BOS05	7,380	0.196	0.276	0.170	1.000	0.643	0.596	0.566	0.531	0.502	0.474	0.367	0.359
BOS06	6,066	0.219	0.303	0.161	0.782	1.000	0.755	0.685	0.638	0.603	0.562	0.444	0.431
BOS07	6,642	0.183	0.265	0.144	0.663	0.690	1.000	0.751	0.685	0.628	0.589	0.477	0.462
BOS08	6,339	0.183	0.267	0.138	0.659	0.655	0.787	1.000	0.815	0.732	0.673	0.548	0.525
BOS09	6,420	0.172	0.254	0.134	0.611	0.603	0.708	0.805	1.000	0.796	0.723	0.591	0.568
BOS10	6,186	0.178	0.253	0.132	0.598	0.591	0.675	0.750	0.826	1.000	0.808	0.656	0.623
BOS11	6,126	0.166	0.241	0.127	0.571	0.556	0.638	0.697	0.758	0.816	1.000	0.715	0.676
BOS12	5,586	0.167	0.234	0.117	0.484	0.482	0.568	0.621	0.679	0.726	0.784	1.000	0.845
BOS13	5,820	0.156	0.220	0.110	0.455	0.449	0.527	0.572	0.626	0.662	0.712	0.811	1.000

Firm counts random rounded base three in accordance with Statistics NZ confidentiality procedures. Reading horizontally, each column reports the share of permanent enterprise numbers which have usable data for the survey referenced in the column heading, conditional on having usable data for the survey referenced in the first row. Surveys are: Business Operations Survey (BOS); Innovation Survey (INN); Business Finance Survey (BFS); Business Practices Survey (BPS). For example, of the 2,988 firms with data for the Innovation Survey 2003, 34.2 percent also have data for BPS01 and 68.2 percent have data for BOS05. For BOS, usable data is defined as observations with response code of R (responded) or F (panel top-up).

Table 10: Panel dimension of BOS (2005-2013)

Usable BOS responses	No. of businesses
1	3,834
2	1,923
3	1,383
4	1,089
5	942
6	945
7	1,383
8	921
9	1,437
Total	**13,857**

Firm counts random rounded base three in accordance with Statistics NZ confidentiality procedures. Permanent enterprise number (PENT) used. Usable data is defined as observations with response code of R (responded) or F (panel top-up).

Table 11: Population of active, private-for-profit businesses

	Number of businesses					**Total FTE employment**		
	Non-employing firms		Employing firms, by FTE			Employing firms, by FTE		
Year	L=0	WP only	(0,10]	(10,50]	$(50,\infty)$	(0,10]	(10,50]	$(50,\infty)$
2000	107,403	195,063	125,136	10,620	2,088	244,400	209,900	433,100
2001	118,497	193,968	125,646	11,034	2,136	247,200	217,900	444,700
2002	123,225	188,364	127,275	11,451	2,193	253,600	225,600	457,400
2003	130,287	186,699	130,722	11,943	2,268	265,100	235,300	470,700
2004	141,138	185,610	134,889	12,504	2,367	275,300	244,200	492,800
2005	150,477	182,988	138,768	12,939	2,505	284,900	252,500	515,300
2006	156,621	182,727	141,669	13,275	2,607	291,600	258,900	539,600
2007	165,711	181,479	143,358	13,521	2,655	295,300	263,800	549,200
2008	170,595	180,252	145,047	13,737	2,715	300,900	268,800	567,500
2009	172,548	176,637	141,867	13,458	2,724	296,000	262,400	572,400
2010	173,526	172,530	136,413	12,855	2,619	286,500	250,700	550,000
2011	171,174	170,949	135,234	12,984	2,616	286,300	254,400	553,200
2012	168,810	167,562	134,586	13,131	2,673	287,000	257,500	561,600
2013	169,806	159,570	134,979	13,365	2,751	290,400	261,800	573,800

Random rounding (base three) and graduated random rounding have been applied to firm counts and FTE counts respectively, in accordance with Statistics NZ confidentiality procedures. Permanent enterprise number (PENT) used. Year is *dim_year_key* (eg, 2000 is the year ending March 2000, labelled 200003 in the LBD). Column 1 reports the number of active (positive GST sales/purchases) firms with no observed labour input. Column 2 reports the number of non-employing firms which have labour input provided by one or more working proprietors (WPs). WPs are excluded from the calculation of firm size in columns 3-8. Columns 6-8 report total FTE employment.

