GEOG

Geographic Information Science in Political Science Research

Methods Workshop for UMD Department of Government and Politics

October 27, 2021

Presented by Henry Overos (GOVT), Jeff Sauer (GEOG)

CENTER FOR GEOSPATIAL INFORMATION SCIENCE

Agenda

- Why Geographic Information Science (GIS)?
- Recent discourse between Political Science and Geography
- What is ‘geographic’ data?
- Quick overview of GIS tools
- Illustrating GIS in contemporary research: working through Mildenberger 2020 exposure calculation
- Building your own geospatial datasets
- A few key considerations (projection, spatial mismatch)

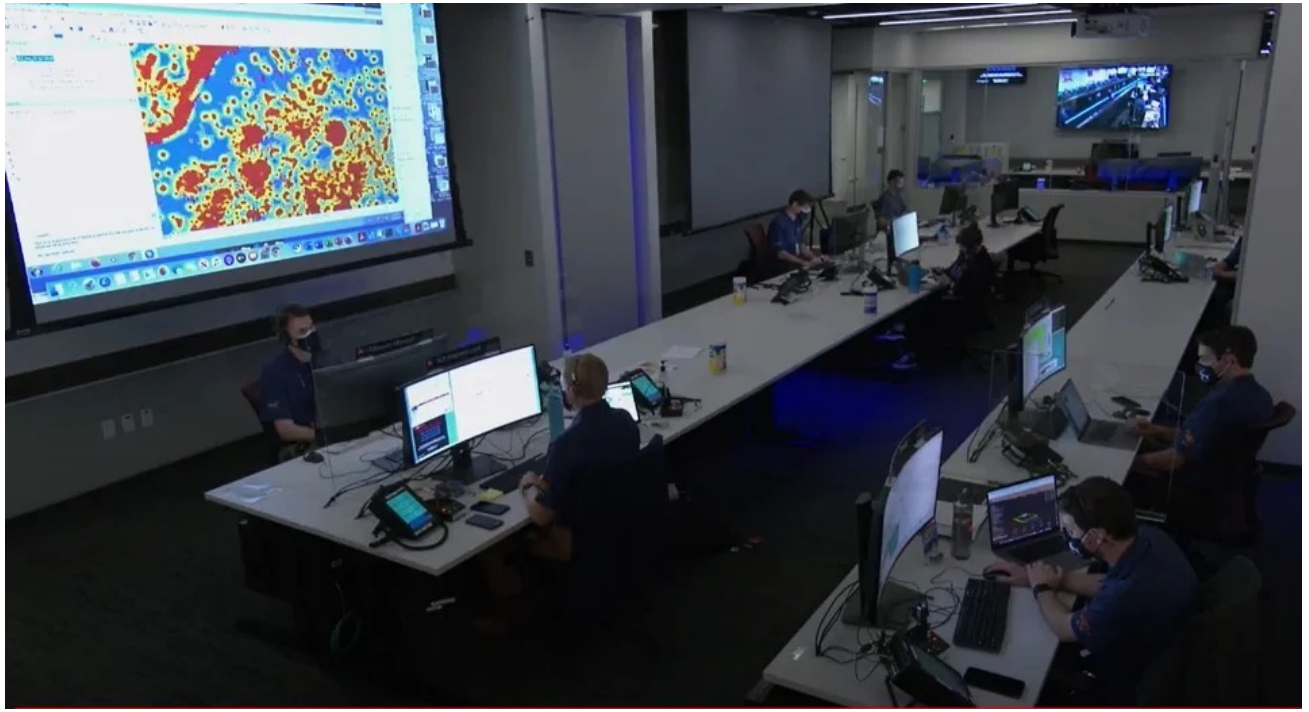
Why Geographic Information Science (GIS)?

"…about 1 in 3 of queries that people just type into a standard Google search bar are about places, they are about finding out information about locations. …this isn't Google Maps, just people normally looking at Google."

Ed Parsons, Chief Technologist at Google, in 2012 interview. More info, PinPoint 2012 talk.

Why Geographic Information Science (GIS)?

LANDSAT-9 satellite preparing for September 27, 2021, launch



NASA monitoring Mars Perseverance mission using AcGIS

Why Geographic Information Science (GIS)?

Commercial giants like Uber rely on state-of-the-art GIS technology, sometimes releasing it to the public to speed up innovation. Example: kepler.gl

Why Geographic Information Science (GIS)?

State Rep. John Szoka of North Carolina examining a redistricting map for 2020 legislative elections. Source: Associated Press.

Discourse between Political Science and Geography

Gary King. 1996. "Why Context Should Not Count." *Political Geography*, 15, Pp. 159–164.

King argued that:

1. Context rarely counts
2. Goal of political researchers should be to show that context does not count
3. Theoretical analyses of empirical questions is not useful
4. Similarly, empirical analyses of theoretical questions is not useful

Published at a turning point in both King's own work and geographic information systems (GIS)

- King would publish 'A Solution to the Ecological Inference Problem' in 1997
- Geography increasingly focused on potentials and applications of GIS as personal computing expands

Discourse between Political Science and Geography

Perhaps a *reversal* in recent years?

Political scientists increasingly using GIS

- Charnysh and Finkel. 2017. ‘The Death Camp Eldorado: Political and Economic Effects of Mass Violence’. *American Political Science Review.*
- Hazlett and Mildenberger. 2020. ‘Wildfire Exposure Increases Pro-Environment Voting within Democratic but Not Republican Areas.’ *American Political Science Review*.

Geographers focusing again on demonstrating the importance of ‘context’

- Fotheringham et al. 2021. ‘Scale, Context, and Heterogeneity: A Spatial Analytical Perspective on the 2016 U.S. Presidential Election.’ *Annals of the American Association of Geographers.*

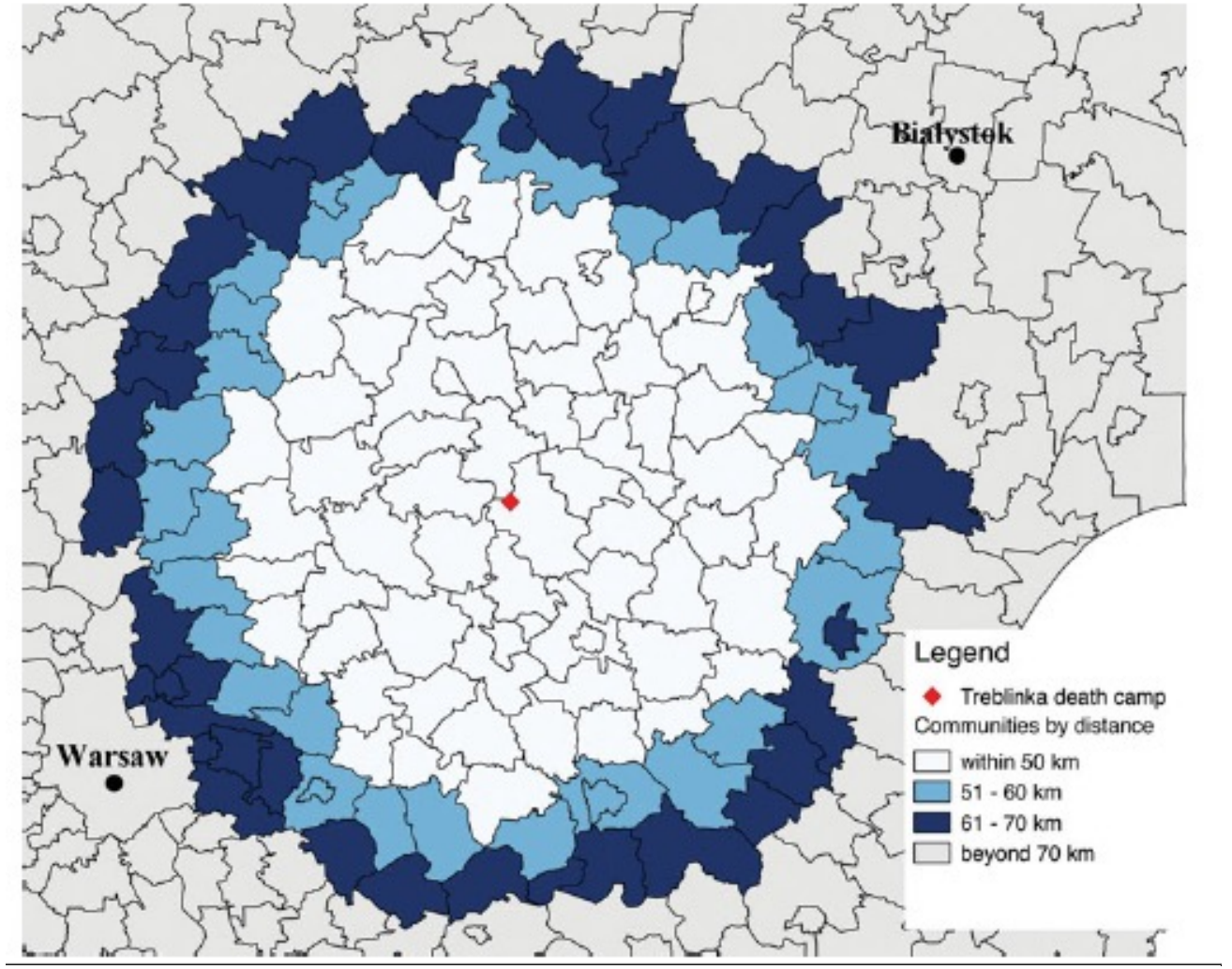
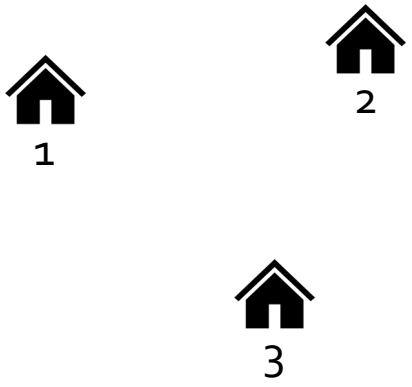


