Methodological News

A Quarterly Information Bulletin by ABS Methodology Division

MARCH 2018 EDITION

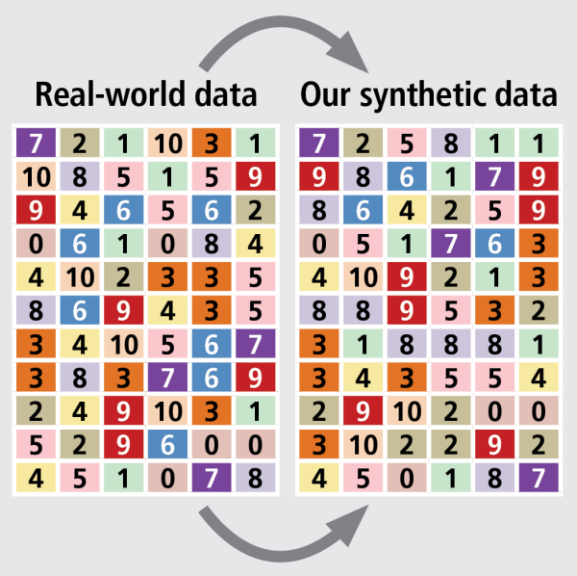
Contents

An Evaluation of the Feasibility of Producing a Prototype Synthetic Dataset using Innovative Methodologies

The ABS is committed to increasing the accessibility of its' valuable microdata while ensuring data confidentiality. In addition to existing microdata protection techniques, the ABS is undertaking exploratory work into synthetic microdata as an alternative approach to allow researchers to analyse the microdata while maintaining ABS's legislative obligations. An important benefit will be an expansion of ABS's confidentialised product suite to better accommodate the needs of the growing market niche of sophisticated users.

In a partially synthetic dataset, some of the original records - especially those sensitive ones - are treated as 'missing values' and then are replaced by synthetic values generated from an imputation model. A high quality synthetic dataset ensures that the confidential information is disguised so that the disclosure risk is kept minimal and valid inferential results of some estimates of interest are preserved from the original data.

A case study to examine the possibility of a partially synthetic data has recently been explored using the 2006-07 to 2010-11 Business Longitudinal Database (BLD). A good imputation model of the BLD should be able to mimic the data generating process well, in terms of capturing the complex dependencies among the variables of mixed type and the longitudinal structure. Three classes of multivariate imputation models are investigated: sequential regression, sequential random forest and Bayesian copula model, with random effects added to take into account the business specific effects. In the sequential regression and sequential random forest approaches, each variable to be synthesised is specified a generalised linear mixed model and a mixed effects random forest model respectively. In the copula based model, all the variables are modelled jointly, such that the latent variables transformed from the observed data are assumed to follow a Gaussian distribution.



The utility of the synthetic data sets is evaluated through comparisons of the parameter estimates in some models from published research, with those from the synthetic data and the original data. Simulation results suggest that the Bayesian copula model leads to the best utility in some longitudinal data analyses on the BLD, followed by the sequential random forest approach.

The ABS intends to build upon this research exploring the potential of developing synthetic data as a dissemination tool to enhance public access to microdata while ensuring

confidentiality. Some considerations include: the demanding computation in big volume data sets, the costs in building more complicated imputation models, evidence that synthetic datasets meet ABS's confidentiality requirements and the trade-off between utility and confidentiality.

For more information:
Jiali Wang
u5298171@anu.edu.au
Bernadette Fox
Bernadette.Fox@abs.gov.au

Complementary Measures for Non-Response Follow Up

Response rate alone is not an ideal measure to work with for understanding the quality of a data collection process. When the response rate is high we can be confident in the quality of outputs but it can be very expensive and operationally inefficient to strive to achieve high response rates. When the response rate is not as high it is not obvious what the implication is for output quality as lower levels of response don't always correspond to lower quality outputs. Blindly seeking increased response rates through extra acquisition effort can have impacts on the data acquisition workforce and other surveys in the field. A number of alternate quality measures have evolved within the research field of Adaptive and Responsive Design. The ABS is currently looking into whether we can use alternate measures to complement the response rate and organise our non-response follow up procedures around a suite of measures. High response rates would still justify ending the follow up process but lower response rates could also be accepted provided other measures hit target values that give us confidence in the quality of outputs.

