Accessing Data with the Census Bureau API

Alex Shum

February 27, 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).

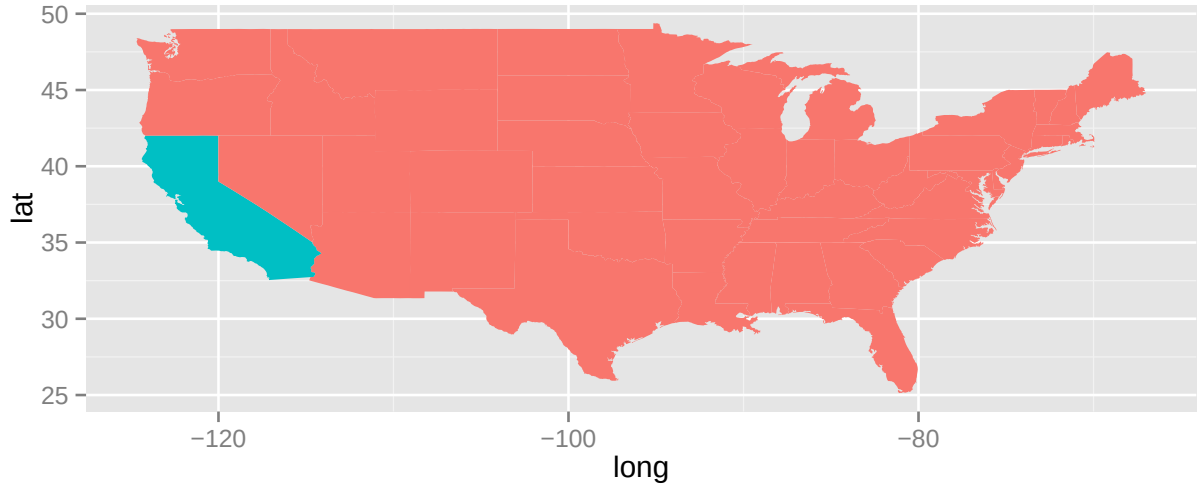
The ACS is an ongoing yearly survey that collects additional demographic variables including but not limited to age, sex, race, income and education. Similar to the decennial census long-form, the American Community Survey is distributed based on a random selection of addresses every year. This data is meant to provide more up to date information than the Census Bureau's decennial census and both the decennial census and the American Community Survey are required by law.

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.

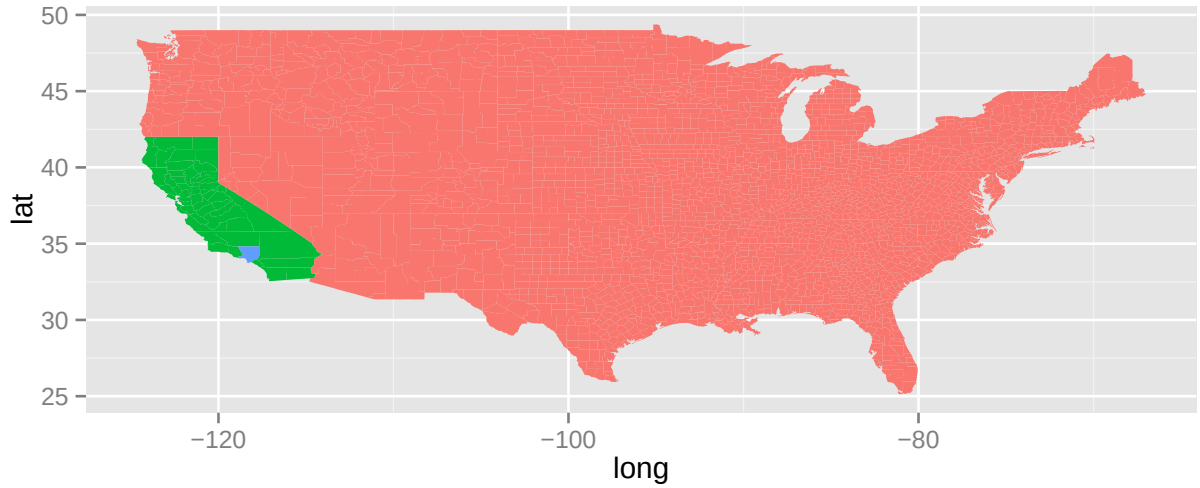
We will discuss how the ACS data is organized and how to access data from the Census Bureau's online developer's API. We will also discuss what kind of variables are available and some limitations with the API. We will use "Taking a Chance in the Classroom: The American Community Survey", an article 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

