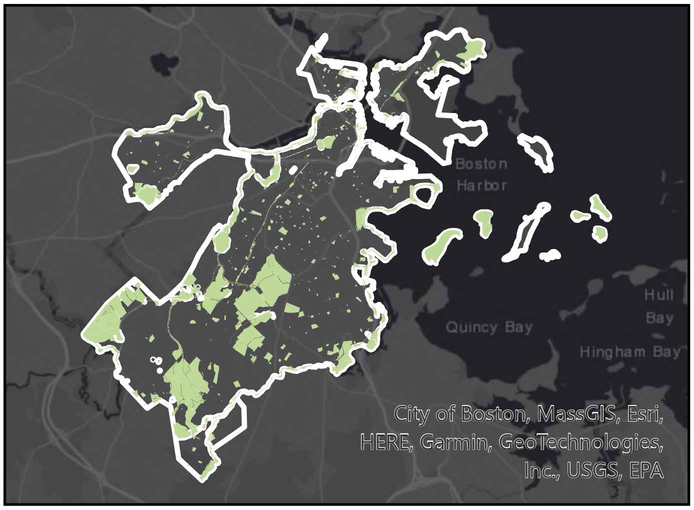
Map Portfolio

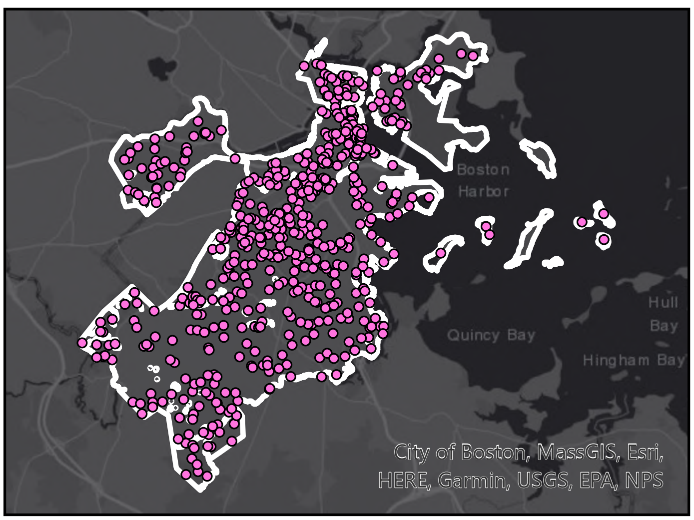
EHS 568: Introduction to GIS for Public Health
Submitted December 1, 2025

Michael Willen
Yale College, Class of 2026

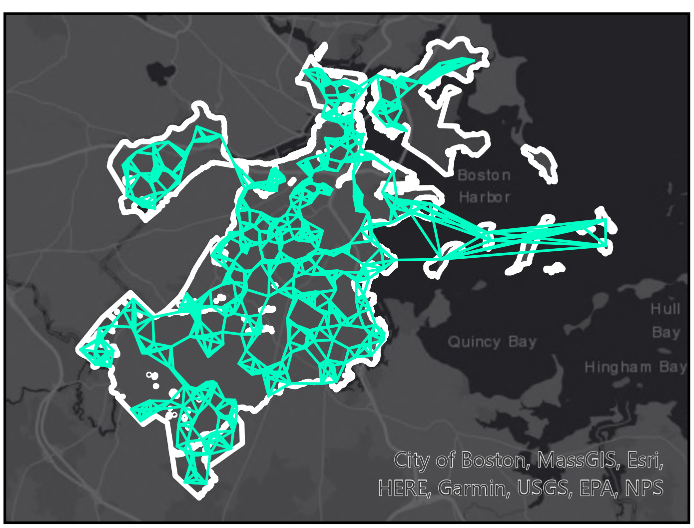
(1) Boston Green Space: Polygons, Points, Lines



Open Spaces: Polygons
City of Boston



Open Spaces: Points (Centroid)
City of Boston



Open Spaces: Lines to 5-Nearest Neighbors
City of Boston

Map 1: Points, Lines & Polygons

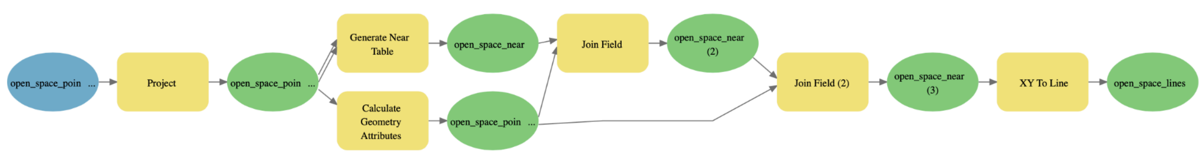
Open Space in the City of Boston

For my first set of maps, I chose to examine open spaces in the city of Boston and look for coverage gaps. Like many American cities, Boston releases a lot of its geospatial data through an open data portal.

The Analyze Boston site contains an "Open Space" dataset[1] containing polygons of every open space in the city — both recreational and conservational — as well as a "City of Boston Outline Boundary (Water Excluded)" dataset,[2] which was important to include as it shows the scope of our observation area. These were the only datasets I used for this set of maps.

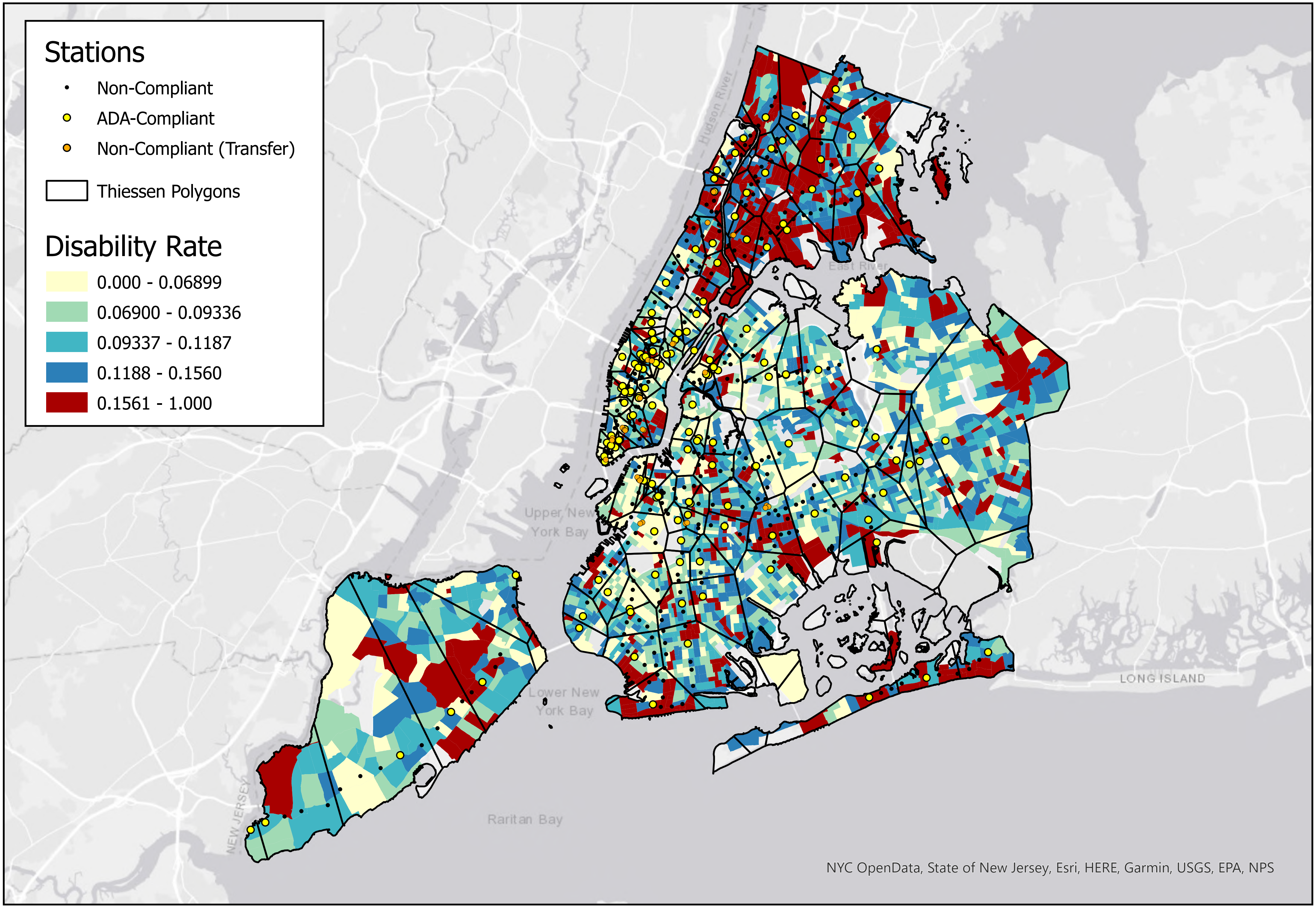
For the polygons, I simply added both layers to the map. For points, I created a point at the centroid of each open space polygon. For the line layer, my goal was to create lines between the point marking each open space and its k-nearest neighbors. I made a table of each open space and its nearest neighbors, using my open space points layer as both the input and near features, settling on five neighbors, which thinned out the web but reduced the number of lines crossing areas of the city without open spaces.

