Segmentation of contagion zones in New York State

Fanirisoa Rahantamialisoa

28 avril 2021

1 Introduction

1.1 Background

COVID-19 has turned into a global crisis, evolving at unprecedented speed and scale. It is creating a universal imperative for governments and organizations to take immediate action to protect their people.

1.2 Problem

One of the most important properties of epidemics is their spatial spread, a characteristic which mainly depends on the epidemic mechanism and human mobility. In this study, we are developing a geospatial and spatio-statistical analysis of the geographic dimension of the 2019 coronavirus disease pandemic (COVID-19) in New York state. The restrictions put in place to limit the diffusion and impacts of Covid-19 have had a widespread impact on people's lives.

1.3 Interest

The aim of this study is to investigate a geographic and geospatial analysis to understand the locations and distribution patterns of COVID-19 in order to be able to define and segment the zones of contagion. The study seeks to highlight the mobility dynamics of the urban population as the process of leaving from home, traveling to and from the activity locations, and engaging in activities the urban transportation system may alter the fundamental dynamics of the infectious disease, change the number of secondary infections, promote the synchronization of the disease across the city, and affect the peak of the disease outbreaks.

2 Data acquisition and cleaning

In this report we mostly focus our attention on New York state. The poject collects, analyses and uses a variety of data from multiple sources. The main types of counties that are collected and managed include :

— Geographic informations for each county,
— Metrics and indicators for each groupe of venus inside county,
— Official Statistical data on the COVID19 situation in each county.

As the sources used to collect such data are heterogeneous and diverse in nature, an Extract-Transform Load (ETL) process is needed in each case to extract the raw data from the respective source and transform them appropriately to the corresponding representation and schema needed for being used into the analysis.

More specifically, a collection of ETL scripts has been designed and implemented for cleaning, converting and importing all collected datasets into the project repository. In general, this process proceeds in the following successive stages :

— **Extract data** : mainly, three types of cases can be distinguished, requiring different methods with varying levels of difficulty :
  1. US Zip Code Latitude and Longitude (using Webscraping)
  2. Foursquare location data,
  3. Statistical data on the COVID19 situation (using Webscraping).
— **Transform data** to fit internal representation format and schema. This stage involves selecting attributes, joining or aggregating information from multiple records, splitting/merging values from attributes, etc.
— **Load data** into the repository. At this stage, previously transformed data are imported into the repository. This involves updating or replacing existing information with more recent one.

Such ETL operations for collecting data into repository were implemented as custom, parametrized scripts written in Python, developed specifically to manipulate each type of information. If wished for, our codes can be made available on Github to wider audience. In the rest of this section, we list the different types of data that have been collected and imported.

2.1 Geographical Zones

A custom Python script is used to extract by using Webscraping from US Zip Code Latitude and Longitude and to load the records into table zones in pandas dataframe :

	State	County	Latitude	Longitude	geometry
0	New York	St. Lawrence	+44.4881125	-075.0743110	"{""type"": ""Polygon"", ""coordinates"": [[[-...
1	New York	Onondaga	+43.0065163	-076.1961336	"{""type"": ""Polygon"", ""coordinates"": [[[-...
2	New York	Monroe	+43.4644839	-077.6646584	"{""type"": ""Polygon"", ""coordinates"": [[[-...
3	New York	Schoharie	+42.5912940	-074.4381718	"{""type"": ""Polygon"", ""coordinates"": [[[-...
4	New York	Kings	+40.6350451	-073.9506398	"{""type"": ""Polygon"", ""coordinates"": [[[-...
5	New York	Nassau	+40.7296118	-073.5894144	"{""type"": ""Polygon"", ""coordinates"": [[[-...
6	New York	Rensselaer	+42.7104206	-073.5138454	"{""type"": ""Polygon"", ""coordinates"": [[[-...
7	New York	Oswego	+43.4614431	-076.2092618	"{""type"": ""Polygon"", ""coordinates"": [[[-...
8	New York	Otsego	+42.6297762	-075.0288410	"{""type"": ""Polygon"", ""coordinates"": [[[-...
9	New York	Clinton	+44.7527120	-073.7056429	"{""type"": ""Polygon"", ""coordinates"": [[[-...

FIGURE 1 – US Zip Code for New York County :

As illustrated in Figure 2 , this dataset contains 62 counties polygons :

FIGURE 2 – Counties of New York.

