Data-driven versus traditional definitions of household membership and household composition in demographic studies: does latent class analysis produce meaningful groupings?

Estelle McLean, Alison J Price, Luigi Palla, Emma Slaymaker, Abena Amoah, Amelia C Crampin, Albert Dube, Fredrick Kalobekamo, Rebecca Sear

Introduction

'Household' is a commonly used unit of measurement in demographic and epidemiological studies. Assigning individuals to households for analytical purposes is useful for avoiding double counting, generating sampling frames, and assessing non-individual interventions and exposures. However, there is no universal, standardised definition for household and its meaning varies across different cultures(1). Household composition (the relationship of members to each other) is often used in analyses but with definitions varying from detailed, study specific composite definitions(2), to simplistic, binary definitions (e.g. nuclear family vs. extended)(3). The nuclear family, which includes only parents and children, is often inappropriately held up as the 'ideal' even if it is not meaningful in other contexts, which not only makes data interpretation difficult, but also perpetuates western-centrism which has damaging consequences beyond research conclusions(4). Simplistic definitions also fail to capture nuance and diversity. We investigate whether latent class analysis can be used to produce meaningful household composition groupings with different household membership definitions using data from Northern Malawi.

Methods

The Karonga Health and Demographic Surveillance Site (HDSS) was established in 2002 in northern Malawi(5). It covers an area of 150km2 and by 2016 had over 40,000 people under surveillance. Births and deaths are captured monthly, while migrations are captured annually. Where possible all participants are linked to their parent and/or spouse ids. The area is largely rural and the majority of the population engage in subsistence farming or fishing. The main ethnic group are Tumbuka, who have followed patrilineal and patrilocal custom since the 19th century: women tend to move to their husband's village when they marry(6). Polygyny is widespread: at the end of 2016 about 15% of households in the HDSS were headed by men with more than one wife.

The continuous HDSS data were reduced to include one data point per year (15 June each year) per person. Households were defined from the perspective of adolescents: all aged 12-18 (inclusive) were included except those who were already married or had a child, or who were not linked to their parents' identifiers.

Two household membership definitions were used:
1. Immediate: this is the standard HDSS definition defined by the participants with guidance from trained fieldworkers: all household members must usually live in the dwelling/compound together and recognise the same household head.
2. Expanded: households living close enough that they are likely to be sharing facilities are grouped. The cut-off distance for linkage varies by population density.

Latent class analysis (LCA) was used to generate household composition groups, it is a statistical technique which groups observations in otherwise unobserved, based on a set of categorical variables (here the presence or absence of types of relatives). For each record, the probability of membership to each latent class is calculated before it is assigned to the group for which it has the highest value (maximum probability assignment rule). LCA was carried out using the poLCA R package(7). The same analyses were carried out separately using data from 2004, 2007, 2010, 2013 and 2016. Models for 3-15 classes were run and the solution selection was guided by the Bayesian information criterion (BIC). For the selected models the average probability of class membership was always over 90% and entropy always over 80%. The classes found through both the 'immediate'

household and 'expanded' household analyses were distinct from each other and contextually appropriate. The independent LCAs conducted on datasets from the 5 separate years produce similar classes, and some similar classes found with both the 'immediate' and 'expanded' household analyses which also gave confidence in the class assignments. However, investigator input was required for choosing and coding the input variables, choosing the number of classes and interpreting and labelling them. Low probabilities of relatives in some classes complicate the labelling process, equally, uncertainty exists even in classes with high probabilities. Due to these complexities the investigators chose to create LCA-guided household composition variables with similar categories to the latent classes (table 1).