Figure 2 from Charnysh and Finkel (2017) *APSR*. ‘Communities at a 50-, 60-, and 70-km Distance from Treblinka’.

What is ‘geographic’ data?

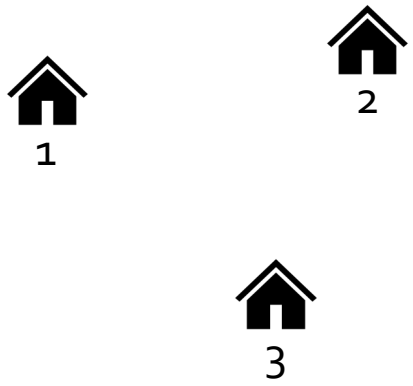
'Geographic’ data attempts to represent the real-world using computers

Imagine you are collecting survey information at various homes



What is ‘geographic’ data?

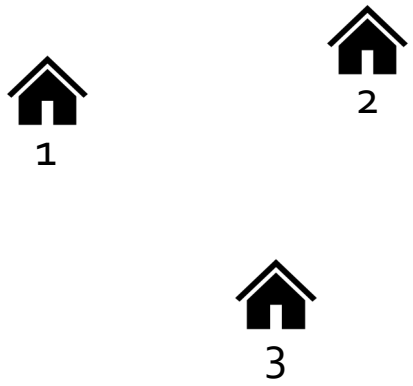
A traditional database for information describing these homes might resemble the following



HouseID	NumPeople	Income
1	4	50,000
2	3	47,000
3	6	102,000

What is ‘geographic’ data?

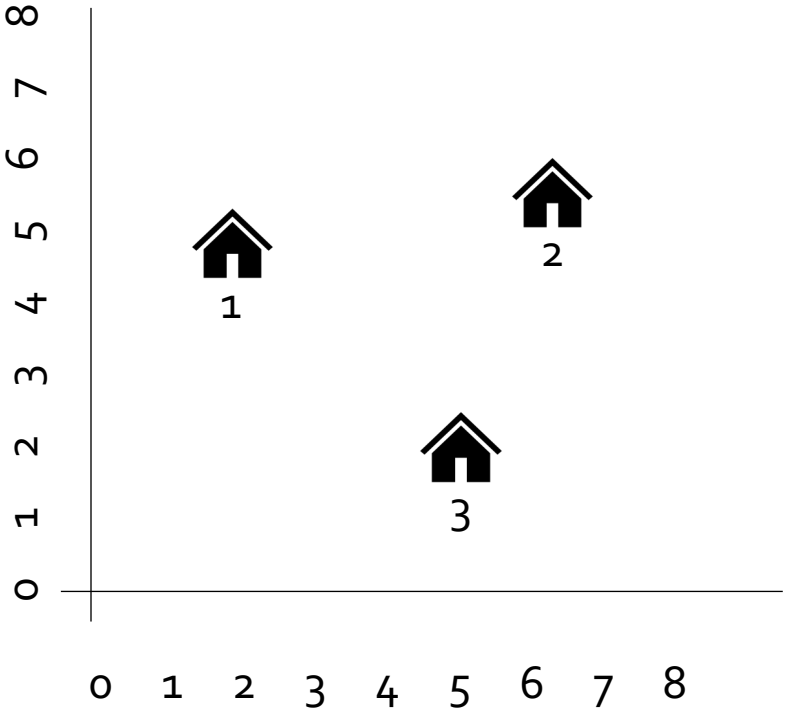
A spatial database describes the location of each observation using (x,y) coordinate pairs



HouseID	NumPeople	Income	Geometry
1	4	50,000	
2	3	47,000	
3	6	102,000	

What is ‘geographic’ data?

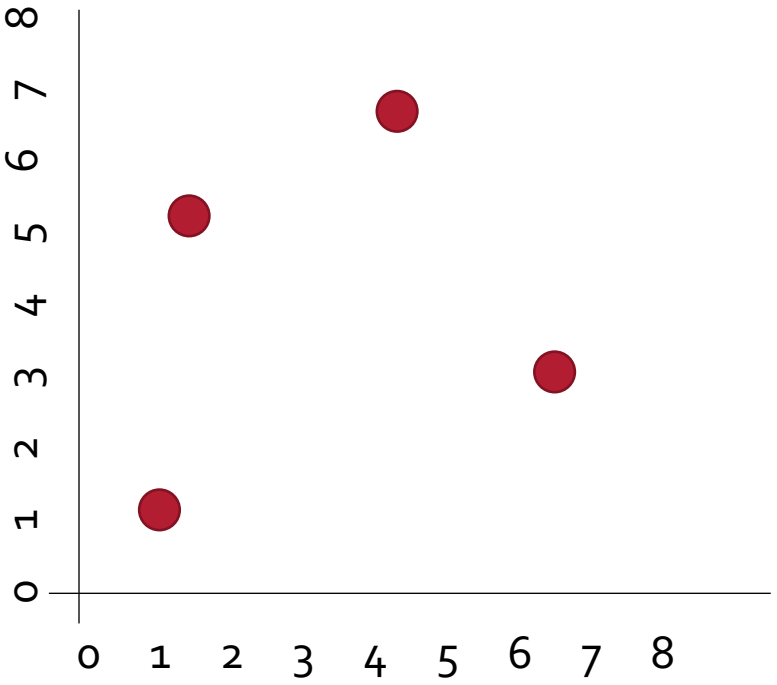
A spatial database describes the location of each observation using (x,y) coordinate pairs



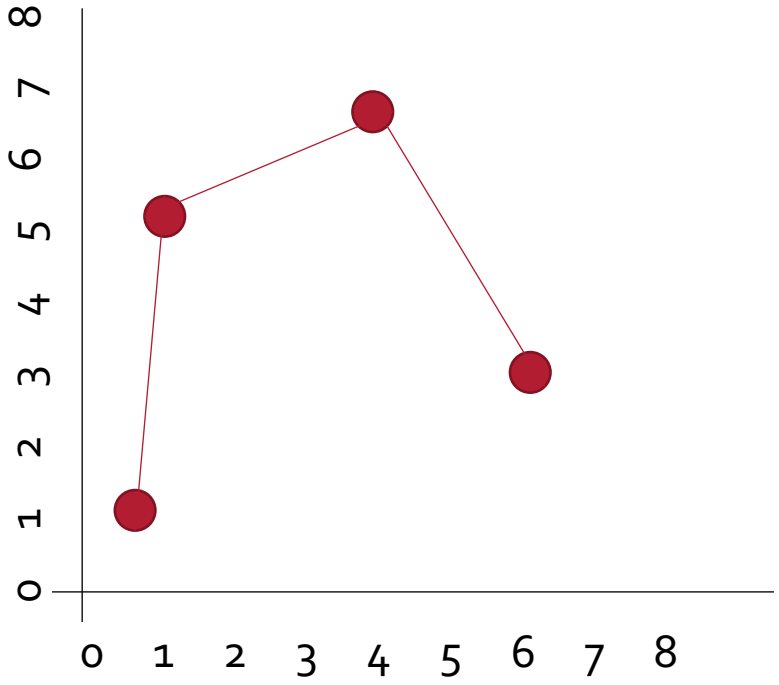
HouseID	NumPeople	Income	Geometry
1	4	50,000	[2.0,4.5]
2	3	47,000	[6.7,5.2]
3	6	102,000	[5.0,2.0]

What is ‘geographic’ data?

