SWIB24: Redlining

Redlining, Health and the Environment

Reading on graphical summaries

- *OI Biostat*: Section 1.6 and earlier material on numerical and graphical summaries.

Overview

We will use data compiled in this tutorial available online to assess the effects of historical Redlining with modern day outcomes. https://github.com/Jayanth-Mani/Redlining_Data_Tutorial/tree/master.

Redlining

Redlining refers to the the 1930s practice by the Home Owners' Loan Corporation (HOLC), which drew lines on maps to grade loan security on an A-D scale. These maps were used to limit access to mortgages in predominantly African American and immigrant neighborhoods. Previous research has shown that formerly redlined neighborhoods have continuing modern-day impacts on health, environment such as worse air pollution and lower economic activity.

Please read this New York Times article https://www.nytimes.com/interactive/2020/08/24/climate/racism-redlining-cities-global-warming.html

Data

The compiled data provided merges information from several data sources listed below:

1. The American Community Census
2. University of Richmond HOLC dataset (2010 census)
3. Diversity Data Kids HOLC grading mapped to the 2010 census
4. EJScreen Census Tract level data
5. Center for Air, Climate and Energy Solutions (CACES) Air Pollution data
6. USDA Food Access (2019 Food Atlas)
7. Diversity Data Kids Child Opportunity Index
8. Opportunity Atlas
9. Open Park Area from the National Neighborhood Data Archive (NaNDA)

Some code to help you get started

1. The file we are reading is a geojson file which includes spatial information corresponding to census tracts in the US. We will use specific R packages to read this type of data and define some important variables that can help you in your analysis.

```
## load data
# install.packages("sf")
# install.packages("spdep")

library(sf)
```

```
## Linking to GEOS 3.10.2, GDAL 3.4.2, PROJ 8.2.1; sf_use_s2() is TRUE
```

```
library(spdep)
```

```
## Loading required package: spData

## To access larger datasets in this package, install the spDataLarge
## package with: `install.packages('spDataLarge',
## repos='https://nowosad.github.io/drat/', type='source')`
```

```
## it might take a little time to load
mydat <- st_read("final_df.geojson")
```

```
## Reading layer `final_df' from data source
##   `/Users/rbalasub/Library/CloudStorage/GoogleDrive-rbalasub@umass.edu/My Drive/Summer_Institute_
##   using driver `GeoJSON'
## Simple feature collection with 13425 features and 426 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -122.7752 ymin: 25.69572 xmax: -70.94519 ymax: 47.71767
## Geodetic CRS:  WGS 84
```

```
dim(mydat)
```

```
## [1] 13425   427
```

```
## Removing census tracts with invalid spatial coordinates
table(st_is_valid(mydat))
```

```
##
## FALSE  TRUE
##    62 13363
```

```
analysis.dat <- mydat[st_is_valid(mydat),]

## Defining key variables
## HOLC grade: A=best, D=worst

grade <- analysis.dat$class1
table(grade)
```

```
## grade
##    A    B    C    D
##  895 2751 6044 3673
```

```r
## Some key demographic variables at the census tract level
## These are demographics are from 2015 ACS survey, which releases
## aggregated data collected over the previous 5 year period from random
## samples of households across the US.

## total population
population <- analysis.dat$ACSTOTPOP

## median age
age <- analysis.dat$median_age

## % minorities
minoritypct  <- analysis.dat$minority_pct.x

## % unemployed
unemp <- analysis.dat$UNEMPPCT

## median family income
income <- analysis.dat$MedianFamilyIncome
income <- as.numeric(income)
```

```
## Warning: NAs introduced by coercion
```

Data on Food Deserts

The LILA variables (eg. LILATracts_1And10) are indicators for food deserts using a variety of definitions. For example, LILATracts_1And10 is a flag for food desert when considering low accessibilty at 1 and 10 miles.

Labels for all columns starting with "la.." can be found here: #https://data-dictionary.regenstrief.org/iadc/catalog/s#howDataset/IADC/4

```r
## LILA10

lila10 <- analysis.dat$LILATracts_1And10
```

Health outcomes

Census tract level summaries of specific health outcomes shown in the code below was obtained from https://dsl.richmond.edu/panorama/redlining/data

```r
## Life expectancy
le <- analysis.dat$life_exp

## Prevalence of high blood pressure
bp <- analysis.dat$highbp_pct

## Prevalence of cancer
cancer <- analysis.dat$cancer_pct
```

```r
## Prevalence of asthma
asthma <- analysis.dat$asthma_pct

## Prevalence of coronary heart diseae
chd <- analysis.dat$chd_pct

## Prevalence of COPD
copd <- analysis.dat$copd_pct

## and others in columns 20:24
```

Air pollution estimates from 2010

Air pollution estimates including carbon monoxide, nitrogen dioxide, ozone, PM10 and sulphur dioxide are from the Center for Air, Climate and Energy Solutions (CACES) Air Pollution (https://www.caces.us/data)

```r
# carbon monoxide
co <- analysis.dat$co

# nitrogen dioxide
no2 <- analysis.dat$no2

# ozone
o3 <- analysis.dat$o3

# PM10 fine particular matter
p10 <- analysis.dat$pm10

# And other pollutants
```

Green space

Data on open park lands is from the Open Park Area from the National Neighborhood Data Archive (https://www.openicpsr.org/openicpsr/project/117921/version/V1/view)

Variable names

There are 427 variables (columns) in this dataset. A guide to the variables and their source is in the excel spreadsheet Column_Labels.xlsx.

