Geospatial Analysis of Accidents Involving Bicycles and Bicycle Route Locations in Chicago

1st Caleb Griffy
Department of Computer Science
Middle Tennessee State University
Murfreesboro, TN, USA
cgg3a@mtmail.mtsu.edu

2nd Dr. Arpan Man Sainju
Department of Computer Science
Middle Tennessee State University
Murfreesboro, TN, USA
arpan.sainju@mtsu.edu

***Abstract*—Bicycle routes are designed to create a safe place for bike riders to use on the road. There has been a vast amount of previous research into traffic crash analysis, but focusing on bicycle routes and crashes involving bicycles has used basic methodology. This project uses an Empirical Bayes estimate to predict the expected crash value at different locations in an example city, Chicago. The difference between this expected value and the real count of crashes is used to identify hotspot locations for bicycle crashes. This experiment is repeated based on only crashes that take place on a bike route and only crashes that take place off a bike route. Using these results, maps are created to showcase locations where a bike route would be useful as well as locations where an existing bike route could be improved.**

***Index Terms*—anomaly detection, data analysis, Empirical Bayes, geospatial analysis, road safety, traffic crash analysis**

I. Introduction

Bicycles leave riders far more vulnerable than passengers in a car. If these two vehicles get into an accident, the bicycle rider is far more likely to get injured. Despite this, many people ride bikes along the same roads as cars to reach their destinations. Bicycle routes are created to help keep them separated and safe, but that does not eliminate the issue. Cars can still hit a bike on a nearby bike route, and not every road has a bike route. There is a need for finding places where bike routes could be improved or where they may be necessary in general. These locations can be useful for city officials to know where to look and make improvements. Bicycle riders would benefit from improved safety in these dangerous locations.

This project aims to use geospatial analysis on bicycle crash locations and their severity to answer three questions. First, where are locations a bike route should be added? Second, where are locations a bike route exists but should be improved? Third, do bike routes help lessen the severity of crashes? To accomplish this, a major city with heavy traffic was chosen as a data source. In the future, this approach could be copied with a different dataset to analyze another city.

When searching previous work for crash hotspot detection, the Empirical Bayes estimate, or EB-Estimate, proved to be a useful metric when compared to other means without delving into something as complicated as a machine learning algorithm [8]. The purpose of this estimate is to use the surrounding locations to predict what the number of crashes in a location should be. However, it also factors in the real value if the surrounding locations vary significantly. The EB-Estimate represents what crash count would be normal. Subtracting this from the real crash count at a location would get the ”local difference” at that location. If an area has a high local difference, it is a hotspot with a higher than normal crash rate.

Previous work has also been done to specify differences in crash data based on factors like vehicle type. Research by Lee et al. [5] takes this approach. Crashes are considered in groups by vehicle and also by the severity level (such as the level of injury or if there was a fatality). However, the model only identified hotspots by average crash frequency and severity. Using only crash frequency means areas with high traffic would be marked as the hotspots, even if each location has a low ratio of crashes given the traffic in that area. The main purpose of the paper was to demonstrate splitting the data into categories. The more advanced analysis techniques were not attempted in this way. In addition, the presence of bicycle routes were not considered.

This project aims to build on what’s come before by applying the EB-Estimate methods to exclusively crash data involving bicycles that is also separated on the attributes of crash severity and being on or off a bike route. In this research, crash severity will be specified by whether a crash involved a fatality or not. The presence of a bike route will be determined for a point based on the datasets used for the project.

II. Problem Definition

Chicago was chosen as the city to analyze for this initial project. The city contains some bike routes both in the downtown area and slower suburban regions. Downtown Chicago is an area with heavy traffic and a large number of crashes as well. In addition, Chicago provides up to date and free for use datasets to help with this research.

The input data for this project comes from the Chicago Data Portal, a source for public use data that is provided by different departments for the city of Chicago and is regularly updated. Input for the algorithms will need to be CSV files detailing the crash counts for grid locations in the city. To create these, four specific datasets were used.