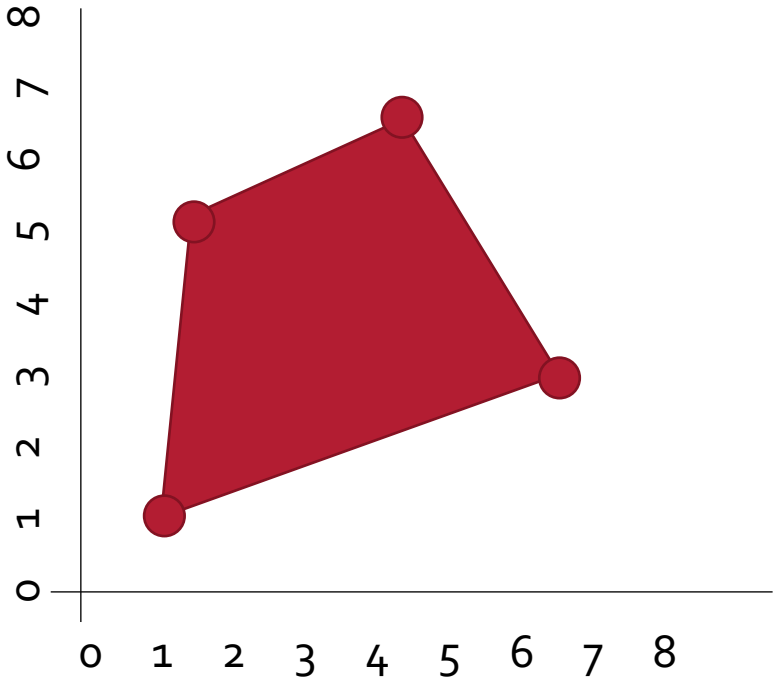
Many types of geographic data evolves from points to form lines, polygons, and other complex representations



N=4 *point* observations



N=1 *line* observation



N=1 *polygon* observation

Geographic data file formats

Geographic data for points, lines, and polygons is commonly referred to as *vector* data

1. **Shapefile (.shp)** – most common, will often come with supplemental files like projection (.prj), database (.dbx), and more. Certain limitations around attribute formatting and size.

2. **Geopackage (.gpkg)** – from the Open Geospatial Consortium (OGC), a special type of SQLite database. Increasingly common and efficient for large spatial datasets.

3. **ESRI File Geodatabase (.gbd)** – proprietary file format from ESRI, commonly encountered when working with data provided by large companies.

4. **Keyhole Markup Language (.kml)/(.kmz)** – also from OGC, originally created for Google Earth, now commonly used across web-based mapping platforms.

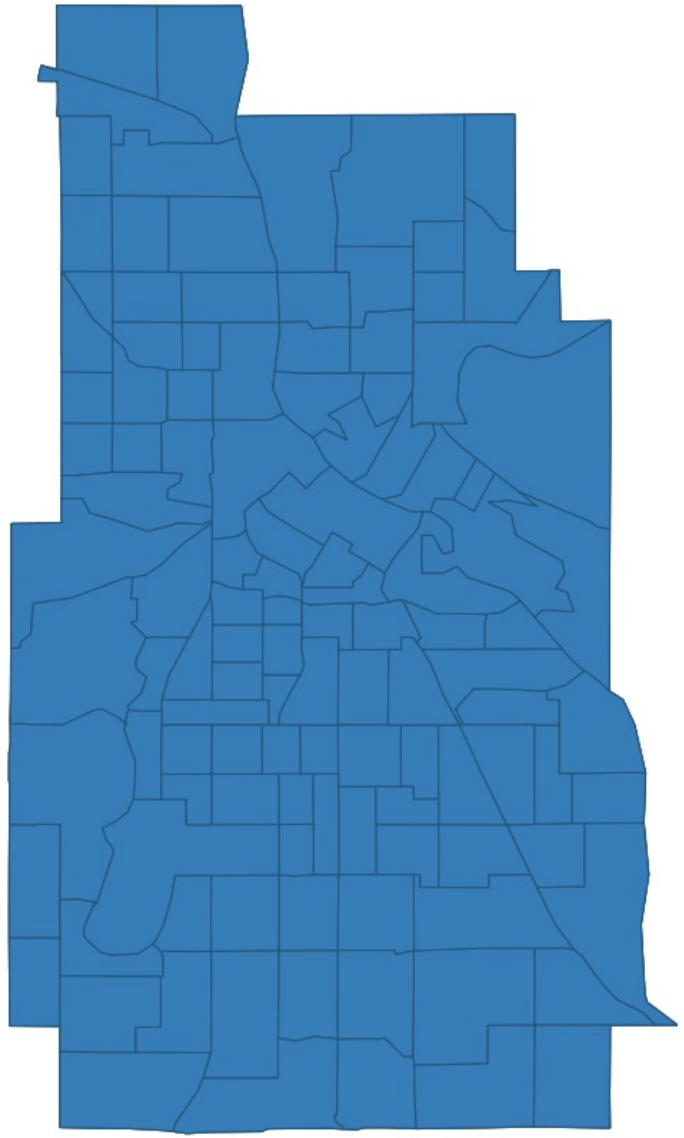
Quick demonstration! Finding ‘geographic’ data

Right: Census Tracts (CT) in Minneapolis, Minnesota. Sourced from the County Government (Hennepin) GIS Open Data Portal (https://gis-hennepin.opendata.arcgis.com)

Each CT is a closed polygon, which has *geometry* and *attribute* information

Geometry can be exploited to calculate other geometric measurements (distance, area, perimeter)

Attributes can be mapped, manipulated, or combined with geometric measurements



Working with geographic data – standalone software

Standalone software:

- **QGIS** – free, open source, standard set of GIS tools, ability to download and work with community-contributed tools, direct interface with command-line tools (python, gdal, grass, among others)

- **ArcMap/ArcGIS Pro** – proprietary ($$$), licenses through UMD, expanded set of validated tools (especially for ML and interfacing with big data), uses custom python interface (ArcPy)

Working with geographic data – languages

R

- **Manipulating data**: sf, sp, rgdal, raster
- **Spatial analysis**: spdep, spatialreg
- **Mapmaking**: tmap, maptools

Python:

- **Manipulating data**: GeoPandas, shapely,GDAL, rasterio, PyProj
- **Spatial analysis**: PySAL, pygis, networkx
- **Mapmaping**: matplotlib (interfaces with geopandas), bokeh, folium

Working through Hazlett and Mildenberger 2020

American Political Science Review

Wildfire Exposure Increases Pro-Environment Voting within Democratic but Not Republican Areas

Published online by Cambridge University Press: **15 July 2020**

CHAD HAZLETT and MATTO MILDENBERGER

Show author details

Article | Supplementary materials | Metrics

Get access | Share | Cite | Rights & Permissions

Working through Hazlett and Mildenberger 2020

Generalizable hypothesis

- individuals who experience a climate-related hazard will alter their political behavior.

Context-specific hypothesis

- people in California who experience climate-related wildfires may be more likely to support pro-environmental ballot initiatives.

Outcome measurement

- Average precinct-level vote support for four pro-environmental ballot initiatives (as %)

Treatment measurement

- precinct-level distance from nearest wildfire (in kilometer, km)

