# Using administrative data to validate income in

# Statistics New Zealand’s household surveys

**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 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 IR for statistical purposes, and is not related to the data’s ability to support Inland Revenue’s core operational requirements.

Abstract

The aim of the experiment was to investigate the use of data integration in the validation of income statistics produced by Statistics New Zealand. Data from Statistics New Zealand household surveys, the New Zealand Income Survey (NZIS) and the Household Economic Survey (HES) were individually linked with tax data provided by New Zealand’s tax agency, Inland Revenue (IR) Department. The IR data used in the linking is available in Statistics New Zealand’s Integrated Data Infrastructure (IDI). The experiment focused on comparing income sourced from wages, salaries and self-employment. The expected outputs of the experiment were to (1) identify issues related to systematically using data from other sources in the validation of official statistics, and (2) recommend potential approaches and modelling techniques.

Description

The experiment carried out was to compare the quality of income sourced from wages, salaries and self-employment derived from the NZIS and HES with income data available from the IR data in the IDI. The IDI brings together linked datasets from different data suppliers including Statistics NZ. Data from the Household Labour Force Survey (HLFS) and HES are individually linked to the IDI. The NZIS was a supplement to the June quarter of the HLFS until 2016.

Statistics NZ’s [quality assessment framework](http://www.stats.govt.nz/methods/data-integration/guide-to-reporting-on-admin-data-quality.aspx) (Statistics NZ 2016) was used to understand the issues related to the use of administrative data in the validation of income statistics. The quality assessment framework is based on Li-Chun Zhang’s [two-phase life-cycle model for integrated statistical microdata](http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9574.2011.00508.x/abstract) (Zhang 2012).

Data characteristics

The NZIS data investigated in this paper covers data collected (from 2013 to 2015) when NZIS was a supplement to the June quarter (1 April to 30 June) of the Household Labour Force Survey (HLFS). From 2016, most of the income data collected from the NZIS will be collected as an income module within the HLFS.

The [HLFS](https://ec.europa.eu/eurostat/cros/content/admindata-sga-1_en) measures quarterly average levels of employment, unemployment, non-participation in the labour force. The HLFS collects responses from around 15,000 households every quarter, amounting to responses from approximately 30,000 individuals aged 15 and over. The HLFS interviews the same respondents over eight consecutive quarters and then replaces them (on a rotating basis) by a new set of respondents. Households are selected using a multistage stratified, clustered design, and all adults in selected households are interviewed.

[HES](https://ec.europa.eu/eurostat/cros/content/admindata-sga-1_en) collects information on household expenditure and (before-tax) income, and a wide range of demographic information on individuals and households. HES is run every year between 1 July and 30 June the following year. HES collects responses from around 5000 households. Households are selected using a multistage stratified, clustered design, and all adults in selected households are interviewed.

Information on personal income is available as administrative data from the New Zealand tax agency, Inland Revenue. The data we have access to in the IDI includes tax returns for the self-employed and records from businesses that deduct tax directly from employees’ regular pay (Pay As You Earn or PAYE tax), withholding payments (usually relating to contractor’s pay), and registers of the main government payments, such as government pensions and unemployment benefits. Investment income and superannuation or pension funds other than the main government pension are not included.

The IDI brings together linked datasets various government agencies (including our own). A key component of the IDI, called the ‘spine’, is a dataset to which all of the other datasets link. The target population for the spine is anybody who has ever resided in New Zealand. At present, the [IDI spine](https://ec.europa.eu/eurostat/cros/content/admindata-sga-1_en) is a single list of individuals created by a union of tax, birth, and long-term visa records, to which all the other datasets, such as income data from administrative sources, can be linked. Statistics NZ has linked HLFS and HES to the IDI spine. The linked datasets provide the opportunity to understand the differences on personal income collected from the tax data, the HLFS and HES. The analysis will focus on income obtained from wages, salaries and self-employment.

Activities

Statistics NZ’s [quality assessment framework](http://www.stats.govt.nz/methods/data-integration/guide-to-reporting-on-admin-data-quality.aspx) provided guidance in carrying out the experiment. This framework enables uunderstanding of the error sources from the individual data sources including those arising from the integrated datasets. Doing so assists in determining the associated methodological and operational issues that may limit the use of administrative data in the validation of official statistics.

The 1st step of the quality assessment is to briefly answer the main questions in the [metadata template](http://www.stats.govt.nz/methods/data-integration/guide-to-reporting-on-admin-data-quality.aspx). The metadata template is a convenient way to record a standard set of information – to compare different input datasets. The most important information about the different datasets are:

* **General information**: Items 1.1–1.6 including source agency, purpose of collection, summary of variables, and time span of the data.
* **Population**: The target population, admin population, and reporting units. The items relating to coverage might not be possible to answer with a quick assessment but note anything you do know.
* **Variables**: A short description of key variables. As work progresses, record the target concepts for the variables under investigation as they become known.
* **Collection**: The timing/delay information and method of collection are important and should be easy to find out and record.

The 2nd step involves the identification and categorisation of errors present in each input dataset. This involves detailed consideration of how the methods, purpose, known issues, and other aspects of the original data collection contribute to each of the specific error categories in the phase 1 of the error framework discussed in the attached file titled *Guide to reporting on administrative data quality*.

