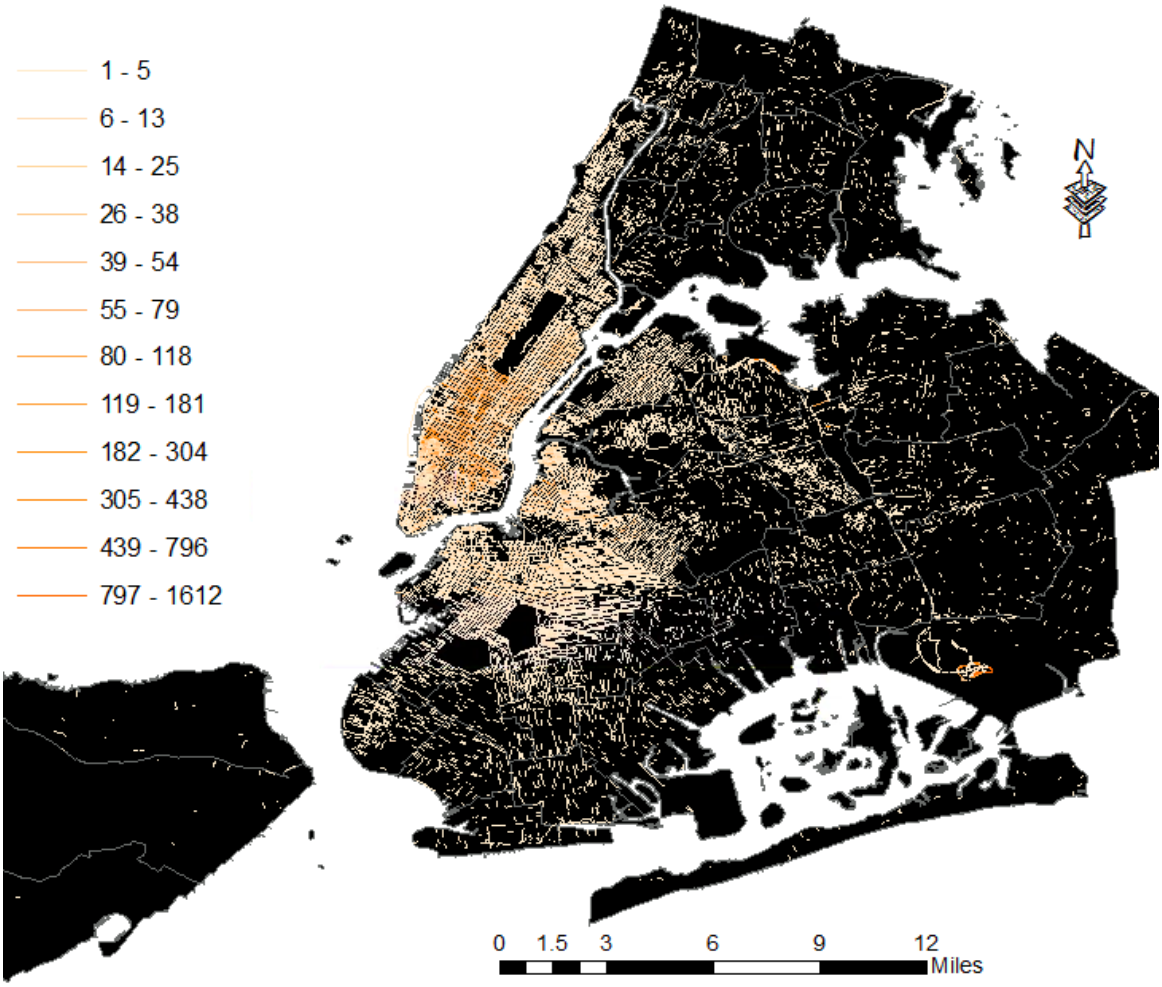

Fig. 1: Standard-scale Uber count between 10PM and 5AM aggregated from 1st ~ 31st of Aug. 2014

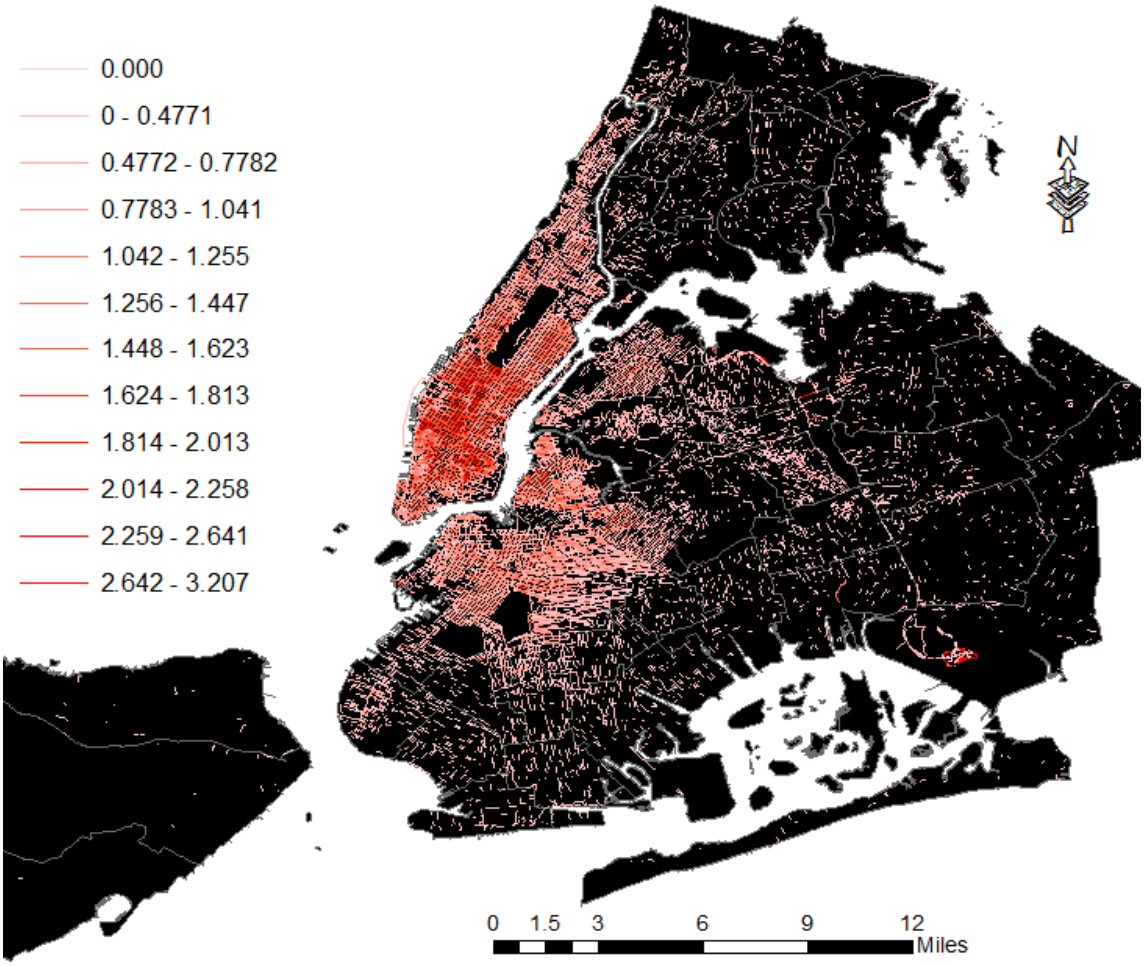

Fig. 2: Log-scale Uber ride counts between 10PM and 5AM aggregated from 1st ~ 31st of Aug. 2014