2.2 Metrics for each Zones

Extract from foursquare, as informations and metrics, we consider :
— the number of bus stops and metro stations (the more trips, the greater the risk of transmission),
— the number of cinemas (the more gatherings, the greater the risk of transmission),
— the number of schools (children are vectors of transmission),
— the number of hospitals,
— the number of shopping centers.
— tips : contains the total count of tips and groups with friends and others as groupTypes. Groups may change over time.
— like : the count of users who have liked this venue, and groups containing any friends and others who have liked it. The groups included are subject to change.
— rating : numerical rating of the venue (0 through 10). Returned as part of an explore result, excluded in search results. Not all venues will have a rating.

As illustrate in the Figure 3, We obtain a dataframe with 62 rows and 19 variables.

	County	Likes_Resto	Likes_Schools	Likes_Bus	Likes_Metro	Likes_Shopping	Likes_Hospital
0	St. Lawrence	63	97	24	87	50	73
1	Onondaga	42	19	49	85	35	91
2	Monroe	78	53	86	83	38	23
3	Schoharie	79	9	59	15	19	15
4	Kings	17	97	83	55	14	42
5	Nassau	7	26	28	91	49	48
6	Rensselaer	36	54	6	30	62	23
7	Oswego	92	40	55	84	18	70
8	Otsego	31	39	95	40	47	52
9	Clinton	11	87	93	85	56	76
10	Erie	48	13	35	31	32	35
11	Chautauqua	87	57	85	40	31	35
12	Dutchess	42	50	19	13	16	30
13	Cortland	85	30	73	71	95	80
14	Richmond	58	38	22	95	79	99
15	Saratoga	31	72	81	74	34	62
16	Hamilton	41	86	9	46	26	97
17	Yates	52	57	14	96	29	42
18	Tioga	75	10	69	65	52	18
19	Tompkins	15	44	51	66	76	98

FIGURE 3 – Counties of New York.

2.3 Statistical data on the COVID19

Imports updated statistical data on the COVID19 situation in each county in Mai 2020 and September 2020 : number new contaminates (J-1), number of deaths and number of cases :

- Cases : The cumulative number of confirmed human cases reported to date. The actual number of infections is likely to be higher than reported.
- Deaths : The cumulative number of confirmed human deaths reported to date. Reporting criteria vary between locations.
- Recov : The cumulative Total of Recovered. May not correspond to actual current figures and not all recoveries may be reported.
- Pop : The total population of the county reported to date.
- Cases 100k : The Ratio of cumulative number of confirmed human cases reported per 100.000 population.
- Deaths 100k : The Ratio of cumulative number of confirmed human deaths reported per 100.000 population.
- Ratio Deaths Cases : The Ratio between cumulative number of cases and cumulative number of cases deaths.

Data will be obtained from Webscraping, we obtain a dataframe with 62 rows and 8 variables.

	County	Cases_sep	Deaths_sep	Recov_sep	Pop_sep	Cases_100k_sep	Deaths_100k_sep
0	Albany	1986.0	108.0	1457.0	305506.0	556.5	30.4
1	Allegany	58.0	2.0	41.0	46430.0	124.9	4.3
2	Bronx	46778.0	3295.0	36.0	1418207.0	3242.1	229.7
3	Broome	657.0	56.0	489.0	190488.0	236.8	20.5
4	Cattaraugus	145.0	4.0	100.0	76117.0	161.6	5.3
...	...	...	...	...	...	...	...
57	Washington	250.0	14.0	212.0	61204.0	362.7	22.9
58	Wayne	137.0	2.0	74.0	89918.0	115.7	2.2
59	Westchester	34385.0	1407.0	21427.0	967506.0	3509.4	143.9
60	Wyoming	95.0	5.0	63.0	39859.0	238.3	12.5
61	Yates	54.0	7.0	45.0	24913.0	188.7	28.1

62 rows × 8 columns

FIGURE 4 – Statistical data on the COVID19 may 2020.

