Accessing NHGIS Data In R

March 5, 2024

Questions and Answers

The following are adaptations of questions received during the live webinar and their answers. If you have additional questions, please contact NHGIS User Support at nhgis@umn.edu

Webinar materials

1. Where can I find a recording of the webinar?

This is the webinar recording; the recording will also be listed on the IPUMS Tutorials site.

2. Where can I find the slides shown during the webinar? I would like to test out the code snippets.

Here is a link to the slides; a copy is also available on the IPUMS Tutorials site. A similar series of code snippets is also provided in the ipumsr article on NHGIS API Requests.

3. What is the citation for the journal article that motivated the example (i.e., homeownership inequity by county)?

Chantarat T, Van Riper DC, Hardeman RR. Multidimensional structural racism predicts birth outcomes for Black and White Minnesotans. Health Serv Res. 2022;57(3):448-457. doi:10.1111/1475-6773.13976C

Ipumsr

1. Does IPUMS provide access to R via a cloud server?

Unfortunately, IPUMS does not provide a cloud server with R that researchers can use. We currently operate under the assumption that users have R installed on their local computer. IPUMS microdata sites include access to an online analysis tool (see the IPUMS USA tool as an example), which allows users to create quick tables and graphs as well as a number of more complex analyses..

2. What does read_nhgis do if you download multiple tables that don't have the same set of columns?

NHGIS delivers one file per dataset, including all tables that were selected for that dataset. All the tabular data files for a single extract will be contained in a single zip archive. `read_nhgis()` will read all of the tables from a single file (i.e., dataset) into a single data frame. NHGIS assigns a unique column code to every table and variable in its collection, so there would never be a conflict in column names when multiple tables are in one file or data frame. For the “source tables” provided in NHGIS datasets, we make no adjustments to harmonize different sets of columns from different datasets. Our time series tables, on the other hand, do have consistent categories and column codes across time.

3. Can you specify the location where you want the zip file to be saved?

Yes, you can. You cannot change the file name, but you can specify the directory where the file will be saved. This code snippet shows how you specify a path (“`data/nhgis`”, as an example) in the `download_extract()` function.

```
define_extract_nhgis(
   description = "2017 ACS Tenure by race and ethnicity",
   datasets = ds_spec(
   "2013_2017_ACS5a", data_tables = c("B25003B", "B25003D", "B25003H", "B25003I"),
   geog_levels = "county"
   )
) |>
 submit_extract() |>
 wait_for_extract() |>
 download_extract("data/nhgis")
```

4. Is the extract also saved locally or only available in R?

The data extract that you download from IPUMS NHGIS will be saved locally. You can access it via R or via some other software package (e.g., Microsoft Excel). The extract definition is typically in your R script; thus, as long as you save your R script, you will have access to the extract definition in the future. You can also save your extract definition to a JSON file with `save_extract_as_json()`.

5. What is the smallest unit of geography for `geog_level` in the extract definition specification?

The smallest unit of geography varies by dataset, but for the Decennial Census, the census block is the smallest unit of geography. In the American Community Survey, the block group is the smallest unit of geography. To find out what geographic levels are available in specific datasets, you can view the table contained in the `geog_levels` field of the metadata for that dataset.

6. How do we use ipumsr to get and use shapefiles?

You can add shapefiles to your extract request using the `shapefiles` argument of `define_extract_nhgis()`. Use `get_metadata_nhgis("shapefiles")` to view shapefile names and descriptions.

You can load shapefiles that you have downloaded from NHGIS with `read_ipums_sf()`, which loads into `sf` format (see the sf package). As with tabular data, if your extract contains multiple shapefiles, you will need to select a single file to load with the `file_select` argument.

7. Will extracts submitted via the ipumsr package show up in our extracts history page on the website along with our extracts created via the NHGIS Data Finder?

Yes, extracts submitted via the ipumsr package will be shown alongside extracts created via the NHGIS Data Finder. From R, you can view your past extracts with `get_extract_history()`.

8. Is there a way to filter your extract before requesting and downloading? It seems like pre-filtering would speed up the extract download process by cutting down on file size.

We currently support geographic extent selection for census blocks and block groups. You can request blocks or blocks groups for one or more specific states. We plan to extend extent selection for other geographic levels (e.g., census tracts, counties, places) in the future.

9. With a view of reproducibility -- do you see it as preferable for people to keep extracts as ZIP files and use these ipumsr functions to work with compressed data? In my normal use of NHGIS I would download the data, unzip, rename etc., but this makes me think I could change my approach

Either way will work for reproducibility. The key thing is to store your extract definition either in your R code or in a configuration file so that you can always re-run the code to generate the same extract. Then, you can either use read_nhgis() to load data into R, or you can unzip. We still use both ways of loading data into R - read_nhgis() and read_csv(). Of course, leaving your extract compressed does also save space on your machine.

10. Building on the data example from the webinar, let's say I want to compare homeownership across income levels at the neighborhood level using the 5-yr ACS data. What would that process look like?

Once you have metadata for a specific dataset, you can find specific data tables (e.g., for median household income) and specific geographic levels (e.g., census tracts or block groups to represent “neighborhoods”) by filtering the output of get_metadata_nhgis(). (In the webinar, we use the dplyr package to filter the metadata output.) You can also save the metadata output as a data frame and scroll through the table, or use the `View()` function to open an interactive

table in RStudio. Ultimately, identifying new tables in the metadata will take some manual exploration.

11. What are the advantages of using ipumsr vs. a product like tidycensus?

We are big fans of tidycensus, and we think ipumsr and tidycensus are complementary products. Ipumsr provides functionality to access IPUMS NHGIS (and other IPUMS products), which provides summary data and GIS mapping files from 1790 - present. Thus, we cover a longer time period and different data types than tidycensus, which uses the APIs provided by the Census Bureau. That provides access to census data only back to 2000 and no integrated time series tables. We also provide access to metadata in a different way from tidycensus, and some users may prefer our presentation over tidycensus.