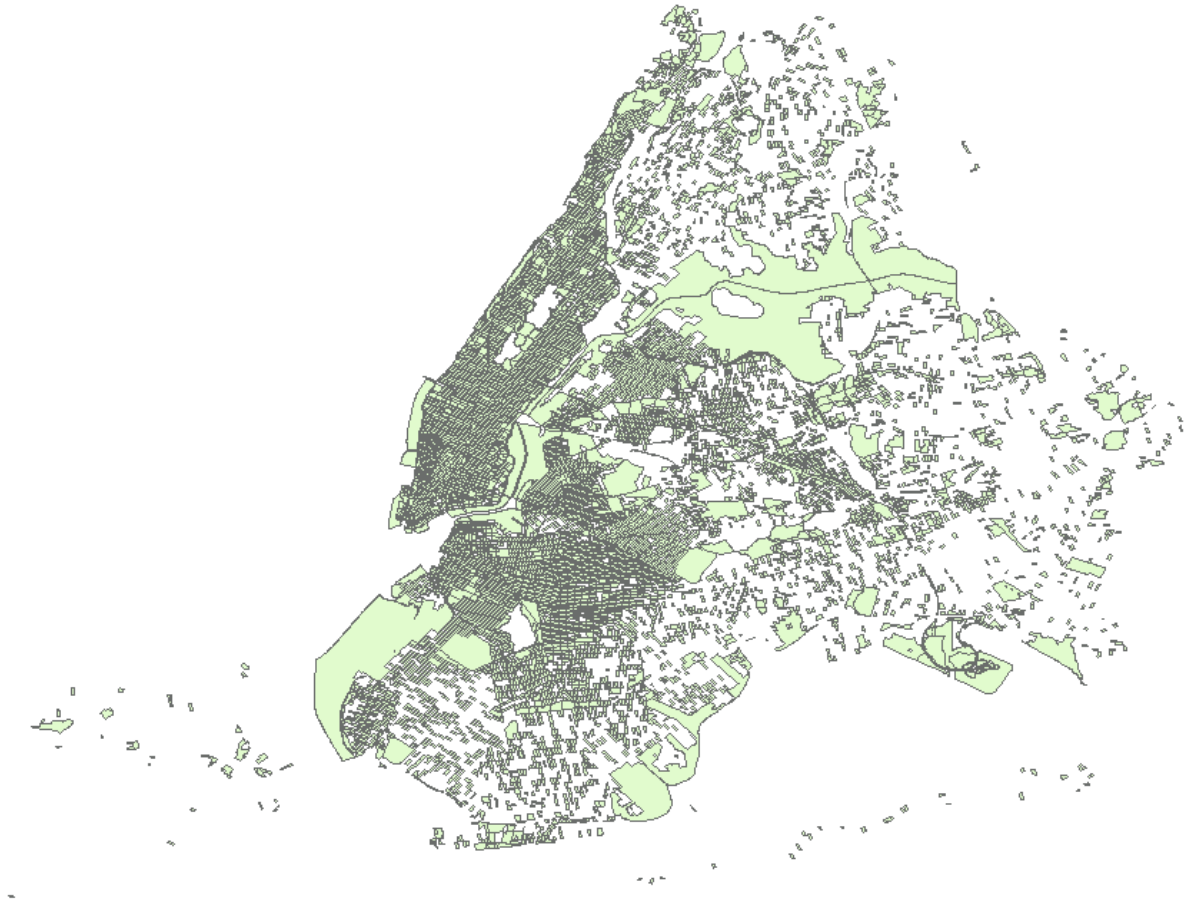

Fig. 3: Polygonized Streets as the spatial bins of Uber ride points

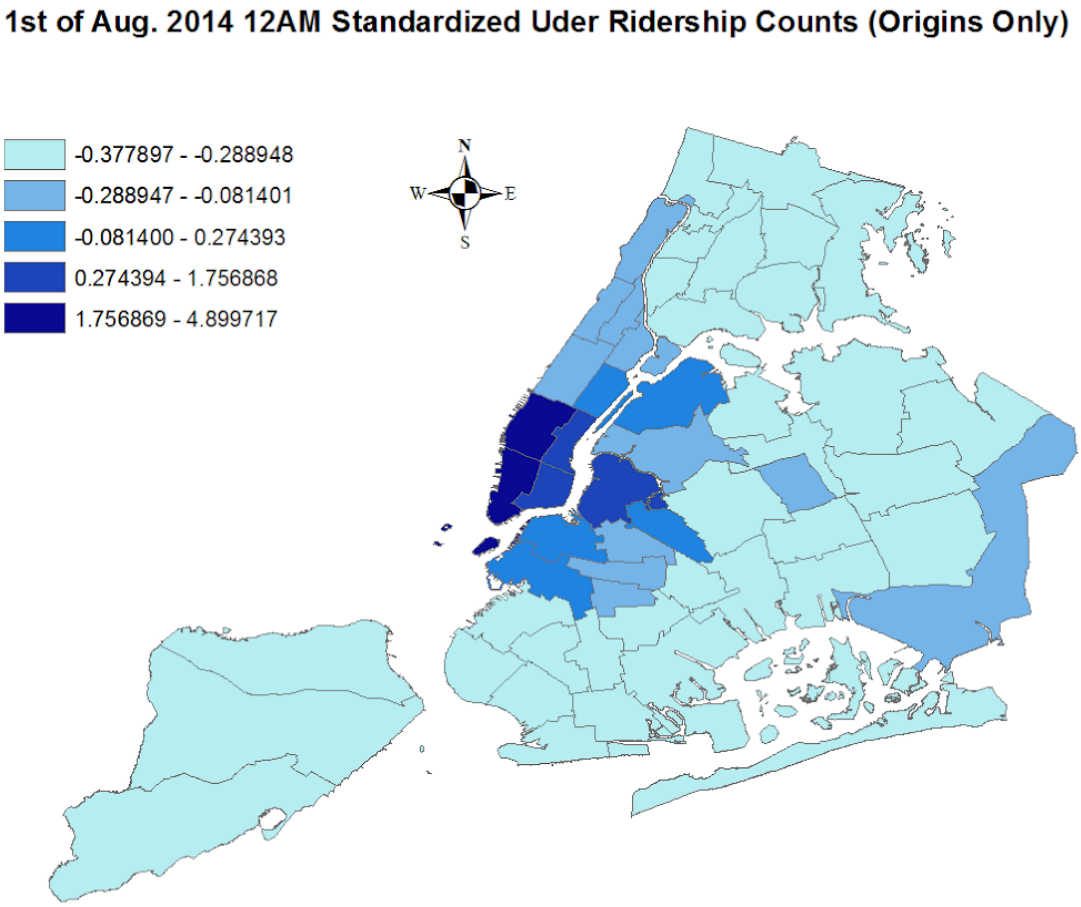

Fig. 4: A choropleth depicting the graduated color scheme for the five boroughs.

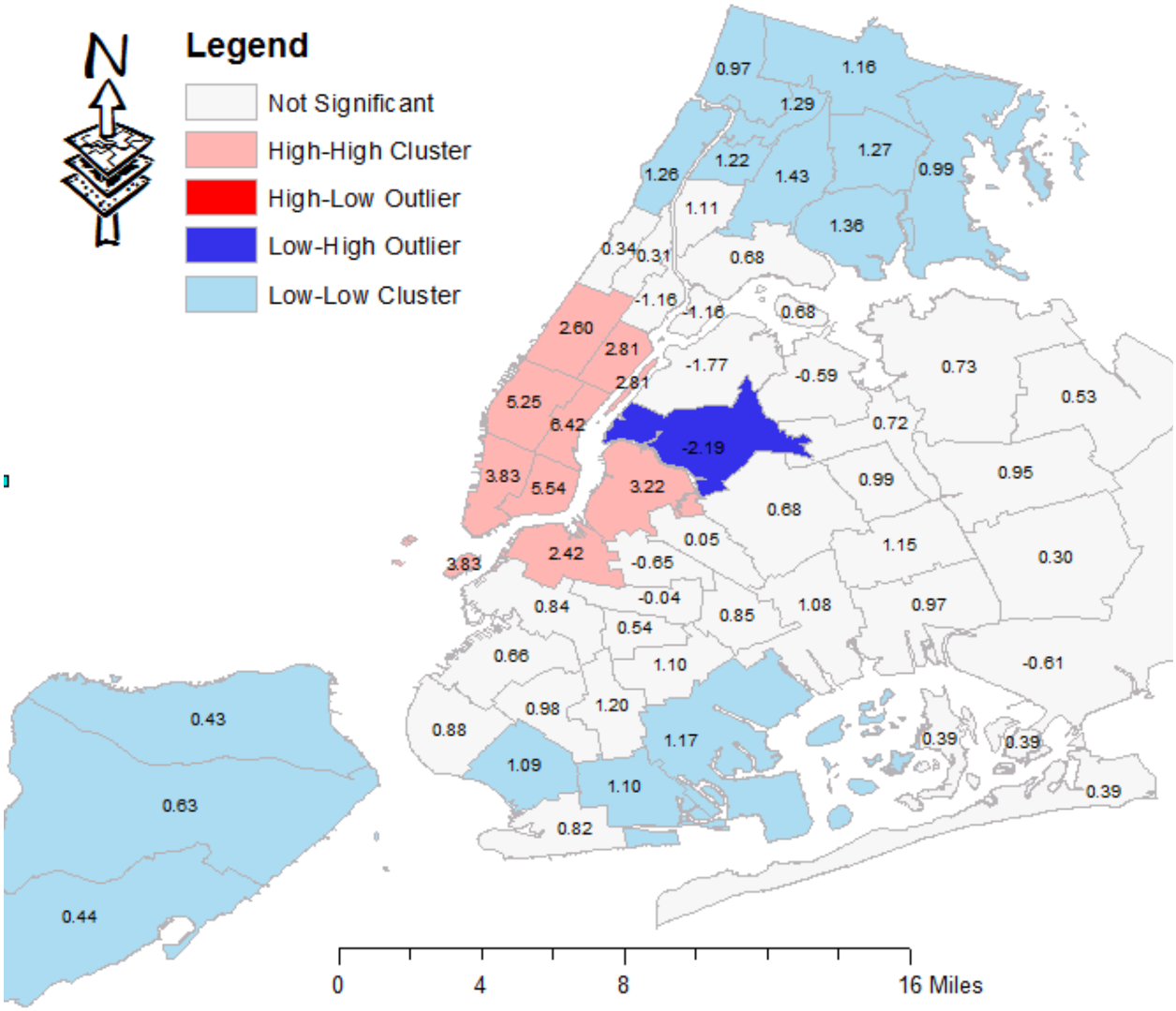

Fig. 5: The result of Anselin Local Moran's I test against PUMA level aggregation

