Accessing Data with the Census Bureau API

Alex Shum

March 13, 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 base this discussion on a paper by Stangl, Rundel, and Morgan [1] 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 Requesting Data

To access data from the census bureau online API we need to construct a proper HTTP GET request. A valid GET request is formed through a constructed web URL and we specify which dataset, year, variable and geographies that we are requesting. The basic structure of an HTTP GET request for the decennial census and for the ACS is as follows:

```
http://api.census.gov/data/[YEAR]/[DATASET]?key=[DEVELOPER'S KEY]&get=[VARIABLES]&for=[GEOGRAPHY]
```

[DEVELOPER'S KEY] is an id code required to perform a valid GET request. A developer's key uniquely identifies everyone who requests data from the API. Requesting a key can be done by registering at

`http://www.census.gov/developers/tos/key_request.html`.

[YEAR] and [DATASET] specify the dataset and year of the data requested. The available datasets include the decennial census and the ACS. The ACS datasets are available in 1-year, 3-year and 5-year timeframes. The [YEAR] variable for the ACS datasets indicates the final year in the timeframe. For example, the 2012 5-year ACS dataset this is the ACS dataset that spans 2008-2012 and the 2012 3-year ACS dataset is the ACS dataset that spans 2010-2012. For the decennial census the [YEAR] indicates the year the census data was collected. When we request a [DATASET] we use abbreviations for the dataset we want: the ACS 5-year dataset is *acs*5. See table 1 to see which timeframes are available for each dataset and the abbreviations.

DATASETS	YEAR	abbr
Decennial Census	1990, 2000, 2010	sf1
ACS 5-year	2010, 2011, 2012	acs5
ACS 3-year	2011, 2012	acs3
ACS 1-year	2011, 2012	acs1

Table 1: Datasets and Years

If we wanted to request some data from the 2010 decennial census we would format our HTTP GET request as follows:

`http://api.census.gov/data/2010/sf1?key=[DEVELOPER'S KEY]&get=[VARIABLES]&for=[GEOGRAPHY]`

Similarly requesting data from the 2011 ACS 3-year dataset would require the following HTTP GET request:

`http://api.census.gov/data/2011/acs3?key=[DEVELOPER'S KEY]&get=[VARIABLES]&for=[GEOGRAPHY]`

[GEOGRAPHY] describes the geographic region and [VARIABLES] specifies the demographic variables of interest. Specifying a correct geographic region and search for available demographic variables require a more detailed knowledge of how the census datasets are organized. We discuss geographies in section 3 and variables in section 4. We also discuss how the data is formatted and structured in section 4.

3 Geography

The Census Bureau has a very sophisticated system of hierarchy for geographic entities. 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 2 for a complete table of geographic entities available on the census API for the 2012 ACS.

From table 2, 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 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"
}
```

Summary Level	Description
010	us
020	region
030	division
040	state
050	state-county
060	state-county-county subdivision
140	state-county-tract
150	state-county-tract-block group
160	state-place
250	american indian area/alaska native area/hawaiian home land
310	metropolitan statistical area/micropolitan statistical area
320	state-metropolitan statistical area/micropolitan statistical area
330	combined statistical area
340	state-combined statistical area
350	new england city and town area
400	urban area
500	state-congressional district
510	state-congressional district-county
610	state-state legislative district (upper chamber)
620	state-state legislative district (lower chamber)
795	state-public use microdata area
950	state-school district (elementary)
960	state-school district (secondary)
970	state-school district (unified)

Table 2: List of valid geographic combinations for 2012 ACS 5-year

This JSON file specifies that for census tract level geography we are always required to specify a state. We are required to specify a county in certain cases. For a valid HTTP GET request corresponding to the above JSON information the [GEOGRAPHY] must be formatted as follows:

```
tract:*&in=state:01
tract:*&in=state:01+county:01
```

In the first GET request we are requesting data for all census tracts in Alabama. The second GET request is for all census tracts in Autauga County, Alabama. Due to how census tracts are labelled, if we are interested in a specific census tract we must specify both a state and county:

