Emerging Hotspot Analysis of MassDOT Impact pedestrian crash data with population density normalization

Karl Tacheron – ktacheron@umass.edu
Final project paper
Geocomputation
Spring 2022

Table of Contents

1. Abstract

Identifying locations dangerous to pedestrians is an important objective of sustainable urban planning. Through the spatio-temporal analysis of traffic crash data, we can identify places where pedestrians are at the greatest risk. The Massachusetts Department of Transportation provides crash data dating back to 2002, which documents 6.5 individual records in crash incidents, and over 77,000 crashes involving pedestrians. This data contains a wealth of spatial information about crashes, but state analyses do not typically normalize data for the density of areas, and represent danger as a function of absolute crash counts only. ESRI ArcGIS's Emerging Hot Spot Analysis tool gives us the ability to detect where crashes are occurring over time, but its results fail to find hot spots in low-density areas. This paper seeks to determine if a method that normalizes crash counts across U.S. Census Block Groups can identify spatio-temporal hot spot patterns in Massachusetts crash data that are otherwise undetected.

2. Background

The Massachusetts Department of Transportation provides crash data dating back to 2002, which documents 6.5 individual records in crash incidents, and over 77,000 crashes involving pedestrians. This data has a wealth of information that is useful to urban planners that seeking improve the safety of their streets. Through the use of spatial statistics we have the potential to identify the most dangerous areas where pedestrian-involved crashes are occurring at the greatest rate.

MassDOT produces yearly reports of the places where the most crashes have occurred[1] using spatial clustering methods. For reasons discussed below, the identified intersections are generally located in the most densely populated areas of the state on arterial roads. Though the absolute counts are helpful to understand where the most accidents are occurring, there is little attention to areas where there is a disproportionately high inherent safety risk to pedestrians.

2.1. Applications of Emerging Hot Spot Analysis upon crash data: strengths and weaknesses

This analysis finds areas with significantly higher values than the rest of a study area and assesses how those values are changing over time.

ArcGIS Pro's Emerging Hot Spot Analysis locates patterns across time by aggregating spatio-temporal data into discrete areas and spans of time and performing statistical analysis on each area within each temporal "bin" in the aggregated set. The statistical analysis produces a Getis-Ord statistic for each area, which identifies the likelihood that its values are significantly higher than the rest of the data set.[2] The results of this step are further analyzed for patterns that occur over time, which allows the Emerging Hot Spot Analysis algorithm to ultimately categorize it as one of 16 types of hot spot.[1]

Emerging Hot Spot Analysis can aggregate points across areas underneath individual units of a hexagonal or square "fishnet" grid, or below areas inside GIS vector layer polygons. In both instances, pattern detection seems to largely correlate with the underlying population density. Most pedestrian crash events occur in cities and towns where population density is the highest and there is the most interaction between cars and pedestrians. Overlaid upon a map of population density, a clear visual correlation between the results of Emerging Hot Spot Analysis and high density becomes apparent (Fig. 1). Hot spots found for aggregate counts below Census Block Groups seem to carry the same high correlation between density and likelihood of being a hotspot (Fig. 3).

3. Methodology

This tendency of Emerging Hot Spot Analysis to identify hot spots only in dense area necessitates a means of correcting for density. The methods described in this paper seek to determine if ESRI ArcGIS Pro's Emerging Hot Spot Analysis' results can be improved by normalizing crash counts across U.S.

1 The classifications take the form of 16 different types of "cold" and "hot" spots with distinct temporal characteristics. These classifications can be seen in the legends for Figs. 2, 3, & 4, and on the documentation page for Emerging Hot Spot Analysis tool.[3]

Census Block Groups, in order to identify spatio-temporal hot spot patterns that are otherwise undetectible where they are spatially diffuse.

To accomplish this, MassDOT Impact pedestrian crash data was aggregated at the Census Block Group level with normalization applied for population density. This was achieved by creating a Block Group vector layer that contained yearly crash totals in units of crashes per 100,000 population.

All table join operations and transformations were performed with the Python libraries `pandas`, `geopandas`, and `numpy`. In addition to the descriptions outlined below, the process is documented in the project code repository (Section 6).

