The Role of Administrative Data in New Zealand's Household Labour Force Survey

Hansen, Christopher
Statistics New Zealand, Statistical Methods
The Boulevard, Harbour Quays
Wellington 6140, New Zealand
E-mail: chris.hansen@stats.govt.nz

Acknowledgements and notes on the data

The opinions, findings, recommendations and conclusions expressed in this report are those of the author. They do not purport to represent those of Statistics New Zealand, who take no responsibility for any omissions or errors in the information contained here.

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 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 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.

Introduction

New Zealand's Household Labour Force Survey (HLFS) displays a certain amount of variability in key outputs, such as the total number of employed persons. This could be due to sampling or non-sampling error, or else it could reflect movements in a labour market that is somewhat dynamic. The Linked Employer-Employee Data (LEED) is a database containing monthly tax data for employees, self-employed persons, and benefits recipients. The discussion here considers the merits of linking the HLFS and LEED for the purpose of validating the HLFS - in particular, during calibration, where the HLFS data is adjusted to ensure consistency with certain derived LEED totals.

Definitions and concepts

The HLFS is a household survey which operates continuously over a 13-week cycle. The sampling frame itself consists only of individuals in private dwellings, while the target population includes all civilian, non-institutionalised usual residents of New Zealand. On the other hand, the raw LEED data used for this study contains information for any individuals who pay New Zealand income tax that is deducted at source. This results in some notable conceptual differences with both the HLFS sample, as well as the HLFS target population. Essentially, any individual issued an IRD number by Inland

Revenue (IRD) could potentially be matched with the LEED database, though not everybody will have one. This is the most obvious difference. In addition, it is difficult to exclude individuals who are not usually resident from the LEED database, whereas such individuals are effectively excluded from HLFS totals, as well as the survey itself. Other groups of individuals that do not appear in either the HLFS sample or the HLFS target population are also difficult to exclude from LEED such as the institutionalised and those living in non-private dwellings such as serviced apartments or hotels.

This work is exploratory, so we make some simplifying assumptions. We assume that persons over 19-years of age will generally have an IRD number allocated to them even if they are not paying taxes for a particular period. Furthermore, little effort is made to ensure the LEED and HLFS populations are consistent. In order to do so, the number of individuals in any particular LEED total not in the HLFS target population would need to be estimable. This itself may well be possible. For example, in the HLFS, the relevant exclusions (permanent armed forces and so forth) are made by applying a proportion to the total population estimates. Similar sets of exclusion ratios could be produced and applied to LEED totals. However, such enhancements are left for further exploratory work.

For further detail regarding LEED visit the Statistics New Zealand (Stats NZ website, for example. More detail about the HLFS can also be found the Statistics New Zealand website.

Matching of unit record data

The HLFS and LEED data were largely matched on individuals' names and birth dates. The match itself was performed with QualityStage with somewhat conservative assumptions. A conservative approach was taken in order to minimise spurious matches. Regardless, a match rate of approximately 80% was achieved quite generally. The author made additional matches, raising the overall match rate by at least 5 percentage points. The latter additions were less conservative than the initial match and possibly increase the number of spurious matches. Initially, data has been matched for the December 2006 to June 2010 quarters, inclusive. These are cycles 85 through 99 of the HLFS, and these cycle numbers are used to index plots in this report.

Note that matching of HLFS and LEED data was subject to a privacy impact assessment in consultation with both the IRD and the Office for the Privacy Commissioner.

Timeliness of LEED data

Quarterly HLFS results are published in the fifth week following the end of the calendar quarter. This makes the HLFS a rather timely collection when compared with LEED. When the HLFS data is published, only 8% of wages and salaries records have been received on average for the final month of the quarter. 90% of the first month, and 77% of the second month will have been received, however.