The 3rd step involves identification and categorisation of errors arising from the integration of the different data sources. These errors are considered with respect to the intended statistical target concepts and population. The effects of phase 1 errors on the integrated dataset must be understood to determine if these flow into phase 2. Phase 2 is discussed in detail in the attached file titled *Guide to reporting on administrative data quality*.

The 4th step involves comparing the results on the distribution of wages and salaries and income from self-employment by age and sex from the linked NZIS-IR and HES-IR datasets.

Outputs

The following table provides the basic information on the input datasets, the NZIS, the HES and tax data in the IDI.

|  |  |  |  |
| --- | --- | --- | --- |
| **Information object** | **NZIS** | **HES** | **IR tax data** |
| **Source agency** | Statistics NZ | Statistics NZ | Inland Revenue Department (IRD) |
| **Purpose of data collection** | Provides a snapshot of income levels for people and households | Collects information on household expenditure and income, as well as a wide range of demographic information on individuals and households | Records tax returns for the self-employed and businesses that deduct tax directly from employees’ regular pay called Pay As You Earn or PAYE tax and withholding payments relating to contractor’s pay |
| **Target set** | All civilian, non-institutionalised usual residents of New Zealand aged 15 years and over | all civilian, non-institutionalised usual residents of New Zealand aged 15 years and over | Any individual who has paid tax |
| **Main variables collected** | Weekly (before-tax) income from various income sources. Main focus for experiment: income sourced from wages and salaries & self-employment | Annual (before-tax) income from various income sources. Main focus for experiment: income sourced from wages and salaries & self-employment  Annual household expenditure | Gross earnings, PAYE, non-zero partnership, self-employment, or shareholder salary income |
| **Mode of collection** | CAPI, CATI ; uses same respondents over eight consecutive quarters | CAPI | On-line and paper forms |

### ****Comparison of income concepts and population definitions****

The [NZIS](http://www.stats.govt.nz/browse_for_stats/income-and-work/Income/NZIncomeSurvey_HOTPJun15qtr.aspx) reports on weekly (before-tax) income and relates specifically to a week during the June quarter; that is, a snapshot in time of income levels for people and households. The weekly earning reported by an individual cannot simply be converted to an annual income since an individual can change jobs in a year resulting in a different weekly pay. Meanwhile, HES collects annual income from all sources – wages and salaries, self-employment, benefits and investments at both the personal and household levels. On the other hand, the tax data in the IDI provides monthly payment for an individual who pays New Zealand income tax that is deducted at source. These indicate notable differences in the concept of personal income across the three data sources. The conceptual differences are also apparent in the linked datasets, HLFS-IDI and HES-IDI.

The target population for NZIS and HES includes all civilian, non-institutionalised usual residents of New Zealand aged 15 years and over. Meanwhile, the target population of the tax data in the IDI includes any individual issued an IRD number by Inland Revenue. New Zealand residents who do not have an IRD number are not included in the tax data. Similarly, not all individuals issued IRD numbers are New Zealand residents. These show different target populations for the three data sources. However, the differences in the definition of the target populations for the linked datasets, HLFS-IDI and HES-IDI disappear.

### ****Comparison of measurement and processing errors across datasets****

Item non-response is an issue in the NZIS and HES responses. Only about 80% of the NZIS respondents and about 67% of the HES respondents have a valid response. Other common measurement errors include: confusing before-tax with after-tax income, recall errors when someone does not remember receiving income from a particular source, approximations made by respondents, such as rating up their latest pay cheque to an annual figure when they also received bonus payments or pay increases; or mistakes when rating up the net pay cheque (e.g., the amount actually on a bank statement) to a before-tax amount. Processing errors arise from edits and imputations carried out.

Measurement errors, including item non-response, are minimal for the tax data because company payroll tax is audited to ensure the correct tax amount is paid to the government. A significant practical issue is that processing for tax data takes considerable time, which means there can be a delay of several months (or more) until full records for a given date are available.

### ****Validating income statistics using the integrated datasets****

The target population for the linked datasets, NZIS-IDI and HES-IDI, is the target population for NZIS and HES. Linkage error is one source of coverage error for the integrated microdata. Missed links, for example if the name of a person is recorded differently on different files, is a potential source of undercoverage. False positives, for example two different people being incorrectly linked if their names and dates of birth are the same, contribute to overcoverage.

The integration of tax in the IDI is also a potential source of linkage error which may flow through in the representation side of the NZIS-IDI and HES-IDI integrated datasets. An example is the duplicate records in the IDI which are caused by an individual who has 2 different IRD numbers.

The target concepts are income sources as defined in NZIS and HES. A source of relevance error in some cases is a person present in the integrated datasets but has no income recorded in the tax data for the reference periods of the NZIS in the NZIS-ID dataset and for the reference periods of the HES in the HES-IDI dataset.

A source of mapping error is caused by the computation of weekly income for the linked NZIS-IDI dataset or annual income for the integrated HES-IDI dataset. A lot of companies pay fortnightly, hence, there could be 3 payment periods reported in a month by an employer. Similarly, some employers pay monthly. These create mapping errors for NZIS-IDI linked data which require the weekly income earned. Another source of mapping error is the computation of income from self-employment. Some components of this income source are paid annually at the end of the tax year. Other components are paid monthly. These create mapping errors for NZIS-IDI linked data which require the weekly income earned from this source as well as the annual income for HES which collects data from July to the 30th June of the following year.

To make the NZIS and HES account for missing tax data (which is equivalent to a person missing an IRD number), an adjusted weight was applied to the final weight of an individual from the linked datasets.

