Methodological News
A Quarterly Information Bulletin
Australian Bureau of Statistics
ABS Methodology and Data Management Division
September 2014

Articles

Creating a Prototype Linked Employer-Employee Dataset, With Example Productivity Analysis

The Advanced Data Analytics section has recently completed work on a prototype linked employer-employee dataset (LEED), and performed preliminary productivity analysis using it. The aim of this project was to assess the feasibility and value to the ABS of potentially producing a full LEED in the future.

The prototype LEED was assembled by linking a number of tax datasets – Business Activity Statements, Business Income Tax, Pay As You Go tax records and Personal Income Tax – with the ABS' Business Longitudinal Database (BLD) and Business Characteristics Survey records, which together contain detailed information on business characteristics. The Pay As You Go records in particular provided the link between employers and employees to connect their records deterministically. The prototype focussed on small and medium size BLD firms with an employee count between 1 to 199 employees for the 2010-11 financial year.

By linking survey and administrative data, a rich prototype dataset emerged, including firm level information on turnover, capital, and non-capital expenses, and employee level information on age, gender, income and occupation. Such linked employer-employee datasets can be used for a number of analytical purposes – for example, to produce labour market statistics on job flows, multiple job holders and employment tenure; or for firm level productivity analysis to incorporate employee characteristics.

As an example of one of these analytical uses, the prototype dataset was used to describe the characteristics of productive firms, in terms of firm and employee level attributes. Multilevel modelling was used to capture the contextual effects in the data, of employees nested within firms, and firms within industries. In particular, both a two-level model (firms in industries) and a three-level model (employees in firms in industries) were investigated to check whether results were consistent across them. The focus was on descriptive analysis rather than causal analysis, and a number of employee and firm characteristics were included. Results were largely in line with expectations, with characteristics like higher capital expenditure, higher operational expenditure, more permanent employees and more high-wage employees associated with more productive BLD firms.

The results of this project were presented at the Asia Pacific Productivity Conference in July, and the paper will be made available on the ABS web site. This project documented the methodological challenges to create a full LEED, provided clear evidence that it would have a number of important analytic uses in answering policy questions, and proposed areas for future research.

Further Information

For more information, please contact Andreas Mayer (02 6252 7140, andreas.mayer@abs.gov.au) or Joseph Chien (02 6252 5917, joseph.chien@abs.gov.au)

Sample and Frame Maintenance Procedures for Census and Household Surveys

Imperfect frames are a well known and inevitable source of non-sampling error in surveys. The aim of Sample and Frame Maintenance Procedures (SFMP) is to reduce the amount of non-sampling error caused by imperfect frames in ABS surveys.

The key design feature of SFMP is to ensure consistent procedures are applied between surveys and to assist staff in making decisions about how to treat differences between a real world unit and the representation on the frame. SFMP has been used in ABS business surveys for many years.

One of the major changes to the Census 2016 enumeration model is to rely on mailing out rather than physically dropping off information to dwellings in some areas. The dwelling address list is the frame extracted from the ABS Address Register. The Address Register is based on the Geocoded National Address File (G-NAF) which contains all physical addresses in Australia, irrespective of their use. The Census counts dwellings rather than addresses, and the relationship between dwelling and address can vary in the real world.

The aim of the dwelling frame is to represent each residential dwelling once and only once. However, some possible scenarios are:

- there is one dwelling at the address
- there are multiple dwellings at the address
- the address is a duplicate of another address on the frame
- there is no dwelling at the address (e.g. commercial property, vacant land, other feature with an address)
- there is a dwelling under construction at the address.

Sample & Frame Maintenance seeks to improve counts and estimates by making the best possible use of information about quality issues in the frame. It is important to have consistent procedures that are able to be applied by staff in data collection and processing so that each dwelling is correctly contributing to Census counts and survey estimates. The current focus of the SFMP project is to develop and test procedures primarily for use in the Census August 2014 tests and the 2016 Census.

Currently Household Surveys use quite different procedures because they use an area based frame, where details of the particular dwellings selected are not known prior to the beginning of the field process. In future, Household Surveys may use the Address Register to create and select a list of addresses to be included in the survey, which SFMP can support. Therefore, while developing Census SFMP, we also aim to develop an enterprise solution to the problem of address register based sample and frame maintenance for both Census and Household Surveys by ensuring that the procedures, principles and statuses developed will support Household Surveys in the future.

Further Information

For more information, please contact Amanda Norton (02 6252 5705, amanda.norton@abs.gov.au) or Pavka

Stevanovic (03 9615 7581, pavka.stevanovic@abs.gov.au)

A More Efficient Sample Design Process for REACS 2013/14

As part of Methodology and Data Management Division's 'Flagship Program' for 14/15, one theme is focussed on 'Harnessing Productivity'. Within this theme there is a key project called 'Survey Productivity Strategies' which has a key element focussing on 'more efficient sample design and collection operations for identified business and household surveys'.

