exam_response.Rmd

Ruici Xia

17:31:19, 14 , 2023

Contents

1 Originality declaration

I, [**Ruici Xia**], confirm that the work presented in this assessment is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

date: 14 , 2023

2 Start your response here

3 Initial project scope

This research wants to find whether certain communities of eviction densities in New York are similar and exhibit spatial autocorrelation in 2020. The author will therefore using the research finding to inform future work on neighborhood planning on New York eviction policies.

Research Question: Whether certain communities in New York have a spatial relationship in 2020 eviction? The null hypothesis that I am going to test empirically is that there is no spatial relationship among densities

of community districts in New York in 2020 eviction. Waldo Tober's first law of geography is that "Everything is related to everything else, but near things are more related than distant things." So we would expect most geographic phenomena to exert a spatial autocorrelation of some kind. Put everything in an academic tone:The null hypothesis is defined as the likelihood of a region receiving a particular value being the same for all locations i, and is independent of what happens in the rest. In the other words, evictions are distributed homogeneously in spatial terms. In contrast to the alternative hypothesis, the likelihood is not the same in all communities and/or the observed level of the vairiable in i is not independent of what happens in the rest, In other words, spatial autocorrelation in seen to exist.

By doing the research, the study aim to offer a spatial perspective to understand why people in those areas could have a higher possibility of eviction and therefore help New York government to prevent more eviction in the future.

3.0.1 Data

(1) List of evictions - (CSV file that contains all the eviction information from 2017 to present within five boroughs such as dates, Scheduled Status police names, spatial coordinates, addresses and other spatial information) https://data.cityofnewyork.us/City-Government/Evictions/6z8x-wfk4 Eviction data is compiled from the majority of New York City Marshals, and Marshals has their own bias as they are independent public officials.

(2) New York City community districts - (shp file that cotains New York spatial information and it can be used for presenting eviction information spatially on the map) https://data.cityofnewyork.us/City-Government/Community-Districts/yfnk-k7r4 New York boundaries of Community Districts is run by NYC Open Data team at the NYC Office of Technology and Innovation(OTI). Although the data is collected and maintained by the City government, it could contains spatial bias, as they do not warranty the completeness, accuracy, content, or fitness for any particular purpose or use of any public data set made available on NYC Open Data, nor are any such warrranties to be implied or inferred with respect to the public data sets furnished therein.

- How will i wrangle the data (based on the previous points) to apply the methods (1)import data, sorted the NAs (2)Data cleaning, coordinate system check (3)basic data visualisation (4)applied methods: Ripley k DBSCAN Dbclustering visualisation Spatial weight matrix Moran's I Global and visualisation (5)Interpretation and reflection
- What are the limitations and assumptions (of either the data or the analysis) Since the data mentioned above has their bias and potential errors, the research conducted from them could also produce errors.The research believe the methods applied in the study could relatively provide a good understanding of the distribution of eviction data in New York, and therefore conduct a basic discussion and policy suggestions.

3.0.2 Data loading

Read in data - note the NA value.

3.0.3 Checking the variable type to make sure there are no character columns that should be numeric due to NAs

```
Datatypelist <- evictions_points %>%
  summarise_all(class) %>%
  pivot_longer(everything(),
               names_to="All_variables",
               values_to="Variable_class")

Datatypelist
```

4 Data wrangling

Check the coordinates on this website for the csv - https://www.latlong.net/. Looks like they are in WGS84. Convert csv to sf object the map

Missing values for coordinates thrown an error so i need to filter them out...

```
Datatypelist <- evictions_points %>%
  summarise_all(class) %>%
  pivot_longer(everything(),
               names_to="All_variables",
               values_to="Variable_class")

Datatypelist


points <- evictions_points%>%
  #also possible to use something like drop_na(Longitdue, Latitude)
  filter(Longitude<0 & Latitude>0)%>%

  st_as_sf(., coords = c("Longitude", "Latitude"),
                   crs = 4326)
```

76,613 features now from 84,556 in the original dataset.