I created x and y columns for each centroid point in the points layer, then joined it to the near table twice in order to get the coordinates and create the final lines. I put all of these steps in a ModelBuilder model in order to more easily test changes as I organized and troubleshooted this process.



[1] https://data.boston.gov/dataset/open-space
[2] https://data.boston.gov/dataset/city-of-boston-outline-boundary-water-excluded

(2) ADA-Compliant Stations and Disability Rates in New York City



Map 2: Put the Public in Public Health

ADA-Compliant Station Access in New York City

I've spent the past two summers in New York City, first doing research on subway station accessibility and then doing data science for the MTA. Less than a third of subway stations have step-free access from the street level to the platform level, while frequent elevator outages mean that numbers are often reduced even further. In 2023, the MTA promised to make 95 percent of stations ADA-compliant by 2055 as part of a class-action lawsuit settlement, but progress is contingent on adequate funding — when congestion pricing was temporarily halted in June 2024, several upgrades were put on hold — and as not every station can receive improvements at once, the MTA must make choices on which stations to prioritize.

The U.S. census has data on disability characteristics available at a census tract level (Table S1810, 2023 ACS 5-Year Estimates Subject Tables),[3] which I was able to download as a CSV. I obtained a New York census tract shapefile[4] from New York City's open data portal and joined the disability data to the shapefile, setting the symbology to the total disability rate in each tract and making the top bucket (>15%) red to further highlight these areas. I then added points marking every subway stop in the city.[5] These included the ADA compliance status of every stop, and I used both color and size to highlight fully ADA-compliant stops.

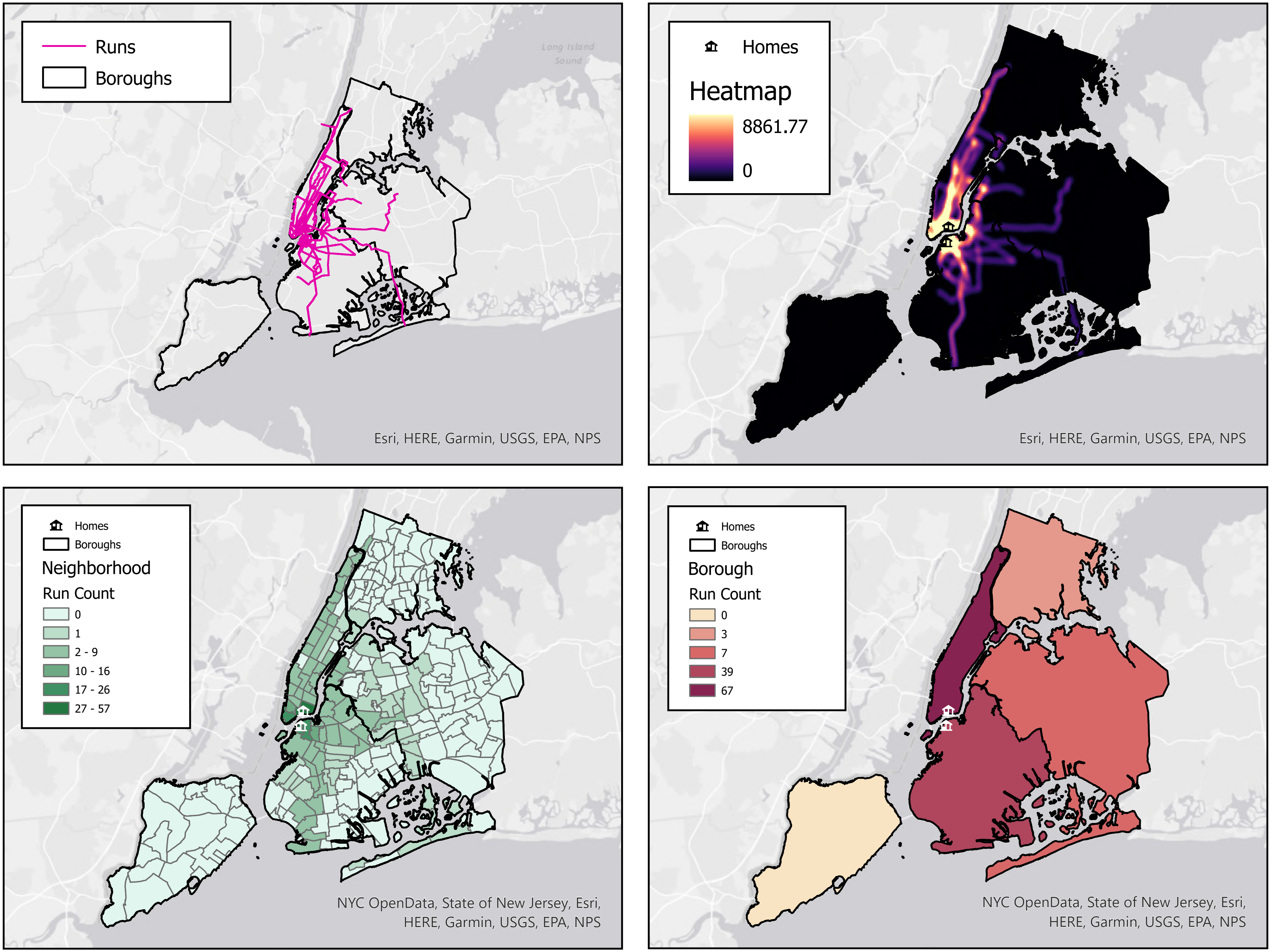
Transfer stations — station complexes where you can move between stations or lines without leaving the area within the fare gates — are particularly important to the movement of people through the subway system. To specifically highlight transfer stations that were not ADA compliant, I aggregated and identified the complexes with more than one stop (each stop has a unique stop ID, but every stop in a complex shares a complex ID), then joined that back to the stop layer and made stops that were both transfer stations and non-ADA-compliant orange. One major example is Broadway Junction in Brooklyn, which services the A, C, J, Z, and L trains (and eventually the IBX) and is currently in the process of receiving accessibility upgrades.

Finally, I created a layer of just the ADA-compliant stations and created Thiessen polygons to better visualize proximity and access to an ADA-compliant station across New York. The largest polygons are those at the eastern edges of Queens, though this is largely because subway service does not extend that far into the borough. However, this visualization can provide insight into where access is currently inadequate and potential focus or priority areas for the next set of upgrades. The intersections of these polygons are the furthest areas from an accessible station, and nearby inaccessible stations could receive priority. This map could be further improved by creating similar polygons by a function of walking or driving distance, rather than simply Euclidean distance.

---

[3] https://data.census.gov/table/ACSST5Y2023.S1810
[4] https://data.cityofnewyork.us/City-Government/2020-Census-Tracts/63ge-mke6/about_data
[5] https://data.ny.gov/Transportation/MTA-Subway-Stations/39hk-dx4f

(4) Running in New York City



Map 4: Create Data

Covering New York City by Foot

As I mentioned above, I've spent the past two summers in New York City. Each summer, I've run nearly every day and usually outside; it's one of my favorite places to run given how much life and activity there is, no matter which direction you go. I track every run I've done on Strava, which allows you to download activities as GPX files, and in total I've completed 79 outdoor runs in the city across the two summers.

I wanted to see how much of New York City I've run at this point — Strava has a feature to do so, but it's locked behind a subscription paywell. Instead, I downloaded the GPX file for each of these runs. I used ModelBuilder to create a pair of models: "BulkGPX" and "MergeBulkGPX." The former takes a folder with all of the GPX files and iterates through them, converting them to features and then collecting the values into a single output. The latter model uses the former as a submodel, taking the set of output files and merging them into a single layer.

