Clustering and Cluster Detection in New York City

Yitao Wu

1. Introduction

Crime is one of the unfortunate inevitabilities in the world, which take place everyday and pretty much everywhere. It is a serious cause for concern and may even impede the development of communities. Since crime behaviour is not well understood, police officers couldn't effectively foresee when and where crimes will take place. Nevertheless, previous research has shown that the distribution of crime is known to have a spatial dimension revealing some spatial patterns[1]. In this project, we conducted spatial analysis for crime clustering detection in New York City(NYC) and provided valuable information for police officers to assess crime patterns, optimize resource allocation, and improve emergency call response.

The NYPD crime Data is a public data set including all crimes reported to the New York City Police Department (NYPD) for all complete quarters [2]. In this project, we focus on the complaint data which included crimes related with valid felony, misdemeanor, and violation. Each record represents a criminal complaint in NYC(Brooklyn, Queens, Manhattan, the Bronx, and Staten Island) and includes information about the type of crime, the location and time of enforcement. In addition, victims and suspects' demographics are also included. For this project , we focus on all the crime records occurring within the 2019 calendar year.

Geographical information can be accessed through the American Community Survey data(ACS)[3]. It is the premier source for detailed population and housing information about the United States. The ACS data profiles also have the most frequently requested social, economic,and demographic data. We used ACS data to get population count within each census tract. Tigris R package was used to obtain shapefiles of NYC from the US Census Bureau

2. Method

We first merged all data together to get population and crime counts for each census tract. We calculated empirical averages (e.g., the SMRs) and mapped them to have an initial look at the distribution of valid felony, misdemeanor, and violation crimes.

We used a non-spatial random effect model to smooth the SMRs and provided more reliable estimates in each of the constituent areas. A Poisson-lognormal non-spatial random effect model is given by

$$Y_i|\beta_0, \epsilon_i \sim_{iid} \ Poisson(E_i e^{\beta_0+\epsilon_i}) \quad (1)$$

$$\epsilon_i|\sigma_\epsilon^2 \sim_{iid} \ N(0, \sigma_\epsilon^2) \quad (2)$$

where Y_i represent the number of crimes in census tract i in NYC and we assume Y_i follows a Poisson distribution with mean μ_i where $\mu_i = E_i \theta_i$. Here E_i is the respective expected numbers and θ_i is the relative risk in census tract i. We also assumed that ε_i follows a normal distribution with mean 0 and a default prior theta (variance).

After fitting the model, we visualized the posterior mean of the relative risk for each census tract. We then fit a spatial hierarchical model to incorporate spatial patterns. We chose the BYM2 spatial random effects model which assigned the spatial random effects an intrinsic conditional autoregressive (ICAR) prior. The Spatial model with the BYM2 parameterization is given by

$$Y_i|\beta_0, \beta_1, e_i \sim_{ind} \ Poisson(E_i\theta_i) \quad (3)$$

$$log(\theta_i) = \beta_0 + x_i\beta_1 + b_i \quad (4)$$

$$e_i|\sigma_e^2 \sim_{iid} \ N(0, \sigma_e^2) \quad (5)$$

$$\boldsymbol{S} = [S_0, ..., S_n]|\sigma_s^2 \sim ICAR(\sigma_s^2) \quad (6)$$

where Y_i represent the number of crimes in census tract i in NYC and we assume Y_i follows a Poisson distribution. We also assumed that the spatial random effect follows a normal distribution with its mean equal to the mean of the neighbor's random effect and variance proportional to one over the number of neighbors. The BYM2 model includes an ICAR spatial random effect with independent and identically distributed terms. We calculated the posterior median for the proportion of the spatial residual variation to assess whether there is any spatial effect.