First, the ”Traffic Crashes in Chicago” sets identify information about all vehicle crashes in the city. The first of these datasets contains the necessary information for its location

as a geometric object and a count on how many fatalities there were if any [2]. However, this dataset contains all types of vehicle crashes. Since this project is only concerned with crashes involving bicycles, the related vehicle version of this data is needed [3]. This dataset contains information about all vehicles that are related to crashes in the previous dataset. This information includes the vehicle type and an identifier for which crash it took part in. A Python script was created to filter the vehicles to only bicycles. Then, a script filtered the crash data to only crashes that were pointed to by an entry in the bicycle list. After doing this, as well as cleaning the data for rows with missing important attributes, the data went from 891,445 crashes to 14,553 crashes.

The output for this project should be several maps of hotspots for the city. These maps should be sorted into all the combinations of ways to split the data, such as by the presence of fatalities or being located on a bike route. These maps will use the local difference score as previously defined, but maps for the intermediate steps such as EB-Estimate and crash count would help demonstrate the usefulness of the local difference value. In addition, using the crash counts for crashes on and off a bike route, a ratio of fatal crashes to total crashes should be found for each so they may be compared.

III. Proposed Solution

The first step in solving the problem is to get the data from the Chicago datasets into a usable format. This was accomplished using the program QGIS. Originally, the crash dataset from the Chicago Data Portal contained the location of each crash as a geometry point, as well as how many fatalities were involved [2]. The related vehicle dataset was used to filter this list to only crashes that involved a bicycle, as well as filtering out rows missing necessary values [3]. The Chicago Data Portal also provided a dataset for bike lane locations stored as geometry [1]. This bike lane data and the bicycle crash data were both imported into QGIS. A grid was created to represent the city of Chicago. A fourth dataset containing a polygonal boundary of the city was overlaid with the grid [6]. Each tile of the grid represents a 200 meter by 200 meter area. If most of a grid tile's area is within city limits, the grid is counted. The crash data was split using the grid into four separate CSV files. Each entry in the file is a count of how many crashes of a given type appeared in the grid corresponding to that tile in the grid. The crashes were split into the following four types: fatal crashes on a bike route, fatal crashes off a bike route, nonfatal crashes on a bike route, and nonfatal crashes off a bike route. While fatal and nonfatal could be separated by values from the dataset, relation to a bike route took a bit more work. Using the bike route data, a buffer of twenty meters was created for a region to be considered "on route". This is enough space to cover the width of the road as well as crashes that have continued off the side of a road. Crash locations were sorted into on- or off-route based on whether they touched this defined buffer. These four files were used to calculate EB-estimates of their type for each tile separately.

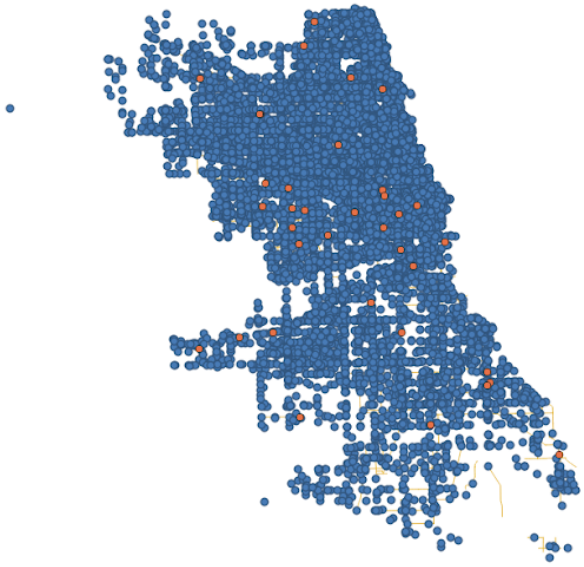

Fig. 1. Points from the Chicago datasets as arranged in QGIS, sorted by fatal and nonfatal.

Fig. 2. Bike routes from the Chicago datasets as arranged in QGIS.

