Submission to Productivity Commission inquiry into Data Availability and Use

The National Centre for Social and Economic Modelling (NATSEM) is part of the Institute for Governance and Policy Analysis (IGPA) at the University of Canberra. NATSEM welcomes the Productivity Commission's preliminary report on data availability and use. Staff at NATSEM have been involved in a number of workshops over the past 3 years on accessing Government datasets, and welcome this report which provides draft recommendations on the way forward for Australia on dataset access.

Over the last 20 years, NATSEM has led the country in modelling social and economic conditions, focussing mainly on developing models of individual behaviour called microsimulation models. NATSEM's Tax/Transfer microsimulation model (STINMOD) has been used extensively to inform the public on the impacts of Government policy, and until 2013 was developed with the Commonwealth Government and informed Commonwealth Government policy. Since the Commonwealth left the partnership in 2013, NATSEM, IGPA and the University of Canberra have continued to use and develop the STINMOD model, using it to inform the Australian community of the impact of Commonwealth budgets and Tax/Transfer policy.

NATSEM has also developed, in collaboration with Australian Government departments, a dynamic microsimulation model called APPSIM, and uses this model, with a group from the University of Sydney, to model the impact of health issues like diabetes and depression (Schofield et al., 2014; Veerman et al., 2015).

Microsimulation models are very powerful, allowing the impact of policies on subgroups of the population to be identified (Tanton et al., 2009), as well as allowing small area estimates to be derived (Tanton, 2011). However, they are data intensive. All microsimulation models require confidentialised unit record files to run. The small area microsimulation models used by NATSEM (SpatialMSM) require a confidentialised unit record file (CURF) from a survey; and reliable small area data from the Census. These CURFS mainly come from the Australian Bureau of Statistics (ABS), although the HILDA dataset has also been used by NATSEM for spatial microsimulation purposes.

The development of static microsimulation models in Australia so far has not been hampered by data availability, as the only data required are a confidentialised unit record file from a survey like the ABS Survey of Income and Housing. So far, this confidentialised (de-identified) unit record file has always been available from the ABS, and for this modelling that informs the Australian public on the impact of changes in tax/transfer policy, this availability must continue.

Dynamic microsimulation models are much more data intensive, and developing NATSEM's dynamic microsimulation model was a 5 year process that was significantly hampered by data availability. NATSEM's dynamic microsimulation model APPSIM requires more than one CURF,

as dynamic models require dynamic processes like marriage, illness, death, labour force status, etc, to be modelled. This then requires extensive longitudinal data, to be able to identify transitions into these various states. With even more complex data requirements are dynamic spatial microsimulation models, as each of these transitions needs to be modelled for each area where estimates are required (Rephann, 2004). One of the few countries able to develop a dynamic spatial microsimulation model is Sweden, where large amounts of administrative data are available.
Any data that we could use was incorporated into APPSIM, including unit record files from the ABS (the Household Sample File), the unit record file from HILDA for transition probabilities, and administrative data. This model was 'modularised', which meant that modules were developed for particular purposes, eg, a health module; a demographic module; etc. This also meant that each module was only as good as the data we could get for the module.

While these models are data intensive, NATSEM also conducts other research which does not use microsimulation models, but is also highly data intensive. Much of this research is hampered by a lack of data – for example, we recently developed an Index of Wellbeing for Older Australians for the Benevolent Society, and we couldn't get any data on small area health outcomes for older people. We applied to get data for hospital admissions from the Commonwealth, but by the time we had gone through the process of filling out the form and getting a response, the index had been finalised. For this project, we weren't even asking for unit record data – we required data for hospital admissions for standard ABS geographies.

In another example, we required a confidentialised unit record file from the ABS Indigenous Social Survey to model spatial indigenous disadvantage, estimates which would greatly assist indigenous communities and researchers. This CURF was not available from the ABS due to data confidentiality reasons. As NATSEM felt that the estimates were important for indigenous Australian communities, and the estimates would be widely available from the Australian Urban Research Infrastucture Network (AURIN), NATSEM used a microsimulation process and aggregate data from the Indigenous Social Survey to create a synthetic CURF to then use in our spatial microsimulation model (Vidyattama et al., 2015). This is not as reliable as using the official CURF from the ABS, but when the CURF is not available, then a synthetic CURF is the next best tool.

We would argue that we have seen many situations where important research could have been conducted, but was hampered due not to a lack of data, but a lack of availability to data, or a complex process to access the data. The data were there, they were just not available to researchers.

Examples of datasets which would have significantly assisted our research are spatial data on health and hospital admissions by age for indexes of disadvantage; data on domestic violence by area for a risk of homelessness index we have published (we did get estimates for some States, and created a separate index for the States where these indexes were available); unit record file for indigenous survey for a microsimulation model of indigenous disadvantage; and cadastre data for buildings in cities to assign synthetic populations to for analysing public transport use, road use, service provision, etc. All these data are held by Commonwealth or State Government departments, but either not released, or released after a great deal of bureaucracy.

Looking internationally, there seems to be much more openness about data availability overseas. As an example, the CURF for the US General Social Survey can be downloaded by anyone, including international persons, directly from http://gss.norc.org/Get-The-Data with no sign up requirements. In many cities, up to date mapped data are also widely available – for example, the Chicago Data Portal allows mapping of crimes to block and street level (https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present-Map/c4ep-ee5m). In the UK, large linked health datasets like the 500,000 respondent UK Biobank are contributing to a greater understanding of health (https://www.ukbiobank.ac.uk/), and large funded research institutes like the Consumer Data Research Centre (https://www.cdrc.ac.uk/) and the Administrative Data Research Centre – England (https://adrn.ac.uk/about/research-centre-england), both funded by the Economic and Social Research Council (ESRC), the ARC equivalent in the UK, are opening up Government and Commercial data sources. Overseas countries have recognised the potential of turning large amounts of data into information to create knowledge for the community, and have put significant funding into these initiatives, something that has not happened as yet in Australia. The productivity commission recognises this in their report, and from a user perspective, we would support this impression.

