Methodological News
A Quarterly Information Bulletin
Australian Bureau of Statistics
ABS Methodology and Data Management Division
December 2013

Articles

Constructing an experimental household level socio-economic index

The ABS produces Socio-Economic Indexes for Areas (SEIFA) which seek to summarise the socio-economic conditions of an area using relevant data collected from the Census. The indexes provide information about the area in which a person lives, but within any area there are likely to be smaller groups with characteristics different to the overall population of that area.

Constructing socio-economic summary measures for finer units such as households would enable researchers and policy makers in Australia to better differentiate between areas with varying concentrations of advantage and disadvantage.

Research was conducted into the construction and dissemination of an experimental household level index. The results were presented to the Methodology Advisory Committee (MAC) in June 2013. Mesh Block, household, family and individual levels were all considered as units of output for the experimental index.

A household level index was chosen as it complements the area level rankings by adding more depth to the information given by SEIFA, as well as providing its own valuable insights by allowing users to make more accurate inferences about smaller units. Households are a clearly definable fine-level unit in which individuals tends to share socio-economic conditions. Furthermore, a household level index maximises the population included in the index given the Census data available.

The summary measure was derived from binary indicator variables constructed from Census data which measured household advantage and disadvantage. Principal Component Analysis (PCA) was used to specify weights for the variables. Occupied private dwellings were the base unit for analysis. Households which had more than 10 non-responses for the relevant Census input data items were excluded, which accounted for 2% of the population. Remaining records with missing responses were coded to the neutral response (0) rather than the advantaging or disadvantaging characteristic (1). These decisions were deemed appropriate based on conceptual validity, a literature review and user familiarity with previous SEIFA research and products.

Avenues identified for disseminating an experimental household level index include using Census TableBuilder to enable cross-classifications with other Census variables and including counts of households with current SEIFA output to pinpoint diversity within areas.

There was interest from the MAC members to see the ABS continue pursuing the release of an experimental household level index product. There is great demand externally, especially from service providers who currently use SEIFA to plan and monitor their services and would benefit from a measure at a finer level. Currently the SEIFA team is investigating producing a finer-level index at the household level in response to MAC feedback and user demands.

The paper presented to MAC is available here Research Paper: Building on SEIFA:

Finer Levels of Socio-Economic Summary Measures (Methodology Advisory Committee) (ABS cat. no. 1352.0.55.135).

Further Information

For more information, please contact Phillip Wise (02 6252 7221, phillip.wise@abs.gov.au) or Courtney Williamson (07 3222 6031, courtney.williamson@abs.gov.au)

The Semantic Web and Official Statistics

On 22 October, 2013, Dr Siu-Ming Tam, Chief Methodologist and First Assistant Statistician at the ABS, was invited to deliver a key note address to the *First International Workshop on the Semantic Web and Official Statistics.*

This was the first time that Semantic Web specialists and official statisticians have gathered together to discuss how work on the Semantic Web may advance official statistics and vice versa.

The Semantic Web is also broadly described as Web 3.0. However, if one poses the question "What is the difference between Web 2.0 and Web 3.0", most official statisticians will find it difficult to answer the question, and not surprisingly, a tongue-in-cheek answer would be "1.0".

Dr Tam believes that the Semantic Web, together with Big Data and the active international collaboration work to reform and re-engineer statistical production processes, are three key developments that will significantly change the landscape of official statistics in the next five years.

Putting it simply, the Semantic Web is a certain way to organise, describe and annotate rich content (statistical data, text, imagery, etc) – supported by a framework of international standards – that will facilitate the discovery, exchange, retrieval, processing and analysis of information from many disparate sources.

The Semantic Web aspires to support a global web of data for consumption by people, machines or people assisted by machines, just like the web of documents supported by Web 1.0 and enriched by Web 2.0.

An example may help to illustrate this. In another talk presented in Abu Dhabi earlier this year, Dr Tam posed the following question to a statistical audience "How many Web 3.0 companies are there in Abu Dhabi?" By googling "Abu Dhabi Web 3.0 companies", Dr Tam advised he achieved 65 million hits, and yet there are only about 110,000 companies reported to be registered in Abu Dhabi by the Abu Dhabi Statistics Centre.

So what is not working? The answer lies in the fact that the web-based descriptions of the activities of the companies do not allow for this type of research to take place accurately.

