Accessing Data with the Census Bureau API

Alex Shum

March 5, 2014

1 Introduction

The United States Census Bureau has been conducting a decennial census since 1790. Originally this census was a simply to count the population across the country. More recently the decennial census includes a short-form asking for name, sex, age, and a few other demographic variables. About one in six households also received a long-form that contained additional socioeconomic questions. After the 2000 decennial census many of the long-form questions were collected as part of a new survey: the American Community Survey (ACS).

The ACS is an ongoing yearly survey that collects additional demographic variables including but not limited to age, sex, race, income and education. Unlike the decennial census, the American Community Survey is distributed based on a random selection of addresses every year. Although the ACS is only sent to a sample of all US households, this data is meant to provide more up to date information than the Census Bureau's decennial census. Both the decennial census and the American Community Survey are required to be completed by law; however it should be noted that the Census Bureau has not opted to prosecute anyone for failure to complete the decennial census or the ACS. Despite the lack of enforcement, the ACS still reports a response rate of 97%.

Both the decennial survey and the ACS data are used in part by federal, state and local agencies to allocate state funding and for policy decisions. The Census Bureau has also released some of this data for public use. Many of the data sets are available directly in a compressed format from the Census Bureau's FTP site: `http://ftp2.census.gov/`. Since 2012, the Census Bureau has also included an online developer's API in order to improve accessibility of the ACS and decennial census datasets. The Census Bureau's online API can be accessed online: `http://api.census.gov`.

We will discuss how the ACS data is structured when we request data and how to access data from the Census Bureau's online developer's API. We will also discuss what kind of variables are available and some limitations with the API. We will use "Taking a Chance in the Classroom: The American Community Survey", an article by Morgan, Cetinkaya-Rundel and Stangl as a starting point on some of the limitations of the API. This article explores some multivariate frequency distributions using data from the ACS dataset; however, there are some gaps in what we can access and inconsistencies in the database.

2 Geography

The Census Bureau has a very sophisticated system of hierarchy for geography. For the ACS, at the top level there is the entire nation, followed by region, division, state, county, county-subdivision, tract, block group, place, congressional district, zip code area, school district and a few other geographic divisions. See table 1 for a complete table of geographies available on the census API for the 2012 ACS.

From table 1, the hierarchy for geographic regions is very specific in the combinations of geographic regions that we can specify. Different ACS datasets might have slightly different geographic regions available. Fortunately, each dataset available on the Census Bureau online API include an associated geography file

	Summary Level	Description
1	010	us
2	020	region
3	030	division
4	040	state
5	050	state-county
6	060	state-county-county subdivision
7	140	state-county-tract
8	150	state-county-tract-block group
9	160	state-place
10	250	american indian area/alaska native area/hawaiian home land
11	310	metropolitan statistical area/micropolitan statistical area
12	320	state-metropolitan statistical area/micropolitan statistical area
13	330	combined statistical area
14	340	state-combined statistical area
15	350	new england city and town area
16	400	urban area
17	500	state-congressional district
18	510	state-congressional district-county
19	610	state-state legislative district (upper chamber)
20	620	state-state legislative district (lower chamber)
21	795	state-public use microdata area
22	860	zip code tabulation area
23	950	state-school district (elementary)
24	960	state-school district (secondary)
25	970	state-school district (unified)

Table 1: List of valid geographic combinations

formatted in JSON and a similar file formatted in XML. JSON stands for JavaScript Object Notation and it is a lightweight machine format for sending data. JSON is structured using name-value pairs and is also designed to be human-readable. There are many libraries to generate and process JSON. The JSON formatted file for geographies has the following format:

```
{
  "name": "tract",
  "requires:" [
    "state",
    "county"
  ],
  "optionalWithWCFor": "county"
}
```

This JSON file specifies that for census tract level geography we are required to specify a state. Additionally we can also specify both a state and a county.

Alternatively, the geographic information is available in XML. XML stands for Extensible Markup Language and is similar in structure to the HTML format used for webpages. XML is another format for sending and storing data. It is formatted in a tree-like structure with a hierarchy of categories with associated values. Here is the same geographic area information formatted in XML:

```
<fips name="tract">
  <requires name="state"/>
  <requires name="county" is-optional-with-wcfor="true"/>
</fips>
```

