The Agincourt HDSS Energy Panel Dataset User Guide

Version 1, August 2023

Martin Wittenberg, Mark Collinson, Taryn Dinkelman, Chodziwadziwa W. Kabudula, Takwanisa Machemedze, Wayne Twine, Kathleen Kahn and Stephen Tollman

1 Background

The MRC/Wits Agincourt Unit (www.agincourt.co.za) has been collecting information on health, social and demographic outcomes in the Bushbuckridge area in the north-east of South Africa since 1992. The study site of the Health and Socio-Demographic Surveillance System (HDSS) comprising initially twenty villages was selected because it fell into an area previously designated as a "homeland" (Gazankulu) and thus exemplified many of the health and developmental challenges facing rural South Africa. One distinguishing feature of the area was that it had seen an influx of refugees from Mozambique during that country's civil war in the 1980s, many of which were initially settled in refugee villages near the eastern border of the site. An original aim of the programme was to understand primary health care developments in the context of a health centre, based at Agincourt, and satellite primary health care clinics.

The data collection process involves annual (since 1999) census rounds during which the main demographic and socio-economic variables are captured, in particular births, deaths, in- and out-migrations. Since 2000 more detailed supplementary modules have been fielded, dealing *inter alia* with household assets, education, grants, and employment information. After 2007 the study site was expanded to include another eleven villages. More detailed information on the Agincourt HDSS is provided by Tollman (1999), Tollman *et al*. (1999) and Kahn *et al* (2012).

The location of the Agincourt HDSS in a rural historically deprived area, with a long history of data collection, makes it an ideal place to investigate development processes. In 2015 DataFirst, the UK Data Archive and the MRC/Wits Agincourt Unit received an ESRC/NRF International Centre Partnership Grant[1] to use the data from the Agincourt surveillance system to investigate the quality of rural electrification data and to research the impact of electrification on local communities.

In terms of this grant, a panel dataset was constructed from the Agincourt HDSS to facilitate this research. In line with the ESRC/NRF policies of open access a version of the household level panel is being made available via the DataFirst open data portal (www.datafirst.uct.ac.za). The data were extracted in 2016, so the latest information in this release is from 2015 although the latest asset information is from the 2014 census round. Earlier versions of the panel dataset have been used in several published studies (Wittenberg and Collinson 2007, Harris *et al* 2017, Wittenberg *et al* 2017, Wittenberg & Collinson 2020). They provide further background and examples of how the data may be used.

[1] Ref: ICPC150423117553

This document outlines the core issues that users should be aware of when working with the data. Section 6 contains information on the variables that are available in the public release version of the data, as well as additional information that can be accessed through the DataFirst Secure Research Data Centre.

2 Accessing and referencing the data

There are two versions of the dataset. The public release version does not include village identifiers or individual level information. It is available via the DataFirst open data portal (www.datafirst.uct.ac.za). The full dataset is available via DataFirst's Secure Research Data Centre.

2.1 Citing the data

Users should cite the dataset as follows:

> Agincourt Health and Demographic Surveillance System. 2023. Agincourt Household Energy Panel [dataset]. Version 1.0: 1992-2015. Acornhoek: MRC/Wits Agincourt Unit [data producer]. Cape Town: DataFirst [panel producer]. Cape Town: DataFirst [distributor].

2.2 Citing this document

Readers wishing to cite this document should use the following reference:

> Wittenberg, M., Collinson, M.A., Dinkelman, T., Kabudula, C.W., Machemedze, T., Twine, W., Kahn, K. and Tollman, S., 2023. *Agincourt Household Energy Panel User Guide*. Version 1. Cape Town: DataFirst.

2.3 Accessing the full version

Applications to use the full version of the Agincourt Household Energy Panel in DataFirst's Secure Research Data Centre need to be made via the application form emailed to support@datafirst.org.

3 The structure of the data

To understand the nature of the panel data it is important to comprehend how the information is collected and organised by the Agincourt HDSS and how it was transformed for the purposes of this research.

3.1 Levels of the data

Data collection and organisation happens in terms of the following hierarchy:

Village

As noted above, the core geographic entity through which fieldwork is organised is the village. At present there are thirty-one villages in the study site. In the public release data, the village identifiers have been removed to protect confidentiality, but anonymised village identifiers are available in the secure version of the data. To present some context for the data a **village type** identifier is included in the data. This is to differentiate between the original study site and villages added when the study site expanded; and within the original study site to distinguish refugee settlements from the other villages. Correspondingly the village type variable recognises three types of settlements:

1. "South African" villages from the original study site
2. "Refugee" villages from the original study site
3. Villages added when the site expanded after 2007

Dwelling

The fieldwork teams maintain maps of dwellings or compounds in each of the villages. Households (discussed below) are identified by their connection to a physical location. The anonymised dwelling identifier is in the secure version of the data but has been removed from the public release version of the data.

Household

Households are defined in a census round as the set of individuals at a common location (dwelling or compound) who typically "eat from a common pot". Unlike in a typical household survey, membership is not restricted to individuals who have resided at that address for more than four days of the last week. Instead labour migrants can be kept on the household roster, as long as the key household informants agree that they are still regarded as part of the household. In the census updates the household is asked how many months out of the last twelve the individual has spent at home. Based on these responses, individuals are categorised as either "permanent" members or "temporary migrants". In some census rounds individuals are encountered who are classified as always having been part of the household (despite not being on the roster in the previous year) and so they are inserted into the household retrospectively. Their residency status for previous years will, however, be unknown.

Besides being linked to a physical location, Agincourt households are also linked to a household head. The relationships are captured in more detail than is usual in surveys. The entire chain of connections to the household head is captured. For instance, a grandson will be either the son's son (SS) or daughter's son (DS). A niece could be the brother's daughter (BD), the sister's daughter (ZD) or even the wife's brother's daughter (WBD). Because of the detailed structure captured in the dataset, this variable is not in the public release dataset.

Defining households in a longitudinal study is complicated (Wittenberg, Collinson and Harris 2017). In the Agincourt HDSS a household is defined by overlap of membership at the same physical location. This means that a household will be regarded as the same, even if the household head dies or migrates out, provided that some of the other household members continue living in the same dwelling. It means, however, that if all members of the household migrate out at the same time, even if it is to a new location in the study site, this is regarded as the dissolution of the existing household and the formation of a new one.

Individuals

At each census round information on basic demographic events (births, deaths, in- and out-migrations, union formation and dissolution) is collected. As far as possible the precise date of these events is fixed and, in the case of migrations, the reasons associated with the move and the place of origin/destination. Periodically additional information is collected. For instance, education, employment status and grants information has been collected through additional modules. In the public release version of the dataset, these additional data are not available.