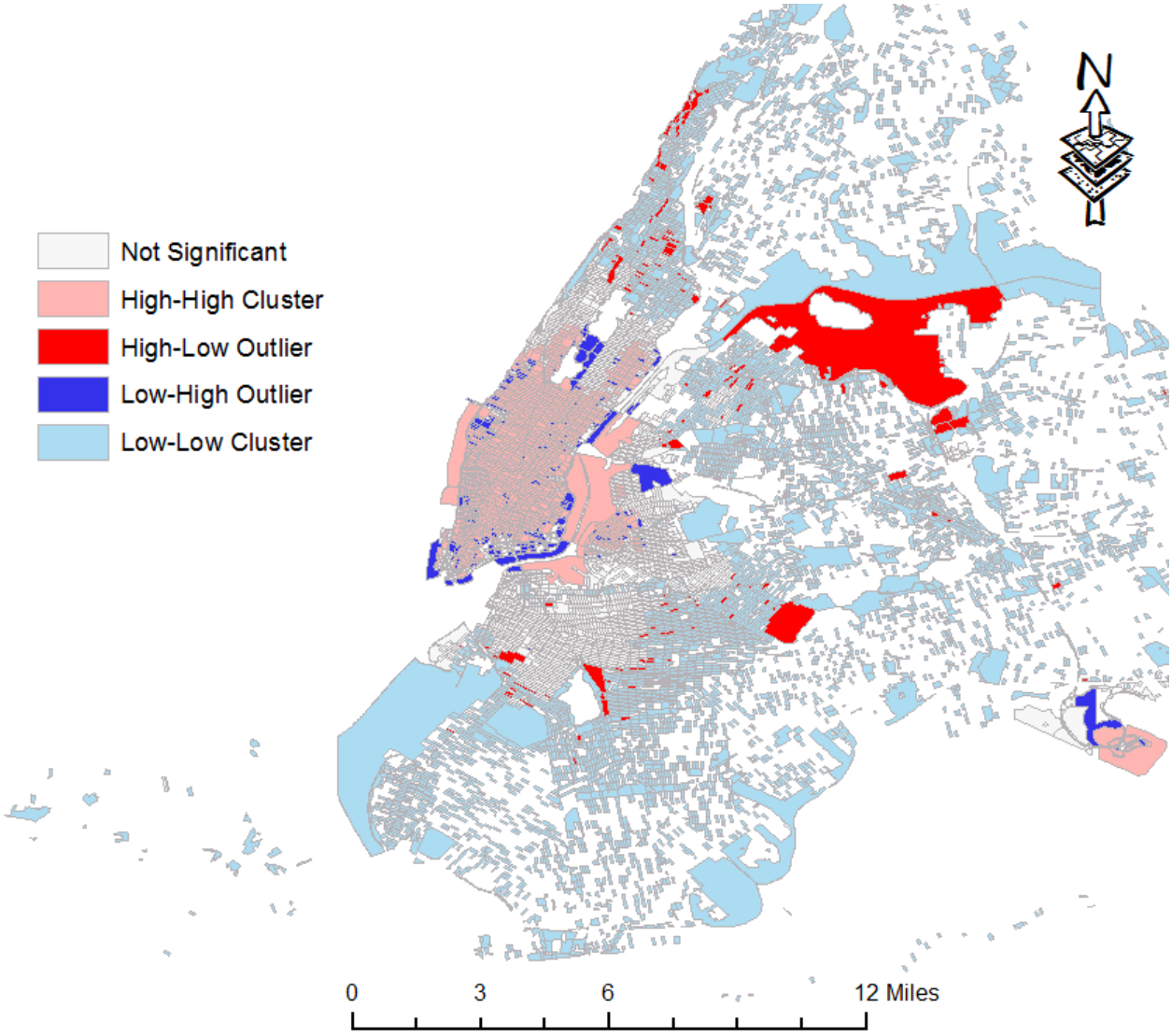

Fig. 6: The result of Anselin Local Moran's I test against Polygonized Streets

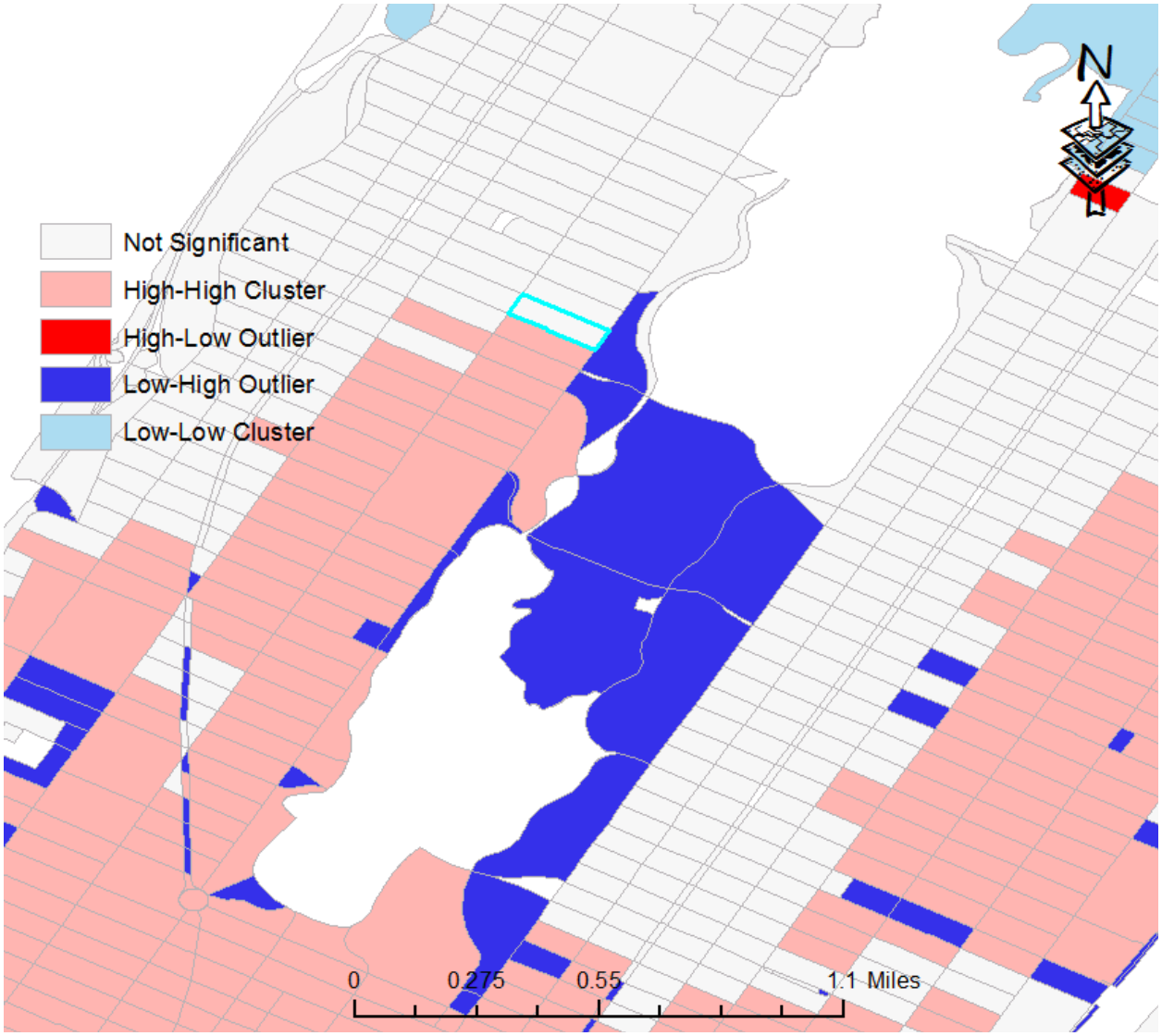

Fig. 7: A zoom-in view in the result of Anselin Local Moran's I test against Polygonized Streets