3.1. Data collection and preprocessing steps

Pedestrian crash data was gathered from the MassDOT Impact Data Portal[4] in one-year extracts of the Person-level table for each year from 2002 through 2019. Each observation within the tables corresponds to a person involved in a crash and contains WGS84 coordinates of the incident. Pedestrians are differentiated from drivers through a `PERS_TYPE` variable with the value "Non-motorist". The files were individually filtered for this value and concatenated into a master table of 77,007 non-motorists involved in crashes.

The "master" pedestrian crash table was used to create the first table used for the Block Group-level hotspot analysis seen in Fig. 3. Individual crash incidents were joined to a vector layer of Massachusetts' U.S. Census Block Group polygons with an "intersection" method that joined each crash to the underlying Block Group's `GEOID` attribute (the unique identifier used across Census data to describe each region surveyed by the Census). The year-by-year crash count summary table was produced containing the total count of crashes within each Block Group using the `DataFrame.groupby()` method. The dimensions of this output table were $19 \times 4{,}979$ – one record per GEOID containing yearly total crash counts for the data set's time period.

The second table of per-year crash counts that was used for the Block Group-level hot spot analysis normalized for population density (as seen in Fig. 4) was produced by dividing the first table by population density. A table of population estimates for each Block Group was obtained from U.S. Census data[5] arranged in the same shape. The population estimates were divided by the area in square kilometers of each Block Group to produce a population density attribute in persons per km^2. The first table of total crashes per year per Block Group was divided by this population density attribute to produce this second table containing a count of total crashes per 100,000 population.

3.2. Analysis

All three tables (the 77,007-record "master" table of pedestrian crash records and the two $19 \times 4{,}979$ tables of yearly crash counts per Block Group) were used for individual analyses with the ArcGIS Pro Emerging Hot Spot Analysis tool[3] , the results of which are displayed for an area of Boston (Figs 2, 3, 4).

The complete pedestrian crash table with all 77,007 pedestrians involved in crashes was used as the input to initial cluster analysis (as seen in Fig. 2). The point data for each was aggregated into 1km hexagonal bins containing aggregate crash counts within that area.

The resultant vector layers were analyzed for state change in the category labels assigned by the Emerging Hot Spot Analysis algorithm, with results provided in Figure 5 in absolute counts of features changed and as a relative proportion of the total features assigned the original tag.

For all Emerging Hotspot Analysis operations, 8 nearest neighbors were considered for the neighborhood area around each polygon being analyzed.

4. Results

The results of Emerging Hotspot Analysis upon the non-normalized yearly crash counts resulted in detection of temporal hot spot patterns in 11.8% of Census Block Groups in the state of Massachusetts. The same analysis upon the normalized set showed a net change in categorization to 417 Block Groups. Of the group of 584 Block Groups where patterns were detected, 274 (46.9%) changed to another categorization when their crash counts were normalized for population density. These changes are summarized as total counts and percentages of the original category size in Figure 5. Maps of the results across the Greater Boston Area for all three analyses are provided in Figures 2, 3, and 4.

Of those categorized initially as having no detectable pattern, 96.7% still did not have a detectable pattern in the normalized set. Of the 146 Block Groups that were assigned categories after normalization, most became “Consecutive” or “Persistent” hot spots.

Hot spots classified as “Intensifying”, “Consecutive”, “Sporadic”, and “Historical” kept their classification in at least 50% of cases, where “New” and “Persistent” hot spots changed to another category. Notably, “Diminishing” hot spots disappeared in the normalized analysis, being entirely reclassified as “Persistent” or “Sporadic” hot spots.

5. Discussion

The normalization of crash counts in the data set produced a significant overall change in the pedestrian crash hot spot assignments of Census Block Groups where a spatio-temporal pattern was detected. 46.9% of Block Groups identified as hot spots in Figure 2 changed classifications.

However, there was little effect upon the classification of low-density areas. Between the non-normalized and normalized crash count data sets, nearly 98% of Block Groups where no pattern was detected in the first analysis did not get assigned a pattern in the latter. Given the initial aims of the process to identify patterns occurring in very low-density areas, the process gave fewer results than hoped, but may provide a useful starting point for the introduction of additional normalization factors.

