Can We Predict Red Light Violations Using Crime Data?

Biqi Lin
New York University
New York, United States
bl2379@nyu.edu

Juanlu Yu
New York University
New York, United States
jy2234@nyu.edu

***Abstract*—**

Reducing the number of traffic violations and crashes at road intersection is a challenging task for the government to make safer roadways for all drivers and pedestrians. Red Light Camera has contributed a significant reduction in red-light violations while it raises a question on the optimal location to install the camera. In this project, we try to predict the corners with high red light violations using comprehensive crime data in Chicago of each zip code. Also, we study the daily correlation of violations and crimes in Chicago. Our analytic work shows the moderate correlation between red light violations and crimes in Chicago of each zip code while the strong correlation between daily incidents of red light violations and crimes in Chicago.

Keywords—analytics, red light violations, crime, safety

I. Introduction

According to the Insurance Institute for Highway Safety (IIHS), in 2012, 683 people died, and about 133,000 people were injured due to red light violations[1]. Increasing road safety by reducing traffic violations and crashes at road intersections is necessary for the government.

Our project is to predict red light violations using crime data. In other words, we want to study the correlation between red light violations and crimes. No evidence showed that crimes have a direct influence on red light violations and vice versa. However, crimes and red light violations may share some common determinant factors: the occurrence of crimes was reported to be affected by economic growth, laws enforcement, climate, and population density; meanwhile, researchers have also extensively explored that various factors—including intersection, traffic feature, pavement and weather conditions[3], social environment, and driver[4]—may influence driver behavior at red light intersections. Therefore, we believe that there's a correlation between the number of crimes and the number of red light violations, and we may use the relationship for the prediction of red light violations based on crime data.

In our project, we focus on two datasets: the daily red light violations in the City of Chicago and the daily crime data in the City of Chicago. We built logical regression models to study the relationship between the number of red light violations and the number of crimes in Chicago in 2016.

II. Motivation

Chicago Police Department maintains a comprehensive database of incidents of crimes that occurred in the City of Chicago since 2001. It contains a vast amount of crime data that are easy to access.

On the other hand, some studies reported that the red light violations and the occurrence of crimes share some common determinant factors. For example, economic growth can decrease the rates of both. However, there is no report that describes the relation between red light violations and the crime rate conclusively.

A report described the red-light running dangers in the United States[7] that 719 people died in red-light running crashes each year and $390 million was lost in cost due to red-light running fatalities each month from 2011-2015. Worse still, total red-light running crash fatalities increased 7% from 2011 to 2015. The good news is that a study shows that installing red-light cameras can reduce red light crashes and thus improve roadway safety. The Federal Highway Administration (FHWA) evaluated the red-light camera programs and found that the intersection crashes effectively decreased by 25 percent after the enforcement of Red Light Camera program[7]. However, several cities are considering shutting down the least profitable red light cameras because sometimes cameras generate so little revenue to justify the cost of running it[6]. Therefore, finding the optimal location to build the red light cameras is important for maximizing the benefit. If a correlation between red light violations and crimes exists, comprehensive crime data in different areas can be a useful reference to help us locate the intersections that need more red light cameras.

The above considerations motivate us to study the potential correlation between red light violations and crimes.

III. Related Work

In a recent study, findings indicate that geospatial factors pertaining to demographic, socioeconomic, and land use characteristics by area type have a close relationship with crashes due to red light violations[4]. Also, the researchers found enforcing traffic laws in some areas would significantly reduce the number of crashes due to red light violations[3]. Another accident analysis on traffic explored the association between cyclists' red light violation and social influence. The result indicates that the social influence such as "more people waiting for the red light" will decrease the rate of risk-taking violations[5]. Our project and these works are all to study the correlation between red light violations and some conditions. To study how social factors influence red light violation, in Fraboni studies[4], they designed a controlled experiment on two groups of people. Our project is through analyzing two public data sets instead of two groups of people to draw the conclusion.

In order to study the relationship between red light violations and crimes in each zip code, we need to convert the (latitude, longitude) pairs in the original data sets to zip codes. To achieve this goal, we applied the Point-In-Polygon Algorithm[8]. The algorithm is designed to determine whether a point is inside a complex polygon: in our case, the algorithm helps us to figure out whether a (latitude, longitude) pair falls inside a zip code area.