In the early period of the Agincourt HDSS individuals were not tracked if they moved from one part of the study site to another. Nowadays, however, they are given persistent identifiers which allows them to be tracked from one household to another. Much of the early history has been reconstructed, although some caution is advisable when analysing data on individuals who are identified as having changed locations without leaving the site permanently.

3.2 Episodes and data relationships

Much of the core information in the Agincourt HDSS is stored as **episodes**, notably residences at a location and membership in a household. For instance, an in-migration may begin a membership episode. This could be ended by a change of headship, since at that stage all the relationships within the household need to be redefined. The information itself is stored in a relational database, with links to key identifiers, e.g. households or individuals. Additional information on those entities (e.g. labour information on individuals, as collected in the labour module) is stored in separate tables with links to the person/household. The information in those tables is itself organised by **observation date**, i.e. the time at which that information was recorded, typically in a census round.

Storing information in this form has various advantages – it is stored more compactly, which is helpful given the size of the database. Secondly it ensures that the data is stored at the appropriate level, e.g. labour information can never be attributed to a household, since the labour module is organised around links to individuals.

Nevertheless, this episodic and relational structure of the data makes it relatively difficult to analyse changes or transitions in the manner that social scientists are used to. Consequently, we transformed the structure of the HDSS information by "flattening" the relational information into household and individual level tables and changing the episodic information into a more conventional panel structure.

3.3 Creating the panel structure

In order to create the panel, we check which individuals and households were present in the study site on 30 November of each year. This is the computer (thought) experiment equivalent of sending out field workers on the 30th November of each year and fully enumerating the individuals and households encountered.

The process is shown schematically in Figure 1. This Figure presents information on five individuals in two households. Individuals 1 and 2 both migrate into Agincourt prior to 30th November in year t. Individual 2 moves out of Agincourt in August of the following year, while individual 1 is still in the site on the 30th November in year t+1. Person 3 is a child born after the 30th November in year t and moves out with individual 2 in August. The other two individuals (4 and 5) both move into the site and out again between the two anchor dates.

In terms of our panel structure only household A and individuals 1 and 2 are recorded, i.e. we will miss individuals and households that move through the site within a year, unless that short episode happens to extend across a 30th November.

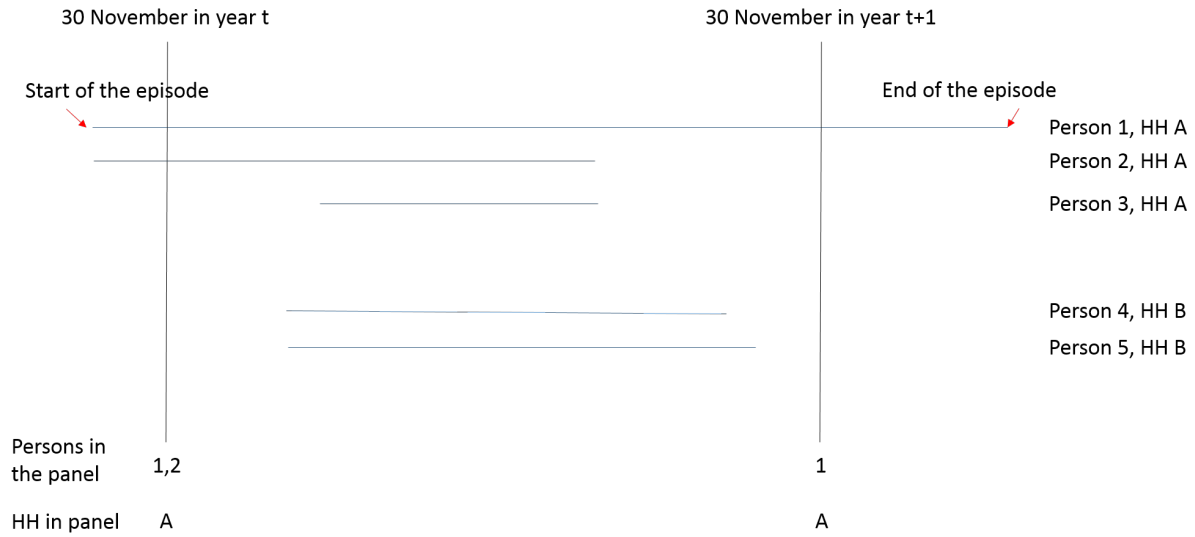

Figure 1 Creating the panel structure from episode data

3.4 Anomalies with the dates

One of the problems with this structure is that in some cases it can lead to a misalignment between the measurement process and the way in which we record this in the data. The measurement process is the "census round" which is typically conducted earlier in the year. Indeed, the reason for picking 30th November as anchor time is to ensure that the census round has been mainly completed. Some of the information (notably the asset schedule, but also special schedules like employment status) is not episodic but linked to the observation date, i.e. we know what the household asset holdings were at the time that the household was interviewed. How do we link this information into the panel structure?

The logical way to do this is to attach this information to the panel in the year of the measurement, i.e., if the measurement was done in September of year t, then it would appear in the panel as the holdings in year t. This produces problems if the household has migrated out of the site by 30 November. To be consistent with the approach above we have discarded such observations since the information belongs to a household that no longer exists on 30 November. A different mismatch occurs if the household enters the site only after 30 November but is still enumerated in that census round. That information is also lost. This means that there can be mismatches between our dataset and the sequence of asset schedules (organised by census round) as shown in Table 1.

Table 1 Number of completed household asset schedules by census round

Census round	Asset schedule available: HH left by 30 Nov	other mismatches	panel + asset info	Total with asset info	no asset info	panel households
2001	92	9	10 961	11 062	861	11 822
2003	88	3	11 489	11 580	555	12 044
2005	60	5	11 425	11 490	965	12 390
2007	70	858	12 984	13 912	1 068	14 052
2009	125	18	14 942	15 085	942	15 884
2011	43	1	13 967	14 011	2 716	16 683
2013	27	0	16 136	16 163	4 640	20 776
2014	43	1	17 359	17 403	3 747	21 106
Total	548	895	109 263	110 706	15 494	124 757

The first column of Table 1 shows the number of households (per census round) that completed asset schedules but where the household had left by 30 November. The second column shows other mismatches, mainly due to households which only entered the study site after 30 November but where the asset schedule was still completed in that calendar year. Almost all of the 858 cases shown for 2007 were from three villages which were part of the process of extending the Agincourt study site and which were interviewed in December 2007. The third column (shaded in grey) represents the households with asset information that do appear in our panel. This is the information which will be used when the asset data in the Agincourt Energy Panel is analysed.

The fourth column is the sum of the previous three, i.e. it represents the total information available on assets in the Agincourt database. Out of the 110 000 records we have lost information on about 1400 (1.3%) due to the way in which the timing of measurement interacts with the construction of the panel. The fifth column has a count of panel households which have no asset schedule in that particular year while the last column is the total number of panel households for the years indicated, i.e. the sum of columns three and five. There is obviously a missing value problem on top of the misalignment one.