In Burlington, MA, the “hex-bin” analysis failed to find patterns (Fig. 2), but Block Group based analyses (Figs. 3, 4) did detect temporal hot spot patterns, possibly due to uneven distribution of crash

points that the hexagonal binning captured poorly. Overall, there are few of these areas relative to the size of the entire data set, but they may provide a starting point for closer analysis on smaller research areas where they occur.

There is the possibility that a spatial analysis of crashes may not be relevant to data that is situated within the bounds of a road network; a network analysis-based method may be necessessitated. My future research seeks to explore normalization methods that rely upon network attributes rather than spatial attributes alone.

5.1. Additional notes about data

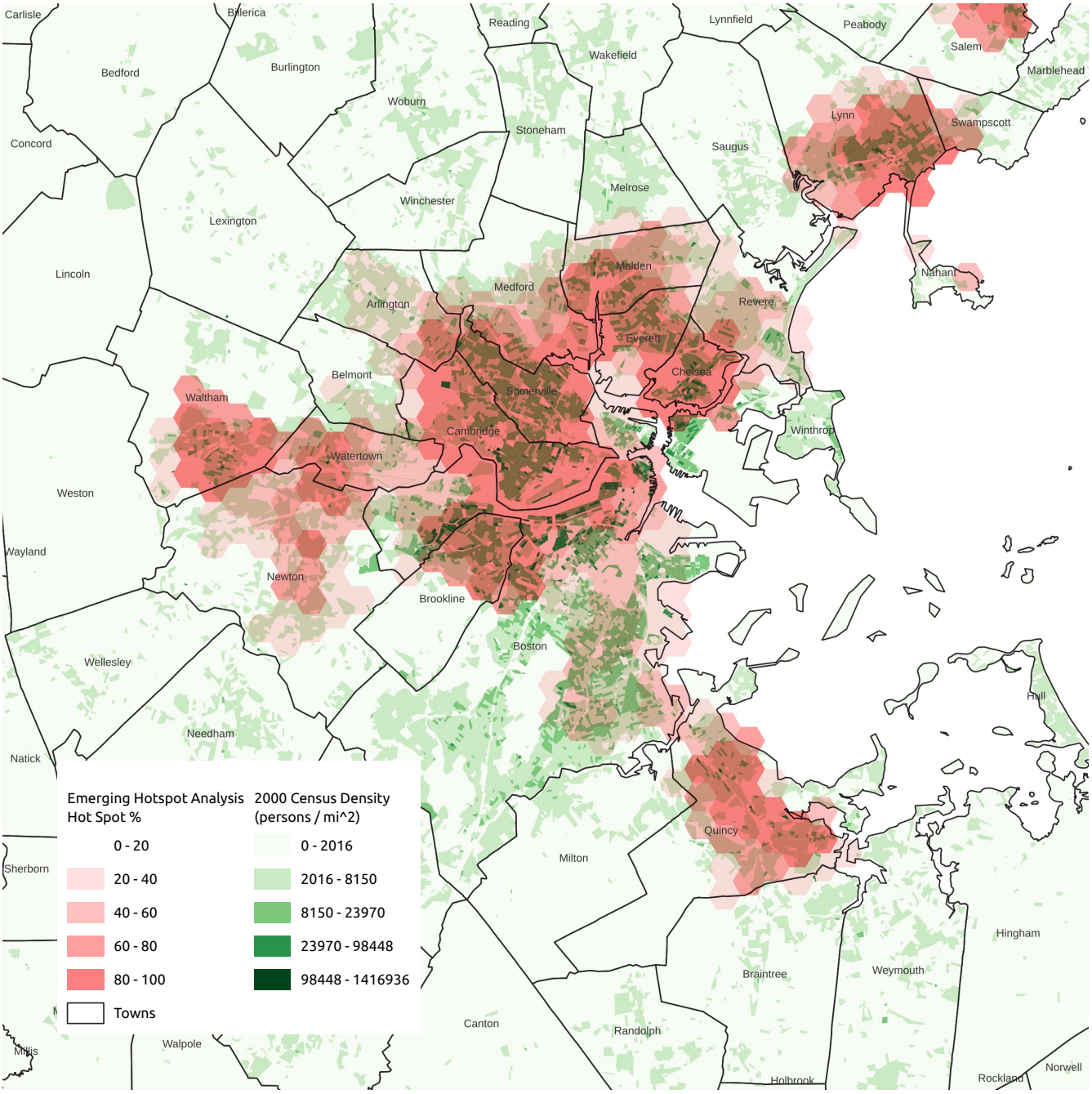
Before processing, the 18 CSV files were checked for data completeness (presence of columns across years, counts of NA values per column) to confirm that filter conditions would successfully capture crash events from the entire data set. Although MassDOT provides data for 2020 and beyond, the 2020-2022 data sets were not included as MassDOT does not currently consider those data comprehensive enough for analysis.