The inverse probability of having an IRD number was multiplied with the final weight to account for individuals with missing IRD numbers. The probability of having an IRD number was predicted using using logistic regression with the following as explanatory variables: highest qualification, labour force status, age group, ethnicity and household composition type.

## Distribution of wage earnings

The following tables provide the correlation coefficients of the same income source between NZIS and the tax data and HES and the tax data using the corresponding linked datasets, NZIS-IDI andHES-IDI, for the years investigated.

Table 1. Correlation coefficients between NZIS and tax data

|  |  |  |
| --- | --- | --- |
| **Year** | **Wages & salaries** | **Income from**  **self-employment** |
| **2013** | **0.827** | **0.606** |
| **2014** | **0.809** | **0.617** |
| **2015** | **0.828** | **0.293** |

Table 2. Correlation coefficients between HES and tax data

|  |  |  |
| --- | --- | --- |
| **Year** | **Wages & salaries** | **Income from**  **self-employment** |
| **2013** | **0.734** | **0.516** |
| **2014** | **0.732** | **0.55** |
| **2015** | **0.762** | **0.393** |

### NZIS and NZIS-IDI wage earnings distributions

#### Wages and salaries

Figures 1, 2 and 3 illustrate the distribution of wages and salaries, by age group, for the unlinked NZIS data, tax data and linked NZIS-IDI dataset.

Figure 1. 2013 distribution of wages and salaries by age group

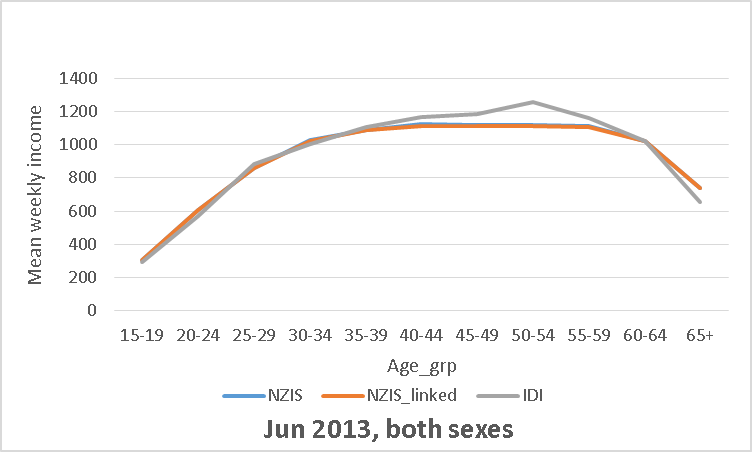
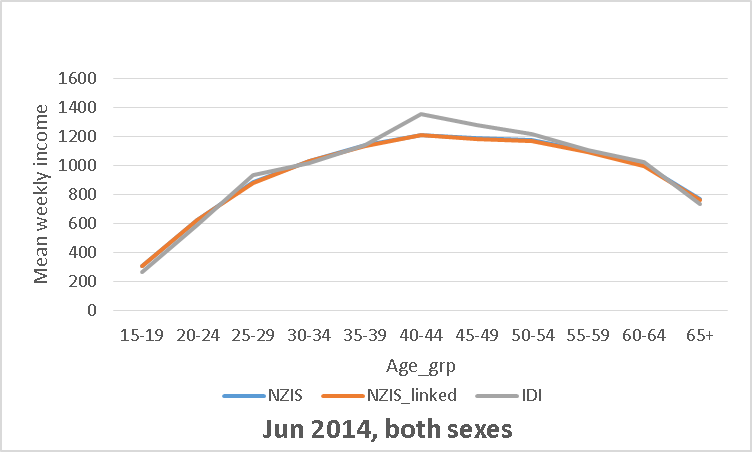
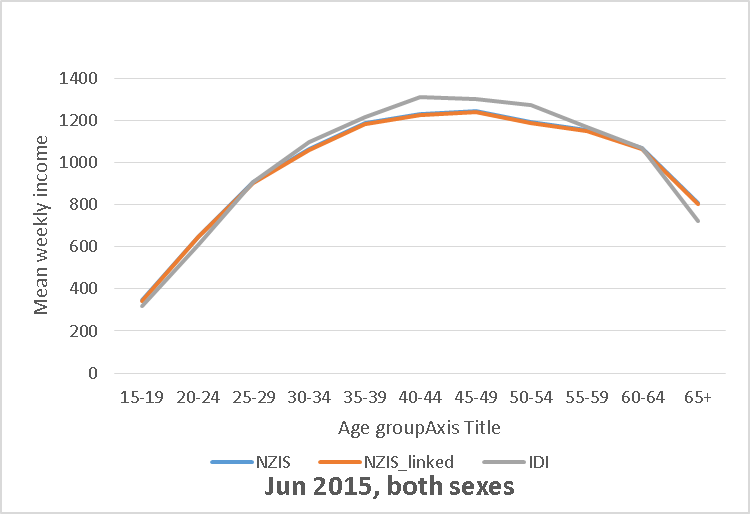
******

Figure 2. 2014 distribution of wages and salaries by age group

******

**Figure 3. 2015 distribution of wages and salaries by age group**

******

**The following summarise results when comparing estimates of mean weekly income from wages and salaries for the years 2013, 2014 and 2015** using the unlinked NZIS data, tax data and linked NZIS-IDI dataset.

