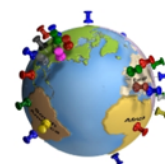


Geocoding

DENISE MALAN, *IRE & INN* | LIZ LUCAS, *IRE & NICAR*



Geocoding

Make that address data work for you.

<https://geoservices.tamu.edu/>

Introduction to geocoding

Most of us have received data at one point or another that includes addresses. Less often we get fields labeled "latitude" and "longitude."

In order to put addresses on a map, any mapping software needs those lat and long coordinates. The process a computer uses to calculate the coordinates for an address so it can place it on a map is known as geocoding.

As reporters we often deal with hundreds or thousands of addresses at once. Thankfully there are geocoding services out there that will geocode hundreds or thousands of addresses very quickly, and plenty will do it for free.

Getting your data ready

For this exercise we'll use a slice of the Small Business Administration's (SBA) 7a loan data: all loans for San Francisco in 2013. The file name is **sba_sf.txt** (download it [here](https://github.com/eklucas/NICAR-Help/raw/master/sba_sf.txt)¹). Before importing data into any program, you should take a look to see how the data are formatted, what geographic fields you want to map, etc. Open up **sba_sf.txt** in a text editor, preferably Notepad++ or Sublime Text, something more powerful than Notepad.

```
Sublime Text  File  Edit  Selection  Find  View  Goto  Tools  Project  Window  Help
sba_sf.txt — data

1 | BorrName  BorrStreet  BorrCity  BorrState  BorrZip  BankName  GrossApproval  ApprovalDate
2 | Martin Castillo & Manuel A. Ca  5800 Third Street #1003 & 100  San Francisco  CA  94124  Redwood CU  635000.00  2013-01-04
3 | FitPFSF, Inc.  340 Sansome Street  San Francisco  CA  94104  Umpqua Bank 1000000.00 2013-01-08
4 | HWA RANG KWAN MARTIAL ARTS CEN 371 5TH ST  SAN FRANCISCO  CA  94107  U.S. Bank National Association 30000.00 2013-01-09
5 | NEILL & LEE GENERAL CONTRACTOR 84 PROSPECT AVE SAN FRANCISCO  CA  94110  Wells Fargo Bank, National Ass 50000.00 2013-01-11
6 | VIP Grooming  4299 24th Street  SAN FRANCISCO  CA  94114  OBDC Small Business Finance 100000.00 2013-01-22
7 | Versus Games LLC  1716 Taraval Street SAN FRANCISCO  CA  94116  Sterling Savings Bank d.b.a Ar 50000.00 2013-01-30
8 | Kent M. Lim Company, Inc.  1260 Egbert Avenue San Francisco  CA  94124  Bank of the Orient 920000.00 2013-01-31
9 | 17th Street Oliveira Chiroprac 3705 17th Street  SAN FRANCISCO  CA  94114  Celtic Bank Corporation 92600.00 2013-02-12
10 | RESTORATION MEMORIES  921 VALLEJO ST  SAN FRANCISCO  CA  94133  Wells Fargo Bank, National Ass 5000.00 2013-02-16
11 | Home Care Options, San Francis 1 Daniel Burnham Court suite SAN FRANCISCO  CA  94109  American National Bank 829300.00 2013-02-1
12 | SBhimani LLC  908 Sutter Street  SAN FRANCISCO  CA  94109  BBN Bank 500000.00 2013-02-19
13 | Lilitab LLC 2339 3rd St. Ste 59 SAN FRANCISCO  CA  94107  OBDC Small Business Finance 250000.00 2013-02-21
14 | Chiwi LLC  4119-4123 24th. Street  SAN FRANCISCO  CA  94114  Bank of San Francisco 1400000.00 2013-02-22
15 | Y Studios, LLC  151 Vermont St. #10 SAN FRANCISCO  CA  94103  U.S. Bank National Association 1000000.00 2013-02-25
16 | JNB FINANCE LLC 1547 CLAY ST APT 301  SAN FRANCISCO  CA  94109  Wells Fargo Bank, National Ass 25000.00 2013-02-27
17 | Amir Akbari and Maryam Emami  4612 - 4614 3rd Street  SAN FRANCISCO  CA  94124  Wells Fargo Bank, National Ass 441100.00 2013-03-0
18 | Growing Tree Child Care 1984 Great Hwy  SAN FRANCISCO  CA  94116  Wells Fargo Bank, National Ass 767000.00 2013-03-01
19 | Five Flavors Herbs  350 SANSOME ST Suite 730  SAN FRANCISCO  CA  94104  U.S. Bank National Association 10000.00 2013-03-06
20 | ENA 323 Grant Ave  SAN FRANCISCO  CA  94108  JPMorgan Chase Bank, National 90900.00 2013-03-12
```

¹ https://github.com/eklucas/NICAR-Help/raw/master/sba_sf.txt

Here our data happens to be formatted just the way TAMU wants it: BorrStreet (borrower street address), BorrCity, BorrState and BorrZip in separate fields. Your data must have this info split into separate fields.