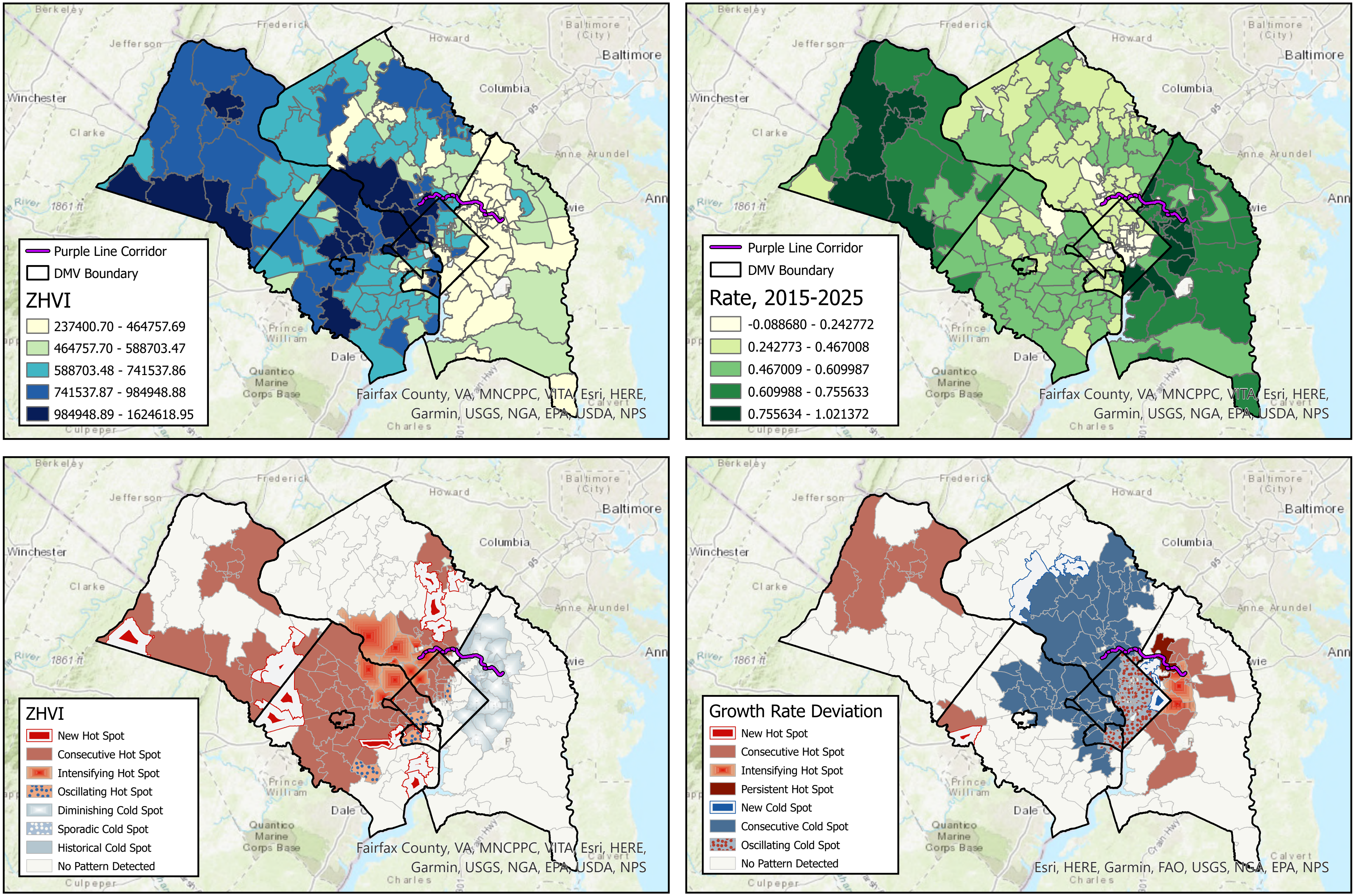
I wanted to see two things using this data: a heatmap of where I ran in the city and the frequency at which I'd visited each neighborhood. While neighborhood definitions in New York City are somewhat fuzzy (though there's a really cool New York Times visualization[6] of how residents view their neighborhood, which is one of my all-time favorites), we can use Neighborhood Tabulation Areas[7] from the census as a close approximation. After adding it to the map, I spatially joined the run paths to the NTAs, giving me a count of the number of routes that went through each neighborhood, and did the same with New York's boroughs.

For each of these maps, I also created a feature class and added markers for my student housing in Brooklyn Heights from summer 2024, and my apartment in the Lower East Side from summer 2025. Unsurprisingly, these were the neighborhoods with by far the most runs, with a clear decrease from there as the distance increased. The furthest reaches of the city remained largely untouched, interrupted by linear stretches of neighborhoods where I just ran far on one path, in one direction. Meanwhile, the only borough I haven't run to is Staten Island.

[6] https://www.nytimes.com/interactive/2023/upshot/extremely-detailed-nyc-neighborhood-map.html

[7] https://data.cityofnewyork.us/City-Government/2020-Neighborhood-Tabulation-Areas-NTAs-/9nt8-h7nd

(5) Emerging Hot Spot Analysis: Home Values and the Purple Line Corridor



Map 5: Other Dimension

A Tale of Two Cities: Home Values (and the Purple Line)

In 2017, construction began on the Purple Line in Maryland's D.C. suburbs after decades of planning. Most metro systems in the United States, including Washington's, operate as a hub-and-spoke model, where riders have to ride all the way into the center of the system (usually downtown) in order to go anywhere else. This makes it difficult to efficiently reach neighboring suburbs that might only be a few minutes away by car, and when lines aren't running, it essentially cuts off riders. This is a major limiting factor in transit ridership; it's unsurprising that so many choose to buy cars and drive when the likelihood of them needing to do so is so high. By creating lines that operate more as a loop (or a segment of one, like this), it gives riders multiple alternatives when needed and allows for easier movement across the suburbs.

Proponents have argued that the new Purple Line Corridor[8] will bring long-term environmental, social, and economic benefits to the surrounding areas.[9] Though I think this will be clearer after the line opens in 2027, I was curious to see if there was already an impact. However, it was difficult to find granular data that would allow me to clearly differentiate and analyze the Purple Line Corridor (business data, for example, was largely available only at state or county levels); instead, I decided to use the Zillow Home Value Index,[10] or ZHVI, which estimates the typical value of a mid-priced home in a given area. This data was available at the ZIP Code level, which I used for this map. As I explained in a later map, I limited my examination of the Washington metropolitan area to the District of Columbia and a few surrounding counties; I reused the same boundaries in this analysis as well. I obtained a shapefile of ZIP Code Tabulation Areas, or ZCTAs, for the entire United States[11] and clipped it to just the focus area.

In order to do the desired analysis, I had to first transpose the ZHVI file so that each ZCTA-date pair was a separate row in the table, as well as other minor adjustments. For the rest of the process, I built two models using ModelBuilder: one submodel to organize the ZHVI file in order to create the desired columns, and a main model to perform the rest of the process. I downsampled the data to years instead of months to focus more on larger trends, and limited the data to 2015 and after (space-time cubes require at least 10 rows of data for each set). For each ZCTA-date pair, I then calculated the percent change since 2015, as well as the deviation from the mean percent change across all ZCTAs that year (if home values in all areas are growing or shrinking, it'd be more useful to see which are changing more or less).

I then joined this final table to the ZCTA shapefile. Using this, I created a space time cube on which I could finally run ArcGIS Pro's emerging hot spot analysis to create the final

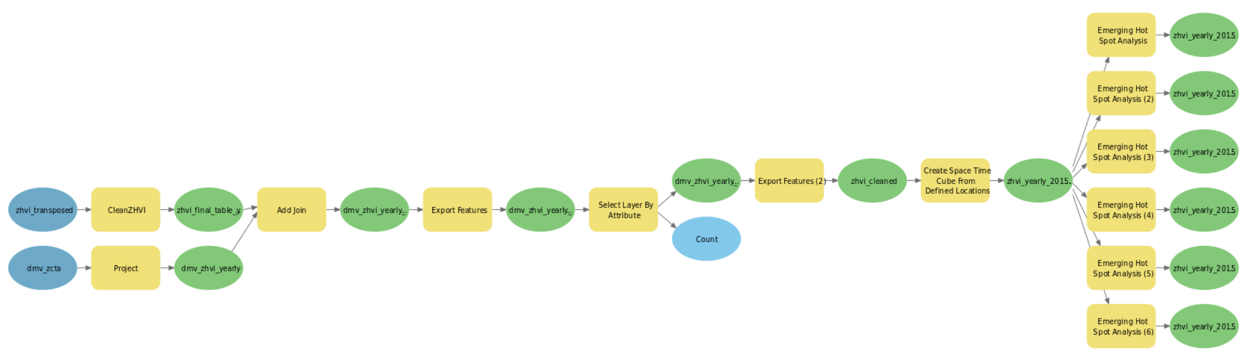
[8] https://data-maryland.opendata.arcgis.com/datasets/maryland::purple-line-alignment
[9] https://www.smartergrowth.net/wp-content/uploads/2015/02/Purple-Line-Factsheet.pdf
[10] https://www.zillow.com/research/data/
[11] https://catalog.data.gov/dataset/tiger-line-shapefile-2022-nation-u-s-2020-census-5-digit-zip-code-tabulation-area-zcta5

output. I ran this six times in total to create output for both the raw ZHVI and for the rate deviations. I also included maps of the overall rate of change from 2015 to 2025, as well as one of the the most recent ZHVI for each zip code.

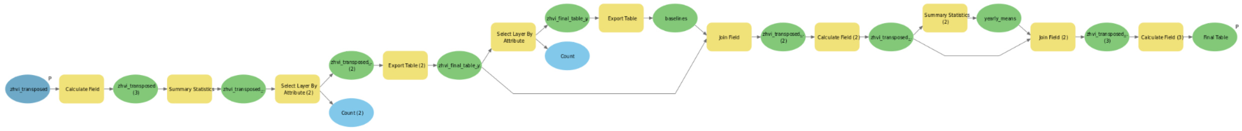
In the end, I think that these maps are inconclusive when it comes to evaluating effects of the Purple Line Corridor thus far. The western parts of the corridor have higher average home values, but have grown slower than the eastern parts. However, both parts appear to be part of larger regional patterns that aren't exclusive to the corridor. ZIP Code areas are still relatively large, and more granular data is necessary to understand how the corridor is impacting the areas directly surrounding it.