API

1. How do I find my API key?

If you have an account for any IPUMS product, you can access your API key at https://account.ipums.org/api_keys. If you have never registered for an IPUMS product, you must do so before you can get an API key. However, note that the IPUMS API is currently only available for certain IPUMS collections.

2. Is there a rate limit for IPUMS APIs?

Yes, the IPUMS APIs have a rate limit of 100 requests per minute. In most cases, you probably won't reach this limit. However, it is possible, particularly if you are iteratively making requests for NHGIS metadata. You can avoid this by manually setting a delay in your iteration code to prevent excessive API calls.

There are also a few ipumsr functions that include a `delay` argument to add a delay in between API calls when hitting a paginated endpoint. However, it is still unlikely that these endpoints would cause you to reach the rate limit.

3. Will you provide a sample extract definition in JSON format?

The IPUMS API developer website provides examples of JSON-formatted extract definitions.

General IPUMS questions

1. How does NHGIS fit into the IPUMS product list? Does NHGIS only provide data for the US?

IPUMS includes a variety of products, including many that provide international data. NHGIS provides summary data and GIS mapping files for the United States. Check out the other IPUMS products to learn more about their data holdings.

2. Can you get American Community Survey-like data for other countries?

Yes, IPUMS IHGIS (International Historical Geographic Information System) provides access to summary data and GIS mapping files for many other countries. The summary data in IHGIS is often similar to what is found in the American Community Survey.

3. What is the timeline for supporting metadata access via ipumsr for other IPUMS products (e.g., IPUMS USA, CPS, or Health Surveys)?

We know this is of interest to API and IPUMS microdata users, and it is currently on our to-do list. Unfortunately, we do not have a definitive timeline for this work.

4. Is there a Python equivalent to ipumsr?

We have a product called ipumspy, and it's under active development. It currently provides support for IPUMS microdata products that have APIs. We don't currently have NHGIS support built into ipumspy, but we have plans to add such functionality in the future.

Originals to delete

5. IPUMS is global correct? NHGIS is US only? is it a subset of IPUMS ?
6. Could you get ACS from other database like International for NHGIS similarities>?
7. What is the timeline for metadata support for IPUMS USA and CPS?
8. Will the functionality to browse metadata in ipumsr for other IPUMS surveys (particularly for the CPS and the Health Surveys) be available at some point in the future?
9. is there a python equivalent to ipumsr?

General NHGIS questions

1. I'm sure there is a good reason, but why is it that NHGIS puts the extract number into the file name? It seems to be more disruptive of workflows than it is useful.

By assigning a unique extract number to the file name, it is easier to track the provenance of the extract and connect the extract to corresponding codebooks. We recommend not renaming your data file and instead using a local macro to call on the extract number and updating this as needed. If we were to instead name these based only on the content or on the description you provide, there is potential for creating duplicate file names, or the file name could become unruly (e.g., it could contain illegal characters or be too long).

2. Does NHGIS have data for Pacific Island territories like Guam and American Samoa?

No, at this time, NHGIS does not provide data for Pacific Island territories.

3. How far back in time do NHGIS data holdings extend?

NHGIS provides data from 1790-present. The first census of the United States took place in 1790.

4. Do you provide any data or tools to map 2020 census tracts onto any prior census tract boundaries?

We have geographic crosswalks that provide interpolation weights for allocating data from 2020 blocks or block groups to 2010 tracts, or vice versa, and similarly, from 2000 or 1990 areas to 2010 tracts. The time series tables also include a selection of data from 1990 through 2020 for 2010 census tracts.

5. Currently, it looks like you have geographically standardized tables of time series data for 1990, 2000, 2010, and 2020 data for 2010 census units. At some point in the future, will you create geographically standardized tables for the yearly ACS 5-year average data for each year between 2005-2020?

We have a grant application currently under review to do exactly that! So, we plan to, but we do not have a timeline yet. In the meantime, you could use our geographic crosswalks to generate your own standardized ACS data.

6. If I have my own data that contains a census tract GEOID for each record, how can I join this to data available from NHGIS? Is there an alternative to a spatial join?

NHGIS data always includes an ID that we identify as the "GISJOIN" ID. Our most recent datasets usually also include the Census GEOID, so you could use that directly. Otherwise, it's possible to transform a GISJOIN to a GEOID for data from recent decades. (GISJOINs add an extra digit to state and county FIPS codes to make room to distinguish historical entities that had never been assigned a FIPS code.)

7. Are the Table and Field names consistent across time, so that Field Name AH39EE001 is the same Census question year over year? Is there a mapping function when the historical census categories subtly change — for example Family Members <5, 5-12, 12-18, >18 shifted at some point to <3, 3-12, 12-18, 18-24, >24? I'm looking for very long longitudinal comparisons across the same (or close to the same) categories.

We don't have a universal solution to this. NHGIS table names are unique across the entire collection and aren't designed to indicate comparability across datasets. Our time series tables, however, are designed to address this exact issue. They provide data for consistent categories

across time. But they're also limited to particular years, subjects, and geographic levels--those that we've completed--so they might or might not include what you're looking for.

8. Is there any in-built linking between these variables and the relevant denominator variables? Such as households vs general population. Like, the education variables are based on 25+ populations, which is provided in the same table by USCB. Are there any controls/backend handling to ensure or facilitate the correct comparisons?

In general, no. As you say, the Census Bureau tables generally include the universe count (like age 25+) in the same table, and NHGIS maintains that for tables we've added after 2000. When you request the table, it includes that count, but we don't do anything in our data extract pipeline to link the universe count.