geogRaphy: an introduction to spatial data in R

Sarah C. Lotspeich

Vanderbilt University

17 Nov 2020

Introduction

spatial data:

xf

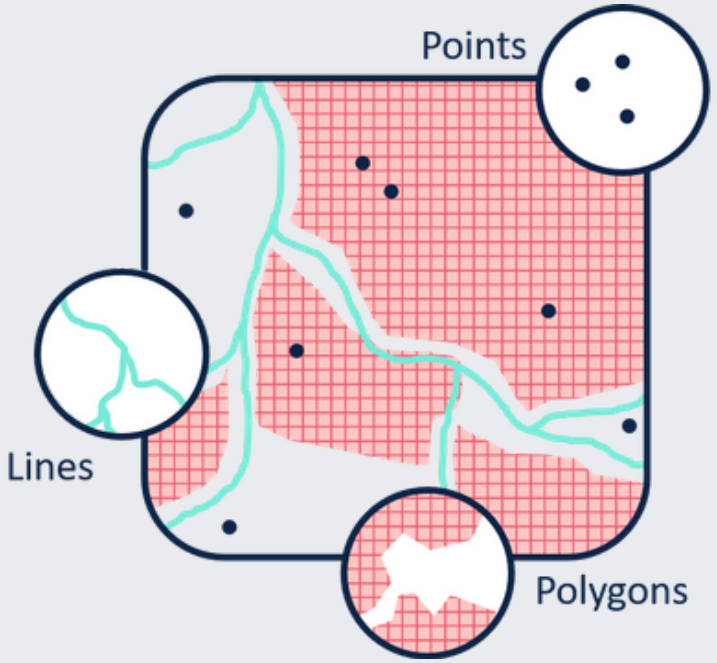
noun • [spay-shull day-ta] • data which is representative of a specific, geographic location on the surface of the Earth.

[1] What Is Spatial Data?

Two main types:

1. Vector data

- Graphical representations of the real world.
- Usually, points, lines, and polygons.
- GIS shapefiles (`.shp`) are typically vector data.



[1] Geospatial Data Models

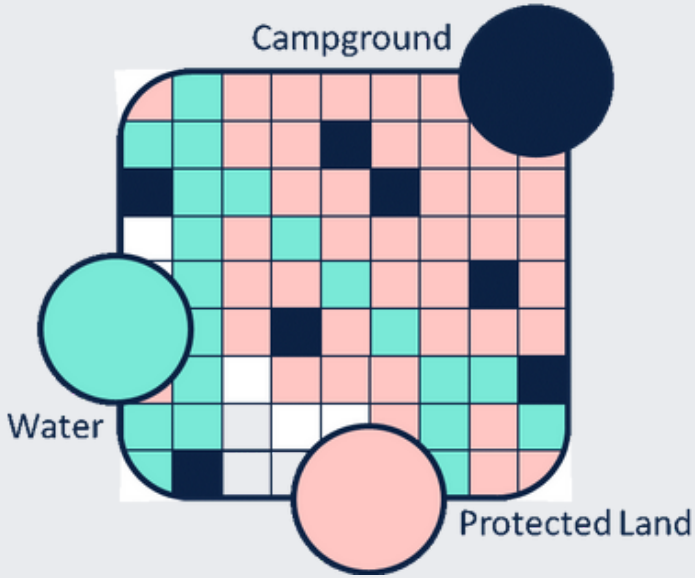
[2] What Is Spatial Data?

[3] Vector and Raster: A Tale of Two Spatial Data Types

Two main types:

2. Raster data

- Continuous surface divided into a grid of cells (pixels).
- Each pixel contains measured value for the area it represents.
- One cell might represent a 10m x 10m area.



[1] Geospatial Data Models

[2] What Is Spatial Data?

[3] Vector and Raster: A Tale of Two Spatial Data Types

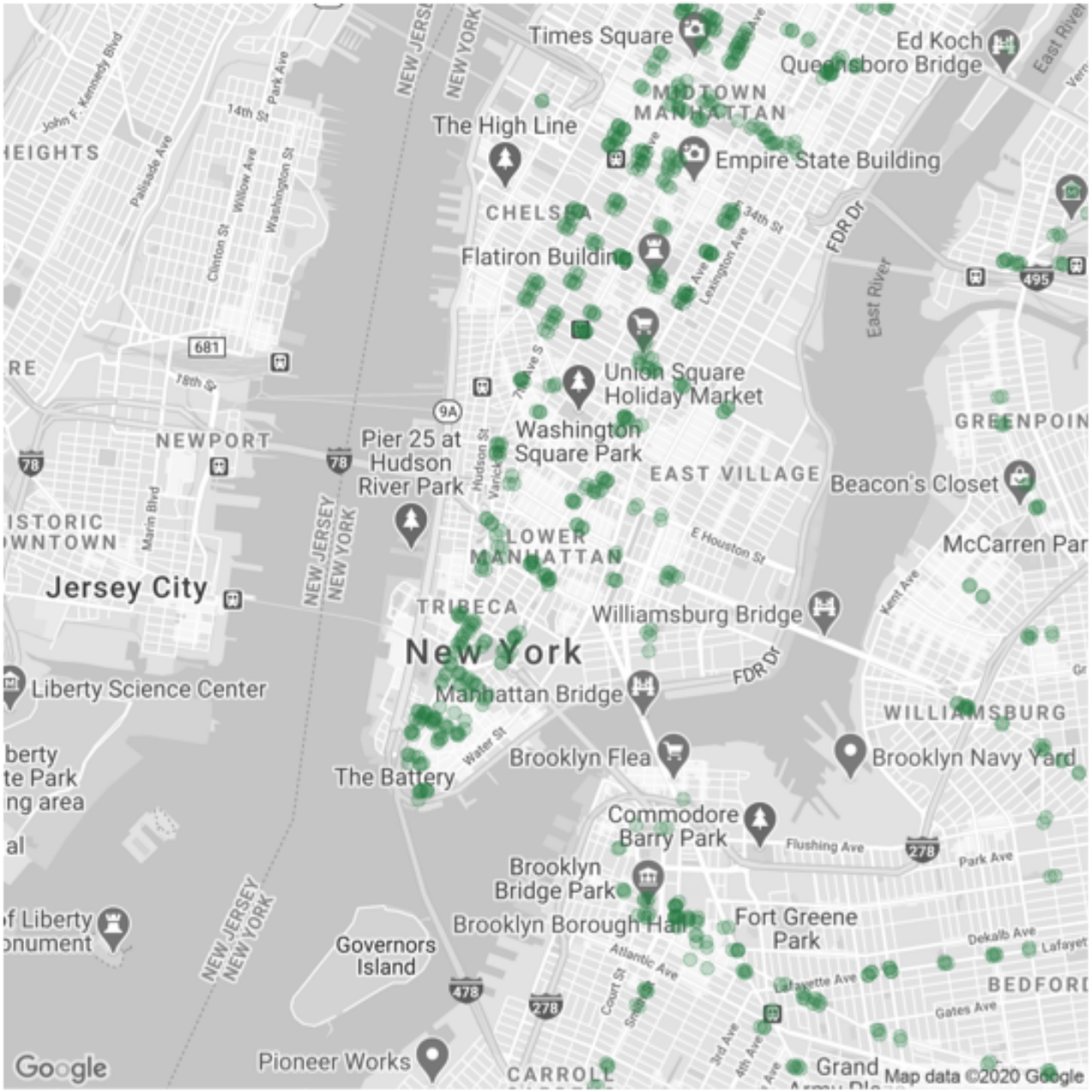
More than the "where"

Spatial data contain more than just location. Additional attributes can be tied to the observation based on its location.