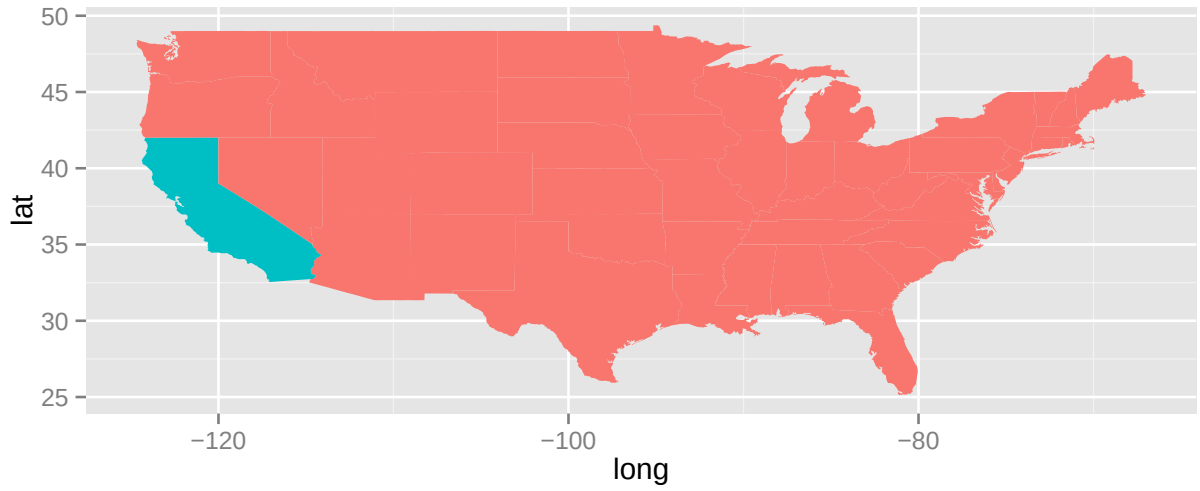


Figure 1: California selected

In the following example, we will examine various geographies available from the online API for the state of California.

At the top of the hierarchy we can view data at the country level; this is data aggregated among all states and corresponds to summary level 010 from table 1. Below that we can view data by state. At the state level we can view data for all states or select a particular state; this is summary level 040. In figure 1 we have selected California.

Below states we can select counties and census tracts. At the county level, after we specify California we can look at counties within California (summary level 050). We can see from figure 2 that we have selected Los Angeles county within California state. There are a few other geographic regions we can subset by within a state such as ZIP code tabulation area; we could have selected 90210 which corresponds to Beverly Hills, California.

There are a number of geographies available below the state and county level. For example, the API also supports school districts, county subdivision, metropolitan statistical areas and legislative districts. The difficulty is that it is not possible to request data for many of the smaller geographic divisions without specifying a state or a county. In figure 3 we have selected Pasadena, a county subdivision, within Los Angeles County. For county subdivisions we cannot simply specify Pasadena and we also cannot specify just California and Pasadena; subdivision alone and state-subdivision are not a valid geographic combinations. Instead, we must specify California, Los Angeles County and then Pasadena. We can see from table 1 that state-county-subdivision is a valid summary level.

Although all the entries in table 1 are valid geographic combinations, some parts of geographic combinations may be optional. For example, if we wanted census tracts we need to specify state and county (summary level 140). In contrast to county subdivisions where we needed to specify both state and county, for census tracts specifying county is optional. This means that state-tract is also a valid geographic combination. It is best to refer to each datasets JSON or XML file for geographic compatabilities.

Another complication is that not all geographic divisions are compatible. That is to say that some of the smaller geographic divisions are not necessarily nested in one of the larger geographic divisions; for example, ZIP code areas are generally used by the United States Postal Service and might span different counties or census tracts. Additionally, Legislative districts do not line up with county borders and school

