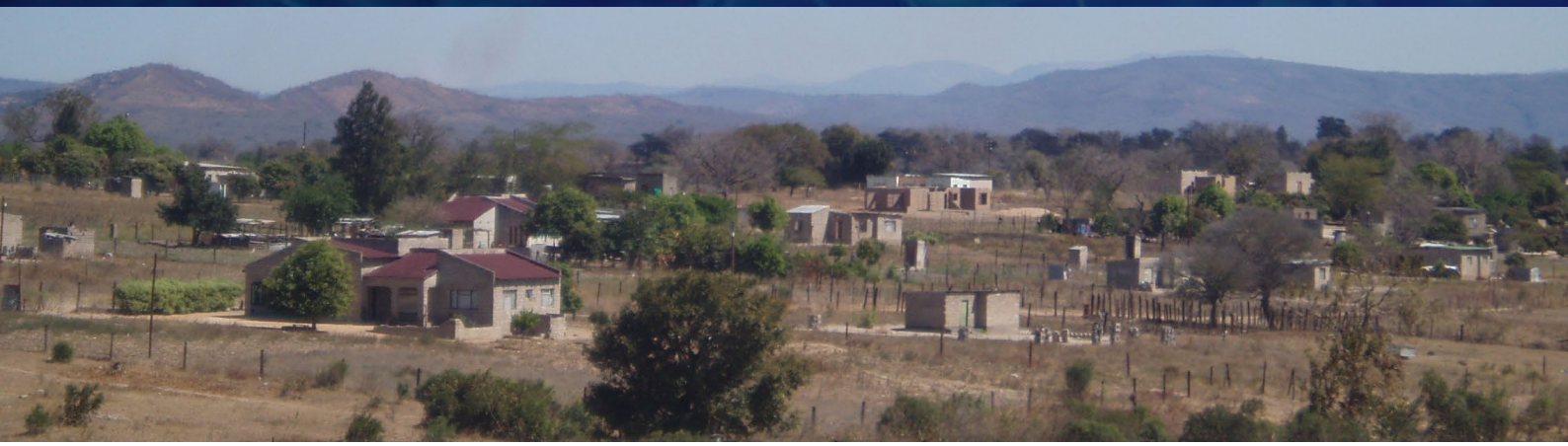
Reshaping a demographic longitudinal population study for socio-economic research



The challenge

Across the world, substantial investments have been made in population and health surveillance sites. These sites record rich longitudinal data on demographics, household composition, health outcomes, migration, employment, incomes and educational attainment. Such programmes are conceptualised with the aim of providing a more accurate and thorough understanding of development processes in various contexts. However, in practice, the sometimes highly complex structure of the recorded data – inherent to the data collections process at these sites – prevent these data from being used and studied widely by social scientists. Such projects can thus fail to have the impact they might, and not provide insights for policy and planning. This case study presents a methodology for restructuring longitudinal population study data to enable a wider spectrum of research questions.

> The electrification of South Africa's rural areas has significantly increased access since the end of apartheid, and contemporary research aims to address what, if anything, has been the impact of this.

Research example

An example of how such complex data can be restructured to provide insights into development processes is provided in the recent work of Mark Collinson, Martin Wittenberg and Tom Harris. Their research focused on electricity provision in rural South Africa.

The electrification of South Africa's rural areas has significantly increased access since the end of apartheid, and contemporary research aims to address what, if anything, has been the impact of this. When looking at available and suitable data sources for studying dynamics of energy provision, social researchers have traditionally depended on census and surveys, but longitudinal data from other disciplines, such as in the health and epidemiology domain, offer the potential for new insights.

The Wits/MRC Rural Public Health and Health Transitions Research Unit has been monitoring demographic and health changes in the Agincourt area since 1992. Since 2000, the team has also monitored changes in infrastructure and household assets as well as labour market outcomes. They have thus amassed a valuable set of longitudinal data, dating back to the 1990s, which has the potential to become a shareable resource to look at the socio-economic impact of the electricity roll-outs on household outcomes in this area.