* The difference in the overall mean weekly income for people in the linked and unlinked NZIS datasets is less than 0.5%.
* In general, mean weekly income from the NZIS is about 2.5% lower than the mean from IRD data. There is more underreporting in the NZIS for higher earning individuals. A reason for this may be that high-income earners tend to “forget” income that they received from additional jobs, overtime, or bonuses.
* NZIS mean weekly income is higher than the IRD data for the younger and older ages. These maybe individuals who hold part-time and/or irregular jobs and tend to report the income they receive in a full week of work which may be not representative of their average weekly income. Consequently, they are likely to over-report their income in the survey.
* Similar results hold for the two different sex groups. These are presented in Figures 4 – 9.

Figure 4. 2013 distribution of wages and salaries by age group for the male population

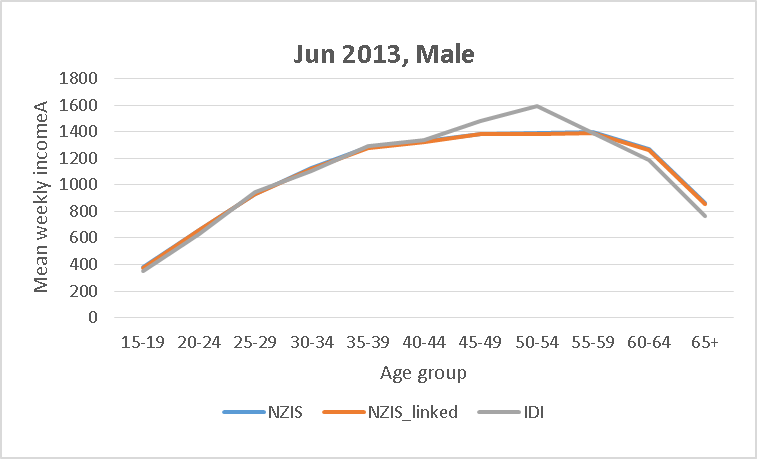
******

Figure 5. 2014 distribution of wages and salaries by age group for the male population

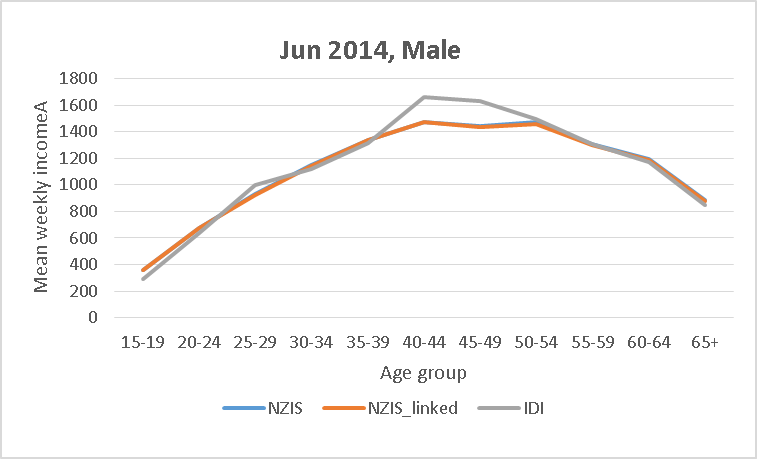
******

Figure 6. 2015 distribution of wages and salaries by age group for the male population

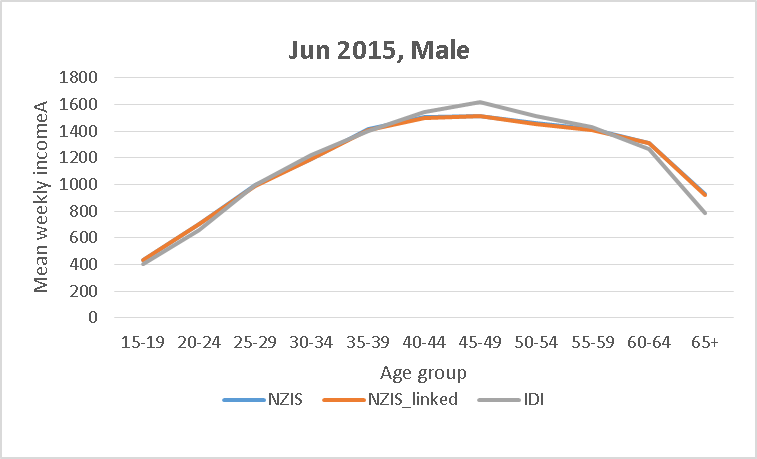
******

Figure 7. 2013 distribution of wages and salaries by age group for the female population

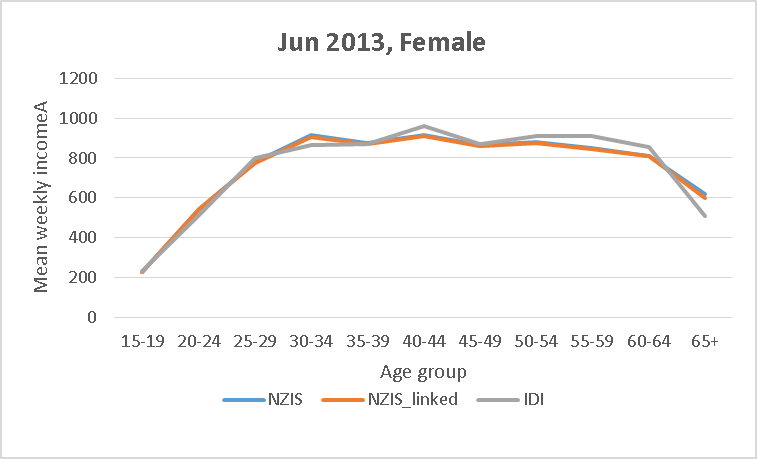
******

Figure 8. 2014 distribution of wages and salaries by age group for the female population