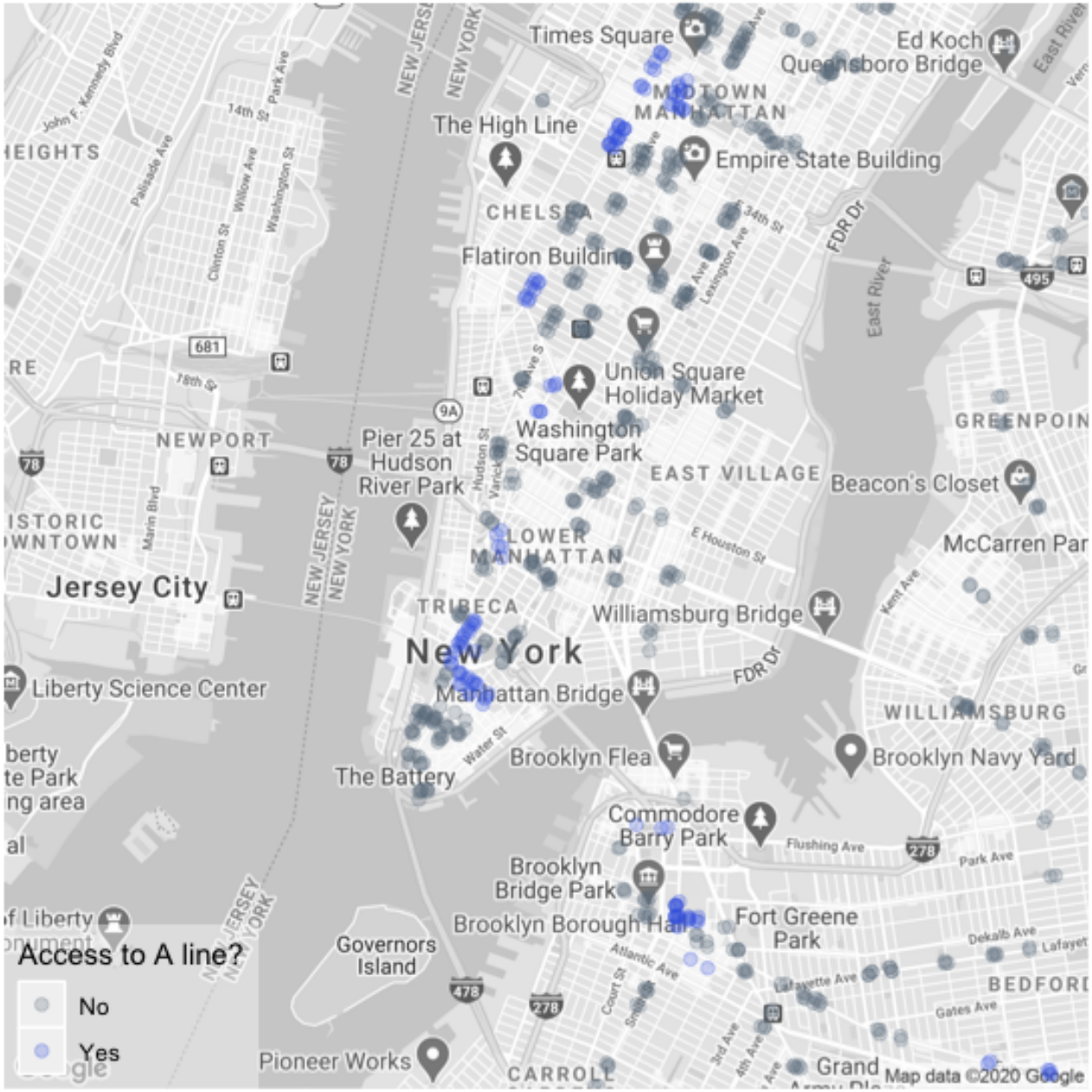
Example: New York City subway entrances.

[1] Data source: City of New York Subway Entrances

New York City subway entrances



New York City subway entrances



New York City subway entrances

Outline

Today's game plan:

1. Geocoding

2. Distance calculations

3. Map-making

Install/load packages

To follow along with my code as written, you will need to install and load the following `R` packages:

- `dplyr*`
- `ggplot2*`
- `ggmap`
- `geosphere`

* You'll need these from the start. The others will be formally introduced later.

Materials available on GitHub

Access the slides and supplementary materials on GitHub at `https://github.com/sarahlotspeich/geogRaphy`.

- Slides
- Datasets
- Maps
- Additional code

I recommend pausing the video to open the `geogRaphy.pdf` lecture file for clickable links.

Geocoding

Geocoding is the process of converting addresses (like a street address) into geographic coordinates (like latitude and longitude), which you can use to place markers on a map, or position the map.

[1] Google Maps Geocoding API

How it works

The algorithm begins by taking a complete street address and breaking it down into its component parts.

Example: 2525 WEST END AVE NASHVILLE TN 37203

- **Street number:** 2525
- **Street name:** WEST END
- **Street type:** AVE
- **Street suffix direction:**
- **City:** NASHVILLE
- **State:** TN
- **ZIP:** 37203

This address is then compared to a reference table that has already been mapped (e.g., GoogleMaps or TomTom).

[1] Blossom (2014), Geocoding Best Practices

Best practices

1. Remove suite/apartment numbers (they will simply create ties).
2. Be mindful of special characters like @, #, ?, etc.
3. Check for misspellings or abbreviations.

original	cleaned
310 25TH AVE S #103 NASHVILLE TN 37240	310 25TH AVE S NASHVILLE TN 37240
2301 VANDERBILT PLC NASHVILLE TN 37240	2301 VANDERBILT PL NASHVILLE TN 37240
2400 BLAKMORE AVE NASHVILLE TN 37212	2400 BLAKEMORE AVE NASHVILLE TN 37212
TWENTY 3RD AVE N NASHVILLE TN 37212	23RD AVE N NASHVILLE TN 37212

[1] Blossom (2014), Geocoding Best Practices

Level of accuracy

Often (but not always) geocoders can pinpoint exact property or building of the address input. When it cannot, a few things can happen:

- a match at a less precise level

  If 170 MAIN STREET, CLEVELAND, OH cannot be found in Cleveland, the address may be matched either with a 170 MAIN STREET in a nearby town or with the city center of CLEVELAND

- address range interpolation along a street

  170 MAIN STREET will be placed 70% of the way along the block of MAIN STREET that ranges from street numbers 100-200

Most geocoders give a level of confidence or matching. You can also confirm matches visually!

[1] Blossom (2014), Geocoding Best Practices

Example: Farmers Markets Directory and Geographic Data

State, address, name, and zip code of farmers markets in the United States, along with attributes such as payment methods accepted and types of goods for sale.

Downloaded from the US Department of Agriculture., but you can read it in from the GitHub: `/data/us_farmers_markets.csv"`.

Setting up your API key

Get a key

First, you'll need to register for a *free* Google Maps API key here: https://cloud.google.com/maps-platform/.

Register your key in R

```
ggmap::register_google(key = "YOUR KEY")
```

You'll need to call the `register_google()` command everytime you open a new session. If you want to set it permanently, instead call:

```
ggmap::register_google(key = "YOUR KEY", write = TRUE)
```

The ggmap package

A collection of functions to visualize spatial data and models on top of static maps from various online sources ... It includes tools common to those tasks, including functions for geolocation and routing.

The `geocode()` function takes inputs:

- `location`: a character vector of street addresses or place names
- `output`: amount of output
  - `output = "latlon` returns latitude, longitude
  - `output = "more"` returns same as `"latlon"` + accuracy measures `type` and `loctype`
- `source`: source of reference table to use (set `source = "google"`)

[1] D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161.

The `ggmap` package

A collection of functions to visualize spatial data and models on top of static maps from various online sources ... It includes tools common to those tasks, including functions for geolocation and routing.