A common practice is to normalize the values of each feature before using them for clustering. Accordingly we use MinMaxScaler, included in the scikit-learn. This computes the minimum and maximum value of each attribute in the training set, and uses these to normalize all values in the training and the test set to the $[0, 1]$ range. In our case, the values of many features are extremely skewed ;

3 Exploratory Data Analysis

3.1 Number of like for restaurant & place for shopping

The current growth of the fashion factory outlets results in the establishment of restaurants and cafes to capture these outside visitors. The relationship between the two variables is not so obvious across all counties as illustrate in the following Figure :

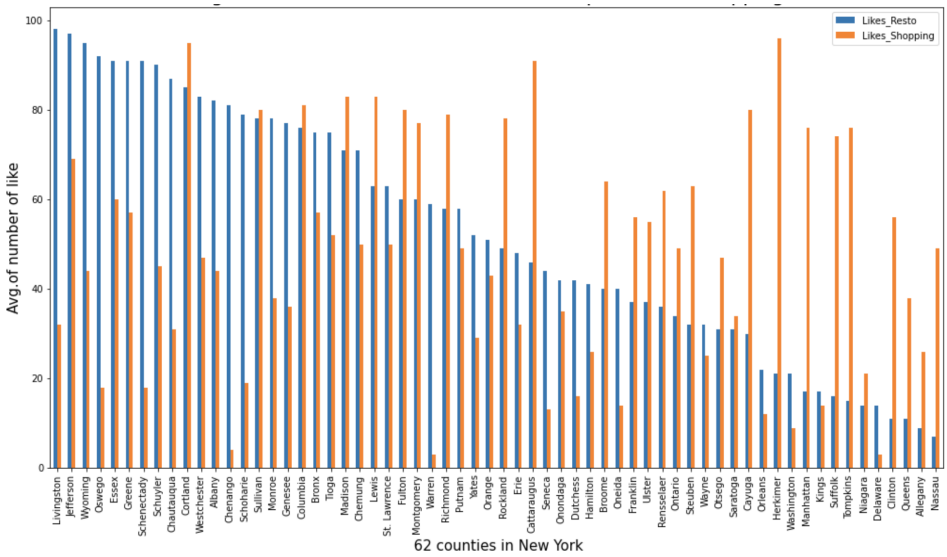

FIGURE 5 – Average of number of like for restaurant and place for shopping.

3.2 Number of cases in May 2020 & in September 2020

About 42% of New York State residents live in New York City, so COVID stats for the entire state will tend to mirror the city's stats. This is true for positivity, which shows the same large increase in April when testing was limited, and for total tests, which have risen steadily since then. However, when looking at positive tests, you see that during the second wave, the state suffered a more pronounced rise than the city, because of large second-wave outbreaks in upstate counties. Similarly, the increase in hospitalizations and deaths in New York State is more pronounced than in New York City, because of these upstate outbreaks.

Queens and Brooklyn have larger populations than the other boroughs, so they tend to have more cases, but when normalized for population, Staten Island and the Bronx have often seen more positives each week. This is particularly

true during the second wave, with Staten Island outstripping the other boroughs.

The statistics obtained over time on the numbers of cases and deaths due to the pandemic, highlight an explicit link between the number of deaths in May 2020 and the number of cases in September 2020. This pattern is more explicit between the numbers of deaths recorded in May and September 2020 as illustrated by the Figure 6.