The problem of aligning the measurement dates with the way in which we have constructed the panel also occurs with some individual level information. Specialist modules like employment status and education also collect information at a point in time, rather than in episodes. Again, we bring the information into the panel matched on observation year and once more this means that we lose measurements for individuals who leave the site between the observation date and 30th November.

3.5 Missing information

The intermittent nature of the specialist modules also creates gaps in the data. For instance, there is no asset information on even numbered years between 2001 and 2013. The employment module was only fielded every fourth year, also leaving many gaps. In the case of the education information the pattern of missingness is more complicated as shown in Table 2.

Table 2 Information recorded on individuals' education levels by year

year	Education information Present	Missing	Total	year	Education information Present	Missing	Total
1992	0	59 482	59 482	2004	5 572	66 128	71 700
1993	0	65 127	65 127	2005	6 080	66 366	72 446
1994	3 526	62 530	66 056	2006	68 037	5 285	73 322
1995	10 537	56 542	67 079	2007	12 678	67 977	80 655
1996	7	67 391	67 398	2008	6 454	80 608	87 062
1997	57 438	10 935	68 373	2009	82 745	6 063	88 808
1998	44	68 546	68 590	2010	7 189	82 886	90 075
1999	11 239	58 902	70 141	2011	8 941	82 833	91 774
2000	5 882	64 691	70 573	2012	87 857	5 336	93 193
2001	6 401	64 376	70 777	2013	103 785	9 923	113 708
2002	67 500	3 432	70 932	2014	110 697	4 499	115 196
2003	6 297	64 899	71 196	2015	0	110 711	110 711

The years in which major updates happened are shaded. The information for the years in between was derived from interviews with new arrivals to whom the education module was administered. Our dataset does not include information before 1997 (except for new arrivals) and also misses the education module from the last update round in 2015.

Tables 1 and 2 also indicate, however, that even in years in which updates are conducted there is missing information. In the case of the asset information around 7% of households in an update year seem to be missed, with a much higher fraction in the last three rounds (2011, 2013 and 2014). We will explore the pattern of missingness in more detail below and have constructed a set of weights to be used with the asset schedule.

3.6 Other data issues

In the process of "flattening" the HDSS data several glitches were encountered. In several cases the start or end dates of episodes were not aligned with the start or end dates of related episodes:

3.6.1 Death dates and change of headship dates

The misalignment of dates could produce two types of problems: in some cases, the change of headship supposedly happened before the death of the current head, leading to two people being identified as head simultaneously. In other cases, the change of headship supposedly happened only after a while, meaning that the household had no identifiable head for the intervening period. This was an issue for our panel dataset only if the period of dual headship or zero headship was active on a 30 November. In most cases it was straightforward to align the dates and resolve the problem. Where the problems could not be resolved in this way, the oldest household member was assigned to headship.

3.6.2 Birth dates and in-migration

There were a few cases where infants were apparently born in the study site before their parents migrated in. These were obviously cases where the "start date" for the baby's residence episode should have been recorded as the date of in-migration instead of their birth date.

3.6.3 Households and temporary migrants

As noted above, continuity of households over time is tied to having members that overlap at the same physical location. In several cases the overlapping members happen to be only temporary migrants. This means that there are households which on 30 November have no permanent members, but only temporary migrants. For purposes of comparing our dataset to typical cross-sectional instruments (like a census or a survey) these households would be "out of scope." They can be excluded by dropping households with a household size (of permanent members) of zero. As shown in the next section, especially Table 4, the number of such households has increased in the study site over time.

4 Weights for the asset schedule

As noted in relation to Table 1, from 2011 onwards there is an appreciable gap between the number of households for whom there is an asset schedule and the number of households in the site. To deal with this situation we have constructed a set of weights that are designed to correct for the undercount. In this section we discuss the patterns of missingness, the process of constructing the weights and provide some diagnostics on the weights.

4.1 The pattern of missingness

In Table 3 we show the proportion of households that are missing an asset schedule by our village typology. The increase in "missingness" is very evident for households in "South African villages" in the original study site but also among households in the expanded part of the site. In the case of the "refugee villages" there seems to have been a problem all along. This is a reason why it would be a mistake to simply ignore the missing information, since it is correlated with a variable that is itself associated with various development outcomes.

Table 3 Proportion of households with missing asset schedule

	Original study site		
Year	SA villages	Refugee villages	Additional villages
2001	0.065	0.141	
2003	0.038	0.116	
2005	0.068	0.161	
2007	0.068	0.123	0.104
2009	0.049	0.103	0.083
2011	0.157	0.163	0.191
2013	0.215	0.200	0.248
2014	0.175	0.106	0.203

Another variable that seems to matter is household size. Bigger households are more likely to complete asset schedules (as we will show below). One particularly problematic category is households without any permanent members on 30 November. It is possible for residence episodes of the permanent members not to overlap that date, so that in fact nobody is "home" although the household still exists in the site. Every year there are a few hundred such cases in our data, although the numbers have increased towards the end of the study period.

In Table 4 we show the proportion of missing asset schedules by household type, where we distinguish between households which had at least one permanent member and households which,

on 30 November, did not have any. In that table we also show in the final column the count of households which had no permanent members. It is evident that this rose rapidly at precisely the time when the fraction of these households that did not complete the asset schedule also rocketed.

Table 4 Households with zero permanent residents and missing information

Year	Household type (number of permanents)		Count of households with no permanents
	At least one	None	
2001	0.069	0.148	549
2003	0.044	0.088	633
2005	0.073	0.144	803
2007	0.072	0.143	767
2009	0.055	0.124	969
2011	0.120	0.742	1139
2013	0.153	0.904	1949
2014	0.089	0.907	2292

4.2 Generating the weights

The weights are created using a typical inverse probability weighting approach, although we modify the procedure. We begin by estimating a probability model (probit) for each year that we have an asset schedule where we estimate the probability of the household returning an asset schedule using the following explanatory variables:

- A set of dummy variables for each village in the study site[2]
- A dummy for female headed households
- A dummy for refugee households
- The log of the full household size (i.e. household size including temporary migrants)
- An indicator variable for a household with zero permanent members.

A summary of the probit results is given in Table 5. The magnitudes of the coefficients are difficult to interpret directly, with the rule of thumb suggesting that 0.4 times the coefficient being approximately the change in probability. By that metric the "log household size" coefficients are very big and the effect of having no permanent household members from 2011 onwards is so enormous that it would effectively reduce the probability to almost zero. Interestingly, the "Refugee household" variable seems to matter only up to 2009.