For clustering analysis, we first evaluated Moran's test for spatial autocorrelation using the "W" style weight function as well as the binary "B" weight option. The "W" style weight method standardizes the weights so that for each area the weights sum to 1 while the B style uses 0/1 corresponding to non-neighbor/neighbor. LISA(Local Measure of Spatial Association) approach was also conducted to visualize the local clustering. In general, we mapped the residuals to get a visual on the clustering and evaluate whether the conclusions, evidence of spatial autocorrelation, were consistent when using various methods. cluster detection: detecting areas or contiguous collections of areas that appear to be at elevated risk.

For cluster detection, we concentrated on the SatScan method. Potential clusters were defined as circles centered on the centroids of the areas. For a given circle, we assumed the number of crimes inside and outside the circle following different poisson distributions. The poisson models are given by

$$Y_1 \sim \ Poisson(E_1\theta_1) \quad (7)$$

$$Y_0 \sim \ Poisson(E_0\theta_0) \quad (8)$$

where Y_0 and Y_1 are the numbers of cases outside and inside the circle, □ E_0 and E_1 the respective expected numbers, and□ θ_0 and θ_1 the relative risks. The Satscan method is based on a likelihood ratio statistics, which evaluates whether the null hypothesis $H_0 : \theta_1 = \theta_0$ is true for each circle. The significance level of the overall test statistic is assessed by carrying out a Monte Carlo procedure. We set 40% as the upper bound on the proportion of the population to be contained in any one potential cluster.

3. Result

In total, 8029 crimes were reported in 2019 within NYC. The number of felony, misdemeanor and violation were 3912, 3369 and 748 respectively. As for each borough, 31.3% of the crimes occured in Brooklyn, followed by Manhattan(24.2%), Queens(22.7%) and Bronx(17.1%). The Stanten island had the lowest frequency of crimes(4.5%).We mapped each crime record to visually check the distribution of valid felony, misdemeanor, and violation crimes in NYC (**Fig.1**). Most of the crimes occurred within central NYC such as downtown Manhattan. The Crown Heights district in Brooklyn was another area where crimes took place frequently. Based on this map, we assumed that there might be spatial patterns and even clusters especially in the Manhattan area.

	Overall **(N = 8029)**
Crime type	
Felony	3912(48.7%)
Misdemeanor	3369(42%)
Violation	748(9.3%)
Borough	
Bronx	1375(17.1%)
Brooklyn	2514(31.3%)
Manhattan	1941(24.2%)
Queens	1824(22.7%)
Stanten island	360(4.5%)
Missing	15(0.2%)

Table.1 Descriptive statistics of crime records collected by NYPD.

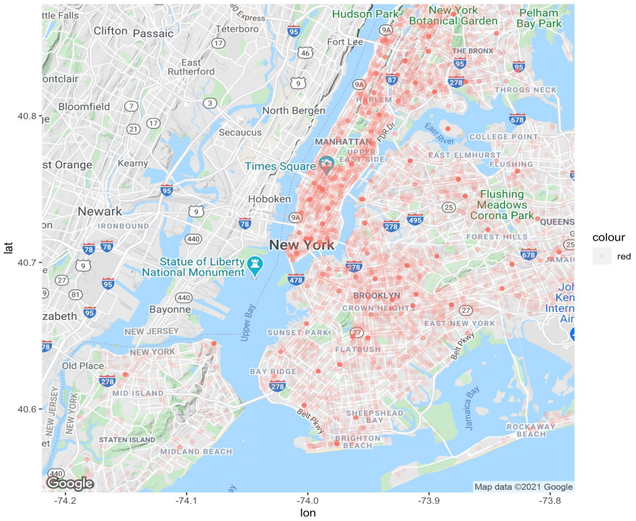

Fig.1 Mapping of felony, misdemeanor, and violation crimes in NYC. A red dot represents one crime record and the color reflects the frequency of crimes.