```
tract:020400&in=state:01+county:01
```

In this GET request we are requesting data for census tract 020400. Note we must specify both state and county because there can be multiple census tracts labelled "020400" among different counties and states.

The same geographic information is also 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>
```

In the following examples, we will examine various geographies available from the online API and their corresponding HTTP GET requests. We will use the United States as a whole and various geographic areas

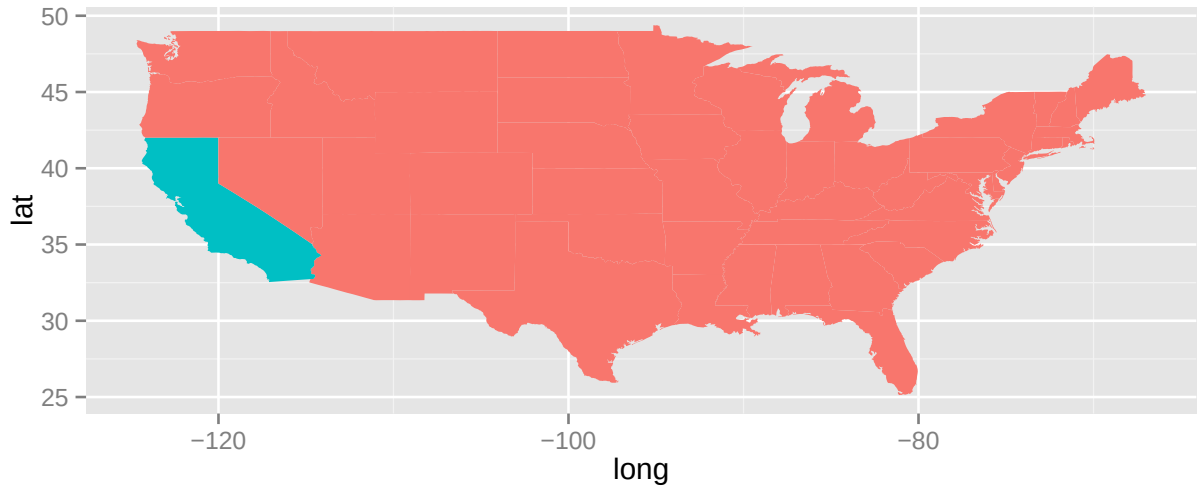


Figure 1: California Selected

within the state of California for our examples.

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 2. 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 with the and this corresponds to a [GEOGRAPHY] of *state:06*.

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. For a corresponding HTTP GET request the [GEOGRAPHY] is *county:037&in=state:06*. 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. The required [GEOGRAPHY] for selecting 90210 is *zip+code+tabulation+area:90210*.

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 2 that state-county-subdivision is a valid summary level.

Although all the entries in table 2 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