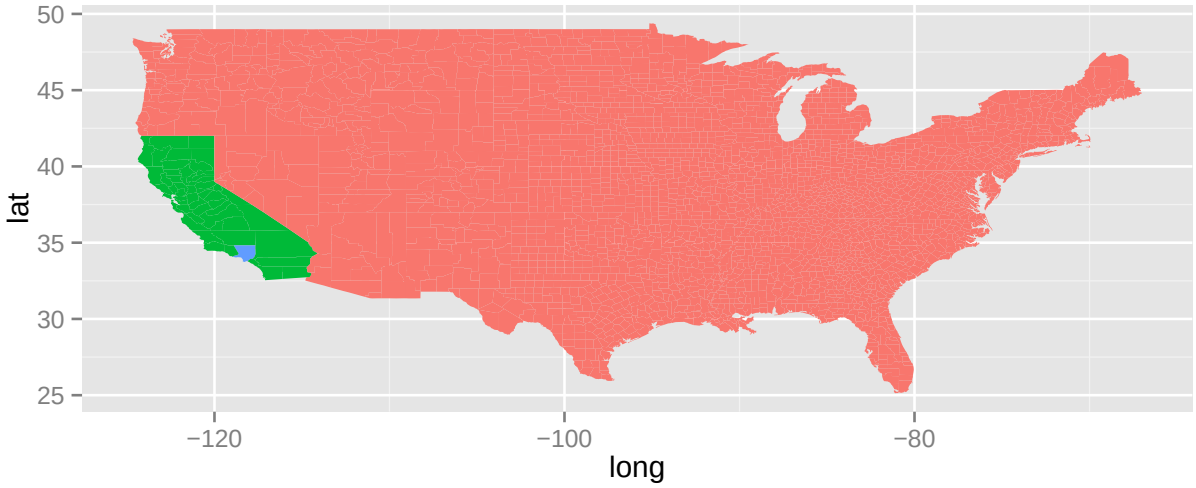


Figure 2: Los Angeles county selected

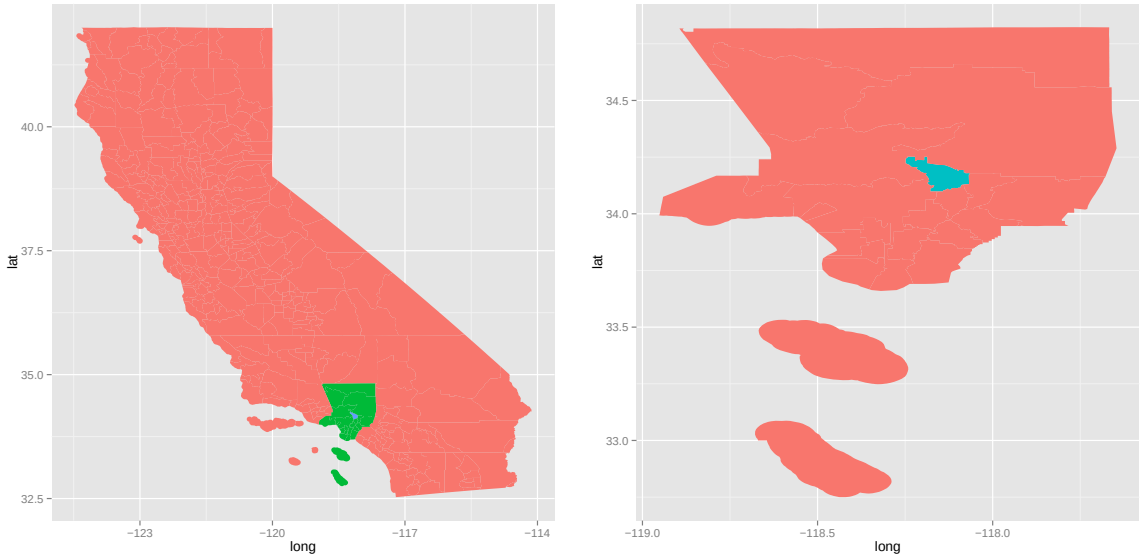


Figure 3: Pasadena county subdivision selected

districts often do not line up with either legislative districts or county borders.

For a more detailed look at which geographies need to be specified, refer to the census bureau list of summary levels for each dataset or the XML or JSON geography files. For the 2012 ACS dataset this is located at `http://api.census.gov/data/2012/acs5/geo.html`, the XML file is located at `http://api.census.gov/data/2012/acs5/geography.xml` and the JSON file is located at `http://api.census.gov/data/2012/acs5/geography.json`.

3 Finding Data Sets and Tables

A recent addition to the Census Bureau API is the inclusion of a master index of available dataset formatted in both JSON and XML.

4 Requesting Data and Table Structure

5 Limitations

The dataset used by Morgan, Cetinkaya-Rundel and Stangl in "Taking a Chance in the Classroom: The American Community Survey" is a random subset of the 2010 ACS public use microdata sample. This article contains a number of classroom exercises that ask the reader to calculate some basic proportions about various demographic data. We will attempt to use the Census Bureau online API to examine the same demographic variables.

Using the Census Bureau's online API for these exercises presents us three main problems. The first problem is the structure of how the data is organized in the public use microdata sample versus how the data is organized in the online API. The second problem is that the census does not provide proportions and finding standard errors for these proportions requires a bit more work. Finally, the third problem is that certain combination of variables are simply not available from the online API.