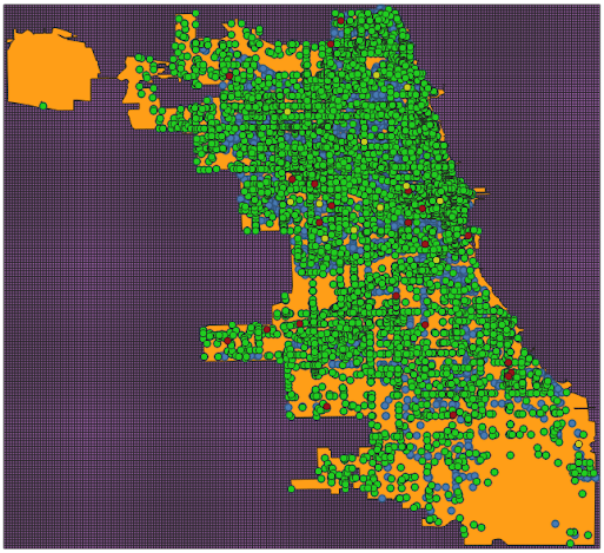

Fig. 3. The grid object and the points from the Chicago datasets as arranged in QGIS, sorted by fatality and also by the presence of a bike route.

Each of the datasets were redownloaded and rerun as of November 10th, 2024, meaning the crash and vehicle data was accurate up to that day. The bike route data had last been updated on the Chicago Data Portal on October 4th, 2024, and the city boundary data had last been updated on April 10th, 2024.

A. Empirical Bayes Estimate

To calculate the EB-Estimate for each tile, an algorithm was created in Python using the Pandas module. Two formulas are needed to calculate the EB-Estimate. First, the weight for the expected crash frequency is found:

$$wi = \frac{E[\lambda i]}{E[\lambda i] + VAR[\lambda i]} \tag{1}$$

The weight (wi) represents how much the EB-Estimate should trust the predicted value ($E[\lambda i]$). This predicted value is simply the average value of all neighboring cells. $VAR[\lambda i]$ represents the variance of these cells. If the neighboring cells vary heavily, the prediction is likely to be off, so the next formula should put more weight behind the historical crash frequency. This formula uses the weight to balance between these values:

$$\lambda i = wiE[\lambda i] + (1 - wi)xi \tag{2}$$

In this equation, xi is the historical crash count, or the original value at a location. λi is the final EB-Estimate for that location.

Each tile in the grid was accessed and run through this equation, with the eight surrounding neighbors being counted. There was no need to check for cells at the edge of the grid, because cells with a score of -1 were skipped and got an EB-Estimate of -1. This is the value used for cells outside the boundary of Chicago. The grid was made slightly larger than the city so that the edge of the grid would all contain values of -1. This means every cell for which an EB-Estimate was found was guaranteed to have all eight neighbors. When averaging the neighbors, scores of -1 were left out of the calculated average as well. This program was run with the input of each separated crash count CSV file, and generated a corresponding EB-Estimate CSV file.

To find the local difference of a tile, each grid tile only needed to subtract the value of its EB-Estimate from the original crash count. A CSV for these values was created for both on a bike route and off a bike route. However, as shown later, this map was hard to display due to having both over-predictions and under-predictions. Separate CSV files were created for the positive values and absolute values of the negative values as well.

IV. Evaluation

The first thing to check with the calculated data is the severity ratio, or how likely a crash was to be fatal given it taking place on or off a bike route. There were 5868 crashes on a bike lane, 17 of which were fatal. There were 7363 crashes off a bike lane, 22 of which were fatal. While there were more crashes off a bike lane than on, this metric does not immediately prove the hypothesis, as there are also more roads without a bike lane than with. However, 0.2897% of crashes on a bike lane were fatal, while 0.2988% of crashes off a bike lane were fatal. With a difference of 0.0091%, crashes off a bike lane were slightly more likely to be fatal. The hypothesis is supported, although the two ratios were incredibly similar. Given how close these values are, and how few fatal crash points of data there were, it's hard to call this answer definitive. It's possible that even if bike lanes do not make crashes less severe, they could make them less likely to occur. To answer this, more testing would need to be done using overall bicycle traffic patterns.

