RGeocoding

Patty Frontiera

December 4, 2015

Geocoding in R

Getting Started: Download the zipfile for this tutorial from `https://github.com/dlab-geo/RGeocoding/archive/master.zip`

Overview

- What is Geocoding
- A simple example in Google Maps
- Why Geocode
- Geocoding in Detail
- How to Geocode in R
  - with GGMAPS
  - with Yahoo Placefinder
  - with TIGER
- Now what

What is Geocoding

Determine the geographic coordinates of a named place, street address, or zip code.

- city, building,
- street address, intersection,
- mountain, landmark,
- crime or other event location,
- zip code, etc.

Try It!

maps.google.com

Geographic Coordinates

Latitude	+/- 90 degrees	*how far north or south of equator*
Longitude	+/- 180 degrees	*how far E/W of prime meridian*

Decimal Degrees (DD)
37.870145, -122.25952

Degrees, minutes, seconds (DMS)
37° 52' 12"N, 122° 15' 36" W

Why Geocode?

- Display locations on a map
- Link locations to other data
- Spatial analysis
  - Calculate distance, direction, area, etc.
  - Identify patterns & relationships:
    - clusters, outliers, neighbors

We will cover the first two

Address Geocoding

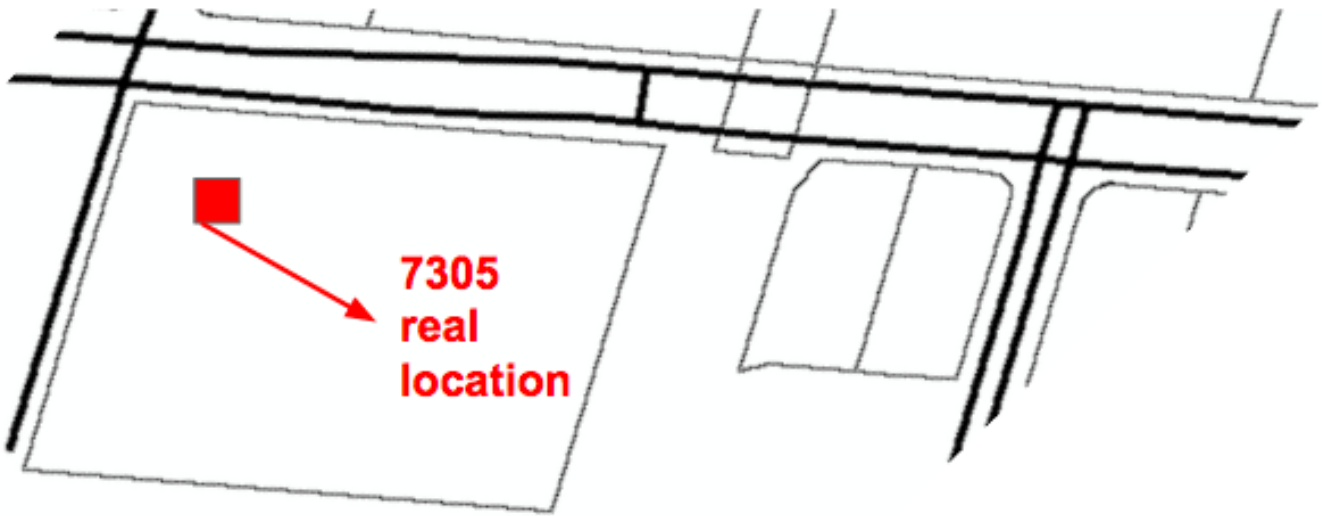
Where do Addresses come from?

- Extract from text documents
- File or Database
- Web Scraping

Process Details

Evaluation

Reference database extremeley important



Geocoder Output Comparison

Considerations

- Geographic scope
- Time period
- Output Quality
- Number of addresses
- Data Privacy/security
- Cost
- ease of use

Local Geocoding

ESRI ArcGIS with local reference database

- Benefits
  - highly customizable
  - fast & robust software
  - easy to use GUI
- Issues
  - free reference database circa 2009
  - limited to USA & Canada

Remote Geocoding Services

- ArcGIS Online
- Google
- Yahoo
- OpenStreetMaps
- Data Science Toolkit (DSTK)
- *and many others*

Geocoding in R

Access an online Geocoder using an API *Application Progromming Toolkit*

- Via a package or script

Geocoding in R with

- GGMAPS
  - Google
  - DSTK
- RYDN & Yahoo Placefinder
- US Census TIGER Geocoding Service

Geocoding with GGMAP

- Created by David Kahle and Hadley Wickham, ggplot2 developer
- Includes functions for Geocoding using:
  - the Data Science Toolkit (DSTK) geocoding service
  - Google's Geocoding service
- Also has functionality for fetching map data from Google and other online services
  - you can use these with your data to create custom maps

Geocoding with GGMAP

- The Data Science Toolkit (DSTK) geocoding service
  - default, unlimited usage
  - FOSS: free and open source softare (& data)
  - good, not great output quality
  - older data, limited geographic coverage
  - *service sometimes unavailable*

Geocoding with GGMAP

- Google's Geocoding service
  - fantastic accuracy, easy, fast,
  - worldwide coverage, up to date
  - limited to 2500 addresses per day
  - other limits may also apply!

