Geospatial Data Acquisition and Evaluation

Center for Geographic Analysis

Harvard University

Nicole Alexander, Ph. D.

nalexander@cga.harvard.edu

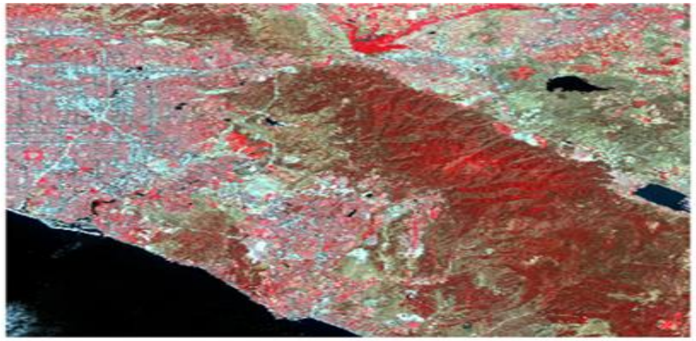
Harvard DataFest 2018

Geospatial Data Acquisition and Evaluation

- Geospatial Data Sources
- Data Transfer
- Metadata
- Geocoding Data Science Datasets in R using Google’s API

Geospatial Data Acquisition and Evaluation

- How data are captured determines the quality of decisions that can be made from analyzing the data
  - *Primary* sources: obtained through direct measurement
  - *Secondary* sources: derived from other sources
- Data accuracy can more reliably be determined from primary sources

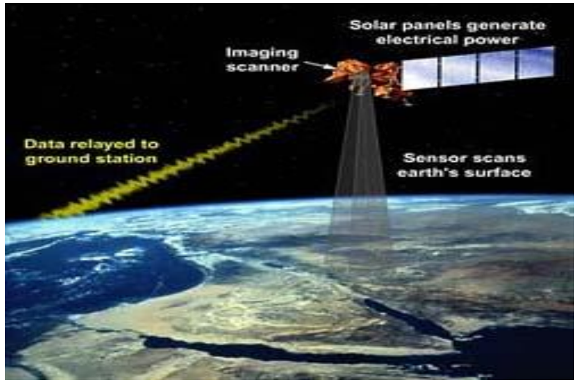


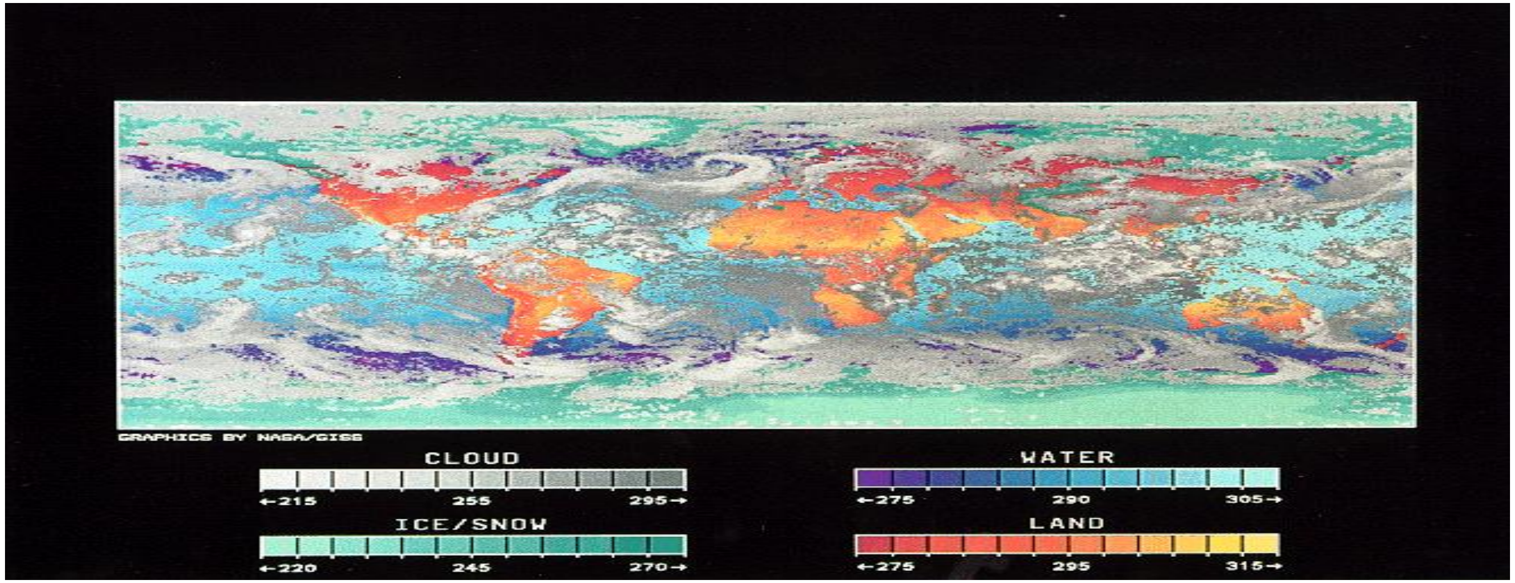
Geospatial Data Sources

	RASTER	VECTOR
Primary	• Digital satellite remote-sensing images • Digital aerial photographs	• GPS measurements • Field survey measurements • LiDAR
Secondary	• Scanned maps and photographs • Digital elevation models from topographic map contours	• Topographic maps • Toponymy (place-name) databases • Geocoding

Primary Raster Data Capture

- Remote sensing
  - Satellite
  - Aircraft
- Image Resolution
  - Spatial
  - Spectral
  - Temporal


