Intro to Spatial Data

Ciencia de Datos para la toma de decisiones en Economía

Ignacio Sarmiento-Barbieri

Universidad de los Andes

September 10, 2022

Agenda

Motivation

- In Big Data volume was only a part of the story
- Big Data are data of high complexity: anarchic and spontaneous
- They are the by product of an action: pay with credit card, tweet, move from point A to point B, buy a house, etc.
- Now we are going to center on spatial data

Types of Spatial Data

Spatial data comes in many “shapes” and “sizes”, the most common types of spatial data are:

- Points are the most basic form of spatial data. Denotes a single point location, such as cities, a GPS reading or any other discrete object defined in space.
- Lines are a set of ordered points, connected by straight line segments
- Polygons denote an area, and can be thought as a sequence of connected points, where the first point is the same as the last
- Grid (Raster) are a collection of points or rectangular cells, organized in a regular lattice

Types of Spatial Data: Points

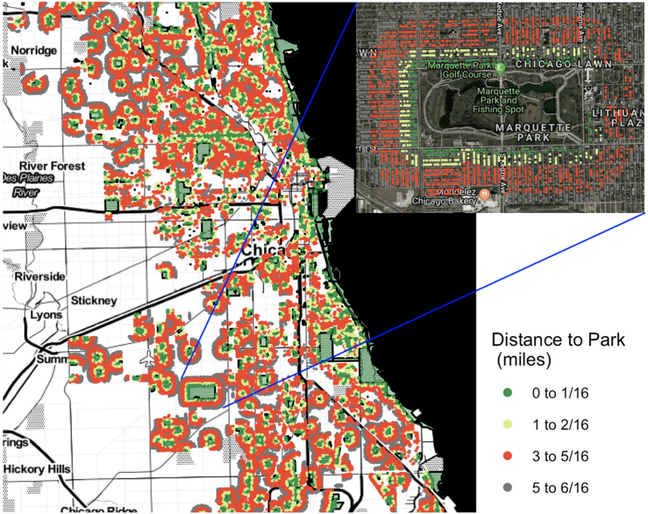


Fig. 1. Housing transactions around parks: neighborhood distance intervals. Notes: The following figure shows transactions within 3/8 miles of the nearest park in Chicago. The

Types of Spatial Data: Lines

D. McMillen, I. Sarmiento-Barbieri and R. Singh — Journal of Urban Economics 110 (2019) 1–25

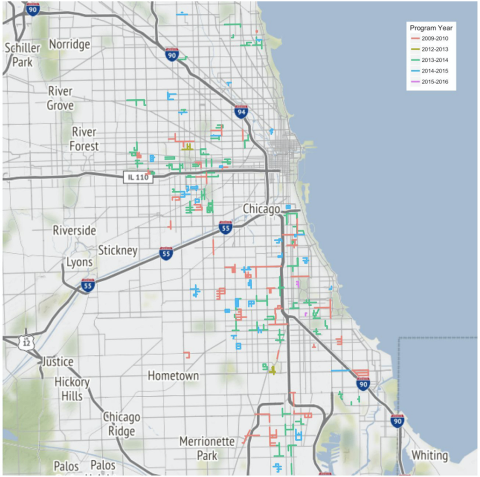
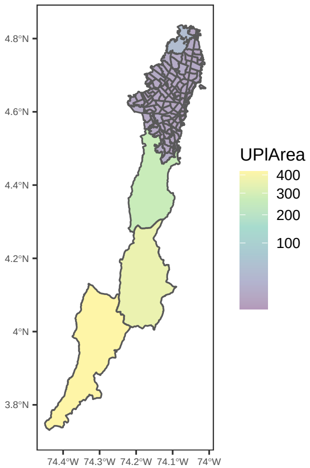


Fig. 1. Safe Passage Routes, by year of program adoption.
Note: Shapefiles with Safe Passage shape and location where obtained from the Chicago Data Portal and year that the program was launched at each location through a FOIA request.

