Exploring the Probabilistic Link between the Household Labour Force Survey and the Linked Employer-Employee Dataset

Paper presented at the New Zealand Association of Economists (NZAE) Conference,
at Wellington, New Zealand,
29 June – 1 July 2011

A. McLeish Martin

Statistical Analyst, Work, Knowledge and Skills Unit, Statistics New Zealand
P O Box 2922
Wellington, New Zealand
Mcleish.Martin@stats.govt.nz
(04) 931 4590

www.stats.govt.nz

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

Citation
Martin, A. M., III, (2011, June). *Exploring the probabilistic link between the Household Labour Force Survey and the Linked Employer-Employee dataset.* Paper presented at NZAE annual conference 2011, Wellington, New Zealand.

Abstract

Statistics New Zealand's quarterly Household Labour Force Survey (HLFS) is used to produce official estimates of the numbers of employed and unemployed people, those not in the labour force, and the official unemployment rate for New Zealand. A recent Statistics NZ project created a probabilistic link between this survey data and administrative data in the Linked Employer-Employee Data (LEED) dataset. The link covers respondents in the HLFS from the December 2006 quarter to the March 2010 quarter, and those same individuals' LEED records from April 1999 to March 2010. This linked data has allowed the creation of research datasets comparing, at a unit record level, a respondent's labour force status in the HLFS with their employment and benefit status in LEED. This paper explains how the data was linked, explores the results of these comparisons, and discusses the possible use of administrative data in the HLFS if a live link between the two datasets were to be created.

JEL Code: J21

Keywords: Labour market, unemployment, employment

Acknowledgements and notes on the Data

The opinions, findings, recommendations, and conclusions expressed in this report are those of the author. They do not represent those of Statistics New Zealand, which takes no responsibility for any omissions or errors in the information. Access to the data used in this study was provided by Statistics New Zealand under conditions designed to give effect to the 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 or firm. The data in this paper contain information only about groups of individuals so that their confidentiality is protected. The results are based in part on tax data supplied by Inland Revenue to Statistics New Zealand under the Tax Administration Act. These tax data must be used only for statistical purposes, and no individual information is published or disclosed in any other form, or provided back to Inland Revenue for any administrative or regulatory purposes. Any discussion of data limitations or weaknesses is in the context of using the Linked Employer-Employee Dataset for statistical purposes, and is not related to the ability of the data to support Inland Revenue's core operational requirements. Careful consideration has been given to the privacy, security, and confidentiality issues associated with using tax data in this project. Any person who had access to the unit record data has certified that they have been shown, have read and have understood Section 87 (Further Secrecy Requirements) of the Tax Administration Act. A full discussion can be found in the LEED Project Privacy Impact Assessment paper, available on the Statistics New Zealand Website.

Purpose of the project

Statistics New Zealand undertook the project with the primary purpose of investigating the feasibility and potential of linking household surveys to LEED administrative data, particularly in light of improvements that are being planned for the HLFS. The project was also viewed as a means to compare estimates of employment from the HLFS survey with LEED administrative employment records to assess the quality of the two data sources. A privacy impact assessment and consultation with the office of the privacy commissioner was undertaken before any linking was conducted.

The HLFS dataset

The HLFS is the quarterly labour force survey run by Statistics New Zealand. The primary purpose of the survey is to estimate the number of people employed, unemployed, and not in the labour force (NILF), and from them, the unemployment rate for the New Zealand labour market. For this project we linked quarters from the December 2006 quarter to the June 2010 quarter. However, LEED data is not yet available for the June 2010 quarter, as the source data is not as timely as the HLFS; thus for the purposes of this report we only use quarters up to March 2010. We began the linking with the December 2006 quarter as that was the first quarter in which computer assisted interviewing was fully integrated into the survey.

The LEED dataset

For the purposes of matching unit records we used the IR Client Register, which is Inland Revenue's historical list of every IRD number that has been assigned. Once we matched to an individual's IRD number we were then able to pull records of their employment from monthly Employer Monthly Schedule (EMS) filings from April 1999 to March 2010. EMS data includes information from employers listing each employee and their earnings in a given month. Earnings are then used as a proxy for employment in that same month. EMS can also be used to identify separate receipt of income based benefits, e.g. paid parental leave, New Zealand Superannuation, ACC payments, and student allowances.