	Sex	Age	Married	Income	HoursWk	Race	USCitizen	HealthInsurance	Language
1	0	31	0	60.00	40	white	1	1	1
2	1	31	0	0.36	12	black	1	1	0
3	1	75	0	0.00		white	1	1	0
4	0	80	0	0.00		white	1	1	0
5	1	64	1	0.00		white	1	1	0
6	1	14	0			white	1	1	0

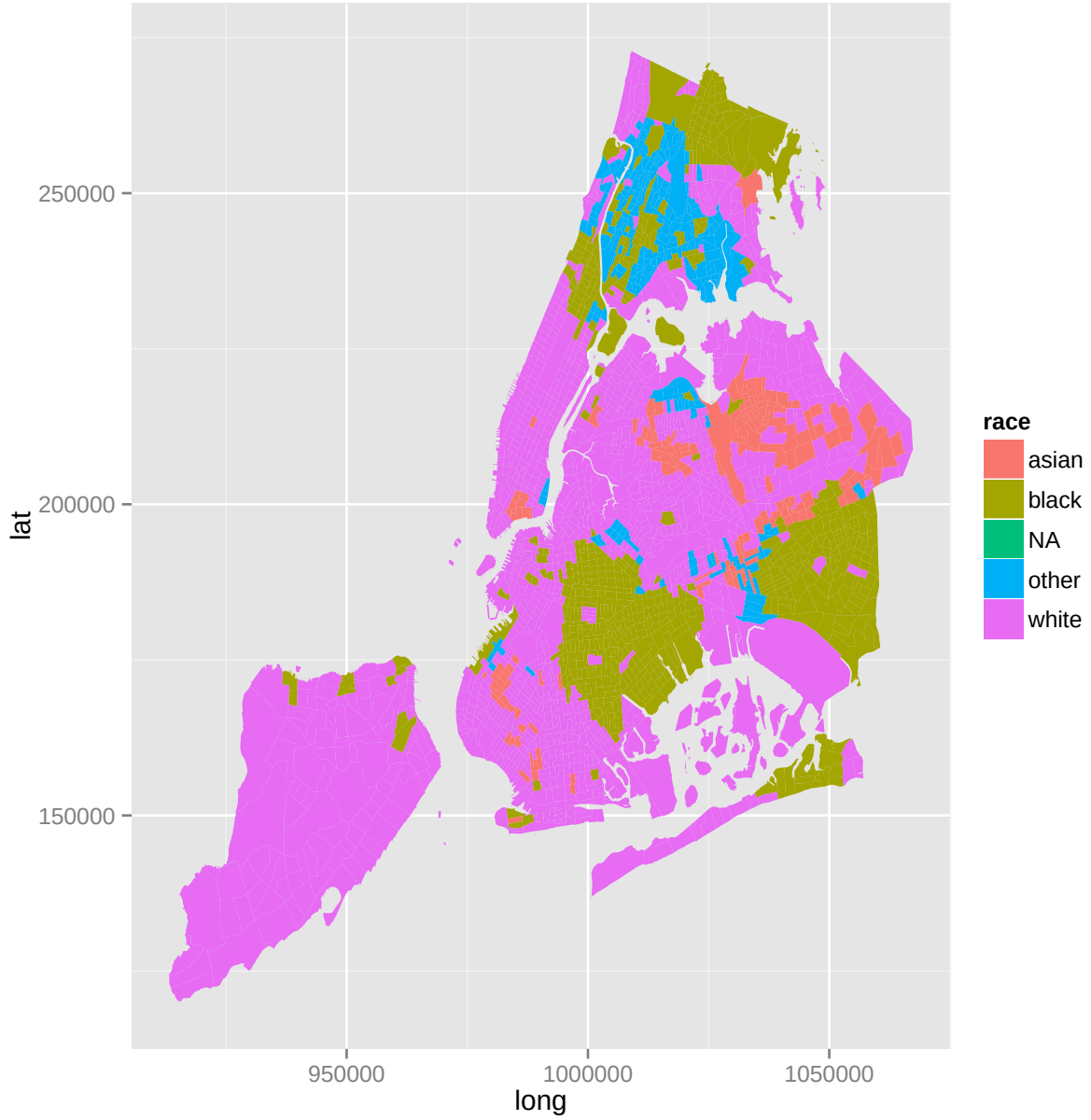
Data from the ACS public use microdata sample is organized to describe individuals. Each row of the dataset describes an anonymized individual and each column represents a different demographic variable. See figure **??** for a small subset of the data used by Morgan et al.

	B27001_001	B27001_002	B27001_003	B27001_004	B27001_005
Alabama	4693822	2256713	186155	176591	9564
Alaska	686905	349855	33053	29026	4027
Arizona	6304406	3099407	279297	249163	30134
Arkansas	2862023	1394466	120828	115053	5775
California	36783532	18138870	1561623	1461559	100064
Colorado	4949633	2457605	210948	193227	17721

By contrast the online API does not provide individual level data. When we perform a get query we must specify what geographic level of detail we want. In table **??** we specify state level summaries for table variable B27001. Table variable B27001 is related to health insurance coverage status for various age groups grouped by gender. Each column in table **??** is a sub-table on B27001.

6 Other Issues

7 Examples from ACS



8 Conclusion