If you want to add latitude/longitude columns to your existing `farm` dataset, you can use `ggmap::mutate_geocode()`. It takes the address column as an input (`location = address`).

[1] D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161.

Checking accuracy of `ggmap::geocode()`

The `type` and `loctype` outputs tell us about the geocoded match. Let's look at a few examples...

Less precise matches:

- There were a couple of addresses that only gave city, state, and zip. These yielded `approximate` matches.
- Others gave street, city, state, and zip, which were matched with `geometric_center`.

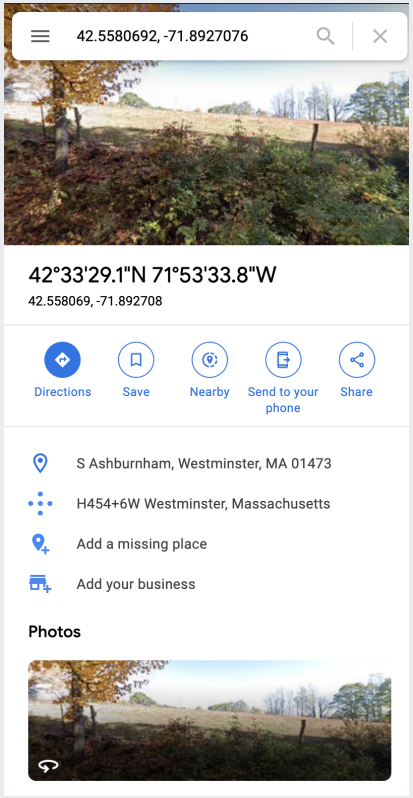
More precise matches:

- Ideally, addresses would be either `rooftop` level matched or `range_interpolated`.

Manual check

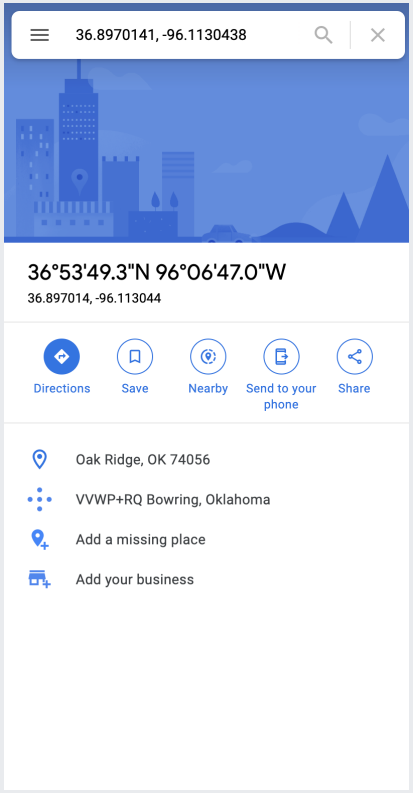
Address: Westminster, MA, 01473

Geocoded: (42.5580692, -71.8927076)



Address: Oak Ridge, OK, 74056

Geocoded: (36.8970141, -96.1130438)



Distance calculations

Great-circle distance calculations

"The **great-circle distance**... is the shortest distance between two points on the surface of a sphere..." (Wikipedia)

Given two `(lat, lon)` coordinates, there are two common options here:

1. Spherical Law of Cosines
2. Haversine Formula

Both calculations have been implemented in the `R` package `geosphere`.

[1] Great-circle distance calculations in R

The `geosphere` package

Spherical trigonometry for geographic applications. That is, compute distances and related measures for angular (longitude/latitude) locations.

Functions `distHaversine()` and `distCosine()` take the same inputs:

- `p1` and `p2`: longitude/latitude of points (in that order!)
- `r`: radius of the earth (set `= 3958.8` for output in miles and `= 6378137` for output in meters)

and give individual distance calculations.

[1] R. Hijmans. Introduction to the "geosphere" package

The `geosphere` package

Spherical trigonometry for geographic applications. That is, compute distances and related measures for angular (longitude/latitude) locations.

The `distm()` function takes two sets of points (rather than two points) and returns a matrix of the distances between the sets.

- `x` and `y`: longitude/latitude of point(s). If `y` is null, `distm(x)` returns the pairwise distances between each point in `x`.
- `fun`: which function to use to compute distances (set `fun = distHaversine` or `fun = distCosine`)

Note: Since `distm()` does not have a `r` input, the output will always be in meters by default.

[1] R. Hijmans. Introduction to the "geosphere" package

The `geosphere` package

Spherical trigonometry for geographic applications. That is, compute distances and related measures for angular (longitude/latitude) locations.

The `distm()` function takes two sets of points (rather than two points) and returns a matrix of the distances between the sets.

- `x` and `y`: longitude/latitude of point(s). If `y` is null, `distm(x)` returns the pairwise distances between each point in `x`.
- `fun`: which function to use to compute distances (set `fun = distHaversine` or `fun = distCosine`)

Note: Since `distm()` does not have a `r` input, the output will always be in meters by default.

This function is really quick, even when `x` and/or `y` are of larger dimensions.

[1] R. Hijmans. Introduction to the "geosphere" package

Defining "distance"

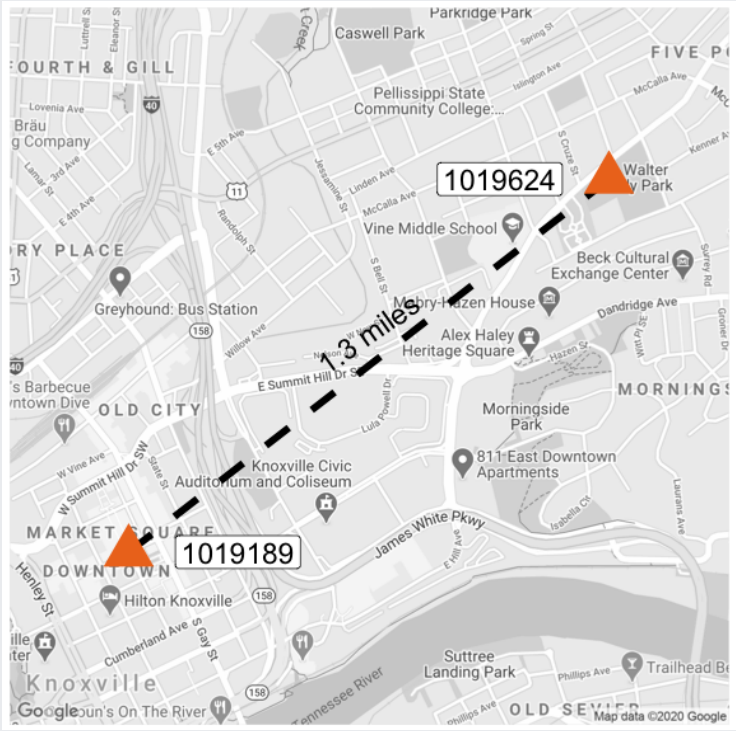
The **great-circle distance** measures assume the shortest possible route between two points, accounting for the approximate curvature of the earth.