Solar panels generate electrical power
Imaging scanner
Data relayed to ground station
Sensor scans earth's surface

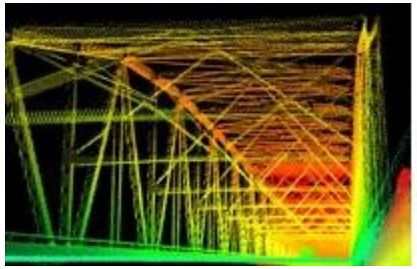


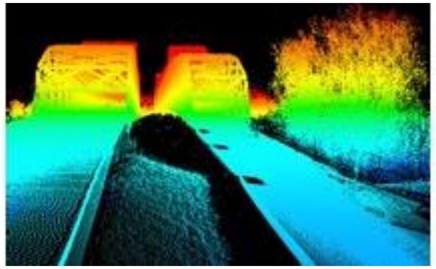
Remote Sensing Data Capture

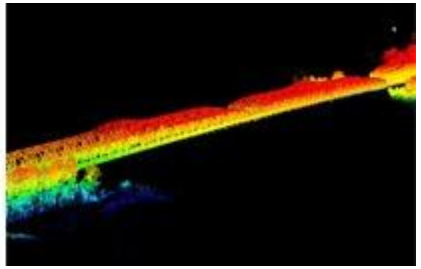
- Captures data over a large geographic areas
    - Total ground coverage range from 9 x 9 km – 200 x 200 km
- Pixel size determines spatial resolution of an image
    - Spatial accuracy of features increases as pixel size decreases
- Satellite systems capture data in the range of 0.5 m – 1 km pixel size
- Camera systems capture data in the range of 0.01 m – 5 m pixel size
- Costly compared with other methods of data capture
- Data volumes can be very large

Primary Vector Data Capture

- Main sources
    - GPS
    - Surveying
- Remote Sensing
    - LiDAR (Light Detection And Ranging)
        - a “cloud” of points that reflects the surface







Primary Vector Data Capture

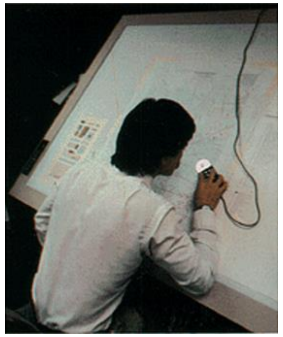
- GPS
  - Recreational: low precision 6 – 12m
  - Mapping and GIS: medium precision 30cm – 5m
  - Surveying: high precision 5mm – 1cm
- Surveying
  - Used for large scale mapping of small areas and property boundaries
  - Capable of 1 mm accuracy
  - Equipment and crews are expensive
- LiDAR
  - 30,000 points per second at an accuracy of around 15cm
  - Often rasterized to create DEMs

Secondary Raster Data Capture

- Scanning of hardcopy media
  - Building plans, CAD drawings, property deeds, film, paper maps, aerial photographs, images, etc.
  - Spatial resolution of scanners in the range of 400 – 900 dpi (16 – 40 dots per mm)
- DEM generation from topographic map contours or LiDAR

Secondary Vector Data Capture

- Vectorizing raster data
- Digitizing
- Geocoding
- Photogrammetry
- COGO – Coordinate Geometry