4.0.1 Make a map

Check how the eviction points spreading in New York.

```
tmap_mode("plot")
tm_shape(community_areas) +
  tm_polygons(col = NA, alpha = 0.5) +
tm_shape(points) +
  tm_dots(col = "blue")
```

A lot of points!

4.0.2 Check the are all within the boundaries through a spatial subset

```
community_areas <- community_areas%>%
  st_transform(., 4326)

points_sub <- points[community_areas,]
```

Still have 76,613! So all were intersecting the boundary.

Now focus on 2020 EXPLAIN...I have used string detect here to find the rows that have 2020 within the column executed_date

```
points_sub_2020<-points_sub%>%
  clean_names()%>%
  filter(str_detect(executed_date, "2020"))%>%
 # filter(eviction_legal_possession=="Eviction")%>%
  filter(residential_commercial=="Residential")
```

This has reduced it to 2,859 points, if i remove the legal possession/eviction line then it's around 2,000

```
tmap_mode("plot")
tm_shape(community_areas) +
  tm_polygons(col = NA, alpha = 0.5) +
tm_shape(points_sub_2020) +
  tm_dots(col = "blue")
```

4.1 Data analysis

Let's do some point pattern analysis...

error that only projected coordinates can be used for ppp object! let's project - https://epsg.io/2263. Note that this is in feet.

A better one might be https://epsg.io/6538 as it uses meters

```
community_areas_projected <- community_areas %>%
  st_transform(., 6538)

points_sub_2020_projected <- points_sub_2020 %>%
  st_transform(., 6538)


window <- as.owin(community_areas_projected)
plot(window)

#create a sp object
points_sub_2020_projected_sp<- points_sub_2020_projected %>%
  as(., 'Spatial')
#create a ppp object
points_sub_2020_projected_sp.ppp <- ppp(x=points_sub_2020_projected_sp@coords[,1],
                          y=points_sub_2020_projected_sp@coords[,2],
                          window=window)
```

Ripley k

Using Ripley K to compare the observed distribution of points with the Poinsson random model for a whole range of different radii.

From the graph performed below, the value of K(black line) falls above the line in red(the theoretical value of K for each distance r window under a Poisson assumption of Complete Spatial Randomness), it means the eviction points appear to be clustered from 0 to 5000 metres. There is now value of K below the line, there is no dispersed data.

```
K <- points_sub_2020_projected_sp.ppp %>%
  Kest(., correction="border") %>%
  plot()
```

Density-based spatial clustering of applications with noise: DBSCAN Quadrat and Ripley's K analysis are useful exploratory techniques for telling us if we have spatial clusters present in our point data, but they are not able to tell us WHERE in our area of interest the clusters are occurring. To discover this we need to use DBSCAN for discovering clusters for New York Eviction in 2020.

DBSCAN requires you to input two parameters: 1. Epsilon - this is the radius within which the algorithm with search for clusters 2. MinPts - this is the minimum number of points that should be considered a cluster

Using kNNdistplot() from the dbscan pacakge to find a suitable eps value based on the 'knee' in the plot.This plot shows for each point the average distance to the k neighbours, which are then plotted in ascending order. The knee is where this value (of distance to neighbours) increases. eps=1000, because this is a proper distant

for a city like New York that allow more distant points to be considered as neighborhood. Then, it needs a bigger MinPts value respectively.