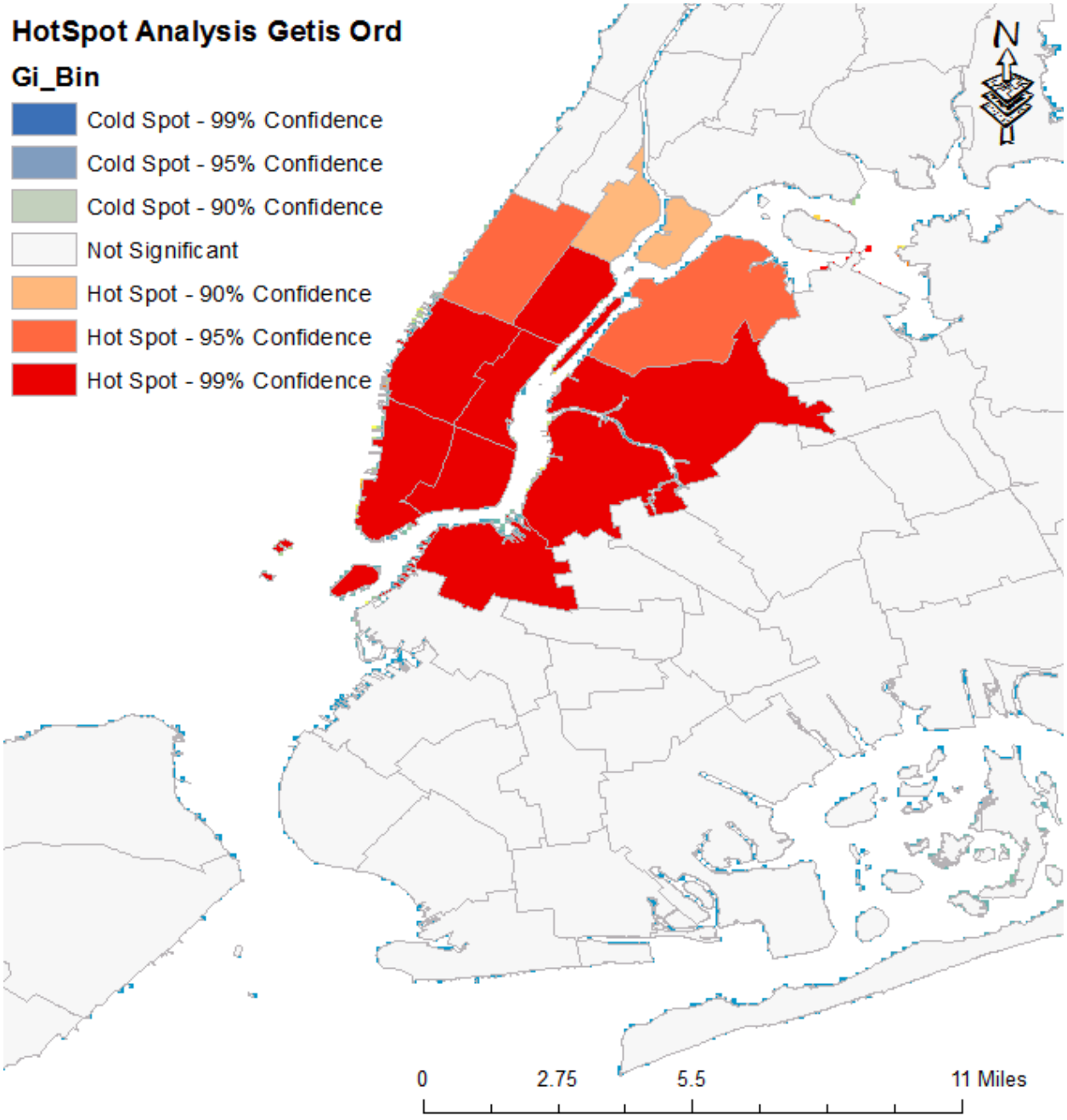

Fig. 8: Getis-Ord Gi* test result in PUMA level aggregation

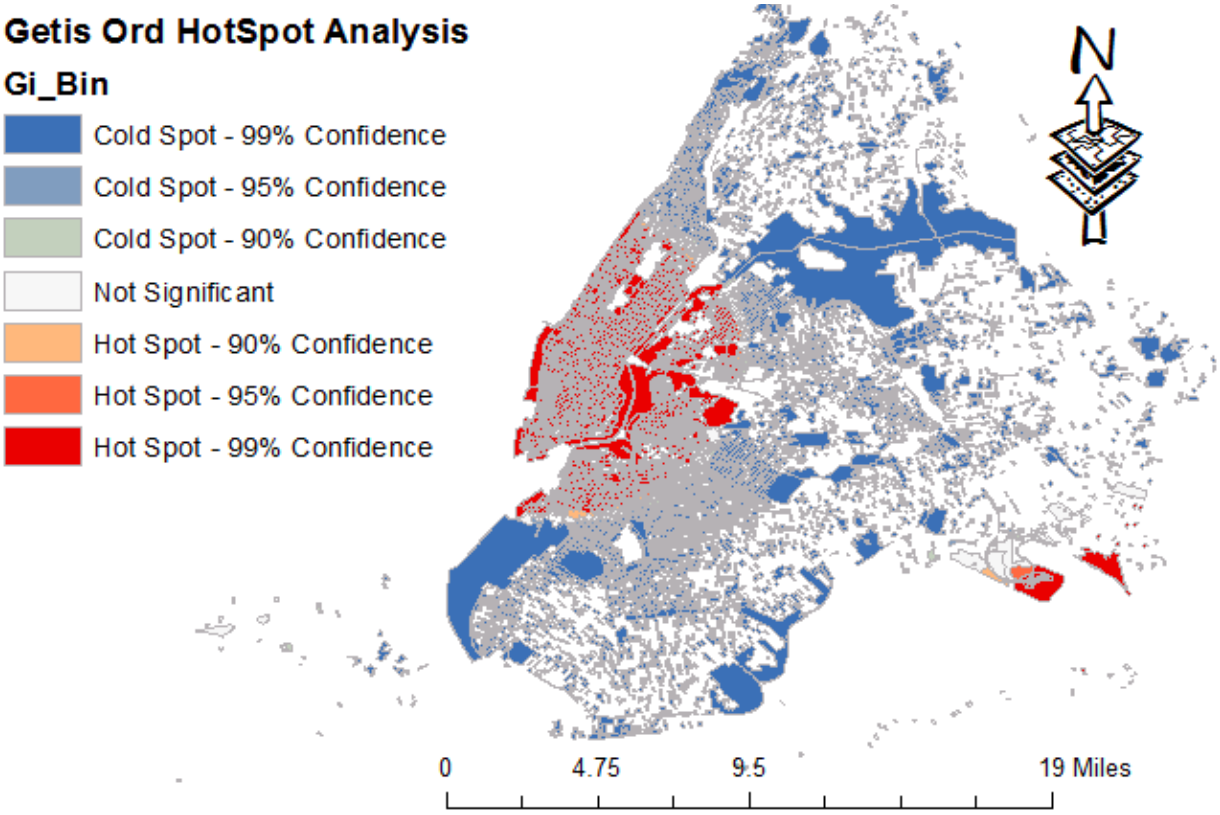

Fig. 9: Getis-Ord Gi* test result in Polygonized streets

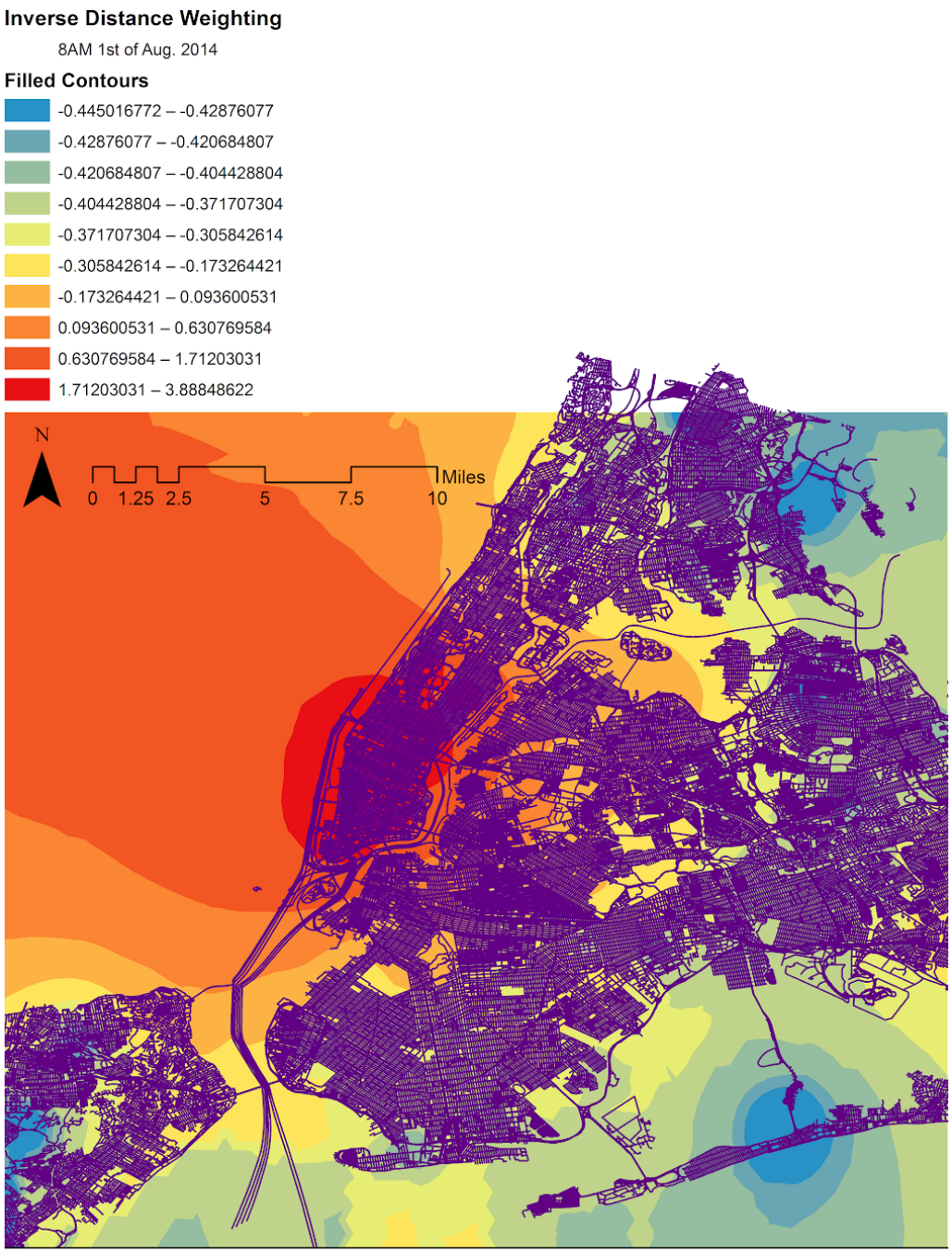

Fig. 10: IDW interpolation result using PUMA level aggregation

Fig. 11: A visualization of peaks of IDW interpolation in the centroids of each PUMA