However, these maps reveal interesting patterns about Washington's metropolitan area as a whole. There is a clear east-west divide running almost directly down the center of the District of Columbia, with higher home values to the west and an intensifying hot spot in the area surrounding Bethesda, Maryland. While the eastern half has lower home values, this analysis shows that the large cold spot is gradually diminishing. The rate deviation map backs this, showing a complete flip in hot and cold spots, with ZHVI in the east consistently growing at a faster rate than the west. This growth is strongest around the eastern end of the Purple Line Corridor, which may be a good sign for proponents of the new infrastructure.

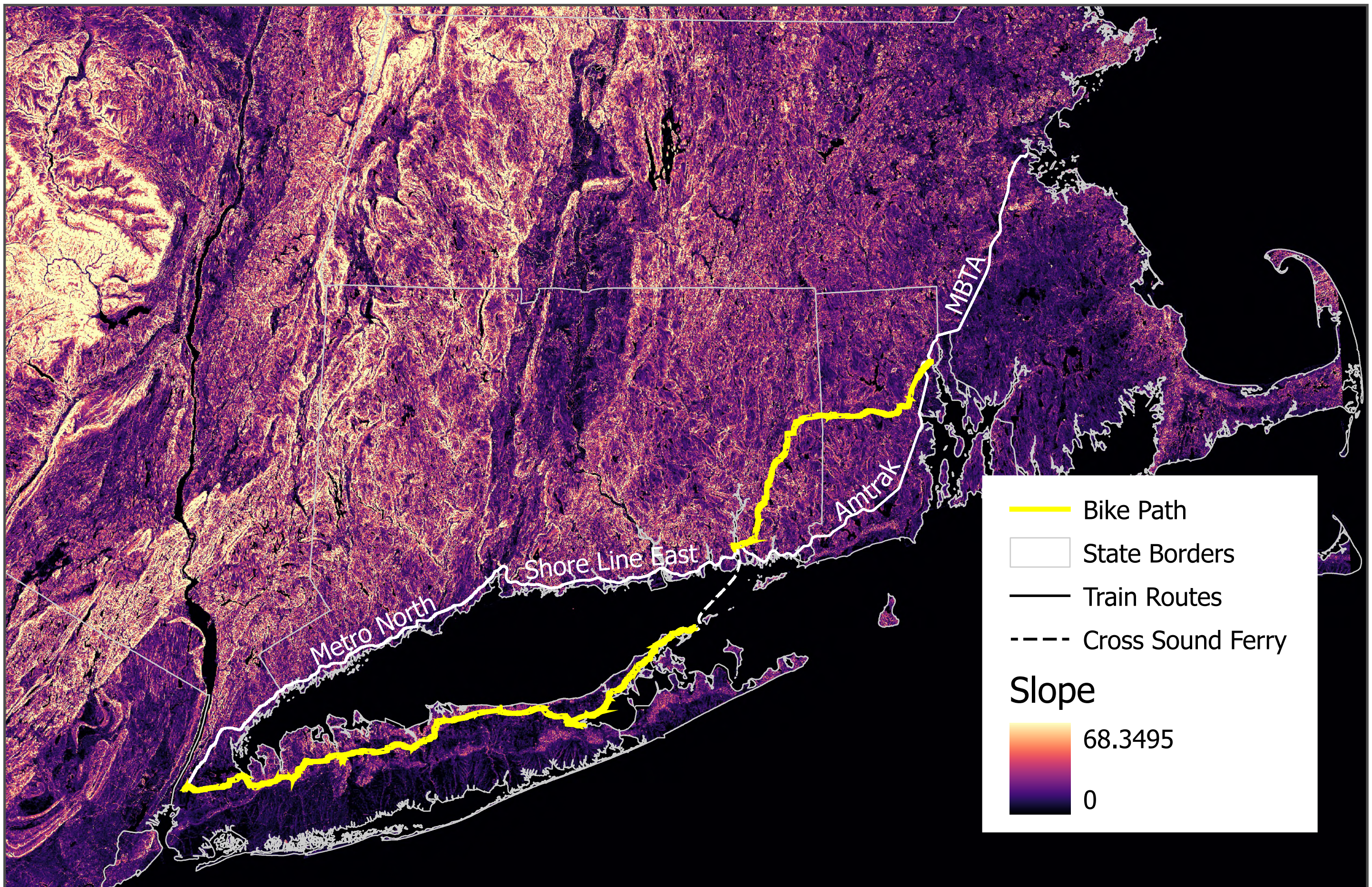
BuildEHSA Model:

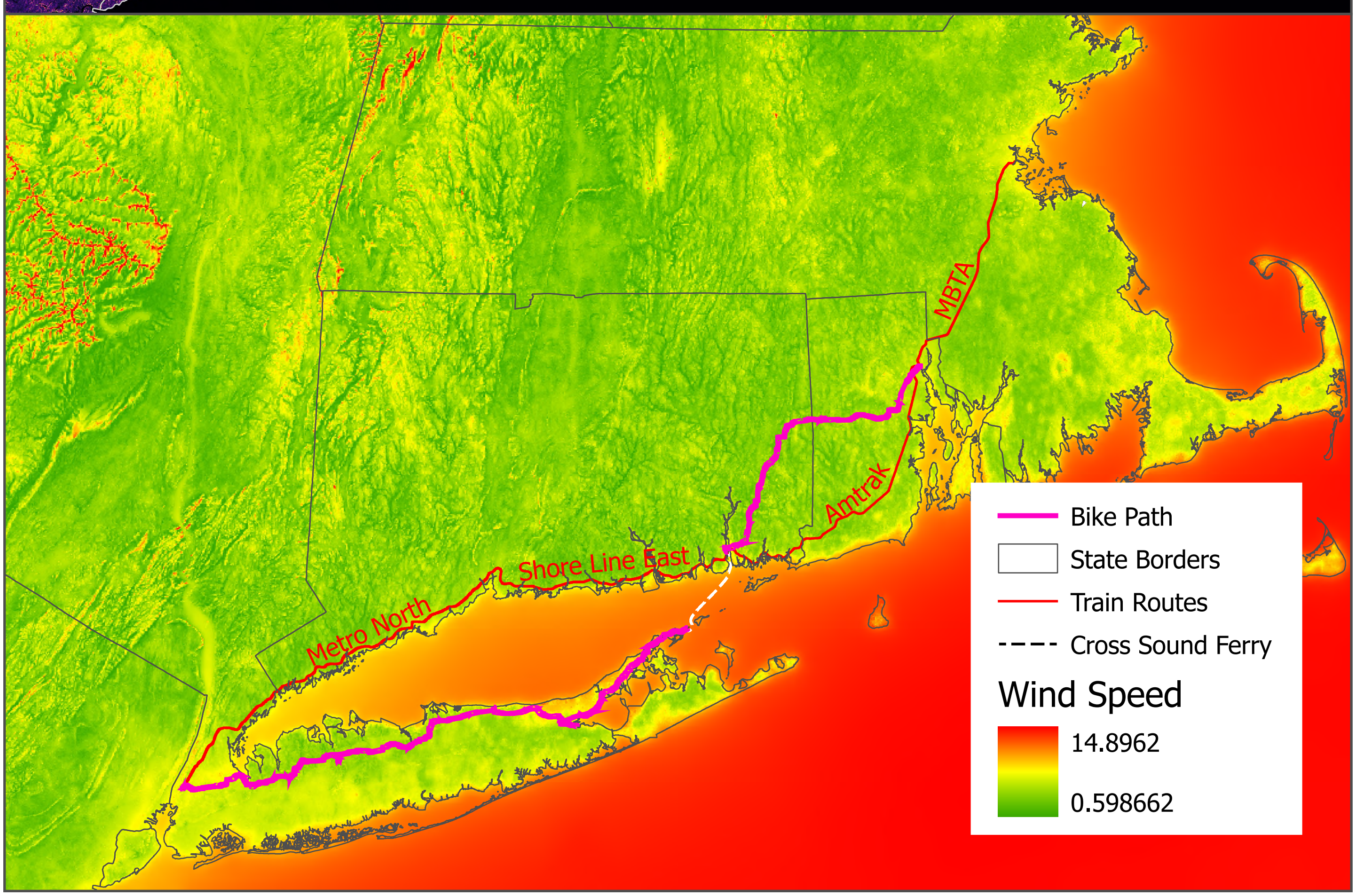


CleanZHVI Submodel:



(6) Slope and Wind Speed, Fall 2025 Bike Trip





Map 6: Raster

Battling the Elements in the Northeastern United States

Over Fall Break this year, I biked over 200 miles across the Northeast — first across parts of Connecticut and Rhode Island and then the entire length of Long Island from Orient Point to Grand Central Terminal. The trip took four days, and each day I faced different elements that sometimes made it tough to continue. For this map, I chose to examine those forces and see if my experiences were backed by data.