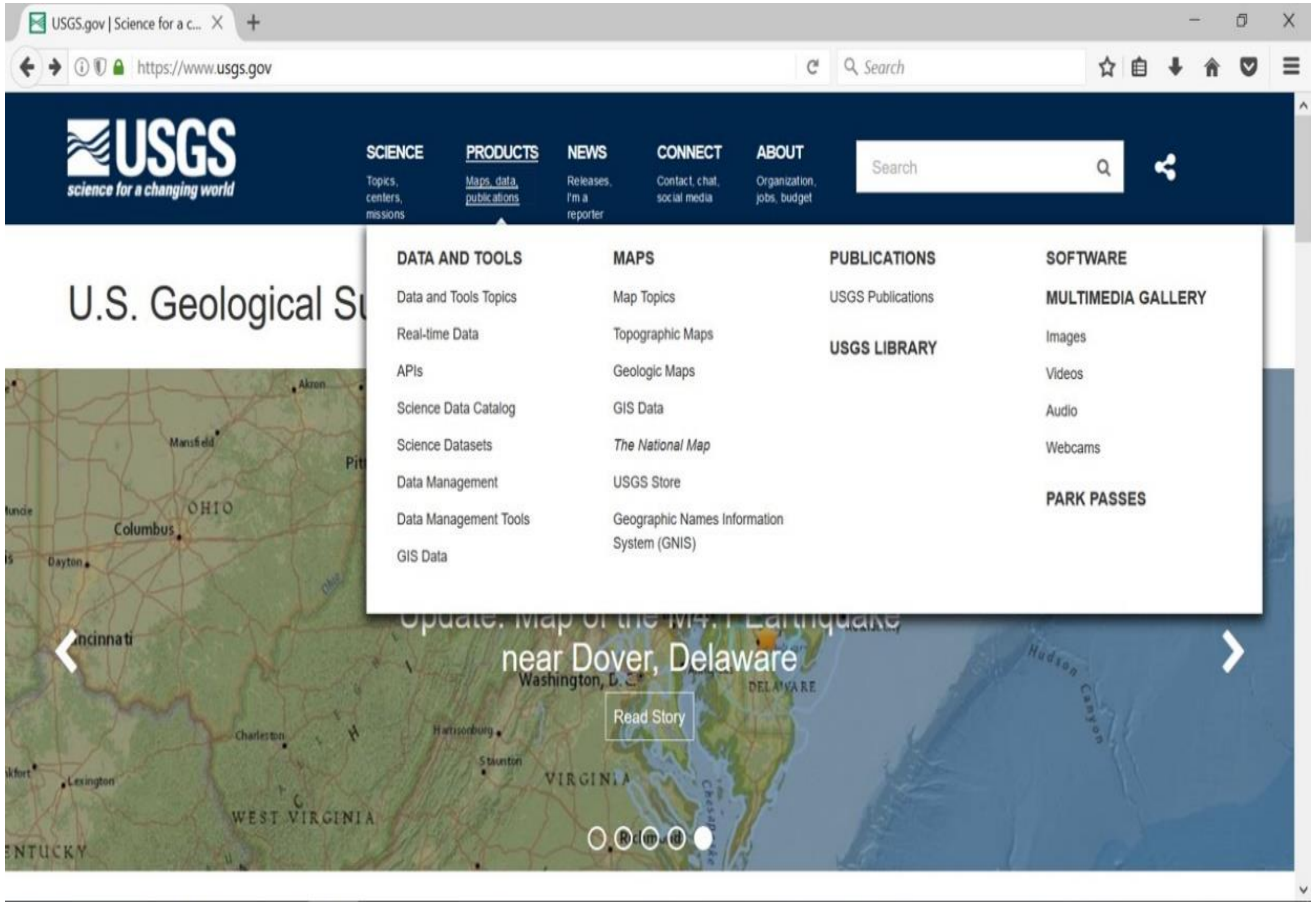


Data Transfer: Obtaining Data from External Sources

- U.S. Geological Survey
- U.S. Census Bureau
- OpenStreetMap
- GeoNames
- Other Geospatial Data Sites

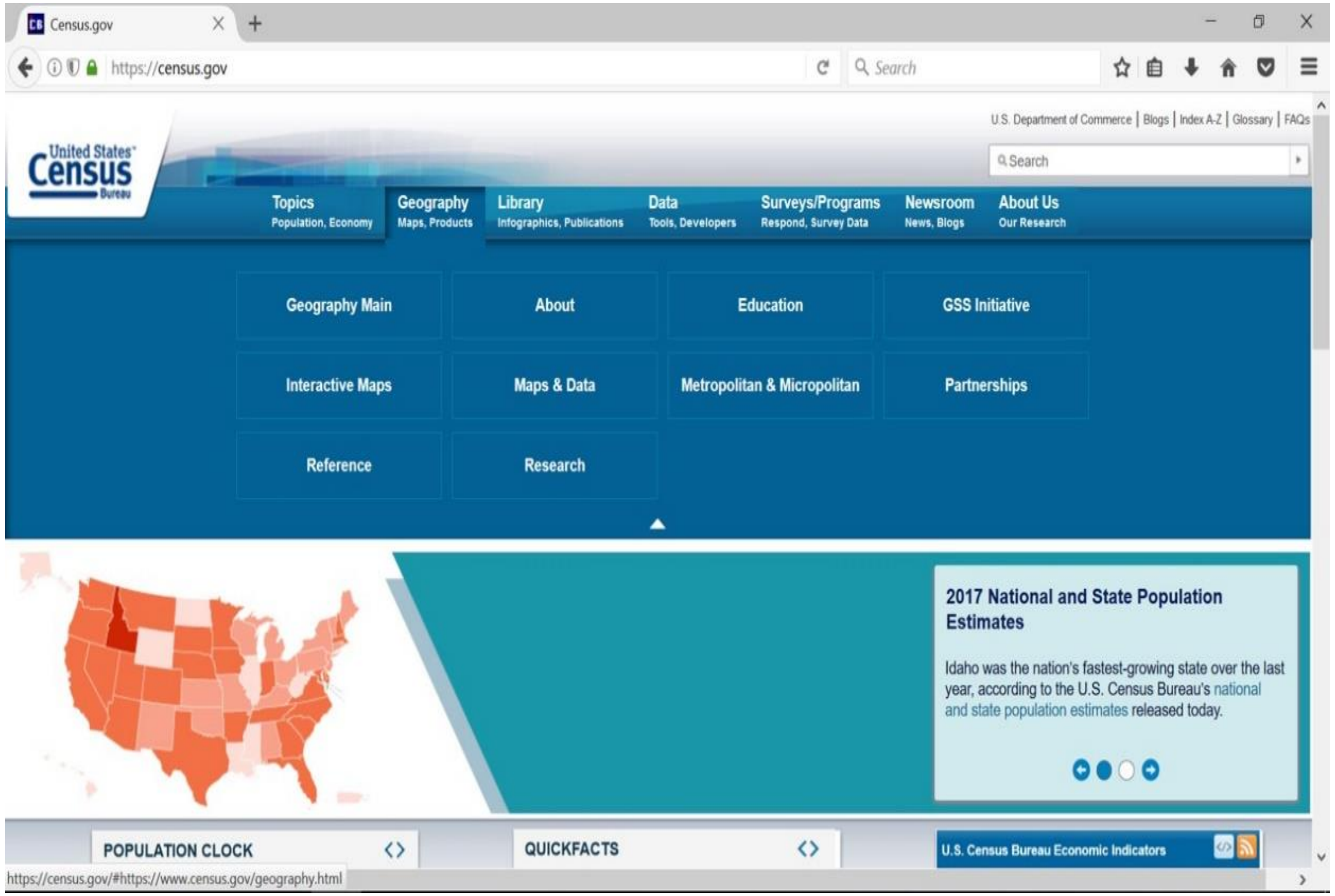
U.S. Geological Survey (usgs.gov)

- The major provider of geospatial data in the US



U.S. Census Bureau (census.gov)