However, as with most surveillance site data, the structure of the data gathered is quite complex and does not exist in a form that social scientists are used to dealing with. Wittenberg and Collinson sought to address this challenge in an initial research paper (2014) and reshaped the original site relational database into a data form more appropriate for longitudinal research. Their methodological approach enabled them to track longitudinal changes in various household-level outcomes, such as household size.

Using this new approach, Wittenberg, Collinson and Harris were able to investigate the dynamics of rural households' electricity access, countering the deficits they saw in the existing literature in this field. The researchers argue that most research has focused on investigating changes in an individual's access to electricity, and does not take into account the importance of the household unit and changes in household access. Thus existing studies tend to portray the process of electricity roll-out as a simple, monotonic progression, while it is often more a far more complex picture. Furthermore, recent literature suggests a strong association between electricity access and poverty, and conventional analyses ignore the complexities of access transitions among the poor – that is, the typically used aggregate data sources do not offer rich enough information on changes over time. The inherent value of the reshaped surveillance site data is that it can fill these gaps and provide deeper insights on these complex transitions.

Data and data issues

Agincourt HDSS data overview

The Agincourt Health and Demographic Surveillance System (Agincourt HDSS) monitors key demographic events and socio-economic variables in the Agincourt sub-district in north-eastern Mpumalanga Province, South Africa. A baseline census was conducted in 1992 with annual census rounds being conducted since 1999. Key variables measured routinely by the HDSS include: births, deaths, in- and out-migrations, household relationships, resident status, refugee status, education, antenatal, and delivery health-seeking practices. Additional modules have also been added over time; but these modules are not collected every year. For example, a household asset module, which includes information on household access to services such as electricity, was added in 2000 and has been run every second year since. 'Temporary migrants' are defined as non-resident members who retain significant contact and links with the rural home and 'share a common pot', and are included on the household grid. The data is stored in a relational database constructed in Microsoft SQL Server.

Issue 1: data shape and structure

Despite the systematic and comprehensive data collection process, these data are not collated in a form that social scientists are familiar with when undertaking longitudinal analyses. Rather than a simple 'wave-based' architecture common to most social surveys, the data take the form of a relational database. Information for different socio-economic categories appeared in various 'data tables', with unique identifiers connecting information for the same units across tables. The core data appear in three primary formats: object tables, event tables or episode tables. As the names suggest:

- Object tables list identify objects (e.g. individuals) along with key pieces of collected information associated with the given object
- Event tables list events that have occurred (within particular categories) and provide information related to the event, including the date of the event
- Episode tables organise the data according to unique identifiers and the reported start- and end-dates of each episode (e.g. recoding an individual's membership within a particular household)

Relationships between these tables are shown in Figures 1 and 2.

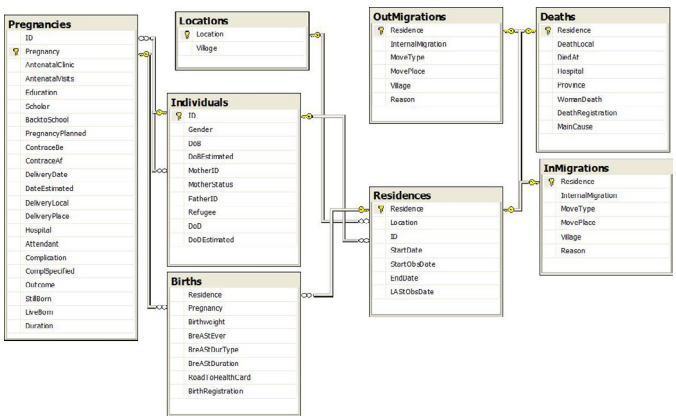


Figure 1. Diagram of links between data tables: Individuals, residences and events

The data in any additional modules (such as the asset status module) are stored as status observations. These status observation modules take a familiar wave-based form, with observations organized according to unique identifiers and identified observational years. However, unlike the core data, additional modules are not collected in every year. Another limitation of the dataset is that the Agincourt HDSS does not have the capacity to retain a respondent's identity number if he or she moved within the study site.

