Michael Borowski

GIS for Health Data Analysis

05/04/2020

Final Project

Coronavirus Clustering in Philadelphia

**Introduction:**

Coronaviruses are a type of virus that comes in many different forms (Sauer, 2020). COVID-19 is the coronavirus that has caught the world by surprise. COVID-19 is highly contagious and leads to a bad cough, shortness of breath, fever, muscle aches, sore throat, loss of taste or smell, headache, diarrhea, and even kidney failure or death in extreme cases (Sauer, 2020). Two strands of COVID-19 have been identifies as SARS-CoV-1 and SARS-CoV-2 ("New coronavirus stable for hours on surfaces", 2020). Information is continuing to expand on the virus and researchers and scientists are working to understand and contain this contagious disease. Research has found that the virus that causes the COVID-19 disease is stable for several hours to days on surfaces and in aerosols ("New coronavirus stable for hours on surfaces", 2020). The Centers for Disease Control and Prevention (CDC) has issued many guidelines to help protect people from, and limit the spread of this virus. Since there is no known cure for the virus as of yet, the CDC suggests practicing social distancing and limiting contact with other humans as much as possible. It is suggested to wash hands as often as possible, especially when in contact with any foreign surfaces that could potentially be infested with the virus. The CDC has reported that wearing masks to cover the mouth and nose are helpful with controlling the spread and many governments have required masks to be worn when entering stores and facilities. The CDC advises these safe practices and to clean and disinfect homes frequently ("How to Protect Yourself & Others", 2020)

This virus has been impactful globally. Government stay at home orders have been made to reduce the spread of the virus and has forced many businesses to temporarily close their doors. As of April 30th 2020, more than 30 million Americans have filed for initial unemployment claims since mid-March (Hammond & Hayes, 2020). All of the 67 counties in Pennsylvania are under a Stay at Home Order that is set to last until May 8th. These counties are all categorized under a “red phase” but some are transitioning to a “yellow phase” to reintegrate daily life following May 8th ("Coronavirus in Pennsylvania", 2020).

This project looks specifically at Philadelphia COVID-19 test results. The spatial extent is spaced by Zip Code regions throughout the county. A cluster analysis is performed in ArcMap 10.6.1 and a python 3.7 code is built to display data acquired from the City of Philadelphia.

**Methods:**

The Data for this project was acquired through three different sites. OpenDataPhilly provided a Zip Code polygon layer, the COVID-19 test results were downloaded as .csv files from the City of Philadelphia’s website, and population data was retrieved for the Census Bureau’s website.

The Census Bureau’s population data provided the populations of every Pennsylvania Zip Code which was limited later in the analysis.

The City of Philadelphia provides updated test results daily. This analysis captured data from March 31st 2020 and April 21st 2020 to show the change over this three week span.

To begin the analysis, the Zip Code polygon layer, test results .csv files, and census data was added to an ArcMap document. The .csv files data was exported as .dbf files in order to create and change the attribute table fields. In order to join the test data to the Zip Code polygon file, a string field was required to match the Zip Code field of the polygon layer to the field contained in the .dbf files. The data from the City of Philadelphia was saved in different tables as they continued to update the page. March 31st data contained all long type fields whereas the April 21st table already contained the Zip Code field as a string. Adding fields to these two tables was required. For March 31st, a field was created as a string type and the field calculator mimicked the long type data as string type. The April 21st data replicated the string Zip Code field as a long type to serve as a unique ID in the spatial autocorrelation process. The April 21st file then used a select by attribute to limit the data as only positive test results due to this file contain data for positive and negative results. This data was exported as a new file and added to the map. Once the fields were created, the layers were able to be joined to the polygon layer and used the ‘save only matching attributes’ to limit the population data to include only the desired Zip Codes in Philadelphia. A new field was created for each date that divided the total positive tests by the population field using a field calculator. The positive results were labeled differently in the March and April dataset according to how the City of Philadelphia provided the data. For March 31st the positive test results were saved under a field called ‘Positives’ and the April 21st positive tests field was labeled “Count”. These two fields were divided by the population data and saved with a meaningful name to help differentiate between the two.