Next, the collected data for each tile was displayed in a heatmap for Chicago. Each tile in the below figures represents a 200 meter by 200 meter area in the city. Some of the maps have been repeated with a differently scaled legend to make them easier to read. Before comparing the visuals between two different figures, the legends should be examined to ensure they are scaled in the same way. The first five figures are all crash frequency of each tile. The next five display the EB-Estimates for each tile. Finally, the local difference of each tile is displayed once before being separated into maps for negative values only and positive values only.

To create these maps, Matplotlib's Pyplot module was used. The values from the grid were printed using a color map directly. The grid already works as a map since the values are stored based on their location in the city. Values of -1 still represent areas outside city boundaries, and are colored gray to differentiate them.

The maps for the EB-Estimates have a blurrier look than those of the actual counts. This shows the values are being influenced by their neighbors. The same general shapes of lines across roads exist, but high-traffic areas like downtown

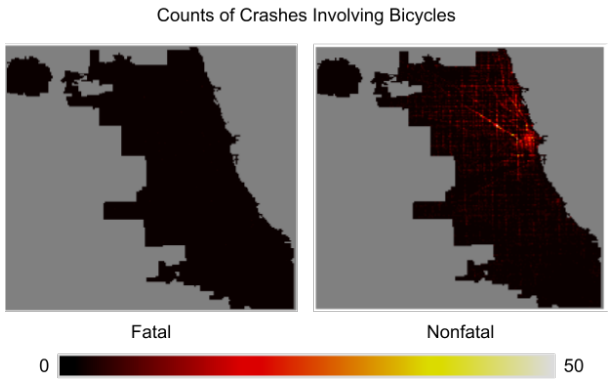

Fig. 4. Counts of Crashes Involving Bicycles.

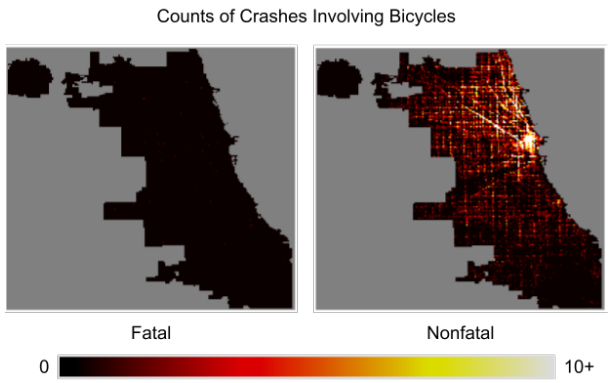

Fig. 5. Counts of Crashes Involving Bicycles with an Adjusted Legend.

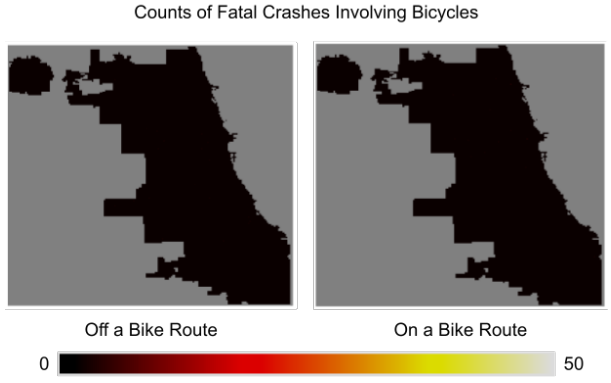

Fig. 6. Counts of Fatal Crashes Involving Bicycles.

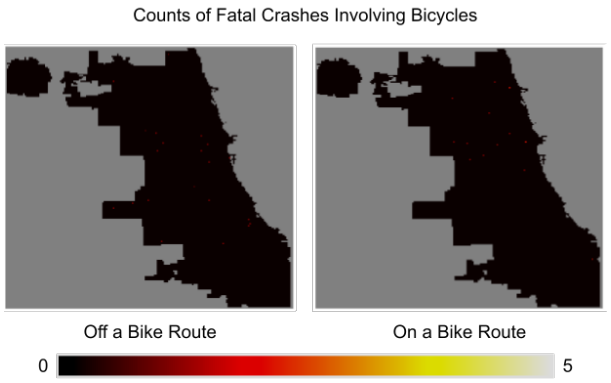

Fig. 7. Counts of Fatal Crashes Involving Bicycles with an Adjusted Legend.