One measure for understanding the extent of bias that non-response can introduce is the R-indicator (Schouten et al 2009). This measures the variation in response propensities where the propensities can be estimated using correlates of both the response indicator and the key output variables. Partial R-indicators allow for identification of where the sample is most unbalanced and hence inform decisions on

where to prioritise follow up efforts (Schouten & Shlomo 2017). Unfortunately the response propensities that the R-Indicator relies on have been difficult to estimate accurately. To avoid having to estimate these we have been developing a Balance Indicator that makes use of past data to partition the sample into homogeneous groups. The indicator only requires the average propensity for the group and the variation in these averages tells us about the potential for non-response bias.

We found that the indicator works well for household surveys where each responding unit is of roughly equal importance to estimates. However, this is not the case for business surveys, where skewed populations ensure groups cannot be of both equal size and equal importance to estimates. Weighting the group by its expected share of the population estimate can address the skewness but makes the indicator specific to the variable used in the weighting. Explicit imputation is another challenge for business surveys. Explicit imputation is used extensively to deal with non-response and cannot be ignored when assessing non-response bias. Multiple Imputation has been proposed as a way of understanding the extra variation that non-response is causing (Wagner 2010). If all the non-response is in parts of the sample where we can impute well then we are less concerned than if it is in parts where we can not impute well. The multiple imputation captures the impact to sampling error but we will still need to understand how imputation is influencing our measure of non-response bias.

The non-response follow up strategy will need to work together with the weighting and estimation procedures and data acquisition operational requirements. If we can accurately estimate the extent of non-response bias it would be sensible to try and remove this via a weighting adjustment. The balance between removing bias through data collection and removing bias through weight adjustment will depend on the size of the data collection budget and a careful assessment of data quality to ensure the output is fit for purpose. For a given budget, the goal in data collection is to allocate follow up effort in a way that will reduce the reliance on the subsequent weight adjustment. The non-response follow up strategy aims to use a suite of measures to define an acceptable region for stopping follow up and then prioritise units for follow up in a way that gives us the best chance of moving into that region in the most efficient manner. This strategy poses a challenge for business processes because data collection procedures will need to be sufficiently flexible to ensure dynamic adjustments to collection requirements can be absorbed in a timely manner.

For more information:
Carl Mackin
Carl.Mackin @abs.gov.au
Daniel Fearnley
Daniel.Fearnley@abs.gov.au

Correcting Biases in Estimation when Linkage Errors are Present in a Probabilistically-Linked Dataset under one-to-one Linking Scenario

Computerised probabilistic record linkage (CPL) attempts to link records belonging to the same individual in multiple datasets when unique identifiers are absent. Linking multiple sets together allows more statistical analysis to be performed because the linked dataset contains more analysis variables than in each individual dataset. CPL may link two records in two different datasets if they share similar values across several attributes (referred to as linking fields), such as date of birth, age and gender. Even so, the two linked records may sometimes correspond to different individuals. Therefore, linkage errors (records

in different dataset not belonging to the same individual are linked) are likely to be present in the linked dataset, leading to estimation bias.

To correct estimation bias, the ABS is investigating a weighting approach. Specifically, a weighting matrix assigns each possible link a weight, where the weight takes into account the linkage error process. Analysts can then perform unbiased standard analysis using the weights. The main issue is modelling the linkage error process, called a Linkage Error Model (LEM), so that the effects of linkage error can be reversed. The LEM in Chambers et al. (2009) is consistent with the assumption that the probability of a link does not depend upon the values of the observed linking field. However, this requirement does not always hold in practice. In my work I We have considered relaxing this assumption by conditioning on observed linking fields (e.g. males are less likely to be linked to females). To estimate our LEM and relax the above assumption, we use a latent model to simulate the linkage error process. The whole process is fully computerised.

