Interagency Work Zone Traffic Data Modeling and Analysis for New York City

Collier, John | Han, Seunggyun

Jaber, Linda | Singh, Akhil Kumar

New York University | Center for Urban Science and Progress

CUSP-GX 5006: Urban Science Intensive II

Mentor: Prof. Kaan Ozbay

Sponsor: HDR

Final Report

July 20, 2020

Abstract

Road construction events are a necessary part of keeping road infrastructure in good condition but can pose significant safety problems when implemented. Transportation authorities in NYC seek a better understanding of the type, severity, and extent of mobility impacts associated with work zones. This project proposes using a k-means clustering approach to predict the probability of a vehicle collision occurring in the proximity of a road construction event (i.e work zone). The proposed clustering method is applied to over 20,000 construction and emergency construction events of relatively short duration in New York City to identify types of work zones that may present greater safety risks. This methodology builds upon the existing body of research by utilizing only publicly available datasets and by applying the methodology to roads and highways in the five boroughs of New York City. The results of this project are in service of enabling practitioners to employ appropriate mitigation strategies during project programming, design, and in the development of effective transportation management plans. This project is part of the Center for Urban Science and Progress capstone process with capstone sponsor HDR and a consortium of transportation authorities in the New York City Metro Area.

Keywords: Work zones, vehicle collisions, clustering, collision probability.

Introduction

Temporary work zones for roadway constructions have the potential to significantly impact mobility and safety for all roadway users. An increase in the number of people using local streets because of work zone diversion plans may increase the likelihood of crashes, including crashes involving vulnerable populations (e.g., cyclists, seniors, and individuals with disabilities). This project brings together disparate data sets to allow for a more comprehensive overview of collisions and work zones. Our aim is to help agencies optimize construction schedules, improve safety and mobility, and adjust travel diversion plans when schedule overlap is unavoidable. We asked, **can we use characteristics of streets and roadway construction events to predict collision probability in future work zones?**

The first task of this project was to produce a pipeline for wrangling disparate spatial data sets – street networks, collisions, and construction events – that allowed for a high-resolution and accurate mapping of the different geometries (Figure 10). Since the data sets are created from different agencies with different formats and scales, it was essential to join these data sets before modeling any analytical algorithms. The outcome was a merged data set and a dashboard of crashes and construction events mapped to the unified street network. With a final data set of more than 20,000 short-term work zones in New York City, we moved to the second task of implementing a clustering algorithm to group similar work zones into cohorts based on their characteristics. A crash probability was calculated for each cluster. These clusters were used to predict the likelihood of a crash happening in a future work zone with similar characteristics and produce a predictive web-based application for future work zones.

Literature Review

According to (Hébert, et al., 2019) the emergence of the concept of public data and the advancement in big data analytics have resulted in a surge in traffic accidents research. Machine learning has played a significant role in recent industry research to derive insights into the probabilities of a vehicle collisions and the characteristics of roadways that cause them. In the public sector, the American Association of State Highway and Transportation Officials published a predictive model to estimate the effect of applying geometric design and traffic control features on the average crash frequency (AASHTO, 2010; 2014). Crash modification factors (CMFs) – site-derived multiplicative factors used to compute the expected number of crashes after implementing a given counter measure – are used to adjust the model's result.

(Hébert, et al., 2019) used machine learning algorithms to infer the causes and predict the probabilities of hourly crashes on a specific road segment in the city of Montreal. They merged the Montreal vehicle collision data set, historical climate data set, and national road network data set and implemented the Balanced Random Forest algorithm. The model detected 85% of crashes, and identified temperature, visibility, the hour, the day, and historic crash count to be the most important predictors.

More specific to this project, extensive research has been done on the increased risk of vehicle collisions in the presence of road construction. (Yang, et al., 2014) modeled highway crash risk for relatively short duration work zones and found that unlike long-term work zones that have many crashes, work zones with relatively short durations usually have no crash or one crash during the entire active work period. Yang el al. also found a rare-event logistic regression to be the most appropriate model to predict the likelihood of a crash in a short-term work zone. For this study, we adopted a similar approach of analyzing short term work events.

Our methodology also builds upon previous research that has utilized clustering to identify different cohorts of work zones by applying it to construction events in New York City. (Tay, et al., 2018) used large amounts of historical data about work zones to predict the likelihood of a crash in a future work zones with similar features. Researchers used characteristics about historical work zones such as time of work, construction type, and work zone duration to cluster work zones into cohorts of similar characteristics. Within these cohorts, researchers calculated collision rates so that the probability of a crash in a future work zone could be calculated given its cluster assignment.

