



Exercise: De-identification of quantitative data

In order to make microdata collected as part of the Northern Ghana Millennium Villages project evaluation publicly available via the UK Data Service, to encourage research and ensure openness and transparency, all variables in the household survey were assessed for disclosure risk, with recommendations for action.

The table below shows some assessed variables (variables commonly assessed for disclosure risk such as age, community, but also variables for which local knowledge is essential to indicate risk) with identified disclosure risk. Which action would you take to reduce the disclosure risk?

Source: University of Sussex. Institute of Development Studies, Columbia University (New York). Earth Institute. (2017). Millennium Village Impact Evaluation in Northern Ghana, 2012-2015: Special Licence Access. [data collection]. 3rd Edition. UK Data Service. SN: 7734, <http://doi.org/10.5255/UKDA-SN-7734-3>

Variables	Disclosure risk	Action
Community	Low frequency counts for all named communities, respondents who gave answers very easily identifiable (especially in combination with other variables).	
Age	Low counts of older respondents over 75 years old	
Main occupation during last 12 months	Low counts of very specific occupations.	
Ethnicity of the Household Head	Low counts of specific ethnicities.	
Household's primary type or energy/fuel used for cooking: Firewood, Electricity-based, Charcoal, Electricity-solar panel, Gas/LPG	Very low counts for 'Gas/LPG' and 'Electricity-solar panel' responses may lead to household identification (especially if combined with other datasets)	
Main material of the wall of the house, e.g. Cane/palm/trunks, Dirt/mud, Wood/bamboo with mud, Stone with mud, Uncovered adobe, Cardboard, Cement blocks, Stone with lime/cement, Bricks, ...	A number of low-frequency responses; exterior features (households/buildings easily identifiable).	
Crops grown on plots	A number of low-frequency specific responses for each variable.	

