Case Study #3 - Bivariate Mapping

PUT YOUR NAME HERE

October 09, 2023

Case Study #3 - Bivariate Mapping (100 points)

This markdown document provides instructions, code, and data to create a bivariate map of environmental quality indicators and demographic information. We will use data from the 2020 5-year ACS and EPA's EJ screening tool known as EJSCREEN.

Familar Steps - Install R, Rstudio, and Github (0 points)

This case study requires the use of the *R* programming language. *R* is available for download here. While you can use *R* on it's own, downloading *RStudio* provides a welcoming environment for replication and allows use of the Rmarkdown functionality. It is available here.

Github is free and available for download here. *Github* is used to house the course repository. Installing and using it to clone the repository will simplify the replication procedure. However, you could also simply download a zipped file version of the repository (instructions below) and unzip in the desired location on your machine.

Cloning the Repository

To clone or download the course repository, navigate to the main page and click on the green "code" button in the top right corner of the page and follow the instructions (this will also provide you with an option to download a .zip file of the repository). Alternatively, navigate in the command line terminal to where you would like to clone the repository and then type:

```
git clone https://github.com/adamtheising/environmental_economics
```

If you chose to download a .zip file of the repository, simply unzip it to wherever you would like to have it and proceed with the following steps in that directory.

Getting started

After installing *R*, *RStudio*, and cloning/downloading the repository, navigate in your file explorer (or equivalent) to `case_studies\case_study_3` and open the markdown file `bivariate_mapping.Rmd`. This will open the file in *RStudio*. This markdown document includes all the instructions and code to create a bivariate map.

Part 1 - Opening and Inspecting the Data (30 points)

Load the Texas data (0 points)

Before getting to the mapping part of this assignment, let's start by opening the data, an R spatial data frame of census and EJScreen information. Census tract-level variables are pulled from the 2020 5-year American Community Survey. EJScreen variables relate to socioeconomic and environmental vulnerability.

```
# Load the Census and EJSCREEN data - Note you might need to adjust your directory.

load('tx_census_ejscreen.RData')
```

Socioeconomic Variables (15 points)

Let's inspect socioeconomic variables in the dataset. These primarily come from EJSCREEN. In order, these variables represent percent low income, percent minority, percent unemployed, percent with linguistic isolation (% non-English speaking), percent with less than a high school degree, percent under 5 years old, and percent over 64 years old.

```
# Removing the spatial aspect of the dataset for speed.
summary_data <- tx_census_ejscreen %>%
  st_drop_geometry

# Calculate summary statistics
summary_data <- summary_data %>%
  summarise(
    across(c(lowincpct, pct_minority, unemppct, lingisopct, lesshspct, under5pct, over64pct),
           list(mean = ~mean(., na.rm = TRUE),
                sd = ~sd(., na.rm = TRUE),
                max = ~max(., na.rm = TRUE),
                min = ~min(., na.rm = TRUE))
    )
  ) %>%
  mutate(across(where(is.numeric), round, 2))  # round to 2 decimal places

# Reshaping and then re-order the summary statistics
summary_data_long <- summary_data %>%
  pivot_longer(cols = everything(),
               names_to = c("variable", "statistic"),
               names_pattern = "(.+)_(mean|sd|max|min)",
               values_drop_na = TRUE)

summary_data_long <- summary_data_long %>%
  pivot_wider(names_from = statistic, values_from = value) %>%
  select(variable, mean, sd, min, max)
```

```
kable(summary_data_long, caption = "Summary Statistics - Socioeconomic Variables ")
```

Table 1: Summary Statistics - Socioeconomic Variables

variable	mean	sd	min	max
lowincpct	0.34	0.20	0	1.00
pct_minority	0.57	0.28	0	1.00
unemppct	0.05	0.04	0	0.56
lingisopct	0.08	0.10	0	0.80
lesshspct	0.16	0.14	0	1.00
under5pct	0.07	0.03	0	0.31
over64pct	0.13	0.08	0	1.00

1. (15 points) Select one variable above and explain how it can be considered an indicator of socioeconomic vulnerability to environmental pollution.