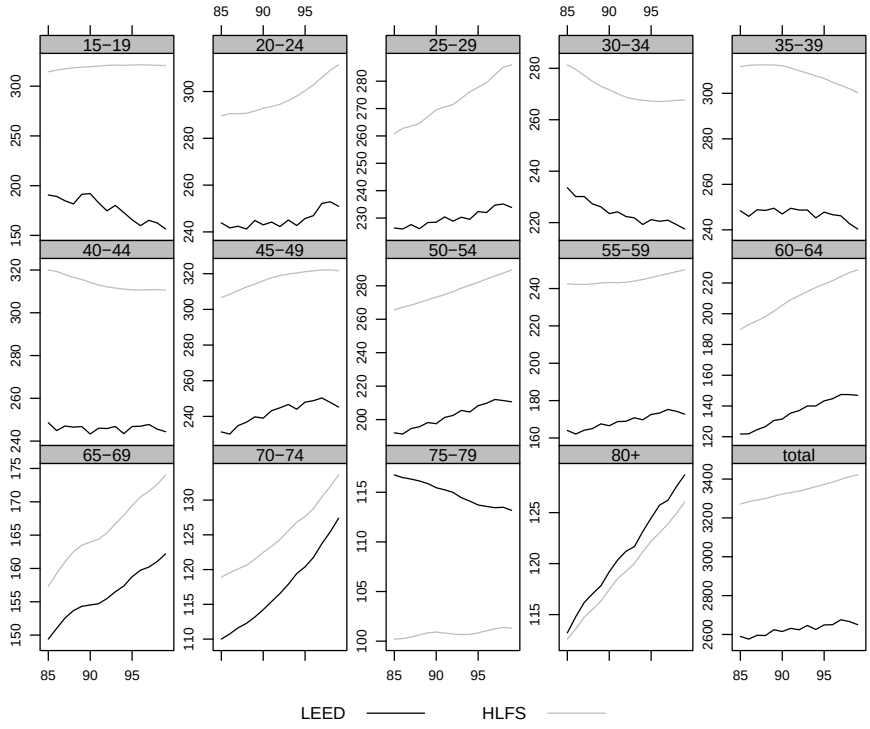
On the other hand, at least 90% of wages and salaries for any month in the quarter will have been processed after a lag of one calendar quarter. Thus, as the process currently stands, to make use of LEED the HLFS would likely need to be published as provisional and then be revised one quarter later. That is, each quarterly release would consist of a provisional release of the most recently completed quarter as well as a revision of the preceding quarter.

Population comparisons

The HLFS is conducted continuously over a 13-week period, with each respondent answering with respect to a single designated week in the quarter. So, population estimates are essentially an average for the quarter in question. On the other hand, as far as it is relevant here, LEED contains data for a particular month. If we were to count the number of individuals appearing over an entire

quarter we would over-count the population. We mitigate this somewhat by instead counting the number of individuals in each of the 3 months in a quarter, then taking the average. Figure 1 provides a comparison of the total population estimates. The HLFS totals are estimates of the working-age population provided by Stats NZ's demography area, while LEED totals are simply counts of unique IRD numbers appearing in the monthly tax data.

Figure 1. LEED and HLFS populations by age group.



Now, assume that the HLFS could be completely and accurately matched with LEED. Furthermore, assume that the HLFS sampling frame is not missing residents in non-private dwellings or who were temporarily overseas, and that we are able to accurately identify usual residents in the LEED population. Let

$$(1) \quad \gamma_i = \begin{cases} 1 & \text{if } i \text{ is matched} \\ 0 & \text{otherwise} \end{cases}$$

and

$$(2) \quad \delta_i = \begin{cases} 1 & \text{if } i \text{ is matched and observed} \\ 0 & \text{if } i \text{ is matched and not observed} \\ \bar{\delta}_i & \text{otherwise} \end{cases}$$

where

$$(3) \quad \bar{\delta}_i = \frac{\sum_{\forall i} \gamma_i \delta_i w_i}{\sum_{\forall i} \gamma_i w_i}$$

and w_i are the usual calibrated HLFS survey weights, matched means a respondent is matched with the LEED database, and observed mean a respondent was observed in the LEED database in the same month as surveyed in the HLFS. Then assuming a proportion α of the HLFS sample is matched we might expect that