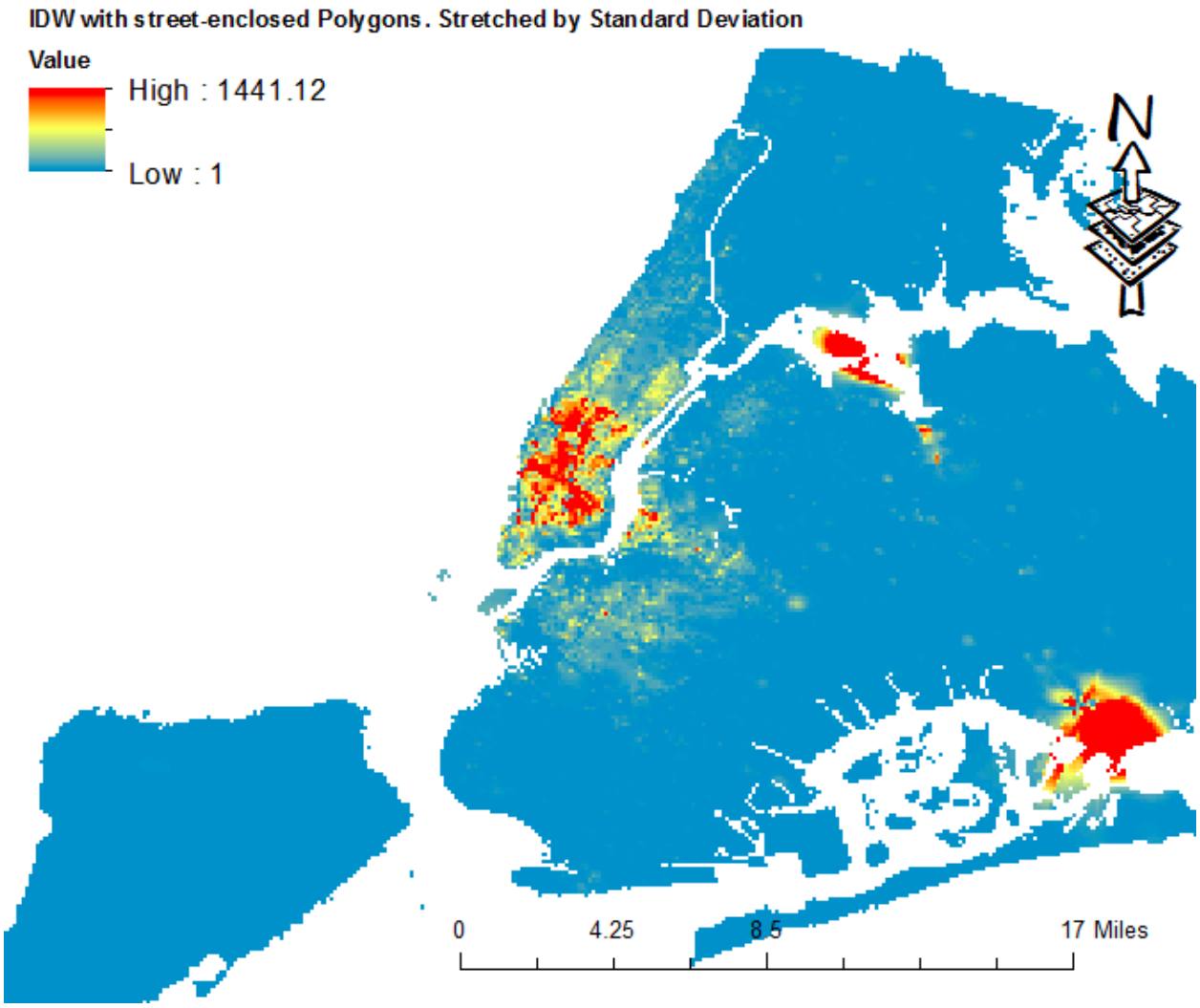

Fig. 12: Polygonized Street - Inverse Distance Weighting Interpolation Stretched by standard deviation

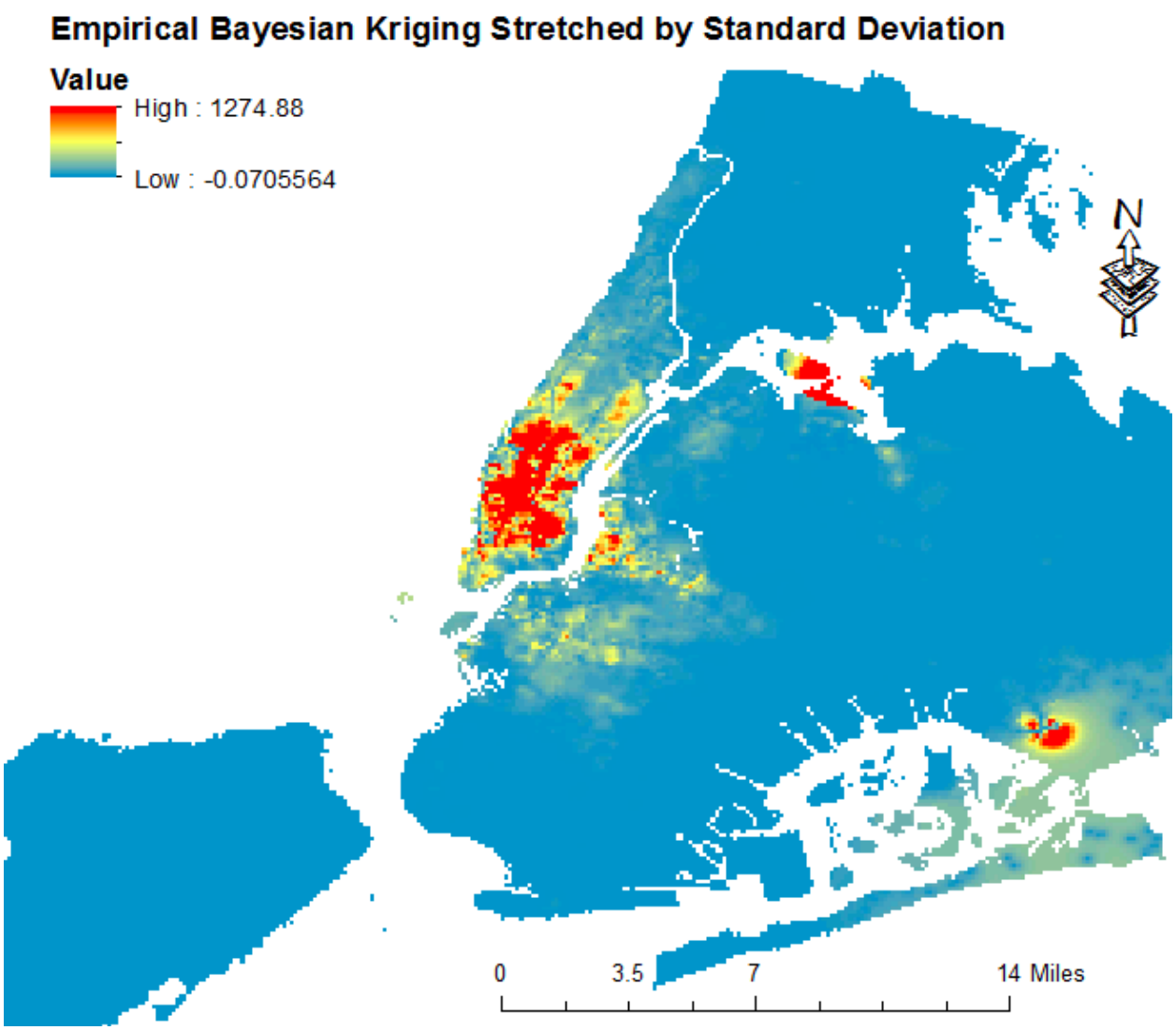

Fig. 13: Polygonized Street - Empirical Bayesian Kriging Interpolation Stretched by standard deviation