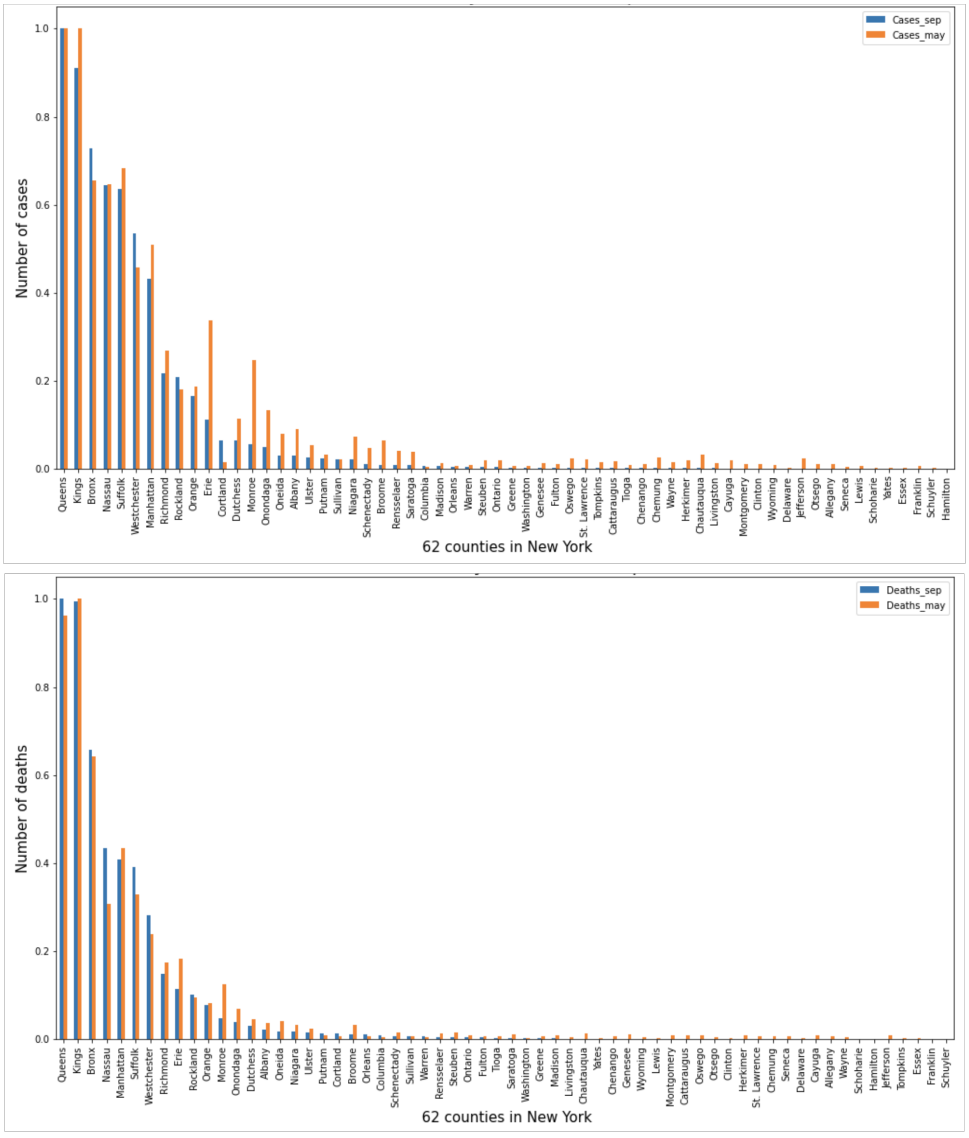

FIGURE 6 – Number of cases and deaths in may 2020 and in september 2020.

3.3 Cases, deaths density per km^2 & population density per km^2

We used a Bubble Plot to highlight the relationship between the number of confirmed cases, the number of deaths and the number of population per km2 :

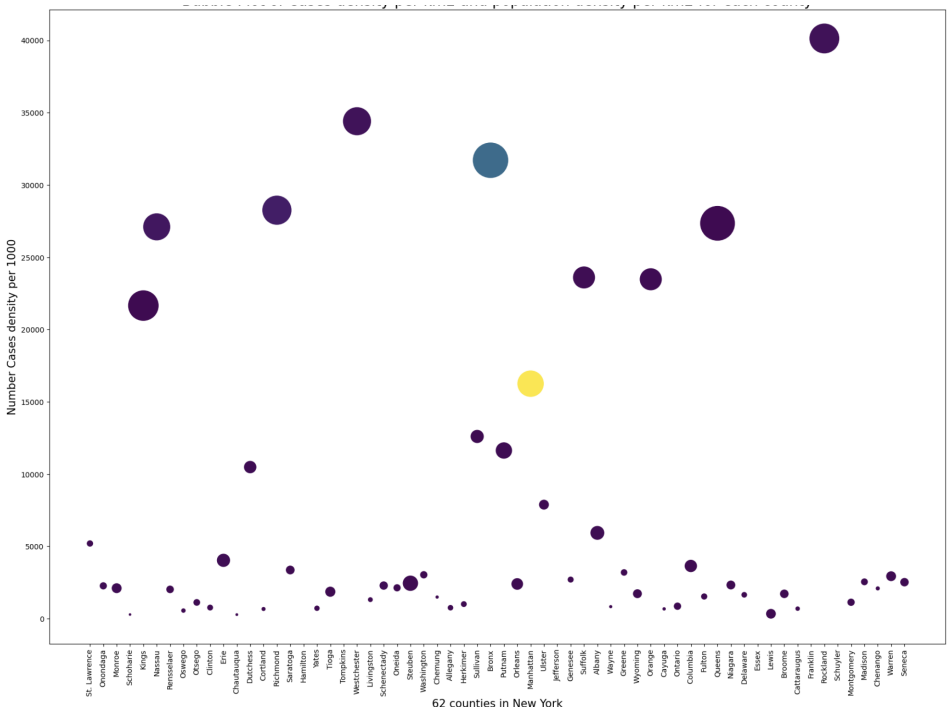

FIGURE 7 – Bubble Plot of Cases, death density per km2 and population density per km2 for each county.