Table 12: Dataset coverage, by firm size

		Number of businesses					**Total FTE employment**		
		Non-employing firms		Employing firms, by FTE			Employing firms, by FTE		
Dataset	Year	Zero L	WP only	(0,10]	(10,50]	(50,∞)	(0,10]	(10,50]	(50,∞)
AES	2012	0.007	0.003	0.020	0.218	0.765	0.040	0.273	0.892
IR10	2012	0.499	0.630	0.684	0.628	0.373	0.692	0.607	0.192
BAI	2012	1.000	0.890	0.974	0.994	0.989	0.988	0.993	0.987
IR4 (all)	2000	0.460	0.085	0.442	0.839	0.815	0.616	0.843	0.685
(all)	2012	0.546	0.251	0.699	0.873	0.745	0.800	0.868	0.541
(companies)	2012	0.768	0.975	0.898	0.915	0.765	0.918	0.906	0.568
GAP	2009	0.006	0.003	0.019	0.059	0.088	0.026	0.066	0.125
	2012	0.001	0.000	0.002	0.021	0.085	0.004	0.027	0.126
OMT - X	2012	0.011	0.005	0.036	0.183	0.403	0.062	0.207	0.538
OMT - M	2012	0.033	0.017	0.088	0.284	0.558	0.130	0.313	0.667
IPO	2012	0.003	0.001	0.008	0.040	0.140	0.013	0.045	0.303
BOS	2012	0.000	0.000	0.011	0.164	0.675	0.031	0.214	0.714
R&D	2012	0.001	0.000	0.007	0.065	0.204	0.016	0.076	0.287
APS (all)	2012	0.048	0.115	0.102	0.036	0.034	0.074	0.035	0.068
(Ag, sample)	2010	0.153	0.213	0.240	0.155	0.100	0.244	0.159	0.180
(Ag, census)	2012	0.397	0.546	0.588	0.439	0.548	0.558	0.453	0.667
ITSS	2012	0.000	0.000	0.002	0.024	0.154	0.004	0.031	0.327
AIIS (census)	2009	0.003	0.000	0.006	0.053	0.133	0.012	0.066	0.100
(sample)	2012	0.001	0.000	0.001	0.013	0.055	0.003	0.017	0.045
QIIS	2012	0.001	0.000	0.000	0.005	0.076	0.001	0.007	0.204
PROD	2012	0.000	0.519	0.623	0.617	0.726	0.633	0.621	0.794

Permanent enterprise number (PENT) used. Year is *dim_year_key* (eg, 2000 the year ending March 2000, labelled 200003 in the LBD), and the 2012 year is reported for all available sources. Where there have been substantial changes in coverage over time (eg, due to changes in the population (IR4) or a rolling sample/census approach (AIIS, APS)), additional years are presented for comparison. IR4 and APS data are presented both as a share of the full population and as a share of a more appropriate target group (companies and agricultural firms, respectively). Group level collections (ITSS, QIIS, AIIS) are analysed at the reporting permanent enterprise level (as opposed to, say, counting all members of a reporting group as being included in the coverage). PROD is the Fabling-Maré productivity dataset. Other acronyms are defined in table 1.

Table 13: Overlap between data sources (2012), all businesses pooled

	Total	AES	IR10	BAI	IR4	GAP	OMT - X	OMT - M	BOS	R&D	APS	ITSS	QIIS	AIIS	PROD
AES	9,294	1.000	0.562	0.990	0.856	0.032	0.199	0.303	0.262	0.082	0.012	0.068	0.035	0.029	0.686
IR10	291,036	0.018	1.000	0.971	0.603	0.002	0.022	0.055	0.010	0.005	0.100	0.001	0.000	0.001	0.612
BAI	464,733	0.020	0.608	1.000	0.513	0.002	0.023	0.054	0.012	0.006	0.090	0.002	0.001	0.001	0.375
IR4	241,830	0.033	0.726	0.986	1.000	0.003	0.038	0.087	0.018	0.009	0.044	0.004	0.001	0.002	0.416
GAP	987	0.298	0.541	1.000	0.827	1.000	0.502	0.541	0.264	0.505	0.033	0.122	0.036	0.046	0.568
OMT - X	10,971	0.168	0.584	0.995	0.842	0.045	1.000	0.740	0.133	0.101	0.025	0.047	0.016	0.029	0.558
OMT - M	25,452	0.111	0.626	0.992	0.826	0.021	0.319	1.000	0.082	0.054	0.017	0.026	0.009	0.015	0.515
BOS	5,415	0.450	0.548	0.989	0.825	0.048	0.269	0.384	1.000	0.172	0.043	0.076	0.034	0.037	0.672
R&D	2,670	0.287	0.543	0.994	0.833	0.187	0.413	0.512	0.349	1.000	0.064	0.106	0.035	0.036	0.626
APS	41,757	0.003	0.700	0.998	0.257	0.001	0.007	0.010	0.006	0.004	1.000	0.000	0.001	0.001	0.550
ITSS	1,110	0.570	0.324	0.992	0.803	0.108	0.468	0.597	0.370	0.254	0.016	1.000	0.149	0.111	0.665
QIIS	432	0.757	0.118	0.979	0.688	0.083	0.410	0.521	0.424	0.215	0.056	0.382	1.000	0.063	0.563
AIIS	633	0.431	0.379	1.000	0.867	0.071	0.498	0.611	0.313	0.152	0.043	0.194	0.043	1.000	0.517
PROD	180,780	0.035	0.985	0.965	0.556	0.003	0.034	0.073	0.020	0.009	0.127	0.004	0.001	0.002	1.000