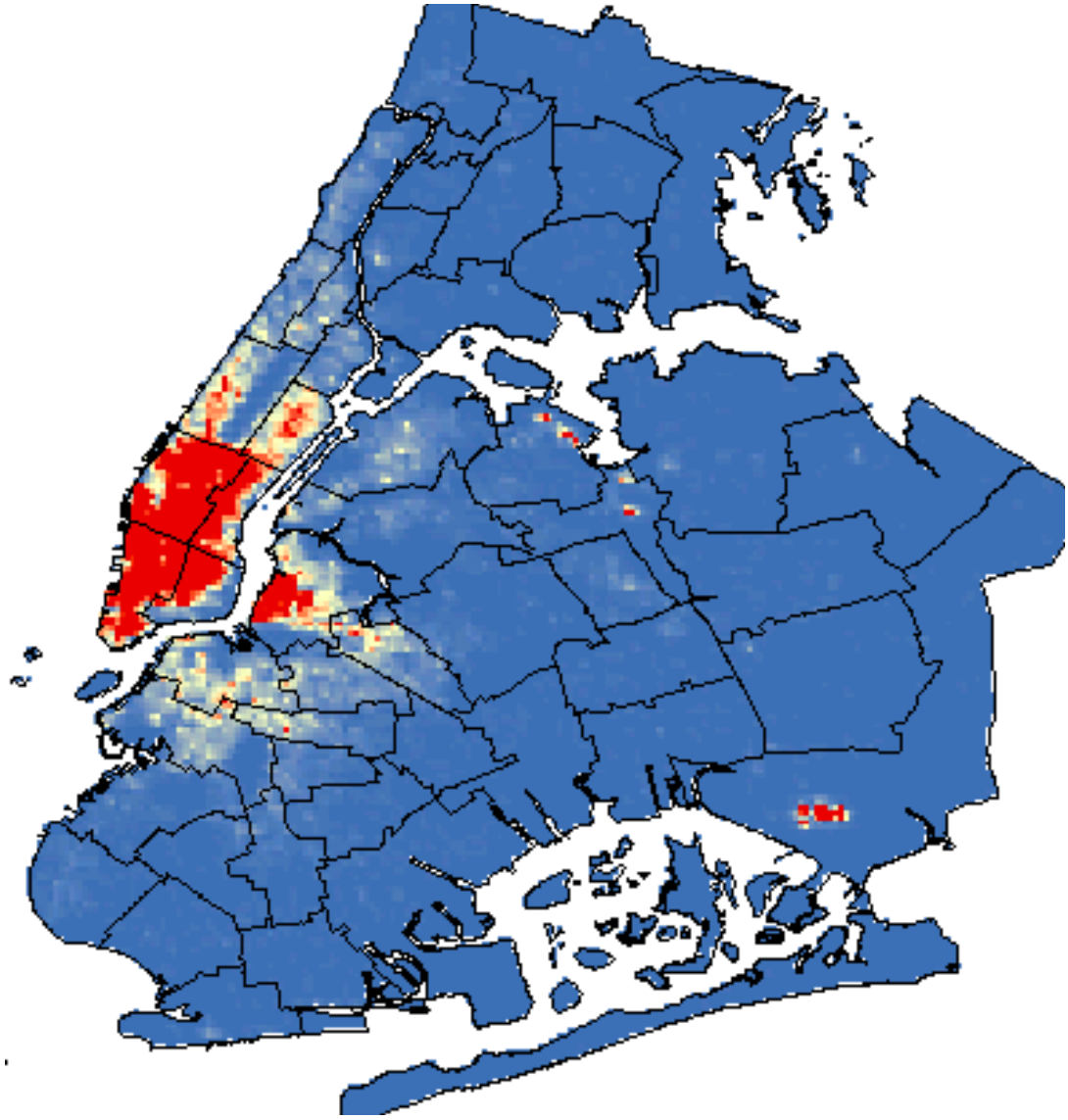

Fig. 14: Fishnet polygon IDW stretched by standard deviation

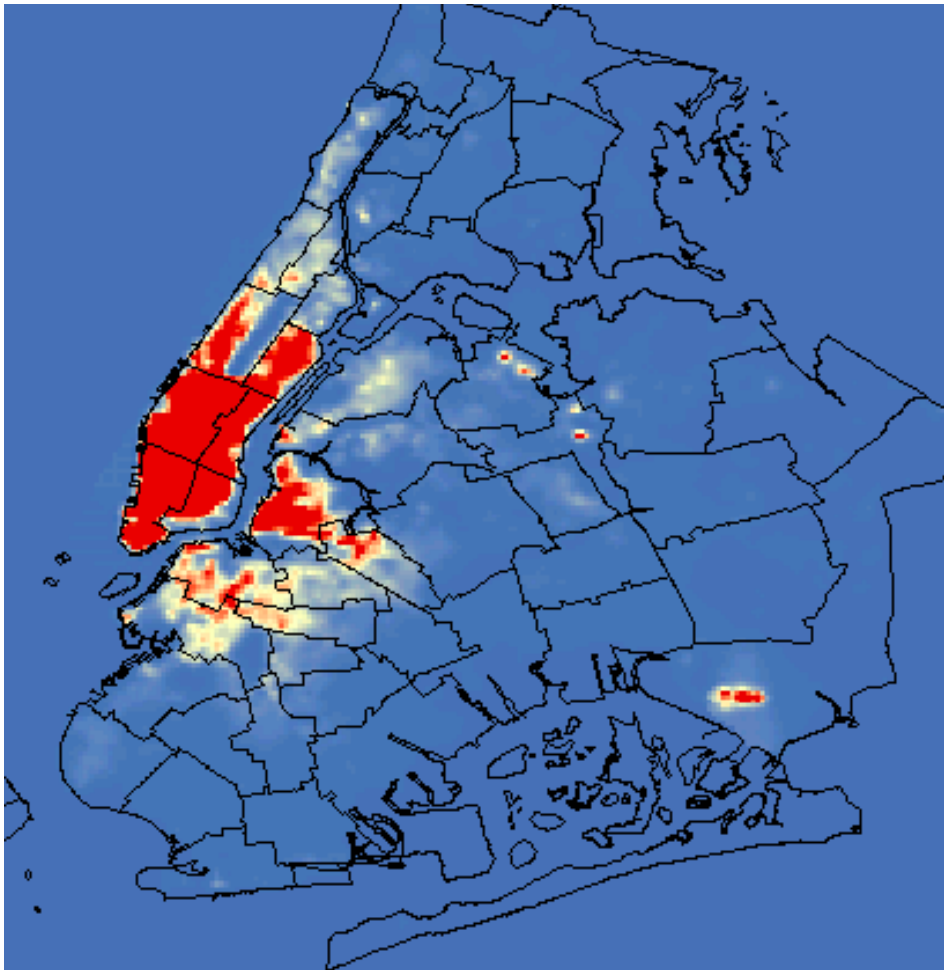

Fig. 15: Fishnet polygon Empirical bayesian kriging stretched by standard deviation