How the link was created

The link between unit records in the HLFS and LEED data was created using the QualityStage matching software by matching on four variables: first names, surnames, dates of birth, and sex. A series of ‘passes’ were conducted using date of birth as the primary means to sort records and then using different combinations of first and last names from the two datasets. The first and last names were also manipulated to pick up additional matches. From these variables a combination of exact matching and probabilistic matching was used (probabilistic matching creates weights for each match based on the uniqueness of the variables and how closely the two records match). Sequential passes were conducted starting with exact matches, matches manipulating names, and then different combinations of first names and last names recorded in the datasets. An 80+ percent match was achieved across the 14 quarters of HLFS records that we attempted to match to LEED.

Figure 1
An illustrative example of names from the two datasets and how they might be matched

HLFS names		LEED names		Match type
Daffy	Duck	Daffy	Duck	Exact match
Mickey	Mouse	Micky	Mouse	Inexact match
Bugs	Bunny	Bugsey	Bunny	Inexact match
Roger	Rabbit	Rog	Rabbit	Inexact match
Charlie	Brown	Charles	Brown	Inexact match

Comparing the matched HLFS and LEED data

This paper discusses only the comparison of wage and salary employees in LEED with paid employees in the HLFS. This is because these two populations are the most similar in definition across the datasets, meaning that the comparison from the two datasets is the most certain. While LEED does include other populations, such as people on the unemployment benefit, the difference in the definitions of the officially unemployed and those on benefit mean that direct comparisons are virtually meaningless (though you could model a relationship between the variables). For example, someone can legitimately claim an unemployment benefit in LEED even while working one hour a week or more and therefore be considered officially employed in the HLFS. Thus, when the number of officially unemployed in the HLFS differs from those on the unemployment benefit, there is no way to determine whether this difference is due simply to definitional differences or to other causes. To a lesser extent, the problem of definitional differences also occurs between the self-employed in both LEED and HLFS. However, a larger problem with comparing self-employed is that many of the self-employed in LEED only file annual tax returns, which means they cannot be accurately compared with quarterly HLFS employment data. LEED annual self-employment information is also the least timely, with waits of over a year between date supplies.

