Spatial Analysis of COVID-19 cases in New York City

Introduction

By December 18, the United States had 17,391,270 COVID-19 cases and 312,636 deaths since its initial detection. (CDC, 2020)New York State was once an epicenter with 828,200 cases. (*COVID-19 United States Cases by County*, n.d.) New York City (NYC) is responsible for the majority of the cases in New York State, with more than 378,000 cases. Many factors might lead to the situation, as previous traditional studies demonstrated in other communicable diseases. (Kelly et al., 2016; Simpson et al., 2019; Tung et al., 2019; Wiewel et al., 2019) Therefore, I have come up with three assumptions and want to test them. If I falsify some of the assumptions, I will try to figure out the potential reasons and limitations in this project.

The hypotheses are as followed,

1. The higher spatial density leads to a higher incidence rate.
2. Having difficulties in access to hospitals causes a higher case fatality rate.
3. The lower hospital capacity results in a higher case fatality rate.

Methods

Software

All the maps were spatialized using ArcGIS 10.7 and ArcCatalog 10.7.1. The data was compiled using the Vlookup function in Excel. The statistical analysis was conducted by R studio.

Data sources

The New York City Zip Code Boundaries shapefile was obtained through *NYC Open Data.* (Calgary, n.d.) The COVID-19 case number, incidence rate, death number, case fatality rate, and population by zip code were collected from *NYC Health*, including all data from the first confirmed cases diagnosed on February 29 to December 17. (*COVID-19: Data Totals - NYC Health*, n.d.) The NYC hospital location by zip code shapefile was downloaded from *NYU Spatial Data Repository*, which includes hospital addresses, the number of certified hospital beds (capacity) reported by the NYS Department of Health December 2014. (*2015 New York City Hospitals - NYU Spatial Data Repository*, n.d.) The 2019 NYC average family size by zip code was obtained from the *US census*. (Bureau, n.d.)

The relationship between NYC population density and incidence rate (Map 1)

The NYC Zip Code Boundaries was added into the layer named *"NYC_Zipcode"*. The population and incidence rate were imported to the attribute table of the layer *"NYC_Zipcode"* by **Joining**. A new field was added into the attribute table named *"Area,"* and **Calculate Geometry** was used to calculate each zip code area. The population density (population normalized by zip code area) was displayed with graduated color classified by Quantile. The incidence rate was displayed with graduated symbols classified by Natural Breaks.

Identification of hotspots and coldspots (Map 2)

First, the incidence rate was displayed with graduated color classified by Natural Breaks. Then, **Getis-Ord Gi*** was utilized to identify the hotspots and coldspots.

The relationship between hotspots/coldspots and population density (Map 3)

To outline hotspots and coldspots, two new shapefiles were created in ArcCatalog. The outlines were sketched utilizing creating new features and sketching tools in ArcMap. The population density layer with graduated color in the previous step was added. The hotspots and coldspots outline layers were added above.

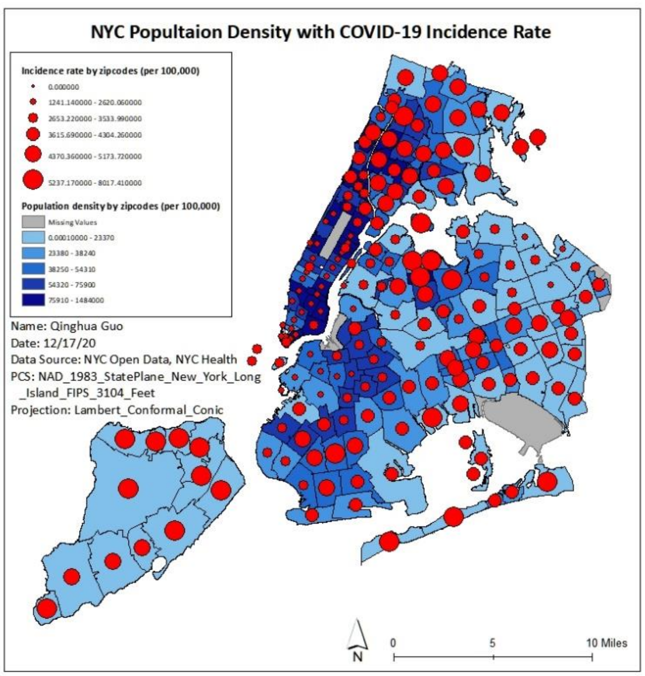
The relationship between hotspots/coldspots and average family size (Map 4)

The average family size data was imported to the attribute table of the layer *"NYC_Zipcode"* by **Joining**. The average family size was displayed with graduated color classified by Natural Breaks. The hotspots and coldspots outline layers were added above.