After excluding areas where the population estimates were missing or equal to zero, we selected 2117 census tracts for this analysis. We calculated the SMRs based on the observations of crimes and expected number of crimes. The SMRs ranged from 0 to 63 where 25%, 50% and 75% quartiles were 0.3, 0.8 and 1.5 respectively. The map of the SMRs (Fig.2) showed a number of census tracts with high relative risks (the risk relative to the city wide risk). There were some spatial trends since the extreme value occurred in the north and central NYC.

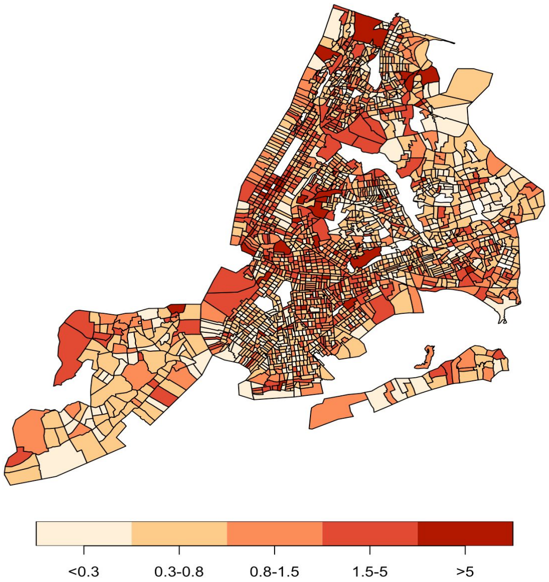

Fig.2 Mapping of SMRs in NYC. The SMRs were categorized into five intervals based on quantiles and extreme values

We first fit a non-spatial random effects model to examine the posterior mean of the relative risk. The residuals ranged from 0.3 to 24 where 25%, 50% and 75% quartiles were 0.6, 0.8 and 1.2 respectively. To better visualize the posterior median, we set a binary indicator of whether the posterior median was greater than 1.5. According to Fig.3, there were a number of census tracts which had high

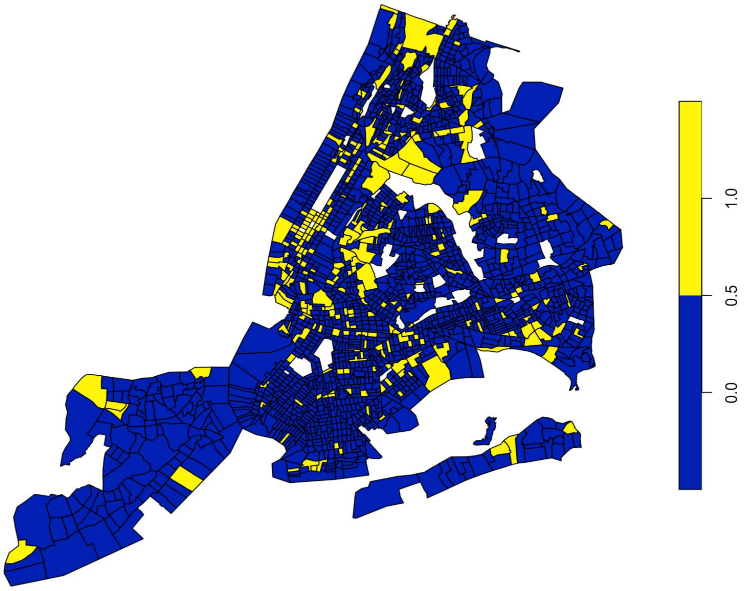

Fig.3 Mapping of posterior median calculated by non-spatial random effects model. The posterior medians were dichotomized using 1.5 as the threshold.

mean values. After having a rough idea about non-spatial effects, we constructed a BYM2 model to evaluate the variances of the spatial and non-spatial random effects. Fig.4 showed a very consistent result with the non-spatial model, that is, there were a number of census tracts which had high mean values. This method also gave a larger collection in the central and the north of New York city though not obvious. The posterior median for the proportion of the spatial residual variation was 0.38 (95% interval: 0.26, 0.5).We can conclude that there were spatial random effects within NYC and we further explored the potential clustering of crimes in neighboring areas.