$$(4) \quad \sum_{\forall i} \gamma_i \delta_i w_i \approx \alpha X$$

and

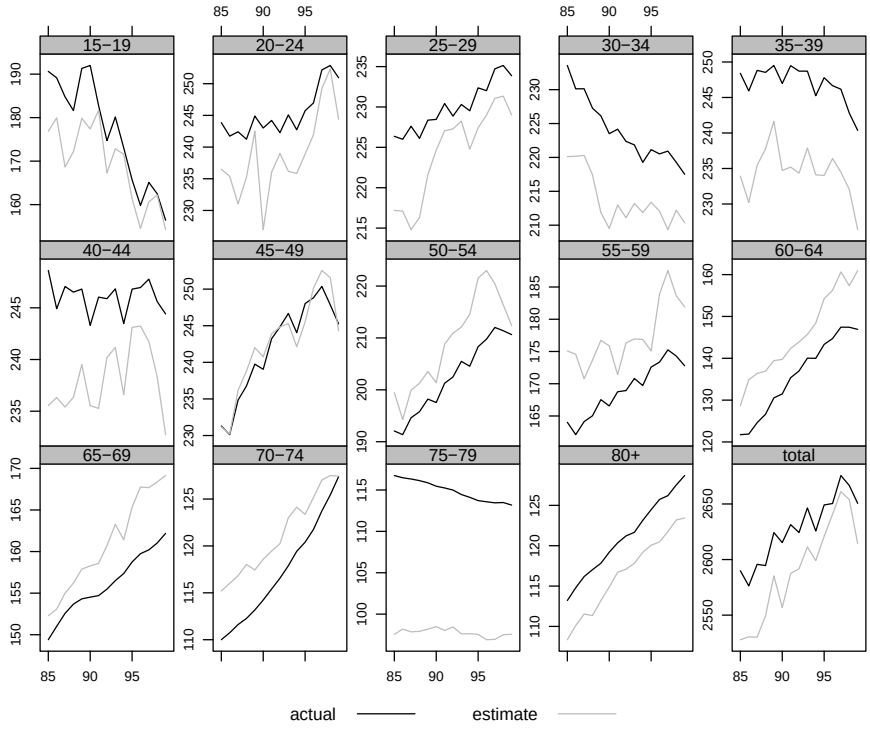
$$(5) \quad \hat{X} = \frac{1}{\alpha} \sum_{\forall i} \gamma_i \delta_i w_i$$

where X is the LEED population total. This latter identity can be shown to be equivalent to

$$(6) \quad \hat{X} = \sum_{\forall i} \delta_i w_i$$

This provides a way of producing approximations of LEED totals from HLFS unit record data in a rather general way simply by replacing δ_i with a different indicator variable - whether an individual received wages or salaries, say. Figure 2 shows the total LEED population and the corresponding HLFS estimate of the same total by age group.

Figure 2. LEED population by age group - actual and HLFS estimate.



Implicit in the above formulation is the assumption that the matched and unmatched individuals have similar characteristics. For example, a proportion $\bar{\delta}$ of the matched sample are observed in the LEED data in the same month as surveyed by the HLFS. We assume that the unmatched sample are indeed in the LEED database, and that had we managed to match them that $\bar{\delta}$ would have been observed in the same month as surveyed - just like the matched sample. There are a number of reasons why this might not be reasonable, not least of which is that there may be a proportion of individuals without IRD numbers. That is, there may be a subset of respondents for whom there is no chance of making a match.

For several age groups, restricting estimates to the matched sample only still results in HLFS estimates that exceed the known totals. Possible explanations for such 'over-coverage' are:

- the HLFS population benchmarks exceed the true working-age population totals
- the LEED database is subject to undercoverage, potentially due to clerical issues or under-reporting

- the HLFS target population is larger than the LEED population.

The HLFS target population includes residents temporarily overseas, but one could argue that even then such residents are likely to appear somewhere in the LEED database. Regardless, the HLFS target population is certainly larger than the survey population for the same reason; and this could certainly lead to some over-representativeness in some sense. We note that the over-coverage discussed is largely a female phenomenon, and it may well be that middle-aged females are relatively likely to be without IRD numbers.

Coverage issues aside, the approach outlined provides us with a relatively simple formulation where the main goal is not to provide the best possible enhancement of HLFS estimates, but rather to test the feasibility of the approach. We do, however, make the following exclusions:

- 15-19: For individuals aged 15 to 19-years of age, the assumption that the majority of the resident population could be matched is quite unreasonable, so we exclude confrontations of population totals for the group.[1]
- 65-years and over: For various reasons, the LEED totals for some groups 65-years and over exceed estimates of the resident population (pensions can be paid overseas, for example, and a number of older New Zealanders do not live in private dwellings), so we avoid confrontation of population totals for those aged 65-years and over.

Calibration

The HLFS data is calibrated using generalised regression (GREG). That is, an estimate of a total can be expressed as follows

$$\hat{Y}_{GR} = \hat{Y}_{HT} + (\mathbf{t} - \hat{\mathbf{t}})'\hat{\mathbf{b}} \tag{7}$$

where

$$\hat{\mathbf{b}} = \left(\mathbf{X}\mathbf{W}\boldsymbol{\Sigma}^{-1}\mathbf{X}'\right)^{-1}\mathbf{X}\mathbf{W}\boldsymbol{\Sigma}^{-1}\mathbf{y} \tag{8}$$

and $\hat{Y}_{GT}$ and $\hat{Y}_{HT}$ denote the GREG and the usual Horvitz-Thompson estimates, respectively. $\mathbf{t}$ and $\hat{\mathbf{t}}$ are actual and estimated control totals, while $\mathbf{X}$ is the design matrix containing the sampled values of the variables being summed.[2] The control totals themselves are the working-age population by sex and 5-year age group, as well as Māori by 2 broad age groups.