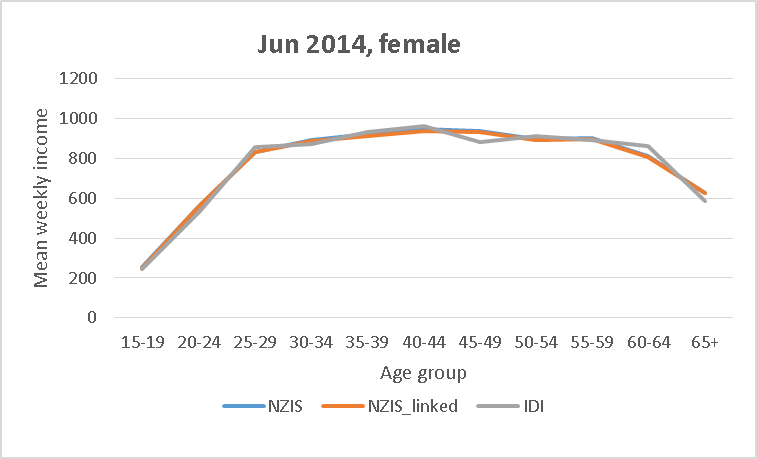
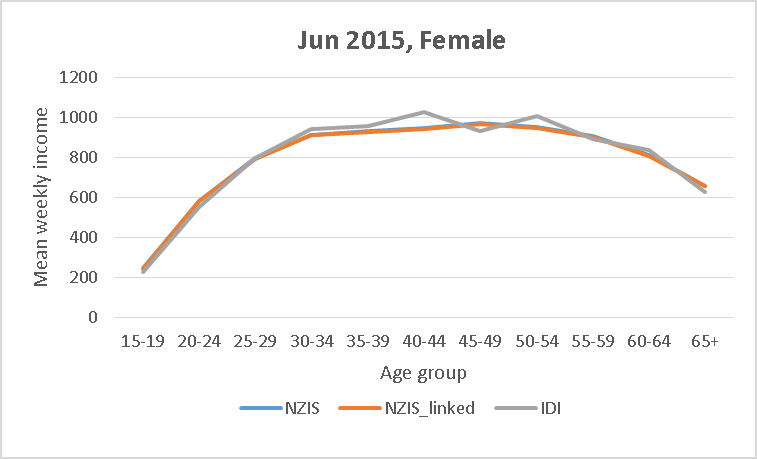
******

Figure 9. 2015 distribution of wages and salaries by age group for the female population

**

**Income from self-employment**

**Figures 10, 11 and 12 illustrate the distribution of income from self-employment, by age group, for the unlinked NZIS data,** tax data and linked NZIS- IDI data. Data for some age groups have been suppressed due to quality issues.

Figure 10. 2013 distribution of income from self-employment by age group

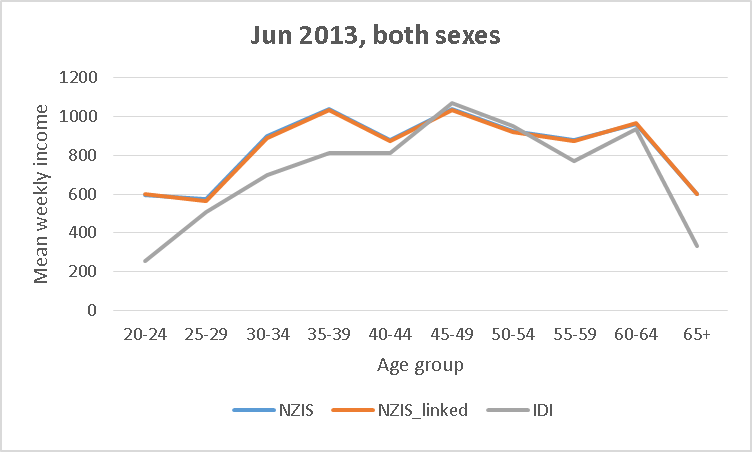
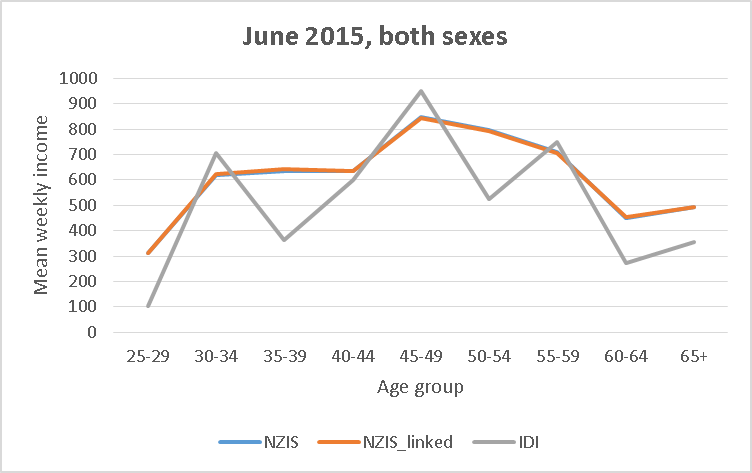
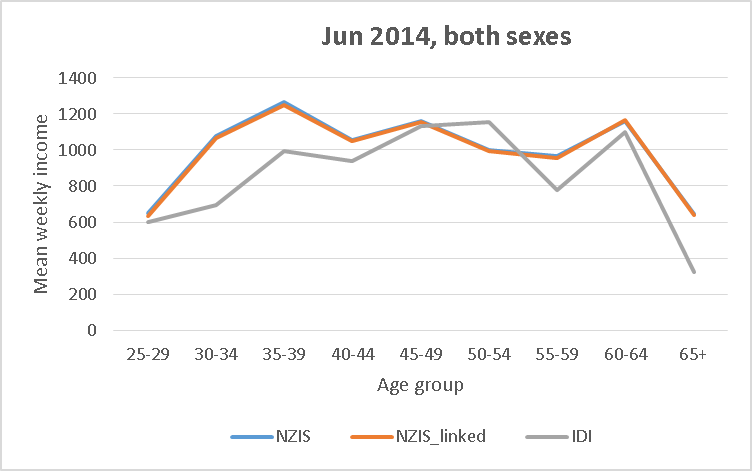
******