Ripley's K suggests a higher eps, but doesn't consider the min points. I tried a few values and these seemed to give a reasonable result - it is a limitation and other methods (HDBSCAN) can overcome it.

```
library(sp)

#first extract the points from the spatial points data frame
points_todf <- points_sub_2020_projected_sp %>%
  coordinates(.)%>%
  as.data.frame()

#now run the dbscan analysis
points_todf_DBSCAN <- points_todf %>%
  fpc::dbscan(.,eps = 1000, MinPts = 50)

points_todf%>%
  dbscan::kNNdistplot(.,k=50)

#now quickly plot the results
plot(points_todf_DBSCAN, points_todf, main = "DBSCAN Output", frame = F)
plot(community_areas_projected$geometry, add=T)
```

DBSCAN offers a spatial angel to the see the clustering on a map and compare to the New York poverty map in 2007, we can see a similar concentration between eviction and poverty.

Add the cluster information to our original dataframe

```
points_todf<- points_todf %>%
  mutate(dbcluster=points_todf_DBSCAN$cluster)
```

Convert our original data frame to a sf object again

```
tosf <- points_todf%>%
  st_as_sf(., coords = c("coords.x1", "coords.x2"),
                   crs = 6538)%>%
  filter(dbcluster>0)
```

Map the data - remember we are adding layers one by one

```
ggplot(data = community_areas_projected) +
  # add the geometry of the community areas
  geom_sf() +
  # add the geometry of the points - i have had to set the data here to add the layer
  geom_sf(data = tosf, size = 0.4, colour=tosf$dbcluster, fill=tosf$dbcluster)
```

```
library(tmap)
library(sf)

#tmaptools::palette_explorer()
library(RColorBrewer)
library(tmaptools)
colours<- get_brewer_pal("Set1", n = 19)

tmap_mode("plot")
tm_shape(community_areas) +
  tm_polygons(col = NA, alpha = 0.5) +
```

```
tm_shape(tosf) +
  tm_dots(col = "dbcluster",  palette = colours, style = "cat")
```

Now, what could be related to this? After a very quick Google it appears there locations have a higher percent of the population living in poverty: https://www.visualizingeconomics.com/blog/2007/09/22/new-york-city-poverty-map

Although, this map is from 2007 and poverty itself is not defined. But it appears these clusters identified are related to some underlying variable that we haven't accounted for here.

5 Clustering

Check eviction density on the map.

```
library(sf)

#basic join but gives a new row for each point (e.g. 10 points in borough 1 then 10 rows)
check_example <- community_areas_projected%>%
  st_join(tosf)

# spatial join that counts the points per borough
points_sf_joined <- community_areas_projected%>%
  mutate(n = lengths(st_intersects(., tosf)))%>%
  janitor::clean_names()%>%
  #calculate area
  mutate(area=st_area(.))%>%
  #then density of the points per ward
  mutate(density=n/area)


# quick map
tm_shape(points_sf_joined) +
    tm_polygons("density",
        style="jenks",
        palette="PuOr",
        title="Eviction density")
```

So, from the map, it looks as though we might have some clustering of eviction in Bronx NY so let's check this with Moran's I and some other statistics.

Before being able to calculate Moran's I and any similar statistics, we need to first define a Wij spatial weights matrix.

A spatial weight matrix represents the spatial element of our data, this means we are trying to conceptualize and model how parts of the data are linked (or not linked) to each other spatially.

```
library(spdep)
#First calculate the centroids of all communities
coordsW <- points_sf_joined%>%
  st_centroid()%>%
  st_geometry()

plot(coordsW,axes=TRUE)

# make neighbors list
```

```
community_nb <- points_sf_joined %>%
  poly2nb(., queen=T)

summary(community_nb)

#Here it is telling us that the average number of neighbours is 4.39.

#plot them
plot(community_nb, st_geometry(coordsW), col="red")
#add a map underneath
plot(points_sf_joined$geometry, add=T)
```

From the weights list we must now make a spatial weight matrix.

```
# make weight matrix
community_nb.lw <- community_nb %>%
  nb2mat(., style="W")

sum(community_nb.lw)

# make weight list for Moran's I

community_nb.lw <- community_nb %>%
  nb2listw(., style="W")
```

Summing the binary (1/0) shows that we have 71 neighbours.