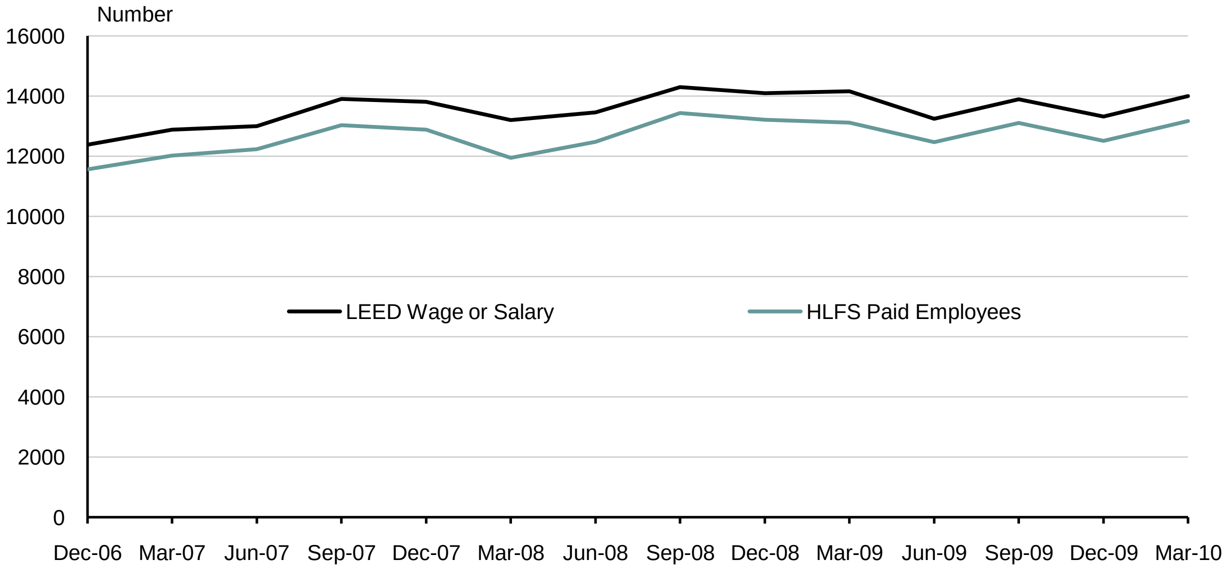
Validating the HLFS employment series

The linked unit records consist of matches made between complete HLFS records and LEED IRD numbers. A comparison of unit records in the HLFS that were recorded as paid employees with EMS records in LEED that were listed as wage and salary earners, shows that both the total numbers[1] and the quarterly movements across the 14 quarters correspond very closely. (See the graphs below – the red lines are the matched HLFS Paid Employees, and the blue lines are the matched LEED Wages and salary earners.) In addition, a correlation between the two series indicates that over 90 percent of the

[1] Note that the most likely reason for the level difference between LEED and HLFS is that the smallest reference period in LEED is a month, whereas the HLFS reference period is only one week. This means that there will be people employed in the first week of a month and thus count as employed in LEED, but then leave their job later that same month and be counted as unemployed or NILF in the HLFS.

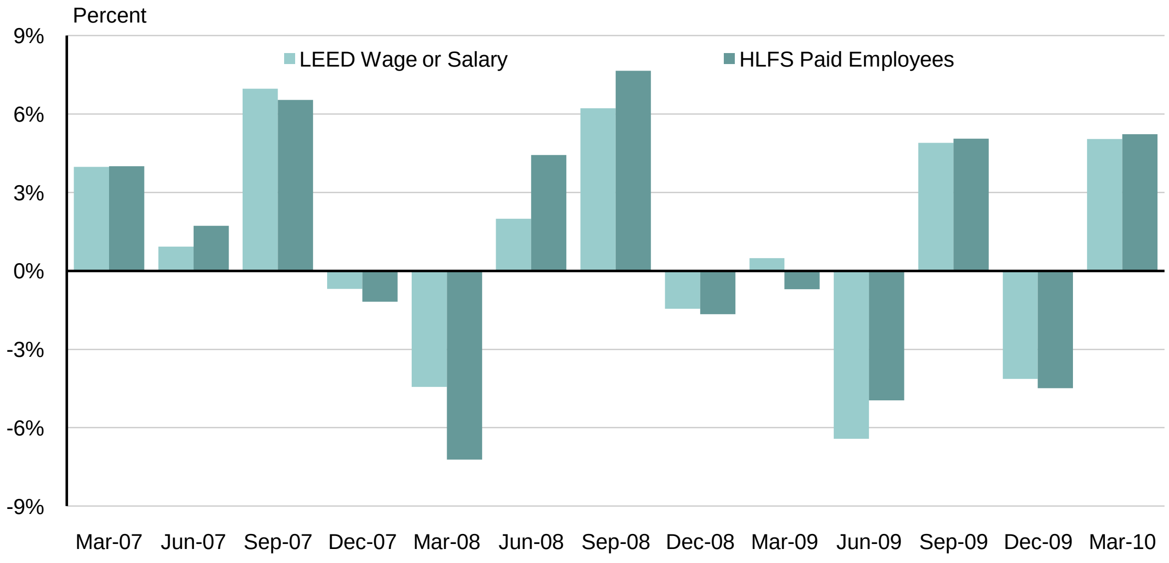
variance in the HLFS paid employees series can be explained by the matched LEED wage and salary series (Correlation Coefficient = 0.95). That is, comparing responses from the HLFS survey with tax information from LEED shows that when definitions are similar, as is the case with paid employees and wage and salary, the HLFS and LEED tell the same story about the New Zealand labour market.