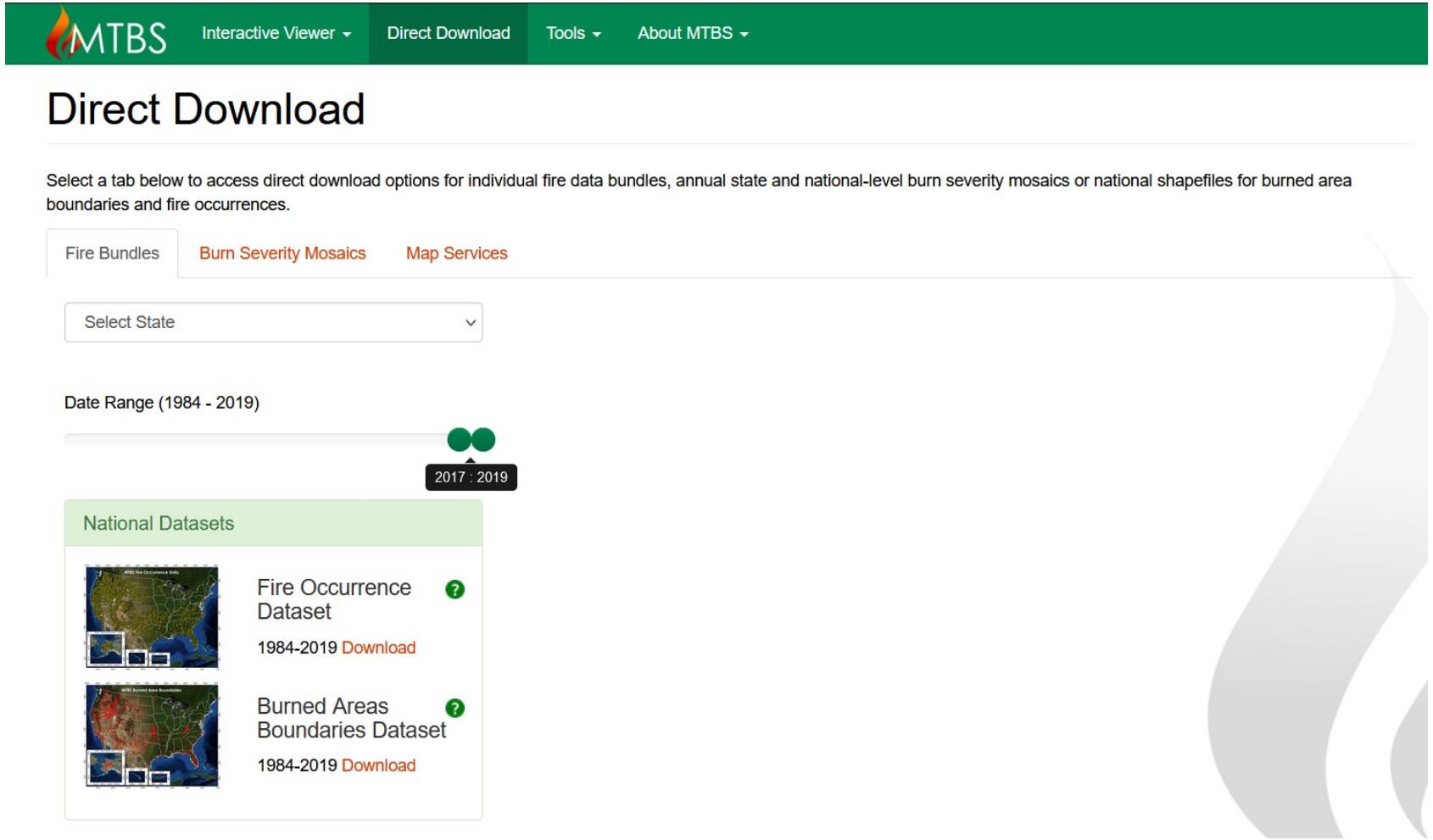
Working through Hazlett and Mildenberger 2020

Treatment Measurement

We extract **wildfire perimeter data** from the **Monitoring Trends in Burn Severity dataset**, an interagency US government effort tracking large fires via Landsat satellite data. We then **spatially merge** the wildfire perimeter data to the census block group data to **determine each block group's distance from wildfires**.

(from Hazlett and Mildenberger 2020; emphasis added)

Working through Hazlett and Mildenberger 2020

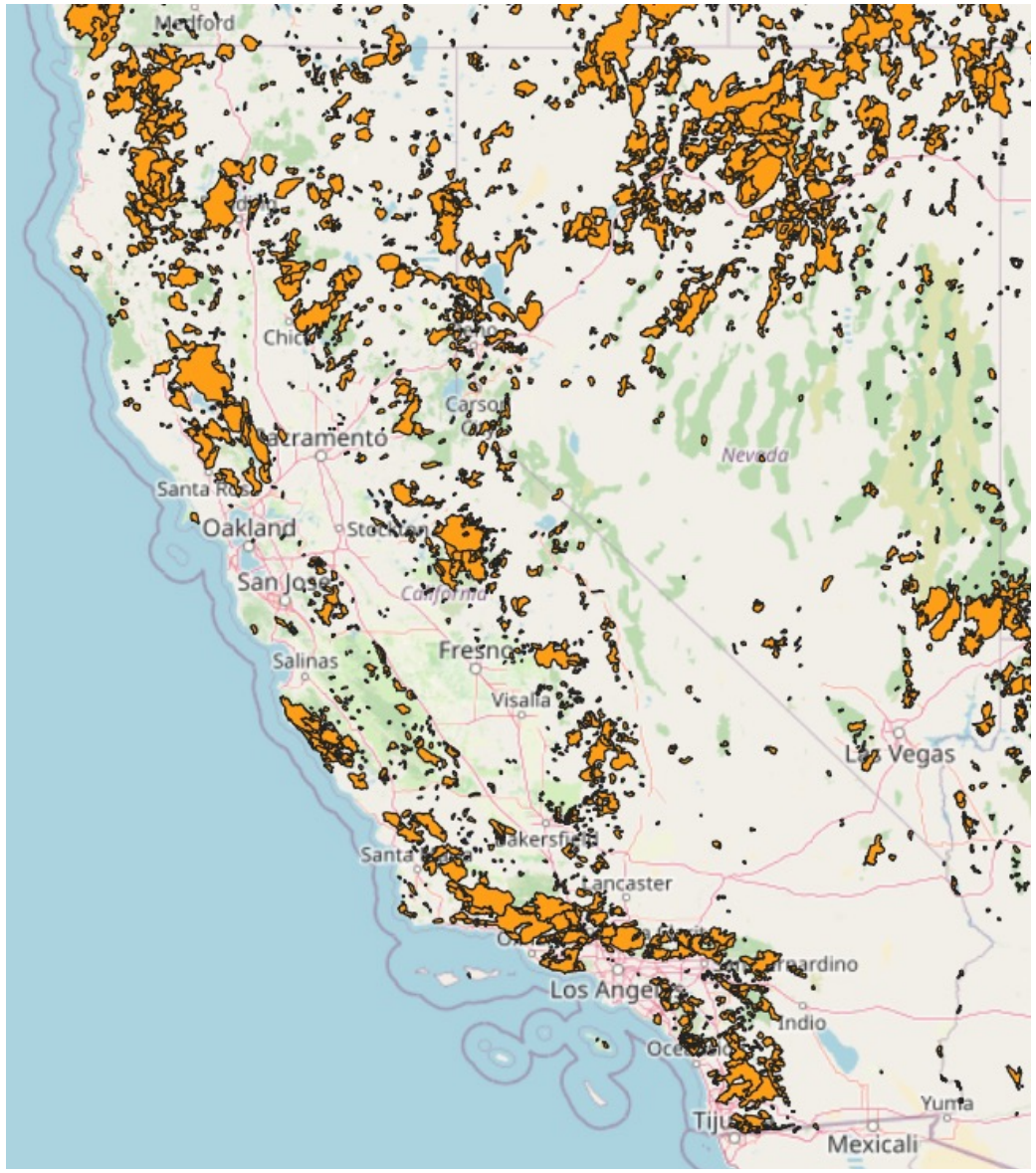


https://www.mtbs.gov/direct-download

Working through Hazlett and Mildenberger 2020

Right: MTBS fire perimeter data loaded into QGIS with an Open Street Map (OSM) basemap

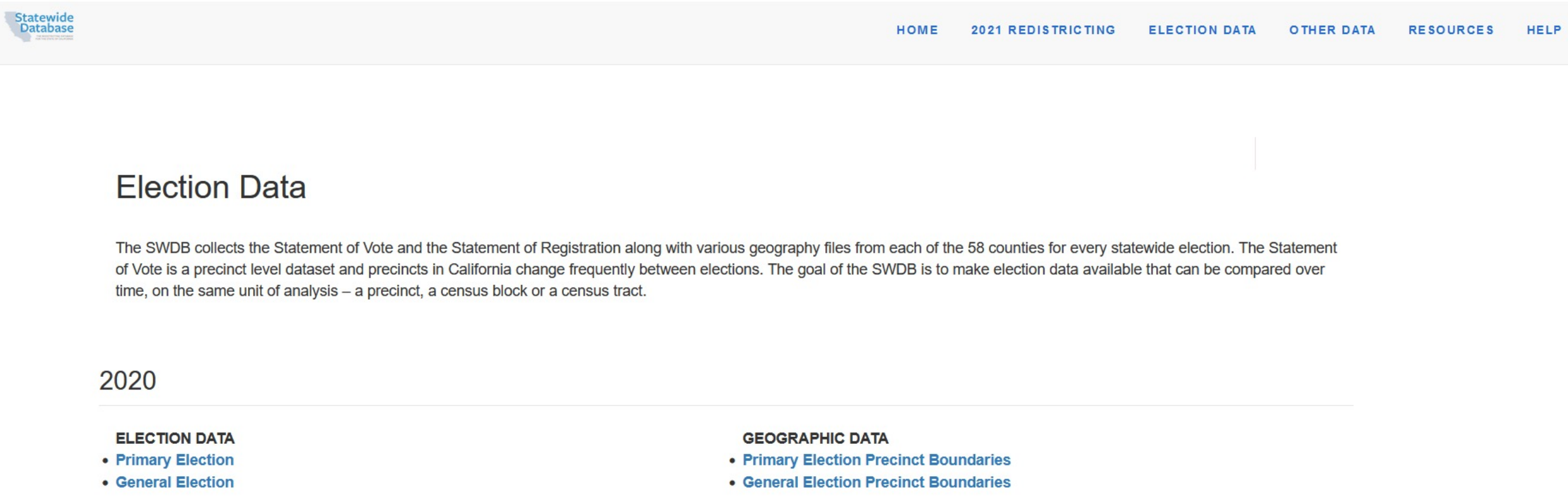
Lower left: Attribute table associated with each fire perimeter (ID, Name, Type, etc.)