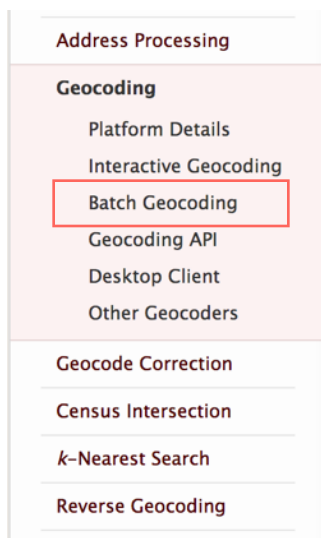
Note also that our file is “tab-delimited”, which means that each value in a row is separated by a tab. There are also no “text qualifiers”, which are usually single or double quotes that surround text fields.

Many times you will get data where the addresses have typos, misspellings, or are missing the directional prefix (such as N. for North) or the suffix (such as St. for Street). In these cases it is best to try and do some cleaning before you geocode. If you put garbage in, you get garbage out.

It’s also often important to have a zip code; many geocoders will not work or give you bad results without them. If you didn’t get zip codes, consider whether it’s feasible to put them in.

Using Texas A&M GeoServices

[Signing up for an account](#) with Texas A&M is simple and will take you only a few minutes. Once you get an account, you have 2500 free credits to use; 1 credit = 1 address geocoded.



If you want to geocode a bunch of addresses all at once, you want to **Batch Geocode**. A “batch” is just a bunch of commands or jobs strung together, in this case geocodes (like a batch of cookies).

Once you’re signed in, click to Geocoding icon (see above) and from the options in the left sidebar, choose “Batch Geocoding.”

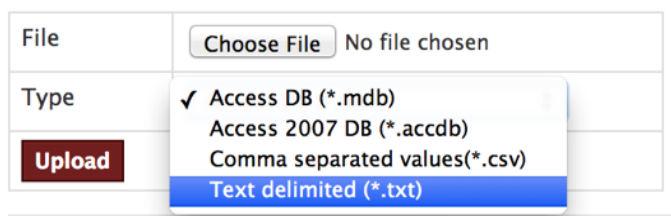
This page outlines the steps we’ll go through to geocode our data:

In a nutshell, to upload and geocode a database you will need to perform four steps:

- **Step 1** – Upload your data files and validate that we can open and read them
- **Step 2** – Choose which data in your files you want to process
- **Step 3** – Identify the fields of your data so we know which column is what
- **Step 4** – Choose your processing options and start the process

Start – Step 1 >>

Click the Start button. On the next page, click “Add New Database,” and then “Upload New Database.” TAMU allows you to upload either an Access database (.mdb or .accdb) or a delimited text file (.csv or .txt). We’ll choose “Text delimited” and then click the “Choose File” button.



Navigate to your desired file, sba_sf.txt. You are familiar with how your text file is structured, and here is where we tell TAMU the specifics. In this case, our file is tab-delimited, there is no text qualifier, and there are headings in the first row:

File	<input type="button" value="Choose File"/> sba_sf.txt
Type	Text delimited (*.txt) ▾
Columns	<input checked="" type="checkbox"/> First row contains column headings
Text Separator	tab ▾
Text Qualifier	none ▾
<input type="button" value="Upload"/>	

Click “Upload.” Once the data is loaded, you will be asked to “Validate Database.” Since we uploaded a flat file there will be only one table, so click “Validate Table.” Here we can see the first 10 rows of our data, and TAMU gives us a record count so that we can be sure everything was uploaded properly. Our record count, 116, is accurate, and the data snapshot looks good. Scroll below the table until you see this:

You may now use your database in the following services:

Address Processing

Services for processing postal addresses including address parsing, normalization, standardization, and validation

Geocoding

Services for turning postal addresses into geographic coordinates including parsed, non-parsed, and batch postal address database geocoding

Click “Geocoding.” This takes us to Step 2, which is just choosing the data we want to geocode. We only have one table uploaded, so go ahead to Step 3.