- Answer:

Environmental Variables (15 points)

Next, let's summarize environmental burden variables from EJSCREEN. In order, the variables below are particulate matter 2.5 concentrations, proximity to traffic, toxic air emissions cancer risk, percent of housing with probable lead paint, underground storage tanks, proximity to Superfund National Priorities List sites, proximity to hazardous waste treatment storage and disposal facilities, and wastewater discharge into surface waters.

```
# Removing the spatial aspect of the dataset for speed, also re-initializing the summary_data
summary_data <- tx_census_ejscreen %>%
  st_drop_geometry

# Calculate summary statistics
summary_data <- summary_data %>%
  summarise(
    across(c( pm25,ptraf,cancer,pre1960pct,ust,pnpl,ptsdf,pwdis),
           list(mean = ~mean(., na.rm = TRUE),
                sd = ~sd(., na.rm = TRUE),
                max = ~max(., na.rm = TRUE),
                min = ~min(., na.rm = TRUE))
    )
  ) %>%
  mutate(across(where(is.numeric), round, 3))  # round to 3 decimal places

# Reshaping and then re-order the summary statistics
summary_data_long <- summary_data %>%
```

```
  pivot_longer(cols = everything(),
               names_to = c("variable", "statistic"),
               names_pattern = "(.+)_(mean|sd|max|min)",
               values_drop_na = TRUE)

summary_data_long <- summary_data_long %>%
  pivot_wider(names_from = statistic, values_from = value) %>%
  select(variable, mean, sd, min, max)

kable(summary_data_long, caption = "Summary Statistics - Environmental Variables ")
```

Table 2: Summary Statistics - Environmental Variables

variable	mean	sd	min	max
pm25	9.144	1.039	4.913	10.811
ptraf	149.101	209.350	0.000	4075.189
cancer	27.566	11.896	10.000	400.000
pre1960pct	0.152	0.197	0.000	0.925
ust	2.231	2.335	0.000	29.930
pnpl	0.085	0.158	0.000	3.128
ptsdf	0.763	1.170	0.000	13.839
pwdis	0.845	16.206	0.000	723.590

2. (15 Points) These variables have some odd values that aren't immediately interpretable. In some cases, the variable has a straightforward value relating to measured pollution. In other cases, underlying data is being transformed so that it can be easily computed for an entire census division. Please pick one environmental variable above and explain what the numbers represent. You can use the technical documentation for EJSCREEN for help. Variables are described starting on page 13.

- Answer:

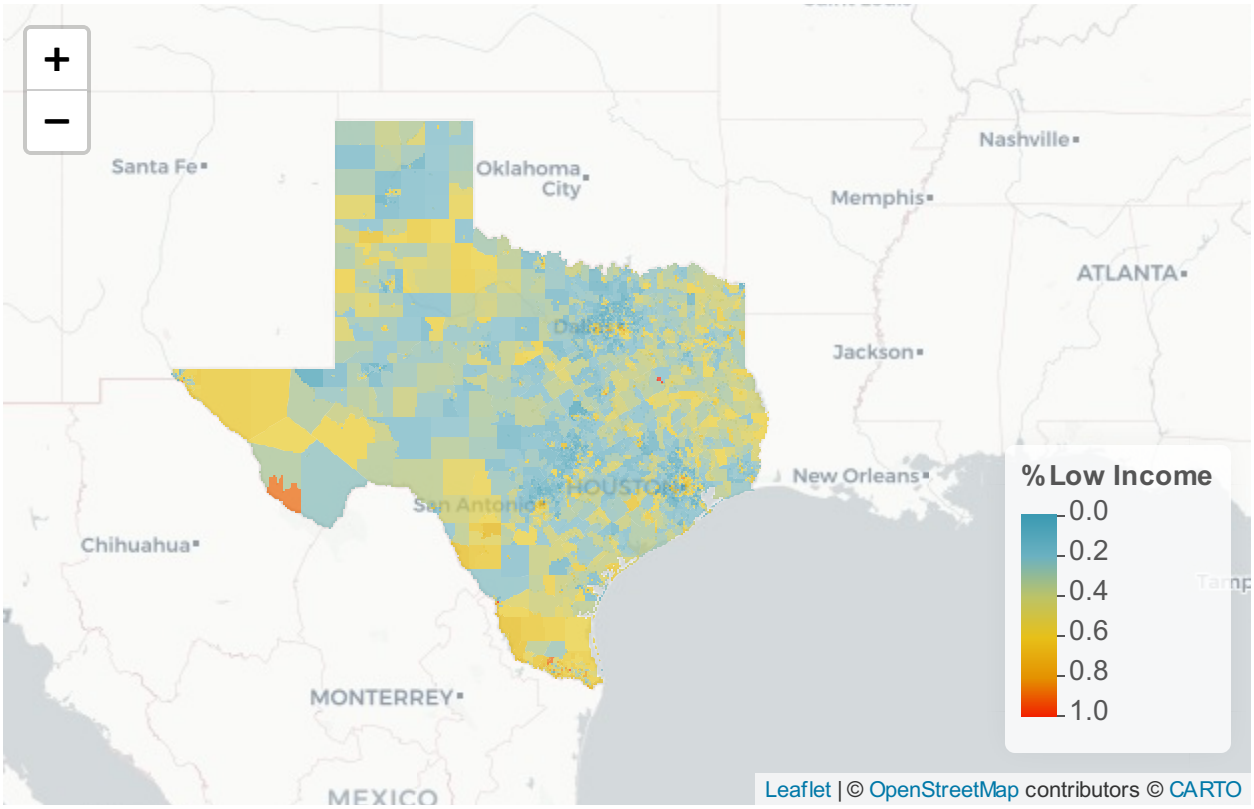
Part 2 - Mapping the Data (20 points)