IV. Design

We have altogether three diagrams to show the design of our project.

The first diagram [IV. 1] shows the progress that we used Map Reduce to do the data clean, and, then, we used Hive to select a subset of the clean data.

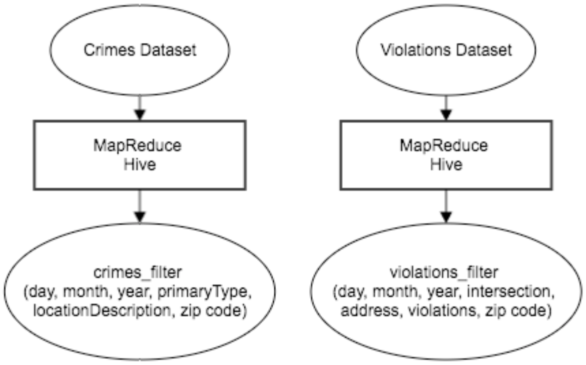
Crimes Dataset
Violations Dataset
MapReduce
Hive
MapReduce
Hive
crimes_filter
(day, month, year, primaryType, locationDescription, zip code)
violations_filter
(day, month, year, intersection, address, violations, zip code)

Figure IV. 1: Data clean diagram

The second diagram [IV. 2] shows how we analyzed the data based on the zip code. We calculated the total number of violations and the total number of cameras group by month, year, and zip code. Then, we got the average number of violations of each zip code and combined the average violations data with the sum crimes data. We used the result to find the relationship between violations and crimes based on the zip code.

The third diagram [IV. 3] shows how we analyzed the data based on the date. We calculated and combined two data sets to get a table. The table has these columns. We used the result to find the relationship between violations and crimes based on the date.

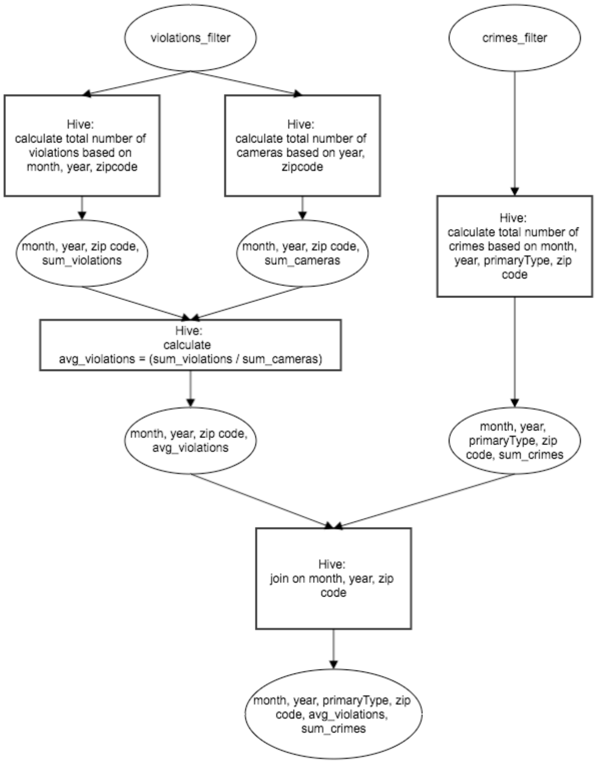
violations_filter
crimes_filter
Hive:
calculate total number of violations based on month, year, zipcode
Hive:
calculate total number of cameras based on year, zipcode
Hive:
calculate total number of crimes based on month, year, primaryType, zip code
month, year, zip code, sum_violations
month, year, zip code, sum_cameras
Hive:
calculate
avg_violations = (sum_violations / sum_cameras)
month, year, zip code, avg_violations
month, year, primaryType, zip code, sum_crimes
Hive:
join on month, year, zip code
month, year, primaryType, zip code, avg_violations, sum_crimes