Firm counts random rounded base three in accordance with Statistics NZ confidentiality procedures. Permanent enterprise number (PENT) used. Year is *dim_year_key* and the 2012 (201203) year is used. PROD is the Fabling-Maré productivity dataset. Other acronyms are defined in table 1 (OMT - X and OMT - M refer to exports and imports respectively). Group level collections (ITSS, QIIS, AIIS) are analysed at the *reporting* permanent enterprise level (as opposed to, say, counting all members of a reporting group as being included in the coverage).

Table 14: Overlap between data sources (2012), businesses with (0,10] FTE

	Total	AES	IR10	BAI	IR4	GAP	OMT - X	OMT - M	BOS	R&D	APS	ITSS	QIIS	AIIS	PROD
AES	2,757	1.000	0.681	0.986	0.931	0.007	0.095	0.181	0.036	0.017	0.005	0.023	0.014	0.015	0.757
IR10	92,112	0.020	1.000	0.985	0.760	0.002	0.034	0.086	0.011	0.007	0.105	0.001	0.000	0.001	0.903
BAI	131,067	0.021	0.692	1.000	0.706	0.002	0.037	0.089	0.011	0.007	0.105	0.002	0.000	0.001	0.631
IR4	94,119	0.027	0.744	0.983	1.000	0.003	0.046	0.109	0.013	0.009	0.061	0.003	0.000	0.002	0.676
GAP	285	0.063	0.600	0.989	0.895	1.000	0.379	0.453	0.032	0.432	0.000	0.042	0.011	0.032	0.589
OMT - X	4,872	0.054	0.642	0.994	0.881	0.022	1.000	0.740	0.031	0.057	0.023	0.017	0.001	0.017	0.636
OMT - M	11,799	0.042	0.670	0.992	0.872	0.011	0.306	1.000	0.023	0.032	0.014	0.010	0.001	0.009	0.645
BOS	1,446	0.068	0.701	0.988	0.876	0.006	0.104	0.189	1.000	0.081	0.052	0.010	0.002	0.006	0.627
R&D	978	0.049	0.644	0.994	0.880	0.126	0.282	0.383	0.120	1.000	0.098	0.037	0.003	0.012	0.626
APS	13,791	0.001	0.702	0.998	0.413	0.000	0.008	0.012	0.005	0.007	1.000	0.000	0.000	0.000	0.686
ITSS	273	0.231	0.484	0.989	0.912	0.044	0.308	0.429	0.055	0.132	0.011	1.000	0.044	0.110	0.582
QIIS	66	0.591	0.091	0.955	0.682	0.045	0.091	0.182	0.045	0.045	0.000	0.182	1.000	0.136	0.409
AIIS	183	0.230	0.459	1.000	0.885	0.049	0.443	0.590	0.049	0.066	0.016	0.164	0.049	1.000	0.541
PROD	83,781	0.025	0.992	0.987	0.759	0.002	0.037	0.091	0.011	0.007	0.113	0.002	0.000	0.001	1.000

Firm counts random rounded base three in accordance with Statistics NZ confidentiality procedures. Permanent enterprise number (PENT) used. Year is *dim_year_key* and the 2012 (201203) year is used. PROD is the Fabling-Maré productivity dataset. Other acronyms are defined in table 1 (OMT - X and OMT - M refer to exports and imports respectively). Group level collections (ITSS, QIIS, AIIS) are analysed at the *reporting* permanent enterprise level (as opposed to, say, counting all members of a reporting group as being included in the coverage).