mtbs_perims_DD — Features Total: 28584, Filtered: 28584, Selected: 0

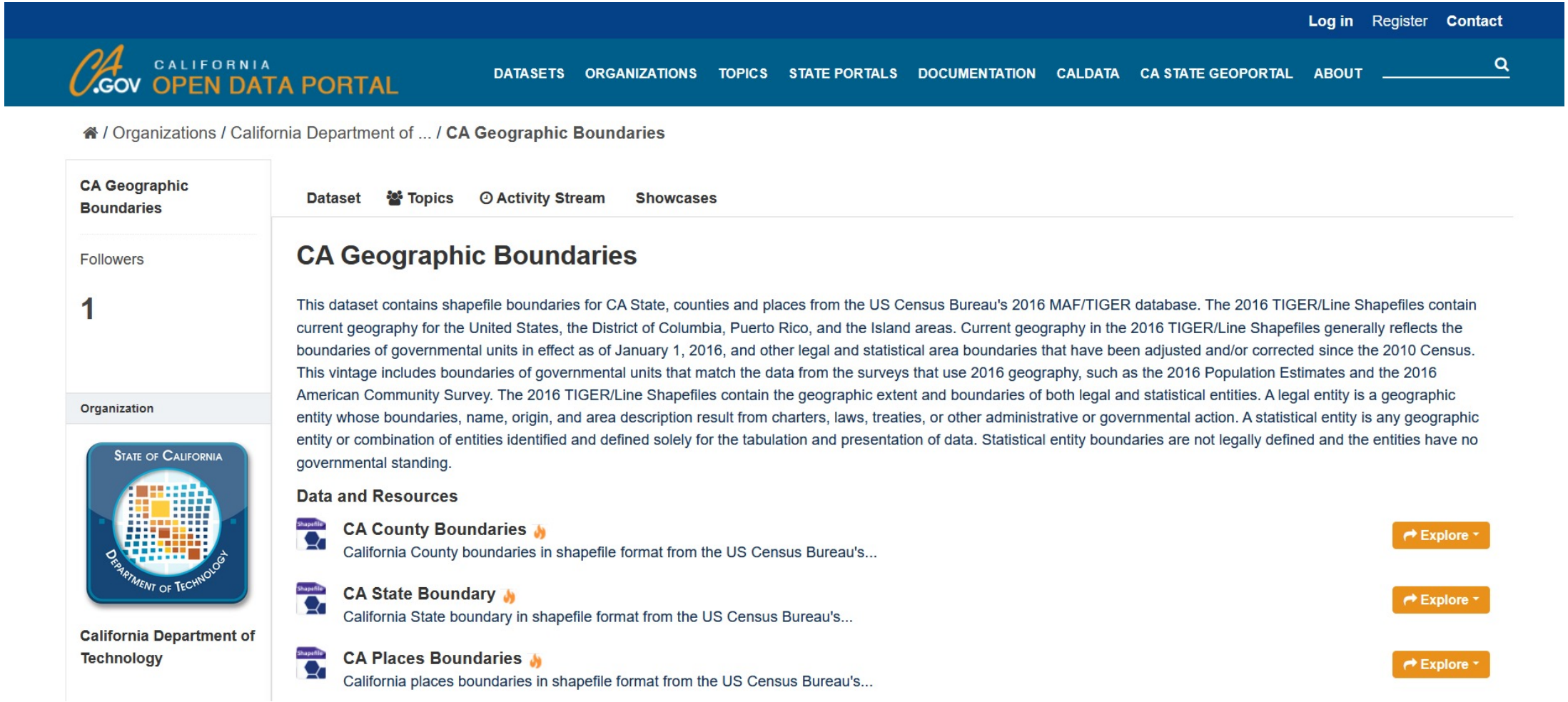
	Event_ID	irwinID	Incid_Name	Incid_Type	Map_ID	Map_Prog	Asmnt_Type	BurnBndAc	BurnBndLat	BurnBndLon
1	UT41764111488...	NULL	HELLS HOLLOW	Prescribed Fire	414	MTBS	Extended	4758	41.766	-111.507
2	WA4796812059...	NULL	KLONE PEAK	Wildfire	441	MTBS	Extended	1634	47.972	-120.582
3	WA4797012073...	NULL	BASALT	Wildfire	444	MTBS	Extended	1705	47.967	-120.746
4	WA4805112058...	NULL	PYRAMID	Wildfire	449	MTBS	Extended	1962	48.044	-120.581
5	WA4788612026...	NULL	FIRST CREEK	Wildfire	454	MTBS	Extended	1434	47.895	-120.259
6	WA4774112024...	NULL	BYRD	Wildfire	460	MTBS	Initial	14145	47.771	-120.215
7	WA4806011985...	NULL	CRANE ROAD	Wildfire	461	MTBS	Initial	12720	48.009	-119.826
8	WA4793011990...	NULL	ANTOINE 2	Wildfire	463	MTBS	Initial	7166	47.965	-119.909
9	FL28760082394...	NULL	UNNAMED	Prescribed Fire	497	MTBS	Initial (SS)	600	28.761	-82.394
10	FL28585082253...	NULL	UNNAMED	Prescribed Fire	505	MTBS	Initial	974	28.585	-82.253

Working through Hazlett and Mildenberger 2020



https://statewidedatabase.org/election.html

Working through Hazlett and Mildenberger 2020



https://data.ca.gov/dataset/ca-geographic-boundaries

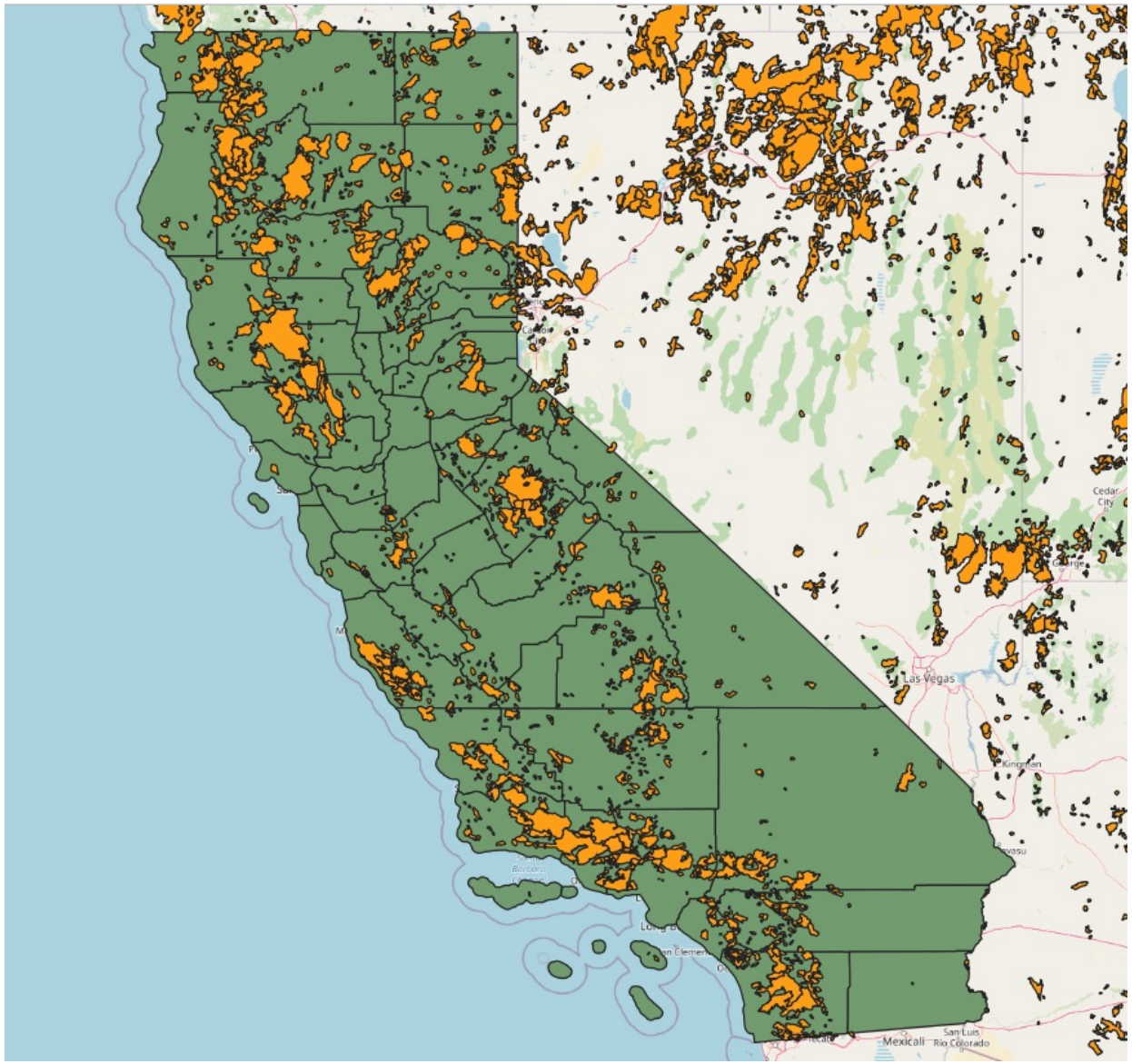
Working through Hazlett and Mildenberger 2020