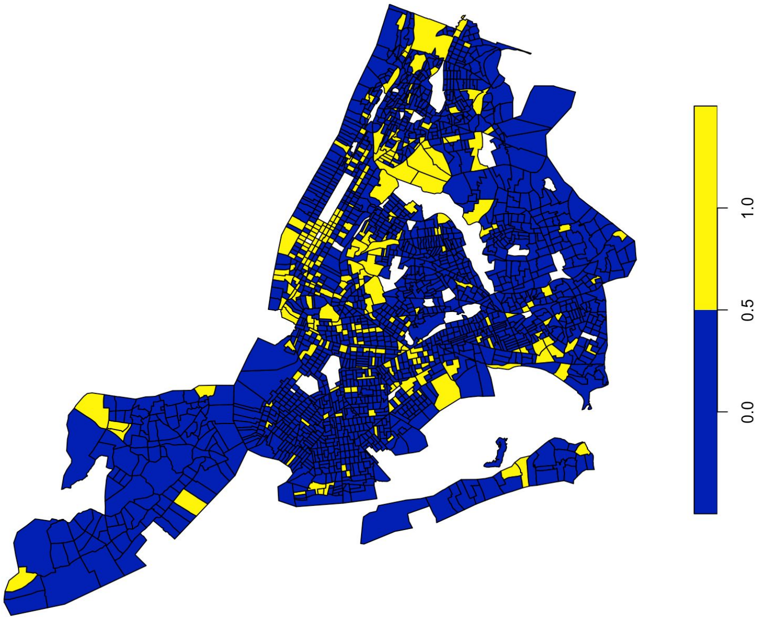

Fig.4 Mapping of posterior median calculated by BYM2 model. The posterior medians were dichotomized using 1.5 as the threshold.

Moran's test as well as Geary's test were performed for spatial autocorrelation using the "W" style weight function as well as the binary "B" weight option. We detrend SMRs from the 2117 census tract using a simple intercept Poisson model. For Moran's I statistic, we noticed that p-values were much lower than 0.05 for both the "B" style and "W" style weights indicating significant spatial clusters . For Geary's C statistic, the p-value for "W" style still indicated that there was significant clustering. Though testing results were not completely consistent, there was sufficient evidence that spatial effects within neighboring areas contributed to residuals.

	Moran's I	Geary's C
"B" style weight	<0.001	0.09

“W” style weight	<0.001	<0.001

Table.2 Results of Moran’s test and Geary’s test using “W”style weight and “B” style weight

We also used the LISA method to explore the spatial autocorrelation(Fig.5). The identified clusters were areas of high crime counts surrounded by other areas of high crime reports. The red area contributed significantly to a positive global spatial autocorrelation result. This result was also consistent with our previous finding that most of the potential clusterings were in central NYC.

Finally, we used the Satscan method for cluster detection aiming to detect areas or contiguous collections of areas that appear to be at elevated risk. We visualized the results in Fig. 6, where red color indicated the most likely clusters and the other color indicated secondary likely clusters. The most likely cluster included 262 census tracts with a total population of 1128142. The blue area represented there was another crime cluster on the other side of New York central park. This result was consistent with our previous findings that the potential cluster was in central NYC.