[2] There were a handful of observations where the village was unknown or where there were too few observations in the village to allow us to incorporate them into the model. There were altogether 302 such household-year combinations among which only 7 asset schedules were returned. No attempts were made to reweight these observations, i.e. the households that returned the schedules got the default weight of one and the other 295 were effectively dropped from the population.

Table 5 The probability of a household returning an asset schedule: probit model coefficients

Variable	2001	2003	2005	2007	2009	2011	2013	2014
FemaleHead	-0.067	-0.093*	-0.046	-0.051	-0.107**	0.010	0.066**	0.043
RefugeeHH	-0.398***	-0.163**	-0.241***	-0.094*	-0.137**	0.022	0.047	0.055
Logsize	0.460***	0.460***	0.408***	0.316***	0.429***	0.508***	0.570***	0.420***
zerohh	-0.061	0.032	-0.091	-0.141*	-0.159**	-1.538***	-2.000***	-2.477***
Village dummies	Y	Y	Y	Y	Y	Y	Y	Y

legend: * p<0.05; ** p<0.01; *** p<0.001

The inverses of the predicted probabilities are our first cut at creating weights, i.e.

$$w_{it} = \frac{1}{p_{it}}$$

where w_{it} is the weight of household *i* in year *t* and p_{it} is the probability of completing an asset schedule. The weights obtained in this way no longer sum to the overall target population, let alone reproducing the village populations. The reason for this is that there are a few probabilities that are too close to zero, producing enormous weights and hence unstable results. We could force the weights to add up to the known village population totals by rescaling them, but that reduces the median weight to below one. Conceptually it is unattractive to have any weights less than one, given that every responding household should count at least for itself. Consequently, we only adjust the "excess weight" (i.e. the portion of the weight above one) to ensure that the deficit is covered. Mathematically

$$w_{it}^{*} = 1 + (w_{it} - 1)\frac{T_{pop} - T_{measured}}{\sum w_{it} - T_{measured}}$$

Here w_{it} is the inverse probability weight (as given above). T_{pop} is the total population that we would like to reproduce, $T_{measured}$ is the total of households with asset schedules, which is less than T_{pop} due to the missing observations. The advantage of this algorithm is that it weights households that have a low probability of completing asset schedules more heavily, while ensuring that none of the weights drop below one. Table 6 shows that we do a good job of reproducing the population. The first two columns reproduce columns five and three from Table 1. Column three is the total population counts that we would like to reproduce. Note that this total includes households with zero permanent members.[3] The weighted count in the last column is the weighted sum over households with an asset schedule using the adjusted weights w_{it}^{*}. The reason why it does not reproduce the total population entirely is that we have built the total population up from village populations, but there were around 300 cases where a household-year observation could not be assigned to a village.

[3] Indeed, in many cases we do have asset schedules for such households.

Table 6 Comparing the weighted counts to the population

Year	No asset schedule	With Schedule	Total Population	Weighted count of schedules
2001	861	10961	11822	11820
2003	555	11489	12044	12042
2005	965	11425	12390	12385
2007	1068	12984	14052	14040
2009	942	14942	15884	15881
2011	2716	13967	16683	16677
2013	4640	16136	20776	20622
2014	3747	17359	21106	20939

4.3 Impacts of the weights

Since one of the main concerns of the Agincourt energy project has been to measure electrification in the Agincourt area, we show the impact of the weights on the use of electricity for lighting. This is a reliable proxy for the availability of electricity to the household. In Table 7 we show the fraction of households that have access to electricity (i.e. the mean of the indicator variable) as well as the total number of connections. Unsurprisingly the weights have a much bigger impact on the totals than on the means, although in the early period the extent of electrification is underestimated by around a percentage point. In the case of the totals the raw counts suggest that there was actually a net loss of connectivity in 2011 (the cell highlighted). This is purely an artefact of the poorer reporting in that year, as is evident when we look at the sequence of weighted counts. In short, we would advise use of the weights when working with the asset data, although in the calculations of means we would not expect to see a substantial impact.

Table 7 The impact of the weights on measures of electrification

	Mean		Total	
	unweighted	weighted	unweighted	weighted
2001	0.691	0.683	7 579	8 077
2003	0.765	0.757	8 791	9 115
2005	0.894	0.889	10 218	11 011
2007	0.883	0.880	11 471	12 357
2009	0.933	0.928	13 948	14 742
2011	0.939	0.927	13 111	15 455
2013	0.959	0.955	15 473	19 702
2014	0.976	0.975	16 938	20 405

5 Generated variables

Besides the weights for the asset schedule, there are some additional variables that have been added to the dataset, to provide additional information about access to energy or the types of households.

5.1 Night-time lights

The Agincourt household energy project extracted the average night-time brightness for villages in the study site between the years 1992 and 2012 (Machemedze *et al* 2017). The data comes from

satellite information recorded in the DMSP-OLS Nighttime Lights Time Series.[4] The brightness per pixel is recorded as a digital number (DN) ranging from 0 (absence of light) to 63 (saturation). Each pixel corresponds to about a square km on the ground. The brightness values within pixels corresponding to each village were averaged. The average values reported in the dataset have been rounded to the nearest half a unit. The maximum brightness measured in our dataset is 15, in keeping with the rural nature of the area. Nevertheless, there is a marked increase in brightness levels over time, which broadly parallels the electrification of the area.

5.2 Household structure

The information about the structure of the relationship to the head of the household, as recorded in the AHDSS, was used to characterise households in terms of who was a member on 30 November of each year. It is worth noting that this was done using all members, including the temporary migrants. As noted earlier, relationships are recorded as chains; a grandchild of the head of the household can be recorded either as DD (daughter's daughter), DS (daughter's son), SD (son's daughter) or SS (son's son). This information was used in two ways: to count the number of generations within a household and to classify the household according to a typology based on that discussed in Wittenberg and Collinson (2007).

5.2.1 Number of generations

This is based on the number of steps up (a father F or mother) or down (a son S or daughter D) in the chain of relationships. A cousin might be recorded as FBD (father's brother's daughter). This would involve a step up and a step down making the cousin of the same generation as the head, even if there could be a substantial age difference. The accuracy of this count obviously depends on the accuracy with which the relationships are recorded. There are several types of problems that were encountered in this exercise:

- In some cases where there were multiple heads of households the relationships could be incompatible (e.g. a mother that was younger than the head).
- In some cases, the codes seem to have been incorrectly entered. In one case (which would otherwise have yielded a six-generation household) the oldest member was coded as FM, i.e. father's mother or grandmother of the Head. The age gap between her and the head was 26 years so it was more plausible to assume that she was simply the mother of the head, who might have been referred to as the "Father" of the household.

