HW3: Discrete, fixed spatial index

STAT 574E: Environmental Statistics

DUE: 10/17 11:59pm

Homework Guidelines

Please submit your answers on Gradescope as a PDF with pages matched to question answers.

One way to prepare your solutions to this homework is with R Markdown, which provides a way to include mathematical notation, text, code, and figures in a single document. A template `.Rmd` file is available through D2L.

Make sure all solutions are clearly labeled, and **please utilize the question pairing tool on Gradescope**. You are encouraged to work together, but your solutions, code, plots, and wording should always be your own. Come and see me during office hours or schedule an appointment when you get stuck and can't get unstuck.

I. Mathematical [6 pts]

(1) [6 pts] Do *either* of the following:

(Option 1) Let $\boldsymbol{x}$ be a mean-zero multivariate normal random vector of length n with precision matrix $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$ such that $\boldsymbol{x} \sim \mathrm{N}(\mathbf{0}, \mathbf{Q}^{-1})$. Show that

$$\mathrm{E}(x_i | \boldsymbol{x}_{-i}) = \sum_{j \neq i} -\frac{Q_{ij}}{Q_{ii}} x_j.$$

Hint: You may use the fact that $x_i|\boldsymbol{x}_{-i}$ is normally distributed with density function proportional to the density function of $\boldsymbol{x}$.

(Option 2) Ch. 7.9 Exercise 12. (a) from Zimmerman and Ver Hoef:

12. Construct figures analogous to Figure 7.3, showing the marginal correlations corresponding to adjacent site-pairs in 3×3 and 6×6 square lattices, for each of the following, and comment on each figure:

(a) SAR models with $\mathbf{W}$ a binary adjacency matrix and $-1 \leq \rho_{\mathrm{SAR}} \leq 1$

I. Exploring the data [17 pts]

For this homework, you'll be analyzing data related to cyclist injuries in San Diego County. The County reports counts of cyclist injuries for "subregional areal" (SRA) units. SRAs are defined by The County for their own use. Roughly, each areal unit is made up of a union of several census tracts (Figure 1). The shapefiles for this assignment were obtained **here** (the **San Diego Regional Data Warehouse** has several other publicly available spatial datasets). Download the file `Subregional_Areas_2020_shapefile.zip` from D2L and unzip it in your working directory.

(2) [2 pts] **Provide executable code** (**PEC**) that uses the `sf` package to read the shapefiles in the `Subregional_Areas_2020_shapefile/` directory into R as a simple features object called `SRA`. How many SRAs are there? Verify your `SRA` boundaries look like the ones in Figure 1.

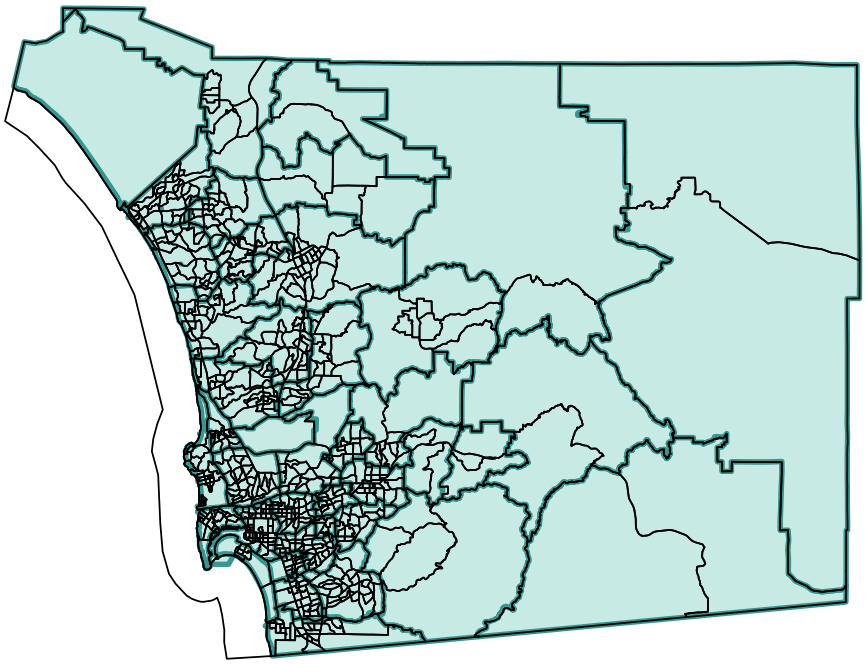

Figure 1: SRAs (large, blue polygons) and Census tracts (black polygons).