Table 15: Overlap between data sources (2012), businesses with (10,50] FTE

	Total	AES	IR10	BAI	IR4	GAP	OMT - X	OMT - M	BOS	R&D	APS	ITSS	QIIS	AIIS	PROD
AES	2,859	1.000	0.587	0.991	0.895	0.029	0.226	0.350	0.281	0.083	0.008	0.057	0.017	0.028	0.806
IR10	8,244	0.203	1.000	0.997	0.948	0.020	0.162	0.262	0.157	0.059	0.036	0.013	0.001	0.007	0.869
BAI	13,050	0.217	0.630	1.000	0.875	0.021	0.183	0.284	0.163	0.065	0.036	0.024	0.005	0.014	0.619
IR4	11,469	0.223	0.682	0.996	1.000	0.021	0.189	0.294	0.164	0.066	0.032	0.024	0.004	0.014	0.658
GAP	270	0.311	0.611	1.011	0.900	1.000	0.644	0.689	0.267	0.600	0.056	0.122	0.011	0.033	0.700
OMT - X	2,400	0.269	0.558	0.998	0.905	0.073	1.000	0.871	0.215	0.169	0.026	0.061	0.009	0.050	0.656
OMT - M	3,723	0.269	0.581	0.996	0.907	0.050	0.562	1.000	0.199	0.134	0.028	0.052	0.009	0.035	0.664
BOS	2,151	0.374	0.603	0.990	0.874	0.033	0.240	0.344	1.000	0.155	0.039	0.040	0.013	0.031	0.640
R&D	852	0.278	0.570	0.996	0.884	0.190	0.475	0.585	0.391	1.000	0.039	0.106	0.014	0.035	0.665
APS	468	0.051	0.628	0.994	0.782	0.032	0.135	0.224	0.179	0.071	1.000	0.000	0.006	0.013	0.615
ITSS	321	0.505	0.346	0.981	0.850	0.103	0.458	0.598	0.271	0.280	0.000	1.000	0.056	0.131	0.654
QIIS	69	0.696	0.174	0.913	0.696	0.043	0.304	0.478	0.391	0.174	0.043	0.261	1.000	0.043	0.696
AIIS	177	0.458	0.322	1.000	0.881	0.051	0.678	0.746	0.373	0.169	0.034	0.237	0.017	1.000	0.576
PROD	8,106	0.284	0.884	0.997	0.930	0.023	0.194	0.305	0.170	0.070	0.036	0.026	0.006	0.013	1.000

Firm counts random rounded base three in accordance with Statistics NZ confidentiality procedures. Permanent enterprise number (PENT) used. Year is *dim_year_key* and the 2012 (201203) year is used. PROD is the Fabling-Maré productivity dataset. Other acronyms are defined in table 1 (OMT - X and OMT - M refer to exports and imports respectively). Group level collections (ITSS, QIIS, AIIS) are analysed at the *reporting* permanent enterprise level (as opposed to, say, counting all members of a reporting group as being included in the coverage).

Table 16: Overlap between data sources (2012), businesses with $(50,\infty)$ FTE

	Total	AES	IR10	BAI	IR4	GAP	OMT - X	OMT - M	BOS	R&D	APS	ITSS	QIIS	AIIS	PROD
AES	2,046	1.000	0.350	0.988	0.736	0.091	0.424	0.584	0.748	0.220	0.023	0.183	0.094	0.063	0.823
IR10	996	0.720	1.000	0.997	0.979	0.075	0.307	0.449	0.660	0.151	0.033	0.063	0.012	0.036	0.864
BAI	2,643	0.765	0.376	1.000	0.753	0.086	0.405	0.561	0.675	0.204	0.034	0.156	0.077	0.056	0.730
IR4	1,992	0.756	0.489	0.998	1.000	0.078	0.398	0.556	0.660	0.185	0.032	0.143	0.069	0.062	0.768
GAP	228	0.816	0.329	1.000	0.684	1.000	0.816	0.855	0.789	0.737	0.079	0.316	0.105	0.118	0.816
OMT - X	1,077	0.805	0.284	0.994	0.735	0.173	1.000	0.950	0.735	0.345	0.042	0.256	0.134	0.097	0.805
OMT - M	1,491	0.801	0.300	0.994	0.742	0.131	0.686	1.000	0.712	0.288	0.040	0.225	0.113	0.087	0.783
BOS	1,803	0.849	0.364	0.990	0.729	0.100	0.439	0.589	1.000	0.268	0.040	0.170	0.085	0.068	0.752
R&D	546	0.824	0.275	0.989	0.676	0.308	0.681	0.786	0.885	1.000	0.071	0.286	0.137	0.099	0.813
APS	90	0.533	0.367	1.000	0.700	0.200	0.500	0.667	0.800	0.433	1.000	0.133	0.167	0.067	0.733
ITSS	411	0.912	0.153	1.000	0.693	0.175	0.672	0.818	0.745	0.380	0.029	1.000	0.292	0.117	0.847
QIIS	204	0.941	0.059	1.000	0.676	0.118	0.706	0.824	0.750	0.368	0.074	0.588	1.000	0.059	0.809
AIIS	147	0.878	0.245	1.000	0.837	0.184	0.714	0.878	0.837	0.367	0.041	0.327	0.082	1.000	0.816
PROD	1,941	0.867	0.444	0.994	0.788	0.096	0.447	0.601	0.699	0.229	0.034	0.179	0.085	0.062	1.000