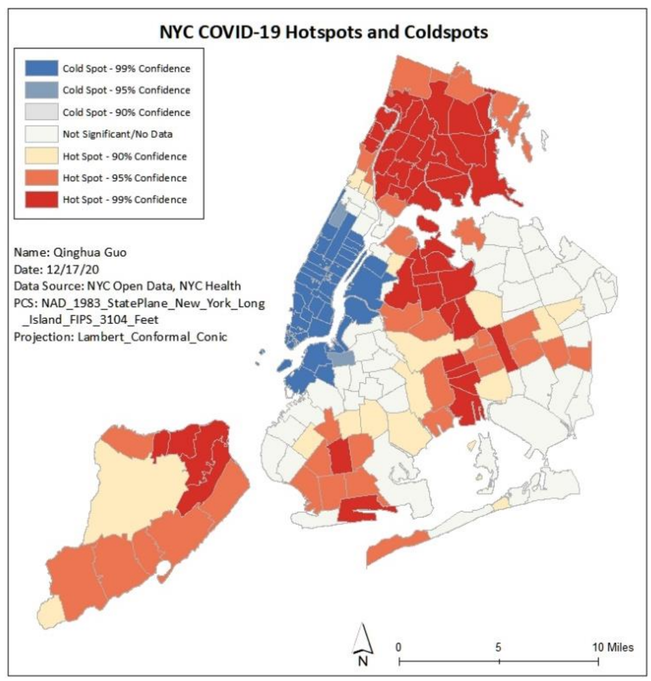
The relationship between hospital locations and case fatality rate (Map 5)

The hospital location by zip code was added into the layer and displayed by symbols. The case fatality rate was imported to the attribute table of the layer *"NYC_Zipcode"* by **Joining**. The case fatality rate was displayed with graduated color classified by Quantile.

