Subject: Re: What is the DHS position on the use of cluster-level data?
Posted by user-rhs on Sun, 27 Mar 2016 20:47:07 GMT
View Forum Message <> Reply to Message

The key issue is that some of the clusters may be too small for the analysis to be meaningful. For example, how would you interpret the average from 3 households? You weren't clear in your original question about how you were using the aggregates, but since you cited the Storey & Kaggwa and Kravdal papers and quoted Reduced-For(u)m's responses, it looks like you want to enter them as covariates in a regression model.

There are no hard and fast rules about what to do, and determination should be done on a case-by-case basis and depends on: 1.) the overall sample size, 2.) the number of clusters, 3.) the number of observations in the cluster. If you have many clusters with a sufficient number of observations, your results will be less biased than if you many clusters with a small number of observations. For example, I would be fairly comfortable entering cluster-level averages into a model from the Indonesia 2012 DHS into a regression model, because 1.) it's huge (>45,000 observations), 2.) it has a lot of clusters (1,832 clusters), 3.) the clusters are sufficiently "large" (around 90% of the clusters have 20 or more people in them, and only about 80 people live in clusters with <10 observations each), but I would have less confidence doing it with a dataset with 5,000 observations and 1,200 clusters where the average size is 10 (I worked with a dataset like that once, and I ended up aggregating up to the district level to get respectable sizes) .

Second, cluster-level analysis can still be useful, depending on the level of inference. I think what the DHS team cautions against is making population-level inference based on the clusters, because the survey is not designed for that level of disaggregation. You can make a case for valid inference to the sample in the worst-case-scenario, or at least minimize the population-level implications of your findings.

Third, even the experts are still in disagreement about this, which works to your advantage. You can take one school of thought and justify it with citations from the peer-reviewed literature. At the end of the day, science is about weighing different opinions and evidence and defending your choices.

A good first step is determining the number and size of your clusters. If they are sufficiently "large" and you can make a case for it that's theoretically/empirically/clinically plausible, then why not? Cluster aggregates are less than ideal, but if we had better measures than cluster-level aggregates for whatever construct we were trying to operationalize, surely we would have used them instead of these proxies derived from the data, right? I would make the suggestion to fit the model first with just the individual/household-level variables first and enter the cluster-level aggregates separately to see how things change. Reduced-For(u)m has some good advice, which you have quoted above.

NB: I'm not a DHS affiliate, so I can't offer the official DHS position