Webinar on Working with Geography Variables in IPUMS DHS
August 25, 2020

Questions and Answers

Note: The Stata code from the Webinar presentation is reproduced at the end of this document

Does IPUMS DHS have subnational geographic variables?

Absolutely! Those subnational geographic variables are the subject of this webinar.

How does one get access to IPUMS DHS data?

Information on variables (which samples have which variables, information about individual variables) are freely available to everyone from the IPUMS DHS website at dhs.ipums.org. If you wish to download IPUMS DHS data to analyze on your computer, you must apply for access from The DHS Program at https://dhsprogram.com/data/new-user-registration.cfm. If you have already been approved as a DHS user by The DHS Program, you can log in at the IPUMS DHS website, using that username and password, and start downloading data immediately.

Are there IRB requirements?

There are no additional IRB requirements, because the data are anonymized in the original DHS public use files that IPUMS DHS uses as source data.

Do IPUMS DHS data come in separate files, like the KIDS file or the Individual recode (women's) file?

You create your own data files when downloading DHS data from IPUMS DHS; the files are not pre-constructed. You select the samples and variables that interest you, and IPUMS creates a tailored file for you. If you select more than one sample, your data file will include the data from all the samples, with the variables fully harmonized across them.

You do have to select a unit of analysis when creating your IPUMS DHS data file. The units of analysis in IPUMS generally map onto the file types at The DHS Program website, but with the added flexibility that, within each unit of analysis, you can select many variables from other file types. For example, all of the mother's variables are available if you select Children as the unit of analysis, and all of the household variables are available if you select Women, Children, Births, Men or Couples as the unit of analysis. This means that with IPUMS you don't have to worry about downloading multiple files and either appending or merging them.

The following units of analysis pair with the following original DHS data file types: Women/IR; Children/KR; Births/BR; Household members/PR; Men/MR; and Couples/CR (with the last to be released in IPUMS DHS in 2021). The DHS household or HR files can be created by selecting household members as the unit of analysis and keeping just the rows in which LINENO equals 1.)

The examples in the webinar used women as the unit of analysis. How do we link the geographic data with other DHS data, such as the data on children?

When you first click on GET DATA, you have the option of choosing children as your unit of analysis. You can add any of the contextual or geographic variables to the DATA CART at that point, and these variables will appear in your customized data file. Put simply, all of the geography variables are available for any of the units of analysis.

What should be the configuration of one's computer to download IPUMS DHS data, and how long does it take to download a dataset?

You can access the IPUMS DHS website and select the samples and variables for your customized dataset using any common web browser, such as Chrome or Firefox. When you complete the specifications for your dataset, you can indicate whether you want the data as a SAS, SPSS, or Stata file, or as an ASCII or CSV file. Because you download a dataset consisting of just the samples and variables you need, leaving out many thousands of DHS variables you don't need, it should be faster to download your customized IPUMS DHS dataset than to download an original public use dataset from The DHS Program. How long the download takes will depend on your Internet connection.

You will receive an email when your customized data file is ready to download. After you click the button to download the data, they will most likely appear in your Downloads folder. Your data file will be zipped (with a .gz extension); you need to unzip the file before you analyze the data. Directions for unzipping a file and accessing decompression software are available here: https://www.idhsdata.org/idhs/extract_instructions.shtml

Most people find the downloading, unzipping, and opening files to be an easy process. If you run into difficulty though, there are some helpful tutorials on opening your data file here: https://www.idhsdata.org/idhs/tutorials.shtml

Is IPUMS DHS's geographic integration only at the region level?

If the original DHS public use files include other geographic variables, such as a variable identifying provinces of a country, then we also include that variable in IPUMS DHS. You can see an overview of the IPUMS DHS geography variables here: http://www.idhsdata.org/idhs/gis.shtml

Do you provide shape files for single surveys, such as Tanzania 2015?

Shape files for single samples are already available directly from The DHS Program's Spatial Repository.