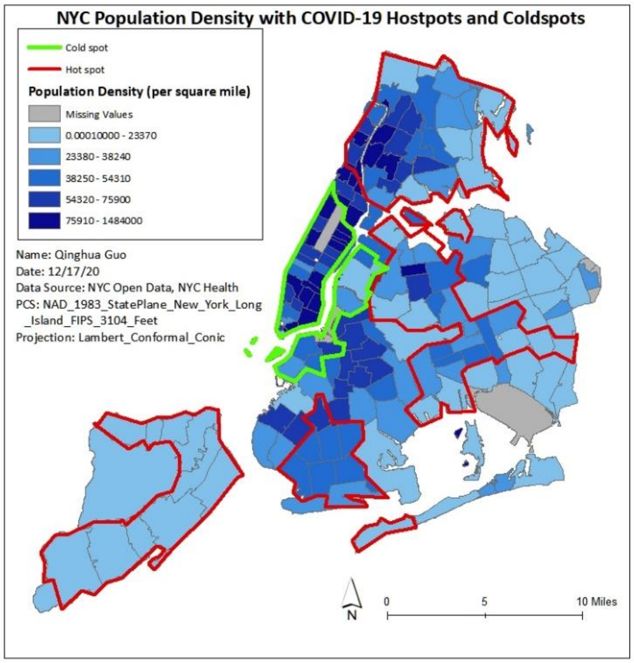
Results



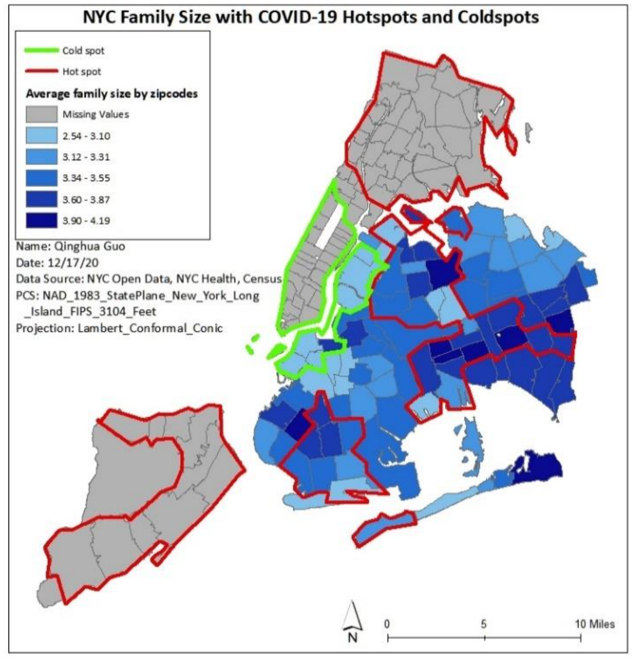
Map 1



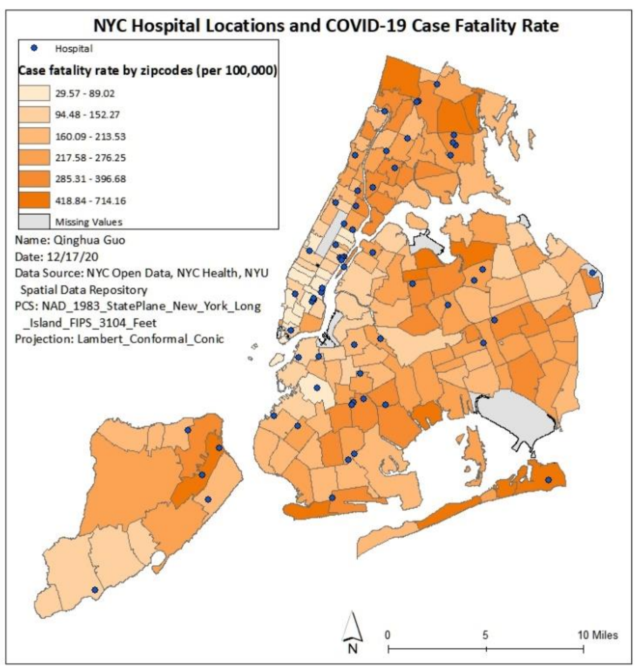
Map 2



Map 3



Map 4



Map 5

The relationship between NYC population density and incidence rate

From Map 1 we cannot see a clear relationship between the NYC population density and incidence rate. For example, we can see clustering in the population density in Manhattan, but the incidence rate is relatively lower than in other places. Therefore, I decided to conduct a hotspot analysis to figure out more accurate relationships. Map 2 is the product of the hotspot analysis. Clusters over 95% confidence were included and outlined in Map 3. A high z-score and a low p-value indicate a significant hotspot. In contrast, a low negative z-score and a low p-value indicate a significant coldspot. A bin value stands for the confidence level (3, 2, 1 refer to 99%, 95%, 90%, respectively). There are 4 main hotspots located in Brooklyn, Queens, Staten Island, and the Bronx. The 2 main coldspots are located in Manhattan and the west coast of Queens and Brooklyn.

Obviously, we cannot figure out a strong association between population density and incidence rate. This might be because of the different socioeconomic status of people clustering in Manhattan, who are more likely to afford health care and more active to prevent the disease due to their higher income and education levels.

The relationship between hotspots/coldspots and average family size

Family size is a potential indicator of living density. As we can see in Map 4, areas with higher average family size (3.60-4.19 people per family) are mostly associated with the hotspots. We can conclude that the average family size might be a better indicator than population density to reflect the spatial density. This project has come to a similar conclusion as other studies. The spatial density is associated with infectious and communicable diseases. Notably, the spatial density should be the variable related to the household size or shared living spaces instead of the population density in an area. (Simpson et al., 2019) Because even though many people are living in the same neighborhood, they cannot transmit the virus by living alone and not getting in touch with each other.

The relationship between hospital locations and case fatality rate

Map 5 represents the hospital locations and the case fatality rate in its respective zip code. Two statistical analyses were conducted. Two assumptions were tested in this section. The first is to compare the case fatality rate in places with hospitals versus places without hospitals using the Chi-square test. The hypothesis is that places with hospitals have a lower case fatality rate than those without hospitals. The result shows weak associations (X-squared = 175, df = 174, p-value = 0.4644). The second is to see whether higher hospital capacity causes lower case fatality rates using linear regression. No significant association is found (p-value = 0.394)

Therefore, we cannot tell the association between hospital locations and case fatality rate or hospital capacity and case fatality rate through these studies. There might be several reasons. Due to the dramatic rise of COVID-19 cases and deaths, new temporary hospitals with more beds have been established, which are not listed in this project. On the other hand, the huge number of cases have far more exceeded the hospital's actual capacity, making the result not adaptable in this case.