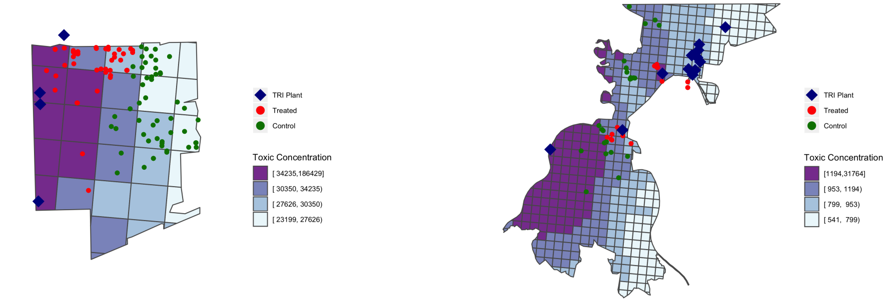
McMillen, Sarmiento-Barbieri & Singh, 2019

Types of Spatial Data: Polygons



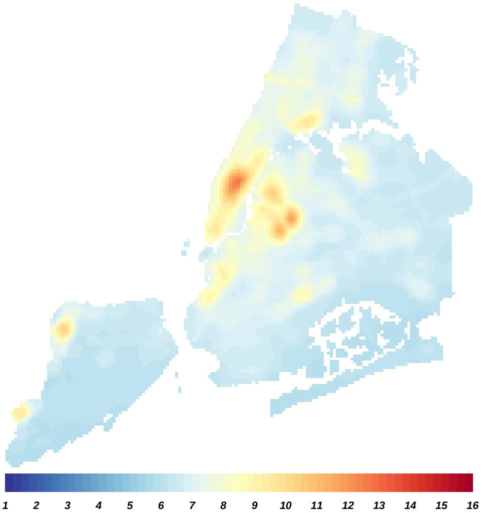
Source:https://datosabiertos.bogota.gov.co/dataset/unidad-de-planeamiento-bogota-d-c

Types of Spatial Data: Combination



Christensen,Sarmiento-Barbieri & Timmins (2022)

Types of Spatial Data: Rasters



Source:https://data.cityofnewyork.us/Environment/NYCCAS-Air-Pollution-Rasters/q68s-8qxv

Reading and Mapping spatial data in R

- Spatial data in various formats.
- One of the most used format are `shapefiles`
- This type of files stores non topological geometry and attribute information for the spatial features in a data set
- Files
  - Main file: file.shp
  - Index file: file.shx
  - dBASE table: file.dbf
  - Other files
    - file.prj
    - file.sbn , file.sbx
    - file.shp.xml
- Today I'm going to use data from `https://datosabiertos.bogota.gov.co`

Reading and Mapping spatial data in R

- Another one of the most used format are `json` o `geojson`

```
{
  "type": "Feature",
  "geometry": {
    "type": "Point",
    "coordinates": [-74.066391, 4.601590]
  },
  "properties": {
    "name": "Universidad de Los Andes"
  }
}
```

Reading shapefiles in R

- Basic Packages
  - Read and handle spatial data

```
require("sf")
```

  - Plotting and data wrangling

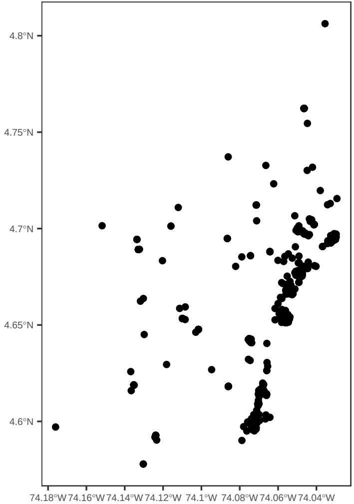
```
require("ggplot2")
require("dplyr")
```

```
bars<-st_read("egba/EGBa.shp")
```

```
## Reading layer ‘EGBa’ from data source ‘egba/EGBa.shp’ using driver
## ‘ESRI Shapefile’
## Simple feature collection with 515 features and 7 fields
## geometry type:  POINT
## dimension:      XY
## bbox:           xmin: -74.17607 ymin: 4.577897 xmax: -74.02929 ymax: 4.806253
## CRS:            4686
```

Visualizing Points

```r
ggplot()+
  geom_sf(data=bars) +
  theme_bw() +
  theme(axis.title =element_blank(),
panel.grid.major = element_blank(),
panel.grid.minor = element_blank(),
axis.text = element_text(size=6))
```