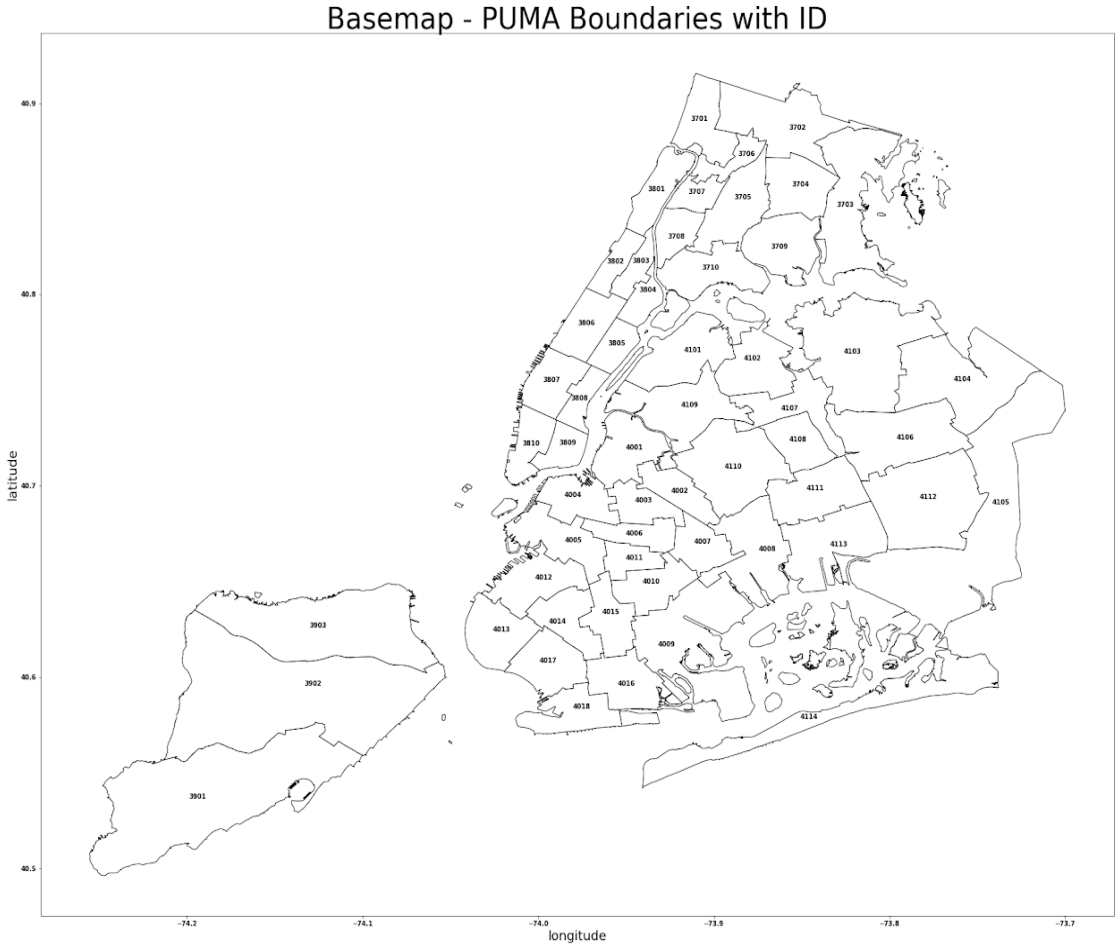

Fig. 16: PUMA Region By ID

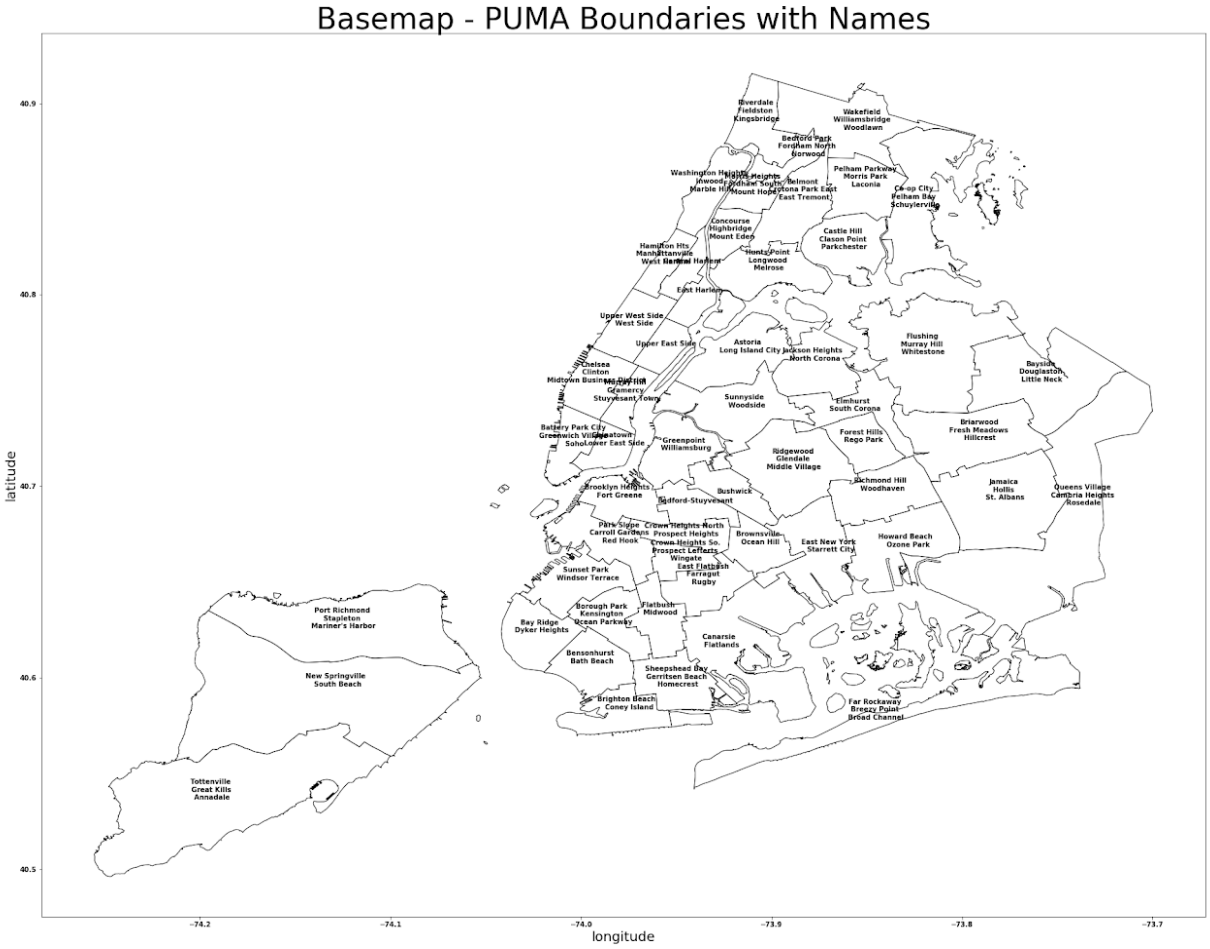

Fig. 17: PUMA Region By Neighborhood Description

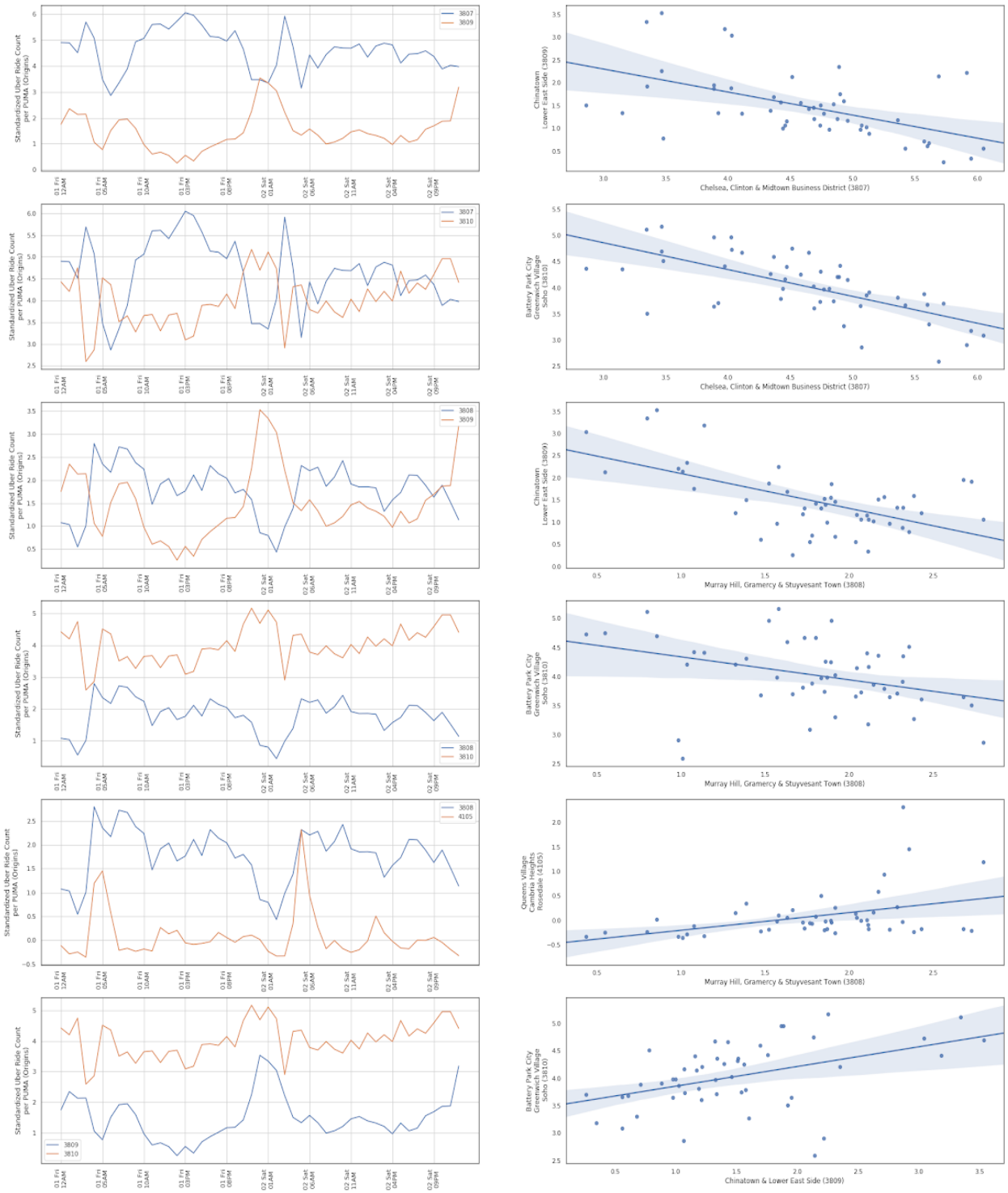

Fig. 18: Time Series Plot and Linear Regression on per PUMA Uber Counts

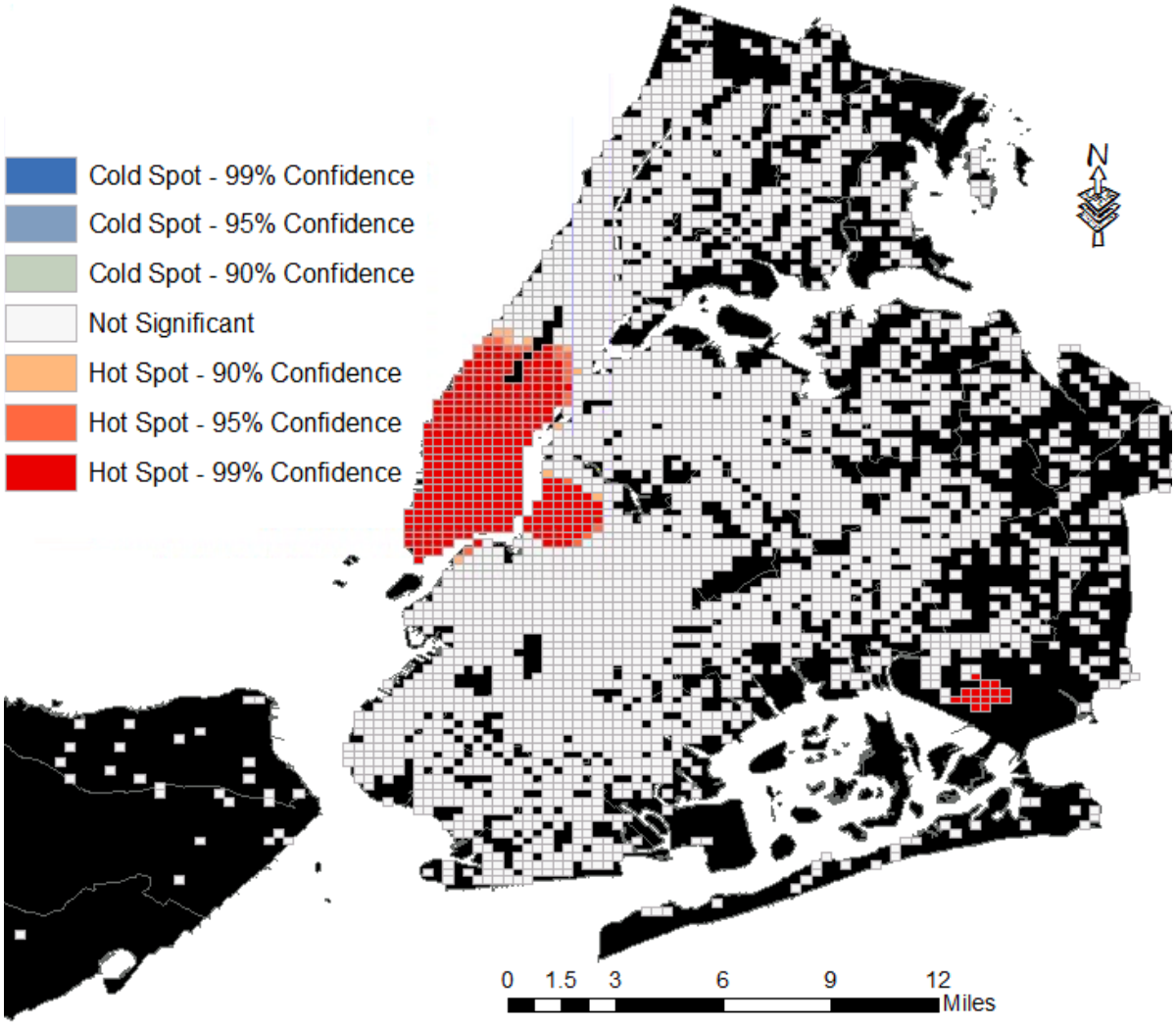

Fig. 19: Optimized Hot Spot Analysis Result of Uber ride counts between 10PM and 5AM aggregated from 1st ~ 31st of Aug. 2014

In this applied spatial analysis, the following question is examined; ***Can we use spatial interpolation to find hot- and cold-spots in Uber ridership across five boroughs in New York City during night time?*** Answering this question is made possible because Uber opened its ridership data for August 2014.[1] Data comprises of the latitude, longitude and timestamp of the starting point of individual rides. Preprocessing filters data such that the analysis treats data between 10PM to 5AM, the hours defined as *Night Time* in the scope of this project. The preliminary exploration of data is done by plotting points against the shapefile of New York City in a timelapse.[2] Recurring clusters in specific locations reveal the hotspots, ranging from the airports, spots in Manhattan such as the school district of East Houston and Chelsea Market, and Williamsburg. **Fig. 1** and **Fig. 2** aggregates above finding along the time dimension. Uber ride points are grouped and counted with respect to the closest street. All of Manhattan South of the Central Park serves as the hotspot and it sprawls along both sides of Central Park reaching 90th street. The other notable hotspot is JFK airport, whose entrance roundabout is clearly highlighted. Both figures show a presence of distance-decay from Lower Manhattan, where the hotspot extends across the East River to Williamsburg and Greenpoint, and experiences the distance-decay as it sprawls into the hearts of other boroughs. The project first demonstrates a misrepresentation by aggregation using a coarse granularity: PUMA demarcation. The problem is remedied by introducing a demarcation strategy based on natural and most granular bins, the streets, that can give rise to a better explanation and presentation of data than PUMA. Finally, using the street-level aggregation of Uber ride counts, the project culminates in examining the fit

[1] Fivethirtyeight. "Fivethirtyeight/Uber-Tlc-Foil-Response." *GitHub*, 14 Jan. 2016, github.com/fivethirtyeight/uber-tlc-foil-response.
[2] https://youtu.be/q0JjOO_AcmU

of the data to Inverse Distance Weighting(IDW) and Empirical Bayesian Kriging(EBK) spatial interpolation models. By improving the spatial granularity of data, the spatial interpolations that are originally designed to estimate the surface of underlying activities can be used to measure spatially discrete, point-based heatmap. Finally, binning data points by fishnet polygons is also discussed as another alternative.

The street shapefile from Open Data NYC[3] is used to generate *Polygonized Streets*. Converting the street polylines to polygons in Esri ArcMap allows for the generation of polygons that are enclosed by streets. A spatial join between street-enclosed polygons and uber rides as points aggregates point data to the closest polygon, which results in **Fig. 3**. The project undertakes the hotspot recognition using spatial statistics and interpolation with the polygonized street layer as well as PUMA polygons as aggregator, and compares the interpolation results.

The first view of data in **Fig. 4** reveals much about the overarching trend of the data. The standardized Uber count values are the highest in the four PUMA regions in the lower Manhattan where they exhibit staggeringly high Uber rides. As also depicted in **Fig. 1 and 2**, the choropleth captures the general trend of distance decay in the east of the East River that spread out to Northern Manhattan, Brooklyn Heights and West Queens (Williamsburg, Sunnyside, and Long Island City, etc.). It fails to identify the hotspot formed by JFK international airport.

Fig. 5 depicts the result of Anselin Local Moran's I analysis. It corroborates the observation of the emergence of the hotspot centered in lower Manhattan by marking all four PUMAs High-High cluster, whose periphery is formed by PUMAs across the East River. Surrounding this area from North and South, the two regions, the majority of Bronx, Southern

