Data Service as a Platform

Measuring household dynamics in South Africa: constructing a household panel

The challenge

Household surveys are invaluable because they allow social scientists to explore the relationships between demographics, labour-market outcomes, incomes, household composition, service delivery (such as access to electricity) well-being, and other crucial details of people's lives. However, cross-sectional household surveys offer only a snapshot of where individuals are living within a given period and of their outcomes at that specific time: they thus fail to capture how people move between households and locations, or how their outcomes change over time. While longitudinal household panel surveys do track individuals over time, they do not track the household units themselves. Thus, despite their advantages, they do not allow insights to be drawn on longitudinal changes in household-level variables, or longitudinal changes in the structure (or existence) of the household units themselves. This case study sets out a methodology for reworking data to enable the observation of household change.

> Our aim was to support researchers in understanding and analysing large and complex datasets, focusing on using the power of popular statistical software like R in a big data environment.

The UK Data Service approach

At DataFirst, based at University of Cape Town (UCT), research by Martin Wittenberg, Tom Harris and Mark Collinson (2017) sought to address this challenge, and developed a new methodological approach, using typical household panel survey data, that would enable them to track longitudinal changes in particular household-level outcomes; in this case, transitions in household electricity access.

They highlight two key gaps in the literature in this regard. Firstly, existing research on service delivery issues, such as on electricity use/access, tends to focus on comparing cross-sectional statistics across different periods. Where longitudinal studies do exist, these studies focus on investigating longitudinal transitions in energy outcomes for individuals, but do not take into account the importance of the household unit or changes in household-level access. Existing studies have also not considered the complexities of access transitions in developing contexts, or recognized that aggregate data sources do not offer rich-enough information on the time dynamics of these outcomes. The research team suggested that a richer picture can be gained by investigating the longitudinal dynamics of household electricity access outcomes using large-scale longitudinal survey data.

Secondly, they observe that the dynamics of household structures and household composition are not well understood, and are even less well documented. Research that aims to develop a better understanding of household formation and dissolution processes is therefore crucial to understanding and predicting the process of economic development, as well as to understanding the dynamics of individual well-being.

UK Data Service

While cross-sectional survey data can be used to calculate net household formation statistics, these are limited in three respects:

- they provide no data on the actual numbers or proportions of households that dissolve, form, or continue to exist
- they provide little insight on the determinants of household formation and household composition
- they fail to explain the changes that occur within those household units that continue to exist, and how these may relate to individual well-being

In order to explore such questions, we require longitudinal data that enables one to track household units over time, that is, a panel of household units, in other words, a household-level panel.

Data and data issues

Most national household panel survey data can, in fact, be sourced in a standard individual-level format, and then restructured so that they take the form of a panel of households. When investigated through this new lens, these rich data will provide new insights on household-level dynamics. This approach also offer enormous potential to researchers to investigate the longitudinal dynamics of household structures, such as formation and dissolution activity, and the dynamics of various components of household service delivery, such as electricity. As part of the highlighted project, the researchers focused specifically on applying this novel restructuring or reshaping approach to the National Income Dynamic Study (NIDS), a household survey data source that is described below.

National Income Dynamics Study (NIDS) overview

The National Income Dynamics Study (NIDS) is a panel study commissioned by the national Presidency of South Africa in an effort to track long-run poverty and well-being. It focuses on income, expenditure, labour market participation, education, health and household well-being, such as access to services. The baseline sample was designed to be nationally representative and consists of around 7,300 households and about 28,000 individuals, who became Continuing Sample Members (CSMs) for the subsequent waves. Babies born to CSM women become CSMs themselves, and individuals who were co-resident with CSMs were also interviewed as Temporary Sample Members (TSMs). The data are available as public use files from the South Africa DataFirst archive, as shown in Figure 1.