As part of the work on this flagship project, Business Survey Methodology (BSM) have recently reviewed the design process for the Rural Environment and Agricultural Commodities Survey (REACS). The original intent of this review was to make the process undertaken for the design more efficient, with the hope that a more efficient process would potentially lead to a decrease in sample as more time could be devoted to analysing the design rather than simply running the process.

The main focus of the review was reducing the large number of design constraints being used in the design. With a large number of commodities being designed at national, state and Natural Resource Management (NRM) level, the number of design constraints quickly reaches over 1,000. This complicates the design in terms of analysing design results to improve the design as well as increases computing time to run through different options.

The review concentrated on whether designing at national level would naturally achieve reasonable results at the state and NRM level with the logic that in order to achieve the national Relative Standard Error (RSE), sample would need to be devoted to the particular states and regions which contributed the most to the national estimate, thereby also achieving reasonable RSEs for these important states and regions.

The results suggested that the number of constraints could be cut to just over 100, with these being mainly at the national level, with some broad level information (e.g. total area of holding) for all states and regions. A small number of commodities were added into the design at the request of the Rural Environment and Agriculture Statistics Branch (REASB) where there were some concerns regarding state or regional RSEs.

This more efficient design process enabled the REACS 13/14 design to be conducted in a much shorter timeframe. This also allowed BSM to improve the design to a point where sample size savings were achieved with a reduction in sample size from 36,531 for REACS 12/13 to 34,694 for REACS 13/14. Following completion of processing of the 13/14 survey, an evaluation will be undertaken to determine whether productivity savings have been achieved and the scope for further savings in the future.

Further Information

For more information, please contact Brett Frazer (07 3222 6028, brett.frazer@abs.gov.au)

Uniqueness Analysis

Data Integration, Access and Confidentiality Methodology Unit (DIACMU) is currently developing methods to evaluate the feasibility of linking datasets prior to the actual linking process and to help identify disclosure risks in linked datasets. One method recently developed is a “uniqueness analysis” on the input datasets.

Data linking involves bringing together records from two or more datasets belonging to the same unit. The process produces a unit record file containing analysis fields from the input datasets for the common population. It is a cost-effective method of acquiring more comprehensive statistics. The recent release of the Australian Census Longitudinal Dataset (ACLD) was an important milestone for data linking in the ABS.

Ideally, datasets should be linked with a high degree of accuracy and coverage. Data linking is only feasible if there are linking variables on datasets that can uniquely identify individual record pairs belonging to the same unit. The more record pairs uniquely identified by a combination of linking variables, the more likely that high quality links are established. It is important to ascertain the likely success of a linking project before undertaking the project.

Uniqueness analysis determines the proportion of records on a single file which are uniquely identified by their values on a combination of variables. It provides a guide to the upper bound of records that could be uniquely linked using the available variables (Conn and Bishop, 2005). For example, if one could uniquely identify 80% of records on File A, but only 50% on File B, then the upper bound for the match rate would be 50%. This is considered an upper bound as errors or changes in linking fields can occur across the two datasets. This analysis helps inform whether a linking project is feasible, and furthermore, provides insight into the optimal linking strategy. This method extends the work of Conn and Bishop in the following ways:

1. investigating the marginal improvement in the percentage of uniquely identified records by increasing the number of variables in the combination of potential linking variables
2. taking into account non-response in linking variables in calculating that percentage.

It is envisaged that a uniqueness analysis will be conducted on linked datasets to discover the relationship between the percentage of uniquely identified records and the linkage accuracy.

Besides data linking, DIACMU is also investigating methods to more efficiently mitigate disclosure risks in disseminating data on TableBuilder and DataAnalyser. Linked datasets released on TableBuilder and DataAnalyser include the ACLD and the Australian Census and Migrants Integrated Dataset. A uniqueness analysis on linked datasets can help quickly identify disclosure risks prior to their release. Thus, the uniqueness analysis can potentially have multiple applications besides determining the feasibility of linking datasets. It also gives DIACMU a guide to the best way in ensuring the relevance of linked datasets while maintaining confidentiality.

References
Conn, L & Bishop, G (2005). Exploring Methods for Creating a Longitudinal Dataset, cat. no. 1352.0.55.076, **Australian Bureau of Statistics**, Canberra.

Further Information

For more information, please contact Charles Au (02 6252 5990, charles.au@abs.gov.au)

The ABS Privacy Policy outlines how the ABS will handle any personal information that you provide to us.

How to Contact Us and Email Subscriber List

Methodological News features articles and developments in relation to methodology work done within the ABS Methodology and Data Management Division. By its nature, the work of the Division brings it into contact with virtually every other area of the ABS. Because of this, the newsletter is a way of letting all areas of the ABS know of some of the issues we are working on and help information flow. We hope the Methodological Newsletter is useful and we welcome comments.

If you would like to be added to or removed from our electronic mailing list, please contact:

Peter M Byron
Methodology & Data Management Division
Australian Bureau of Statistics
Locked Bag No. 10
BELCONNEN ACT 2617

Tel: (02) 6252 6804
Email: p.m.byron@abs.gov.au