[3] Calgary, Open. "NYC Street Centerline (CSCL)." *NYC Open Data*, data.cityofnewyork.us/City-Government/NYC-Street-Centerline-CSCL-/exjm-f27b

Brooklyn and Staten Island are Low-Low clusters. The analysis identifies Sunnyside & Woodside as the Low-High outlier. All of Its adjacent PUMAs have higher standardized values than the outlier, whose effect is in turn amplified by the effect of contiguous Williamsburg.

The street level aggregation naturally returns a more granular result of Anselin Local Moran's I test (**Fig. 6**). One notable emerging pattern is that the major hotspots, Lower Manhattan and JFK Airport, are surrounded by Low-High Outliers. This is because the hotspots dwarf the neighbors, forming a ring of Low-High Outliers. The test results in a clearer demarcation of hotspots. This can be observed in the Lower Manhattan Hotspot, as it stretches northward on the either wings of central park, reaching 90th street. **Fig. 7** is the zoom-in view of this area. Anselin Local Moran's I test reflects in its result the improved granularity. A zoom-in view reveals clusters and outliers in micro-scale.

Fig. 8 and **9** are the results of Getis-Ord Gi* hotspot Analysis with PUMA and street-level aggregation respectively. Getis-Ord Gi* does not benefit as much from the improvement of granularity because the local statistics are calculated against the entire data set, which works as a smoothing effect. The concentric hotspot around in the Lower Manhattan visualizes this effect.

The majority of ride hailing happens in areas with big pull effects, and between these peaks are noises of random Uber rides. The result of the IDW interpolation using the data set aggregated by PUMA area is shown in **Fig. 10**. While the interpolation succeeds in identifying the hotspot and the general trends of Low-Low clusters (marked by blue troughs), the peaks and troughs are located in the centroids of respective PUMA polygon, which is not representative of the real location of the ride-hailing events. This misleading visualization is made more obvious

when grayscale is used over black shapefile, as shown in **Fig. 11** The illuminating spheres concentric with the host polygons are the peaks of the interpolation.

Fig. 12 is the result of the IDW interpolation that uses more granular polygons. It results in more granular centroid points and thereby more granular *heatmap* that better represents the reality. The Lower Manhattan, LaGuardia, and JFK are the main hotspots, while Brooklyn Heights, Greenpoint, and Williamsburg form the periphery of the Lower Manhattan hotspot. The visualization of the EBK (**Fig. 13**) is very similar to the outcome of IDW, and exhibits an improved granularity of the interpolated surface, thereby expressing more local hotspots.

The project may be extended to comparing the performance of street-level granularity against randomly generated fishnet. A spatial join outcome between a fishnet generated using Esri ArcMap and points of Uber ride starting points is performed to create the data for interpolation. The IDW and EBK interpolations resulted in **Fig. 14** and **15** respectively. Polygonized streets can include erratic polygons, such as Grand Central Parkway tangential to La Guardia airport in **Fig. 12,** which is over-represented, because the adjacent body of water is falsely included in the polygon. A fishnet can avoid this problem by finding the *pixels* of polygons that cover the extent of the hotspot areas. A fishnet may serve as a benchmark for the optimization of polygonized street generation.

The project examines how the inaccuracy of coarse spatial aggregation can be improved by using a more granular spatial demarcation. When it is provided, EBK and IDW interpolations improve its recognition of hotspots. The resulting spatial interpolations generate surfaces whose pattern generally agree with Anselin Moran's I and Getis-Ord Gi* results in both PUMA and street level. Results for the statistics tests also improve with the granularity spatial aggregation.

However, Getis-Ord Gi* Hotspot analysis fails to benefit as much from the improvement because it calculates the local statistic against entire data set, as opposed to its adjacent data points.