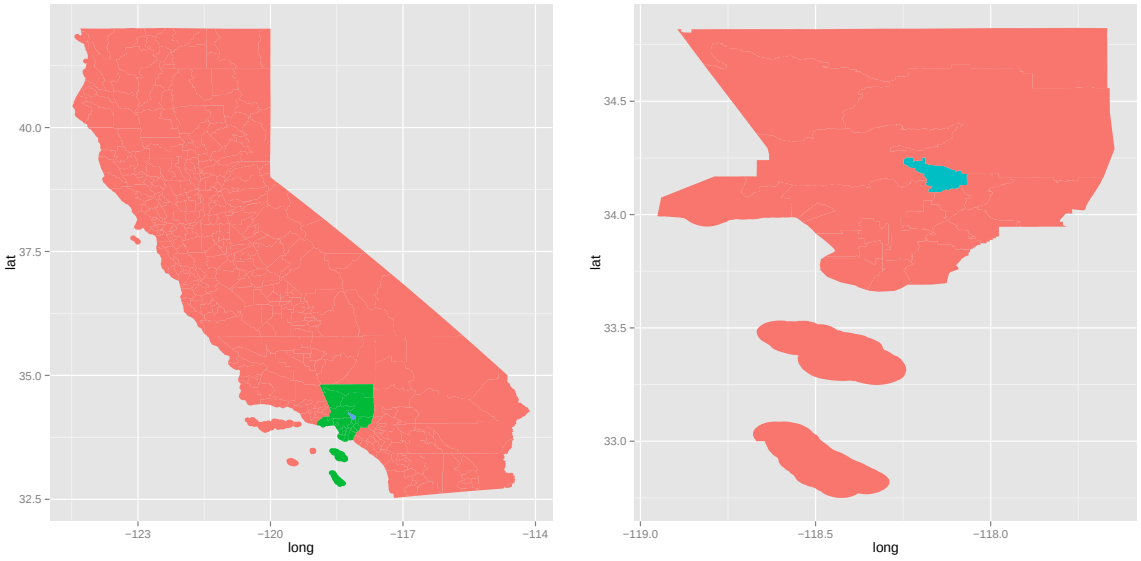
Before we take a look at how the variables in the Census Bureau API are organized, we first look at how the Census Bureau handles the various geographic borders. The Census Bureau has a very sophisticated system of hierarchy for geography. For the ACS, at the top level there is the entire nation, followed by region, division, state, county, county-subdivision, tract, block group, place, congressional district, zip code area, school district and a few other geographic divisions. Outside of the ACS, the Census Bureau employs at least a dozen other geographic concepts.



Near the very top of the hierarchy we can view data from all states or specify a state. In the above plot we have selected California.



Alternatively we can also view data from all counties in all states, we can specify all counties in a specific state or we can select a specific county within a specific state. Above we have selected Los Angeles county within California. Congressional district and ZIP code areas are the other geographic divisions that we can view simultaneously for all states.



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. For many of these geographic divisions it is not possible to pull data without specifying a state or a county. Above we have selected Pasadena, a county subdivision, within Los Angeles County.

Below the county level it is quite complicated which items need to be specified. For example, if we wanted census tracts we need to specify state and county but we cannot specify state, county, census tract and county subdivision. Another complication is 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 districts often do not line up with either legislative distrcits 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. For the 2012 ACS dataset this is located at `http://api.census.gov/data/2012/acs5/geo.html`. Fortunately, the census bureau has also included JSON and XML files that can be parsed to insure that a valid geography is requested.

Each dataset available on the Census Bureau online API include an associated geography file formatted in JSON as follows:

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

Here is the corresponding geographic area formatted in XML:

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

The JSON and XML indicate for this dataset requesting a census tract level information we are required to specify a state. We may also specify a county if we wish to request information for some or all census tracts in a county but if we do not specify a county we can have a similar request for some or all census tracts in a state.