5.0.1 Moran's I DBSCAN focuses on density-based clustering analysis, and Moran's I evaluates the spatial autocorrelation in a dataset.

Now we want to use Moran's i as the current methods does not permit us to ascertain whether the eviction in New York are influenced by the evictions located in neighboring communities, forming what would be termed a spatial locational cluster. Nor did we consider dynamic spatial spillovers. In this vain, global Moran's i allows us to pinpoint the existence of a spatial dependence pattern and its sign. Moran's I test tells us whether we have clustered values (close to 1) or dispersed values (close to -1).

```
I_LWard_Global_Density <- points_sf_joined %>%
  pull(density) %>%
  as.vector()%>%
  moran.test(., community_nb.lw)

I_LWard_Global_Density
```

Moran I statistic: 0.63874955. this value usually ranges between -1 (perfect negative correlation) and +1 (perfect positive correlation). A value close to +1 indicates a trend towards similarity or clustering in the spatial distribution, while a value close to -1 indicates a trend towards heterogeneity or dispersion in the spatial distribution. This value (around 0.64) indicates that the data exhibit positive spatial autocorrelation, i.e. similar values tend to cluster spatially.

Standard Deviation: 8.508. This value is the standard deviation of the Moran's I statistic and is used to help determine the significance of the Moran's I statistic.

P-value: $< 2.2e\text{-}16$. this is a very small value indicating that the test results are highly statistically significant. In traditional hypothesis testing, a p-value less than 0.05 is usually considered statistically significant. The p-value here is much smaller than this threshold, strongly suggesting that the observed spatial autocorrelation did not arise by chance.

Expected value: -0.01428571. with a random distribution, the expected value of Moran's I is close to 0. The expected value of this test is slightly negative, but close to 0. The expected value of Moran's I is close to 0.

Variance: 0.00589142. this is the variance of the Moran's I statistic, which is used to calculate the standard deviation and the p-value.

In summary, the results of this Moran's I test indicate that the data are spatially significantly positively autocorrelated. This means that spatially close observations tend to have similar values, indicating that there may be some kind of spatial pattern or spatial aggregation.

Global Moran's I is: - a global statistic that provides us with a single value for our entire dataset to describe if a variable of interest corresponds to the first law of geography - everything is related, but things that are closer together are more related than things further away. - Global Moran's I operates by comparing how similar every object (such as a census tract) is to its neighbors, and then averaging out all of these comparisons to give us an overall impression about the spatial pattern of the variable.

Local Moran's I is: - To dig a little deeper and understand exactly which objects are similar or different to the objects in their neighborhood.The Local Moran's I statistic is relatively similar to the Global Moran's I in that it is providing a measure of how similar locations are to their neighbors. However, the difference is that each location, i, receive its own I value, as well as its own variance, z value, expected I, and variance of I.

- The difference between a value and neighbours * the sum of differences between neighbours and the mean
- Where the the difference between a value and neighbours is divided by the standard deviation (how much values in neighbourhood vary about the mean)
- Z-score is how many standard deviations a value is away (above or below) from the mean

```
I_LWard_Local_density <- points_sf_joined %>%
  pull(density) %>%
  as.vector()%>%
  localmoran(., community_nb.lw)%>%
  as_tibble()
```

```
points_sf_joined <- points_sf_joined %>%
  mutate(density_I =as.numeric(I_LWard_Local_density$Ii))%>%
  mutate(density_Iz =as.numeric(I_LWard_Local_density$Z.Ii))
```

Map

```
breaks1<-c(-1000,-2.58,-1.96,-1.65,1.65,1.96,2.58,1000)

library(RColorBrewer)
MoranColours<- rev(brewer.pal(8, "RdGy"))

tm_shape(points_sf_joined) +
    tm_polygons("density_Iz",
        style="fixed",
        breaks=breaks1,
        palette=MoranColours,
        midpoint=NA,
        title="Local Moran's I, Evicitions in New York")
```

This map shows areas in the Bronx that have relatively high scores, indicating areas has similarity(eviction) between nearby observations.