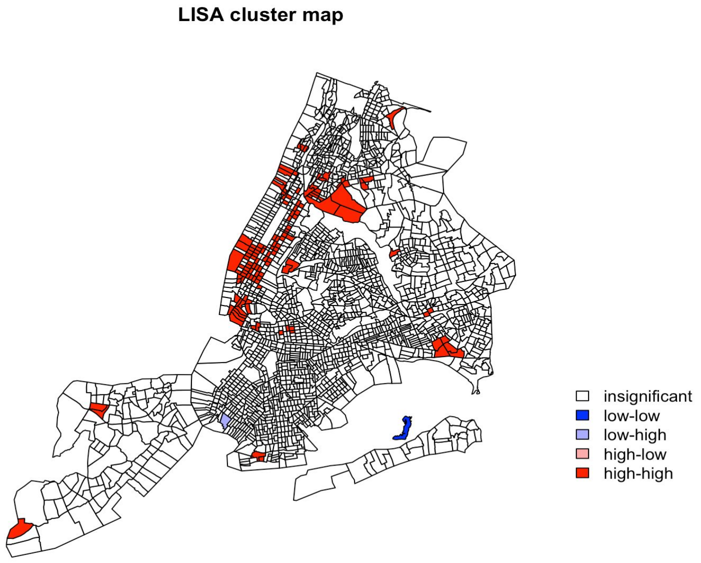

Fig.5 LISA cluster map. Red areas have high values of the residuals and have neighbors that also have high values (high-high). Pink areas have high values of the residuals and have neighbors that have low values (high-low). Light blue areas have low values of the residuals and have neighbors that have high values (low-high). Dark blue areas have low values of the residuals and have neighbors that also have low values (low-low)

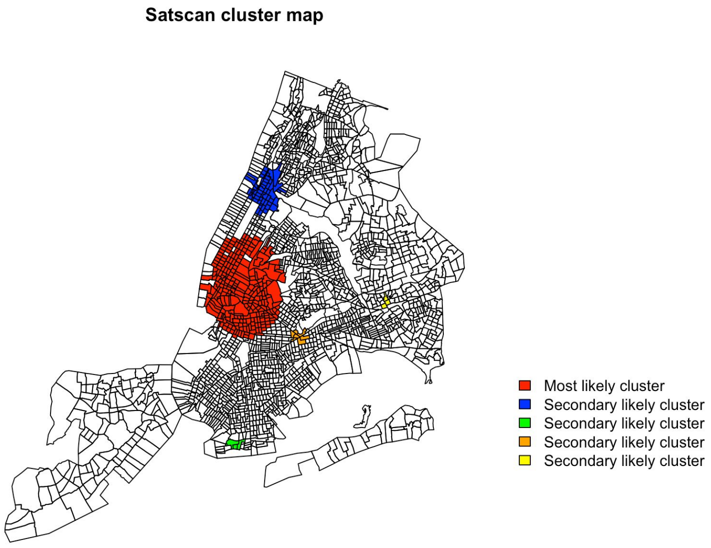
Satscan cluster map
🟥 Most likely cluster
🟦 Secondary likely cluster
🟩 Secondary likely cluster
🟧 Secondary likely cluster
🟨 Secondary likely cluster

Fig.6 Satscan cluster map. Red color indicated the most likely clusters. Blue, green, orange and yellow indicated secondary likely clusters respectively.

4. Conclusion & Discussion

To sum up, we investigated the potential crime clusters in NYC. Using the BYM2 model, we measured that the proportions of the spatial residual variation is 38%. We performed Moran's tests and Geary's test to determine the level of clustering in a set of data. The spatial autocorrelation was significant when using "W" style for both test and "B" style for Moran test. The LISA statistics also reflected that areas of high crime counts were surrounded by other areas of high crime reports. We conducted the Satscan method for cluster detection and we identified the most likely cluster was in central NYC. Overall, all these analysis results suggested that there were crime clusters in NYC, especially in downtown and near central park. It enabled police officers to assess crime patterns and optimize their resource allocation.