Research involving work zones, however, is prone to a wide array of errors. (Ozbay, et al., 2013) writes about the lack of work zone data and how data derived from existing sources are usually subject to several uncertainties and measurement error. For example, research using work zone duration as a feature in its analysis is likely to have significant errors since the exact starting or end date of a specific work zone may not be readily available and most transportation authorities do not report on the times where actual work is being completed on the site. Similarly, work zone length is subject to significant error. Work zone lengths can vary according to the progress of the project and the length information for each stage is not usually available from project plans. These measurement errors – that we are cognizant of while designing our research methodology – point to transportation authorities' need to improve data collection around work zones and work zone collisions.

Data

Data describing work zones (WZs) and collisions were collected from the New York State Department of Transportation, and the New York City Police Department, respectively. Data from these two sources were merged with the LION data set from the Department of City Planning (DCP). The study focuses on WZs with a duration less than 24 hours, which account for more than 80% of all the WZs observed during 2016-2019. The data set constructed for this study includes 20,717 WZs (Figure 1) and 3,727 WZ collisions. In most cases, a WZ occurred on a road with three lanes (Figure 2). The average duration of a work zone is five hours (Figure 3). Around 3.8% of the construction work was done during the peak-periods (7-9 AM and 4-6 PM). Moreover, about 33% of the work was done in daylight (Figure 3), and WZs were typically on weekdays (Figure 4).

The observed 3,601 WZ collisions occurred on a small subset (14.5%) of all the recorded WZs. Among the 3,005 WZs with collisions, 81.5% witnessed one collision and 96% witnessed two or less collisions (Figure 5). The highest number of WZ collisions took place on Fridays, while far fewer occurred on Saturdays and Sundays, which could be explained by the lower number of WZs active on weekends (Figure 4). The seasonal distribution of WZ collisions implies that most of the collisions took place in the summer which is not surprising if compared the WZ activity per season (Figure 4). Following this exploratory data analysis, twelve attributes were selected as possible input variables (Table 1) based on their ability to influence the occurrence of a collision. The correlation between a subset of these variables and the number of WZ collisions is provided in (Figure 6), and indicates relatively weak correlation between attributes.

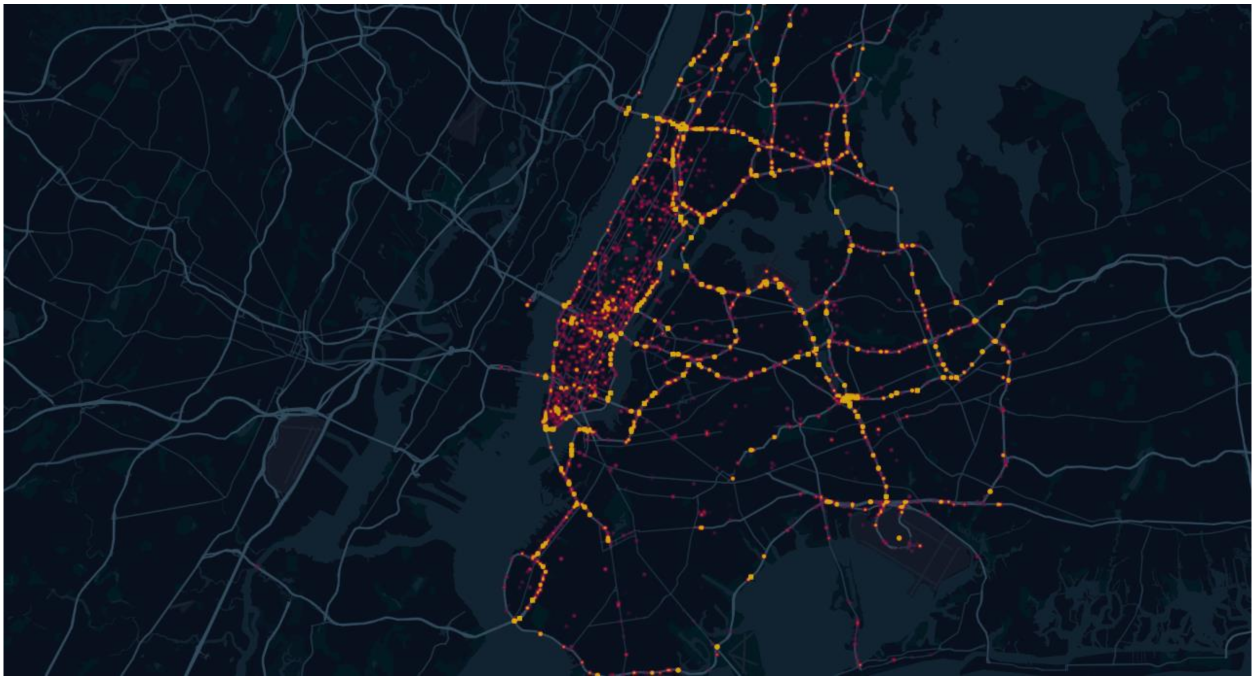


Figure 1: Heatmap showing locations of work zones that took place in New York City during 2016-2019. Higher density of work zones can be seen in yellow. We notice a higher number of work zones in Manhattan. We also notice a concentration of work zones on highway roads.