This ignores things like:

- Roads
- Traffic
- Other real-world obstacles

Alternatives obtainable using `ggmap`:

1. Route-based travel distance
2. Estimated travel time



Calculating routes in ggmap

The `route()` function uses Google Maps API to calculate a path between two points. Key inputs include:

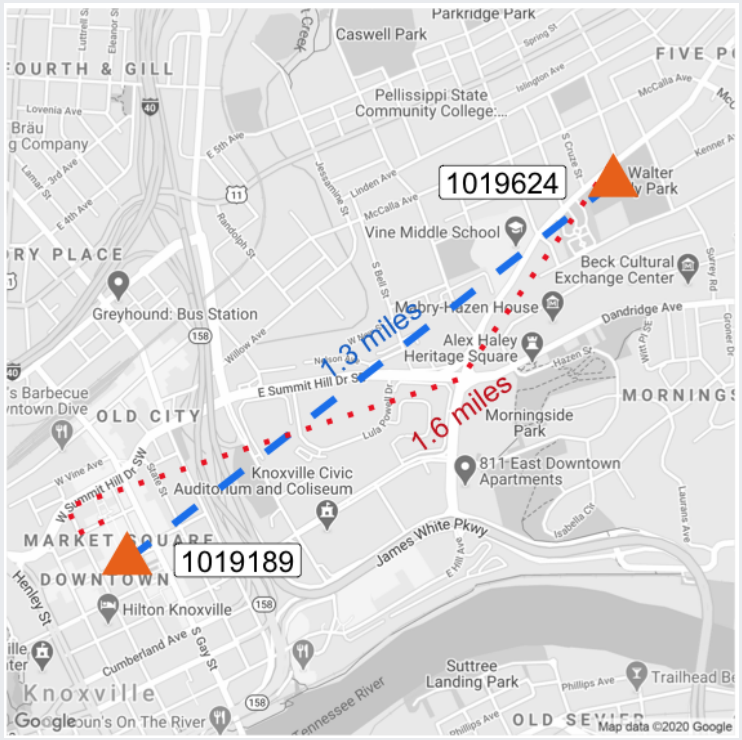
- `from`/`to`: addresses of the start/end point of the route
  - need to be a character inputs in the form `"latitude, longitude"` or addresses like `"street, city, state, zip"`
- `mode`: mode of transportation (options include `"driving"`, `"bicycling"`, `"walking"`, or `"transit"`)
- `structure`: structure of the output (options include `"legs"` and `"route"`)
  - if you want to map it, you'll want `structure = "route"`

[1] D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161.

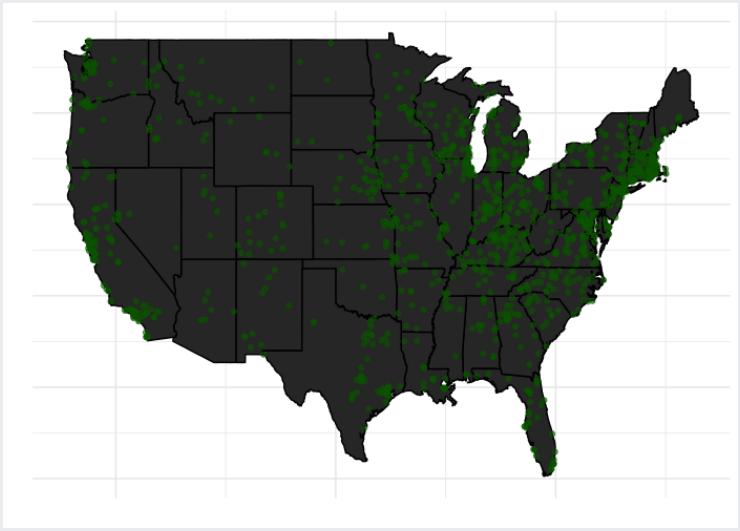
Revisit two Knoxville farmers markets

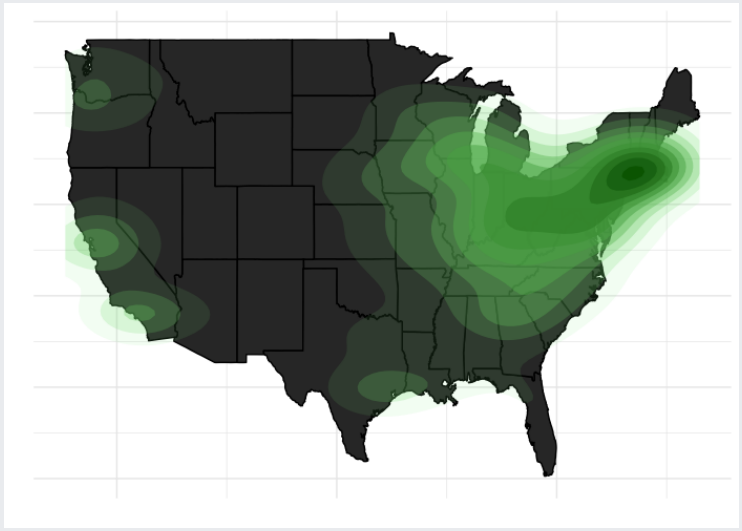
While the great-circle distance indicated that these farmers markets are **1.3 miles apart**, the true travel distance was found to be slightly farther at **1.6 miles** apart. This drive is estimated to take you **6 minutes**.

In fact, the great-circle distance was closer to the walking distance calculated using `ggmap`: **1.5 miles**.

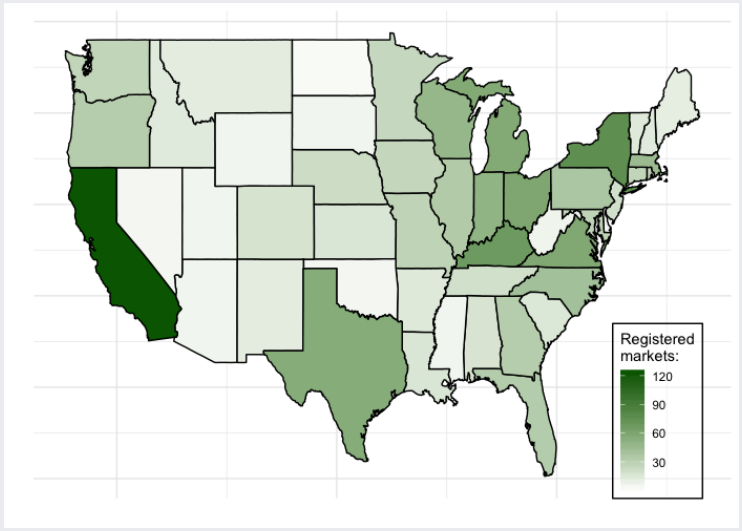


Map-making





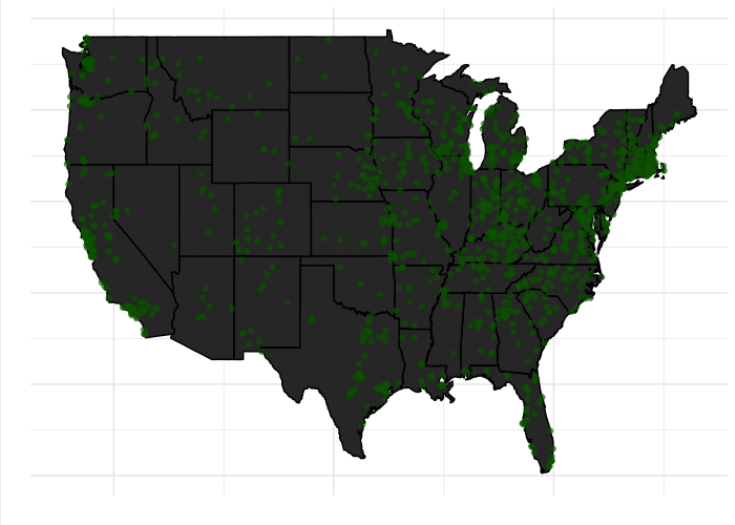
Popular map types



Mapping point data

Data consist of **points/coordinates** (lat, long) of distinct locations.

- Examples: home addresses
- Point can have attributes tied to them.
- Attributes can be incorporated using coloring or shapes.



In the US Farmers Market Registry, we can plot the locations of each individual market based on the geocoded latitude and longitude. This is pictured above.