Here we identify data fields that are pertinent to the geographic information. Match up the borrower information (the Id field was automatically generated when we uploaded):


Input Fields

AddressData	
Id	AUTO_UNIQUE_ID_2014-06-20_eklucas_sba_sf ▾
StreetAddress	BorrStreet ▾
City	BorrCity ▾
State	BorrState ▾
Zip	BorrZip ▾

Leave the rest of the defaults and go to Step 4, leave the defaults and click “Start Process.”
Now stand up and stretch! TAMU is doing all the hard work.

You will get an email about the start and finish of the process, but you can also check the status by clicking “View Process Status.”

Since we’re only geocoding 116 addresses, this should go pretty quickly. When it’s done, we’ll see this:

Start	Service	Database	completed / total	status
6/20/2014 2:56:40 PM	Geocoding	sba_sf.txt sba_sf	116 / 116 	Completed

Checking your results

So, that was the easy part. TAMU has done most of the work up until this point, but we still need to evaluate the results. The tricky thing about geocoding is that it is far from an exact science, and almost all geocoders (if not all) make some errors. We need to do a little fact-checking.

Thankfully TAMU provides some helpful pointers on its geocodes that can guide us. First, in the left sidebar, go to “Databases,” then “Current Databases.”

Click on the link to “download” the data:

name	type	date	size	status	do not auto delete	actions
eklucas sba_sf.txt	Text File	6/20/2014 2:40:56 PM	13.18 KB	ready	<input type="checkbox"/>	download delete share

Save it as **sba_sf_geocoded.txt** so we can tell it apart from our original file.

TAMU adds 27 fields, listed below. We’re mostly concerned with the ones highlighted:

AUTO_UNIQUE_ID_2014-06-20_eklucas_sba_sf
Source
TimeTaken
UpdatedGeocoding
Version
ErrorMessage
TransactionId
naaccrQualCode
naaccrQualType
FeatureMatchingResultType
MatchedLocationType
RegionSizeUnits
InterpolationType
RegionSize
InterpolationSubType

FeatureMatchingGeographyType
MatchScore
FeatureMatchingHierarchy
TieHandlingStrategyType
FeatureMatchingResultTypeTieBreakingNotes
GeocodeQualityType
FeatureMatchingHierarchyNotes
FeatureMatchingResultTypeNotes
FeatureMatchingResultCount
Latitude
Longitude
MatchType

“**FeatureMatchingGeographyType**” is the type of geography that TAMU used to place the address. We want our addresses to be placed using “Street Segments” or “Parcels” because they are fairly exact. If something like “USPS Zip” is used, the lat/long will place the address in the *center* of the zip code, and that is probably not where the address actually is. Sometimes the geography type used can be a county, a state, or even the United States. These geocodes should be thrown out without question. In some cases if you clean the data a bit you can get more accurate geocodes.

Many times, if a geocoding service can't find an address, it will return the lat and long of the center of the geographic area you're working in (usually city or state). Another common error is placing the point at (0,0).

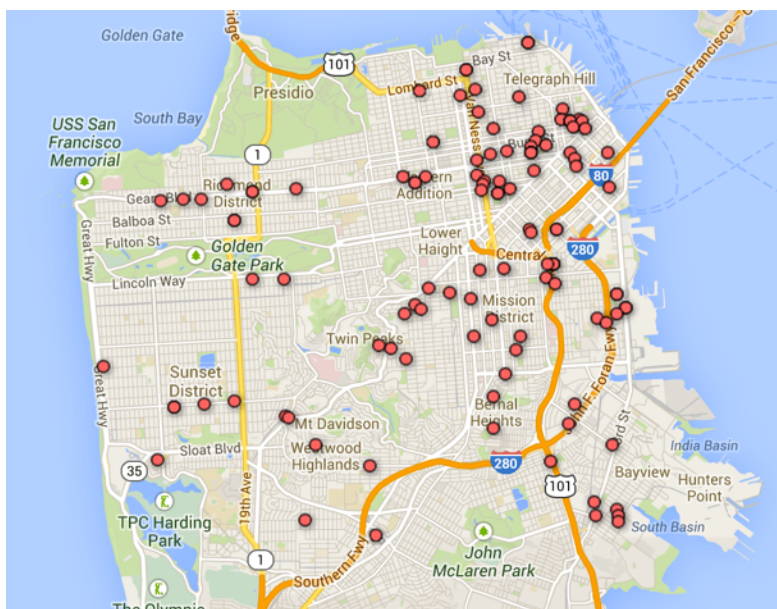
A good check is to sort your results by latitude then longitude to check for duplicates all placed at the same point or at (0,0). If you find some, those addresses might need more work for the geocoding service to place them correctly.