Table 1: Logical rules used to create LCA-guided categories

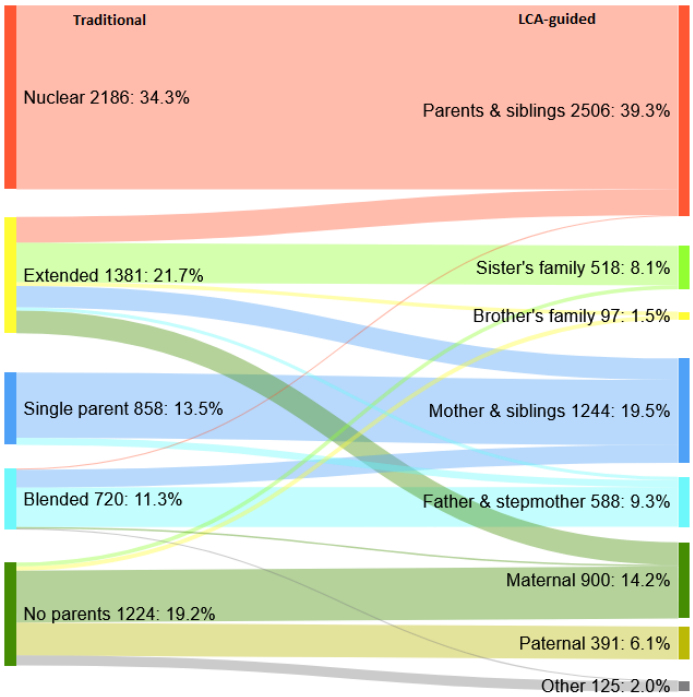
'Immediate' household categories	'Expanded' household categories
Parents & siblings: Both parents present and does not fit into any of the non-other categories	*Parents & siblings*: Both parents present in immediate household, brother and family and paternal family not present in expanded household and does not fit into any of the non-other categories
Sister's family: At least 1 over-18 sister or her family, sister+family larger than brother+family, and mother or father present or no maternal or paternal family present	*Sister's family*: At least 1 over-18 sister or her family, sister+family larger than brother+family, and mother or father present or no maternal or paternal family present, no brother and/or family in the expanded household
Brother's family: as above but with brother instead of sister	*Brother's family*: As above but with brother and sister reversed
	Parents & siblings->paternal: Both parents present in immediate household, paternal family present in expanded household and does not fit into any of the other non-other categories
Mother & siblings: mother present, no father, father's other wife nor maternal family	*Mother & siblings*: mother present, no father, step-mother or maternal family in expanded household
	Mother & siblings->maternal: mother present, no father, step-mother nor maternal family in immediate or expanded household
Father & stepmother: mother not present, father or father's other wife present	*Father & stepmother*: mother not present, father or step-mother present
	Polygynous: mother and another father's wife present in immediate or expanded household
Maternal: father and father's other wife not present, at least 1 maternal relative present and maternal relatives larger than paternal	*Maternal*: No mother, father nor other father's wife, at least 1 member of maternal family in immediate or expanded household and maternal family larger than paternal family
Paternal: mother and father's other wife not present, at least 1 paternal relative present and paternal relatives larger than maternal	*Paternal*: as above but with paternal rather than maternal
Other: Does not fit into any of the above categories	*Other*: Does not fit into any of the above categories

A ‘traditional’ household composition variable was also created based on definitions commonly used in the literature: nuclear (both parents present and only under 18 siblings), extended (both parents present plus others), blended (one parent and one step-parent present), single parent (only one parent and no step-parent present) and no parents (no biological parent present). The ‘traditional’ variable was also further simplified to nuclear vs. non-nuclear, where non-nuclear includes extended, blended, single parent and no parents.

Results

Using the ‘traditional’ household composition definitions and the ‘immediate’ household membership definition, the largest group of adolescents lived in a ‘nuclear’ household, however the percentage was relatively low at 34.3%. The largest category using the LCA-guided definition was ‘parents & siblings’ (39.3%). The correspondence between these two definitions is shown in a Sankey diagram in figure 1. There is not a one-to-one relationship between any of the categories, showing that the LCA-guided variable is not simply a more complex version of the traditional variable.

Figure 1: Sankey diagram showing the correspondence between the 5-level ‘traditional’ (left) and LCA-guided (right) household composition definitions



In total, 1462 (23%) adolescents change category from ‘immediate’ to ‘expanded’ household. Some categories had almost no changes (‘father and step-mother’, ‘paternal’) while ‘parents & siblings’, ‘sister’s family’, ‘mother & siblings’ and ‘maternal’ had higher rates of change. Those in ‘parents & siblings’ households moved to ‘brother’s family’, ‘parents & siblings ->paternal’ or ‘polygynous’, from ‘sister’s family’ the most common move was to ‘brother’s family’, from ‘mother & siblings’ people moved to ‘mother & siblings->maternal’ or ‘polygynous’ and from ‘maternal’ the most common move was to ‘mother & siblings->maternal’. The value of including nearby relatives is demonstrated here, providing more evidence for the complexity of living arrangements as even fewer adolescents can be categorised as living in a ‘nuclear’-type family (‘parents & siblings’), and also aids in distinguishing between those living in households which may be vulnerable (i.e. ‘mother & siblings’) and those in that category who may be less vulnerable due to proximity to other support networks (i.e. ‘mother & siblings->maternal’).