The approach taken in this analysis is to simply augment both the control totals and the design matrix with information derived from LEED and the HLFS-LEED match. For example, recall (6) and let $\boldsymbol{\delta} = \{\delta_i\}$. Then we produce the usual GREG estimate but replace $\mathbf{t}'$, $\hat{\mathbf{t}}'$, and $\mathbf{X}'$ with

$$\left[\mathbf{t}'|X\right], \quad \left[\hat{\mathbf{t}}'|\hat{X}\right], \quad [\mathbf{X}'|\boldsymbol{\delta}], \tag{9}$$

respectively.

[1] While an individual may well be in the 15-19-year age group in a particular HLFS quarter, the further in the past the quarter, the more likely an individual will appear in LEED and be matched. This appears to be confirmed by the data - the match rate steadily declines for the age group over time, with the match rate being the lowest for the most recent of the HLFS quarters being considered.

[2] Actually, the HLFS employs integrated weighting which requires that all individuals in a household are assigned equal weight.

An example - LEED wage and salary earners

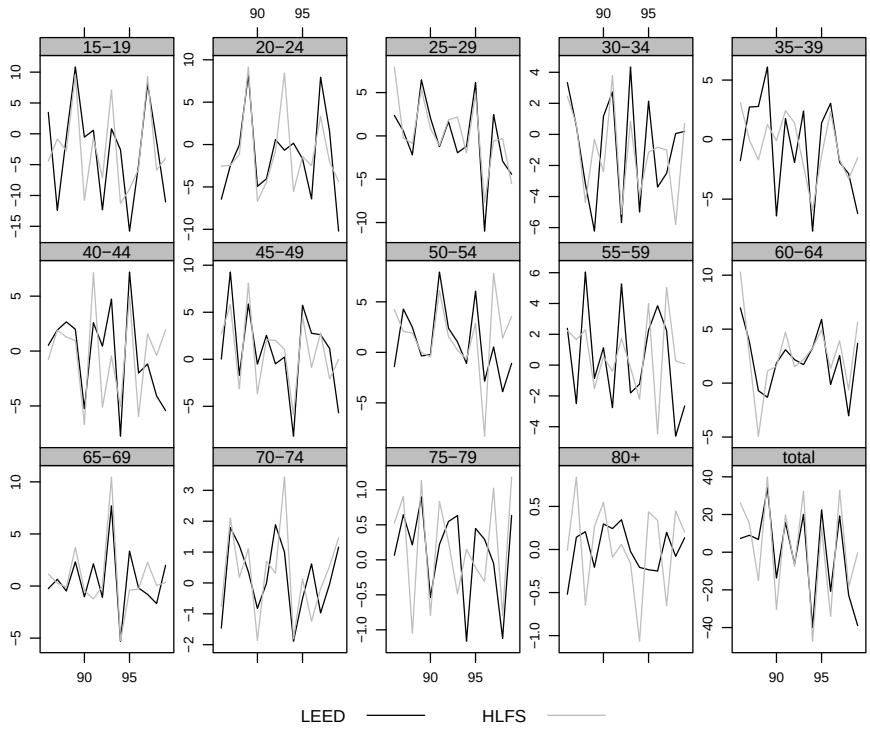
We can enforce the LEED population totals above through calibration, though doing so ought to have a limited effect on HLFS estimates. Instead, we seek LEED variables that have a stronger relationship to labour force statistics.

Consider the number of individuals who attract wages and salaries in any one month period. For the matched sample, we assign an indicator which is true if an individual earned wages or salaries in the same month they were interviewed for the HLFS, and false otherwise. Generally, 95% of the (matched) HLFS employed earned wages and salaries according to the LEED data, while the rate was only 7% and 24% for those not in the labour force and those unemployed, respectively. It is worth noting that HLFS employment conceptually includes all self employed, as well as unpaid family workers, while wage and salary totals from LEED do not.

It is of note that the employment rate for the matched sample generally exceeds 60%, while the employment rate for the unmatched sample is generally around 10% lower. Restricted to those between the ages of 15 and 64, the employment rate for the matched sample exceeds 75%, but the rate for the unmatched sample is still some 10% lower. This supports the idea that there is a core of people who do not appear in the LEED database. Hence, assuming that the unmatched sample has the same characteristics as the matched sample, as we have, will result in a mild overstatement of the general level of employment. However, we expect movements to be only marginally affected.

