Midterm Exam

Shisham Adhikari

Math 345

Instructions:

1. You have 3 hours to complete this exam, and you must complete it and upload it to gradescope before class on Tuesday 4/1.

Start time: April 4, 12:15PM

End time: April 4, 3:15PM

2. Due to the time limit and asynchronous nature of the exam, I won't be able to answer questions about it. If you have a question, please pose the question in the exam and state what you assume the answer to be, then solve the problem using that assumption. Reasonable assumptions and correct corresponding solutions will be given full points.
3. You may use any available resources except other people. This includes the book, your problem sets, notes, and online resources (e.g. stack overflow, rspatial, anything on the internet).
4. Make sure that you read and answer the questions carefully. Some questions require both coding and a written response.

By typing my name below, I acknowledge that I did not go beyond the allotted 3 hour time limit and that I did not discuss this exam with anyone else: Shisham Adhikari

Problem 1: Analyzing Spatial Data, in theory

For each scenario, assume you can get any data. Briefly explain what data set(s) you would collect and how you would represent each piece of data. Then provide one (and only one) possible analysis you could do to answer the given question.

(a) Where do traffic accidents tend to occur?

- We used a dataset called pdxcrash for Kelly's Data Science. I would probably use that dataset since it has extensive information on location and various variables including frequency and severity of traffic accident. Or, I would check our the traffic data from The National Center for Statistics and Analysis. I can also combine state-wise data with us_states to define a spatial object that I can use for the analysis.
- I would create a kernel density map using the density function and kernel function of my choice (most probably quadrat density) like we did in HW 4 to answer where traffic accidents tend to occur.

(b) Is bird poop more prevalent in urban areas?

- I would choose a random sample location from urban and rural areas. Then ollect the data on frequency and amount of bird poop in the sample selected. Have two separate dataset for rural and urban area with location variable, and poop variable. Or, we can also use some kind of satelite image data to detect fecal matters, that might be complicated though. If we want to be more specific, I know that Gary Granger from Community Safety has an extensive data on crows population in various parts of Portland. I would check it out and see if I can use the data to divide it into urban and suburban/rural areas.

- Implement the likelihood ratio test using anova() function like we did for HW5 to see if we can reject the null model over alternate model to answer if urban location can explain more prevelant bird poop.

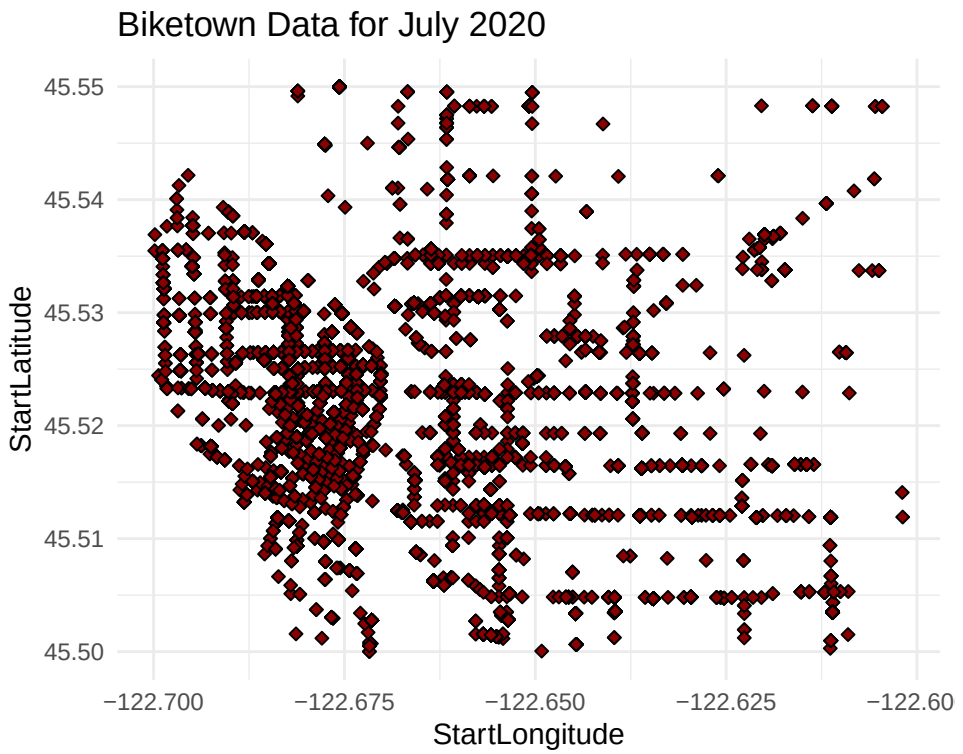
(c) Are homes below flight paths less valuable?