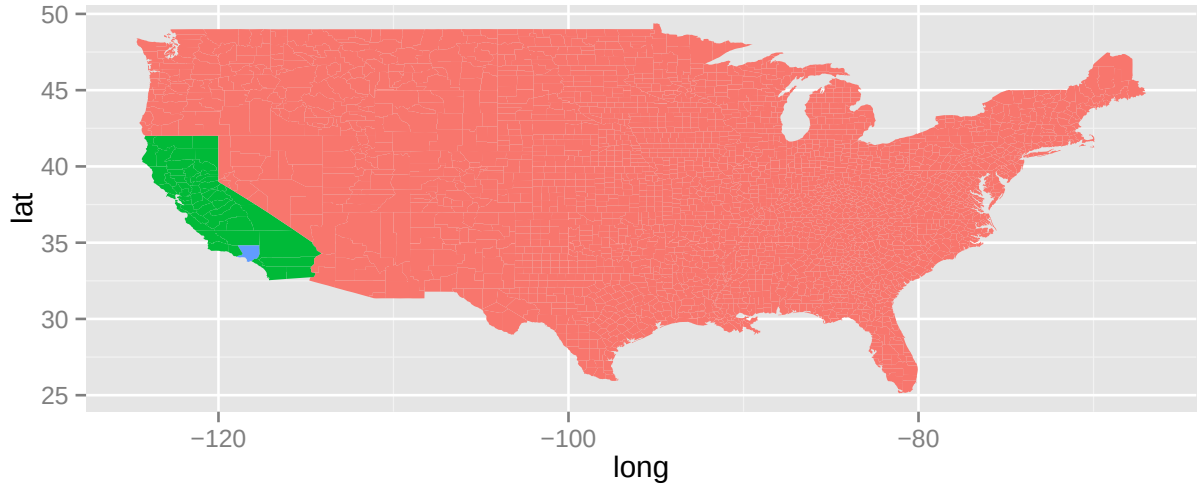


Figure 2: Los Angeles county selected

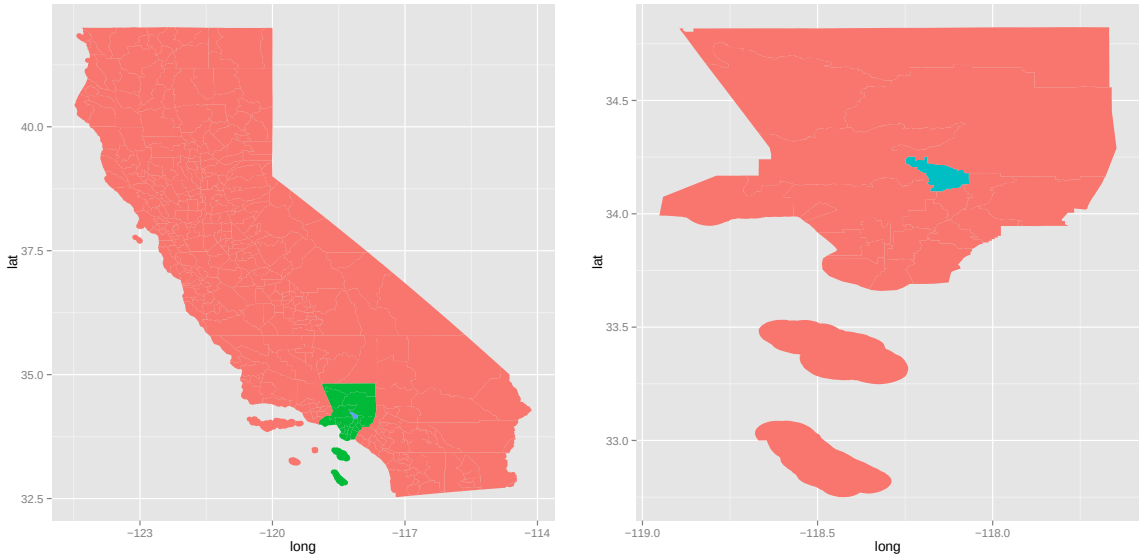


Figure 3: Pasadena county subdivision selected

districts often do not line up with either legislative districts or county borders. ¡EXAMPLE HERE¿

Please note that certain datasets might not have available all geographies and might require a different specification of geography. For example, in the 2010 decennial census if our geographic area is zip code tabulation areas then we are required to specify states. By contrast, the 2012 ACS 5-year dataset has zip code tabulation areas but we do not need to specify states. The 2010 ACS 5-year dataset simply does not have zip code tabulation areas available.

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`.

4 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. The JSON file is available online at `http://api.census.gov/data.json` and the XML file is available online at `http://api.census.gov/data.xml`. This master index includes necessary meta-information about each dataset including description, links to geography and variable information and contact information for maintainer of datasets. The JSON format is formatted as follows:

```
{
  "c_vintage": 2012,
  "c_dataset":[
    "acs5"
  ],
  "c_geographyLink": "http://api.census.gov/data/2012/acs5/geography.json",
  "c_variablesLink": "http://api.census.gov/data/2012/acs5/variables.json",
  "c_tagsLink": "http://api.census.gov/data/2012/acs5/tags.json",
  "c_examplesLink": "http://api.census.gov/data/2012/acs5/examples.json",
  "c_documentationLink": "http://www.census.gov/developers/",
  "c_isAggregate": true,
  "title": "2012 American Community Survey: 5-Year Estimates",
  "webService": "http://api.census.gov/data/2012/acs5",
  "accessLevel": "public",
  "bureauCode": [
    "006:07"
  ],
  "contactPoint": "Census Bureau Call Center",
  "description": "The American Community Survey (ACS) is a nationwide survey...",
  "identifier": "2012acs5",
  "mbox": "pio@census.gov",
  "publisher": "US Census Bureau",
  "references": [
    "http://www.census.gov/developers/"
  ],
  "spatial": "US",
  "temporal": "2012"
  },
}
```