Aggregate comparison from LEED and HLFS matched unit records



Source**:** Statistics New Zealand

Quarterly Percentage changes - LEED and HLFS matched unit records



Source**:** Statistics New Zealand

Exploring why some records did not agree

Our next step was to test for systematic bias in the people whose status as HLFS paid employment and LEED wage and salary did not agree. We began with the people who reported being paid employees in the HLFS and examined what their LEED records indicated. First, across all the quarters, between 91 percent and 94 percent of the people reporting that they were paid employees in the HLFS were also listed as wage and salary workers in LEED. The people whose HLFS records did not agree with LEED did not seem to show any systematic bias towards alternate LEED responses – benefits, withholding, pensions, ACC, paid parental leave, or student allowance. In addition, there were little or no systematic differences in personal characteristics of the people whose records did not agree. We found that there were no differences based on sex, proxy response, or ethnicity, or in five year age bands between 20 and 64. There were small differences related to age for those aged 15–19 years and age bands above 65-years-old. However, as these people are more likely to be transitioning into, and out of, the labour force, disagreement between the data sources is more understandable.

We then took those people who were listed as wage and salary in LEED and examined their HLFS records. Similar to the HLFS paid employees, the vast majority of LEED wage and salary workers were also paid employees in the HLFS – between 91 percent and 93 percent across the quarters. Further, we again found no systematic bias towards alternate responses – self employed, unpaid family workers, unemployed, or not in the labour force. In addition, there were no sex or age differences across the people whose records did not agree.

Further work

In addition to validating the HLFS paid employees series with LEED wage and salary data, we also investigated the potential for new statistics and improvements to current statistics to come out of a link between the HLFS and LEED.

New information through combining LEED and HLFS

There are three main categories of variables in LEED that could be used to provide new information when combined with the HLFS: labour market history and job tenure, labour market dynamics and worker accession and separation, and employer variables. Note that all of these statistics would be produced from the linked HLFS and LEED data, and that the discussion is broken into ‘LEED’ and ‘HLFS’ statistics only because these are where these statistics usually appear.

LEED statistics with new HLFS breakdowns

The first two groups of variables (labour market history and job tenure, and labour market dynamics and worker accession and separation) already form part of the LEED information release. The benefit to linking LEED and the HLFS lies in taking advantage of population breakdowns in the HLFS that are not available in LEED. These breakdowns are ethnicity, occupation, highest qualification, household type, and hours worked. Due to the timeliness of tax return data, these variables would be best suited to a vehicle outside of the current HLFS information release.

HLFS statistics with new LEED breakdowns

The third category of LEED variables are those related to employers. Variables, such as firm size and public versus private sector, provide a population breakdown that is not currently available in the HLFS information release. Adding this LEED information allows for the possibility of new breakdowns in the HLFS variables. For example, statistics on the number of employed by firm size or by sector. Again, the publication of these new statistics would have to work around the longer delay in using tax return data.

Use LEED to enhance the HLFS

A live link between LEED and the HLFS may benefit the HLFS by allowing imputation of missing data. Once a household is selected into the HLFS they are part of the sample for eight quarters. However, due to non-response, not all unit records appear in all eight quarters. Because imputation would rely on observing the matched trends in a person’s HLFS and LEED records, our initial imputation investigations used only records that had both HLFS and LEED information from all the eight quarters. When looking at peoples’ final quarter of response, we found that, for people with only one source of income, changes in their LEED earnings records are matched by changes in their HLFS employment status. From this early work, it would seem that some level of imputation of missing HLFS data from LEED records could be possible. Initial investigations comparing industry coding between LEED and the HLFS also indicate that there is room for improving industry codes.

Future work

There are a number of areas where potential additions to the analyses in this paper could be conducted and Statistics New Zealand is currently working on which areas to pursue. Firstly, the comparison of HLFS and LEED records presented in this paper could be further improved by using LEED job and start dates to recreate the HLFS reference week. Secondly, we could conduct additional analyses comparing industry codes from the HLFS and LEED. Thirdly, as mentioned above, we could try to build a model explaining the relationship between administrative data and HLFS data for those people who were not paid employees (self-employed and beneficiaries). Fourthly, we could extend our investigations into the potential for using LEED to impute missing data in the HLFS. Finally, although the evidence in the paper points towards a strong agreement between the HLFS unit records and EMS records from LEED, further work could be done examining the occasional disagreement between the published statistics of employment in the HLFS and in LEED.