Figure IV. 2: Diagram shows the workflow based on zip code

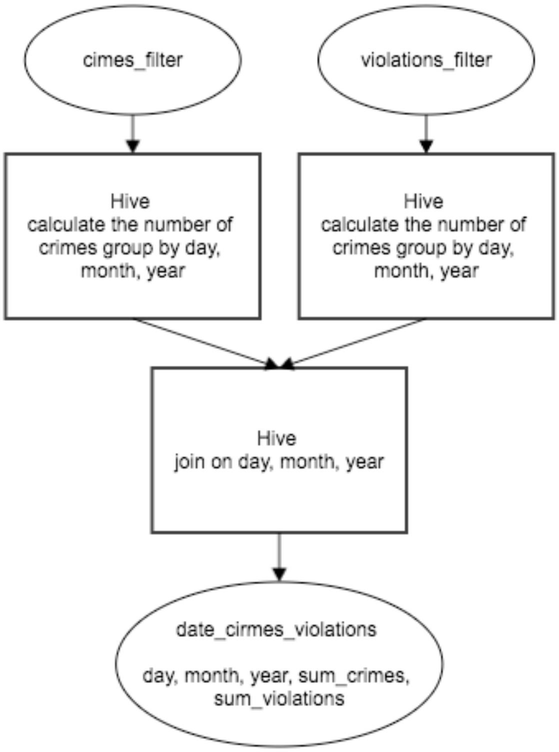
cimes_filter
violations_filter
Hive
calculate the number of crimes group by day, month, year
Hive
calculate the number of crimes group by day, month, year
Hive
join on day, month, year
date_cirmes_violations
day, month, year, sum_crimes, sum_violations

Figure IV. 3: Diagram shows the workflow based on date

V. Results

We investigated the correlation between red light violations and crimes in using two approaches.

In the first approach, we plotted the trends of the red light violations and crime occurrences between the date Jul/1/2014 and Jul/1/2017 as shown in Figure [V. 1]. We observed at the intuitive level that the trends of the two observables are clearly correlated. To obtain of quantitive understanding, we collected data of the form (x_t, y_t) where t is a date in the year of 2016, x and y are the numbers of instances of red line violations and crimes occurred on the date t respectively. Then, we performed a linear regression to this data set and found that the correlation coefficient is 0.733629, which indicates a strong correlation. This result confirms our belief on the relation between red light violations and the crime.

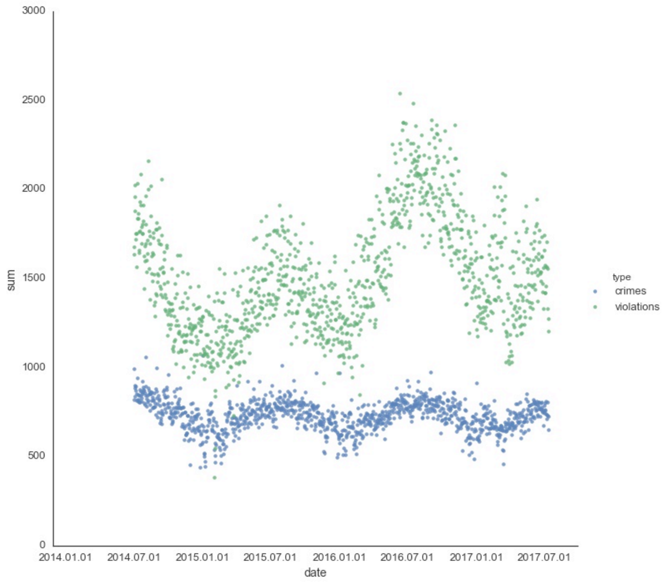

Figure V. 1: The trends of red light violations and crimes between Jul/1/2014 and Jul/1/2017

Moreover, we used the corresponding data in 2015 to test the accuracy of the predictive model we built. More precisely, using crime data in 2015 as input to the linear model we obtained above from linear regression, we calculated the predictive daily volumes of red light violations in 2015. The actual mean value of daily red light violations in 2015 is 1657.065, and the expected value is 1656.012. The Normalized Root Mean Square Error (NRMSE) between the actual value and the predictive value is 0.13, which indicates the accuracy of our prediction.