6. Github repository

The Jupyter notebooks used to create crash data joins can be accessed via my Github repository: https://github.com/karltach/geocomp22-final-massdot-impact

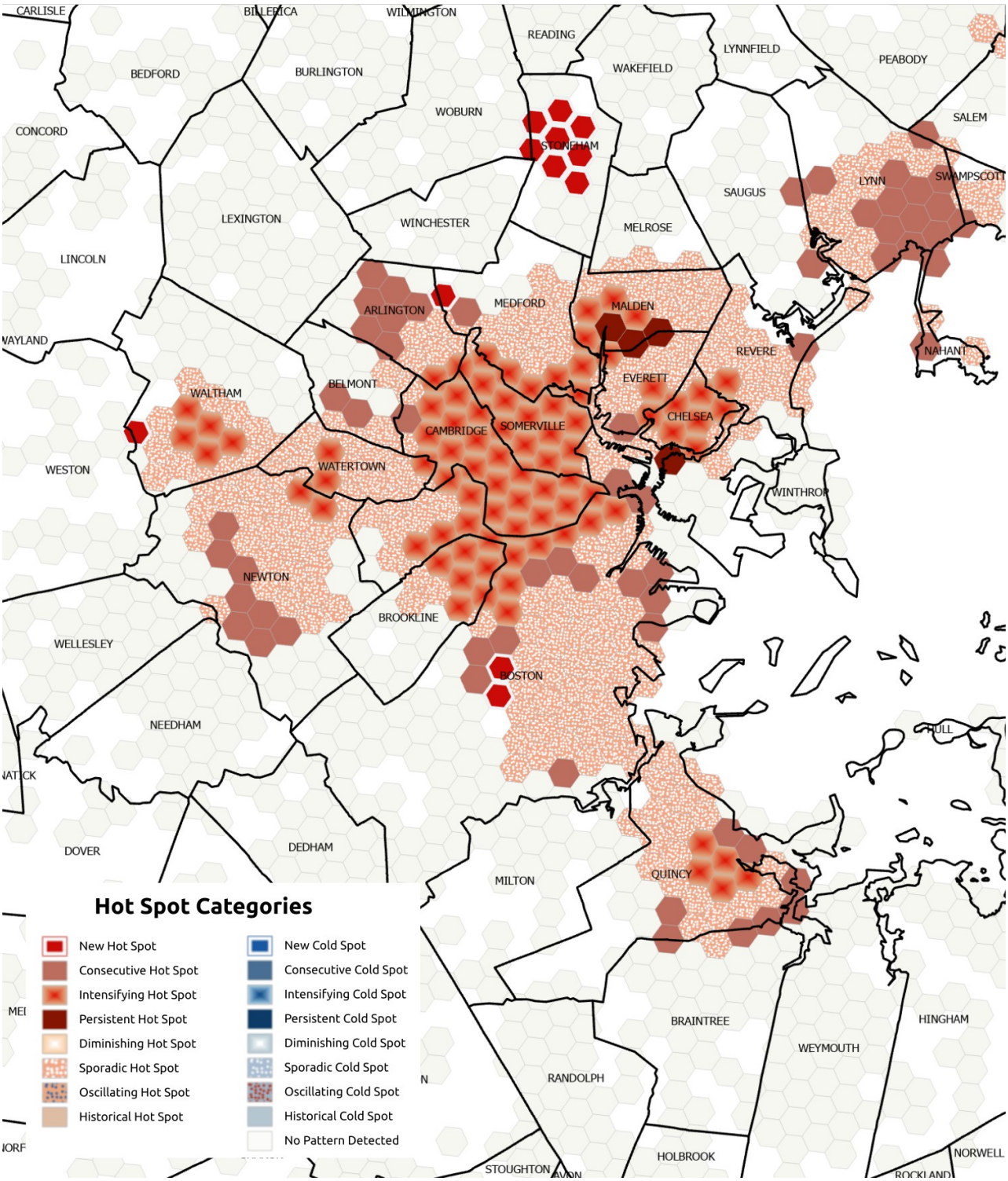
7. Figures

Figure 1. Emerging Hotspot Analysis versus Census Block population density