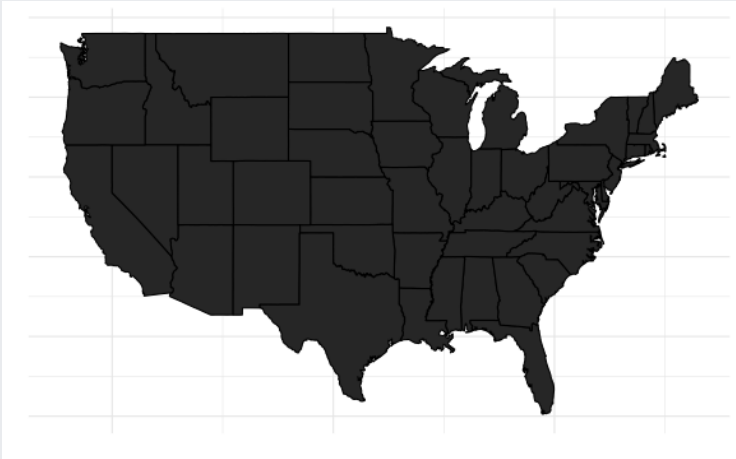
Let's walk through how to build it...

We say (lat, long), but we *plot* (long, lat)

In other words, `x = long` and `y = lat`

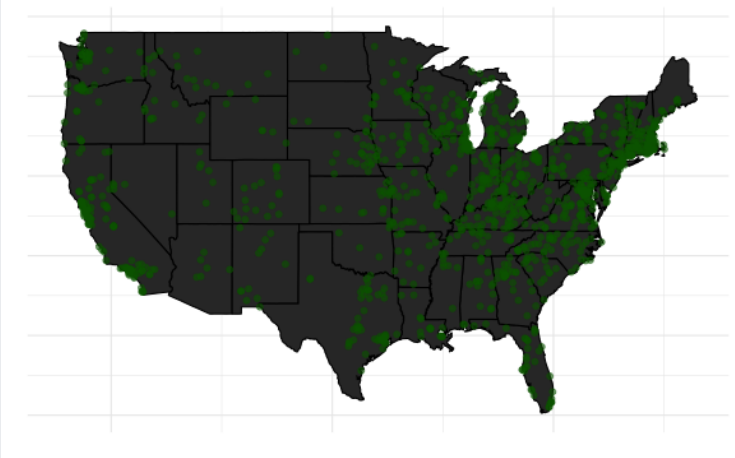
Start with a map of the US

Each pair of coordinates (`lat`, `long`) denotes a point along the polygon's outline. The `order` variable tells it which order in which to connect the points. The `group` variable (in this case the state) specifies individual polygons.



Layering point data

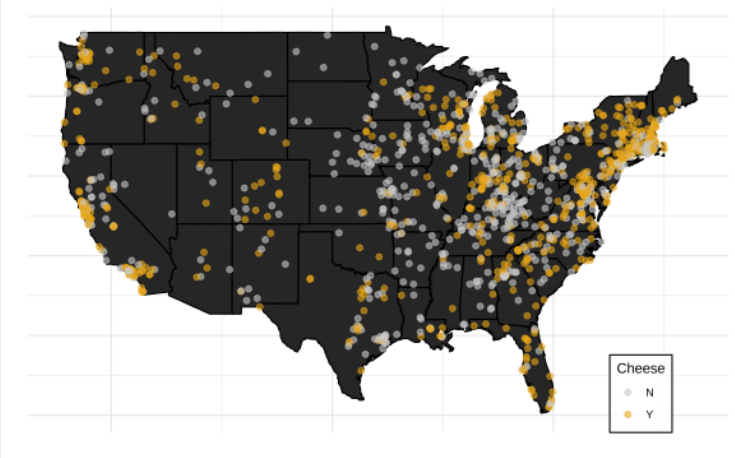
We can overlay the basic map with the latitude and longitude of the markets using `geom_point()` in `ggplot2`. Recall that in the aesthetic, `x = lon` and `y = lat`.



Adding point attributes

The `farm` data include additional indicators on the offerings at each market. We can include this in the map by coloring the points by whether or not the market offers a specific product.

Let's look at national `Cheese` offerings.



Alternative background maps using `ggmap`

Instead of the plain polygon map background, we can also use the `ggmap::get_map()` function to build upon a variety of more detailed views. Some important inputs:

- `location`: a string address, (`long`, `lat`) pair, or (`left`, `bottom`, `right`, `top`) coordinates of the bounding box
- `zoom`: map zoom (integer values) where `zoom = 3` is for continent, `zoom = 21` is building level, and the default is `zoom = 10` for city.
- `source`: where to pull the map from, options include Google Maps (`"google"`), OpenStreetMap (`"osm"`), and Stamen Map (`"stamen"`)
- `color`: options are `"color"` or black-and-white (`"bw"`)

[1] D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161.

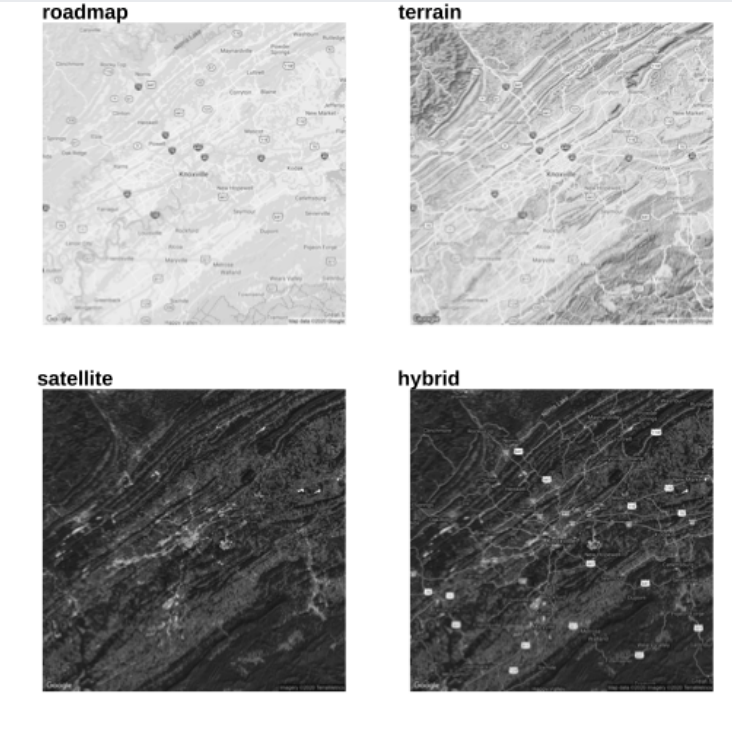
Alternative background maps using `ggmap`

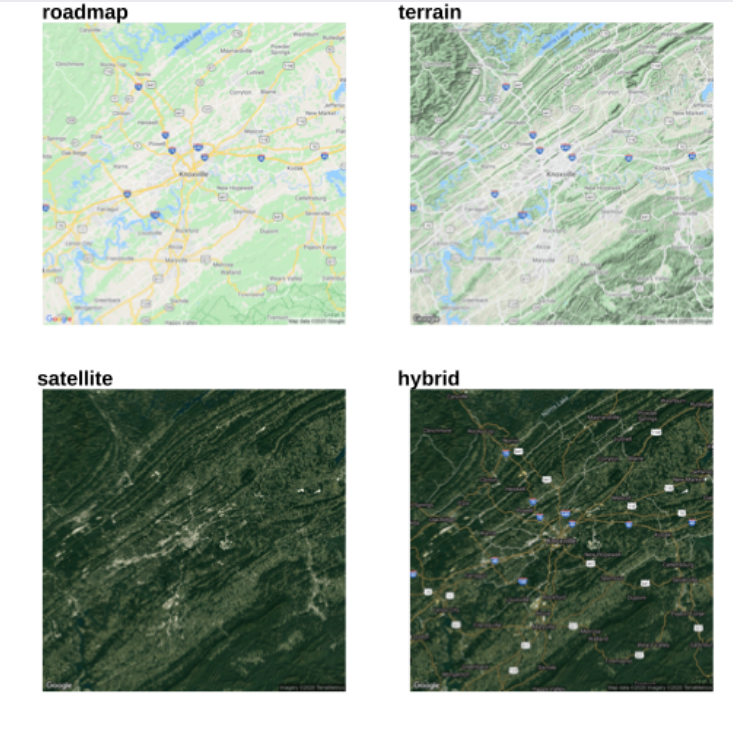
Instead of the plain polygon map background, we can also use the `ggmap::get_map()` function to build upon a variety of more detailed views. Some important inputs (continued):

- `maptype`: map theme/aesthetic! Different `source` inputs have different options...
  - when `source = "google"`: `"terrain"`, `"terrain-background"`, `"satellite"`, `"roadmap"`, and `"hybrid"`
  - when `source = "stamen"`: `"terrain"`, `"watercolor"`, and `"toner"`

[1] D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161.