The above excerpt from `http://api.census.gov/data.json` is the meta-information about the 2008-2012 ACS 5-year dataset. For this dataset there are links to the associated geography file and to another JSON

file *variables.json* which is a list of variables available from this dataset. For the above example the same meta-information is formatted in XML as follows:

```
<dataset vintage="2012"
  geographyLink="http://api.census.gov/data/2012/acs5/geography.xml"
  variablesLink="http://api.census.gov/data/2012/acs5/variables.xml"
  tagsLink="http://api.census.gov/data/2012/acs5/tags.xml"
  examplesLink="http://api.census.gov/data/2012/acs5/examples.xml"
  documentationLink="http://www.census.gov/developers/"
  pod:webService="http://api.census.gov/data/2012/acs5" isAggregate="true"
  pod:accessLevel="public"
  dcat:contactPoint="Census Bureau Call Center"
  dct:identifier="2012acs5"
  pod:mbox="pio@census.gov"
  dct:publisher="US Census Bureau"
  dct:spatial="US"
  dct:temporal="2012">
<dataset-name> <part name="acs5"/> </dataset-name>
<dct:title>2012 American Community Survey: 5-Year Estimates</dct:title>
<pod:bureauCode> 006:07 </pod:bureauCode>
<dct:description>
The American Community Survey (ACS) is a nationwide survey...
</dct:description>
<pod:reference link="http://www.census.gov/developers/"/>
</dataset>
```

After the user has selected the dataset of interest, they will need to lookup what variables are available for that dataset. The JSON and XML master index of datasets contain the location of *variables.json* and *variables.xml* respectively which list available variables available for that dataset. Available variables are organized into tables referred to as "concepts"; a "concept" is a combination of factors. For example, "Health Insurance Coverage Status by Sex by Age" is a concept from the 2012 ACS. Although health insurance coverage, sex and age are all different factors, this "concept" contains information that links these three factors together.

Within each table there are sub-tables that provide information on different levels of each of the factors. For example, under the "Health Insurance Coverage Status by Sex by Age" concept there is a table for males under 6 with health insurance. For this concept there are tables for each combination of levels for sex (male, female), age group (under 6, 6 to 17, 18 to 24, 25 to 34, 35 to 44, 45 to 54, 55 to 64, 65 to 74 and 75 and over) and health insurance coverage (with and without health insurance). Additionally there are a number of total columns: total number of males, total number of females and totals for males/females in each age group. For more details on how the data is organized within this table structure see the next section.

To lookup what variables are available, the *variables.json* and *variables.xml* files contain a list of all sub-tables along with a description. The JSON formatted *variables.json* is formatted as follows:

```
"B27001_056E": {
  "label": "Female:!!75 years and over:!!With health insurance coverage",
  "concept": "B27001.  Health Insurance Coverage Status by Sex by Age"
},
```

This describes sub-table 056E of concept B27001. This is a table of 75 and older females with health insurance. The associated XML formatted version is formatted as follows:

```
<var xml:id="B27001_056E"
  label="Female:!!75 years and over:!!With health insurance coverage"
  concept="B27001. Health Insurance Coverage Status by Sex by Age"/>
```

5 Limitations

The dataset used by Stangl, Rundel, and Morgan [1] 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