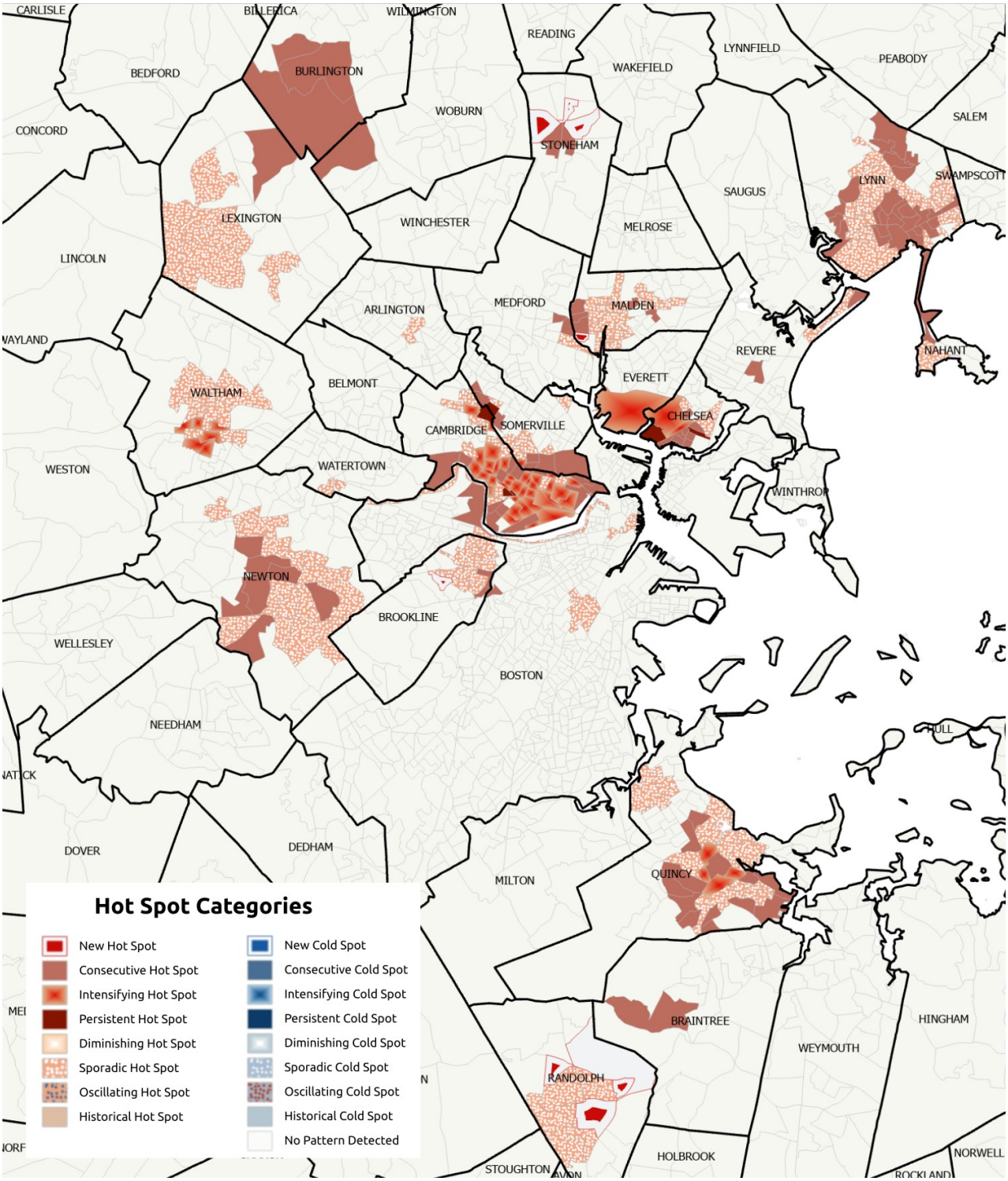
A comparison of Boston-area population density with the areas identified as hot spots by Emerging Hot Spot Analysis (seen with hot spot category labels in Figure 2). There is a high visual correlation between hot spot areas and high density.