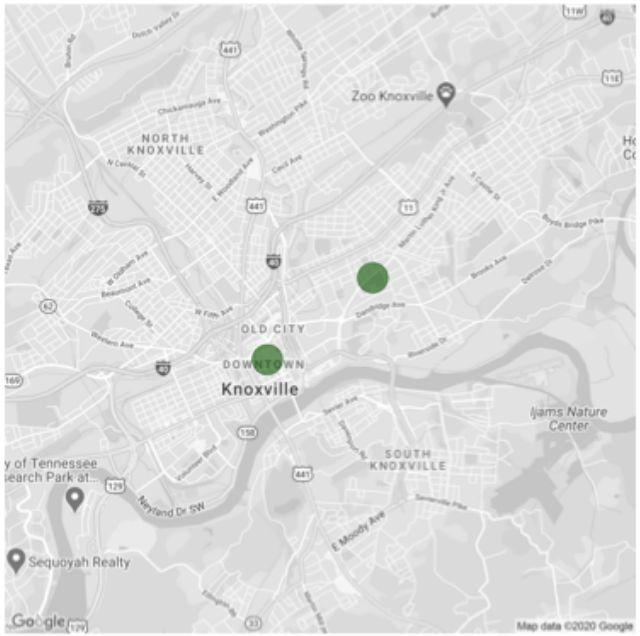
Map types from Google Maps



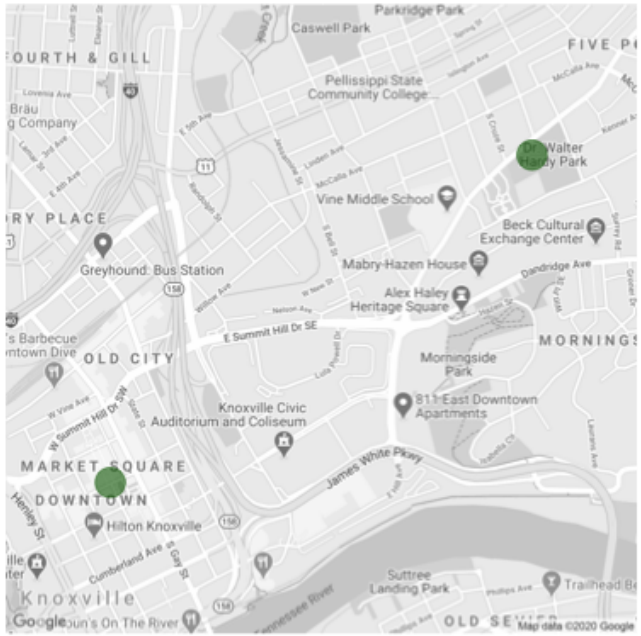


Finetuning it with `zoom`

zoom = 13



zoom = 15



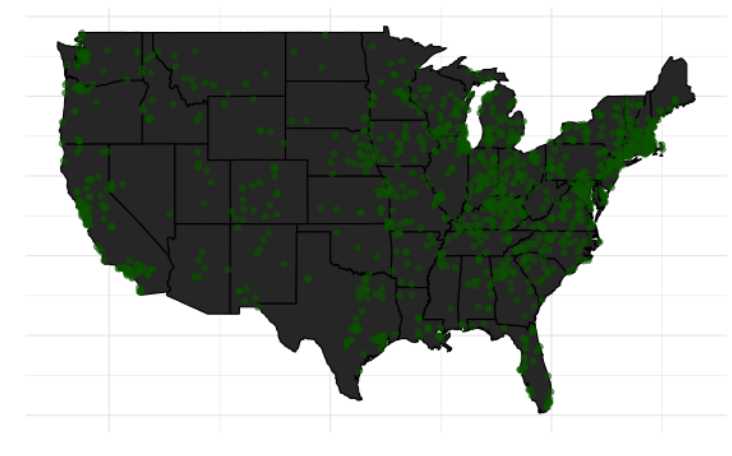
Want to find the balance between zooming too far (and losing context or your data) and being too far away to see details like roads.

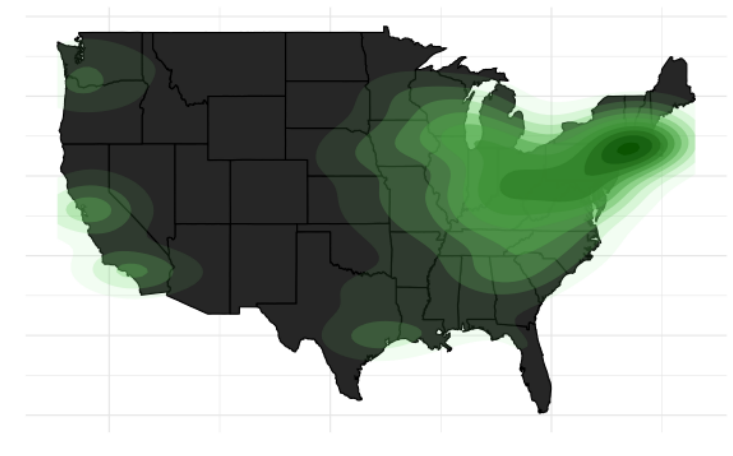
Incorporating `route` data

Layer the `route` over a map using `geom_path()` based on the `lat` and `lon` coordinates.

Density maps

Often, we are less interested in individual locations of each point in our dataset. More so, we inspect potential **clusters** or areas of high concentration.