In this section, we'll map some of the socioeconomic information across Texas. This helps us to ensure that the data has no obvious errors and also provides a visual intuition for the information. For these maps, we'll use the Leaflet R package, which allows us to zoom in and inspect various regions.

Mapping % Low Income (10 points)

Map the percent of the population with low income (please feel free to pick any other socioeconomic variable). Low income in this case is defined as households living with income below twice the federal poverty limit. As such, this measure captures the percent of the population with income below $28,000.

```r
# Define a color palette for the map. Note I'm using the color scheme from
  # The Life Aquatic with Steve Zissou, but you can pick the colors from any
  # Wes Anderson film. See more schemes here: https://github.com/karthik/wesanderson
zissou_colors <- wes_palette("Zissou1")
color_func <- colorRampPalette(zissou_colors)
pal <- colorNumeric(palette = color_func(100),
                    domain = tx_census_ejscreen$lowincpct,
                    na.color = 'whitesmoke')


# Map Texas population across census tracts.
tx_census_ejscreen %>%
  st_transform(crs = "+init=epsg:4326") %>%
  leaflet(width = "100%") %>%
  addProviderTiles(provider = "CartoDB.Positron") %>%
  addPolygons(popup = ~ str_extract(GEOID, "^([^,]*)"),
              stroke = FALSE,
              smoothFactor = 0,
              fillOpacity = 0.7,
              color = ~ pal(lowincpct)) %>%
  addLegend("bottomright",
            pal = pal,
            values = ~ lowincpct,
            title = "% Low Income",
            opacity = 1)
```



3. (10 points) Describe what the map tells us about the geospatial distribution of low income residents and interpret the implications for socioeconomic vulnerability across the state of Texas.

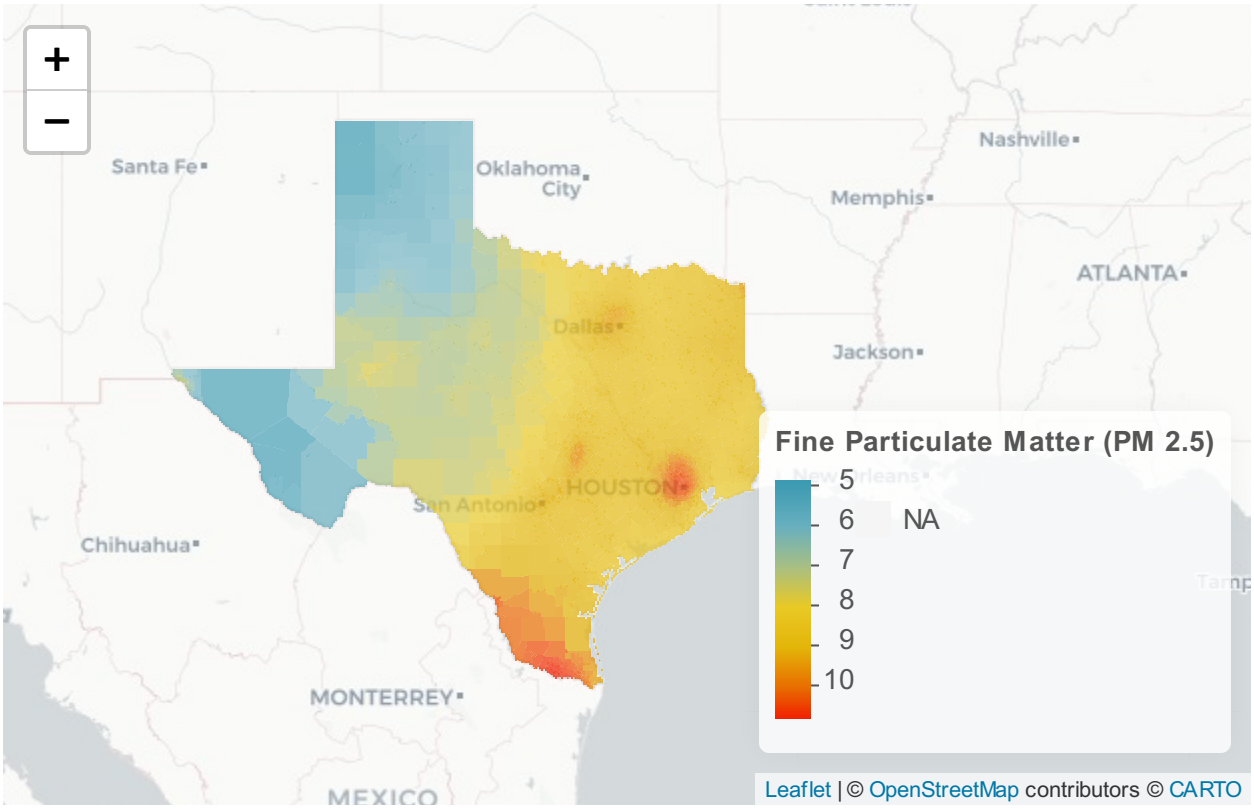
- Answer:

Mapping Environmental Indicators (10 points)

Next, let's map an environmental indicator. You can pick any of them, but below I'm displaying fine particulate matter (i.e., pm25). Fine particulate matter refers to small air pollution particles that can travel deep into the lungs. PM 2.5 has been causally implicated in a wide variety of health impacts including all-cause mortality.

```
# Define a color palette for the map.
pal <- colorNumeric(palette = color_func(100),
                    domain = tx_census_ejscreen$pm25,
```

```
                    na.color = 'whitesmoke')


# Map PM 2.5 across Texas Censu Tracts
tx_census_ejscreen %>%
  st_transform(crs = "+init=epsg:4326") %>%
  leaflet(width = "100%") %>%
  addProviderTiles(provider = "CartoDB.Positron") %>%
  addPolygons(popup = ~ str_extract(GEOID, "^([^,]*)"),
              stroke = FALSE,
              smoothFactor = 0,
              fillOpacity = 0.7,
              color = ~ pal(pm25)) %>%
  addLegend("bottomright",
            pal = pal,
            values = ~ pm25,
            title = "Fine Particulate Matter (PM 2.5)",
            opacity = 1)
```



4. (10 points) Describe hotspot areas of PM 2.5 in Texas. Explain why this information alone is not sufficient for identifying a pocket of environmental justice concern.

- Answer:

Part 3 - Regression Output (20 points)

Next, we'll run a basic regression to determine associations across sociodemographic and environmental information. In the following code, we regress percent minority in a census tract on environmental quality indicators from EJSCREEN. This regression shows the correlation between the percent of a census tract that is people of color and metrics of environmental quality. Note that I'm converting the information to a scale from 0-100 instead of 0-1.

```
tx_census_ejscreen <- tx_census_ejscreen %>%
  mutate(pct_minority=100*pct_minority ) %>%
  mutate(lowincpct=100*lowincpct )
```

```r
# Pick either the low-income regression or the pct-minority regression for this question.

reg_min  <- lm(pct_minority  ~ pm25 + resp + ptraf + cancer + pre1960pct + ust + pnpl + ptsdf +
               data = tx_census_ejscreen) # lm stands for linear model. This is a basic ordina
summary(reg_min)
```

```
##
## Call:
## lm(formula = pct_minority ~ pm25 + resp + ptraf + cancer + pre1960pct +
##     ust + pnpl + ptsdf + pwdis, data = tx_census_ejscreen)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.214 -19.669  -0.197  19.151 165.623
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.108573   3.034980  -3.660 0.000254 ***
## pm25          7.397687   0.340144  21.749  < 2e-16 ***
## resp        -47.675344   3.819181 -12.483  < 2e-16 ***
## ptraf        -0.005660   0.001701  -3.328 0.000879 ***
## cancer        0.107568   0.027281   3.943 8.13e-05 ***
## pre1960pct   17.316422   1.688114  10.258  < 2e-16 ***
## ust           3.106492   0.159161  19.518  < 2e-16 ***
## pnpl         12.623498   2.056850   6.137 8.87e-10 ***
## ptsdf         2.871190   0.299752   9.579  < 2e-16 ***
## pwdis        -0.026151   0.018515  -1.412 0.157863
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.73 on 6779 degrees of freedom
##   (107 observations deleted due to missingness)
## Multiple R-squared:  0.2139, Adjusted R-squared:  0.2129
## F-statistic:   205 on 9 and 6779 DF,  p-value: < 2.2e-16
```

```r
# reg_inc  <- lm(lowincpct ~ pm25 + resp + ptraf + cancer + pre1960pct + ust + pnpl + ptsdf +
#                data = tx_census_ejscreen)
# summary(reg_inc)
```

5. (20 points) This regression is associational and does not represent causation. The way to interpret the point estimates is that a one unit increase in the independent variable is associated with an x unit decrease in the dependent variable (percent minority). For example, the point estimate 7.4 on PM 2.5 suggests that an additional concentration unit of PM 2.5 in the air is associated with 7.4 percentage points more people of color in a census tract. A percentage point increase of 7.4 refers to moving from, for example, 50% to 57.4%; it is not the same as a percent increase.