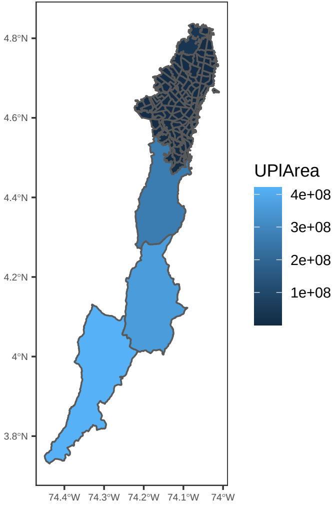
Visualizing Lines

```r
ciclovias<-read_sf("Ciclovia/Ciclovia.shp")
ggplot()+
  geom_sf(data=ciclovias) +
  theme_bw() +
  theme(axis.title =element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.text = element_text(size=6))
```

Visualizing Polygons

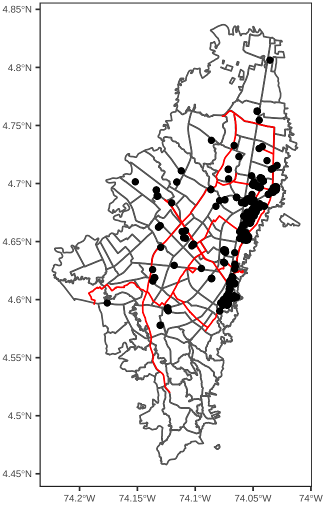
```
upla<-read_sf("upla/UPla.shp")

ggplot()+
  geom_sf(data=upla, aes(fill = UPlArea)) +
  theme_bw() +
  theme(axis.title =element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.text = element_text(size=6))
```



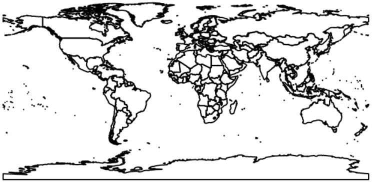
Visualizing Points, Lines, and Polygons

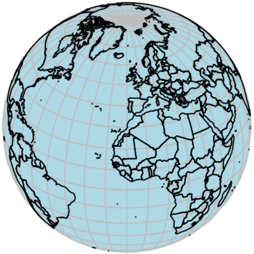
```r
ggplot()+
  geom_sf(data=upla
%>% filter(grepl("RIO",UPlNombre)==FALSE),
fill = NA) +
  geom_sf(data=ciclovias, col="red") +
  geom_sf(data=bars) +
  theme_bw() +
  theme(axis.title =element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.text = element_text(size=6))
```



The earth ain't flat

- The world is an irregularly shaped ellipsoid, but plotting devices are flat
- But if you want to show it on a flat map you need a map projection,
- This will determine how to transform and distort latitudes and longitudes to preserve some of the map properties: area, shape, distance, direction or bearing





The earth ain't flat

- For example, sailors use Mercator projection where meridians and parallels cross each other always at the same 90 degrees angle.
- It allows to easy locate your self on the line showing direction in which you sail
- But the projection do not preserve distances

Source: https://www.geoawesomeness.com/all-map-projections-in-compared-and-visualized/

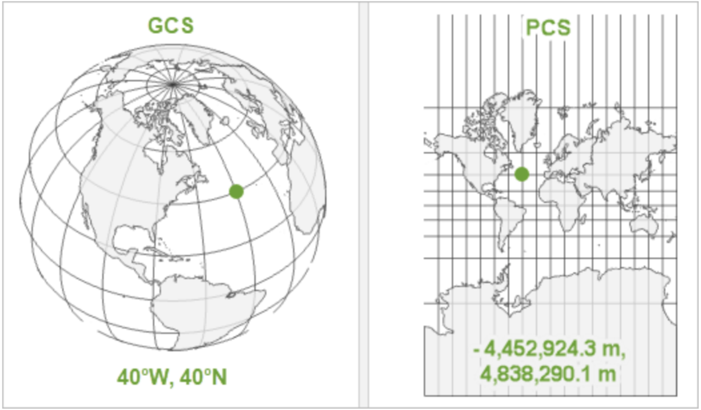
Which projection should I choose?