I tracked the entirety of my trip on Strava, and like with map 4, I could download the path I took as GPX files, then convert them to features and merge them into a single feature. There were several times along the journey when I rode the train to make the plan work logistically: I rode Shore Line East[12] from New Haven to New London to start the trip, MBTA's commuter rail[13] from Providence to Boston for a place to stay for the night, Amtrak[14] from Boston back to New London to get to Long Island, and finally Metro North[15] from New York back to New Haven at the end of the trip. Each file took some work, using several tools to trim them to just the segments I rode. I took the ferry from New London to Orient Point, the very tip of the North Fork of Long Island; I obtained this path from a line in a larger U.S. ferry routes shapefile.[16] I also added state borders to the map for extra context, though I had to use a file[17] from the National Weather Service due to issues with accessing U.S. census data.

The first day, the major barrier was the sheer hilliness of New England. Connecticut was very hilly, though once I got to Rhode Island I was able to follow a largely flat rail trail. My total elevation gain that day was over 3,000 feet — far more than any other day. While Long Island was far flatter, I faced extreme amounts of wind along the North Fork on day two of the trip.

I was able to get a wind speed raster for the entire United States from the Global Wind Atlas,[18] which I then clipped to the desired focus area. This result backed up my experience, as the North Fork had notably higher wind speeds than the rest of Long Island and the rest of the Northeast (outside of mountain peaks further north).

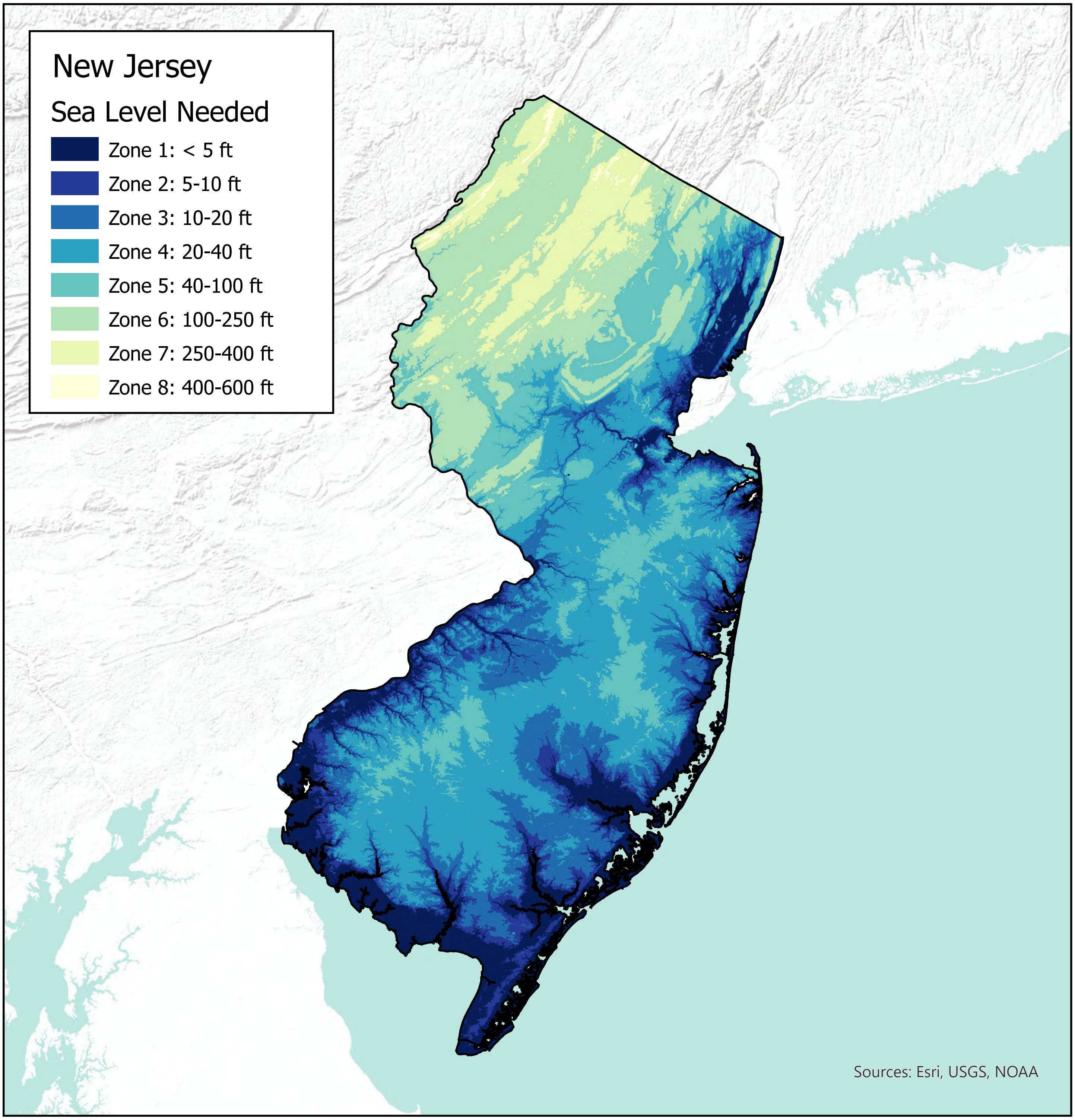
To examine the first day's hilliness, I created a slope raster. I was able to download elevation rasters from the U.S. Geological Survey,[19] but these were divided into 1-arc-second segments and so I needed 13 to cover the desired area. I again used ModelBuilder to iteratively upload each raster and create slope rasters for each one, finally mosaicing all 13 rasters into one slope raster for the target area.

---

[12] https://ct-deep-gis-open-data-website-ctdeep.hub.arcgis.com/datasets/CTDEEP::connecticut-railroads
[13] https://www.mass.gov/info-details/massgis-data-trains
[14] https://data-usdot.opendata.arcgis.com/datasets/usdot::amtrak-routes
[15] https://hub.arcgis.com/documents/DCP::metro-north-railroad-lines-mta
[16] https://geodata.bts.gov/datasets/usdot::ferry-routes
[17] https://www.weather.gov/gis/USStates
[18] https://globalwindatlas.info/en/download/gis-files
[19] https://apps.nationalmap.gov/downloader/

There were no rasters for the portions that were entirely ocean but I wanted a rectangular output, so I created a raster of zeroes that matched the desired output extent and added it to my slope raster to fill in the missing areas. Once again, the result backed my observations: I faced more and larger slopes in New England, while Long Island was generally much flatter but got hillier as I got closer to New York City. I was following the North Shore, which is generally hillier than southern Long Island.

(7) How Quickly Can We Sink New Jersey?



Map 7: Hazards

How Quickly Can We Sink New Jersey?

As a Philly kid, I'm somewhat obligated to engage in some level of New Jersey hate. It's a frequent target of my jokes, especially when directed towards a friend from the state. And if it were to sink below the waves, the United States might be one state better off. So how quickly could this happen?

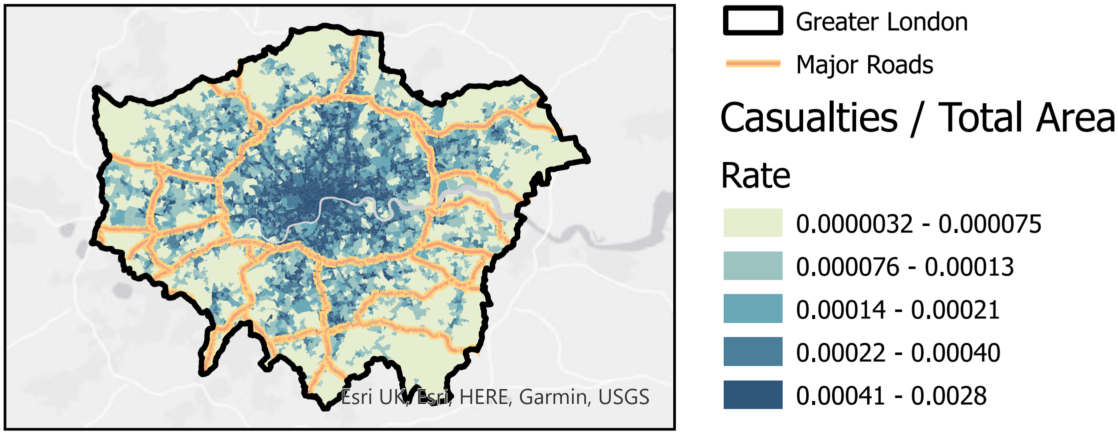
I once again used the U.S. Geological Survey to obtain elevation rasters[20] that covered all of New Jersey — 10 were needed in total — and extracted just the state from a shapefile containing every U.S. state.[21] I then mosaiced the component rasters into one larger raster and clipped it to just the state of New Jersey. Finally, I reclassified the values into eight zones with cleaner dividing values, thus giving the final result.