Our analysis had several limitations. Firstly, there were 43 census tracts with no people living there . In our exploratory analysis we just simply excluded those areas even though some areas had crime records. The rationale for doing this was that it may cause problems for the calculation of SMRs. We also noticed that there were 64 census tracts with population estimates less than 200. One example is census tract 109 representing New York central park. Usually census tract contains population around 2000 but the total population was only 172 within census tract 109. One possible reason was that those park residents were homeless or parks department employees. Keeping these low population areas in our analysis led to extreme value of SMRs and residuals. These outliers also turned to inconsistent results in Moran's test and Geary's test. For example, the p-value for Moran's test was <0.001 while the p-value for Geary test was 0.48. After excluding census tract with residuals larger than 50, the result became more reasonable. Further analysis could be explored to carefully consider the extreme population areas rather than simply excluding them. The socio-economic characteristics could also be considered to provide more profound analysis of spatial correlation and assist police officers to assess crime patterns.

5. Reference

[1]Murray, Alan T., et al. "Exploratory spatial data analysis techniques for examining urban crime: Implications for evaluating treatment." British Journal of criminology 41.2 (2001): 309-329.

[2]https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243

[3]https://www.census.gov/acs/www/data/data-tables-and-tools/data-profiles

6. Appendix(code)

```
###Load in data

nyc_tracts <- tracts(state = '36', county = c('061','047','081','005','085'))

NYPD_complain <- rgdal::readOGR(dsn = "." , layer =
"geo_export_af75edd8-edf3-4940-8fd1-b91d42e44702")

NYcrime <- NYPD_complain@data

head(NYcrime)

table(NYcrime$date_cmpln)

###select 2019 data

NYcrime$time.label<- unlist(lapply(NYcrime$date_cmpln, grepl, pattern = "2019*"))

table(NYcrime$time.label)

NYcrime.2019 <- NYcrime %>% filter(time.label == TRUE)
```

```
head(NYcrime.2019)

###generate google figure

nyc <- get_map(location = c(lon = -74.00, lat = 40.71), maptype = "terrain", zoom = 11)

ggmap(nyc) +

  geom_polygon(data=nyc_neighborhoods_df, aes(x=long, y=lat, group=group), color="blue", fill=NA)
+ stat_density2d(aes(x = longitude, y = latitude, fill = ..level..,alpha=..level..), bins = 20, geom =
"polygon", data = NYcrime.2019) +

  scale_fill_gradient(high = "red")+

  ggtitle("Map of Crime Density in NYC")

ggmap(nyc)+ geom_point(aes(x = longitude, y = latitude, colour = "red"), data = NYcrime.2019, alpha
= 0.1)

colnames(NYcrime.2019)

###mapping to each census tract

points.frame <- as.data.frame(NYcrime.2019[,c("longitude", "latitude")])

points.frame <- SpatialPoints(points.frame)

spd <- sf::as_Spatial(st_geometry(nyc_tracts), IDs = as.character(nyc_tracts$GEOID))

proj4string(spd)

proj4string(points.frame) <- proj4string(spd)

admin2.key <- over(points.frame, spd)