Zip Codes 19109 and 19112 had 0 population and 13 population respectively. To limit outliers, these areas were excluded from the analysis. A select by attribute excluded these two Zip Codes and exported the resulting data as a new file.

The cluster analysis began with creating a Spatial Weigh Matrix for each of the dates included in the report. In ArcMap, the “Generate Spatial Weight Matrix” tool was used. For this tool, the input file was the joined shapefile layer. The next parameter required a unique integer field of each Zip Code. The March 31st columns contained data in a long type which worked for this requirement. The April 21st data required creating the long type field as mentioned above, to pass into this process. The “Output Spatial Weights Matrix File” required naming the file according to which date’s matrix is being used. The “Conceptualization of Spatial Relationships” used ‘CONTIGUITY\_EGDES\_CORNERS’ which is best suited for touching polygons, in this case the Zip Code areas of Philadelphia. Using EUCLIDEAN in the “Distance Method” was the file step in creating weight matrixes. The matrixes were then generated and were ready to be used in the following processes.

The next step calculated Moran’s I. The ArcMap tool “Spatial Autocorrelation (Moran’s I)” was opened. The “Input Feature Class” used the joined shapefile similarly to creating the spatial weight matrixes. The input field was the resulting field that was created by dividing the positive tests by population created previously. The “Generate Report” box was selected to view the resulting outcome. The “Conceptualization of Spatial Relationships” used the pull-down option of ‘Get Spatial Weights from File’ and the ‘Weight Matrix File’ opened the according weight matrix of the date being analyzed. In the “Geoprocessing” tab, the “Results” menu allowed the spatial autocorrelation .html file to be viewed. An issue with ArcMap did not allow this process to be done twice in one session. In order to do both dates, the map document had to be saved and reopened before running the second attempt.

The “Cluster and Outlier Analysis (Anselin Local Moran’s I)” was then used to calculate local indicators of spatial autocorrelation statistics. Again, the joined shapefile was passed to the “Input Feature Class.” This window used the same parameters of the Spatial Autocorrelation (Moran’s I) tool using the positive test results divided by population as the “Input Field.” The “Conceptualization of Spatial Relationships” used the pull-down option of ‘Get Spatial Weights from File’ and the ‘Weight Matrix File’ opened the according weight matrix of the date being used. This process resulted with a new shapefile map being created in the ArcMap document. In the properties tab, the options were set to display only high-high and low-low clusters.

The joined shapefile was then able to be used to create an interactive web map. This report used Spyder to write the python 3.7 script although other python scripting software could be used. The interactive map used folium and required an install. In a prompt window, a new environment was created that installed folium. To install folium, the anaconda window used

conda install folium -c conda-forge

With folium now accessible, the script that was created read:

import folium

import os

import geopandas as gpd

os.chdir('C://Users//tuf97085//Gis\_Health//Final\_project')

corona = 'pos\_pop\_full.shp'

corona\_gdf = gpd.read\_file(corona)

m = folium.Map(location=[40.005, -75.155338], zoom\_start=11, tiles= 'OpenStreetMap', control\_scale=True)

march\_31 = folium.Choropleth(

geo\_data=corona\_gdf.to\_json(),

name='March 31st',

data=corona\_gdf,

columns=['CODE', 'pos\_pop'],

key\_on='feature.properties.CODE',

fill\_color='YlOrRd', #YlGn#PuBu ### https://github.com/dsc/colorbrewer-python

fill\_opacity=0.7,

line\_opacity=0.2,

#threshold\_scale=[0, 0.001, 0.003, 0.005, 0.007, 0.009, 0.011, 0.013],

legend\_name='March 31st Postive accounts of Covid19 by Zip-code population',

highlight = True

).add\_to(m)