Conclusion

The population density does not affect the COVID-19 incidence rate. People living in a larger family are more likely to get infected with COVID-19. The access to hospitals and the hospital capacity have no strong association with the case fatality rate.

Discussion/Limitations

There are some missing data values, including population, incidence rate, case fatality rate, and average family size. Specifically, there are 63 missing data in the population and 71 missing data in the incidence rate. All the data, except for the COVID-19 cases and deaths, are not collected in 2020, which might be out of date. The COVID-19 case and death number largely depend on the diagnosis's capacity, so these numbers are likely underreported or false positive/negative. To analyze the relationships between hotspots/coldspots and spatial density, advanced statistical analysis is needed. The map could only be a facilitator to visualize the result. The way to sketch the hotspots and coldspots' outline is through manual outlining, which might be imprecise. Therefore, it needs to be more accurate and precise through some special functions in ArcGIS. Average family size data are only available in Brooklyn and Queens. However, most of the cold spots are located in Manhattan, which affects the analysis of the relationship between cold spots and spatial density.

If we want to assess whether the hospital's accessibility affects the case fatality rate, the buffer function or calculating service area can be applied to measure hospitals' distance. However, it's hard to decide the distance. Other papers used 5 miles and 8 miles to measure the accessibility to diagnostic hospitals and other health facilities in NYC. In this case, I have tried 3 miles, there are too many overlaps covering the whole of NYC, so I do not think it could be the right choice. Furthermore, due to the extensive transportation system in NYC, including subways and buses, hospital locations might not be an appropriate contributor to hospitals' accessibility. Instead, more socioeconomic factors should be considered to access the hospital.

References

2015 New York City Hospitals—NYU Spatial Data Repository. (n.d.). Retrieved December 19, 2020, from https://geo.nyu.edu/catalog/nyu-2451-34494

Bureau, U. C. (n.d.). *Census.gov*. Census.Gov. Retrieved December 19, 2020, from https://www.census.gov/en.html

Calgary, O. (n.d.). *Zip Code Boundaries*. NYC Open Data. Retrieved December 19, 2020, from https://data.cityofnewyork.us/Business/Zip-Code-Boundaries/i8iw-xf4u/data?no_mobile=true

CDC. (2020, March 28). *COVID-19 Cases, Deaths, and Trends in the US | CDC COVID Data Tracker*. Centers for Disease Control and Prevention. https://covid.cdc.gov/covid-data-tracker

COVID-19: Data Totals—NYC Health. (n.d.). Retrieved December 19, 2020, from https://www1.nyc.gov/site/doh/covid/covid-19-data-totals.page

COVID-19 United States Cases by County. (n.d.). Johns Hopkins Coronavirus Resource Center. Retrieved December 19, 2020, from https://coronavirus.jhu.edu/us-map

Kelly, C., Hulme, C., Farragher, T., & Clarke, G. (2016). Are differences in travel time or distance to healthcare for adults in global north countries associated with an impact on health outcomes? A systematic review. *BMJ Open*, *6*(11). https://doi.org/10.1136/bmjopen-2016-013059

Simpson, P. L., Simpson, M., Adily, A., Grant, L., & Butler, T. (2019). Prison cell spatial density and infectious and communicable diseases: A systematic review. *BMJ Open*, *9*(7). https://doi.org/10.1136/bmjopen-2018-026806

Tung, E. L., Hampton, D. A., Kolak, M., Rogers, S. O., Yang, J. P., & Peek, M. E. (2019). Race/Ethnicity and Geographic Access to Urban Trauma Care. *JAMA Network Open*, *2*(3). https://doi.org/10.1001/jamanetworkopen.2019.0138

Wiewel, E. W., Borrell, L. N., Jones, H. E., Maroko, A. R., & Torian, L. V. (2019). Healthcare facility characteristics associated with achievement and maintenance of HIV viral suppression among persons newly diagnosed with HIV in New York City. *AIDS Care*, *31*(12), 1484–1493. https://doi.org/10.1080/09540121.2019.1595517