summary(admin2.key)

NYcrime.2019$GEOID <- nyc_tracts$GEOID[admin2.key]

### exclude zero estimates for each census tract

crime.sp = SpatialPoints(NYcrime[, c("longitude","latitude")])

all_crime <- NYcrime.2019 %>% group_by(GEOID) %>% dplyr::summarise(num_points = n())

all_crime

nyc_popc <- sp::merge(x = nyc_pop, y = all_crime, by = "GEOID", all.x = TRUE)
nyc_popc <- nyc_popc %>% filter(estimate > 0)

nyc_popc[is.na(nyc_popc$num_points),"num_points"] <- 0

### calculate and plot SMRs

nyc_popc <- nyc_popc %>% filter(estimate >= num_points)

summary(nyc_popc$num_points)

Y <- nyc_popc$num_points

E <-  nyc_popc$estimate * sum(nyc_popc$num_points)/sum(nyc_popc$estimate)

nyc_popc$SMR <- Y/E

nyc_popc$EXP <- E

pal1 <- brewer.pal(9,"OrRd")

library(RColorBrewer)

pal <- brewer.pal(5, "OrRd")

nyc_popc$SMR_INT <-cut(nyc_popc$SMR, breaks =c(-0.0001, 0.3, 0.8, 1.5, 5, 100), labels
=c("<0.3","0.3-0.8","0.8-1.5", "1.5-5", ">5"))


plot(nyc_popc["SMR_INT"], pal = pal, key.pos = 1)

### non-spatial model

library(INLA)

head(nyc_popc)

m0 <- inla(num_points ~ f(GEOID, model = "iid"),family = "poisson", E = EXP, data =
as.data.frame(nyc_popc), control.predictor = list(compute = TRUE))

head(m0$summary.fitted.values)

nyc_popc$RRpmean0<- m0$summary.fitted.values[, 1]

summary(nyc_popc["RRpmean0"])
```

```
plot(nyc_popc["RRpmean0"],breaks = seq(0,10,0.5))

nyc_popc$RRpmean0_INT <-cut(nyc_popc$RRpmean0, breaks =c(-0.0001, 0.5, 0.8, 1.2, 5, 100), labels
=c("<0.5","0.5-0.8","0.8-1.2", "1.2-5", ">5"))

plot(nyc_popc["RRpmean0_INT"], pal = pal, key.pos = 1)


###generate neighbor

nyc_popc$GEOID2 <- as.numeric(nyc_popc$GEOID)

nyc_popc$id1 <- 1: length(nyc_popc$GEOID)

nyc_popc$id2 <- 1: length(nyc_popc$GEOID)

nyc.adj <- poly2nb(nyc_popc)


B.nyc <- nb2mat(nyc.adj, style = "B", zero.policy = TRUE)

W.nyc <- nb2mat(nyc.adj, style = "W", zero.policy = TRUE)


col.W <- nb2listw(nyc.adj, style = "W", zero.policy = TRUE)

col.B <- nb2listw(nyc.adj, style = "B", zero.policy = TRUE)

nyc_popc$GEOID2 <- as.numeric(nyc_popc$GEOID)


### BYM2

formula <- num_points ~ 1 + f(id1, model="bym2", graph=W.nyc)

dim(W.nyc)

m2 <- inla(formula, data=as.data.frame(nyc_popc), family="poisson",E=EXP,
control.predictor=list(compute=TRUE),control.compute=list(config = TRUE))

m2$fit2fitted <- m2$summary.fitted.values$`0.5quant`

m2$summary.hyperpar[,1:5]
```

```
###generate plot for report

nyc_popc$RRpmean1<- m2$summary.fitted.values[, 1]

summary(nyc_popc["RRpmean1"])


plot(nyc_popc["RRpmean1"])

plot(nyc_popc["RRpmean1"],breaks = seq(0,28,2), key.pos = 1)

nyc_popc$RRpmean1_INT <-cut(nyc_popc$RRpmean1, breaks =c(-0.0001, 0.5, 0.8, 1.2, 5, 100), labels
=c("<0.5","0.5-0.8","0.8-1.2", "1.2-5", ">5"))

plot(nyc_popc["RRpmean1_INT"], pal = pal, key.pos = 1)


summary(m1$summary.fitted.values[, 4])

nyc_popc$RR0_bin <- m0$summary.fitted.values[, 4] > 1.5

nyc_popc$RR1_bin <- m2$summary.fitted.values[, 4] > 1.5

table(nyc_popc$RR0_bin)

table(nyc_popc$RR1_bin)

plot(nyc_popc["RR0_bin"])

plot(nyc_popc["RR1_bin"])

###Moran test

quasipmod <- glm(num_points ~ 1, offset = log(EXP), data = nyc_popc,family = quasipoisson())