After some cleaning of the data, the maximum number of generations observed in our dataset is five. This happens in the case of 253 household-year observations, to 74 distinct households. In one case, checked at random, it was a household that included a great-great-grandchild (DDDS). One might wonder whether there might have been an extra "D" recorded by accident. In this case the relationship is believable. The head is aged 79 and there is an 18-year-old great-grandchild (DDD) in the household and the great-great-grandchild seems to be her 3-year-old son. Given the age gap between the 18-year-old and the head, it is not strange to suppose that she really could be his great-grandchild, making her son a great-great-grandchild.

Nevertheless, there will be errors in the relationship chains. Consequently, some caution is in order when using the generation count.

[4] http://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html

5.2.2 Household typology

The household types are also based on the relationship codes, so the same warning applies. Table 8 records the household types that have been defined in the dataset. It should again be noted that this is based on the full membership including all temporary migrants. The distribution of household types would undoubtedly look a bit different if based on the *de facto* household membership on 30 November. Another point to note is that in many cases the relationship of an individual to the head of household is unclear (coded as an "X") or not recorded at all. Households in which even one relationship was unknown are not classified, unless we know that it includes at least one "unrelated" individual. There are altogether 78 107 household-year combinations with missing types, roughly a quarter of all household-year records. There are, however, no missing values in the generation count since individuals with missing relationships were ignored in the count. If all the other relationships were accurately captured, then the generation count would be a lower bound of the true value if all relationships were known.

Table 8 Household types defined in the Agincourt Household Energy Panel

Type	**Description**
Single person	There is only one person in the household according to the full household count
Couple	A head plus spouse (coded as H – husband or W – wife)
Nuclear	A head plus spouse plus children of the head
Single parent	A head plus children of the head
Three generation linear	Either: a) A head (with or without spouse), plus children of the head, plus parent(s) of the head or parent(s)-in-law of the head (but not both) b) A head (with or without spouse), plus children of the head, plus grandchildren of the head
Three generation skip	A head (with or without spouse), plus grandchildren of the head, but no children
Multi-generation linear	A household (not classified as one of the types above) with only the following types of relationships: head, spouse, child, parent, grandchild, grandparent, great-grandchild, great-great-grandchild
Siblings	A head with siblings
Nuclear + stepkids	A head plus spouse plus children of the head plus children of the spouse (identified as WD, WS, HD or HS)
Three generation + stepkids	Defined like the "three generation linear" household, except that in addition to children there could be stepchildren (defined as above), or in addition to grandchildren there could children's stepchildren or children of stepchildren.
Complex, related	Any other type of household in which all the members are related to the household head in some way. This includes households containing cousins, nephews and nieces and various in-laws
Complex, plus unrelated	Any household in which one of the members has been flagged as being unrelated to the head of household

This typology was influenced by the concerns in the early 2000s about the impact of the AIDS pandemic on household structure. For instance, there was the worry that "three generation skip" households would become more prevalent, as the middle generation was succumbing to the disease. In the article in which Agincourt household dynamics were analysed (Wittenberg and Collinson 2007), the typology was simplified, with the categories "nuclear + stepkids" amalgamated with "nuclear;" "three generation + stepkids" included with "three generation linear;" and "siblings"

incorporated in “complex, related.” The more complex typology is preserved here, although some of these categories are not all that common in the data. Furthermore, the stepchildren classification should probably be approached with caution.

6 The datasets and variables

The information is contained in three datasets:

- AgincourtHHPanel
  This is the public use panel of **households**. It contains information on household size and information from the asset schedules
- AgincourtIndivPanel
  This is a panel of **individuals** which can be linked to the household panel via the unique (anonymised) household identifiers. This dataset can only be accessed in the DataFirst Secure Research Data Centre. This dataset contains mostly time-varying information.
- AgincourtIndivData
  This is a dataset with time-invariant data on the individuals in the individual panel. It can be linked to the panel via the unique (anonymised) individual identifiers. It is available only in the DataFirst Secure Research Data Centre.

6.1 Variable naming convention

Variables in the AHDSS begin with an upper-case letter and are mixtures of upper- and lower-case letters (e.g. Household_Anon). Variables created or transformed for the Agincourt Energy Panel (e.g. by converting from string to numeric format) are all in lower case. Some of the original AHDSS variables have been edited, e.g. some of the relationship chains in the HHRelation variable. Those cases are noted in the metadata below. Except for those cases, variables that are capitalised initially can be assumed to be taken from the AHDSS tables as supplied to us.

6.2 AgincourtHHPanel

Variables marked with an asterisk are not in the open access version of the dataset. The public use file contains 326 247 records, i.e. household-year observations. The combination of Househhold_Anon and year uniquely identifies observations.
Comment: Many of the variables from the asset schedule have been converted from string to numeric formats. To avoid confusion they have the same names as in the original Agincourt HDSS system but all in lower case.

Household_Anon

This is the anonymised household identifier. There are 29 285 distinct households in the dataset.

year

Valid range: 1992 to 2015

Village_Anon*

This is the anonymised village identifier. It takes on 36 distinct values, not all of which are valid villages.

villagetype

This takes on the following values:
1. Original study site, South African villages
2. Original study site, refugee villages
3. Villages added to the study site after 2007

ExternalID*

This is the dwelling identifier within a village.

HHStart*

Household start date. Range 15 Jan 1982 to 28 Dec 2014.

HHEnd*

Household end date. Range 1 Jan 1994 to 31 Dec 2100. The date of 31 Dec 2100 indicates that the household was still present in the study site at the end of the period covered in this dataset.

HouseholdHead_Anon*

This is the anonymised individual identifier (ID_Anon) of the household head on 30 November of the year. It can vary within households. It can be used to link to the individual information of the head (in the AgincourtIndivPanel or AgincourtIndivData files). There are 34 927 distinct heads. There are also 1 049 Household-year records where this information is missing.

Note: This is not determined in the first instance by the "HouseholdHead_Anon" identifier in the Agincourt HDSS tables, since that identifier does not vary. In the HDSS tables only the most recent household head is kept. In the Agincourt Energy Panel, the identifier is time varying and points to the person identified as head on 30 November of that year. The headship history is reconstructed from the relationship codes, particularly from individuals identified as Head ("T" for "Tatane").

headflag*

This indicates whether the information in HouseholdHead_Anon is derived from the data "as is" or whether there was a data problem which necessitated an imputation process. It takes on the following values:

0. No imputation
1. Duplicate Heads – oldest "Head" imputed as Head
2. Missing Head
3. Missing Head – fixed by shifting episode end date to later
4. Missing Head – fixed by shifting start date of next episode sooner
5. Missing Head – no "T" in the household on 30 November, but the "HouseholdHead_Anon" identifier in the Agincourt HDSS table identifies a head

hhsizefull

Household size including temporary migrants. Range: 1 to 45

hhsize

Household size counting only permanent residents. Range: 0 to 41. There are 21 831 household-year observations where hhsize is zero. See discussion in section 2.6.3.