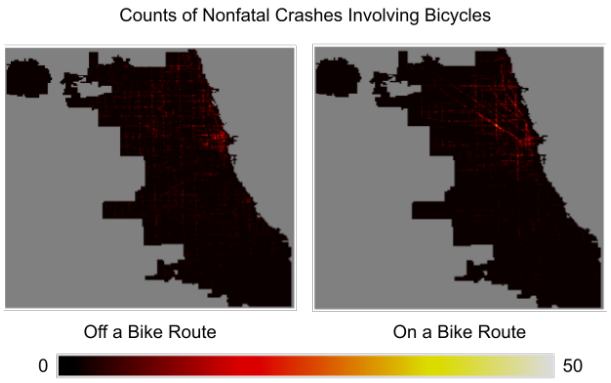

Fig. 8. Counts of Fatal Crashes Involving Bicycles.

have been smoothed out into a filled shape. Specific dots where crashes spiked in frequency are softened immensely in the EB-Estimate map.

The maps for both metrics that specify fatal crashes are very sparsely populated. There were not as many fatal crashes involving bicycles in the dataset as there were nonfatal ones, so most locations have only one crash if any. There are very few locations with only more than one fatal crash. In the EB-Estimate map, these locations are even more scarce. The predictions for the locations say there should be hardly any fatal crashes.

Exploring the crash frequency and EB-Estimate maps shows that the downtown area has far more crashes than other areas in Chicago. Main roads leading out of downtown also have high values. This is due to the heavier amount of traffic in that area. The local difference maps take into account this expected level of traffic and filter out that area's bias.

The negative local difference map almost works as iden-

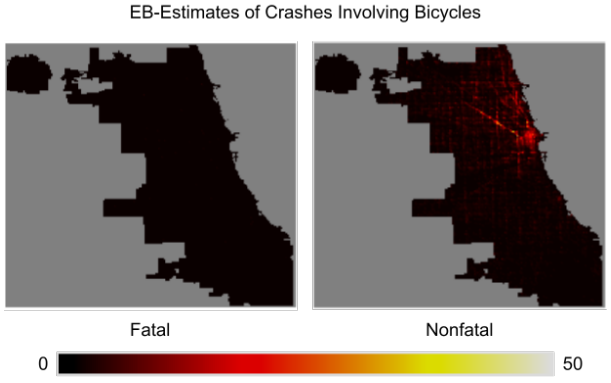

Fig. 9. EB-Estimates of Crashes Involving Bicycles.

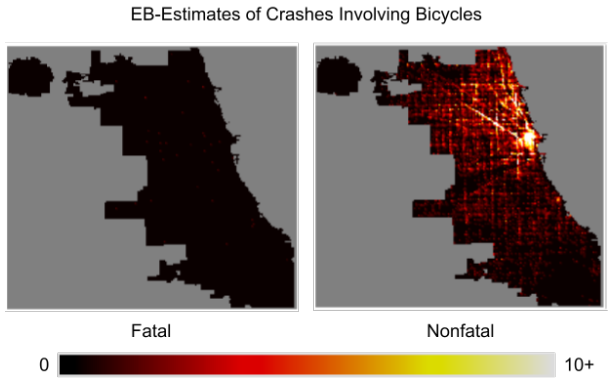

Fig. 10. EB-Estimates of Crashes Involving Bicycles with an Adjusted Legend.

