IPUMS International: Using Geographic Variables and Shape Files Webinar
June 27, 2019 (11:00 a.m.-12:00 p.m. CST)

QUESTIONS AND ANSWERS

The following are the questions received during the live webinar and their answers.
For more user support, email IPUMS at ipums@umn.edu.

Can you review what the primary and secondary levels of geography are? What determines what falls into the primary or secondary level of geography?

In common geographic terminology used by the United Nations and many other institutions, within country, administrative level-1 (primary level of geography) represents the largest sub-national division that exhaustively partitions the country. For example, level 1 would mean, states in the United States, Germany, Brazil or provinces in Kenya, Pakistan, etc. The second administrative level exhaustively partitions level-1 units into smaller units. For example, counties in the United States, municipios in Brazil or districts in Kenya. Most countries have progressively lower level units of geography (third, fourth and beyond). The divisions tend to correspond to geopolitical divisions indicating some kind of administrative control.

How would one use first or second level geographic data to study regions that have undergone federal/disputed border changes? (e.g Crimea, former Yugoslav countries, Czechoslavakia)

IPUMS distributes data entrusted to us by the National Statistical Offices (NSO). If the NSO has microdata for a disputed territory, IPUMS includes it within the respective country. Disputed territories in most cases result in overlapping polygons in the world map in cases where multiple countries enumerate the same geographic area.

Could you explain European NUTS data? How do those data differ from IPUMS administrative units? Except for Spain, is there no level 2 information for the NUTS 3 regions?

European NUTS variables identify the Nomenclature of Territorial Units for Statistics (NUTS) within Europe in which the household was enumerated. NUTS is a standard administrative division of the European Union (EU) and was developed by the EU. The European Free Trade

Association extends the NUTS system to several additional countries outside of the EU and they are also incorporated into the IPUMS European NUTS variable. NUTS is organized into NUTS1, NUTS2, and NUTS3 at increasing geographic detail, and correspond the ENUTS1, ENUTS2, and ENUTS3 variables available on the IPUMS International database. The code labels for ENUTS variables include the standard code for the NUTS system and the name of the NUTS region, separated by a slash. IPUMS distributes GIS files associated with the NUTS variables and they can be found here.

Austria, Greece, Ireland, Portugal, Romania, Slovenia, Spain, and Switzerland have NUTS 3 regions in IPUMS data. More on the NUTS variables can be found here.

What is spatial harmonization?

Spatial harmonization is the process of creating a single geographic variable with consistent boundaries across all sample years to facilitate comparisons over time. Where geographic boundaries of modern units do not align with historical census units because of boundary changes, larger aggregated units are created that remain stable over time. If units split or merged, the harmonized unit will have the boundaries of the largest version of the unit; if a territory is redistributed between two or more units, the units are combined. We refer to this process as harmonization of geographic boundaries or spatial harmonization. More on the process of harmonization is addressed in our working paper.

What do you mean by "confidentialization" or regionalization? What makes 20,000 the magic number for the minimum population in defining regions?

IPUMS distributes microdata about individuals and households only by agreement of collaborating national statistical offices and under the strictest of confidence. Limiting geographic detail is one of the primary means statistical offices employ to ensure confidentiality. If geographical units have less than 20,000 populations, they are grouped with a contiguous unit until they exceed that threshold. We refer to this process as regionalization. Regionalization is not required for samples whose total populations at the first and second level of geography are greater than 20,000 persons. More on the process of regionalization is addressed in our working paper.

When regionalizing, do you take into consideration that both areas belong to the same higher spatial level (i.e. all municipalities belong to the same region)?

Regionalization is always performed within the corresponding geographic hierarchy provided by the country. If lower level geographic units need to be combined for confidentiality reasons, they are only combined with units also within their parent regions.

How is confidentiality handled over time when the population is changing? That is, what if an area grows to over 20,000 in a later year and you want to compare regions?

When performing regionalization for a set of samples, the populations from the most recent sample year is used to create the harmonized variable. In situations where we receive additional data from partner countries for a later census, the process of harmonization and regionalization is repeated to create new variables and GIS shapefiles that include any new boundary changes, and using the populations of the most recent sample to create the new harmonized and regionalized variable.