It would be superfluous for IPUMS DHS to distribute those single-sample share files. IPUMS DHS instead distributes shape files for integrated geographic variables, available here: https://www.idhsdata.org/idhs/gis.shtml

Sometimes there is more than one integrated geographic variable in IPUMS DHS for a country, such as GEO_TZ1991_2015 and GEO_TZ1996_2015 for Tanzania. How should a researcher choose between these variables?

First, you should determine which samples include the variable you want to study over time. To identify which samples include a variable of interest, you can use the Search tool on the IPUMS DHS website and see which samples show an X (indicating variable is included in that sample) for that variable in the

"Select Data" part of the IPUMS DHS website. If a variable that interests you is only available for 1996 forward, in the example of Tanzania, then use GEO_TZ1996_2015. If the variable is also available for Tanzania 1991 and you want to look at a longer time series, then you should use GEO_TZ1991_2015. A second factor to keep in mind is how much geographic detail matters to you. The oldest DHS samples generally have the least geographic detail, and integrated geographic variables that include older samples will generally have less geographic detail.

Is there a place where the user can find the documentation for boundary changes and descriptions of those changes?

IPUMS DHS includes variable-specific documentation for each (integrated) variable in the dataset. Go to the "Select Data" part of the IPUMS DHS website and click on the name of a variable to pull up this documentation (which we call "variable descriptions"). One tab, headed "Comparability," covers boundary changes for the integrated geographic variables.

Why does IPUMS DHS not include administrative level 2 (e.g., districts) geography for some countries?

The geographic detail included in a particular DHS sample is a decision made by the government agency administering the survey in each country, in collaboration with ICF. IPUMS DHS has no control over the level of geographic detail available. We start with the public use files from The DHS Program when integrating the data, and we can include only the level of geography included in those source files.

Do the DHS data allow small area analysis (for example, districts in Tanzania or Wards in the Lake Victoria region)?

The DHS is not representative at very small levels of analysis, so that would be a concern. That said, people do sometimes use small levels of analysis to construct their own contextual variables. They do this by aggregating information at the cluster level or by connecting the DHS to census data. Examples of articles that do this include:

Boyle, E. H., & Svec, J. (2019). Intergenerational transmission of female genital cutting: Community and marriage dynamics. *Journal of Marriage and Family*, *81*(3), 631–647.

Hayford, S. R., & Trinitapoli, J. (2011). Religious differences in female genital cutting: A case study from Burkina Faso. *Journal for the Scientific Study of Religion*, *50*(2), 252–271.

Kravdal, Ø. (2002). Education and fertility in sub-Saharan Africa: Individual and community effects. *Demography*, *39*(2), 233–250.

In September, 2020, IPUMS DHS will begin to include linking keys to facilitate the use of small-area geography variables constructed from the censuses in IPUMS International.

You mentioned the possibility of combining DHS data with IPUMS International census data. The DHS cluster sample points are displaced before the data are supplied to researchers. Given this displacement, how can we link DHS data to census data at a lower administrative level?

Displacement of DHS cluster points ranges from 0 to 2 kilometers in urban areas to 0 to 5 kilometers for rural areas, with an additional 1 percent of rural clusters displaced up to 10 kilometers. Clusters are not displaced across first-level administrative survey regions (or across second-level administrative regions

for some recent surveys) or national boundaries. The linking code to link to census data from the IPUMS International database are calculated keeping in mind the values within the radius of displacement.

Is there an available list of all the variables included in PUMS DHS?

IPUMS DHS currently includes over 15,000 integrated variables, and more integrated variables are added whenever we process another sample. It would therefore not be very helpful to display such a list, and the list would not indicate which variables are in which samples. To find variables of interest, use the drop-down menu for topics of interest or search for a variable name. Note that you can restrict the display to only the samples/countries that interest you. You can also use the original DHS variable name (for standard variables), as well as IPUMS DHS mnemonic variable names and likely keywords, in the search engine on the IPUMS DHS website. You can find a summary of the geographic variables in IPUMS DHS, by country, here: https://www.idhsdata.org/idhs/gis.shtml

How do you pool IPUMS DHS data or create weights? Are the weights pre-defined by The DHS Program?