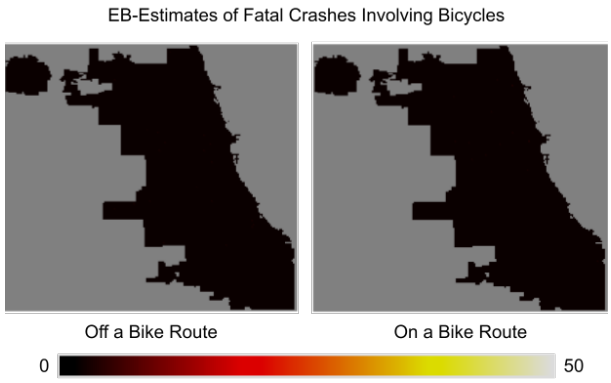

Fig. 11. EB-Estimates of Fatal Crashes Involving Bicycles.

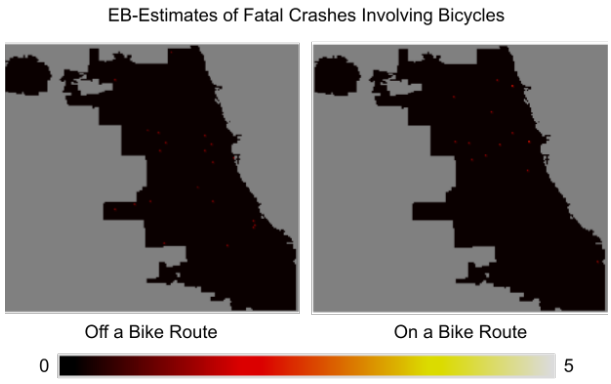

Fig. 12. EB-Estimates of Fatal Crashes Involving Bicycles with an Adjusted Legend.

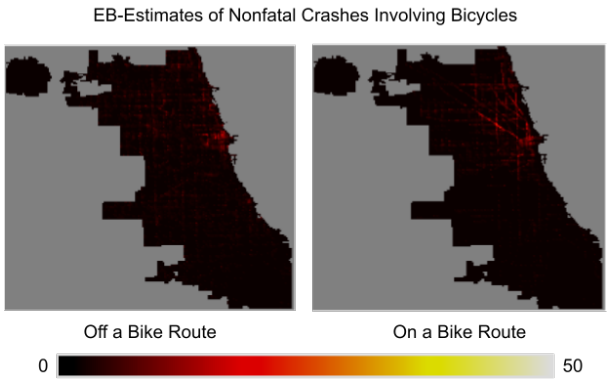

Fig. 13. EB-Estimates of Fatal Crashes Involving Bicycles.

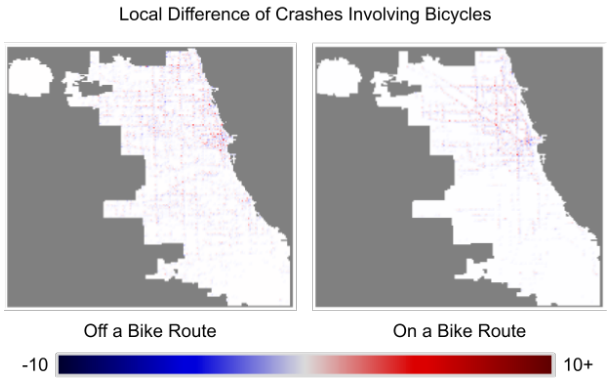

Fig. 14. Local Difference of Crashes Involving Bicycles.

tifying boundaries of roads. There are a few roads that are specifically outlined by two lines on either side of it while the road itself is left blank on the map. This is because this map shows where the prediction for crash counts was higher than the actual value. This means there are more crashes nearby, but the specific location tends to not have as many crashes. This is why it typically occurs near the edges of roads or close to actual hotspots. This map is not an effective map for identifying hotspots, but a byproduct of creating the positive

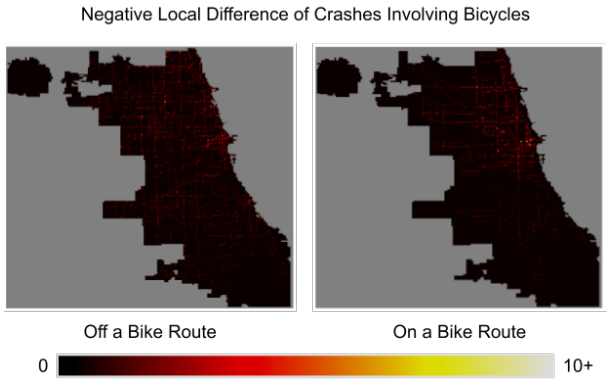

Fig. 15. Negative Local Difference of Crashes Involving Bicycles.