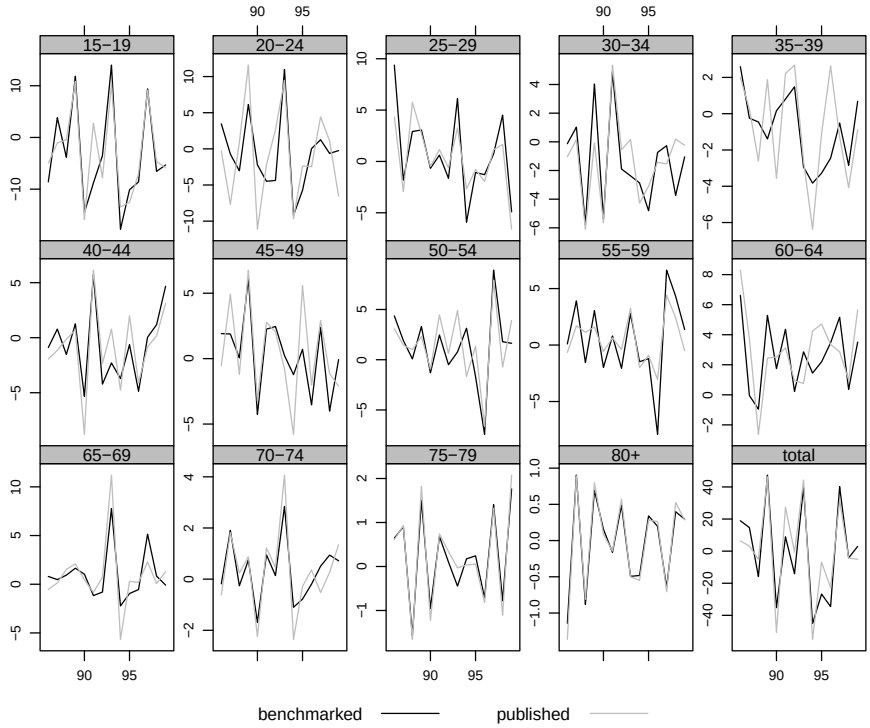
Figure 3 demonstrates the relationship between the LEED wage and salary total and employment in the HLFS as a time series. To account for differences in levels, quarterly change is plotted. It has been suggested that movements in HLFS employment data do not reflect reality, so validating such movements is of particular interest. It is noteworthy, then, that the LEED wage and salary totals show levels of variation similar to those in the HLFS as well as broadly similar movements in the total all ages category.

Figure 3. LEED wage and salary total and HLFS employed by age group, quarterly change.



Finally, figure 4 shows the effect that calibrating to LEED wage and salaries totals has on HLFS total employed. In fact, besides some expected level changes, calibrating has had limited effect - perhaps because the control totals are themselves quite variable, and perhaps also because the HLFS already does a reasonable job of reproducing those control totals without calibration.

Figure 4. HLFS total employed by age group, quarterly change.



Summary, recommendations and future work

Analysis so far suggests that for related variables, the HLFS and LEED are quite agreeable at a high level. This is of interest in itself and could be interpreted as validating the HLFS data. Currently, calibrating as described results in some large level shifts of time series data, though correcting LEED totals to account for conceptual differences with the HLFS population ought to resolve this. Calibration also yields significant reductions in sampling variation which could prove advantageous if bias can be controlled. For example, absolute sampling errors for employed totals were reduced by up to 30% when using a control total related to employment. Similar improvements are observed for those not in the labour force but, owing to the small sample of unemployed, no improvement is observed for that group.

Further investigation would see the application of additional constraints on the HLFS. For example, both unemployment and not in the labour force totals could be improved by introducing controls for certain types of benefits. Moreover, while the initial analysis here focuses on age by sex and labour force, constraints could also be applied to other estimates such as industry-level totals.

Finally, the direct use of LEED data has not been mentioned up to this point. LEED variables not currently collected by the HLFS could be appended to the HLFS data - personal income measures, say. A separate analysis has already confirmed high-level agreement between the New Zealand Income Survey (which is a supplement to the HLFS) and LEED. In addition, the LEED unit record could be used to assess the accuracy of HLFS responses in certain cases. For example, inconsistencies between the HLFS and LEED record could be assessed, and HLFS responses altered if required. Initial investigations have shown that such discrepancies are largely explainable thus far, however.