Right: California counties (green), with MTBS fire perimeters superimposed (orange)

Note that MTBS fire perimeter dataset is nationwide

To ease calculations, we want to *restrict* the MTBS fire data to California

This operation is known as *clip*



Working through Hazlett and Mildenberger 2020

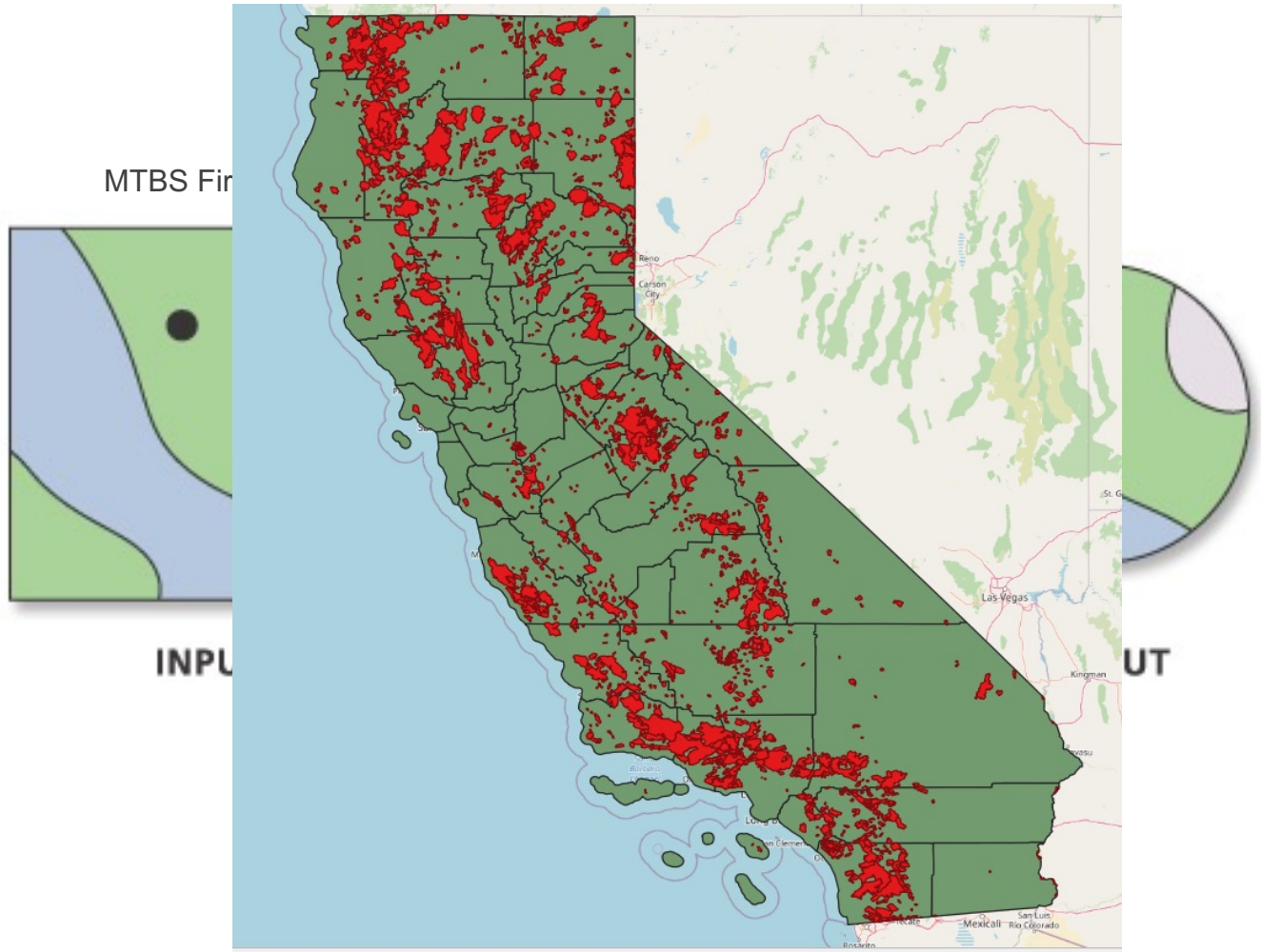
Clip is a common vector processing operation

Available in standard QGIS

Accessible via the following menus:

- Vector > Geoprocessing Tools > Clip

Result shrinks MTBS fire dataset from n=28584 fire perimeters to n=1895



Working through Hazlett and Mildenberger 2020

We can now calculate the distance from each political unit (precinct, county, or other) to the nearest wildfire perimeter

This is known as a *nearest neighbor* analysis

Several approaches to executing this analysis:

- QGIS: standard QGIS only supports point-to-point nearest neighbor analysis, so we could use the popular community tool NNJoin. You can install NNJoin with the following steps: **Plugins > Manage and Install Plugins > Search 'NNJoins' > Install**
- R: use a combination of **sf::st_nearest_feature()** and **sf::st_distance()**

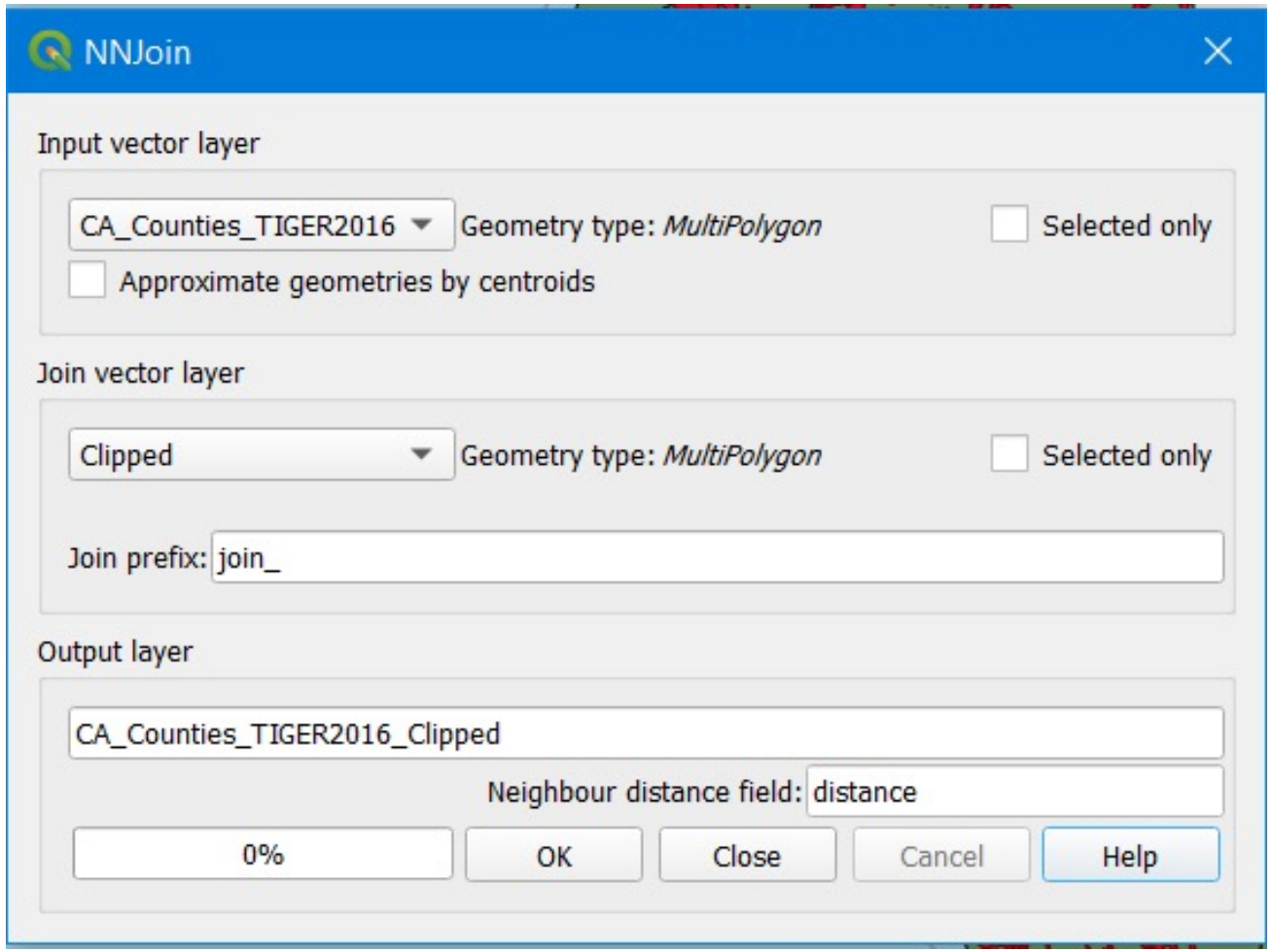
You can speed up the analysis by simplifying polygons, merging small polygons into large polygons, and other computational tricks