Figure 2. Hotspot analysis with hexbins



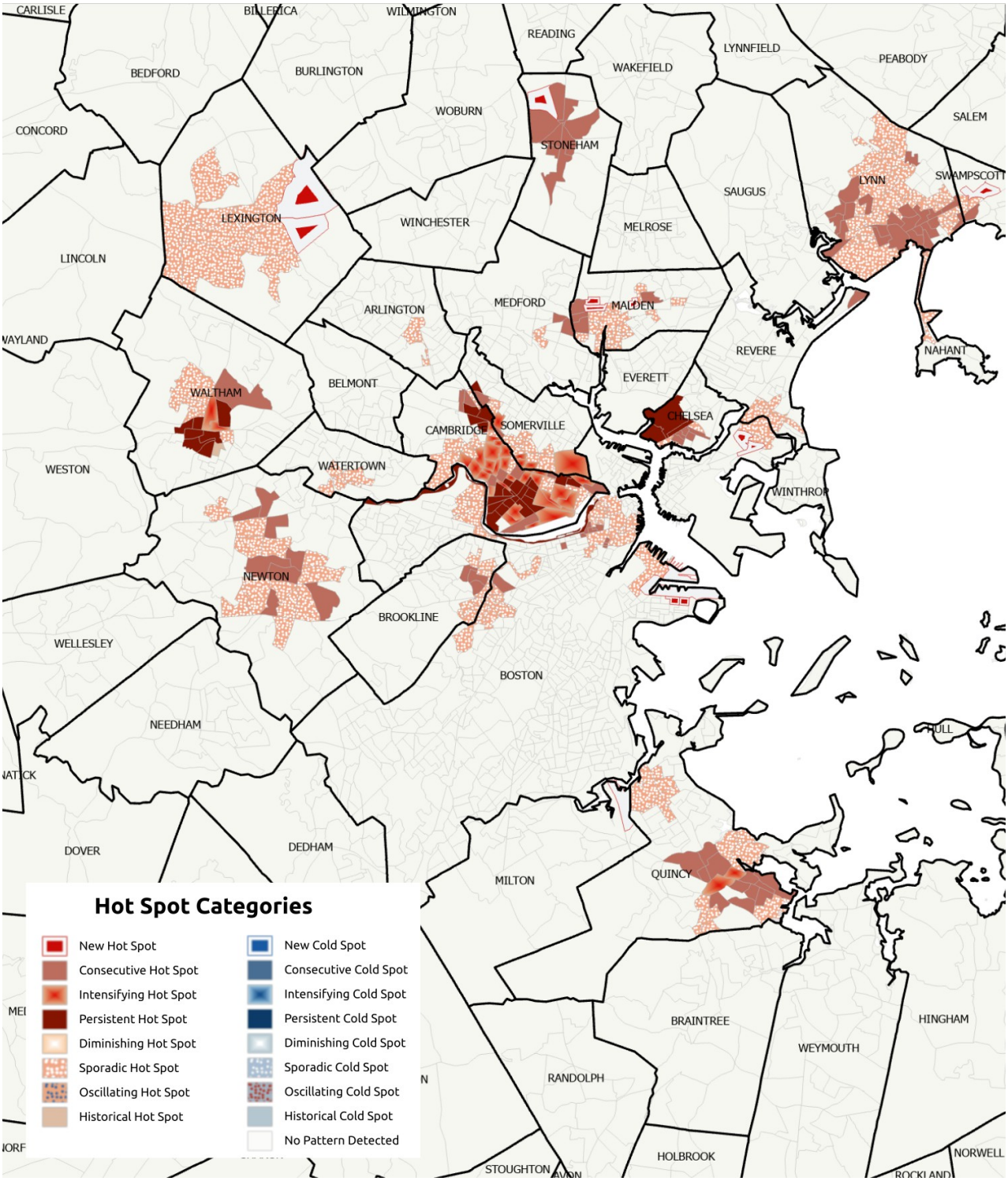
Initial Emerging Hot Spot Analysis performed on the “master” table, aggregated by the year of the incident. Each hexagonal bin has a diameter of 1 kilometer.