An advantage of getting data from IPUMS DHS is that, if you include multiple samples in your customized data file, the IPUMS DHS system creates a pooled dataset "on the fly." Pre-selected variables automatically added to your dataset allow you to distinguish between the pooled samples, with variables specifying the country, sample year, country-sample year identifier, and weights. To denormalize the weights, that is, to weight each sample up to its full population, see this User Note.

If one is doing a fixed effects regression in a panel setting, which variable would you recommend using, so the region fixed effects are consistent over time?

Note that the DHS and IPUMS DHS data are repeated cross-sections, so these data cannot be used as panel data at the individual level. You could, however, include an integrated geography variable as a fixed effect in a regression analysis. This is one advantage of the geographic integration in IPUMS DHS.

Are there any plans for adding geographic data on road traffic accidents or other contextual variables?

IPUMS DHS is interested in adding more contextual variables to the database. We welcome suggestions for additional contextual data to include. Most useful are suggestions that direct us to a publicly available dataset available for multiple countries; please send suggestions for such sources to include in future to ipums@umn.edu. Our immediate priority is adding to IPUMS DHS the contextual variables created and distributed through The DHS Program. When considering which contextual variables to recommend, keep in mind the displacement of sampling clusters by The DHS Program. Such displacement makes adding contextual variables on distance to something--say, a major road or a market or a healthcare center--less useful, since the actual distance may be off by 5-10 kilometers in rural areas.

Do you plan to release DHS calendar data in future?

Yes, IPUMS DHS will be releasing calendar data in spring 2021. Calendar data are retrospective data collected on the timing of births, pregnancies, pregnancy terminations, and contraceptive use over the five years prior to the time of a woman's DHS interview. We will release calendar data as a new unit of analysis, woman-months, in which each observation consists of one month at risk of reproductive events. Because the calendar data are retrospective data, they support longitudinal analysis, such as

event history, of reproductive events and contraceptive use over five years. The calendar data released by IPUMS DHS will simplify analyzing these valuable and under-used data.

You said that IPUMS DHS includes standard, continuous, and interim DHS surveys. Will you be integrating other DHS data files, such as the AIDS monitoring (AIS), malaria monitoring (MIS) and service provision (SPA) datasets?

We will be seeking another five years of funding from NICHD in 2021, so we can add data for additional regions (East Asia, Latin America, Eastern Europe, Caribbean, Oceania) as well as newly released samples for countries already in IPUMS DHS. Our annual survey of registered IPUMS DHS users indicates that the highest demand is for including standard and continuous DHS samples for all regions of the world. There is also substantial, but not as great, demand for adding additional types of samples, including the AIS, MIS, and SPA samples. We will do as much as we can in the next 5-year phase of IPUMS DHS. Ultimately, we hope to include all public DHS samples in IPUMS DHS.

Is there a web API to access the data or documentation for IPUMS DHS?

This is coming soon (sometime over the next 12 months). Our IT team is working on this for all IPUMS projects.

Do you plan to harmonize the MICS geography with the DHS regional geography?

We are happy to report that NICHD recently approved funding for an IPUMS MICS database. One of the goals laid out in our grant proposal was maximizing comparability between IPUMS DHS and IPUMS MICS data. How much cross-project geographic harmonization can be done will depend on the specific geographic information included in the MICS samples.