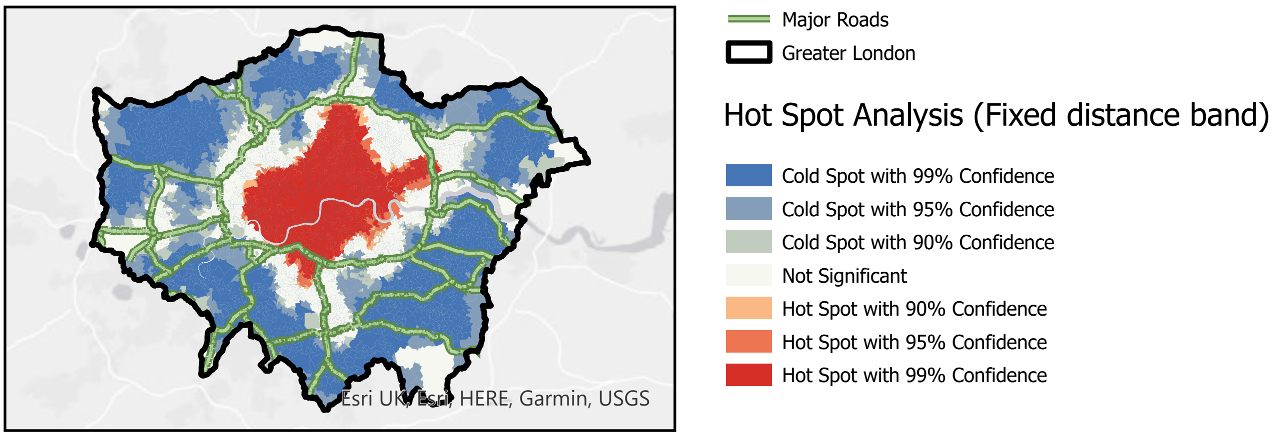
This map was largely straightforward, but — tongue-in-cheek New Jersey jokes aside — does serve a useful purpose. While the very north of New Jersey has higher elevations, most of the state is a flat coastal plain and is highly susceptible to flooding and future sea level rises. The state already takes a lot of preventative action to ensure these risks are mitigated; one such example are the sand dunes that line the Jersey Shore, preventing flooding on the barrier islands and in areas further inland. A map like this could be used by governments and planners to take preventative measures, while keeping residents informed on the potential risks they face depending on where they live. Further steps for this map could be to add key amenities and services, as well as current preventative measures to better determine where more action is needed.

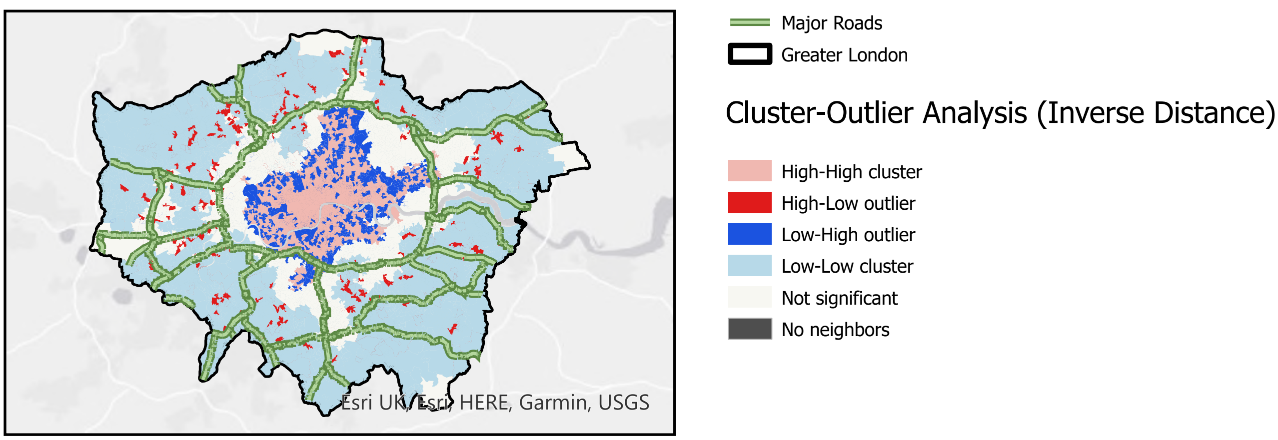
---

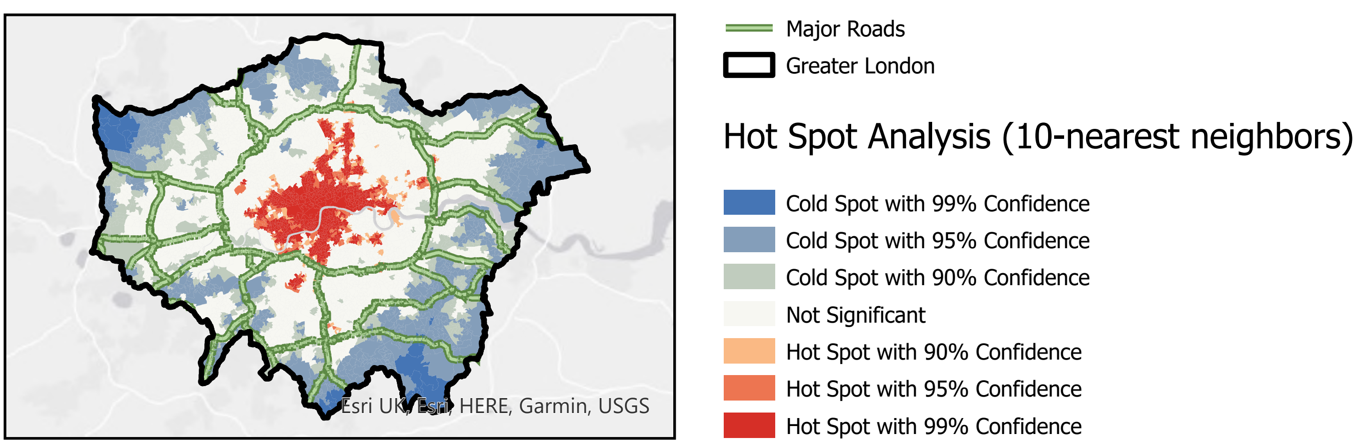
[20] https://apps.nationalmap.gov/downloader/
[21] https://www.weather.gov/gis/USStates

(8) Road Casualty Hotspots in London, 2010-2018









Map 8: A New Tool

Hot Spot and Cluster Analysis in Greater London

I'm very interested in the hot spot, cluster, and outlier analysis tools from the Spatial Statistics package, and chose to explore them further as new geoprocessing tools not taught in class.

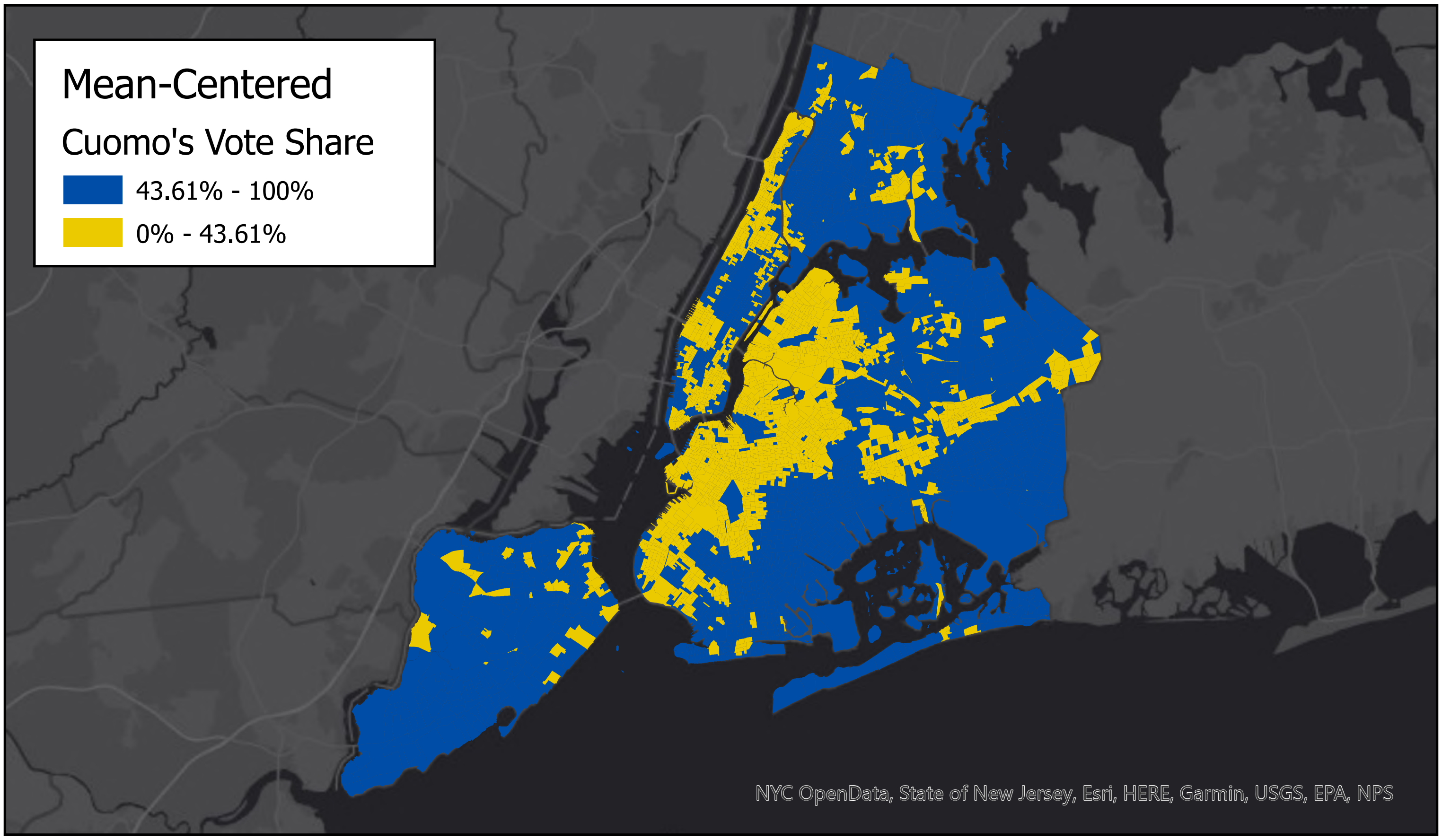
For this, I found a data set with the number of yearly road casualties in London between 2010 and 2018,[22] divided by "Lower Super Output Areas," or LSOAs. As the LSOAs changed in 2021 after the decennial census, I had to find a shapefile containing the 2011 LSOA map.[23] The file was divided into different feature layers for each borough, so I first merged them all into a single layer.