US States represented

Not all states are equally represented. If modern day census tracts were not mapped to HOLC areas, they won't be included in these data. If you would like to focus your analysis within a state, this will be helpful. You can see that California, New York and Illinois have data on more than 1000 census tracts

```
sort(table(analysis.dat$STATE_NAME))
```

```
##
##    Mississippi       Arkansas South Carolina  New Hampshire        Arizona
##             18             21             22             28             30
##  West Virginia           Utah       Oklahoma         Oregon   Rhode Island
##             55             71             93             93             93
##       Colorado         Kansas        Alabama North Carolina       Nebraska
##            100            103            114            117            127
##       Kentucky     Washington           Iowa        Georgia      Louisiana
##            133            139            142            145            173
##    Connecticut      Tennessee       Virginia      Minnesota        Florida
##            185            194            200            225            249
##       Maryland       Missouri      Wisconsin        Indiana  Massachusetts
##            255            276            321            358            393
##          Texas     New Jersey       Michigan           Ohio   Pennsylvania
##            433            718            755            792            811
##       Illinois     California       New York
##           1163           1777           2441
```

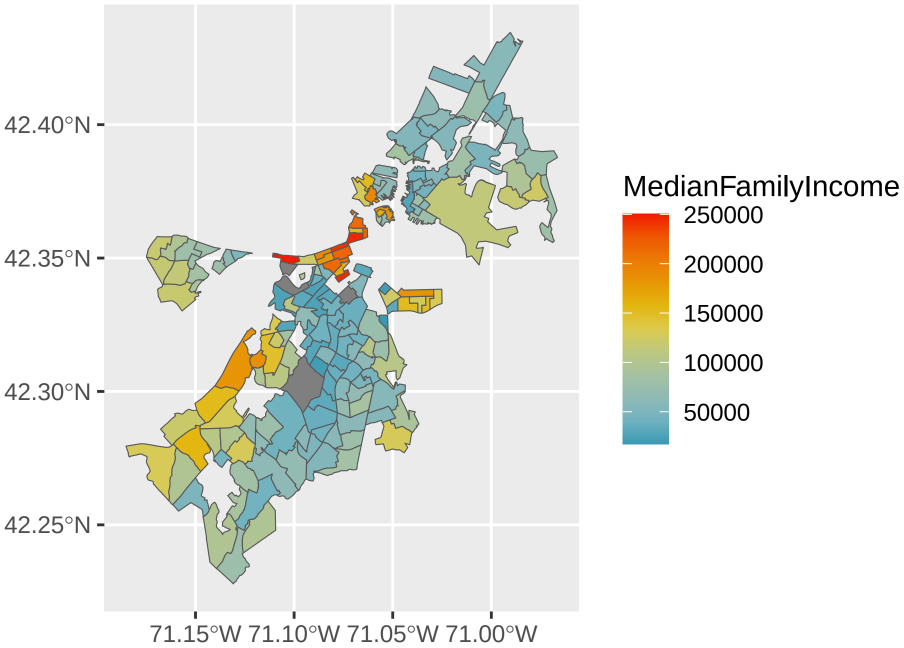
Code for a spatial plot

You can make cool spatial plots using ggplot2 in R. See map of Sussex County, MA (Boston and surrounding cities/towns).

```
library(ggplot2)
#install.packages("wesanderson")

## subset data for MA
datMA <- analysis.dat[analysis.dat$STATE_NAME == "Massachusetts", ]

## fill in missing values for income by setting it to equal median income across MA
datMA$MedianFamilyIncome <- as.numeric(datMA$MedianFamilyIncome)
```

```
## Warning: NAs introduced by coercion
```

```
datMA[is.na(datMA$MedianFamilyIncome)] <- median(datMA$MedianFamilyIncome, na.rm=T)

## ggplot2 for spatial plot (the wesanderson pkg gives you the cool colors!)
ggplot(data=datMA[datMA$County == 25,]) + geom_sf(aes(fill=MedianFamilyIncome)) + scale_fill_gradie
```



```
## compare to plot colored by HOLC grade
```

Your turn..

Here are some ideas to get started. But feel free to explore the data and come up with your own..

Project idea 1

- Pick a health outcome (eg diabetes, CHD) or overall life expectancy.
- Does historical HOLC grade still influence modern day health outcomes?
- If there is an association, does it persist after adjusting for confounders (eg age, socioeconomic status, education, etc.)
- Can you compare if this association varies between New York and California?

Project idea 2

- Focus on air pollution, say PM10, ozone etc.
- Does historical HOLC grade still influence modern day neighborhood characteristics?
- If there is an association, does it persist after adjusting for confounders (eg socioeconomic status, etc.)
- Can you compare if this association varies between New York and California?

Project idea 3

- Using machine learning, can you build a classifier for historical HOLC grade using modern-day demographics, health, neighborhood, economic outcomes
- Perhaps build a classifier using data from one state and predict grade for another state
- Which factors are the most important in predicting HOLC grade?