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 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

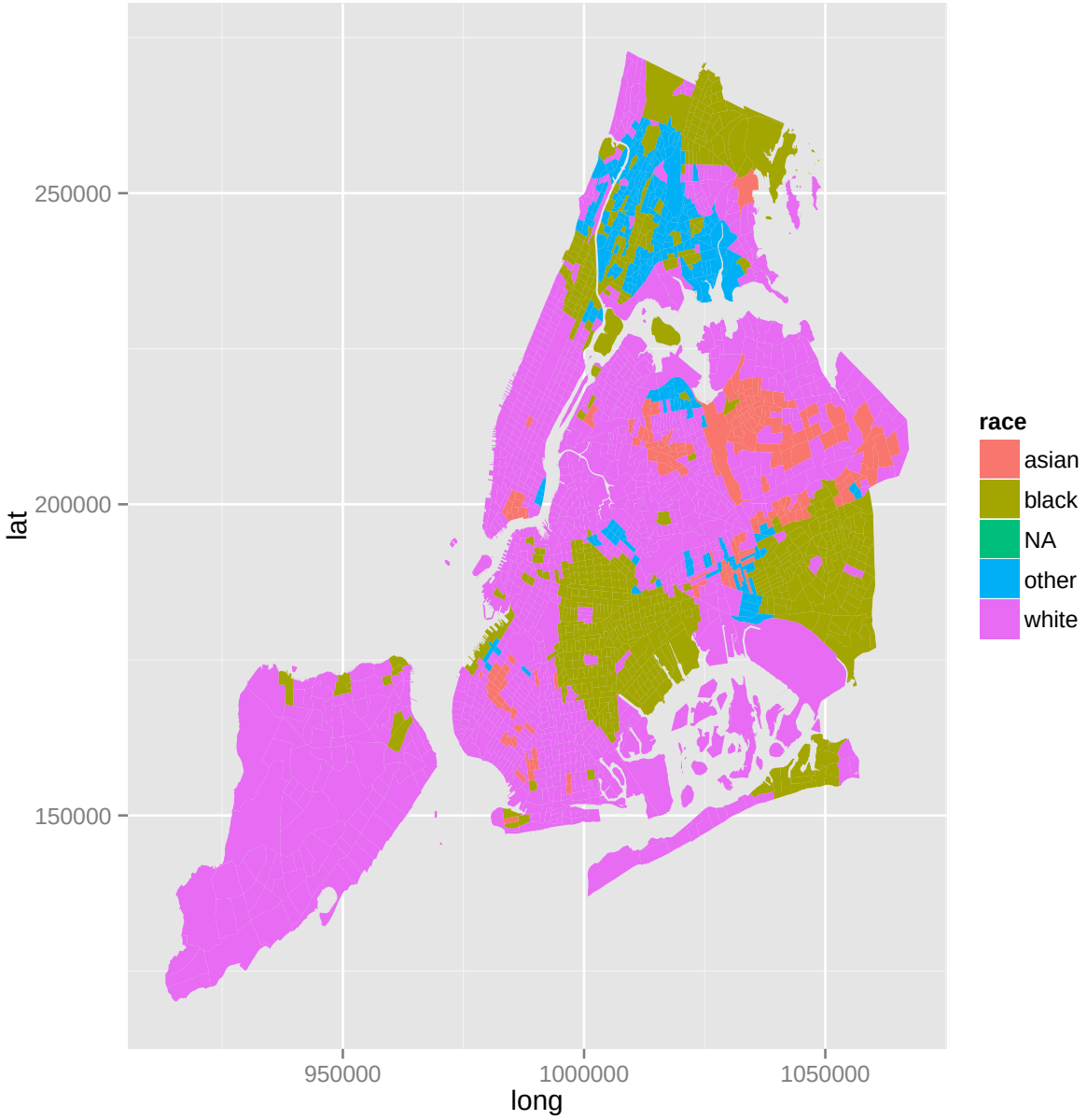
Data from the ACS public use microdata sample is organized to describe individuals. Each row of the dataset describes an anonymized individual and each column represents a different demographic variable. See figure **??** for a small subset of the data used by Morgan et al.

	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 request information we must specify what geographic level of detail we want. In figure **??** we specify state level summaries for table variable B27001.

6 Other Issues

7 Examples from ACS



8 Conclusion