- Provides data to support the US decennial census

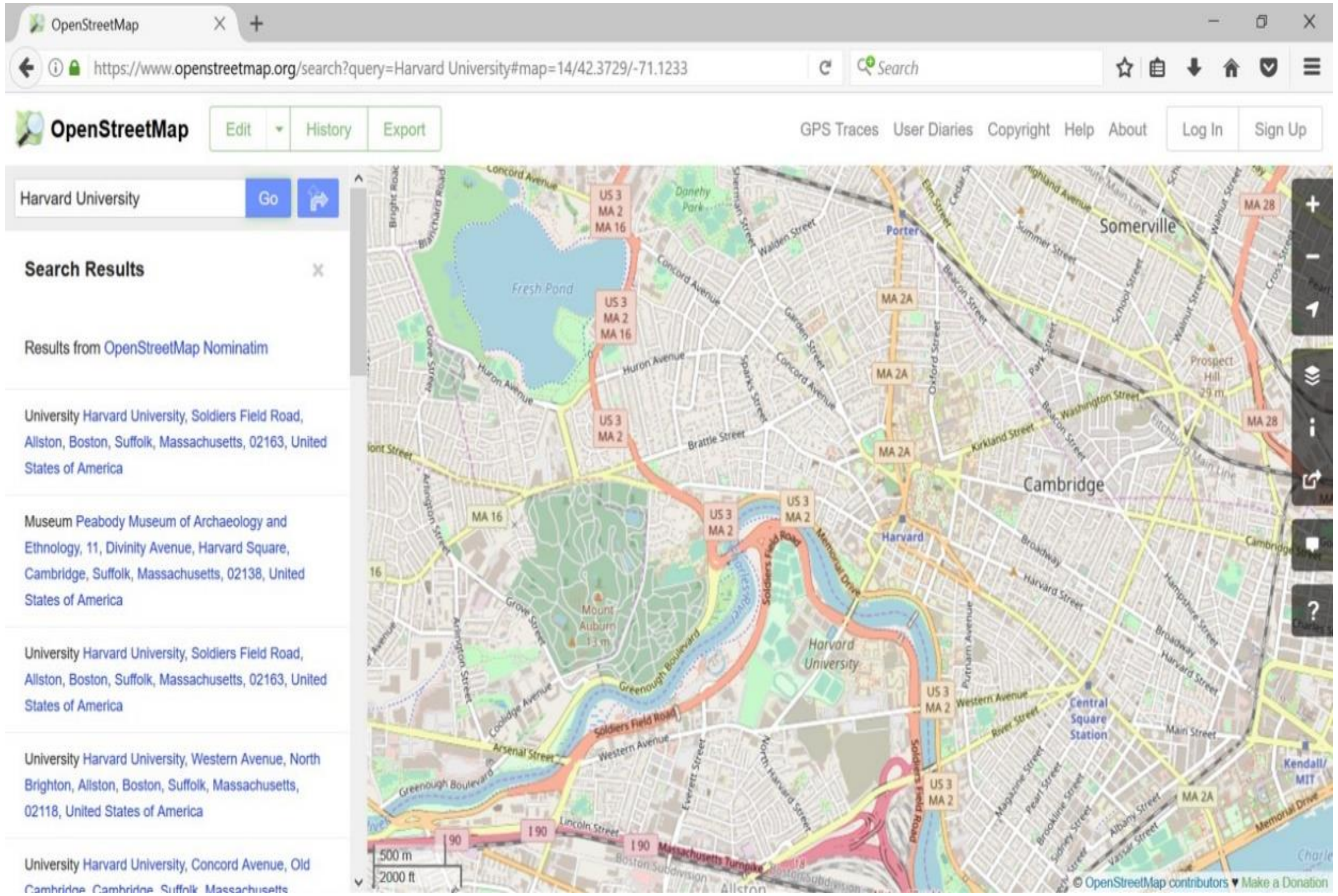


2017 National and State Population Estimates

Idaho was the nation's fastest-growing state over the last year, according to the U.S. Census Bureau's national and state population estimates released today.

OpenStreetMap (OSM) (openstreetmap.org)

- Map data of the world
- Created and maintained by a community of mappers



https://www.openstreetmap.org/search?query=Harvard University#map=14/42.3729/-71.1233

Harvard University

Search Results

Results from OpenStreetMap Nominatim

University Harvard University, Soldiers Field Road, Allston, Boston, Suffolk, Massachusetts, 02163, United States of America

Museum Peabody Museum of Archaeology and Ethnology, 11, Divinity Avenue, Harvard Square, Cambridge, Suffolk, Massachusetts, 02138, United States of America