- "There exist no all-purpose projections, all involve distortion when far from the center of the specified frame" (Bivand, Pebesma, and Gómez-Rubio 2013)
- In some cases, it is not something that we are free to decide: "often the choice of projection is made by a public mapping agency" (Bivand, Pebesma, and Gómez-Rubio 2013).
- This means that when working with local data sources, it is likely preferable to work with the CRS in which the data was provided.

Which projection should I choose?

- Geographic coordinate systems: coordinate systems that span the entire globe (e.g. latitude / longitude).
  - For geographic CRSs, the answer is often WGS84
  - WGS84 is the most common CRS in the world, EPSG code: 4326.
  - For Bogotá the IGAC promotes the adoption of MAGNA-SIRGAS. EPSG code: 4626
- Projected coordinate systems: coordinate systems that are localized to minimize visual distortion in a particular region (e.g. Robinson, UTM, State Plane)

Which projection should I choose?

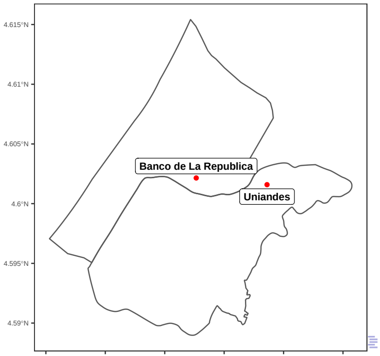


Source: https://www.geoawesomeness.com/all-map-projections-in-compared-and-visualized/

Creating Spatial Objects

```
db<-data.frame(place=c("Uniandes","Banco de La Republica"),
               lat=c(4.601590,4.602151),
               long=c(-74.066391,-74.072350),
               nudge_y=c(-0.001,0.001))
db<-db %>% mutate(latp=lat,longp=long)
db<-st_as_sf(db,coords=c('longp','latp'),crs=4326)
```

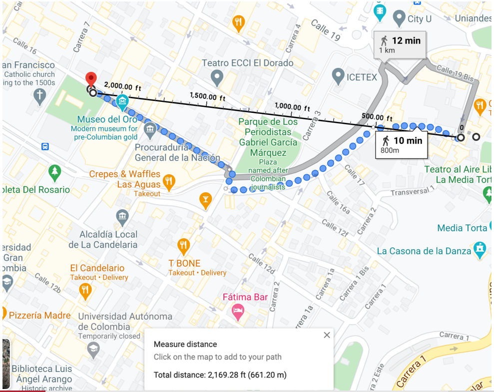
```
ggplot()+
  geom_sf(data=upla
        %>% filter(UPlNombre
        %in%c("LA CANDELARIA","LAS NIEVES")), fill = NA) +
  geom_sf(data=db, col="red") +
  geom_label(data = db, aes(x = long, y = lat,
                              label = place),
        size = 3, col = "black", fontface = "bold",
        nudge_y =db$nudge_y) +
  theme_bw() +
  theme(axis.title =element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.text = element_text(size=6))
```



Measuring Distances

```r
st_distance(db)
```

```
## Units: [m]
##           [,1]     [,2]
## [1,]    0.0000 664.1323
## [2,] 664.1323   0.0000
```



Measuring Distances

```
st_distance(db,ciclovias)
Error in st_distance(db, ciclovias) : st_crs(x) == st_crs(y) is not TRUE
st_crs(ciclovias)
```

```
## Coordinate Reference System:
##   User input: 3857
##   wkt:
## PROJCS["WGS 84 / Pseudo-Mercator",
##     GEOGCS["WGS 84",
##         DATUM["WGS_1984",
##             SPHEROID["WGS 84",6378137,298.257223563,
##                 AUTHORITY["EPSG","7030"]],
##             AUTHORITY["EPSG","6326"]],
##         PRIMEM["Greenwich",0,
##             AUTHORITY["EPSG","8901"]],
##         UNIT["degree",0.0174532925199433,
##             AUTHORITY["EPSG","9122"]],
##         AUTHORITY["EPSG","4326"]],
##     PROJECTION["Mercator_1SP"],
##     PARAMETER["central_meridian",0],
##     PARAMETER["scale_factor",1],
##     PARAMETER["false_easting",0],
##     PARAMETER["false_northing",0],
##     UNIT["metre",1,
##         AUTHORITY["EPSG","9001"]],
##     AXIS["X",EAST],
##     AXIS["Y",NORTH],
##     EXTENSION["PROJ4","+proj=merc +a=6378137 +b=6378137 +lat_ts=0.0 +lon_0=0.0 +x_0=0.0 +y_0=0 +k=1.0 +units=m +nadgrids=@null +wktext +n
##     AUTHORITY["EPSG","3857"]]
```