The solution, offered by Semantic Web technologies, lies in two key enhancements. Firstly, structuring the data in "triples" and relating object to standards through Uniform Resource Identifiers creates a network of web-accessible linked information. Secondly,

the meaning of this data – its "semantics" – is described along with its structure and format in a machine-interpretable way.

The challenges facing data scientists are to better articulate the value proposition, and raise awareness of the Semantic Web; and the challenge for official statistics is in harnessing the opportunities provided by the Semantic Web to improve the business of official statistics.

Further Information

For more information, please visit the SemStats 2013 website or contact Siu-Ming Tam (02 6252 7160, siu-ming.tam@abs.gov.au).

Data Mining and Editing

Various data mining algorithms are being explored with a view to improving our data editing processes, especially for large administrative datasets.

To develop an editing strategy for a new dataset, we need to understand the relationships between the variables, learn how to detect anomalous observations, and develop criteria for detecting and treating errors. This is often an exploratory process, using a variety of ad hoc methods.

As the ABS handles more data every year, there is an increasing emphasis on automating as much of this process as possible. One special challenge is to find appropriate *edit rules:* that is, to describe a set of logical constraints that error-free data should satisfy. Traditionally this has required a great deal of input from subject matter specialists and has been a very labour-intensive process.

Data mining offers a possible solution to this challenge, using machine learning algorithms to identify outliers and to find and describe patterns in the data. So far, we have obtained promising results from three different methods.

Cluster analysis is a collection of methods for sorting records into a number of related groups. *Hierarchical agglomerative clustering* is an iterative method of doing this. Start by finding the two most similar records: this forms the first group. Then look for the next closest pair, and decide whether they should join the first group or form a new group. Repeat the process until all records are joined into a small number of groups. This process is easy to automate, and we can identify which records are quick to join a nearby group, and which ones remain isolated until near the end of the process. The latter records are the ones most likely to be anomalous.

Random forests fall into the class of *ensemble methods*. These use repetition and randomness to improve the decision making process. A large number of decision trees are generated, each one representing a simplistic model of the data, and each one incorporating some randomness. The average of all these models gives more precise results than any of the models singly. Furthermore, each branch of a decision tree represents an edit rule.

Association rule mining was originally applied to supermarket transactions: what items are typically purchased together? There are several algorithms for discovering such

associations, and we are investigating ways to express associations as edit rules. The same concepts can be applied to survey data and administrative data: which characteristics or responses are typically found together in the same record?

These three methods, and other machine learning algorithms, are potentially valuable tools for understanding new sources of data and for streamlining our existing processes. There is also possible application of these methods to other areas, for example modelling for small area estimation, and output validation for population surveys.

Further Information

For more information, please contact Claire Clarke (08 8237 7468, claire.clarke@abs.gov.au), Kevin Mark (08 8237 7631, kevin.mark@abs.gov.au) or Alexander Hanysz (08 8237 7434, alexander.hanysz@abs.gov.au)

Monitoring the Impact of Web Form Introduction in the Monthly Population Survey

As part of the strategy for implementing web-based data collection (also known as web forms) for the Monthly Population Survey (MPS), a controlled experiment was run on the incoming sample to detect any major impacts of the process change on key Labour Force estimates. For the households brought into the MPS sample between May and September 2013, 50% were given the option to complete the MPS via web form while the remainder were only offered the traditional process of personal/telephone interview.

Each month until April 2014 (when the last of these split samples completes its final month of MPS), estimates of the differences in labour force status between the two groups are produced and compared. These estimates are produced via a composite estimator, which includes all data collected from the split samples since the beginning of the experiment; this gives an estimate that makes maximum use of the available data, and helps control for seasonal and time-in-survey effects. The method is based on that used previously to measure the impact of introducing telephone interviewing in 1996-97, and computer-assisted interviewing in 2003-04.

The impact measurement strategy aims to measure the net impact of introducing web forms to the collection, including modal effects, effects related to the offer of a web form option to respondents, and changes to existing procedures that may affect respondents in all modes; other methods to estimate the specific modal effect of web forms are currently being investigated.

Further Information

For more information, please contact Chris Mann, (02 6252 6758, chris.mann@abs.gov.au) or Anna Poskitt, (02 6252 5668, anna.poskitt@abs.gov.au).

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:

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

Tel: (02) 6252 7037
Email: methodology@abs.gov.au