Firm counts random rounded base three in accordance with Statistics NZ confidentiality procedures. Permanent enterprise number (PENT) used. Year is *dim_year_key* and the 2012 (201203) year is used. PROD is the Fabling-Maré productivity dataset. Other acronyms are defined in table 1 (OMT - X and OMT - M refer to exports and imports respectively). Group level collections (ITSS, QIIS, AIIS) are analysed at the *reporting* permanent enterprise level (as opposed to, say, counting all members of a reporting group as being included in the coverage).

Recent Motu Working Papers

All papers in the Motu Working Paper Series are available on our website www.motu.org.nz, or by contacting us on info@motu.org.nz or +64 4 939 4250.

16-02 MacCulloch, Robert. 2016 "Can "happiness data" help evaluate economic policies?"

16-01 Gørgens, Tue and Dean Hyslop. 2016. "The specification of dynamic discrete-time two-state panel data models"

15-20 Maré David C., Ruth M. Pinkerton and Jacques Poot. 2015. "Residential Assimilation of Immigrants: A Cohort Approach."

15-19 Timar, Levente, Arthur Grimes and Richard Fabling. 2015. "Before a Fall: Impacts of Earthquake Regulation and Building Codes on the Commercial Market"

15-18 Maré David C., Dean R. Hyslop and Richard Fabling. 2015. "Firm Productivity Growth and Skill."

15-17 Fabling, Richard and David C. Maré. 2015. "Addressing the absence of hours information in linked employer-employee data."

15-16 Thirkettle, Matt and Suzi Kerr. 2015. "Predicting harvestability of existing *Pinus radiata* stands: 2013-2030 projections of stumpage profits from pre-90 and post-89 forests"

15-15 Fabling, Richard and David C. Maré. 2015. "Production function estimation using New Zealand's Longitudinal Business Database."

15-14 Grimes, Arthur, Robert MacCulloch and Fraser McKay. 2015. "Indigenous Belief in a Just World: New Zealand Maori and other Ethnicities Compared."

15-13 Apatov, Eyal, Richard Fabling, Adam Jaffe, Michele Morris and Matt Thirkettle. 2015. "Agricultural Productivity in New Zealand: First estimates from the Longitudinal Business Database."

15-12:Laws, Athene, Jason Gush, Victoria Larsen and Adam B Jaffe. 2015. "The effect of public funding on research output: The New Zealand Marsden Fund."

15-11 Dorner, Zachary and Suzi Kerr. 2015. "Methane and Metrics: From global climate policy to the NZ farm."

15-10 Grimes, Arthur and Marc Reinhardt. 2015. "Relative Income and Subjective Wellbeing: Intra-national and Inter-national Comparisons by Settlement and Country Type"

15-09 Grimes, Arthur and Sean Hyland. 2015. "A New Cross-Country Measure of Material Wellbeing and Inequality: Methodology, Construction and Results."

15-08 Jaffe, Adam and Trinh Le. 2015. "The impact of R&D subsidy of innovation: a study of New Zealand firms."

15-07 Duhon, Madeline, Hugh McDonald and Suzi Kerr. 2015 "Nitrogen Trading in Lake Taupo: An Analysis and Evaluation of an Innovative Water Management Policy.

15-06 Allan, Corey, Suzi Kerr and Campbell Will. 2015. "Are we turning a brighter shade of green? The relationship between household characteristics and greenhouse gas emissions from consumption in New Zealand" (forthcoming)

15-05 Fabling, Richard and Lynda Sanderson. 2015. "Exchange rate fluctuations and the margins of exports"