Issue 2: confidentiality and ethics

Numerous confidentiality, legal and ethical issues must be considered when accessing and publishing data from HDSS such as the Agincourt project.

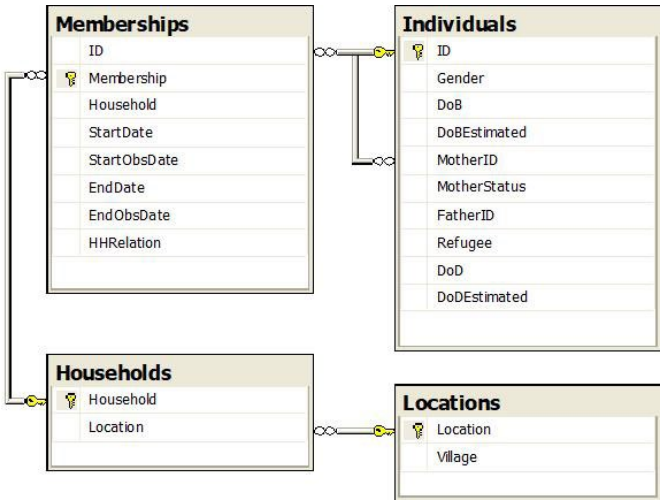


Figure 2. Diagram of links between data tables: Individuals membership in a household

The methodological solution

In order to investigate dynamics and change over time, and overcome some of the limitations of the original HDSS data, Wittenberg and Collinson developed a new methodological approach (2014). They based their approach on traditional panel study methodology, but considered the designation of households as the appropriate units of analysis, rather than individuals. Using a novel household typology, they defined each household unit according to when it formed, and whether it continued to exist or not. Their longitudinal analysis of household outcomes thus rested on their ability to identify the same household in each identified period of the study.

After organizing the HDSS data into one-year intervals, and reshaping and preparing into a format more appropriate for longitudinal research, they applied their new methodological approach to the data. This allowed them to set up a panel of household units for the period of interest – with a set of unique household panel identifiers that could be matched to a unique household in each identified observational period. Longitudinal changes in household electricity access can now be explored (using standard transition matrices, as well as other analytical approaches).

Figure 3 depicts an example of how electricity access could be investigated using the authors' household typology. The example is set across three observational periods: T1-T3. The arrows indicate where households continue to exist between two periods (e.g. H1 continues to exist between T1 and T2). Households marked with an ' ' cease to exist after a given observation period (e.g. H2 in T1), and households marked with an 'N' are new in a given observational period (e.g. H4 in T2).

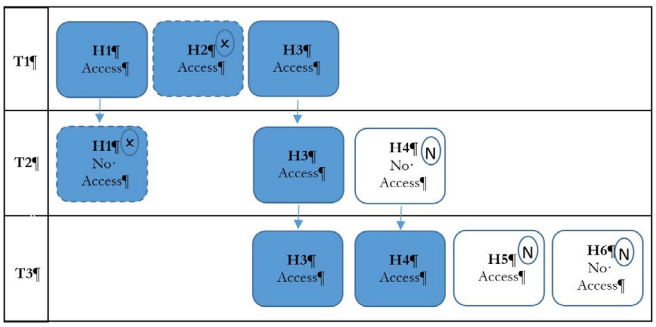


Figure 3. Example of household electricity access by Household Type

A household that initially has access to electricity can either retain access (e.g. H3 between T1 and T2) or lose access (e.g. H1 between T1 and T2). Short-term declines in aggregate access can result when households with electricity access cease to exist (e.g. H2 between T1 and T2) or when new households form in locations that lack access (e.g. H4 in T2).