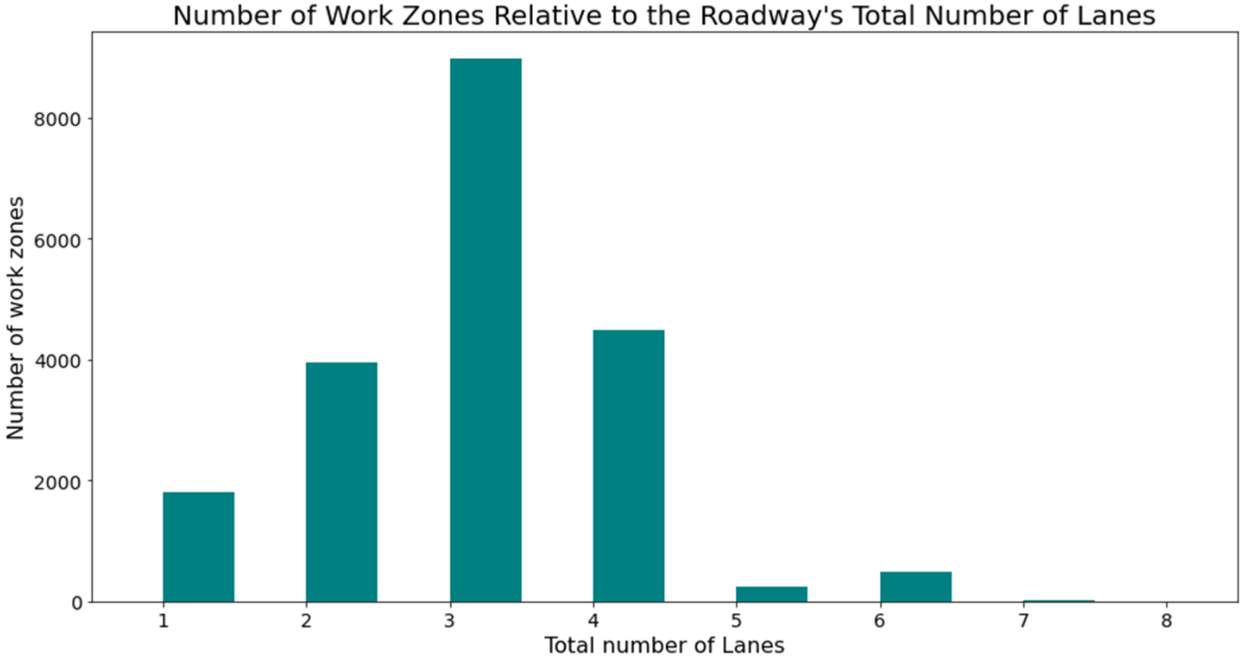


Figure 2: A bar plot of number of lanes of a roadway on which construction occurred shows that more than 8000 work zones occurred on a roadway with three lanes.