march\_31.geojson.add\_child(

folium.features.GeoJsonTooltip(

fields= ['CODE' ,'Positives', 'P001001', 'pos\_pop'],

aliases = ['Zip Code:','Total Positive COVID-19 Cases for Zip Code as of March 31st:','Total Zip Code Population:', 'March 31st Positive Tests per Zip population:']))

april\_21 = folium.Choropleth(

geo\_data=corona\_gdf.to\_json(),

name='April 21st',

data=corona\_gdf,

columns=['CODE', 'april\_posp'],

key\_on='feature.properties.CODE',

fill\_color='YlOrRd', #YlGn#PuBu ### https://github.com/dsc/colorbrewer-python

fill\_opacity=0.7,

line\_opacity=0.2,

#threshold\_scale=[0, 0.001, 0.003, 0.005, 0.007, 0.009, 0.011, 0.013],

legend\_name='April 21st Positive accounts of Covid19 by Zip-code population',

highlight = True,

show = False

).add\_to(m)

april\_21.geojson.add\_child(

folium.features.GeoJsonTooltip(

fields= ['CODE', 'Count','P001001','april\_posp', 'increase'],

aliases = ['Zip Code:','Total Positive COVID-19 Cases for Zip Code as of April 21st:','Total Zip Code Population:', 'April 21st Positive Tests per Zip population:', 'Increase of positive cases from March 31st to April 21st']))

folium.LayerControl().add\_to(m)

m.save("choropleth.html")

The os.chdir() command requires navigating to the directory containing the joined Zip Code polygon layer. In this case, the joined layer was saved as “pos\_pop\_full.shp”. Parameters will need to be changed in order to be compatible with the user defined names that were saved throughout the ArcMap process. The code reads the shapefile and creates a json within the folium script. This json crated the boundaries of Zip Codes and the resulting script assigns fields to create a map based off the field that was created to calculate positive tests divided by population. A tooltip is then added to display zip code name, number of positive test results, the population, and positive test results per population. An addition field was added to the input shapefile that subtracted the March positive test results from the April results in order to show how many more cases were added. This was added to the April 21st tooltip.

**Results:**

The March 31st Spatial Autocorrelation report (Figure 1) showed the analysis to have a z-score of 5.33 signifying a significant cluster than is not likely due to random chance. The April 21st Spatial Autocorrelation report (Figure 2) reported a z-score of 1.81 which is also significant showing there is less than a 10% likelihood that the cluster could result from random chance.

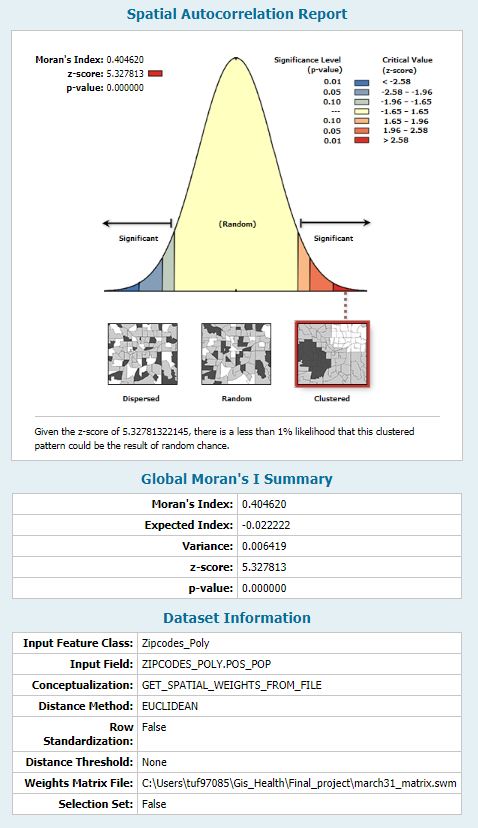
The Cluster and Outlier Analysis (Anselin Local Moran’s I) maps (Figures 3 and 4) showed the low-low clusters to be in similar locations while the high-high were located in different areas.