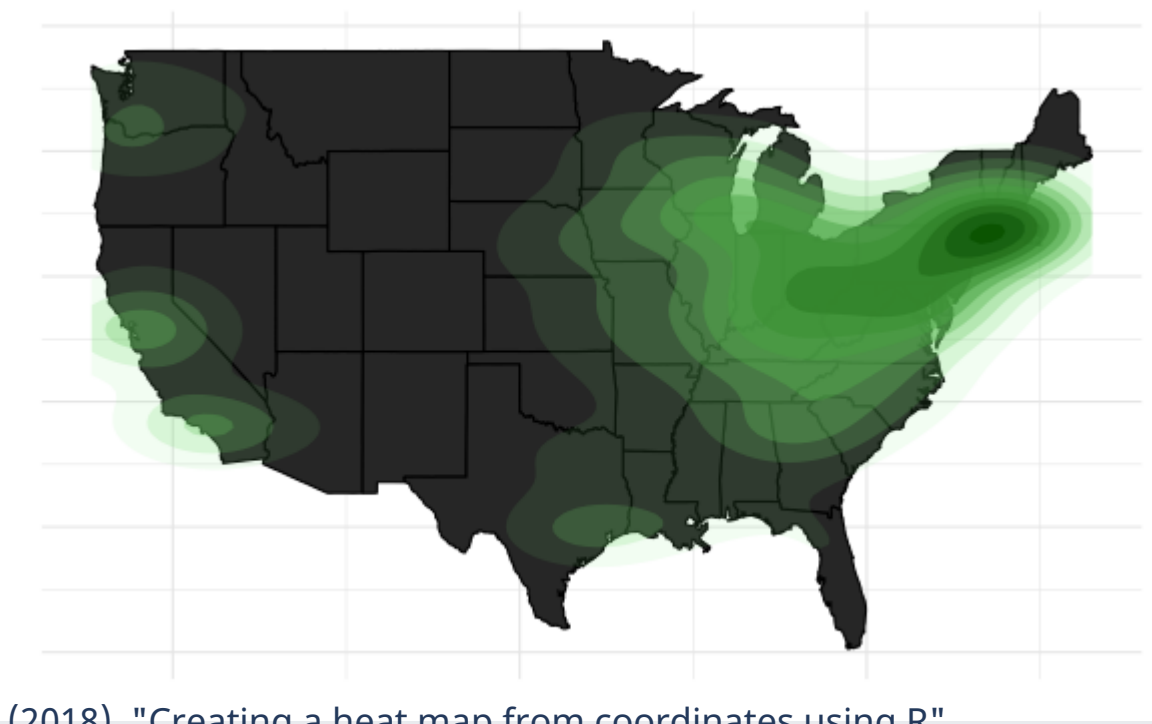


This can be done in `ggplot2` using the `stat_density2d()` function.

[1] A Holder (2018), "Creating a heat map from coordinates using R"

Density maps

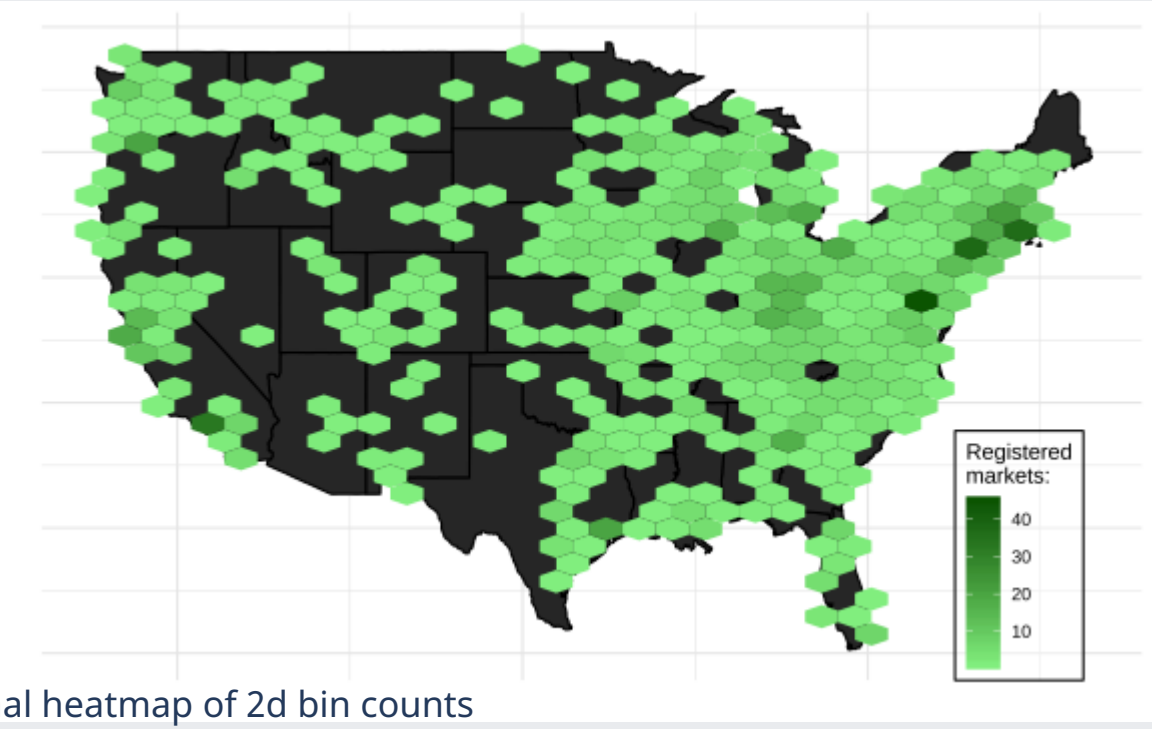
The `state_density2d()` function is effectively *smoothing* over the (`lat`, `lon`) coordinates of the markets.



[1] A Holder (2018), "Creating a heat map from coordinates using R"

Hexagon maps

The `geom_hex()` function... *"divides the plane into regular hexagons, counts the number of cases in each hexagon, and then (by default) maps the number of cases to the hexagon fill."* This is **raster** data representation, divided into uniform hexagons.



[1] Hexagonal heatmap of 2d bin counts

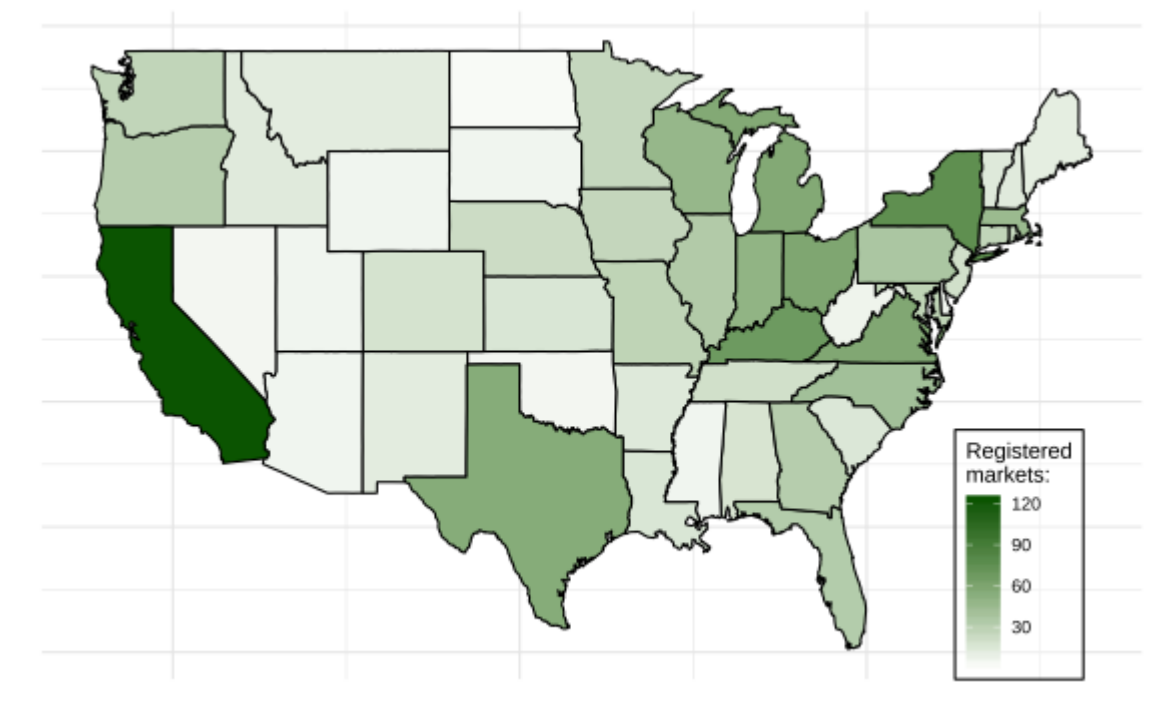
Aggregating point data across space

There are other ways to aggregate point. For example, we might be interested in the **state- or county-level numbers of markets**. This requires two alterations to our `flt_farm` data:

1. Use `dplyr::group_by()` to sum the number of markets by state
2. Use `dplyr::right_join()` to merge the market counts with the US map data

Maps like this are called **choropleth** maps.

State-level counts of farmers markets



Maps can be misleading

California stands out on our state-level market counts, but not on any of the previous maps. In fact, there are 123 registered California farmers markets (the largest of any state, followed by New York with 76).