To assess the effect of the different household composition variables on analyses, logistic regression models were run using a binary education outcome of being delayed or dropped out of school. Each model was adjusted for socio-demographic factors and household size, the baseline group was ‘nuclear’ for the basic and ‘traditional’ variables, and ‘parents & siblings’ for the LCA-guided adolescent variables. There was no association between household composition and the education outcome when using the basic definition (adolescents categorised as living in either ‘nuclear’ or ‘non-nuclear’ households). When using the ‘traditional’ 5-level household composition definitions, for both female and male the ‘blended’ (female aOR=1.7, p=0.004, male aOR=1.3 p=0.042) and ‘no parents’ (female aOR=1.6, p=0.006, male=1.5 p=0.006) categories were associated with increased

likelihood of being behind or dropped out of school. With the LCA-guided 'immediate' household definition, adolescents living in the 'father and step-mother' category had higher odds of experiencing the outcome for both female (aOR=1.8, p<0.002) and male (aOR=1.4, p<0.018), and female adolescents living in 'maternal' households had higher odds of being behind or having dropped out of school (OR=1.7, p=0.003). Male adolescents living in 'paternal' households had higher odds of the outcome (aOR=1.5, p=0.027), while for female adolescents there was less evidence for an association using the 'immediate' household but these adolescents did have higher odds of the outcome when using the 'expanded' household (aOR=1.9, p=0.029). There were few additional associations found with the categories only found with the 'expanded' household, though there was some evidence that female adolescents in 'brother's family' households were more likely to experience the outcome (aOR=1.5, p=0.055).

Conclusions and recommendations

LCA provided a robust approach to better understanding the extent of variation in the data and to guide generation of context appropriate household membership and composition definitions, however we found it most useful as a guide for manually derived variables. Our findings suggest that context appropriate definitions provide a more nuanced understanding of the relationship between household composition and other factors such as educational outcomes. Using the LCA guided variable it was possible to identify variation in the effects of the context specific household types (e.g. 'father and stepmother', 'maternal', 'paternal') that were categorised together as extended family in the 'traditional' 5-level variable. Equally, the finding that for girls, living in a 'maternal' household was associated with being behind or dropped out of school, while for boys the same was only true for 'paternal' households would not have been found using 'traditional' definitions. Using the expanded household definition also provided valuable understanding into how adolescents are living in this area.

We recommend that other researchers consider gathering data to allow development of more nuanced household membership and composition definitions, and if only secondary data are available to analysts should carefully consider whether it has enough meaning in their context before using it in analyses.

References

1. Randall S, Coast E, Leone T. Cultural constructions of the concept of household in sample surveys. Vol. 65, Population Studies. Taylor & Francis Group; 2011. p. 217–29.
2. Pilgrim NA, Ahmed S, Gray RH, Sekasanvu J, Lutalo T, Nalugoda F, et al. Family structure effects on early sexual debut among adolescent girls in Rakai, Uganda. Vulnerable Child Youth Stud [Internet]. 2014 Jul 1 [cited 2018 May 28];9(3):193–205. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25317199
3. Akinyemi JO, Chisumpa VH, Odimegwu CO. Household structure, maternal characteristics and childhood mortality in rural sub-Saharan Africa. Rural Remote Health [Internet]. 2016 [cited 2016 Nov 16];16(2):3737. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27107479
4. McEwen H. Nuclear power: The family in decolonial perspective and 'pro-family' politics in Africa. Dev South Afr. 2017 Nov 2;34(6):738–51.
5. Crampin AC, Dube A, Mboma S, Price A, Chihana M, Jahn A, et al. Profile: the Karonga Health and Demographic Surveillance System. Int J Epidemiol [Internet]. 2012 Jun [cited 2015 Feb 2];41(3):676–85. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3396313&tool=pmcentrez&rendertype=abstract
6. Malawi Human Rights Commission. Cultural Practices and their Impact on the Enjoyment of Human Rights, Particularly the Rights of Women and Children in Malawi. 2006.
7. Linzer DA, Lewis JB. poLCA: An R package for polytomous variable latent class analysis. J Stat Softw. 2011 Jun 14;42(10):1–29.