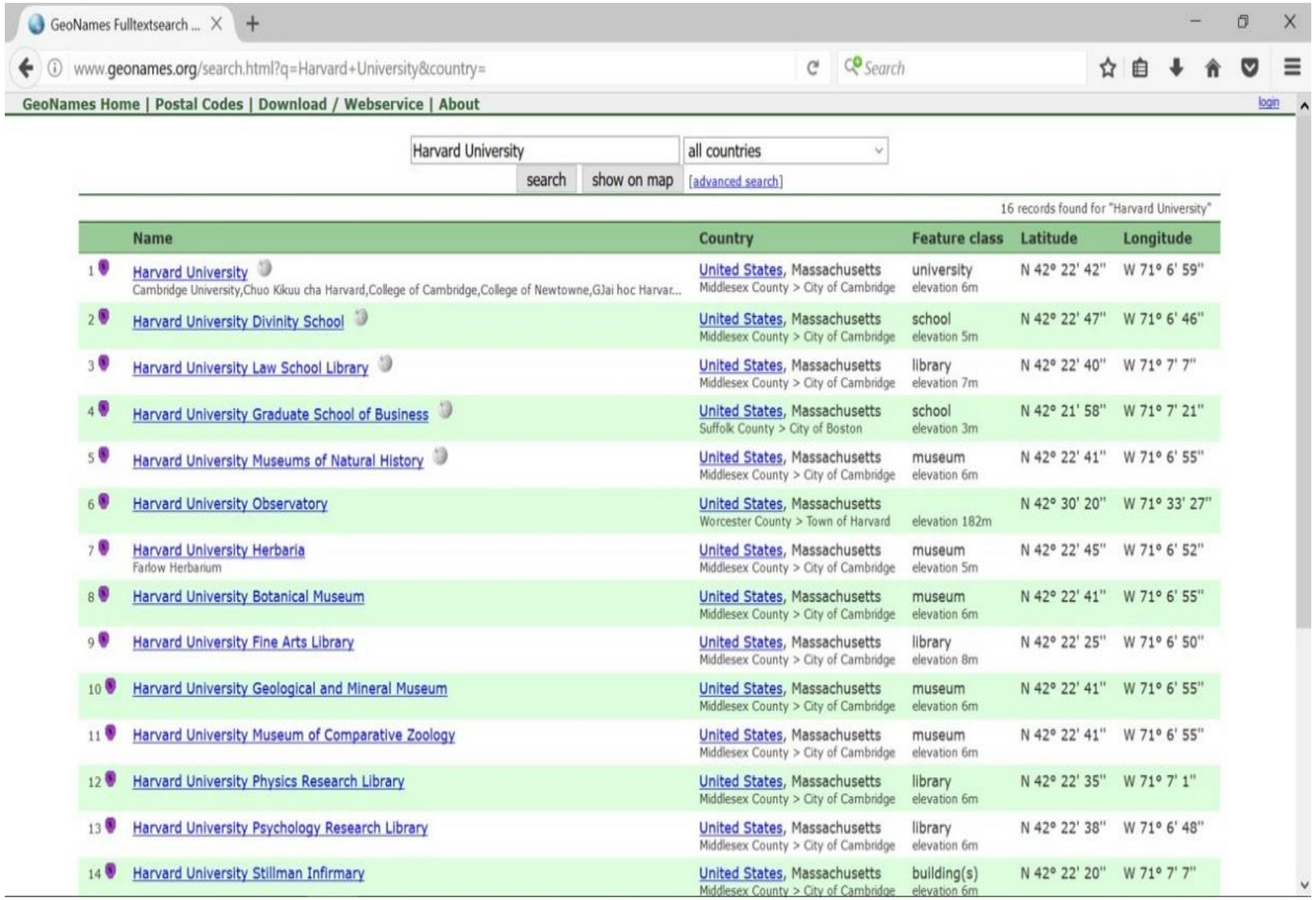
University Harvard University, Soldiers Field Road, Allston, Boston, Suffolk, Massachusetts, 02163, United States of America

University Harvard University, Western Avenue, North Brighton, Allston, Boston, Suffolk, Massachusetts, 02118, United States of America

University Harvard University, Concord Avenue, Old Cambridge, Cambridge, Suffolk, Massachusetts

GeoNames (geonames.org)

- Global geographical database of place names


GeoNames Fulltextsearch ...
Search
GeoNames Home | Postal Codes | Download / Webservice | About
login
search
show on map
[advanced search]

www.geonames.org/search.html?q=Harvard+University&country=

Harvard University

all countries

16 records found for "Harvard University"