**Conclusions / Limitations:**

The results showed that there were clusters of positive test results according to population size per zip code. Using cluster analyses to help assist in times of pandemics can help health officials identify areas that need increased attention and service. Limitations of this analysis would include a lack of knowledge concerning COVID-19. At the time of writing this report, scientists are still working to gain knowledge about the virus. It is uncertain what can come of this pandemic and if a cure will be developed any time soon. Continuing this report would require updating current data and analyzing if any trends can be drawn from future data.

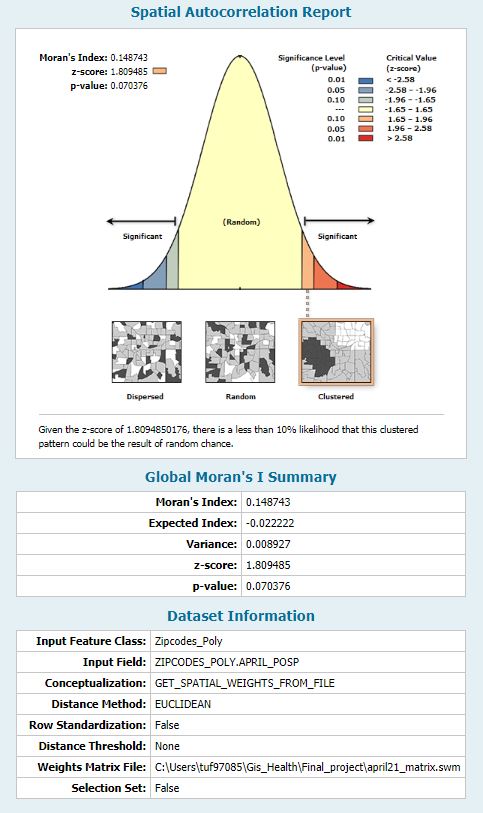
**Figures:**

*Figure 1: March 31st Spatial Autocorrelation Report*

****

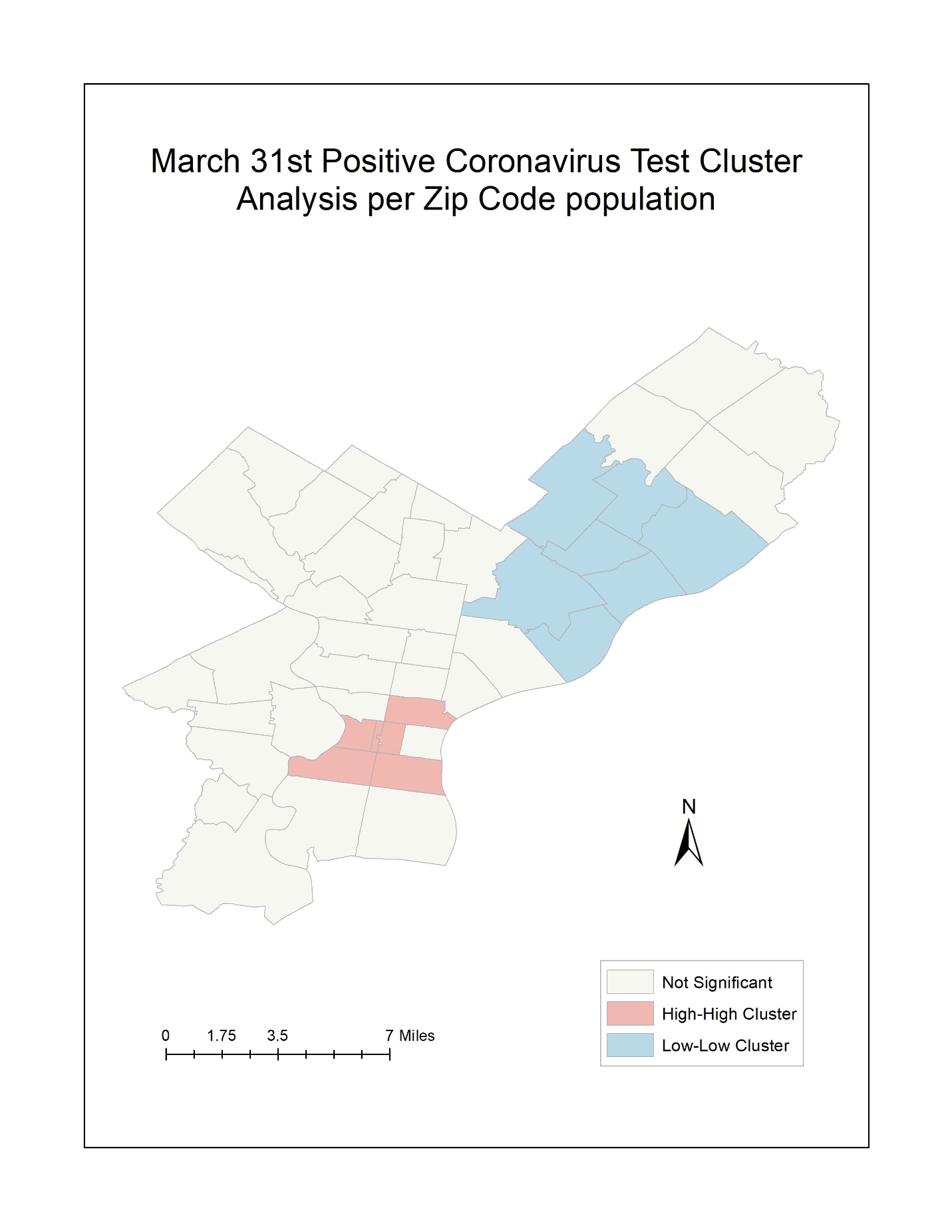
**Figure 1- The March 31st Spatial autocorrelation report that resulted from the “Spatial Autocorrelation (Moran’s I)” process. This result shows a significant cluster that would likely not result from random chance.**