- A home sales dataset, maybe from some online real estate advertising websites or a cleaner dataset like Grisha mentioned in the class. Then assign distances to airports and airport flight paths to every home in the sample. I would try to form a dataset with home variables, their location, their proximity from airport or flight paths, and its sale/current value.
- Use Spatial Regression methods, just a simple regression suffices for a cross-section data, to see if homes below flight paths are less valuable.

Problem 2: Biketown, USA, in practice

(a) Import the provided biketown data for July 2020. We'll use the StartLatitude and StartLongitude variables for our locations. Let's limit our analysis to latitudes between 45.5 and 45.55 and longitudes between -122.7 and -122.6. If you need a CRS and / or projection, you may choose which one(s) to use. Create a plot of the point data. It does not have to be pretty.

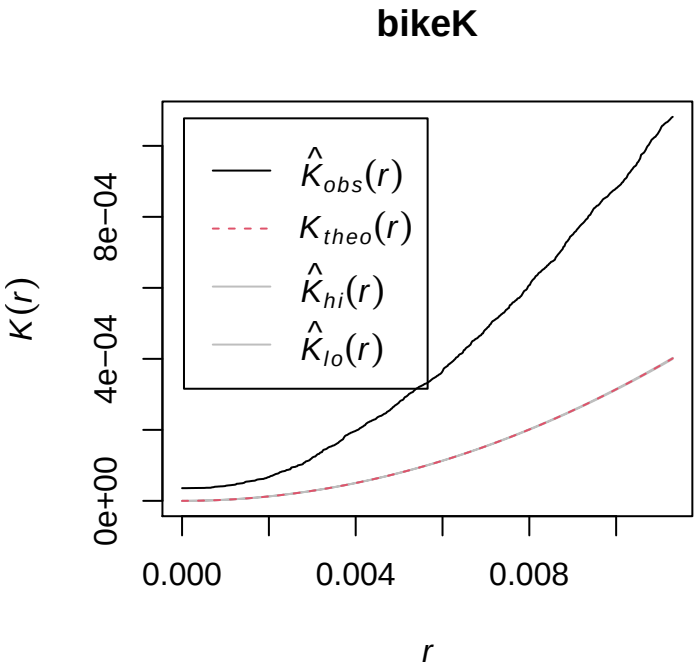
```
pacman::p_load(tidyverse, spdep, maptools, spatstat, rgdal, rspatial, spgwr)
biketown <- read.csv("/home/adhikars/Math 345/Midterm/2020_07.csv")
biketown <- biketown %>%
            filter(between(StartLatitude, 45.5,45.55) & between(StartLongitude,-122.7,-122.6))
#static map
ggplot(data = biketown) +
geom_point(aes(x = StartLongitude, y = StartLatitude),
shape = 23, fill = "darkred") +
labs(title="Biketown Data for July 2020")+
theme_minimal()
```



(b) Graph the K-function and compare it to the theoretical K-function. There may be a limitation on how

many observations will be allowed if you use the spatstat package. You can randomly sample, or use the (automatically) imposed limitation. You do not need to calculate any confidence intervals. Does it seem that the data might be distributed poisson? If not, does it appear to be more clustered or more dispersed?

```
biketown_sf <- st_as_sf(biketown, coords = c("StartLongitude", "StartLatitude"), crs = 4326)
#converting sf to sp
biketown_sp <- as(biketown_sf, 'Spatial')
pts <- coordinates(biketown_sp)
bike_points <-as.ppp(pts, c(-122.6999, -122.6019, 45.50001, 45.54999))
n <- 100
bikeK <- envelope(bike_points, fun=Kest, nsim= n, verbose=F)
plot(bikeK)
```