Figure 11. 2014 distribution of income from self-employment by age group

Figure 12. 2015 distribution of income from self-employment by age group

******

******

**The following summarise results when comparing estimates of mean weekly income from self-employment for the years 2013, 2014 and 2015** using the unlinked NZIS data, tax data and linked NZIS-IDI dataset.

* **There are** large discrepancies between mean weekly income obtained from the NZIS and IRD data. These may be attributed to the computation of this source from the IRD data. Some of the components of this income source are reported to IRD annually at the end of the tax year while the rest are reported to IRD monthly. Therefore, the difference is mostly attributed to conceptual differences in self-employment income.
* The difference in the overall mean weekly income between the linked and unlinked NZIS datasets is less than 0.3%.
* In general, the mean weekly income from the NZIS is higher than the mean from IRD data.

### HES and HES-IDI wage earnings distributions

#### Wages and salaries

Figures 13, 14 and 15 illustrate the distribution of wages and salaries, by age group, for the unlinked HES data, tax data and linked HES-IDI dataset.

**The following summarise results when comparing estimates of mean annual income from wages and salaries for the years 2013, 2014 and 2015** using the unlinked HES data, tax data and linked HES-IDI dataset.

* The difference in the overall mean annual income for people in the unlinked and linked HES datasets around 2%.
* In general, the mean annual income from the HES is about 5.5% higher than the mean from IRD data with higher differences exhibited for the younger and older ages. These maybe individuals who hold part-time and/or irregular jobs and may report income from full-time employment rather than their true income.

Figure 13. 2013 distribution of wages and salaries by age group

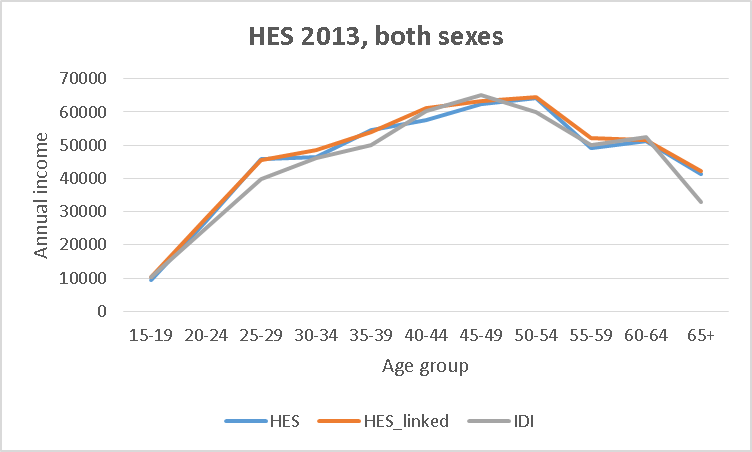
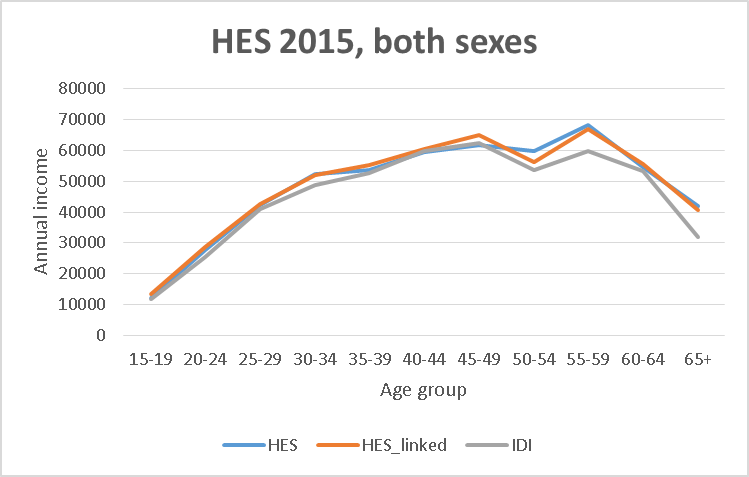
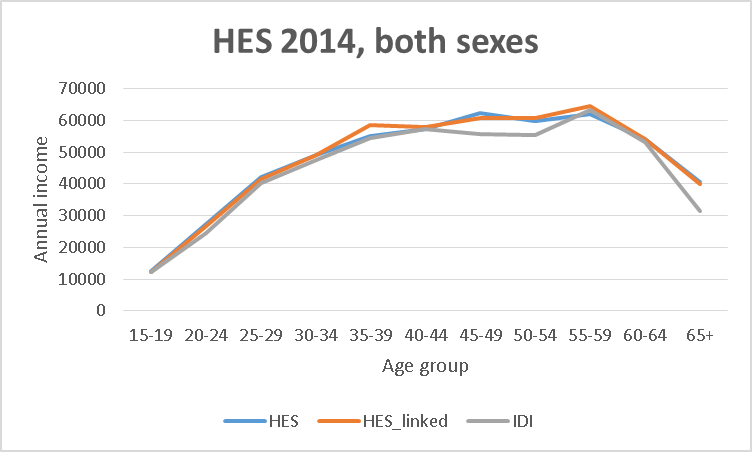
******