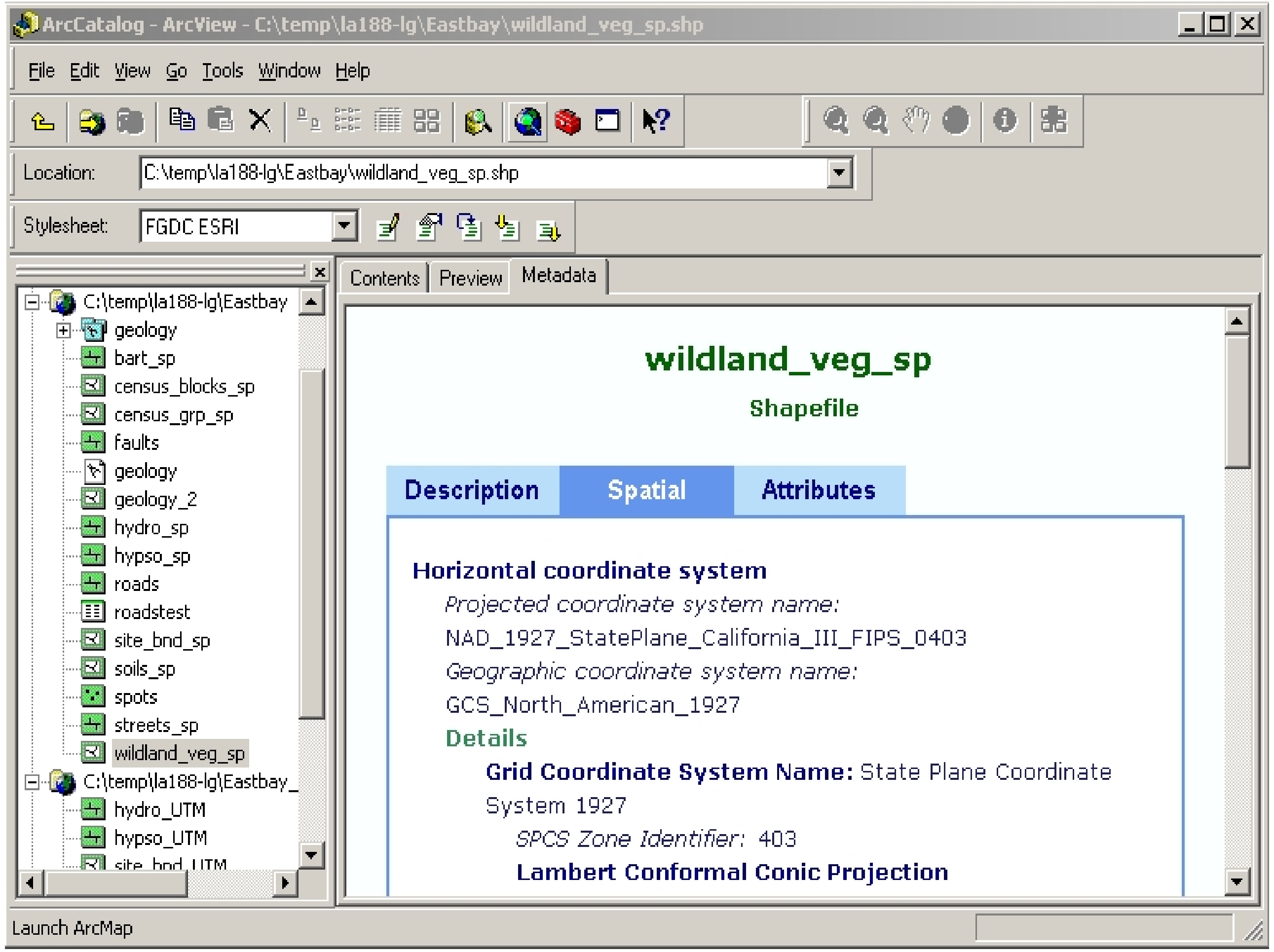
	Name	Country	Feature class	Latitude	Longitude
1	Harvard University Cambridge University,Chuo Kikuu cha Harvard,College of Cambridge,College of Newtowne,GJai hoc Harvar...	United States, Massachusetts Middlesex County > City of Cambridge	university elevation 6m	N 42° 22' 42"	W 71° 6' 59"
2	Harvard University Divinity School	United States, Massachusetts Middlesex County > City of Cambridge	school elevation 5m	N 42° 22' 47"	W 71° 6' 46"
3	Harvard University Law School Library	United States, Massachusetts Middlesex County > City of Cambridge	library elevation 7m	N 42° 22' 40"	W 71° 7' 7"
4	Harvard University Graduate School of Business	United States, Massachusetts Suffolk County > City of Boston	school elevation 3m	N 42° 21' 58"	W 71° 7' 21"
5	Harvard University Museums of Natural History	United States, Massachusetts Middlesex County > City of Cambridge	museum elevation 6m	N 42° 22' 41"	W 71° 6' 55"
6	Harvard University Observatory	United States, Massachusetts Worcester County > Town of Harvard	elevation 182m	N 42° 30' 20"	W 71° 33' 27"
7	Harvard University Herbaria Farlow Herbarium	United States, Massachusetts Middlesex County > City of Cambridge	museum elevation 5m	N 42° 22' 45"	W 71° 6' 52"
8	Harvard University Botanical Museum	United States, Massachusetts Middlesex County > City of Cambridge	museum elevation 6m	N 42° 22' 41"	W 71° 6' 55"
9	Harvard University Fine Arts Library	United States, Massachusetts Middlesex County > City of Cambridge	library elevation 8m	N 42° 22' 25"	W 71° 6' 50"
10	Harvard University Geological and Mineral Museum	United States, Massachusetts Middlesex County > City of Cambridge	museum elevation 6m	N 42° 22' 41"	W 71° 6' 55"
11	Harvard University Museum of Comparative Zoology	United States, Massachusetts Middlesex County > City of Cambridge	museum elevation 6m	N 42° 22' 41"	W 71° 6' 55"
12	Harvard University Physics Research Library	United States, Massachusetts Middlesex County > City of Cambridge	library elevation 6m	N 42° 22' 35"	W 71° 7' 1"
13	Harvard University Psychology Research Library	United States, Massachusetts Middlesex County > City of Cambridge	library elevation 6m	N 42° 22' 38"	W 71° 6' 48"
14	Harvard University Stillman Infirmary	United States, Massachusetts Middlesex County > City of Cambridge	building(s) elevation 6m	N 42° 22' 20"	W 71° 7' 7"

Other Geospatial Data Sites

- Harvard University
  - CGA: http://gis.harvard.edu/resources/data
  - Harvard Geospatial Library: http://hgl.harvard.edu
  - Harvard WorldMap: http://worldmap.harvard.edu
  - Harvard Map Collection: http://hcl.harvard.edu/libraries/maps/collections/digital.html#overview
- Local
  - MassGIS: http://www.mass.gov/mgis
  - City of Boston: https://data.boston.gov/dataset?groups=geospatial
  - Metro Boston Data Common: http://www.metrobostondatacommon.org/
- National
  - US Federal Government: http://data.gov
  - US Geological Survey: http://viewer.nationalmap.gov/viewer/
- Global
  - The ESRI Data and Maps: http://bit.ly/NBoQzQ
  - ArcGIS Online Services from ESRI: http://www.arcgis.com/home/

Metadata

- Data about the geospatial data:
  - Identification
  - Data quality
  - Coordinate system
  - Attributes, etc.
- Especially important when using public data