In terms of the Productivity Commission suggestions, NATSEM would support opening up access to public and private sector datasets, and NATSEM staff have been involved in a number of forums and discussion in the past 3 years on opening up Government data. While there has been much interest from researchers, we would argue that the reaction from Government Departments so far has not been as enthusiastic.

The trusted user model suggested by the productivity commission is based upon the trusted access model now adopted by the ABS for accessing ABS data, which staff at NATSEM have been involved in commenting on through the ABS Web User Forum. Providing additional funding for some organisations that hold data to become Accredited Release Authorities, to be able to implement the trusted user model, appears sensible, as long as this does not imply a new administrative layer for the researchers to negotiate. We would argue that if an academic institute is given a "trusted" status, and has a regular audit of their IT processes, then this trusted staus should be recognised by all ARA's, rather than assessed separately for each ARA.

In terms of data developed through ARC grants, NATSEM would support distribution of data developed through grants, and already distributes our small area data through online maps (http://www.natsem.canberra.edu.au/maps/), downloadable datasets (available through the online maps), and the AURIN portal. NATSEM is a strong believer in data collaboration, and increasing the use of any outputs from our research.

We would also support the recommendation that value added to datasets by researchers should be retained and available for other researchers (Draft Recommendation 5.3), particularly if the value added was funded by an ARC grant. However, we would suggest a timeframe for open availability, as the publication process that researchers go through can take up to a year from beginning to end, and the primary researchers need to have the opportunity to publish their work before the dataset is opened up to other users.

In terms of charging for public data (Daft Recommendations 7.3 and 7.4), we would suggest a process similar to that used by the ABS for academics, in that Universities Australia pays an annual subscription to the ABS for free academic access to basic ABS data. More complex requests for ABS data are dealt with by the ABS on a cost recovery basis, however the ABS does take a relaxed approach to this, and have been known to send trusted academics aggregate datasets for research with no charge. We would generally support low cost or free access to high value datasets with little or no value add, and cost recovery for value added datasets.

A final note we would like to make is that nothing in this area should be rushed. Communication with the public, and getting the public and any privacy groups on board with any changes is essential if Australian data in the future is to be reliable. If a collection agency's reputation is damaged in any way, then the data provided will be under question. Any new legislation drawn up needs to protect the respondents while also making data available to trusted researchers, and this needs to be made known to respondents. The UK experience with the UK Biobank is that there are 500,000 respondents happy to provide their information if it can make the UK a more healthy place while also maintaing their privacy, and this needs to be made clear to respondents of Australian surveys. The productivity commission recognises this trust in Chapter 1 of their draft report, but this then needs to be supported by any new legislation.

Overall, as users of public data, and potential users of a much more opened set of Government data to inform research that then informs Australian policy, NATSEM would support a much more open attitude from the Government (State and Commonwealth) on data release. The trusted user model would enable this, and a regular audit of a Research Institute's IT and data storage processes would support this trusted user model, and would put the onus on the research institute to maintain their systems and processes to keep their trusted user status, and hence their access to data. We would argue for a research institute trusted user model, rather than a University trusted user model, as different Institute's within the same University will have different data requirements and different incentives to maintain a trusted user status.

NATSEM Staff involved in submission: Professor Robert Tanton, Professor Laurie Brown, Dr Jinjing Li, Dr Yogi Vidyattama, Dr XiaoDong Gong, Dr Riyana Miranti

Contact:

Robert Tanton
NATSEM
Institute for Governance and Policy Analysis
University of Canberra
ACT 2601

References

Li, J., O'Donoghue, C. and Dekkers, G. (2014), "Dynamic Models", *Handbook of Microsimulation Modelling*, inbook, , pp. 305–343.

Rephann, T.J. (2004), "Economic-Demographic Effects of Immigration: Results from a Dynamic Spatial Microsimulation Model", *International Regional Science Review*, Vol. 27 No. 4, pp. 379–410.

Schofield, D.J., Cunich, M., Shrestha, R.N., Callander, E.J., Passey, M.E., Kelly, S.J., Tanton, R., et al. (2014), "The Impact of Diabetes on the Labour Force Participation and Income Poverty of Workers Aged 45–64 Years in Australia", edited by Zhang, H.*PLoS ONE*, Vol. 9 No. 2, p. e89360.

Tanton, R. (2011), "Spatial microsimulation as a method for estimating different poverty rates in Australia", *Population, Space and Place*, Vol. 17 No. 3, pp. 222–235.

Tanton, R., Vidyattama, Y., McNamara, J., Vu, Q. and Harding, A. (2009), "Old, Single and Poor: Using Microsimulation and Microdata to Analyse Poverty and the Impact of Policy Change among Older Australians", *Economic Papers: A Journal of Applied Economics and Policy*, Journal Article, , Vol. 28 No. 2, pp. 102–120.

Veerman, J.L., Shrestha, R.N., Mihalopoulos, C., Passey, M.E., Kelly, S.J., Tanton, R., Callander, E.J., et al. (2015), "Depression prevention, labour force participation and income of older working aged Australians: A microsimulation economic analysis", *Australian & New Zealand Journal of Psychiatry*, Vol. 49 No. 5, pp. 430–436.

Vidyattama, Y., Tanton, R. and Biddle, N. (2015), "Estimating small-area Indigenous cultural participation from synthetic survey data", *Environment and Planning A*, Vol. 47 No. 5, pp. 1211–1228.