Yet, improvements in aggregate access rates can also result when households without access dissolve (e.g. H1 between T2 and T3) or when new households form in locations with access to electricity (e.g. H5 in T3). Also, while the number of connected households in T1 and T3 is the same, the simultaneous expansion of the household population contributed to the decline in aggregate access rates declined (from 100% to 75%).

Research outcomes

Using this new method and restructured HDSS data it has been possible to decompose aggregate changes in electricity access within Agincourt, and explore the different longitudinal access transitions that occur across different household types. The research team explored how three factors contributed to changes in aggregate electricity access in the HDSS data:

1. gains and losses in access within continuing households
2. access rate differential between dissolving and new households
3. growth of the household's population (i.e. the extent to which the base of existing connections is diluted due to an increase in the size of the household population)

The research provided new insights on the dynamics of electricity access in rural areas, demonstrating, in particular, the importance of different types of access transitions across different household types. An example of change for the period 2005-2011 is presented in Table 1.

Between 2005 and 2007, net household formation (the 'dilution effect') was the primary contributor to the aggregate decline in electricity access in the rural area of Agincourt, while changes that occurred within continuing households contributed only a third of the overall decline (the 'within effect'). Notably, differences in access rates between households that ceased to exist and new households that formed (the 'replacement effect) worked in the opposite direction – slightly moderating the magnitude of the aggregate decline in access. However, the 'within effect' was the key contributing factor to the aggregate increase in access between 2007 and 2009. In the final period under analysis, 2009-2011, despite large positive 'within effects' and 'dilution effects', a substantial negative 'replacement effect' led to only a small net increase in aggregate electricity access rates.

In summary, the dynamics of these transitions are evidently neither simple nor uniform: with each transitional component working in both positive and negative directions, and varying in importance depending on the period under investigation. Movements are observed both up and down the 'energy ladder' with some household gaining access and other households losing access.

Period	Overall change	Change due to transitions within continuing households	Change due to differential between dissolving and new households	Change due to growth of the household population
2005-07	-0.011	-0.0036	0.0022	-0.0096
2007-09	0.051	0.0633	0.0019	-0.0146
2009-11	0.005	0.0086	-0.0128	0.0094

Table 1. Decomposing changes in electricity access, Agincourt, 2005-2011

Conclusion

Restructuring the Agincourt HDSS data has enabled a richer picture to be gained by investigating the short-term dynamics of electricity access using large-scale longitudinal data.

From a methodological point of view, the authors conclude that aggregate statistics can conceal a considerable degree of the complexity and volatility inherent in the development of electricity access. Viewing access as a time-variant process, rather than assuming linear roll out of services, needs to be appreciated and studied. Further studies into service delivery in low and middle income countries (LMICs) should consider using the longitudinal techniques applied here, taking into account household dynamics. In particular, these techniques show the potential for HDSS-type repurposed data to be used to analyse service delivery outcomes in other contexts.

Of note here is the novel method used to repurpose data from a typical HDSS, undertaken using an economist-centric perspective on data utility by creating a traditional panel dataset. This approach offers an immensely promising avenue for opening up access to other complex HDSS data by considering how data can be restructured and repurposed for use across disciplines. The sometimes highly complex structure in HDSS–type data collections can be re-worked to create new datasets that can be better understood by those outside the population, migration and epidemiology research domains, such as social scientists. In essence, this brings great hope for improving access to complex longitudinal data from multi-million pound population surveillance and intervention /evaluation investments conducted in LMICs. The approach would benefit from being widely showcased and exploited as an inspiration for data-hungry researchers and, of course, for maximising impact opportunities for funders.

Authors:

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

See:

Wittenberg, M., Collinson, M., (2014). Household formation and household size in post-apartheid South Africa: Evidence from the Agincourt sub-district 1992-2003. A DataFirst Technical Paper 27. Cape Town: DataFirst, University of Cape Town

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

Wittenberg, M., Collinson, M., Harris, T., (in press). Decomposing changes in household measures: Household size and services in South Africa. Demographic Research.