Location: C:\temp\la188-lg\Eastbay\wildland_veg_sp.shp

Stylesheet: FGDC ESRI

wildland_veg_sp

Shapefile

Horizontal coordinate system

Projected coordinate system name:
NAD_1927_StatePlane_California_III_FIPS_0403

Geographic coordinate system name:
GCS_North_American_1927

Details

Grid Coordinate System Name: State Plane Coordinate System 1927

SPCS Zone Identifier: 403

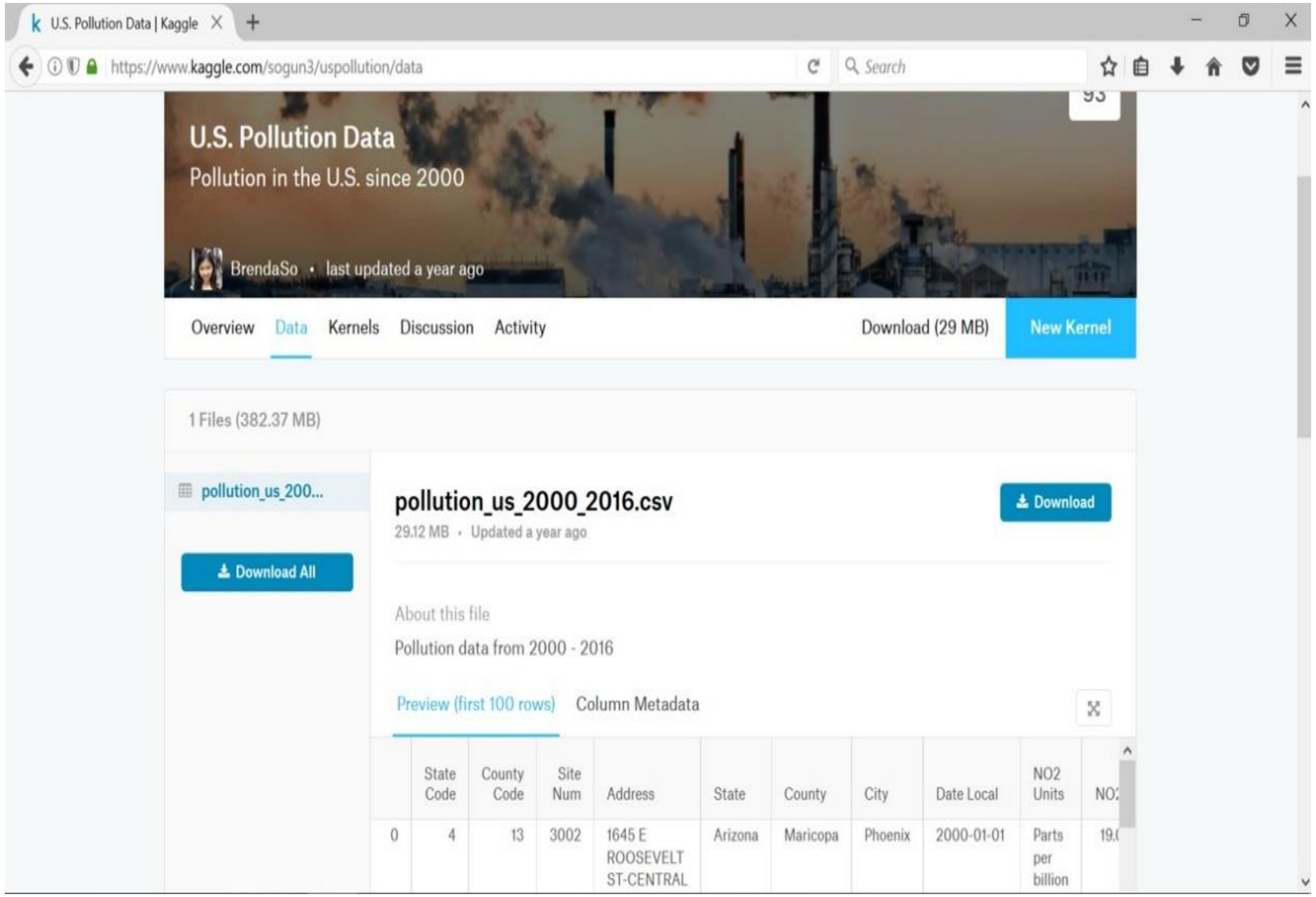
Lambert Conformal Conic Projection

Geocoding Data Science Datasets in R using Google’s API

- Kaggle
- Geocoding a CSV of Addresses using Google’s API

Kaggle (kaggle.com)

- Data science and machine learning site
- Datasets containing place-name or address information can be geocoded to perform spatial analysis