*Figure 2: April 21st  Spatial Autocorrelation Report*

****

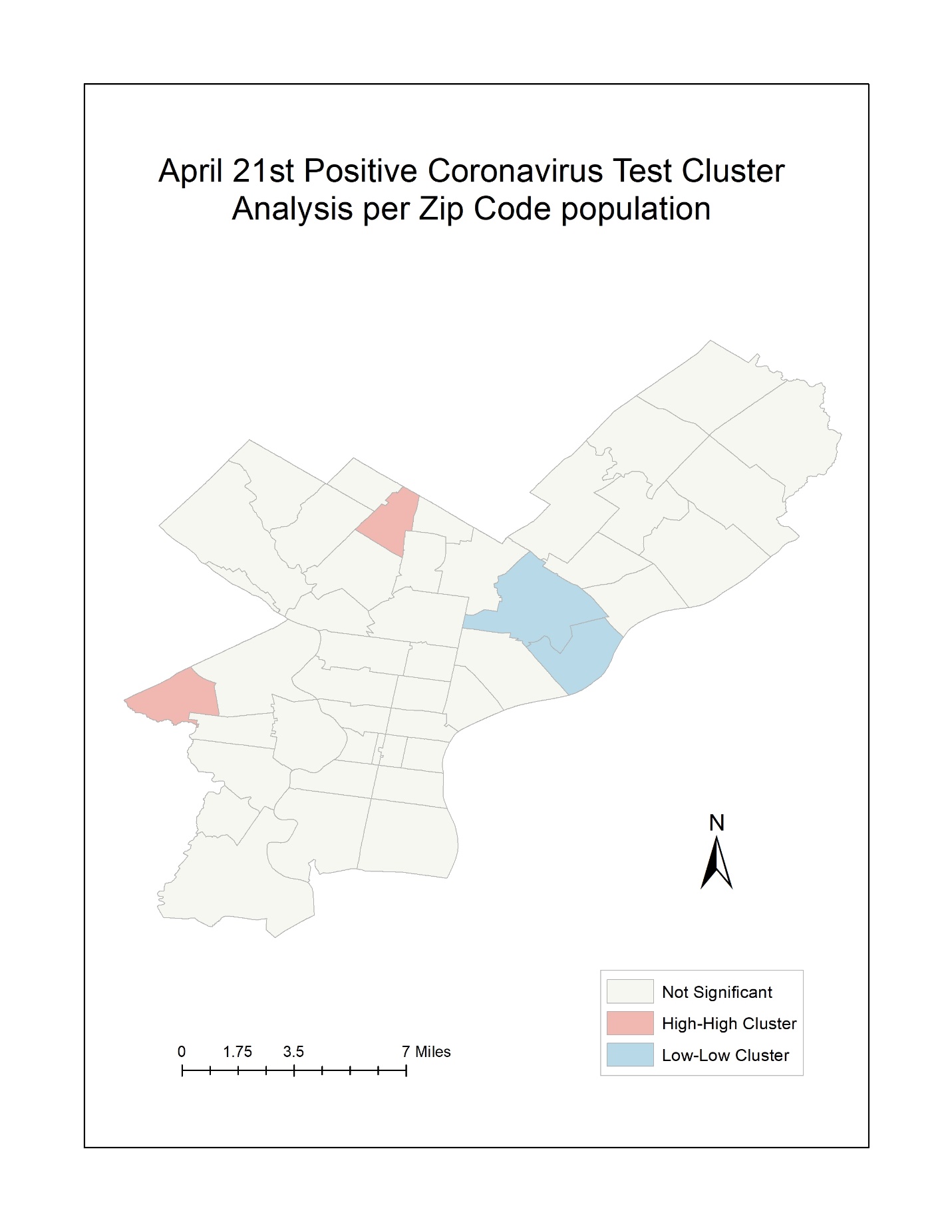
**Figure 2- The April 21st Spatial autocorrelation report that resulted from the “Spatial Autocorrelation (Moran’s I)” process. This result shows a significant cluster that would likely not result from random chance.**

*Figure 3: March 31st Cluster and Outlier Analysis*

****

**Figure 3- The March 31st map resulting from the “Cluster and Outlier Analysis (Anselin Local Moran’s I)” process.**

*Figure 4: April 21st Cluster and Outlier Analysis*

****

**Figure 4- The April 21st map resulting from the “Cluster and Outlier Analysis (Anselin Local Moran’s I)” process.**

**Bibliography:**

Coronavirus in Pennsylvania. (2020, May 4). Retrieved from <https://www.health.pa.gov/topics/disease/coronavirus/Pages/Coronavirus.aspx>

Explore Census Data. (n.d.). Retrieved from <https://data.census.gov/cedsci/>

Hammond, E., & Hayes, M. (2020, April 30). US coronavirus update: Number of cases exceed 1 million. Retrieved from <https://www.cnn.com/us/live-news/us-coronavirus-update-04-30-20/index.html>

How to Protect Yourself & Others. (2020, April 24). Retrieved from <https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/prevention.html>

New coronavirus stable for hours on surfaces. (2020, March 17). Retrieved from <https://www.nih.gov/news-events/news-releases/new-coronavirus-stable-hours-surfaces>

Sauer, L. M. (2020, May 4). What Is Coronavirus? Retrieved from <https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus>

Testing and data: Department of Public Health. (n.d.). Retrieved from <https://www.phila.gov/programs/coronavirus-disease-2019-covid-19/testing-and-data/>

Zip Codes. (2014, December 8). Retrieved from <https://www.opendataphilly.org/dataset/zip-codes>