The “**Match Score**” field gives us a score from 0 to 100, where 100 is a perfect match (in the eyes of the program). Using match scores below 100 is a judgment call, but be careful with lower scores. Note, however, that a geocode based on a zip code can still get a score of 100.

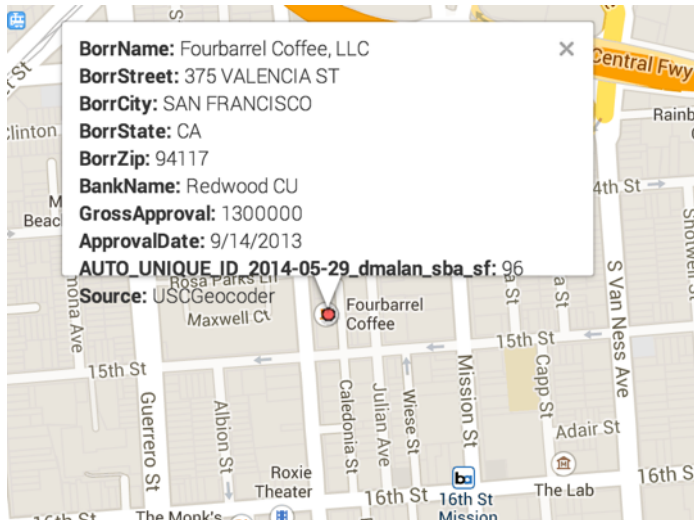
You also will want to inspect your results visually. You can do this with any mapping software, but Google Fusion Tables is a good free option (see below).

Any good geocoder will give you the type of geography an address was match to and the match score; these are important indicators of the accuracy of the latitude and longitude.

Checking your results visually can help you identify any addresses that were placed way outside the boundaries you expected. For example, we would expect all these addresses to be plotted within the city of San Francisco:



This is the result of importing our data into Google Fusion tables. Here we can visually inspect our results to make sure that there aren't stray points out in the water or in a different state or continent, and we can also click on any of the points to check the data:



When geocoding goes wrong

Doing these integrity checks are very important, so that you don't miss something like this:

<http://www.vox.com/2014/4/21/5636040/whats-the-matter-with-kansas-and-porn>

VOX did an analysis of internet porn traffic by getting IP address data from Pornhub.com, one of the largest internet porn sites, and geocoding the IP addresses. As their title suggests, they found that Kansas was the home to a crazy amount of porn traffic, and said so. In fact, it wasn't just Kansas, it was one city in particular: Wichita.

After publishing their results, it came out that Wichita was so off the charts because it is in the center of the country. When the geocoder couldn't place the IP addresses exactly, it used the whole country as its matching geography type and plopped them right in the center of that geography, right onto Wichita. Vox changed the headline and published a correction at the top of the story.

For great commentary on all the things that were not great about this analysis, go [here](#).

Final words

When you get data that contains addresses and you're tempted to jump straight to the geocoder, ask yourself a few questions first:

1. Why do I want to see these addresses on a map? Will it help me understand the scope of the data? Are geographic clusters potentially important? Could these locations have a spatial relationship (close together, far apart, clustered, segregated, etc) that would be meaningful?
2. What does the data look like? Are the addresses neat, containing all the necessary pieces: prefix (such as N. for North), suffix (such as St. for Street), standardized names (without a lot of misspellings or typos), a zip code? If its missing one or more of these things, how hard will it be to clean the data up?
3. How important is the accuracy to the zip code? If you just want to look at your data points zoomed out at the county or state level, you can accommodate more relaxed geocoding; if a point is plotted a block or two away from where it should be, looking at the state level that is not necessarily a big deal. But if you want to really zoom in and see points plotted at the street level, accuracy is more important.

Remember that geocoding is not an exact science and no geocoders will be perfect; diligently check results and look at all the information available to you to assess how well your addresses were geocoded.

A note about projections

It would be wrong to write a whole tipsheet on mapping-related things and not talk about projections. If you are geocoding a dataset with TAMU or another like service and just plotting the points on a Google map, you don't need to worry too much about projections. But if you want to join your points with other geographic information of any kind, projections will become very important.

A projection is simply the way cartographers have put a round world on a flat display (either paper or your computer screen). Think of a round orange peel that you squash flat on the table: there will be gaps in the peel that need to be accounted for, and distortion is inevitable. There are many different ways to deal with that, and hence many different projections: Albers, Mercator and Peters, to name a few of the most popular. Mapping software such as Esri's ArcGIS will allow you to choose a projection for your map and then convert geographic data to that projection. If you intend to do that, you'll need more than this tipsheet. =)