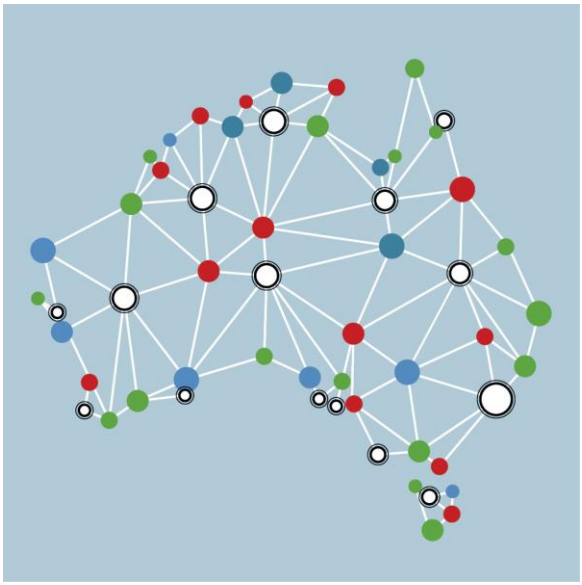
As expected, our simulation result shows that the performance of our LEM and Chambers et al 's LEM model are unbiased if linking fields and covariates in the model are independent. However, when the independence assumption is violated, such as when one of the linking fields is a covariate of the response variable, our LEM leads to estimates with very little bias while Chambers et al' s LEM could be heavily biased. Further directions would be to apply this method to a real linkage situation and to use the method to estimate the proportion of links that are correct in any given linkage exercise.

For more information:
Yue Ma
ym894@uowmail.edu.au
James Chipperfield
James.Chipperfield@abs.gov.au

An Empirical Bayesian Approach to Entity-Based Data Linkage

The ABS's involvement in cross-government data integration initiatives such as the Multi-Agency Data Integration Project (MADIP) and the Data Integration Partnerships for Australia (DIPA) are presenting new opportunities to improve the ABS's data linkage methodology. An important pursuit as part of these initiatives, is the development of an entity-based approach to data linkage, where records in datasets are linked to entities in a population spine. This represents a paradigm shift from the ABS's existing linkage methodology which is tailored to conventional record linkage across pairs of datasets. While the existing methodology can

be adapted to the population spine, the empirical Bayesian approach has the potential to give more statistical validity to inferences about which records constitute an entity in the spine. For the creation of a linkage spine, existing methods such as probabilistic linkage (Fellegi & Sunter 1969) and deterministic linkage (ABS, 2016) need to be combined with a post hoc conflict resolution procedure. A separate issue with these methods is their inability to provide principled estimates of linkage uncertainty for analysis. If such estimates were available, they could be propagated through analyses on the spine-linked data to provide more statistically sound inferences.



With these areas for improvement in mind, the ABS has been investigating alternative, state-of-the-art approaches to entity-based linkage. One method which is well suited to a number of important ABS requirements is the empirical Bayesian approach described in (Steorts 2015), known as ebLink. Unlike many alternative methods, ebLink directly models the entities in the domain (e.g. Australian residents) and the links from records to entities. This makes it a good fit for the population spine, since it can provide a statistical measure of the association between spine entities and dataset records. It also includes the usual benefits of a Bayesian framework, namely: accounting of uncertainty through the posterior distribution, the ability to incorporate prior information, and the facilitation of complex hierarchical models.

In order to assess the feasibility of ebLink, the ABS is collaborating with the University of Melbourne through the APR.Intern programme. The poor scalability of ebLink was quickly identified as an obstacle, but has been somewhat mitigated by a re-parametrisation of the model that incorporates blocking ideas, and enables the inference to be distributed across a compute cluster. As part of the collaboration, a prototype is being implemented in Apache Spark (a distributed computing framework). Early experiments indicate that ebLink slightly outperforms the ABS's established methods in terms of linkage accuracy, while also providing a full posterior distribution over the linkage structure. However, computational efficiency/scalability remains a challenge for future work.

References

Fellegi, Ivan; Sunter, Alan (1969). "A Theory for Record Linkage". *Journal of the American Statistical Association*. **64** (328): pp. 1183–1210.

Steorts, Rebecca C. (2015). "Entity Resolution with Empirically-Motivated Priors". *Bayesian Analysis*. **10** (4): pp. 849-875.

ABS (2016). "Personal Income Tax and Migrants Integrated Dataset (PITMID) 2011-

12 Quality Assessment". *ABS Research Paper*. cat. no. 1351.0.055.060

For more information:
Neil Marchant
Neil.G.Marchant@unimelb.edu.au
Daniel Elazar
Daniel.Elazar@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 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:

Nick Husek
Methodology Division
Australian Bureau of Statistics
Locked Bag No. 10
BELCONNEN ACT 2617

Email: methodology@abs.gov.au

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