This could potentially be explained by:

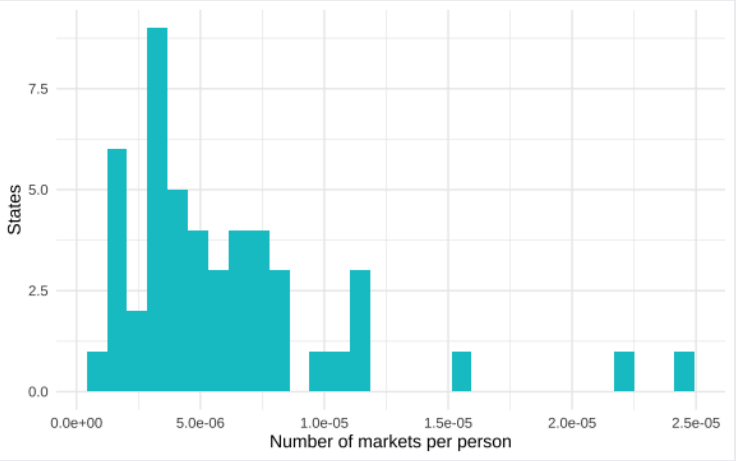
1. larger land coverage
2. greater population

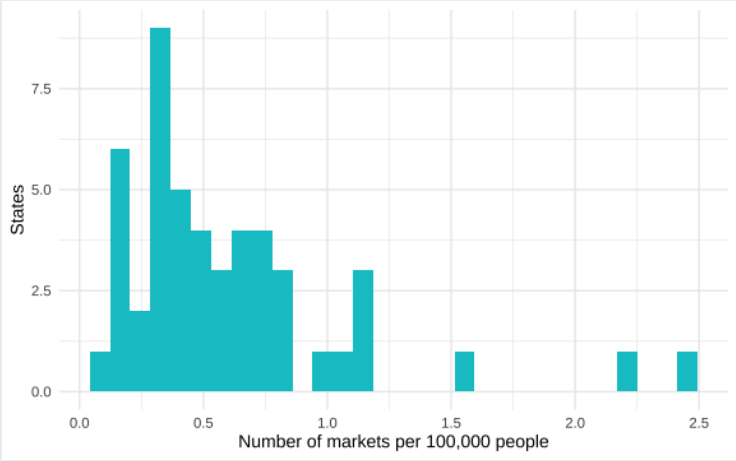
rather than a stronger propensity toward homegrown goods.

For these reasons, it is often advisable to not simply map "crude" counts and measures, because you can get misleading hotspots that are more related to city centers. To even the playing field, one way to correct for this is to look at per capita counts and measures (or perhaps per mile, etc).

State-level counts of farmers markets *(standardized)*

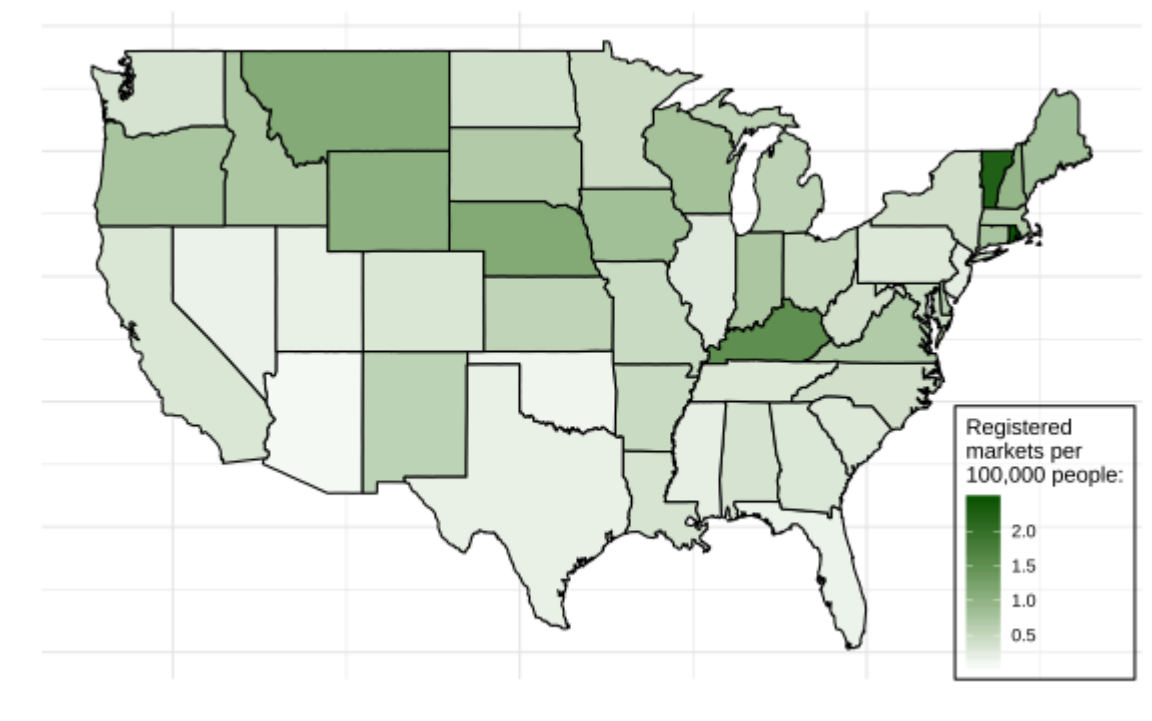
We merge in state population estimates(US Census Bureau) to map state-level rates of farmers markets *per capita* for comparison. Read `pop` in from the `GitHub`.





Color the choropleth by the number of markets per 100,000 people (rather than per person) since it still allows for **fair comparison between states**, but it is easier to see differences between them since the numbers are larger.

State-level counts of farmers markets *(per 100,000 people)*



References

“What Is Spatial Data? The Basics & GIS Examples: FME.” Safe Software, www.safe.com/what-is/spatial-data/.

S. Firmin. “Vector and Raster: A Tale of Two Spatial Data Types.” Velocity Business Solutions Limited, 15 Jan. 2019, www.vebuso.com/2019/01/vector-raster-tale-two-spatial-data-types/.

“Geospatial Data Models.” Humboldt State Geospatial Online, gsp.humboldt.edu/OLM/Courses/GSP_216_Online/lesson3-1/data-models.html.

“Subway Entrances: Map of NYC Subway Entrances.” NYC Open Data, data.cityofnewyork.us/Transportation/Subway-Entrances/drex-xx56.

“Get Started.” Google Maps Platform, Google, developers.google.com/maps/documentation/geocoding/start.

J. Blossom. “Geocoding Best Practices.” Center for Geographic Analysis, Harvard University, 2 Mar. 2015, gis2.harvard.edu/services/blog/geocoding-best-practices.

References (cont.)

D. Kahle and H. Wickham. "ggmap: Spatial Visualization with ggplot2." The R Journal, 5(1), 144-161.

“Farmers Markets Directory and Geographic Data.” Data.gov, Publisher Agricultural Marketing Service, Department of Agriculture, 21 Feb. 2020, catalog.data.gov/dataset/farmers-markets-directory-and-geographic-data.

M. Pineda-Krch. “Great-Circle Distance Calculations in R.” R Bloggers, 12 May 2011, www.r-bloggers.com/2010/11/great-circle-distance-calculations-in-r/.

R. J. Hijmans. "Introduction to the 'geosphere' package (Version 1.5-10)." 25 May 2019, https://cran.r-project.org/web/packages/geosphere/vignettes/geosphere.pdf.

Hodler, Axel. “Creating a Heat Map from Coordinates Using R.” Axel Hodler, 20 Dec. 2018, axelhodler.medium.com/creating-a-heat-map-from-coordinates-using-r-780db4901075.