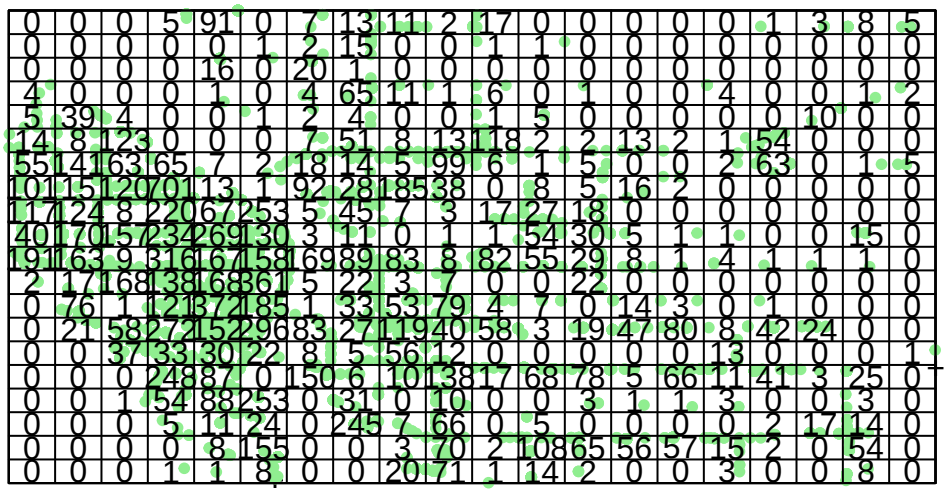


The observed K(r) function falls above the envelope, aka theoretical K-function, indicating the data is not likely poisson distributed and that data is highly clumped at every distance. Although the K function plot for all the data together indicates clumping, perhaps there is some variation by some characteristics like PaymentPlan or something. We can subset the data and invesitage further but we won't need to explore that here.

(c) For our study area, create a 20 x 20 raster where the value for each cell is the count of trips that originated within that cell. Plot the raster. What is another name for what you have created?

We can compute the raster where the value for each cell is the count of trips that originated within that cell using spatstat's quadratcount() functions. The following code chunk divides our study aread into a grid of 20 rows and 20 columns then tallies the number of points falling in each quadrat.

```
Q <- quadratcount(bike_points, nx= 20, ny=20)
plot(bike_points, pch=20, cols="lightgreen", main="Raster with trip counts")  # Plot points
plot(Q, add=TRUE)  # Add quadrat grid
```



Another name is Quadrat density.

(d) Calculate Moran's I using queen's adjacency with the raster cells as the units of observation. Is there evidence of spatial autocorrelation here? Can you reject a null hypothesis of no spatial autocorrelation? You may use the analytical or monte carlo method for your result.

```
#First create polygon
freq <- as.vector(t(Q))
ra <- as.tess(Q)
y <- as(ra, 'SpatialPolygons')

#Create a Queens' case neighborhood object
wr <- poly2nb(y, queen=TRUE)
wm <- nb2mat(wr, style='B')
# Inspect the first order rook's adjacency matrix
wm[1:10, 1:10]
```

```
##                    [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## Tile row 1, col 1     0    1    0    0    0    0    0    0    0     0
## Tile row 1, col 2     1    0    1    0    0    0    0    0    0     0
## Tile row 1, col 3     0    1    0    1    0    0    0    0    0     0
## Tile row 1, col 4     0    0    1    0    1    0    0    0    0     0
## Tile row 1, col 5     0    0    0    1    0    1    0    0    0     0
## Tile row 1, col 6     0    0    0    0    1    0    1    0    0     0
## Tile row 1, col 7     0    0    0    0    0    1    0    1    0     0
## Tile row 1, col 8     0    0    0    0    0    0    1    0    1     0
## Tile row 1, col 9     0    0    0    0    0    0    0    1    0     1
## Tile row 1, col 10    0    0    0    0    0    0    0    0    1     0
```

```
# Calculate Moran's I
ww <- nb2listw(wr, style='B')
i1 <- moran(freq, ww, n=length(ww$neighbours), S0=Szero(ww))$I
i1
```

```
## [1] 0.3712431
```

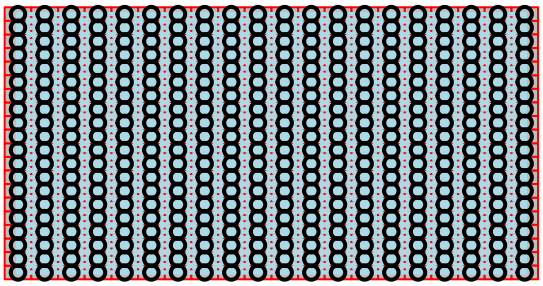
```r
#Now we can test for significance
MC <- moran.mc(freq, ww, nsim=111)
MC
```

```
##
##  Monte-Carlo simulation of Moran I
##
## data:  freq
## weights: ww
## number of simulations + 1: 112
##
## statistic = 0.37124, observed rank = 112, p-value = 0.008929
## alternative hypothesis: greater
```