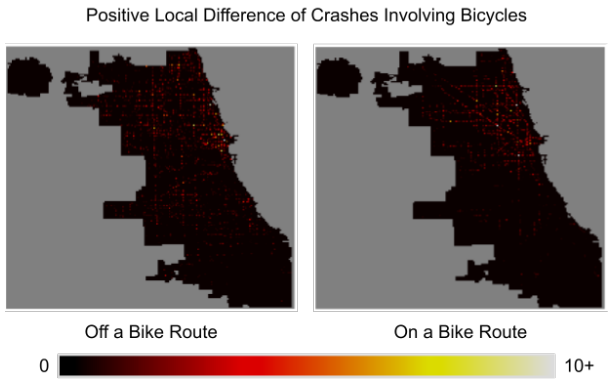

Fig. 16. Positive Local Difference of Crashes Involving Bicycles.

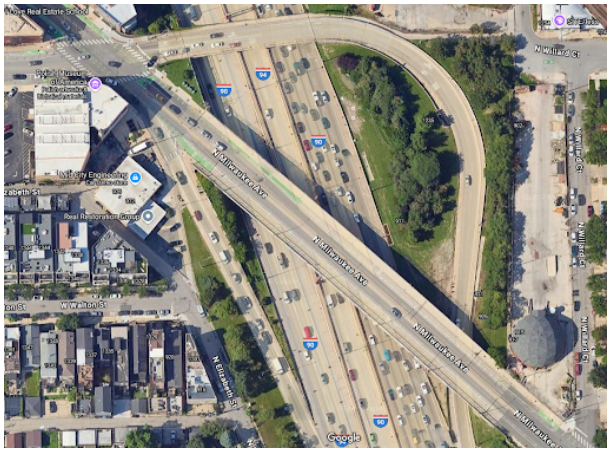

Fig. 17. Map of an intersection between N Milwaukee Ave and I-90 [4].

local difference map. If the positive local difference is the value after removing the bias from local traffic patterns, the negative local difference is the value of that removed bias.

When comparing the positive local difference map to the overall count and EB-Estimate maps, the local difference creates more individual points rather than large groups of high values. This shows that finding the local difference in this way was an effective strategy for singling out hotspots. While a point may not have the highest number of crashes across the city, points on this map had far more crashes than expected based on the point's surrounding neighbors. Many of these points appear to be at intersections, particularly on the map of points on a bike route. Having a place where cars cross a path bicycles will be traveling makes sense to create a spike in bicycle crashes.

There is one diagonal road that appears as a bright line across most maps. It contains several hotspot locations on a bike route in Fig. 16, but is clearly defined as the most bright line in Fig. 4, Fig. 8, and Fig. 9. This road is N Milwaukee Ave. When exploring this road, it was found that it intersects with interstate I-90 at multiple points along the road. Despite this, the road still contains a bike lane. Fig. 17 displays an image of this road where cars exit I-90 and are immediately tasked to cross a bike lane to enter the road. There are multiple points along this road that feature similar crossings. Not only this, the bike path also crosses a bridge across the interstate. Due to all of these factors and the high number of bicycle crashes along this road, it becomes clear that this bike lane should be re-evaluated for safety.

One other area of interest on the maps is the circular region in the northwest with seemingly no crashes. There is exactly one nonfatal crash that takes place off a bike route in this area. This region is almost entirely the Chicago O'Hare International Airport that is considered within city limits. There are no bike routes and likely few bicycle riders. While there may be more crashes of different types that happen at the airport, in the time that the dataset was collecting records, there was exactly one crash that involved a bicycle at this airport.