femalehead

This is a binary variable coded as:

0. Male Head
1. Female Head

There are 1050 missing values.

refugeehh

This is a non-time-varying indicator of whether the household is a "refugee" or a "South African" household. The household is characterised according to the status of the head of the household at the time that the household is first formed. This categorisation assumes that if (for instance) a Mozambican refugee married a South African, set up a household and then subsequently died, this

household would still be considered a “refugee” household even if the South African spouse is now defined as head. Implicit in this approach is the idea that the “social capital” of a household is relatively slow changing and the “refugee” label is intended to capture households that are relatively more marginal within the Agincourt population.
This is a binary variable coded as:
0. Not a refugee household
1. Refugee household
There are no missing values. Cases where the head of household was missing were coded as “not refugee.”

construct

Question: Is the structure still under construction? Codes:
-1 Don’t know/invalid answer
1 No
2 Yes
There are 216 984 missing values

extend

Question: Are there plans to extend the structure in the near future? Codes:
-1 Don’t know/invalid answer
1 No
2 Yes
There are 216 984 missing values

StartYear

Question: In what year was construction started? (Ask only if structure still under construction).
Range: 1959-2015. Special codes:
-2 Invalid year
-1 The household moved into an existing structure (code 9999 in the original HDSS data)
Comment: There were a lot of invalid entries but where the intended answer could be guessed. These have been cleaned. The code is available on request.

FinYear

Question: In what year was the building completed?
Range: 1900-2015. Special codes:
-2 Invalid year
-1 The household moved into an existing structure (code 9999 in the original HDSS data)
Comment: This variable has also been cleaned.

walls

Question: What is the construction material of the walls? Codes:
-1 Don’t know/invalid answer
1 Brick
2 Cement
3 Other modern
4 Stabilized mud
5 Traditional mud
6 Wood
7 Other informal
There are 217 223 missing values

roof

Question: What is the construction material of the roof? Codes:

-1 Don't know/invalid answer
1 Tiles
2 Corrugated iron
3 Other modern
4 Thatch
5 Other informal

There are 216 984 missing values

floor

Question: What is the construction material of the floor? Codes:

-1 Don't know/invalid answer
1 Tiles
2 Cement
3 Modern carpet
4 Wood
5 Other modern
6 Dirt
7 Mat
8 Other traditional

There are 216 984 missing values

Rooms

Question: What is the total number of bedrooms in all structures?
Range: -2 to 707. There are 216 986 missing values.
These numbers are as extracted from the HDSS database. Neither the negative value nor the extreme ones have been edited.

Bedrooms

Question: What is the total number of bedrooms in the main structure?
Range: -2 to 202. There are 216 985 missing values.
These numbers are unedited.

sepkit

Question: Is there a separate kitchen? Codes:

-1 Don't know/invalid answer
1 No
2 Yes

There are 216 984 missing values

sepliv

Question: Is there a separate living/dining room? Codes:

-1 Don't know/invalid answer
1 No
2 Yes

There are 216 984 missing values

toiletfac

Question: Where is the toilet facility? Codes:

-1 Don’t know/invalid answer
1 In house
2 In yard
3 Other house
4 Bush
There are 216 984 missing values

toilettype

Question: What is the type of toilet? Codes:
-1 Don’t know/invalid answer
1 Modern
2 VIP
3 Pit toilet
4 None
There are 216 984 missing values

watersup

Question: What is the main water supply? Codes:
-1 Don’t know/invalid answer
1 Tap in house
2 Tap in yard
3 Tap in street
4 Truck
5 Cement well
6 Traditional well
7 Pond
8 River
9 Dam
10 Rainwater tank
11 Other
There are 216 984 missing values

wateravail

Question: What is the availability of the main water supply? Codes:
-1 Don’t know/invalid answer
1 Always
2 Most of the time
3 Few hours a day
4 Irregular, not every day
5 Very irregular
There are 216 984 missing values

distmetre

Question: What is the distance to the main water source? (Only to be asked if water source is not tap in house or tap in yard). Codes:
-1 Don’t know/invalid answer
0 None (note: this code is not on the questionnaire, but in the data dictionary)
1 Immediate (<50 metres)
2 Nearby, but not immediate (50-200m)

3 Far away (>200m)
There are 222 653 missing values.

distmin

Question: Number of minutes required to walk to the main water supply?
Range: 0 to 190
There are 217 410 missing values
Comment: It appears that this question was asked only in 2001, 2003 and 2005. The value of 0 seems to be a "not applicable" code.

powerlight

Question: What is the primary source of power for light and appliances? Codes:
-1 Don't know/invalid answer
1 Electricity
2 Battery/generator
3 Solar power
4 Paraffin
5 Candles
6 Other
There are 216 984 missing values

powercook

Question: What is the primary source of power for cooking? Codes:
-1 Don't know/invalid answer
1 Electricity
2 Gas bottle
3 Paraffin
4 Wood
5 Other
There are 216 984 missing values.

stove

Question: Is there a functioning stove in the household? Codes:
-1 Don't know/invalid answer
1 No
2 Yes
There are 216 984 missing values

fridge

Question: Is there a functioning fridge in the household? Codes:
-1 Don't know/invalid answer
1 No
2 Yes
There are 216 984 missing values

tv

Question: Is there a functioning TV and/or hifi/stereo in the household? Codes:
-1 Don't know/invalid answer
1 No
2 Yes

There are 216 984 missing values

video

Question: Is there a functioning video machine or DVD player in the household? Codes:

-1 Don't know/invalid answer
1 No
2 Yes

There are 216 984 missing values

satdish

Question: Is there a functioning satellite dish in the household? Codes:

-1 Don't know/invalid answer
1 No
2 Yes

There are 216 984 missing values

radio

Question: Is there a functioning radio (no tape or cd player) in the household? Codes:

-1 Don't know/invalid answer
1 No
2 Yes

There are 216 984 missing values

fixphone

Question: Is there a functioning landline phone in the household? Codes:

-1 Don't know/invalid answer
1 No
2 Yes

There are 216 984 missing values

cellphone

Question: Is there a functioning cell phone in the household? Codes:

-1 Don't know/invalid answer
1 No
2 Yes

There are 216 985 missing values

car

Question: Is there a functioning car or truck in the household? Codes:

-1 Don't know/invalid answer
1 No
2 Yes

There are 216 984 missing values

mbike

Question: Is there a functioning motor bike in the household? Codes:

-1 Don't know/invalid answer
1 No
2 Yes

There are 216 984 missing values

bicycle

Question: Is there a functioning bicycle in the household? Codes:

-1 Don't know/invalid answer
1 No
2 Yes

There are 216 986 missing values

cart

Question: Is there a functioning animal drawn cart or sled in the household? Codes:

-1 Don't know/invalid answer
1 No
2 Yes

There are 216 984 missing values

cattle

Question: How many cattle are owned by the household? Codes:

-1 Don't know/invalid answer
1 None
2 1-3
3 4-10
4 more than 10
5 Cattle owned, but number unknown

There are 216 984 missing values

goats

Question: How many goats are owned by the household? Codes:

-1 Don't know/invalid answer
1 None
2 1-3
3 4-10
4 more than 10
5 Goats owned, but number unknown

There are 216 984 missing values

poultry

Question: How many chickens are owned by the household? Codes:

-1 Don't know/invalid answer
1 None
2 1-10
3 11-40
4 more than 40
5 Chickens owned, but number unknown

There are 216 984 missing values

pigs

Question: How many pigs are owned by the household? Codes:

-1 Don't know/invalid answer
1 None

2 1-3
3 4-10
4 more than 10
5 Pigs owned, but number unknown
There are 216 984 missing values

dn_av

This is the average digital number (DN) for the village of the household during that year. It measures the brightness on a scale from 0 (absence of light) to 63 (saturation). For more information see section 5.1 above and Machemedze *et al* 2017.
Range: 0 to 15
There are 62 567 missing values. These occur because the night-lights data only extends to 2012.

gen

The number of generations within the household counted according to the relationships to the head of the household. The logic is described in section 5.2.1 above.
Range: 1 to 5.

hhtype

This is the type of household, as discussed in section 5.2.2. Definitions are given in Table 8. Codes:
1 single person
2 couple
3 nuclear
4 single parent
5 3 generation linear
6 3 generation skip
7 Multi generation linear
8 Siblings only
9 nuclear with stepkids
10 3 generation with stepkids
11 Complex, related
12 Complex, plus unrelated
There are also 78 107 missing values.

weight

This is the weight to be used with the variables from the asset schedule to correct for missing information.
Range: 1 to 52.83205. The mean weight is 1.13859.
There are 216 984 missing values.

6.3 AgincourtIndivPanel

This is available only in the DataFirst Secure Research Data Centre. The file contains 1 904 374 records i.e. individual-year observations. The combination of Id_Anon and year uniquely identifies observations. Most of the variables in this file have not been modified from those extracted from the original AHDSS tables. Users should consult the Agincourt data dictionary (available on the Agincourt website) for the codes where these are not provided below.

Id_Anon

This is the anonymised individual identifier. There are 234 580 distinct individuals in the dataset. Id_Anon and year uniquely identify records in the panel.

year

Valid range: 1992 to 2015

Village_Anon

This is the anonymised village identifier. It takes on 36 distinct values, not all of which are valid village codes.

ExternalID

This is the dwelling identifier within the village.

Household_Anon

This is the anonymised household identifier. There are 29 285 distinct households. The individual panel can be merged with the household panel on Household_Anon and year.

age

This is calculated from the Date of Birth variable in the AHDSS data and is the age (in years) of the individual on 30 November of the year.
Range: 0 to 117. There are 76 missing values.

gender

The gender of the individual. Codes:

-1 Don't know/invalid answer
1 Female
2 Male

husbandstatus

This is a marital status variable asked of women aged 15 years and older. Codes:

-1 Don't know/invalid answer
1 Never married
2 Married
3 Divorced
4 Not living with husband
5 Widowed

There are 924 286 missing values

ResMonths

Number of months during the 12 months preceding the interview that the individual resided in the study area. This variable has not been edited.
Range: -12 to 1212. There are 544 567 missing values.

resident

This is based on the ResStatus variable in the AHDSS tables. It records the residency status of the individual. Codes:

1 Permanent – the person has lived more than 6 months of the previous year in the area
2 Migrant – member of the household but has lived less than 6 months of previous year in the area
3 The person migrated for the purpose of care or support
4 The person migrated for the purpose of education
5 Visitor – the person is not a member of the household
6 Migrated with a parent
7 Other/Query/Unknown

There are 41 361 missing values.

resstatflag

This variable flags cases where imputations were done to the "resident" variable. These imputations were to fill in gaps, copying a particular residency status forwards (particularly to the year 2015) or backwards (particularly for the years 1992 and 1993). Codes:

0 No imputation done
1 Information stays missing (no imputation done)
2 Value is imputed

HHRelation

This records the relationship to the head of the household. The relationship is described by the chain linking an individual to the head (as discussed in sections 3.1 and 5.2 above). The elements of that chain can be:

M Mother
F Father
B Brother
Z Sister
H Husband
W Wife (also W1, W2 for additional wives of the husband)
S Son
D Daughter
R Relation (indirect relative by marriage)
U Unrelated
X Unknown

The head of the household is identified as "T" (Tatane). After editing (see "relflag" below) there are 282 distinct relationship chains in the dataset.

relflag

This variable indicates whether the HHRelation variable has been edited or whether it is known to be suspect in some other way. Codes:

0 No editing done
2 Head Data suspect –missing head of household
3 Relationship data edited – these observations were changed

The manual edits reset *inter alia* relationship codes (to "unknown") where there were duplicate heads. Altogether 181 545 records were changed for this reason. Another 570 records were edited because the codes did not make sense (in general) or did not make sense in the context of the household. Examples of the former include resetting "ZW" to "WZ" (sister's wife to wife's sister, i.e. sister-in-law) and "BM" to "MB" (brother's mother to uncle). Examples of the latter include the "FM" (grandmother) recoded to "M" (mother) discussed above in section 5.2.1.

relhead

The relationship chains were categorised into types. The codes are:

1 Head
2 Spouse: H or W
3 Child: D or S
4 Parent: F or M
5 Grandchild: DD, DS, SD, or SS
6 Grandparent: FF, FM, MF or MM

7 Great-grandchild: DDD, DDS etc.
8 Great-great-grandchild: DDDD or DDDS
9 Sibling: B or Z
10 Parent-in-law: HF, HM, WF or WM
11 Brother/Sister-in-law: HB, HZ, WB or WZ
12 Stepchild: HD, HS, WD or WS
13 Step-grandchild: e.g. HDD, WSS
14 Child's stepchild: e.g. DHD, SWD
15 Nephew/Niece: e.g. BD, ZS
16 Uncle/Aunt: FB, FZ, MB or MZ, but also MMS (mother's half-sibling) and MZH (aunt's husband)
17 Cousin: e.g. FBD, MZS
18 Daughter/Son-in-law: DH, SW
19 Grandniece/nephew: e.g. BDD or ZSS
20 Granddaughter/son-in-law: e.g. SSW or SDH
21 Half-brother/sister: FD, FS, MD, MS
22 Co-wife: HW, i.e. another wife of the head's husband
23 Co-wife's kids: HWS or HWD
24 Co-wife's grandkids: e.g. HWSS
25 Great-grandniece/nephew: e.g. BDDD
26 Cousin's child: e.g. MZSS
27 Spouse of niece/nephew: e.g. BSW or ZDH
28 Half-sibling's children: e.g. FSD
29 Half-sibling's spouse: e.g. FSW
30 Cousin's spouse: e.g. MBSW or MZDH
31 Step-parent: e.g. FW, MH
33 Sibling's stepchild: e.g. BWD
34 Sibling-in-law's grandchild: e.g. HBDS or WZSS
35 Stepsibling: e.g. FWD or MHS (note these are different from half-siblings)
36 Great-Uncle/Aunt: e.g. FMZ
37 Related: e.g. sister's great-grandchild (ZDSS), nephew's brother-in-law (BSWB)
39 Unrelated
40 Unknown

Education

The highest education level obtained at the time of observation. For codes see the Agincourt data dictionary. There are 1 235 468 missing values.