(3) [1 pt] Use the spdep package to create a nb object that defines neighbors based on adjacency. **PEC**. What is the most neighbors any areal unit has? Fewest?

(4) [2 pts] Choose another method besides spatial adjacency for defining a binary neighborhood structure. Describe your method and **PEC** that creates an associated nb object or adjacency matrix, $\mathbf{W}$. What is the most neighbors any areal unit has? Fewest?

Next, you'll incorporate data reported by the county on the rates of injuries to cyclists caused by vehicles in 2017. These data were obtained from the County of San Diego's public data repository **here**. The following lines of code read the data into R, and then isolate rows corresponding to SRAs and columns corresponding to variables of interest.

```
bikes_full <- read.csv("Motor_Vehicle_Injuries_to_Pedalcyclists.csv")
bikes <- bikes_full[bikes_full$GeoType == "SRA" & !is.na(bikes_full$GeoType),
                    c("OUTCOME", "GeoName", "GeoID", "Total", "TotalRate")]
```

The variable OUTCOME corresponds to one of six outcomes for the cyclist (you will focus on three of them); GeoName gives the name assigned to an individual SRA; GeoID gives a unique integer label to each SRA that matches with the values in SRA$OBJECTID; Total gives the total number incidents that of the corresponding outcome type that occurred in the given SRA; TotalRate = Total/population of SRA * 100,000 gives the number of incidents per 100,000 people. Values of NA in Total and TotalRate correspond to incident counts for the SRA less than 5.

(5) [2 pts] How many total deaths occurred? How many total hospitalizations? How many total emergency room discharges?

(6) [2 pts] Explain in words what the following 7 lines of code do. Make sure you address all 7 lines. What is the meaning of the column TotalRateInjury in SRA?

```r
SRA <- merge(SRA, bikes[bikes$OUTCOME == "Hospitalization", c("TotalRate", "GeoID")],
             by.x = "sra", by.y = "GeoID")
SRA <- merge(SRA, bikes[bikes$OUTCOME == "ED Discharge", c("TotalRate", "GeoID")],
             by.x = "sra", by.y = "GeoID", suffixes = c("_Hospital", "_EDD"))
SRA$TotalRate_Hospital[is.na(SRA$TotalRate_Hospital)] <- 0
SRA$TotalRate_EDD[is.na(SRA$TotalRate_EDD)] <- 0
SRA$TotalRateInjury <- SRA$TotalRate_EDD + SRA$TotalRate_Hospital
```

One potentially relevant variable that could help explain the number of vehicle-caused cyclist injuries is the total length of bicycle routes within each SRA. The file `BIKE_ROUTES.zip` on D2L contains shapefiles for all the bicycle routes in San Diego.

(7) [2 pts] **PEC** that: (i) reads the bike route shapefiles into R as a simple features object called `routes_sf` and (ii) verifies that the CRS for `routes_sf` matches the one used for the SRAs.

(8) [4 pts] **PEC** to make a plot similar to Figure 2. Make sure your map shows the SRAs of San Diego colored according to `TotalRateInjury` and the bicycle routes (hint: you may want to use the `plot()` function twice; once with `reset = FALSE`, and once with `add = TRUE`). Include your map in your submission.