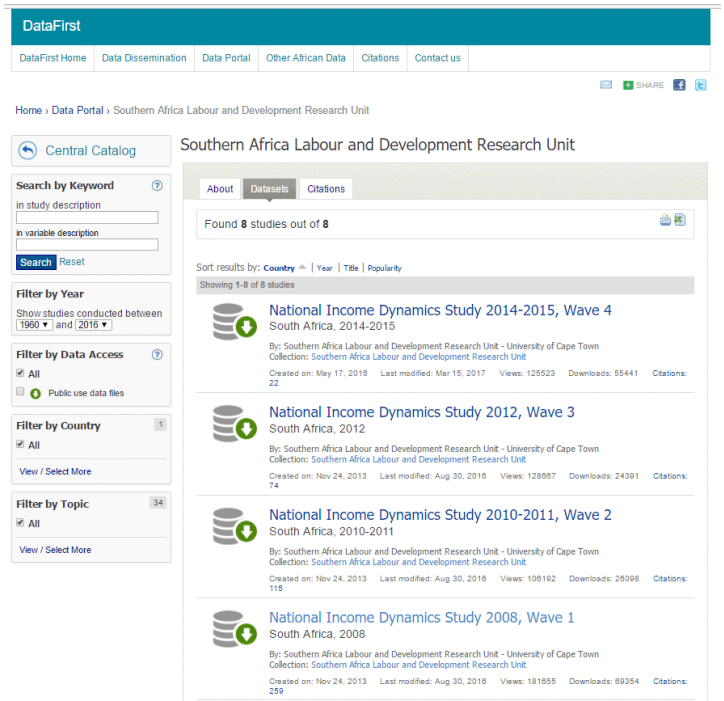

Figure 1. DataFirst catalogue record for NIDS data

Issue 1: Shape of data

As with most household panel survey datasets, NIDS currently exists in the form of an individual-level panel. Individual respondents observed in the data can be followed or tracked from one period of data to the next using unique individual-level panel identifiers. However, with the data in its current form, the same 'tracking' cannot be done for household units. Household units are allocated different households identifiers in each observational period, even when the household structures/ composition remain the same. The data providers/publishers thus remain agnostic as to whether households observed across two periods can be identified as the same unit, where they can be identified as having ceased to exist (dissolved) or where new household can be identified to have formed. There thus exists a challenge in reshaping and transforming the data to enable one to track identified households longitudinally.

Issue 2: Missing information on TSMs

TSMs are only observed in the NIDS data when they are co-resident with a CSM. These individuals therefore transition in and out of the survey sample across periods, and thus certain households (i.e. those households that comprise entirely of TSMs in specific periods) also transition in and out of the panel sample. This presents a large-scale missing information problem in the context of a panel of households, as one is unable to identify/ classify such households that are unobserved and exit/enter the panel sample.

Issue 3: Bias due to chosen household definition

The restructuring of the panel must be based on certain household definitions, yet such definitions generally involve the allocation of arbitrary rules, and thus may introduce bias in the restructured panel.

Issue 4: Data Quality

Overall, the quality of the NIDS data is quite high, especially when compared to other national household panel surveys of a similar nature. NIDS has a markedly low level of non-response at both the household- and individual-level, yet when viewed through the lens of household dynamics and migration, a potential association between location moves and attrition (or between household formation and attrition) presents a concern for data quality.

The methodological solution

In order to render the data suitable for investigating the dynamics and change of households over time, the data must be restructured - from a complex panel survey of individuals into a panel of households. The authors developed and implemented an innovative methodological approach that involved the identification of overlapping members and dwelling-units within household units across periods. They focused on creating household classification rules (based on selected household definitions), which allowed them to identify the same household unit across two observational periods - and thus, when households formed or ceased to exist (dissolved).

A graphical example of how these household classifications work in practice is provided in Figure 2. Where a household continues to exist, an arrow connects the unit across periods. Notably, a continuing unit in period t+1 could be categorised as either a "surviving" or "dissolving" unit with respect period t+2. This is evident by the fact that each household in period t+1 has both a C/N as well as an S/D classification.