Geocoding with GGMAP

Go ahead and stick that in maps.google.com - must be in *lat,lon* order, comma separated!

- Then try using GGMAP to Geocode
  - an address
  - a zipcode

Be sure to specify source="google"

?geocode

Try these Geocode Options for *output=*

- geocode("Barrows Hall, Berkeley, CA", source="google", output="latlon")
- geocode("Barrows Hall, Berkeley, CA", source="google", output="latlona")
- geocode("Barrows Hall, Berkeley, CA", source="google", output="more")
- geocode("Barrows Hall, Berkeley, CA", source="google", output="all")

Output differences

Output differences

Checking Output Quality

Append geocoded info to input data

Create a data frame with three addresses

Geocode the three Addresses

Join output to input

Map it with GGMAP

Try different (or no) zoom levels!

Geocode a file of addresses

We need one column with address (not multiple)

We need one column with address (not multiple)

Irregularity is a Problem

Now geocode that address again.

How to spot problems like that?

Save it!

Know Your limits

Scaling up to more than 2500 records?

- The downloaded data for this tutorial contains an R script showing how to geocoded within the google limits
- scripts/google_geocode_limits.R

Geocoding Output

- With a little preprocessing most reliable geocoders will be able to geocode 80% or more of your addresses within a block of the actual location.
- based on my my experience!
- assumes US addresses.
- Cleaning and standardizing addresses is a lot of work!
  - unlikely to get it perfect
  - extremely important

Standardize Addresses

- provide all components
- remove unnecessary components
- remove duplicates
- remove extra spaces or commas
- remove odd characters like “#” “/”, “@”
- standardize capitalization

Standardize Addresse

- Intersections
  - Corner of Main and Long Ave should be **Main & Long**
- Numbered Streets
  - Fourth St should be **4th St**
- Directional Prefixes
  - North, No, N., etc should be **N**
- Apartment numbers and letters
  - Remove them!
- Use **PO Box**
  - unless you have the address!

Use Standard Abbreviations

Use	For These
HWY	Highway
LN	Lane
DR	Drive
EXPY	Expressway

Problemns in Reference Database

- Incorrect street ranges
- Inaccurate or low quality features
- Inaccurate feature attributes
- Missing streets
- Address changes

Output Quality & Population Density

Address Format Differs by Geocoder!

Google

Census Tiger

Be mindful of commas!

Assessing Output Quality

- map the results
- examine the range of coordinates
  - CA: -124, 32, -114, 42
- review output metadata
- specific to the geocoder

Google Limits

https://developers.google.com/maps/documentation/geocoding/usage-limits

Geocoding with Yahoo Placefinder in R

- RYDN Package
- devtools::install_github("trestletech/rydn")
- 2000 addresses per day limit!
- You need to apply for a YDN API Key
  - Yahoo Developer Network
- See script for example usage
  - **scripts/yahoo_geocoding.R**

US Census TIGER Geocoding API in R

- No limit
- Addresses only
- Returns the Census FIPS code for each geocoded address
-
  - See script for example usage
  - **scripts/tiger_geocoding.R**

Linking to Census Data

U.S. Census Bureau Census Block 15 character FIPS Codes

FCC FIPS API

-
- The first two characters (06) indicate the state (CA),
- the next three (085) indicate the county (Alameda),
- the next 6 indicate the census tract (5046.01)
- and the last four characters indicates the census block group and block number (1175).
  - *The first digit of the block identifies the block group.*

Linking to Census Data

Another method

See this script: **scripts/getFipsForPoints.R**

Next Steps

- Consider hybrid approaches
- Make a D-Lab consulting appointment if you are using RUD
- Take a look at the References

References

- https://cran.r-project.org/web/packages/ggmap/index.html
- https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf
- https://www.nceas.ucsb.edu/~frazier/RSpatialGuides/ggmap/ggmapCheatsheet.pdf
- https://developers.google.com/maps/documentation/geocoding/intro
- http://www.albany.edu/faculty/ttalbot/Geocoding_Lecture_2015.pdf

References

- http://rstudio-pubs-static.s3.amazonaws.com/90665_de25062951e540e7b732f21de53001f0.html
- https://github.com/walkerke/tigris
- http://zevross.com/blog/2015/10/14/manipulating-and-mapping-us-census-data-in-r-using-the
- http://www2.census.gov/geo/tiger
  - Then go to: http://www.census.gov/geo/maps-data/data/tiger-line.html (read how do i choose...)
- http://dlab.berkeley.edu/blog/address-geocoding-options-uc-berkeley-community