sidsres <- residuals(quasipmod, type = "pearson")

nyc_popc <- nyc_popc %>% filter(res <= 30)

nyc_popc %>% filter(estimate < 200)

moran.test(nyc_popc$res, col.W, zero.policy= TRUE)

moran.test(nyc_popc$res, col.B, zero.policy= TRUE)
```

```
geary.test(nyc_popc$res, col.W, zero.policy= TRUE)
geary.test(nyc_popc$res, col.B, zero.policy= TRUE)

###regenerate neighbour after exclusion and reconduct analysis(didn’t repeat here)
nyc.adj <- poly2nb(nyc_popc)
B.nyc <- nb2mat(nyc.adj, style = "B", zero.policy = TRUE)
W.nyc <- nb2mat(nyc.adj, style = "W", zero.policy = TRUE)
col.W <- nb2listw(nyc.adj, style = "W", zero.policy = TRUE)
col.B <- nb2listw(nyc.adj, style = "B", zero.policy = TRUE)
nyc_popc$GEOID2 <- as.numeric(nyc_popc$GEOID)
### LISA
local <- localmoran(x = nyc_popc$num_points, listw = col.W, zero.policy = TRUE)
local
quadrant <- vector(mode="numeric",length=nrow(local))

# centers the variable of interest around its mean
m.crime <- nyc_popc$num_points - mean(nyc_popc$num_points)

# centers the local Moran's around the mean
m.local <- local[,1] - mean(local[,1])

# significance threshold
signif <- 0.1
# builds a data quadrant
```

```
quadrant[m.crime >0 & m.local>0] <- 4
quadrant[m.crime <0 & m.local<0] <- 1
quadrant[m.crime <0 & m.local>0] <- 2
quadrant[m.crime >0 & m.local<0] <- 3
quadrant[local[,5]>signif] <- 0

brks <- c(0,1,2,3,4)
colors <- c("white","blue",rgb(0,0,1,alpha=0.4),rgb(1,0,0,alpha=0.4),"red")
plot(nyc_popc$geometry,border="black",col=colors[findInterval(quadrant,brks,all.inside=FALSE)],main = "LISA cluster map")
legend(x = -73.65, y = 40.65,legend=c("insignificant","low-low","low-high","high-low","high-high"),
    fill=colors,bty="n")
### Satscan
population <- nyc_popc$estimate
cases <- nyc_popc$num_points
n <- length(cases)
centroids <- matrix(0, nrow = n, ncol = 2)

for (i in 1:n) {
centroids[i, ] <- c(nyc_popc$INTPTLON[i], nyc_popc$INTPTLAT[i]) }
geo <- centroids
pop.upper.bound <- 0.4
n.simulations <- 999
alpha.level <- 0.05
```

```
Kpoisson <- kulldorff(geo, cases, population, expected.cases = nyc_popc$Exp, pop.upper.bound, n.simulations, alpha.level, plot = T)

Kpoisson

Kcluster0 <- Kpoisson$most.likely.cluster$location.IDs.included

Kcluster1 <-Kpoisson$secondary.clusters[[16]]$location.IDs.included

Kcluster2 <-Kpoisson$secondary.clusters[[18]]$location.IDs.included

Kcluster3 <-Kpoisson$secondary.clusters[[27]]$location.IDs.included

Kcluster4 <-Kpoisson$secondary.clusters[[28]]$location.IDs.included

nyc.new <- nyc_popc$geometry

plot(nyc.new, main = "Satscan cluster map")

plot(nyc.new[Kcluster0], add = TRUE, col = "red")

plot(nyc.new[Kcluster1], add = TRUE, col = "blue")

plot(nyc.new[Kcluster2], add = TRUE, col = "green")

plot(nyc.new[Kcluster3], add = TRUE, col = "orange")

plot(nyc.new[Kcluster4], add = TRUE, col = "yellow")

colors <- c("red","blue","green","orange","yellow")

legend(x = -73.65, y = 40.65,legend=c("Most likely cluster","Secondary likely cluster","Secondary likely cluster","Secondary likely cluster","Secondary likely cluster"),fill=colors,bty="n")
```