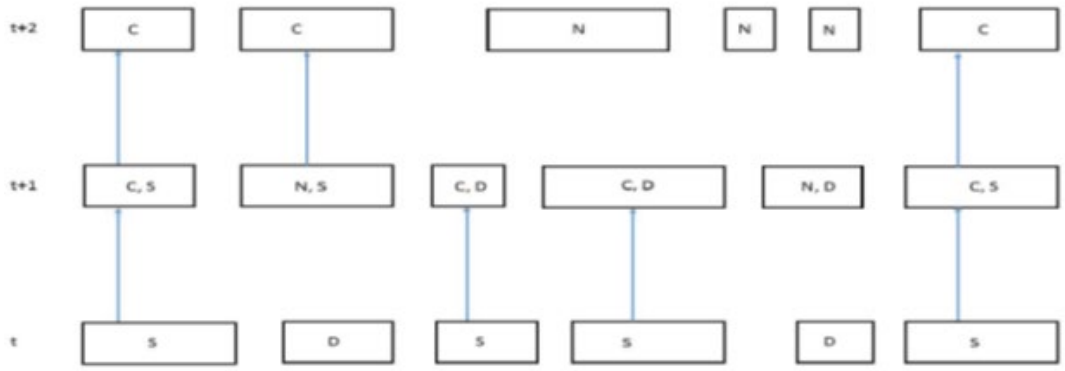

Figure 2. The new household classifications work in practice
Note: Surviving households (S) at time t (or time t + 1) are observed in the following period where they will be classified as continuing (C). They are connected by arrows. Households that don't survive, or cease to exist, are labelled as dissolving (D). New households at time t + 1 or t + 2 are labelled "N".

Using the approach outlined above, the researchers generated a set of unique household-level panel identifiers that enables any data-user to track identified household units from one period to the next, as in Figure 3. As part of this novel method, the team also generated a new set of household panel weights to account for the particular patterns of household-level attrition observed in the new/restructured panel. In order to account for bias resulting from their chosen panel household definition (for example, bias in household formation and dissolution rates) the authors created a comparison panel using the traditional "headship-based" definition. A comparison of these data revealed remarkably similar findings across the two panels.

Panel Analysis: Transitions for Surviving Households, South Africa, 2008-2010
(Number (and proportion) of Households)

		Wave 2		
		No Access	Access	Total
Wave 1	No Access	1 608 484 (0.694)	705 834 (0.306)	2 313 318 (1.000)
	Access	836 298 (0.087)	8 755 202 (0.913)	9 591 500 (1.000)

Figure 3. Panel Transitions in NIDS (2008-2010)

Research outcomes

Measurements that can be performed	Example	Multi-period cross-sectional survey data	NIDS individual-level panel survey data	NIDS household-level survey panel data
Aggregate changes in individual-level outcomes	Aggregate increase in take-home wages between two periods	✓	✓	×
Aggregate changes in household-level outcomes	Aggregate increase in household-level electricity access rate between two periods	✓	✓	✓
Longitudinal changes in individual-level outcomes	Number of individuals that experience an increase in take-home wages between two periods	×	✓	×
Longitudinal changes in household-level outcomes	No. of households that gain access to electricity between two periods	×	×	✓
Aggregate changes in household composition	Aggregate increase in the size of household between two periods	✓	✓	✓
Longitudinal changes in household composition	Identifying the number and characteristics of household that that increase in size	×	×	✓
Aggregate changes to the household population	Net household formation rate across two periods; net growth in household population between two periods	✓	✓	✓
Longitudinal changes to the household population	Identifying the number, and characteristics of household that form, dissolve or continue to exist between two periods	×	×	✓

Table 1. Comparison of potential data applications by survey data type

An analysis of this restructured panel data revealed notable new insights on the dynamics of electricity access.

Using standard transition matrices, the authors showed that, while new household connections are added in every period, the dynamics of disconnections (i.e. household units that lose a connection) can also contribute disproportionately to short-term declines/stagnations in access. They also calculated household formation and dissolution rates using this new panel in order to develop a deeper understanding on the complexities underlying the growth of the household population – including new insights on the characteristics of new and dissolving households.

A summary of the significance of the restructured NIDS household-level panel data is shown in Table 1 (previous page), comparing what this new data structure is able to measure relative to the original NIDS individual-level panel, or other multi-period cross-sectional survey data.

Conclusion

This innovative data restructuring approach of individual-level panel survey data (from a household panel survey) to provide a view of the longitudinal dynamics of household-level outcomes shows significant potential. Given the number of household panel studies across the globe, the application of the novel methodological approach presented offers the potential to draw new insights on the dynamics of both service delivery outcomes and household structures/composition. Data collectors and their funders can benefit from opening up rich data for insight in new domains, taking full advantage of the unexploited potential of standard household survey data.

Authors:

Tom Harris, independent consultant, Martin Wittenberg, Cape Town University, Mark Collinson, University of the Witwatersrand, and Louise Corti, UK Data Service

See: NIDS data

Full paper: Martin Wittenberg, Tom Harris and Mark Collinson (2017) Aiming for a Moving Target: The Dynamics of Household Electricity Connections in a Developing Context, World Development