CurrentEducation

Question: What grade/level is the person currently in? For codes see the Agincourt data dictionary.

SeniorCert

Question: If completed grade 12 or more, was senior certificate written? For codes see the Agincourt data dictionary.

SeniorCertOutCome

Question: If senior certificate was written, what was the outcome? For codes see the Agincourt data dictionary.

EverWorked
Question: Ever worked for pay? For codes see the Agincourt data dictionary.

CurrentlyWorking
Question: Currently working for pay? For codes see the Agincourt data dictionary.

Unemployment
Unemployment code: see Agincourt data dictionary.

UnemploymentSpec
Question: If other unemployment, specify. See Agincourt data dictionary.

PensionStudent
Indicates whether individual receives a pension or is a student.

PWorkType
Describes individual's primary type of work

PWorkCat
Individual's primary work category. See Agincourt data dictionary for codes.

PSector
Individual's primary work sector. See Agincourt data dictionary for codes.

PEmployer
Individual's employer for primary job. See Agincourt data dictionary for codes.

PPeriod
Time period of the primary job. See Agincourt data dictionary for codes.

PPlaceCode
Place of the individual's primary job. See Agincourt data dictionary for codes.

PTax
Indicates whether the individual pays tax for primary job. See Agincourt data dictionary for codes.

SWorkType
Describes individual's secondary type of work.

SWorkCat, SSector, SEmployer, SPeriod, SPlaceCode, STax
These detail the information in PWorkCat-PTax for secondary jobs.

Grant1
Records the type of Grant1 received. For codes see the Agincourt data dictionary.

Grant2
Records the type of Grant2 received. For codes see the Agincourt data dictionary.

6.4 AgincourtIndivData
This file contains basic time-invariant information on individuals which can be merged back into the individual panel (AgincourtIndivPanel) via the individual identifier Id_Anon.

Id_Anon
This is the anonymised individual identifier. There are 234 580 unique individual records in the file.

DoB

This is the individual's date of birth (in daily date format).
Range: 1 January 1800 to 31 December 2014 (from -58438 to 20088). The first date is equivalent to a missing date of birth. There are 12 individuals in this situation.

DoBestimated

This is a string variable to indicate whether the individual's date of birth was estimated.

DoD

This is the individual's date of death (in daily date format).
Range: 2 January 1994 to 4 March 2021 (from 12420 to 22343). The latter date is obviously a mistake. There are 221 034 individuals in our dataset where DoD is missing, i.e. they were still alive at the end of the period (2015) or when they migrated out of the study site.

DoDestimated

This is a string variable indicating whether the individual's date of death was estimated.

StartDate

This is the start date of the **first** membership episode of that individual. For many individuals that will be the date of birth or the date of in-migration. For household members that were already in the site when the surveillance system began in 1992 that will be the date that surveillance began. This will also be the case for household members in villages that were part of the expansion of the Agincourt site after 2007. This variable is also in daily date format.
Range: 15 January 1982 to 31 December 2014.

EndDate

This is the end date of the **last** membership episode of that individual. This will be either the date of death, or outmigration, or will be set to 31 December 2100 if the individual was still alive and member of a household in the study site during the last visit.
Range: 1 January 1994 to 31 December 2100 (from 12419 to 51499).

InitiatingMEventType

This is the type of event that initiated the **first** membership episode of that individual. It is a string variable with the following codes:

A Enumeration (this is where someone has been in the site prior to the start of fieldwork)
B Birth
H Household Head change (these must be data errors, since that type of change should not be coincident with the first time that we see an individual)
M In-migration

TerminatingMEventType

This is the type of event that ended the **last** membership episode of the individual. It is a string variable with the following codes:

C Current – this episode has not concluded
D Death
H Household Head change (again these must be data errors)
M Out-migration

refugee

This is a numeric version of the Refugee variable from the AHDSS database. The codes are:

1 South African
2 pre-93 refugee
3 post-92 arrival
4 other/query/unknown

References

Harris, T., Collinson, M. & Wittenberg, M., 2017. Aiming for a moving target: The dynamics of household electricity connections in a developing context. World Development 97,14-26.

Kahn, K., Collinson, M.A., Gómez-Olivé, F.X., Mokoena, O., Twine, R., Mee, P., Afolabi, S.A., Clark, B.D., Kabudula, C.W., Khosa, A., Khoza, S., Shabangu, M.G., Silaule, B.,Tibane,J.B., Wagner, R.G., Garenne, M.L., Clark, S.J. and Tollman, S.M., 2012. Profile: Agincourt Health and Socio-demographic Surveillance System. International Journal of Epidemiology 41:988–1001 doi:10.1093/ije/dys115

Machemedze, T., Dinkelman, T., Collinson, M.A., Twine, W. and Wittenberg, M., 2017. Throwing light on rural development: Using nightlight data to map rural electrification in South Africa. DataFirst technical paper 38. https://www.datafirst.uct.ac.za/images/docs/DataFirst-TP17_38.pdf

Tollman, S., 1999. The Agincourt field site - Evolution and current status. South African Journal of Medicine 89(8), 855-57.

Tollman, S., Herbst, K., Garenne, M., Gear, J. & Kahn, K., 1999. The Agincourt demographic and health study - Site description, baseline findings and implications. South African Journal of Medicine 89(8), 858-64.

Wittenberg, M. & Collinson, M.A., 2007. Household Transitions in Rural South Africa, 1996-2003. Scandinavian Journal of Public Health 35(suppl 69): 130-137.

Wittenberg, M. & Collinson, M.A., 2020. Household formation and service delivery in post-apartheid South Africa: Evidence from the Agincourt sub-district 1992–2012. Development Southern Africa 37(4):708-726.

Wittenberg, M., Collinson, M.A. & Harris, T., 2017. Decomposing changes in household measures: Household size and services in South Africa 1994-2012. Demographic Research 37(#39), 1297-326.