Figure 14. 2014 distribution of wages and salaries by age group

**Figure 15. 2015 distribution of wages and salaries by age group**

******

******

#### Income from self-employment

Figures 16, 17 and 18 illustrate the distribution of income from self-employment, by age group, for the unlinked HES data, tax data and linked HES-IDI dataset. **Data for some age groups have been suppressed in the graphs due to quality issues.**

**The following can be observed from the graphs:**

* Estimates from the original HES sample are similar to estimates obtained from the linked HES-IDI sample for 2014 and 2015.
* There are large discrepancies between mean annual income obtained from the HES and IRD data. These may be attributed to the computation of income from the IRD data. Although most of the income derived from this source of income is reported to IRD annually, some components are reported monthly and had to be mapped accordingly in the computation for annual income. Therefore, the difference is mostly attributed to mapping errors in self-employment income.

Figure 16. 2013 distribution of income from self-employment by age group

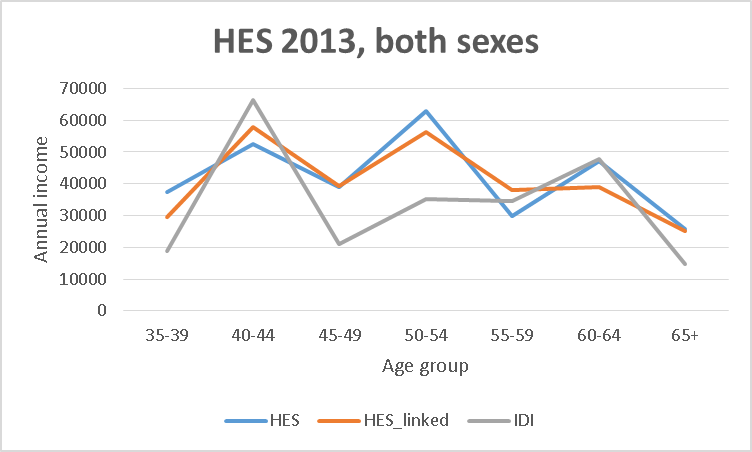
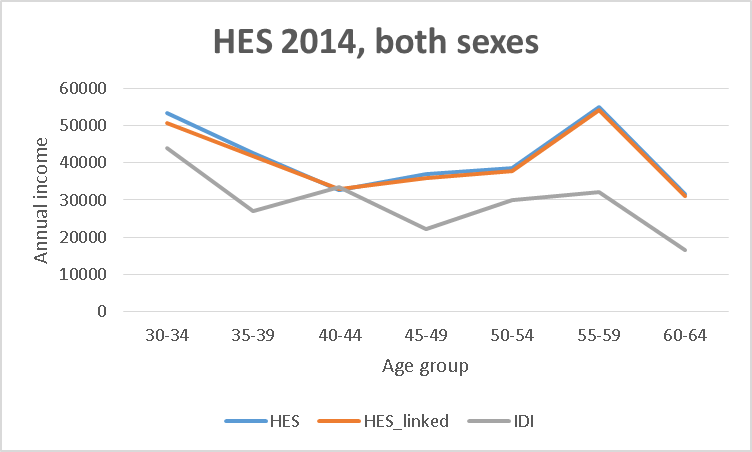
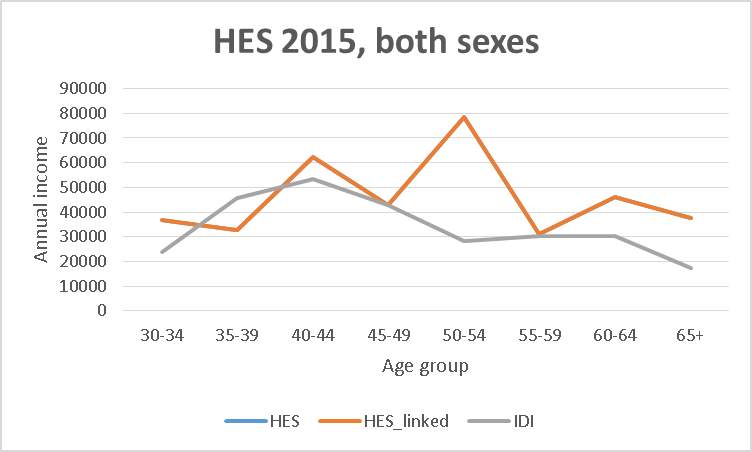
******

Figure 17. 2014 distribution of income from self-employment by age group

******

**Figure 18. 2015** distribution of income from self-employment by age group

******

Findings

Although the link rate between HES and the IR is higher (=84%) than that of the NZIS and IR (=70.1%), the correlation coefficient between NZIS and IR data is higher (=0.81) than that of the HES and the IR (=0.74).