Working through Hazlett and Mildenberger 2020

Input layer: what we are calculating distances from

Join layer: what we are calculating distances to

Output layer: resulting layer where distances will be stored

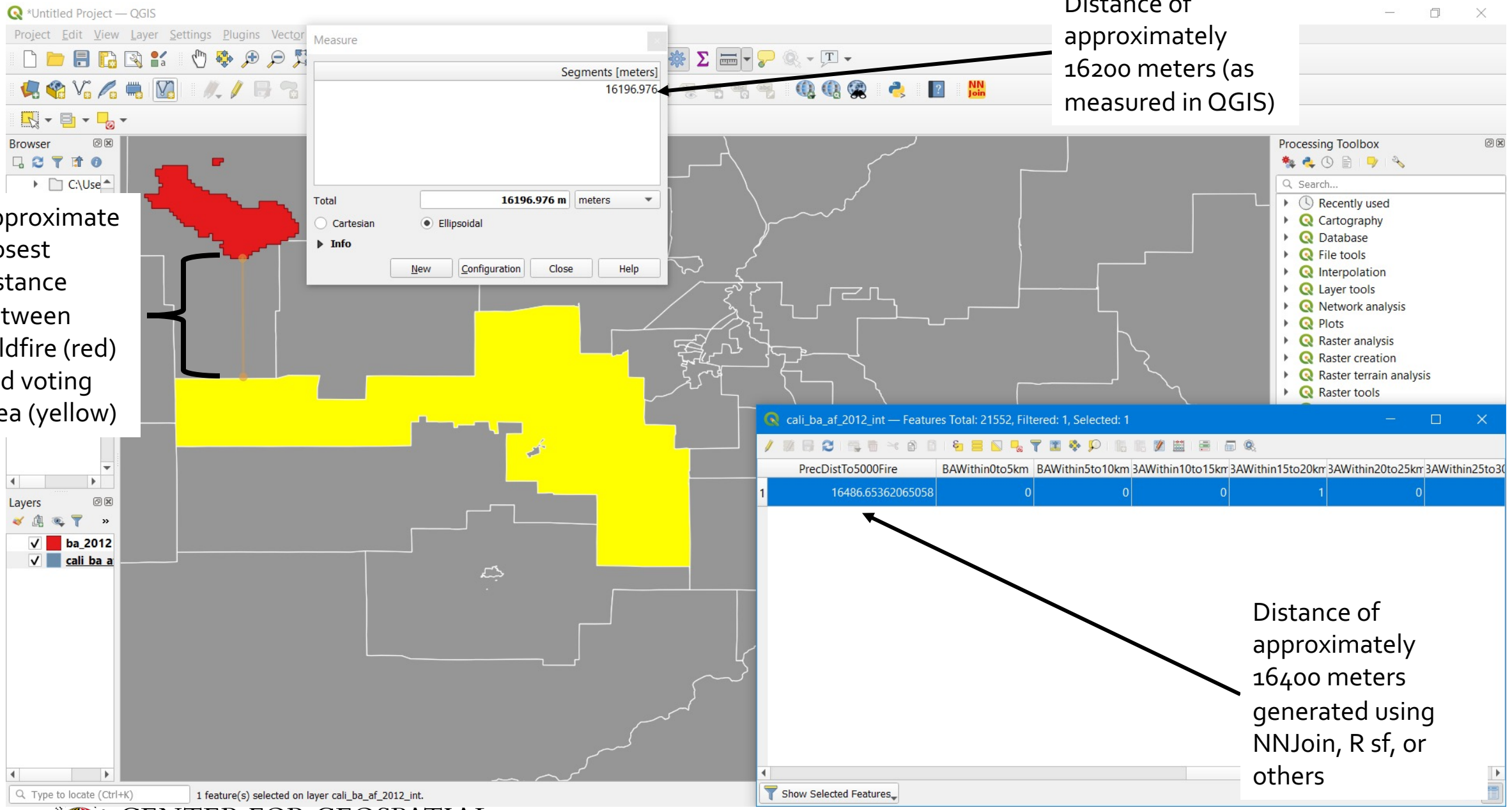


Working through Hazlett and Mildenberger 2020

```
## The resulting object is a list index identifying which fire polygon
## is closest to each precinct. This part takes the longest!
CloseIndex <-
  sf::st_nearest_feature(PoliticalBoundaries, BurnPerimeter)

## Select those fire polygons that made it into the index (ignoring others
## to help speed up calculations)
BurnPerimeterTemp <- BurnPerimeter %>% dplyr::slice(CloseIndex)

## Calculate the minimum distance between each precinct and the associated
## fire polygon. Resulting distance is in meters.
PrecinctToFire <-
  sf::st_distance(PoliticalBoundaries, BurnPerimeterTemp, by_element = TRUE)
```


Distance of approximately 16200 meters (as measured in QGIS)
Approximate closest distance between wildfire (red) and voting area (yellow)
Distance of approximately 16400 meters generated using NNJoin, R sf, or others
Measure
Segments [meters]
16196.976
Total
16196.976 m
meters
Cartesian
Ellipsoidal
Info
New
Configuration
Close
Help
Browser
Layers
Processing Toolbox
Recently used
Cartography
Database
File tools
Interpolation
Layer tools
Network analysis
Plots
Raster analysis
Raster creation
Raster terrain analysis
Raster tools
cali_ba_af_2012_int — Features Total: 21552, Filtered: 1, Selected: 1
PrecDistTo5000Fire
BAWithin0to5km
BAWithin5to10km
BAWithin10to15km
BAWithin15to20km
BAWithin20to25km
16486.65362065058
Show Selected Features
1 feature(s) selected on layer cali_ba_af_2012_int.

Working through Hazlett and Mildenberger 2020

- …and that wraps up the exposure calculation!
- Could now export the dataset as .csv for analysis in your preferred statistical programming software now export the dataset
- Hazlett and Mildenberger then go one to discretize the continuous measure of distance into 5-kilometer bands

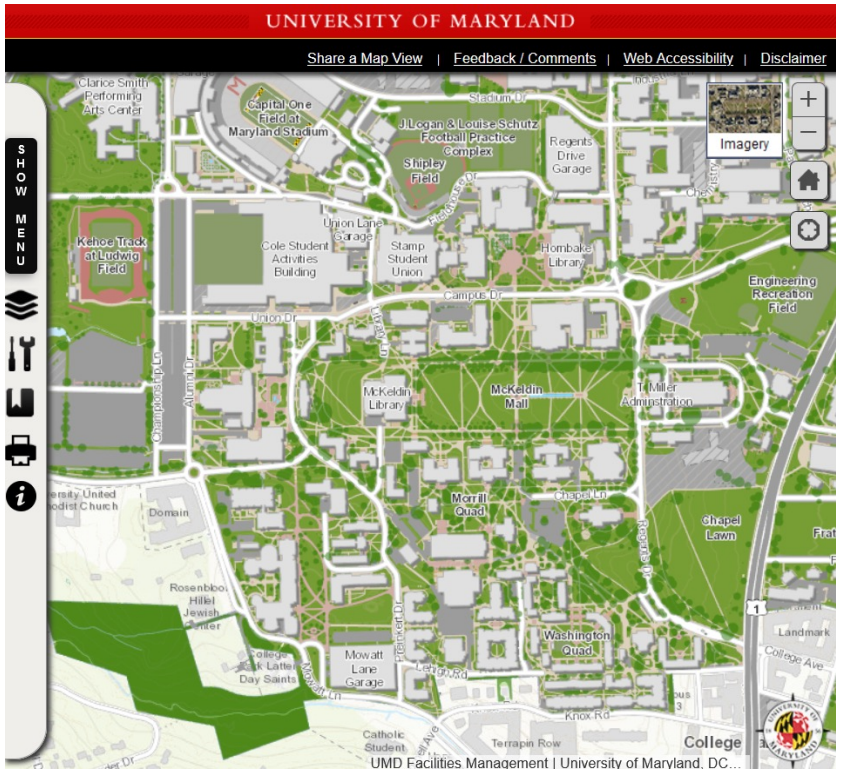
Building your own geospatial datasets

"…the widespread engagement of large numbers of private citizens, often with little in the way of formal qualifications, in the creation of geographic information, a function that for centuries has been reserved to official agencies. They are largely untrained and their actions are almost always voluntary, and the results may or may not be accurate." Michael F. Goodchild, 2007

Wikipedia, Google Maps, Reviews on Yelp, Travel Guides, Stories, Interpersonal knowledge

Building your own geospatial datasets

You can use **beautifulsoup4** in python, **rvest** in R, or other web-scraping tools to collect publicly available information and transform it into spatial data!