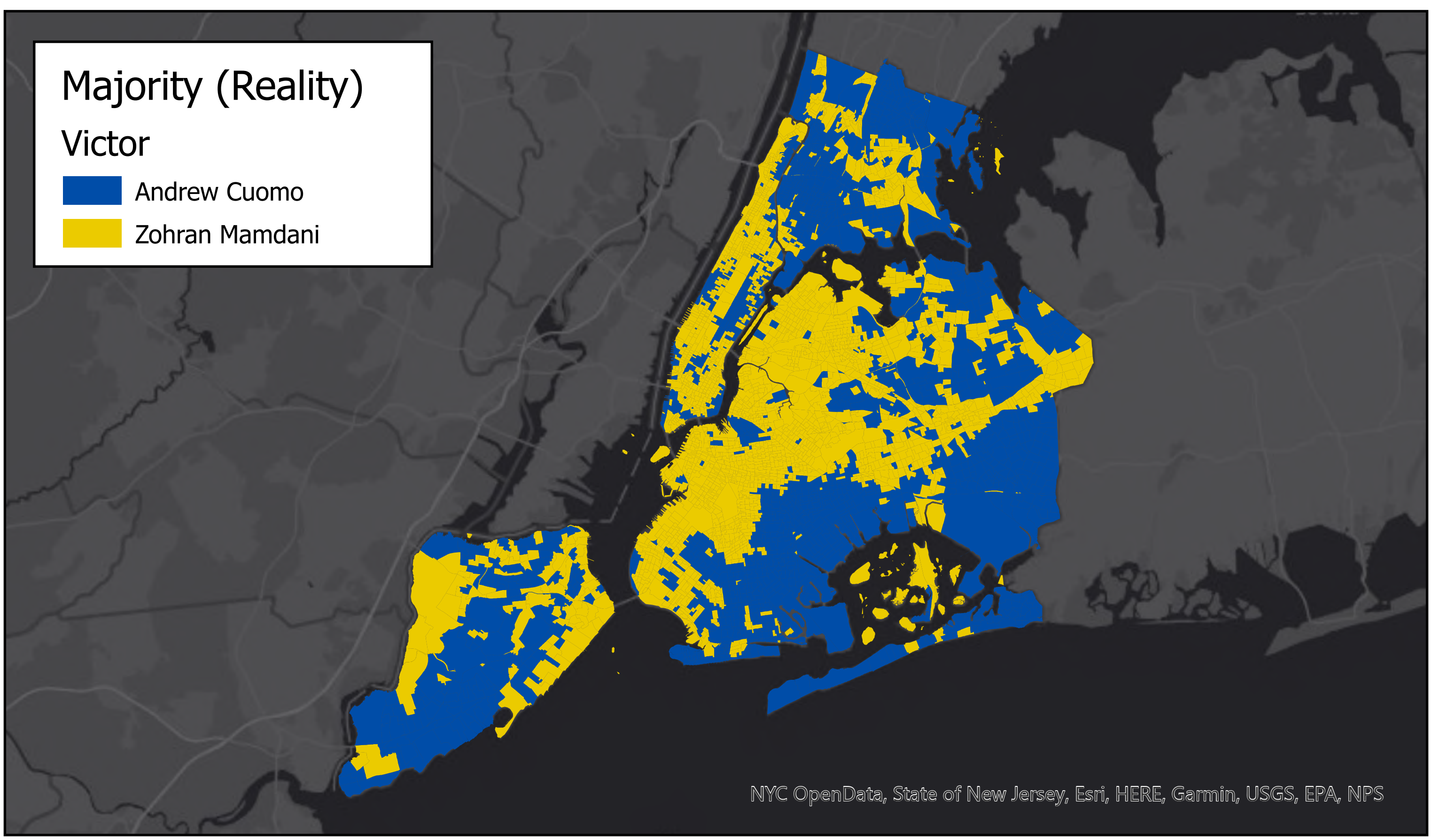
After doing some cleanup of the road casualty CSV file, I joined the table to the LSOA shapefile and created a column that summed the total number of casualties between 2010 and 2018. After initially performing my analysis on the raw totals, I realized that I should adjust for the size of the LSOA (which, yes, may not be exactly proportional to the total road distance in each one — additional fine-tuning for future analysis). With these rate values, I ran the hot sport analysis tool, first using a fixed distance band algorithm and then using k-nearest neighbors, with $k = 10$. K-nearest neighbors gave greater nuance in results than the fixed distance band. I then used the cluster and outlier analysis tool, which marks hot and cold spots generally but also the "high-low" and "low-high" areas within them, where the polygon has a markedly higher or lower value than its surrounding low or high spot, respectively. I used the inverse distance algorithm in order to try out another method.

Finally, I added a layer that highlighted the major roads in the Greater London area.[24] Future steps with this map could be to normalize the data according to traffic volume to better identify places that see unexpectedly high or low numbers of casualties, as the current map and analysis may just be a reflection of traffic volume.

[22] https://data.london.gov.uk/dataset/road-casualties-by-severity-e1z1j/
[23] https://data.london.gov.uk/dataset/2011-census-geography-boundary-files-29jwj/
[24] https://www.data.gov.uk/dataset/95f58bfa-13d6-4657-9d6f-020589498cfd/major-road-network

(9) Political Maps and the Dangers of Multipart Color Ramps





Map 9: A Bad Map

Political Maps and a Grudge

The political arena is rife with bad maps, often intentionally designed to mislead and misrepresent in order to push a given agenda. For example, U.S. election maps make the country seem far redder than it is given the differing population densities; but land doesn't vote, people do. Further, gerrymandering is in some part evil genius and in some part just bad maps that completely ignore underlying demographics in order to achieve a desired result.

This map comes largely as the result of a very specific grudge I hold against a set of visualizations that the Yale Daily News published[25] at the end of my sophomore year. While I have no issues with the content itself, they included a pair of interactive maps which appeared to indicate that a much larger portion of New Haven had voted "Uncommitted" in the 2024 Democratic presidential primary, and that a majority of the city had voted for Bernie Sanders in the 2016 Democratic primary. However, the multipart color ramp that they chose for these maps centered around about 25 percent for the former and 38 percent for the latter, meaning that any ward where Uncommitted or Sanders received greater than those thresholds, even if the other candidate received a majority, was colored in the same color as areas where they received a majority. In the end, it's misleading, especially for the many readers who likely won't examine the maps further and thus miss this nuance.

I wanted to recreate this effect, and decided to use the 2025 New York City Democratic mayoral primary, where Zohran Mamdani defeated Andrew Cuomo in an upset victory, to do so — as data is not yet available from the recent general election. I was able to obtain the Cast Vote Record, or CVR, from the New York Board of Elections website;[26] Unfortunately, New York City publishes its election data in a manner that added several extra steps to the process. The CVR is divided into 32 different files, split by borough as well as by absentee, early, day-of election, affidavit, and emergency ballots. Each row contains a unique voter's ballot and precinct information, as well as columns for their ranked choices for each election, with each candidate given an ID number in place of their names. The BOE did not include any documentation that explained this, but luckily a GitHub post helped break down all of this information.[27] Further, each file was a Microsoft Excel spreadsheet rather than a CSV file.