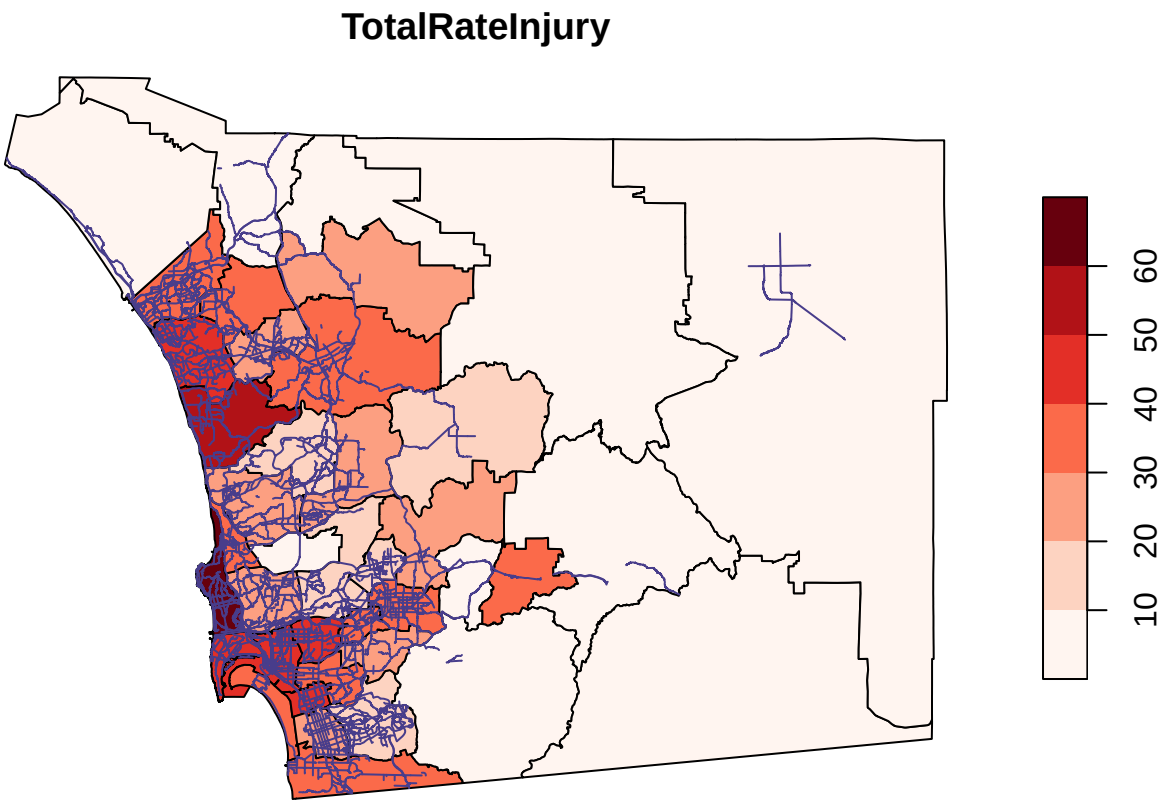

Figure 2: `TotalRateInjury` by SRA. Bicycle routes shown in purple.

Your next goal is to build a model for injury rate as a function of the total length of bicycles routes in a given SRA, so the first thing to do is create that predictor variable. The following code uses the `st_intersection()` and `st_length()` functions from the `sf` package to calculate the length of the bicycles routes in each SRA polygon.

```r
routes_SRA_int <- st_intersection(routes_sf['ROUTE'], SRA['sra'])
routes_SRA_int$length <- units::set_units(st_length(routes_SRA_int), "mi")
SRA_lengths <- aggregate(length ~ sra, data = routes_SRA_int, FUN = sum)
SRA <- merge(SRA, SRA_lengths, by = "sra", all.x = T)
SRA$length[c(38, 40)] <- 0
```

(9) [2 pts] Explain what each of the 5 lines in the code chunk above do. Be sure to explain why the last line about SRAs #38 and #40 is included.

II. Statistical models [17 pts]

(10) [4 pts] **PEC** that fits a traditional linear model for `TotalRateInjury` as function of the miles of bicycle routes in the SRA assuming independent residuals. Give a 95% confidence interval for the effect of miles of bicycles routes. Make a map of the SRAs colored according to the residuals from the linear model. Use a diverging color scheme centered at 0 (hint: you may need to set the `breaks` argument in `plot()`).

(11) [6 pts] **PEC** that fits the same linear model while accounting for residual spatial dependence using a CAR model based on the adjacency network from question (2). Compare the models with and without spatial dependence using AIC and leave-one-out cross-validation. Is there evidence in favor of including the "small" scale spatial effect?

(12) [3 pts] Give a 95% confidence interval for the effect of miles of bicycle routes based on the CAR model from (10). How does it compare to the interval you found using the linear model *without* accounting for residual spatial dependence?

(13) [4 pts] Fit another linear model with spatial dependence (could be another CAR model, but need not be) based on your alternative neighborhood structure from question (3). Compare the two model fits using a predictive score of your choice. Point out differences/similarities you notice in both the "large" and "small" scale effects.