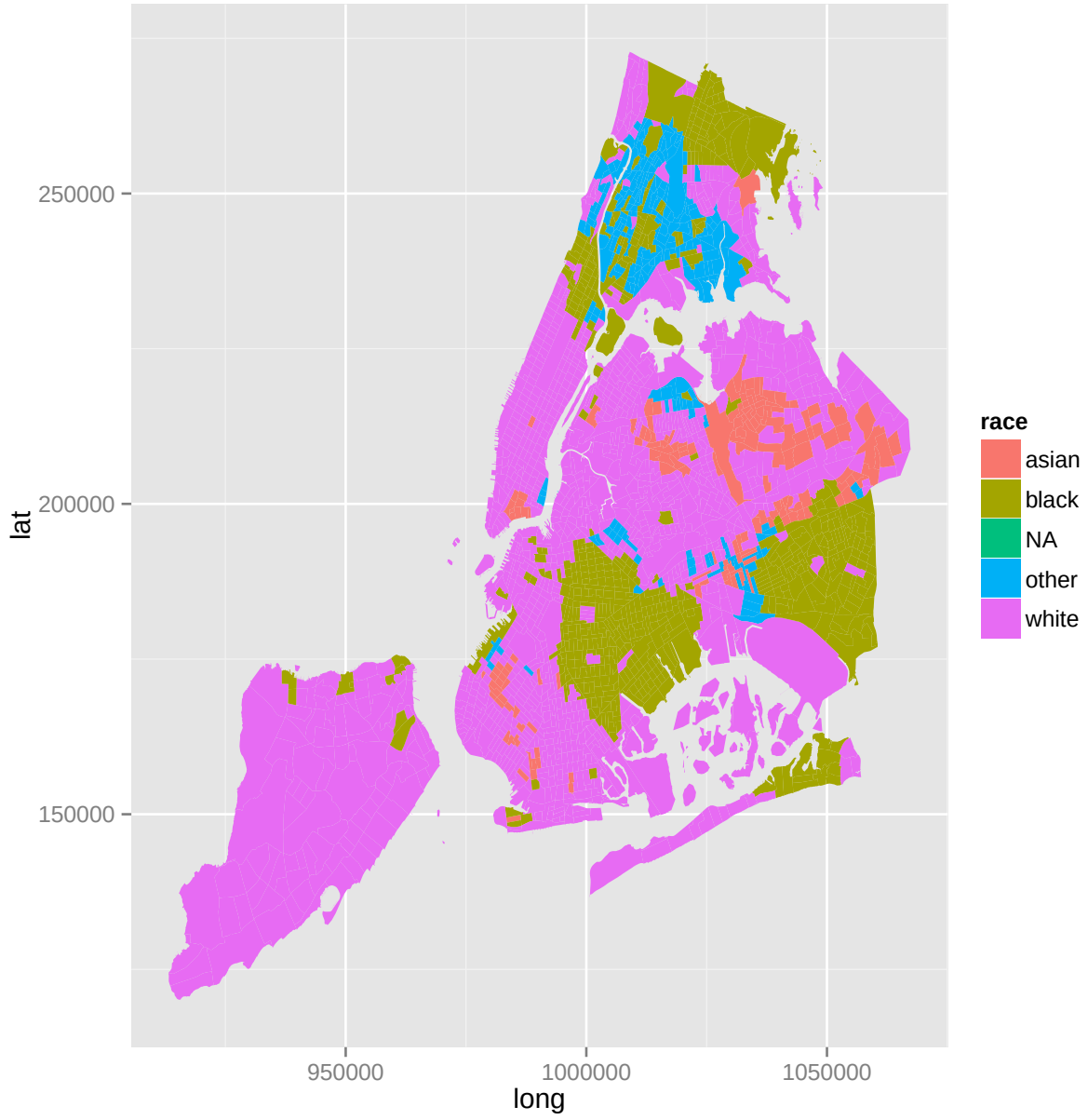
Data from the ACS public use microdata sample is organized to describe individuals. Each row of the dataset describes an anonymized individual and each column represents a different demographic variable. See figure **??** for a small subset of the data used by Morgan et al.

	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 HTTP GET request 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

References

[1] Dalene Stangl, Mine Çetinkaya Rundel, and Kari Lock Morgan. "Taking a Chance in the Classroom: The American Community Survey". In: *CHANCE* 26.1 (2013), pp. 42–46. DOI: 10.1080/09332480.2013.772392. eprint: http://www.tandfonline.com/doi/pdf/10.1080/09332480.2013.772392. URL: http://www.tandfonline.com/doi/abs/10.1080/09332480.2013.772392.