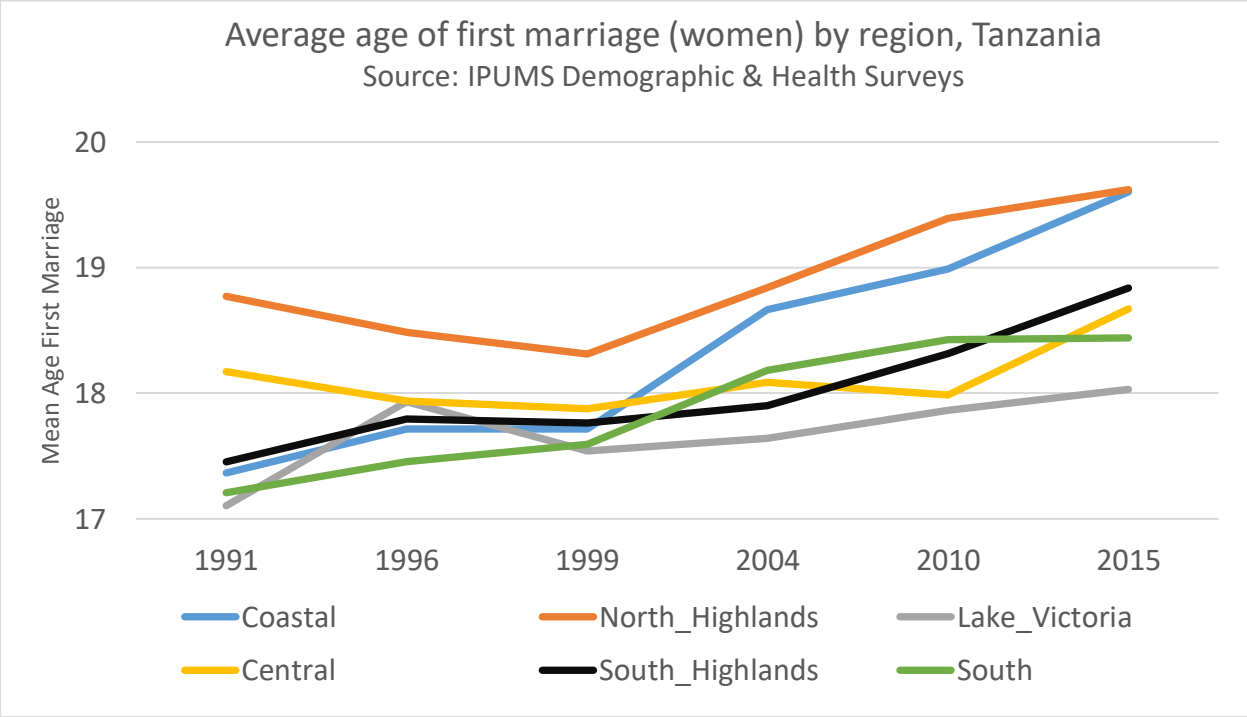
Will you please do a future webinar on how to extract GIS and contextual data, particularly the rainfall and temperature data?

We plan to do further webinars, including a webinar on using the contextual variables in IPUMS DHS. In the interim, you may find the following article on contextual variables in IPUMS DHS useful: Elizabeth Heger Boyle, Miriam L. King, Sarah Garcia, Corey Culver, and Jordan Bourdeaux, "Contextual Data in IPUMS DHS: Physical and Social Environment Variables Linked to the Demographic and Health Surveys," Population and Environment, May 2020. (https://link.springer.com/content/pdf/10.1007/s11111-020-00348-4.pdf)

Stata code

We used Stata 16.

We transferred the results from Stata to Excel to create both of the figures. Cutting and pasting output into Word, saving it as a .txt file, and importing it into Excel is one reasonable way to do this.