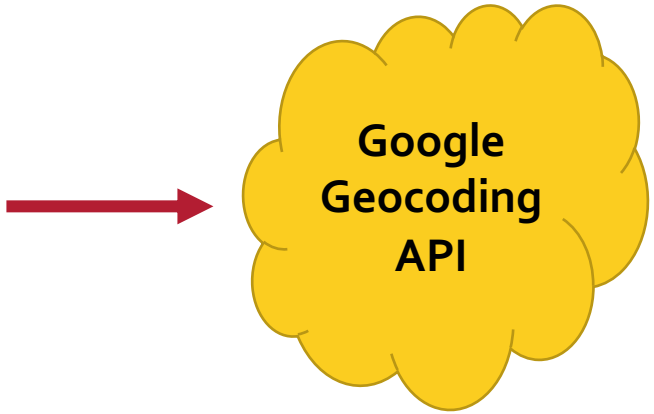


BuildingName	ScrapedAddress
Health Center	3983 Campus Dr., College Park, MD 20742
Jimenez Hall	4125 Library Ln., College Park, MD 20742
McKeldin Library	7649 Library Ln., College Park, MD 20742

Building your own geospatial datasets

Geocoding is the process of obtaining (x,y) coordinate information from street addresses

BuildingName	ScrapedAddress
Health Center	3983 Campus Dr., College Park, MD 20742
Jimenez Hall	4125 Library Ln., College Park, MD 20742
McKeldin Library	7649 Library Ln., College Park, MD 20742



ggmap::geocode() in R

googlemaps.geocode() in Python

Building your own geospatial datasets

Resulting dataset can be transformed into a *point* shapefile in QGIS, ArcMap, Python, R, and others by reading longitude as x and latitude as y

In QGIS: **Layer > Add Layer > Add Delimited Text Layer…**

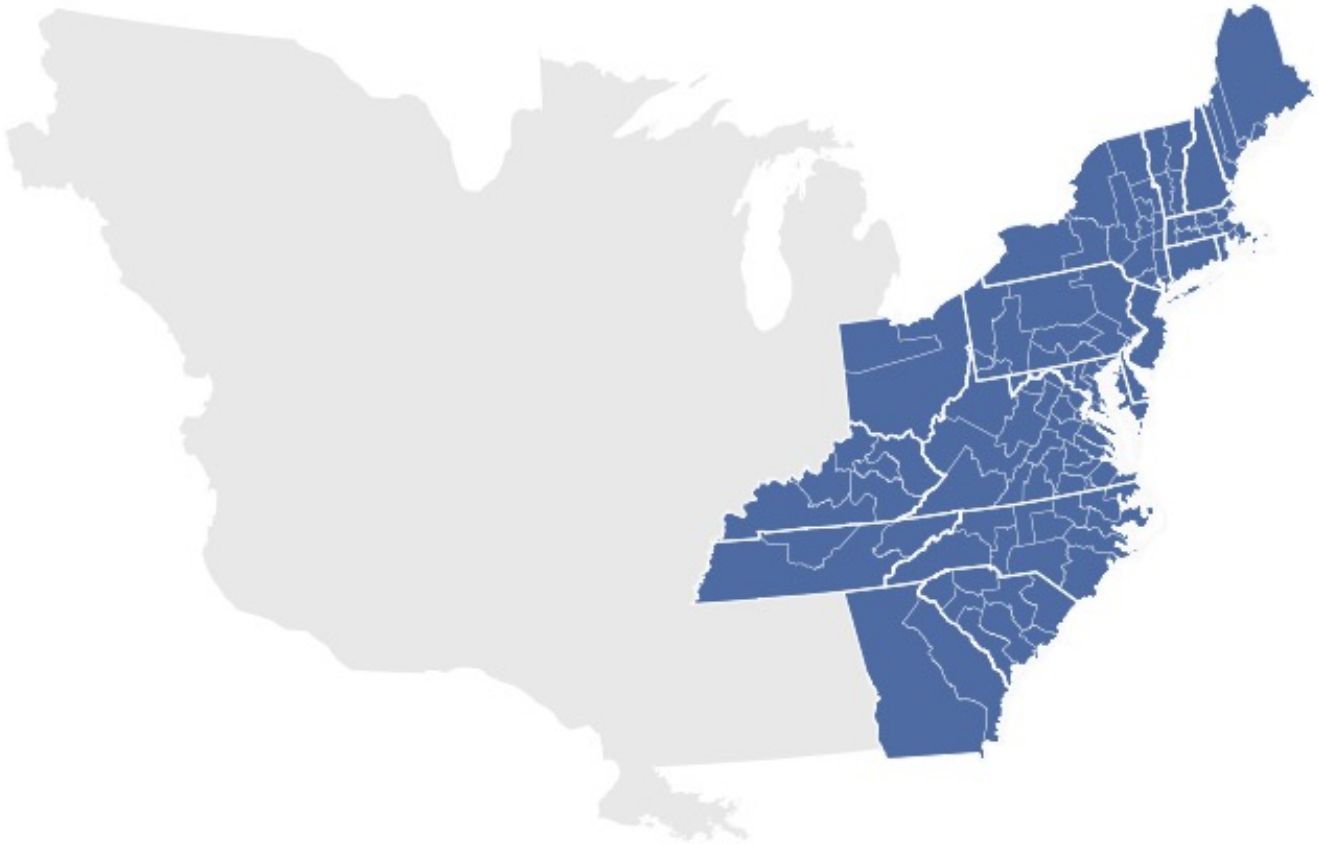
BuildingName	ScrapedAddress	Latitude	Longitude
Health Center	*3983 Campus Dr., College Park, MD 20742*	38.985240	-76.946580
Jimenez Hall	4125 Library Ln., College Park, MD 20742	38.987140	-76.945250
McKeldin Library	7649 Library Ln., College Park, MD 20742	38.987140	-76.945250

A few key considerations: projections

- Projections as special procedures used adjust geographic data for the 3D complexity of Earth
- Always ensure that your analysis datasets are in **consistent** projection
- If analysis datasets are in different projections, *re-project* to a single system
- Especially important if you are **calculating geometric measurements** over large spaces (area, distance, perimeter, etc.)

A few key considerations: spatial mismatch

- Most political areas have shifted boundaries over time
- Shifting boundaries are especially important for geometric calculations, especially if you are working with longitudinal data
- More critically, shifting boundaries can change what your unit of observation means (e.g., the population of a unit in Time 1 may not be the same population in Time 2)
- Conduct visual checks and use crosswalks (if available)



Historical US congressional district shapefiles (built by Jeffrey B. Lewis, Brandon DeVine, and Lincoln Pritcher with Kenneth C. Martis). Available from: https://cdmaps.polisci.ucla.edu/

Working with geographic data – resources

R

- Spatial Data Science with R (by Pebesma and Bivand)
- Geocomputation with R (by Lovelace)

Python

- Geographic Data Science with Python (by Wolf, Arribas-Bel, and Rey)
- Extensive list of python packages for working with geospatial data

Advanced and Command Line Tools

- GDAL
- GRASS
- SAGA

Thank you so much!

Always feel free to reach out!

Jeff Sauer: jcsauer@terpmail.umd.edu

Henry Overos: hoveros@terpmail.umd.edu

GEOG

Geographic Information Science in Political Science Research

Methods Workshop for UMD Department of Government and Politics

October 27, 2021

Presented by Henry Overos (GOVT), Jeff Sauer (GEOG)

CENTER FOR GEOSPATIAL INFORMATION SCIENCE