6 DISCUSS

This result allows us to reject the hypothesis of homogeneity, and we can therefore say the distribution of eviction in New York 2020 displays a deterministic and not a non-random pattern.This confirms that communities exhibit eviction frequencies similar to the participation of their neighboring communities, and that community with high eviction values and those with low location (high or low) influence each other.This finding implies that eviction in neighboring communities are correlated with the eviction in the reference community and tent to group spatially. This triggers spatial spillover effects as well as negative externalities stemming from social influence. This means that positive spatial autocorrelation is greater and, therefore, that spatial interaction is deeply intervened in the society.

As the New York government wants to conduct a study that aim to prevent people being evicted in 2020, this research has set a grounded scientific foundation to understand the spatial autocorrelation in New York eviction. Result shows that there is a positive spatial autocorrelation of evictions and are mainly concentrated in Bronx area with spillover effects in adjacent communities. It is interesting to compare the result to New York City Poverty map, Bronx is not only one of the most poverty concentrated area, but also the most eviction concentrated area. Although there are other poverty concentrated area, it is worth questioning why the evictions is not as intensely happened elsewhere. DBSCAN offers a spatial angel to the see the clustering on a map and compare to the New York poverty map in 2007, we can see a similar concentration between eviction and poverty.While the poverty data was from 2007, and the eviction was from 2020, it demonstrates that over a decade, New York is still has the same deprivation problem. It is urgent for the government to put more focus on targeting preventing people from deprivation.

6.1 Reflection

Firstly, it would be more insightful to carry local Moran's I test closely in Bronx NY area using micro-geographic data to check for further analysis and offer more detailed patterns for depth policy suggestions. Additional regression analysis with factors like poverty, housing price and other deprivation index would offer more information on why eviction would happen in those areas.

Secondly, in many of the events of an social nature, spatial dependence between areas is the result of trends that occur due to a spatial-temporal effect. In this way, the spatial-temporal Moran's It statistic enables to evaluate the time evolution of spatial dependence, and provides information concerning the type of spatial dependence whether instantaneous or contemporary, as well as lagged or non-contemporary. In our case, analysis spans different years can see both spatial and temporal changes and see the efficiency of policy enforcement and social changes across the neighborhood communities before 2020 to understand the possible factors that got people being evicted.

Finally, as the presented eviction data was collected by different police, their territorial bias could reflected on the result, therefore, the policy makers need to take the bias into the consideration.

- Evictions can be complicated as it is likely these include **NO FAULT** evictions - where the landlord wants to remove the tenant for no fault. This year New York have to moved to a good cause eviction - meaning there must be a cause to actually evict someone. How might this influence our data?
- Whilst we have looked at point data we haven't considered the number of housing units in each area.
  - Might we expect more evictions in an area with more housing?
  - And as our min points or distance is set for the whole study area how will this influence our results?
  - What other methods allow the distance parameter to vary?….https://cran.r-project.org/web/packages/dbscan/vign
  - If we were to do spatial autocorrelation we would need to create some sort of rate, see: https://council.nyc.gov/data/evictions/#:~:text=The%20dataset%20is%20updated%20daily,addresses%20are%20cleaned
- E.g. now i have clusters, could i extract the community districts and then look at some other data (e.g. census data) to explore factors that might influence evictions?

6.2 references

https://andrewmaclachlan.github.io/CASA0005repo/spatial-autocorrelation.html# https://data.cityofnewyork.us/City-Government/Evictions/6z8x-wfk4 https://data.cityofnewyork.us/City-Government/Community-Districts/yfnk-k7r4 https://www.visualizingeconomics.com/blog/2007/09/22/new-york-city-poverty-map https://rpubs.com/quarcs-lab/spatial-autocorrelation