Measuring Distances

```r
ciclovias<-st_transform(ciclovias, 4686)
st_crs(ciclovias)
```

```
## Coordinate Reference System:
##   User input: EPSG:4686
##   wkt:
## GEOGCS["MAGNA-SIRGAS",
##     DATUM["Marco_Geocentrico_Nacional_de_Referencia",
##         SPHEROID["GRS 1980",6378137,298.257222101,
##             AUTHORITY["EPSG","7019"]],
##         TOWGS84[0,0,0,0,0,0,0],
##         AUTHORITY["EPSG","6686"]],
##     PRIMEM["Greenwich",0,
##         AUTHORITY["EPSG","8901"]],
##     UNIT["degree",0.0174532925199433,
##         AUTHORITY["EPSG","9122"]],
##     AUTHORITY["EPSG","4686"]]
```

Measuring Distances

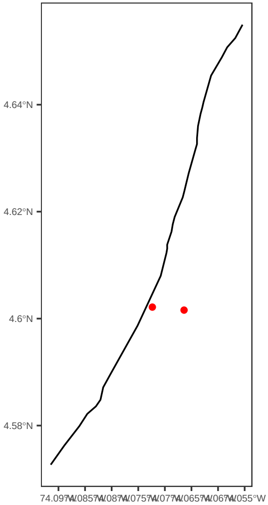
```r
db<-st_transform(db, 4686)
st_distance(db,ciclovias)
```

```
## Units: [m]
##          [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]     [,8]
## [1,] 9514.617 10789.90 6035.283 12855.90 6025.017 8311.922 4579.450 741.6047
## [2,] 9221.998 10686.39 6143.960 13004.84 5871.073 7656.183 4014.993 116.5939
##           [,9]    [,10]    [,11]    [,12]    [,13]    [,14]
## [1,] 1002.8751 6255.692 2385.125 8402.580 8669.030 3788.265
## [2,]  981.1991 5839.565 2425.508 7738.774 8048.108 3436.819
```

Measuring Distances

```r
ciclovias_sp<-ciclovias[8,]

ggplot()+
  geom_sf(data=ciclovias[8,], fill = NA) +
  geom_sf(data=db, col="red") +
  theme_bw() +
  theme(axis.title =element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.text = element_text(size=6))
```



Further Readings

- Arbia, G. (2014). A primer for spatial econometrics with applications in R. Palgrave Macmillan.
- Albouy, D., Christensen, P., & Sarmiento-Barbieri, I. (2020). Unlocking amenities: Estimating public good complementarity. Journal of Public Economics, 182, 104110.
- Bivand, R. S., & Pebesma, E. J. (2020). Spatial Data Science https://keen-swartz-3146c4.netlify.app/ (Chapter 8)
- Bivand, R. S., Gómez-Rubio, V., & Pebesma, E. J. (2008). Applied spatial data analysis with R (Vol. 747248717, pp. 237-268). New York: Springer.
- Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. Science, 350(6264), 1073-1076.
- Christensen, P., Sarmiento-Barbieri, I., Timmins C. (2020). Housing Discrimination and the Pollution Exposure Gap in the United States. NBER WP No. 26805
- Lee, K., & Braithwaite, J. (2020). High-Resolution Poverty Maps in Sub-Saharan Africa. arXiv preprint arXiv:2009.00544.
- Lovelace, R., Nowosad, J., & Muenchow, J. (2019). Geocomputation with R. CRC Press. (Chapters 2 & 6)
- McMillen, D., Sarmiento-Barbieri, I., & Singh, R. (2019). Do more eyes on the street reduce Crime? Evidence from Chicago's safe passage program. Journal of urban economics, 110, 1-25.
- Wasser, L. GIS With R: Projected vs Geographic Coordinate Reference Systems https://www.earthdatascience.org/courses/earth-analytics/spatial-data-r/geographic-vs-projected-coordinate-reference-systems-UTM/ Last Access September 10, 2020