Are you using computing for regionalizing?

We use the Regionalization with Dynamically Constrained Agglomerative Clustering and Partitioning (REDCAP) algorithm and accompanying software. Regionalization is conducted using population density, such that the algorithm combines geographic units that have similar population compactness. REDCAP enforces spatial contiguity and creates regions while optimizing the sum of squared differences.

References:

Guo, Diansheng. 2008. "Regionalization with Dynamically Constrained Agglomerative Clustering and Partitioning (REDCAP)." International Journal of Geographical Information Science 22 (7): 801– 23. doi:10.1080/13658810701674970.

Guo, Diansheng, and Hu Wang. 2011. "Automatic Region Building for Spatial Analysis: Automatic Region Building for Spatial Analysis." Transactions in GIS 15 (July): 29–45. doi:10.1111/j.1467- 9671.2011.01269.x

To confirm, if an administrative level 1 unit splits into two administrative level 1 units at any point between the country's first and most recent census, then in the spatially harmonized first-level geography shapefile, it will appear as one unit or two?

If a country were to have a single unit split into two units between sample years, the harmonized file will contain one unit. The label for said unit will have the names of the corresponding units that were combined to create the stable boundary. For example, as seen in the webinar presentation, Tanzania had a single unit, "Arusha", split into two units, "Arusha" and “Myanara” between the 1988 and 2002 censuses. In the resulting harmonized geography variable, there is one unit, consisting of both regions combined together, with the label, “Arusha, Myanara”.

In your example of literacy rates over three decades, if one were only interested in a VISUAL representation of change in literacy rate over time, is it necessary to use the harmonized data? Presumably using the higher resolution data that is year-specific would potentially be even more informative. Is it correct that harmonized polygons are only needed when some sort of CHANGE statistic is being calculated?

Year-specific variables are ideal for users studying one specific place and time. Year-specific variables provide greater detail than spatially harmonized variables because they do not need to account for changes over time by aggregating units together that otherwise meet the 20,000 population threshold. In the webinar presentation, we were looking at the percent change in literacy in the 1990, 2000, and 2010 census rounds, so we would use consistent boundaries (harmonized data) across all sample years to facilitate comparisons over time.

Some samples have information on migration, and on previous residence (including geolevel1 data), has that information been harmonized too?

We are currently working on migration variables that will have codes matching the GEOLEV1 variable. These new geographically harmonized migration variables will be available with our scheduled 2019 data release.

When do you expect you will move on to harmonize geolevel2 previous residence information?

Currently all available migration variables are matched up by the names of the units and does not take into consideration changing boundaries. Presently we are working on spatially harmonizing household geography codes to the migration codes at the primary level of geography. We are also working on creating global migration units that match GEOLEV1 for place of previous residence, place of residence one year ago, five years ago, and 10 years ago. Once this work is completed, we will focus on harmonizing second level of geography (GEOLEV2) to migration.

About the DHS crosswalks, are these only available for the first level geography (because as you mentioned, there are no district variables in DHS)? Do you also have DHS regions?

Crosswalk geography variables identify administrative units in IPUMS-DHS countries that match those in the IPUMS-International census data. The variables correspond to the primary level of geography in both IPUMS-DHS and IPUMS-International. The geography variables in the two databases allow researchers to summarize census data and attach these data as contextual information to the DHS samples or vice versa. DHS geography is only available for the first administrative level of geography (also known as regions in DHS) and therefore can only be linked to the first level of geography in IPUMS International. More on DHS geography can be found here.

Why do the crosswalk codes not match the codes from both DHS and census datasets (e.g. 1-1-10)?

There is no uniform coding system for geographic locations and it is common for separate projects to code the name of a place differently even though it refers to the same geographic space. From the webinar presentation, in the crosswalk slide, DHS codes their primary level of geographic units from 1 to 5, whereas census uses geographic codes starting at 10. The use of the crosswalk is to streamline these codes into a common coding scheme such that it is easy for researchers to use census data from IPUMS-International in conjunction with health data from IPUMS-DHS. The crosswalk variables not only use a uniform coding scheme, but also takes into consideration changing boundaries.