Pick any variable other than PM 2.5 in the regression model and explain what the coefficient estimate tells us about the association between the environmental variable and the dependent variable (a non-technical explanation of the intuition is fine).

- Answer:

Part 4 - The Bivariate Map (30 points)

A bivariate map jointly displays two variables in a way that facilitates easy location of areas that tend to have high or low values for both indicators simultaneously. This code makes use of the biscale package, which is also used in some of your class readings in modules 3 and 4.

In this example, we plot an environmental indicator alongside a population metric of socioeconomic vulnerability. You can pick other combinations of variables here, but the code below generates a map of % minority plotted against average PM 2.5 concentrations.

First, create the dimensions of the categories for each bin. This command splits up all census tracts into 9 categories corresponding to the lowest third, middle third, and highest third of the PM 2.5 distribution and does the same for the % minority distribution. 3 categories times 3 categories is 9 total combinations of PM 2.5 and % minority.

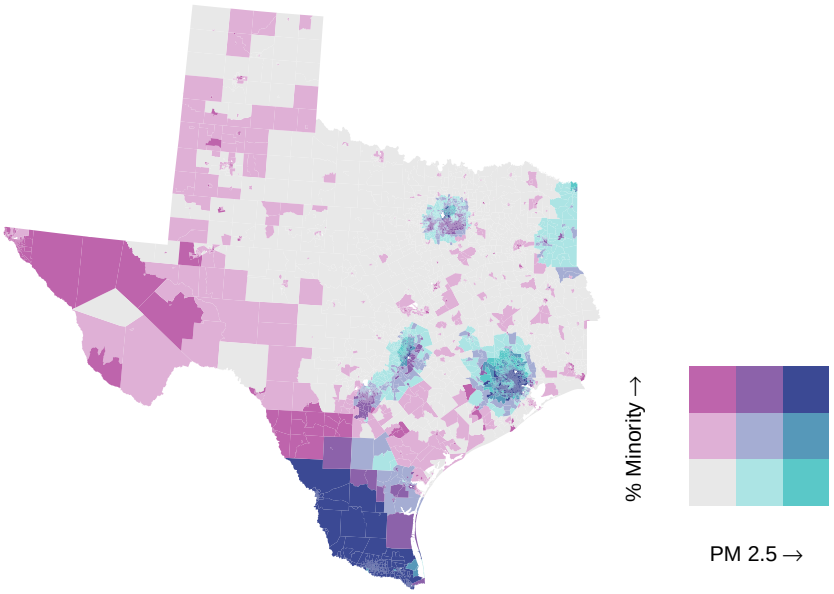
```
data_biscale <- bi_class(tx_census_ejscreen,
                         x = pm25,
                         y = pct_minority,
                         style = "quantile",
                         dim = 3) %>%
  filter(!str_detect(bi_class, 'NA')) # This last step is just in case there are missing value
```

Finally, we can generate the bivariate map for the state of Texas.

```
# This step generates the map's information.
map <- ggplot(data_biscale) +
  geom_sf(data = data_biscale, mapping = aes(fill = bi_class),
          color = NA , size = 0.1, show.legend = FALSE) +
  bi_scale_fill(pal = "DkBlue", dim = 3) +
  labs(
    title = "PM 2.5 and % Minority",
    subtitle = "",
    size = 2
  ) +
  bi_theme()

# This step generates the legend.
legend <- bi_legend(pal = "DkBlue",
                    dim = 3,
                    xlab = "PM 2.5",
                    ylab = "% Minority",
```

```
                    size = 8)

# This step draws the legend and the map in one figure.
finalPlot <- ggdraw() +
  draw_plot(map, 0, 0, 1, 1) +
  draw_plot(legend, 0.4, .08, 0.9, 0.3)

finalPlot
```

PM 2.5 and % Minority



6. (10 points) Great, there's a bivariate map! (You get points as long as your map is not photo-shopped into your document or anything like that.)

7. (10 points) Describe any patterns that seem apparent from the data.

- Answer:

8. (10 points) Explain how a bivariate map provides additional value beyond simply mapping each variable separately.

- Answer: