Subject: Global identifier between BR and GPS shapefiles
Posted by Yonemese on Tue, 09 May 2017 12:20:25 GMT
View Forum Message <> Reply to Message

Greetings,

I have downloaded both the GPS and DHS survey datasets. The shapefiles have been really helpful when using QGIS.
However, as I have tried to join the data from the surveys (BR datasets) with the spatial information from the GPS dataset in Stata, I have encountered myself with a big problem. This refers to the identifier for the merging of these two datasets. As recommended elsewhere, one should merge these two by using as key identifier the "DHSCLUST" figure of the geospatial dataset, in relation with the "v001" detail of the DHS survey data. They, in principle, should be equal and then, the geospatial information would be added to the already DHS survey information. This would be the ideal "masking".

However, I have noticed that these specific numbers given to each cluster within a country, and depending on the phase of the survey, begins with "1" and tends to a finite "n" number. The problem then is that these numbers might repeat for different countries, and in different years, and so the merging could end up linking not matchable characteristics, i.e. we can end up having the geospatial information for some -for example- Nepal cluster linked with the information from a Brazilian cluster -more over, in different years-, while for instance, they both could have as key identifier the number, say, "14".

On the other hand, if we create our unique identifier by, let´s imagine, concatenating the country code (first two letters), number of cluster id (that goes from 1 to n), and the year of the survey -e.g.: for Brazil, cluster 3, in survey made in 2009, we would have:"BR-3-2009"-, we might have a more reliable "joining". However, some times we have surveys made in different years in the same country. And so, we don´t really know which record to keep while survey waves could be done through more than one year, i.e. a survey wave began one year and finished the following one -or maybe even the year after the next one, e.g. 2008-2010. Or just, different waves in different years.

Having said this: Could you please be so kind in telling me how I might efficiently deal with this problem of the identifier?