The result does not make it possible to directly highlight the relationship between the 3 dimensions of confirmed cases, recorded deaths and number of population per km2. In the figure, we have the size of the sphere which represents the number of deaths, and the color the number of population.

4 Analysis of Zone-Based Clustering

Under these updated metrics, the goal is to Identify the set of objects with similar characteristic. Clustering can be considered the most important unsupervised learning problem ; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. Micro-cluster zones will now be determined using K-means.

We can build the K-Means in python using the KMeans algorithm provided by the scikit-learn package. The KMeans class has many parameters that can be used, but we will be using these three :

— Init : Initialization method of the centroids. The value will be : $k-means++$ Selects initial cluster centers for the $k-means$ clustering in a smart way to speed up convergence..

— $n_clusters$: The number of clusters to form as well as the number of centroids to generate. The value will be 5

— n_init : Number of times the k-means algorithm will be run with different centroid seeds

The final results will be the best output of n_init consecutive runs in terms of inertia. The value will be 12 After building the model, we will be fitting and define a variable 'labels' to store the cluster labels of the built model. Let's do it in python. The class of cluster for each county are represented inn the following :

```
[4 4 4 1 2 2 4 1 1 1 4 1 3 1 2 4 1 1 1 1 5 1 4 4 1 4 1 1 1 3 5 3 4 3 4 1 4
 2 4 1 4 1 2 1 1 4 4 2 4 1 1 1 1 1 5 1 1 4 1 4 1]
```

FIGURE 8 – The result of the K-means clustering analysis for all the 62 counties.

According to Figure 9, it can be found that as the value of k equal to 5.

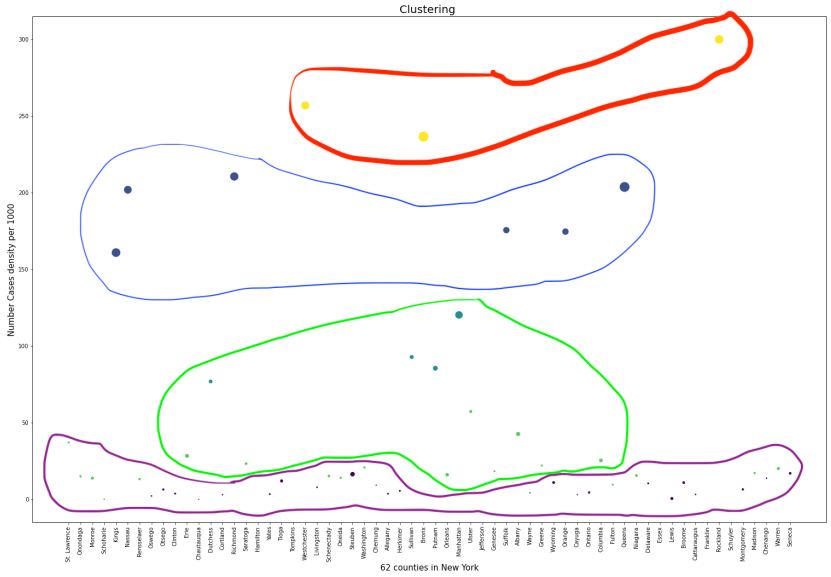

FIGURE 9 – The result of the K-means clustering analysis for all the 62 counties.

A clustering-based approach can be used to solve the problems. Clustering involves grouping of similar elements of a given set of elements. By analyzing

the clusters, we discover common or discriminative factors among the clusters that are likely to explore whether clusters of contacts could better explain the transmission of infectious diseases.

The results of clustering are highly dependent on the features used. The result reveals that 5 region clusters were created from 62 counties, represented using different colors. Generally, geographically adjacent regions were more likely to be grouped into the same cluster. This result is consistent with previous studies. However, we did not find significant differences between clusters, unlike the disease-based clusters or time-based clusters. That is, the difference in the occurrence of infectious diseases was too small to separate into clusters.

5 Conclusion

The k-means algorithm is a well-known partitional clustering algorithm. But the traditional approach selects the number of clusters (k), prior to the clustering process, and randomly selected initial centroids to produce the clusters. Since the initial centroids which is selected randomly will contribute much on the accuracy of this clusters, accuracy cannot be guaranteed.