V. Conclusion

Overall, the map of positive local differences proved to be a very easy to read and useful method for locating hotspots of higher-than-normal crashes. This method and this dataset worked well together to provide valuable insights. When examining the locations of the hotspots, areas with safety issues that could be addressed were found. While the improved safety of bike routes could not be heavily supported by these results, the data also did not compare in any way to bicycle traffic that did not crash. Even given the limited data size, the severity ratio for crashes on a bike route was slightly lower than the severity ratio for crashes off a bike route. For the hotspots themselves, some expected patterns emerged. Downtown has a higher expected crash value than the suburbs. Intersections have a higher local difference than the roads surrounding them. However, the map was able to point out some locations with dangerous conditions that should be looked into. Further research into the traffic patterns at these locations could lead to changes to make these roads safer.

A. Future Work

Improvements could likely be made with the choice of prediction metric. EB-Estimate was a more advanced metric than using only the crash frequency like previous attempts at separating crashes by severity and vehicle type, but there are other formulas to try and compare to. Research can be done into several other parameters as well. The twenty meter buffer around a route could be revisited to see if another value would better represent the data. If smaller grid sizes than the 200 meter by 200 meter area are chosen, more specific hotspot locations could be discovered. The definition of neighboring locations would likely need to expand, however.

A limit of this project was the small size of fatal crashes involving a bicycle. Finding the local difference for fatal crashes alone was not a worthwhile map, as the hotspots were at most crashes. Perhaps with another city's data, a larger dataset could fill this need. Another method may be to use

the data from multiple cities to calculate severity ratios for a more comprehensive measure.

In addition to improving this experiment, repeating this method on data from other cities would provide new outlooks on finding roads that need adjustments. There may be other types of crashes worth exploring beyond bike routes and fatalities as well. For instance, rather than separating severity by fatal or non-fatal, some of the other attributes of the Chicago dataset include the presence of various types of injuries.

Many new avenues could be opened up by using a dataset for bicycle traffic patterns. In this research, assumptions could only be made using how severe a crash was and how many crashes there were. Using the local difference helped to offset the bias of high-traffic areas, but using a dataset containing bicycle trips that did not end in a crash would provide a more definitive answer. This data would also be useful for answering whether bicycle routes help prevent accidents.

REFERENCES

[1] Chicago Department of Transportation. (2024). Bike Routes in Chicago [Data set]. https://data.cityofchicago.org/Transportation/Bike-Routes/hvv9-38ut/

[2] Chicago Police Department. (2024). Traffic Crashes in Chicago - Crashes [Data set]. https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if/

[3] Chicago Police Department. (2024). Traffic Crashes in Chicago - Vehicles [Data set]. https://data.cityofchicago.org/Transportation/Traffic-Crashes-Vehicles/68nd-jvt3/

[4] Google Maps. (n.d.). [N Milwaukee Ave Intersection with I-90]. Retrieved November 18, 2024, from https://www.google.com/maps/@41.8991813,-87.6598125,204m/

[5] Lee, J., Yasmin, S., Eluru, N., Abdel-Aty, M., & Cai, Q. (2018). Analysis of crash proportion by vehicle type at Traffic Analysis Zone Level: A mixed fractional split multinomial logit modeling approach with spatial effects. Accident Analysis & Prevention, 111, 12–22. doi:10.1016/j.aap.2017.11.017

[6] Levy, Jonathan. (2024). City Boundary of Chicago [Data set]. https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-City/ewy2-6yfk

[7] Schallhorn, C., & Duis, P. R. (2024). Chicago. Encyclopædia Britannica. https://www.britannica.com/place/Chicago

[8] Yu, H., Liu, P., Chen, J., & Wang, H. (2014). Comparative analysis of the spatial analysis methods for hotspot identification. Accident Analysis & Prevention, 66, 80–88. doi:10.1016/j.aap.2014.01.017

[9] Ziakopoulos, A., & Yannis, G. (2020). A review of spatial approaches in road safety. Accident Analysis & Prevention, 135, 105323. doi:10.1016/j.aap.2019.105323