Figure 3. Hotspot analysis with crash counts aggregated into Census Block Groups, no normalization



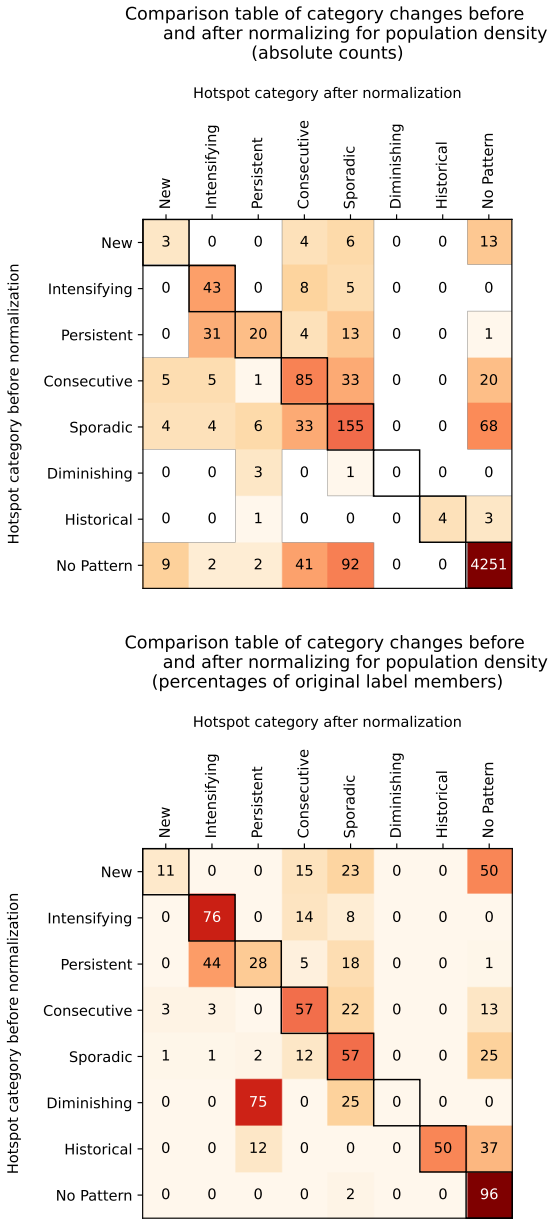
Results of Emerging Hot Spot Analysis performed on the non-normalized table of Block Group crash totals, aggregated by the year of the incident.

Figure 4. Hotspot analysis with crash counts aggregated into Census Block Groups, normalized for population density



Results of Emerging Hot Spot Analysis performed on the table of Block Group crash totals from Figure 3 normalized for population density.

Figure 5. Comparison table of Emerging Hot Spot analysis category changes before and after normalizing for population density



This figure demonstrates the amounts and proportion of Census Block Groups that changed their hot spot categorization before and after normalizing for population density. By comparing the Y-axis label with the X-axis label, one can understand the amount of features that either changed from one state to another or retained their original categorization. For example, we can see that 85 Block Groups labeled as “Consecutive” hot spots before normalization kept that label, or 57 percent of the 149 Block Groups initially given that label.

8. References

1. Massachusetts Department of Transportation. Top crash location reports by year | Mass.gov. https://www.mass.gov/lists/top-crash-location-reports-by-year.
2. E.S.R.I., Inc. How Hot Spot Analysis (Getis-Ord Gi*) works—ArcGIS Pro | Documentation. https://pro.arcgis.com/en/pro-app/2.8/tool-reference/spatial-statistics/h-how-hot-spot-analysis-getis-ord-gi-spatial-stati.htm.
3. Emerging Hot Spot Analysis (Space Time Pattern Mining)—ArcGIS Pro | Documentation. https://pro.arcgis.com/en/pro-app/2.8/tool-reference/space-time-pattern-mining/emerginghotspots.htm.
4. Massachusetts Department of Transportation. MassDOT Impact Data. (1998).
5. U.S. Census Bureau. American Community Survey Population Estimates. (1998).