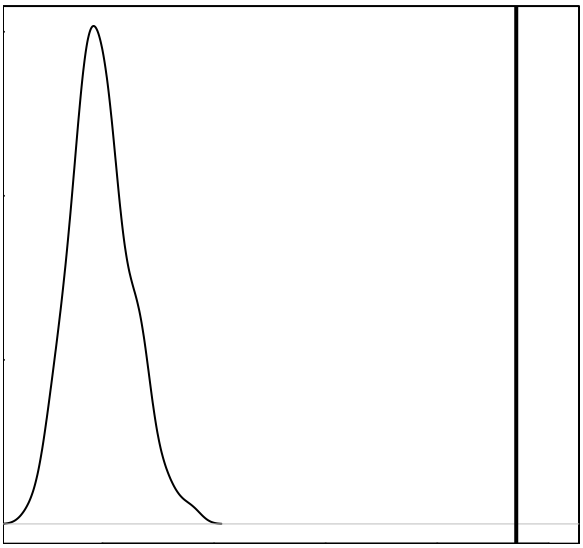
Here, the value of Moran-I is 0.3712 which indicates the presence of the positive spatial autocorrelation. Also, the p-value of the monte-carlo significance test is 0.008 which is less than 0.05. So we can conclude that the result is statistically significant, meaning we can reject the null hypothesis that there is no spatial auto-correlation in the data.

(e) Extra credit [Do not do this until you have completed Problem 3]: Create a nice visualization that showcases one or both of your analyses.

```r
#Plotting Queen's Adjacency
par(mar = c(0, 0, 1, 0))
plot(y, col='gray', border='red')
xy <- coordinates(y)
plot(wr, xy, col='lightblue', lwd=2, add=TRUE)
```



```r
# Plot the distribution (note that this is a density plot instead of a histogram)
plot(MC, main="", las=1)
```



Problem 3: Discussion section

(a) Explain the relationship between the two analyses in Problem 2. Do they suggest the same conclusion? Are there differences that we should think about?

Both analyses in Problem 2 indicate a clustering or spatial autocorrelation in the biketown dataset.

The Moran I analysis is an inferential statistic, meaning its results are interpreted within the context of its null hypothesis. Whereas, the K function is giving results based on the distribution of the data. So, it's really important to choose appropriate conceptualization of spatial relationships to get right answer using Moran I. Also, Moran I assess the overall pattern and trend of our data, so they are most effective when the spatial pattern is consistent across the study area. Whereas, K-function assess the feature of interest within neighboring features' context, so the comparison is between the local context to the global context; so we need to be careful of the distinction.

(b) Suppose we downloaded the biketown data for every month and then made the same raster that we did in part (2c) for each month. What is a name for this type of data?

- Panel data or Longitudinal data

(c) You want to understand what spatial factors explain the changes in the count of bike trips for each cell between 2019 and 2020. Describe what spatial regression you could run in order to investigate this. Again, suppose you could have any datasets you want. It might be helpful to write down your regression equation, but it's not required.

We would need a panel data of the count of bike trips for every month between 2019 and 2020 and run a spatial panel regression/estimation. So, we would have a dataset ordered first by cross-section and then by time period. We can use plm() pacakage in R to run the panel regression in R. We can choose to either run a simple pooling model or fixed-effect model to find understand the spatial factors. For simplicity, we can use the pooling model given by:

$$y_{rt} = \beta_0 + \sum_{i=1}^{j} \beta_i x_{irt} + \epsilon_{rt}$$

, where y is the count of bike trips, x_i is the ith explanatory variable, ϵ is the error term, r is the region index, and t is the time index. We can implement the model in R using plm package. The pseudo methodology would be: start with the panel data, df $\rightarrow$ convert to a pdata.frame object using *pdata.frame(df, c("region","year"))* $\rightarrow$ run the pooled spatio-temporal model using *df.pool = plm(gy gx, data=df, model="pooling")*. Finally use *summary(df.pool)* to see what spatial factors explain the changes in teh count of bike trips for each cell between 2019 and 2020.