Once I converted them all to CSV files, I merged them into one combined table with every ballot — over 1 million rows. In the mayoral race, voters rank their top 5 candidates. To condense the table, I grouped the ballots by both precinct and unique top-5 orderings and got the frequency of each of those pairs. In ranked-choice voting, candidates with the least number of votes are eliminated until a candidate reaches a majority, and when a voter's preference is removed their vote goes to their next-top choice. In this election, Cuomo and Mamdani were the final two candidates, so I calculated whether each unique ballot pattern had ranked Cuomo or

---

[25] https://yaledailynews.com/blog/2024/04/08/analysis-protest-vote-sweeps-younger-neighborhoods-falls-flat-at-yale/

[26] https://vote.nyc/page/election-results-summary

[27] https://gist.github.com/anbnyc/282646575f41fabf41da239cbc9ace1f

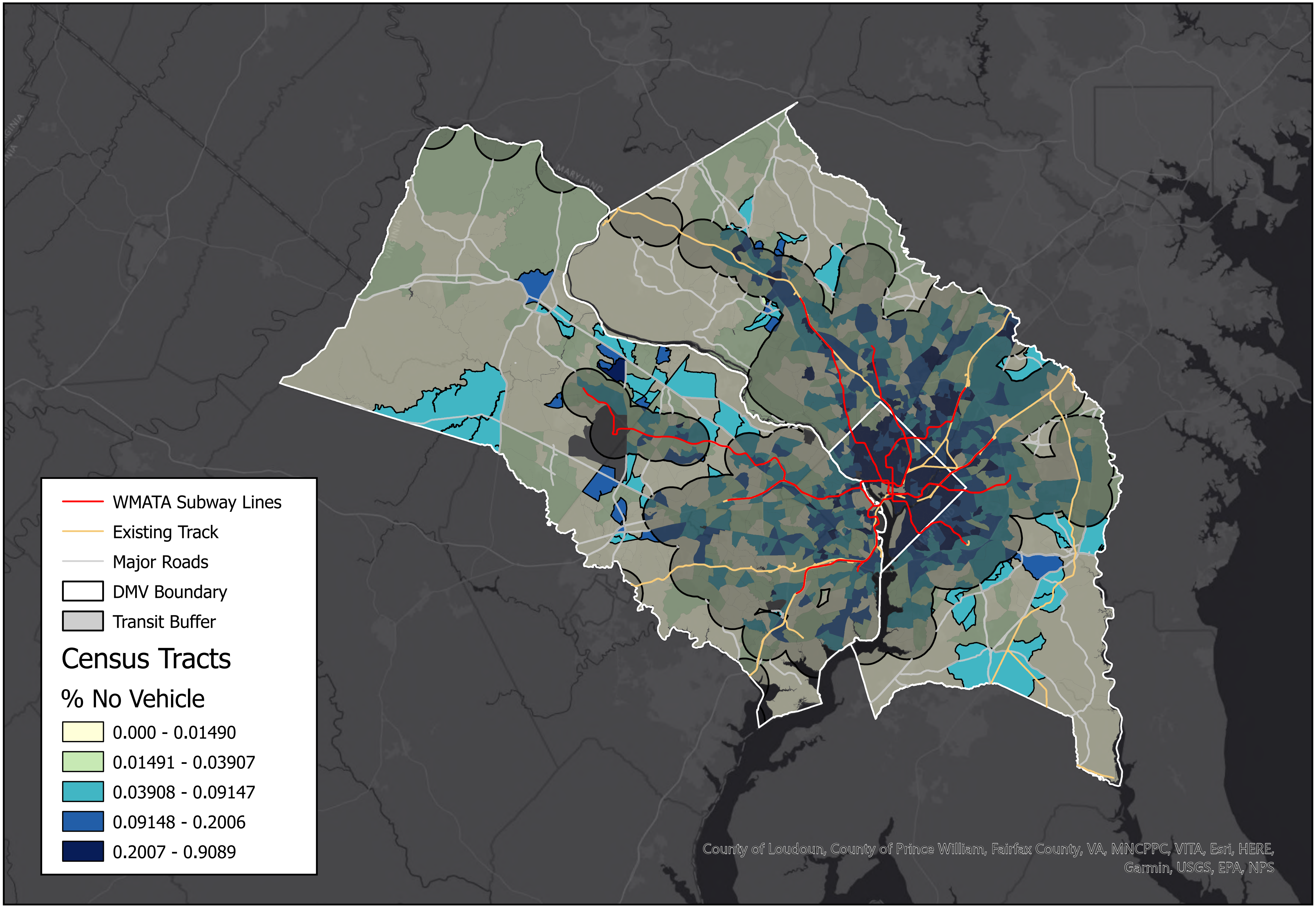
Mamdani higher. Ballots that contained neither were not counted in the final vote tallies, so I kept them separate. I then grouped by precinct and preferred candidate to get the total number of votes by precinct for Mamdani and for Cuomo, then pivoted the table to create a table with just one row for each precinct and the candidate tallies in separate columns. Finally, I calculated the percentage of votes each candidate got in each precinct, and calculated Cuomo's share of the total citywide vote. This number matched publicly available results exactly, which tells me that my data cleaning was accurate.

I was able to join this data to a New York City precinct shapefile,[28] thus creating an electoral results map for the entire city. Similar to the Yale Daily News' map of New Haven, I set the symbology to a multipart color ramp, where the mean was centered at 43.61 percent — Cuomo's total percent of the vote across the city — and colored anywhere he earned more than that his color (blue) and anywhere he earned less Mamdani's color (gold). For comparison, I included a second map showing the actual winner in each precinct.

As Cuomo did earn almost half of the vote, the result was not quite as egregious as it could have been if his tally was in the 30 or 20-percentage point range. However, it does still appear as if he won a much larger proportion of the city than he did in reality, highlighting why it's important to be careful about how you represent your data and be aware of the significance of certain pivot points. Similarly, a multipart color ramp encompassing positive and negative values may not always flip colors at zero, which may cause misleading interpretations.

[28] https://data.cityofnewyork.us/City-Government/Election-Districts/wwxk-38u4

(10) Transit Access and Vehicle Ownership in the DMV



Map 10: Access

Expanding Transit Access in the DMV

For my final map, I chose to examine transit accessibility in the Washington, D.C. metropolitan area, colloquially known as the DMV (D.C.-Maryland-Virginia). Working with data in the DMV proved to be trickier, as the metro area is split between three different jurisdictions, and data sets are often divided by state.

I started by finding shapefiles for WMATA (Washington's public transit agency), adding one for subway lines[29] and one for subway stations.[30] The DMV is also served by the Maryland Area Rail Commuter[31] (MARC) and the Virginia Rail Express[32] (VRE) commuter rails in Maryland and Virginia, respectively, and so I added point feature layers containing the stations for each train. Finally, I added a layer with stations in WMATA's bus network,[33] and a wider track map[34] that could possibly show where train routes could be added without building more tracks. I then merged the stations from all four systems into a single feature and created a 2-mile buffer around the stations.

For the purpose of this assignment, I considered the DMV to be the District of Columbia, Montgomery and Prince George's counties in Maryland, and Loudoun, Fairfax, and Arlington counties in Virginia. While the actual metropolitan statistical area is larger, this area contains the entire extent of WMATA's transportation network and would be the best focus area for potential improvements. I obtained the counties from a shapefile[35] in Maryland's open data portal, then reduced it to just the above counties, dissolving county divisions but keeping the state divisions. I then clipped the transit buffer and the rail map to just the area of interest.

I then added data from the American Community Survey (ACS) on vehicle ownership per household by census tract,[36] using the proportion of households with no vehicles as symbology. I then made a copy of this layer, selecting just the areas with a low vehicle ownership but erasing the areas within the transit buffer. These are the final target areas of our access improvement plan. I then imported the symbology from the vehicle ownership layer, making the original layer more transparent to let these target areas stand out more clearly. Finally I added primary and secondary road shapefiles from Virginia,[37] Maryland,[38] and the District of Columbia to show possible added bus routes, erasing those inside the transit buffer.

---

[29] https://opendata.dc.gov/datasets/metro-lines-regional/about
[30] https://opendata.dc.gov/datasets/metro-stations-regional/about
[31] https://data.imap.maryland.gov/datasets/maryland-transit-marc-trains-stations/about
[32] https://gis-drpt.opendata.arcgis.com/datasets/DRPT::virginia-passenger-rail-stations
[33] https://opendata.dc.gov/datasets/metro-bus-stops/about
[34] https://catalog.data.gov/dataset/tiger-line-shapefile-2019-nation-u-s-rails-national-shapefile
[35] https://planning.maryland.gov/MSDC/Pages/census/Cen2010/maps/metroareamap/zipfile_idx.aspx
[36] https://www.arcgis.com/apps/mapviewer/index.html?layers=9a9e43ec1603446880c50d4ed1df2207
[37] https://catalog.data.gov/dataset/tiger-line-shapefile-2023-state-virginia-primary-and-secondary-roads
[38] https://catalog.data.gov/dataset/tiger-line-shapefile-2023-state-maryland-primary-and-secondary-roads

The resultant map highlights areas that the three governments could focus on in an attempt to improve transit access in the DMV. While I generally support expanding the subway network, added busing is often the easier and faster solution. In some areas, like around Gaithersburg or Centreville, this could likely be addressed with slight changes or extensions to existing bus routes. Meanwhile, added focus should be placed on entire new routes along the Leesburg Pike corridor between Tysons and Leesburg, as well as between Sterling and Centreville past Dulles International Airport. While existing tracks go by target areas east of the city, the map shows that using this track to improve access would create an extraordinarily inefficient route into Washington, and so other options should be prioritized.