Is it possible to merge DHS Kyrgystan (2012) and IPUMSI Census Kyrgystan (2009)?

Although Kyrgyz Republic data is available in IPUMS international, it is not yet available in IPUMS DHS. At this point we are unable to provide a crosswalk variable for the Kyrgyz Republic in IPUMS.

Do all the shapefiles use the same coordinate system? Do they include an attribute table?

All shapefiles provided by IPUMS International use the same geographic coordinate system of GCS_WGS_1984. Attribute tables include information on the country code and the geography code that the users would need to link (or join) the GIS shapefile to the microdata downloaded from the IPUMS International website.

Are shapefiles for India also included in IPUMS?

Both year specific and harmonized geographic variables and shapefiles are available for India in IPUMS. All available geographic variables and shapefiles for countries in IPUMS are listed on the IPUMS International website, under the "Geography and GIS" tab.

Could you the open source GIS packages?

Some of the open source GIS packages that would read GIS shapefiles include QGIS, GVSIG, and R.

Do you use a particular satellite image for representation or did you analyze your data?

We do not use satellite imagery for the creation of our variables, and instead use maps and/or images provided by our partner countries. In cases where no maps or images are provided, we turn to other sources such as published census figures, printed maps, and other information from the countries' statistical offices. In rare cases, such as the 1983 Guinea sample, countries will provide us with geographic data, but we are unable to locate credible maps or images corresponding to the data, and in these cases, no geographic variables or shapefiles are created.

In a country void of third level administrative data, as a researcher going to the field to obtain his/her data, how do you think he/she can go about getting a population sample? What is the optimum sample for clustering?

IPUMS mostly has data on the first and second level of geography, for some countries (like Bangladesh, Cameroon, Mali) there is information on the third administrative level of geography, but that data has not been spatially harmonized (meaning, it does not take into consideration changes in boundaries between censuses). Individual censuses usually suppress microdata at the thirdadministrative level of geography as it becomes highly confidential. Perhaps aggregate data (and not microdata) from the census can be used for research at lower levels of geography.

How can we access the multiple countries' data and how to we get access?

The tutorial page on IPUMS International provides a step-by-step video tutorial on creating an IPUMS account and also creating a customized data extract. The video is available in Chinese, Portugese, Russian, Spanish, and Swahili. Also, please refer to our previous webinars on "Intro to IPUMS International" and "IPUMS International; Una Introducción" in Spanish language for information on international census microdata and how to get access to the data.

What is the authenticity of the data and how can we measure it? How do you check the raw data is accurately collected and recorded when you get it from each country?

IPUMS applies a number of data quality checks when processing input data. We check basic characteristics against published total where possible and work hard to identify household breaks accurately by checking characteristics of the household. We empirically verify and document the universe for each variable, label or recode implausible or stray values, and evaluate frequency distributions for accuracy. IPUMS assesses the quality of age reporting in source files using standard indices such as Whipple's Index, Myers Index and the United Nations age/sex accuracy index. We have also undertaken cohort based coherence tests, measuring the consistency of the distribution of an invariant characteristic in two successive censuses such as education.

Why would we use male and female ratios in discovering literacy; does geography effect our education?

The spatial distribution of education is of interest to several policy makers and researchers. For example, urban areas tend to be more educated than rural areas. Disparity in education is one of the important areas of research as Goals 4 and 5 of the Sustainable Development Goals (SDG) from the United Nation strive for quality education (goal 4) and gender equality (goal 5) by the year 2030.

Do the censuses for Malawi, Mozambique, etc. cover full population or it is just a representative sample? Do you have full-count census data?

For all contemporary censuses (like Malawi and Mozambique), only a sample is accessible through the website. More on the samples and their density can be found in the sample descriptions page. For some historical censuses, full-count data are available. To access the historical censuses, follow the Select Samples page, and then select the Historical tab. Full count data are available for Canada, Denmark, Iceland, Norway, Sweden, United Kingdom, and the United States. Internally, we hold full count data for about 1/3 of the data files available as samples. However, they are in archival storage and not even available to researchers within IPUMS, except for select data checking or methodological work. We are working with some countries to make these higher density and full-count data available via a more secure portal environment, so please keep an eye out for that.