U.S. Pollution Data

Pollution in the U.S. since 2000

BrendaSo • last updated a year ago

1 Files (382.37 MB)

pollution_us_2000_2016.csv

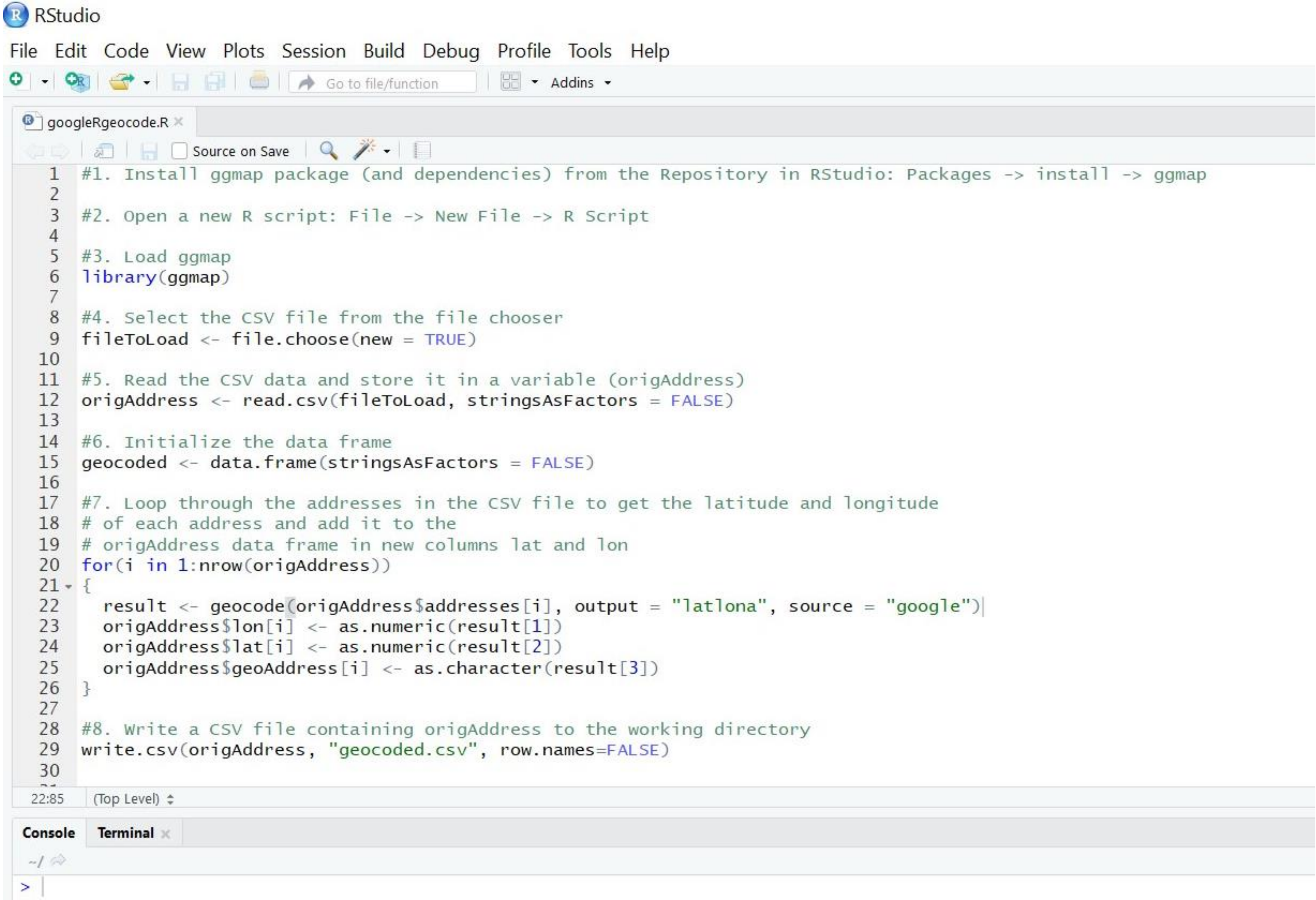
29.12 MB • Updated a year ago

About this file

Pollution data from 2000 - 2016

	State Code	County Code	Site Num	Address	State	County	City	Date Local	NO2 Units	NO2
0	4	13	3002	1645 E ROOSEVELT ST-CENTRAL	Arizona	Maricopa	Phoenix	2000-01-01	Parts per billion	19.(

Demo: Geocoding a CSV of Addresses in R using Google’s Geocoding API



```r
#1. Install ggmap package (and dependencies) from the Repository in RStudio: Packages -> install -> ggmap

#2. Open a new R script: File -> New File -> R Script

#3. Load ggmap
library(ggmap)

#4. Select the CSV file from the file chooser
fileToLoad <- file.choose(new = TRUE)

#5. Read the CSV data and store it in a variable (origAddress)
origAddress <- read.csv(fileToLoad, stringsAsFactors = FALSE)

#6. Initialize the data frame
geocoded <- data.frame(stringsAsFactors = FALSE)

#7. Loop through the addresses in the CSV file to get the latitude and longitude
# of each address and add it to the
# origAddress data frame in new columns lat and lon
for(i in 1:nrow(origAddress))
{
  result <- geocode(origAddress$addresses[i], output = "latlona", source = "google")
  origAddress$lon[i] <- as.numeric(result[1])
  origAddress$lat[i] <- as.numeric(result[2])
  origAddress$geoAddress[i] <- as.character(result[3])
}

#8. Write a CSV file containing origAddress to the working directory
write.csv(origAddress, "geocoded.csv", row.names=FALSE)

```

```
>
```

Data Acquisition and Evaluation Summary

- Data collection can be expensive and time-consuming
  - Main techniques
    - Primary
      - Raster – e.g. remote sensing
      - Vector – e.g. GPS, field survey and LiDAR
    - Secondary
      - Raster – e.g. scanning
      - Vector – e.g. digitizing and geocoding
- Conversion of existing data and online data options available
- Always ask first: **to buy or to build?**

References

- Longley, P. A., M. F. Goodchild, D. J. Maguire, D. W. Rhind (2010). Geographic Information Systems & Science (3rd Ed). John Wiley & Sons, Inc.
- Clarke, K.C. (2010). Getting Started with GIS (5th Ed). Prentice-Hall, Inc., London.
- Chang, K-T. (2010). Introduction to Geographic Information Systems (5th Ed). McGraw-Hill.
- Center for Geographic Analysis (CGA) Harvard: GIS Training Materials.
- Storybench: http://www.storybench.org/geocode-csv-addresses-r/
- Google: https://developers.google.com/maps/documentation/geocoding/start

Q&A