In general, the mean income from wages and salaries in the NZIS is about 2.5 % lower than that from IRD data while the mean income from HES is about 5.5% higher than that from IRD data. Differences are more pronounced for the younger and older ages.

**Relevance**

National statistical institutes are looking for opportunities to make greater use of administrative sources. There are numerous advantages in doing so: cost reduction in statistical production, reduction in respondent burden and potentially more informative datasets, to name a few. These data sources are now used for various statistical uses which include use for sampling frames and stratification, replacement of a data collection, use in data editing and imputation, use for estimation and validation of survey estimates. However, one must be careful in integrating these data sources in statistical production because many of these have not been collected for statistical purposes. Therefore, this experiment is relevant in the use of external data sources in the validation of official statistics.

**Technology**

Statistics NZ uses IBM’s QualityStage to perform probabilistic record linkage. Data analyses utilised SAS and R. These are the tools available to users of IDI data. These tools are sufficient to carry out investigations with administrative data sources.

**Methodology**

The use of Statistics NZ’s [quality assessment framework](http://www.stats.govt.nz/searchresults.aspx?q=guide%20to%20reporting%20the%20quality%20of%20administrative%20data) provides a thorough, structured and methodological way of carrying out investigations involving a potential use of administrative data in the production of official statistics. We have also recommended it as a basis for an analysis of cost/quality trade-off as well in the evaluation of competing designs of a statistical output.

**Source**

The data sources in this experiment were only available in Statistics NZ’s IDI. There are restrictions in accessing these data sources.

**Quality**

The investigations carried out in this experiment enables a statistical analyst from a national statistical institute to make recommendations on the quality requirements to make sound methodological decisions on the use of administrative data in the production of official statistics.

Conclusions

One issue identified with the use of data integration to validate official statistics is the need for a weight to adjust for missed links, especially for individuals who reported non-zero income in the household surveys. More work is needed to improve on the weight adjustment methodology for linked data.

Methods to better estimate linkage errors are required to determine models appropriate to account for these linkage errors. Linkage errors contribute to potential coverage errors in the resulting target population. Care should also be undertaken when creating statistical units from integrated datasets wherein one dataset is sourced from an administrative dataset since the unit may be defined differently in the administrative dataset.

Administrative data may suffer from measurement errors, e.g., validity error, and these errors propagate when the administrative data is integrated with other data sources to produce a statistical output. Hence, target concepts used in an administrative dataset should be well understood before being used in the production of official statistics.

Understanding the error sources from the input data sources, including those arising from the linked datasets, is essential in determining the associated methodological and operational issues that may limit the use of administrative data in the validation of official statistics.

References

Black, A. 2016. [The IDI prototype spine’s creation and coverage](http://www.stats.govt.nz/methods/research-papers/working-papers-original/idi-prototype-spine.aspx). (Statistics New Zealand Working Paper No 16–03). Available from [www.stats.govt.nz](http://www.stats.govt.nz).

Gibb, S, Bycroft, C, Matheson-Dunning, N. 2016. [Identifying the New Zealand resident population in the Integrated Data Infrastructure](http://www.stats.govt.nz/methods/research-papers/topss/identifying-nz-resident-pop-in-idi.aspx) *(IDI).* Available from [www.stats.govt.nz](http://www.stats.govt.nz).

Statistics New Zealand. 2016. [Guide to reporting the quality of administrative data](http://www.stats.govt.nz/methods/data-integration/guide-to-reporting-on-admin-data-quality.aspx). Available from www.stats.govt.nz.

Statistics New Zealand. 2015. [Household Economic Survey (Income): Year ended June 2015](http://www.stats.govt.nz/browse_for_stats/people_and_communities/Households/HouseholdEconomicSurvey_HOTPYeJun15.aspx) (Income): Year ended June 2015. Available from [www.stats.govt.nz](http://www.stats.govt.nz).

Statistics New Zealand. 2015. [Household Labour Force Survey sources and methods: 2015](http://www.stats.govt.nz/browse_for_stats/income-and-work/employment_and_unemployment/hlfs-sources-methods-2015.aspx). Available from [www.stats.govt.nz](http://www.stats.govt.nz).

Statistics New Zealand. 2014. [Linking methodology used by Statistics New Zealand in the Integrated Data Infrastructure project](http://www.stats.govt.nz/browse_for_stats/snapshots-of-nz/integrated-data-infrastructure/idi-resources/linking-methodology-statsnz-idi.aspx). Available from [www.stats.govt.nz](http://www.stats.govt.nz).

Statistics New Zealand. 2015. [New Zealand Income Survey: June 2015 quarter](http://www.stats.govt.nz/browse_for_stats/income-and-work/Income/NZIncomeSurvey_HOTPJun15qtr.aspx). Available from [www.stats.govt.nz](http://www.stats.govt.nz).

Zhang, L-C. 2012. [Topics of statistical theory for register-based statistics and data integration](http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9574.2011.00508.x/abstract). *Statistica Neerlandica, 66,* 41–63. doi: 10.1111/j.1467-9574.2011.00508.x