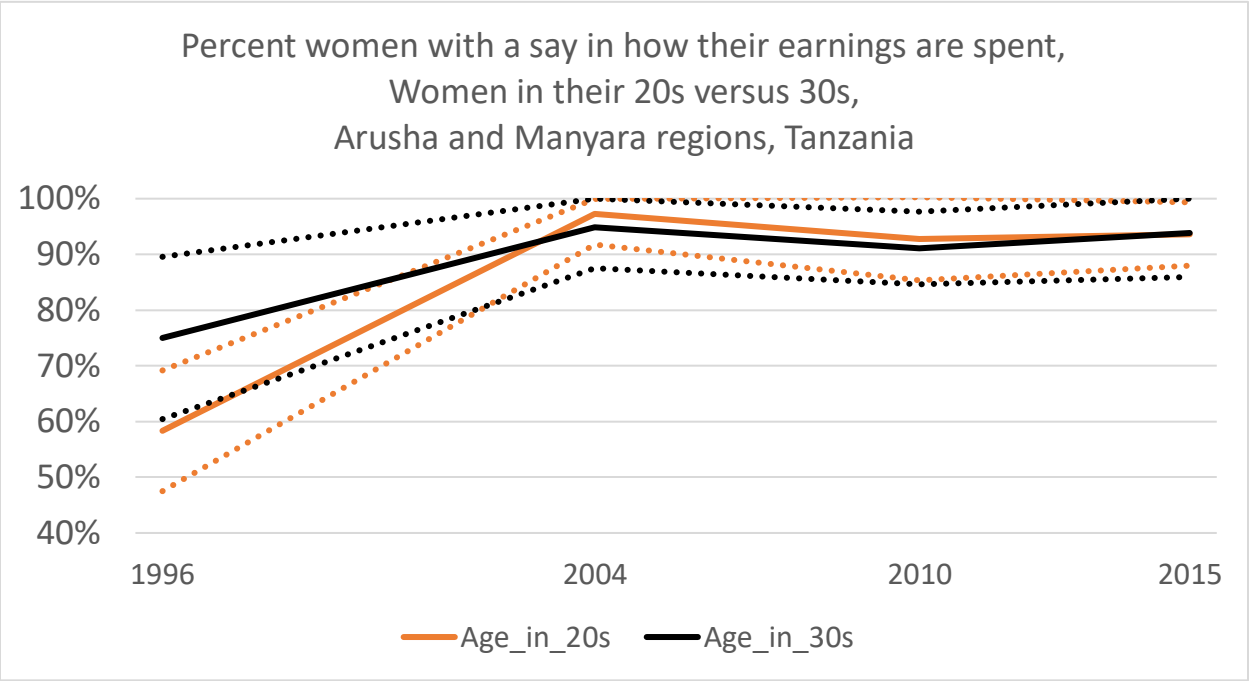
This is the Stata code to produce the output that's the basis for the figure above.

```
* Make the region labels shorter
label variable geo_tz1991_2015 "Harmonized regions"
label define region 1 "Coastal" 2 "North_Highlands" 3 "Lake_Victoria"
   4 "Central" 5 "South_Highlands" 6 "South", modify
label value geo_tz1991_2015 region

* Define missing values
recode agefrstmar (99=.)

* Don't forget to weight!
svyset, clear
svyset idhspsu, strata(idhsstrata) weight(perweight)
vce(linearized)///
  singleunit(centered)

* Calculate the mean for each region in each survey year
svy: mean agefrstmar, over(geo_tz1991_2015 year)
```



This is the Stata code to produce the output that's the basis for the second figure above:

```
* Create "married" variable to make samples comparable
recode marstat (21/22=1 "Married")(98=.)(else=0 "Not_married"),
gen(married)

* Recode decision-making over women's earnings
recode decfemearn (40/50=0 "Someone_else")(10/30=1
"Woman_has_say")(98/99=.), gen(rdecfemearn)

* Recode age
recode age (20/29=20 "Age_in_20s")(30/39=30 "Age_in_30s")(else=.),
gen(agecats)

* Aggregate the variable of interest by region and year. Don't forget
to weight!
svyset, clear
svyset idhspsu, strata(idhsstrata) weight(perweight) vce(linearized)
singleunit(centered)

* The calculation
svy: mean rdecfemearn if geo_tz1996_2015==2 & married==1, over(agecats
year)
```