In the second approach, we first collected data of the form (x_z, y_z) where z represents an area in Chicago of each zip code, x is the number of red light violations, and y is the number of crimes occurred in the zip code z in the year of 2016. Since there are only 47 zip codes in our data, we expanded the (x_z, y_z) pairs from 47 to 564 (47×12=564) by replacing the pairs of the same zip code by the pairs of the same zip code and the same month. We calculated the Pearson coefficient of correlation of the crimes and violations based on these 564 pairs; however, we didn't get the result we expected: the coefficient is only 0.0993121. We then thought that the type of the crimes may also have some effects on the result. Therefore, we calculated the coefficients of each type of crime and violations in each zip code. Among all 31 types of crimes, deceptive practice and theft are the two most related types with the coefficients being 0.4304357 and 0.4084930.

We tried to find out the reason behind the discrepancy between the correlation we computed using this two approaches. Although no conclusive answer has been reached, we believe that the following factor plays a substantial role: the accuracy of the crime data grouped by zip code is higher than the accuracy of the red light violations data grouped by zip code. Ideally, every crime will be documented regardless of the zip code; however, not every red light violation is recorded due to the absence of certain red-light cameras. This fact may introduce noise when determining the correlation of the two observables. We calculated the correlation coefficient between the number of cameras and red light violations in Chicago of each zip code. The value is 0.647, which indicates a strong correlation between the collections of red light violations and the number of cameras. Therefore we believed that our first approach is more accurate.

VI. Future Work

We summed up three possible respects to further research into this project.

First, our datasets were collected from Chicago area. To draw a more solid conclusion, we can collect the crime and red light violation data across the whole country.

Furthermore, we didn't reach a conclusive answer of the correlation between red light violations and the number of crimes in Chicago of each zip code due to the limitation of the red light violation dataset. Crime data are usually more comprehensive because people would report it to the police once it happens while red light violations wouldn't be recorded if there is no camera in the area. In order to overcome this issue, we can analyze the red light running crash at the signalized intersection to avoid the limitation of data collection, since every crash would be recorded in details by police.

Finally, from Figure [V. 1], we can observe that both violations and crimes follow the pattern that, during the year, the number of violations and crimes goes up and down. The highest point of each year locates on one of the hottest days in summer. Since we have already learned the fact that the temperature can influence the occurrence of crimes[2]: warmer weather leads to higher crime rate and cooler weather leads to lower crime rate. Maybe the number of violations is also influenced by the temperature. More efforts are needed to verify this assumption.

VII. Conclusion

In this paper, we analyzed the correlation between the red light violation and crimes based on data collected in Chicago of 2016. Our methods include comparing the tendency graph as well as linear regression analysis. We also used the data in

2015 to validate our model. Our result of the first approach supports our hypothesis that the two observables are strongly correlated.

We believe that our finding can be a reference for designing the camera system at road crossings since the crime data are comprehensive and available. However, we have to point out that our analysis is not strong enough to draw a firm conclusion due to the scale of the data we used and the fact that no advanced data analysis techniques were applied. We expect to improve the analysis in future work.

Acknowledgment

Thank NYU HPC for the support.

References

[1] Traffic Safety Facts 2012, National Center for Statistics and Analysis US Department of Transportation, Washington, DC (2014), p. 812032

[2] Anderson CA, Anderson DC. Ambient temperature and violent crime: Test of the linear and curvilinear hypotheses. J Pers Soc Psychol. 1984;46:91–7.

[3] H. Li, H. Rakha, I. El-Shawarby. Designing yellow intervals for rainy and wet roadway conditions.Int. J. Transp. Sci. Technol. (2012) 171-190

[4] Fraboni et al. Social Influence and Different Types of Red-Light Behaviors among Cyclists. Front Psychol. (2017).

[5] Srinivas et al. Due to Red Light Violation: Locations and Geospatial Influential Factors. (2017) 17-05051.

[6] Red Light Safety Program: City of Dallas.(2009 Oct.9). [online] Available at:

http://www.dalascityhall.com/pubsafe/safelight.html [Accessed 8 Jul. 2017].

[7] Fact Sheets American Traffic Solutions Smarter, Safer Transportation. (2017). [online] Available at:

http://www.easybib.com/reference/guide/apa/website [Accessed 8 Jul. 2017].

[8] Alienryderflex.com. (2017). Determining Whether A Point Is Inside A Complex Polygon. [online] Available at:

http://alienryderflex.com/polygon/ [Accessed 20 Jul. 2017].