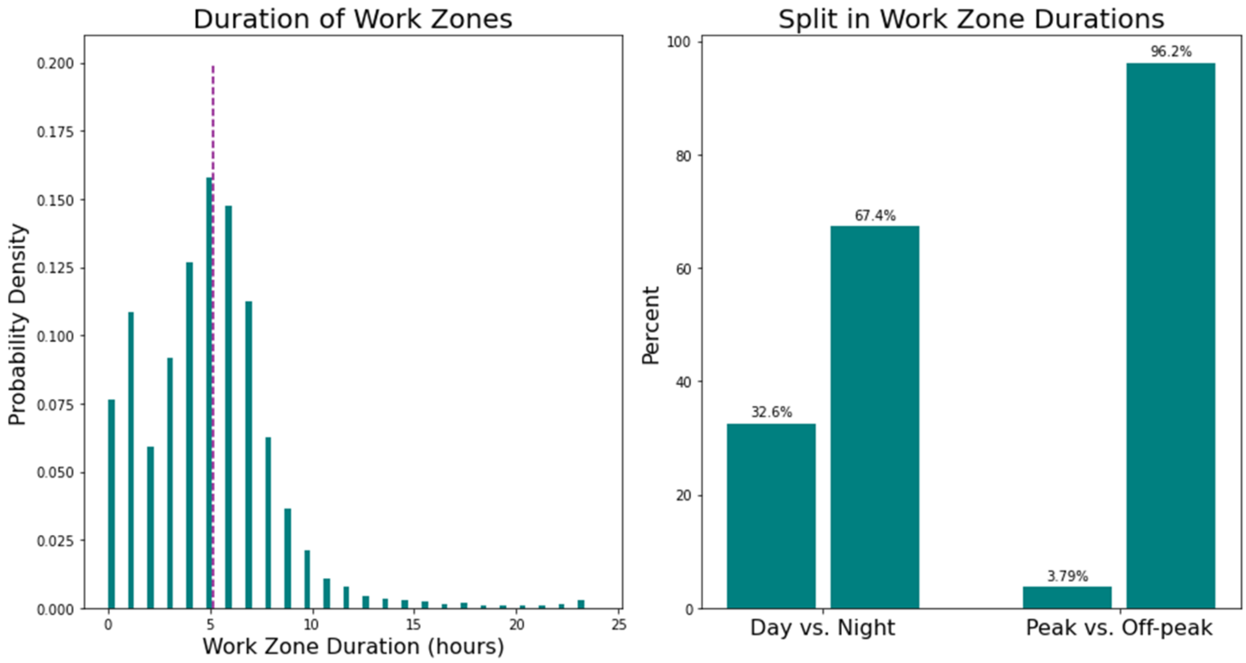


Figure 3: The distribution of WZ duration in hours (left) and the split in active hours between day/night and peak/off-peak. The figures show an average total WZ duration of 5hrs. 67% of the work is done at night and most of the work (96%) is done during off peak hours. Note: peak hours (7-9 AM and 4-6 PM).

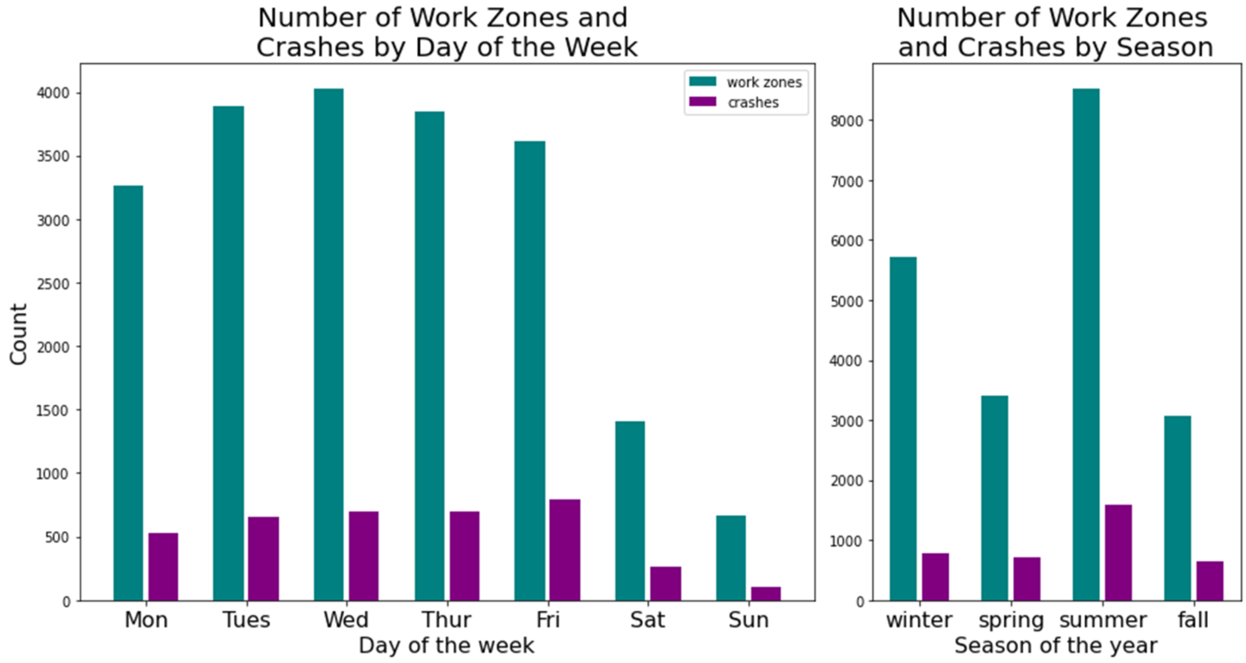


Figure 4: The number of WZs and WZ collisions per days of the week (left) and seasons of the year (right). The majority of WZs and collisions happen on a weekday and during the summer.

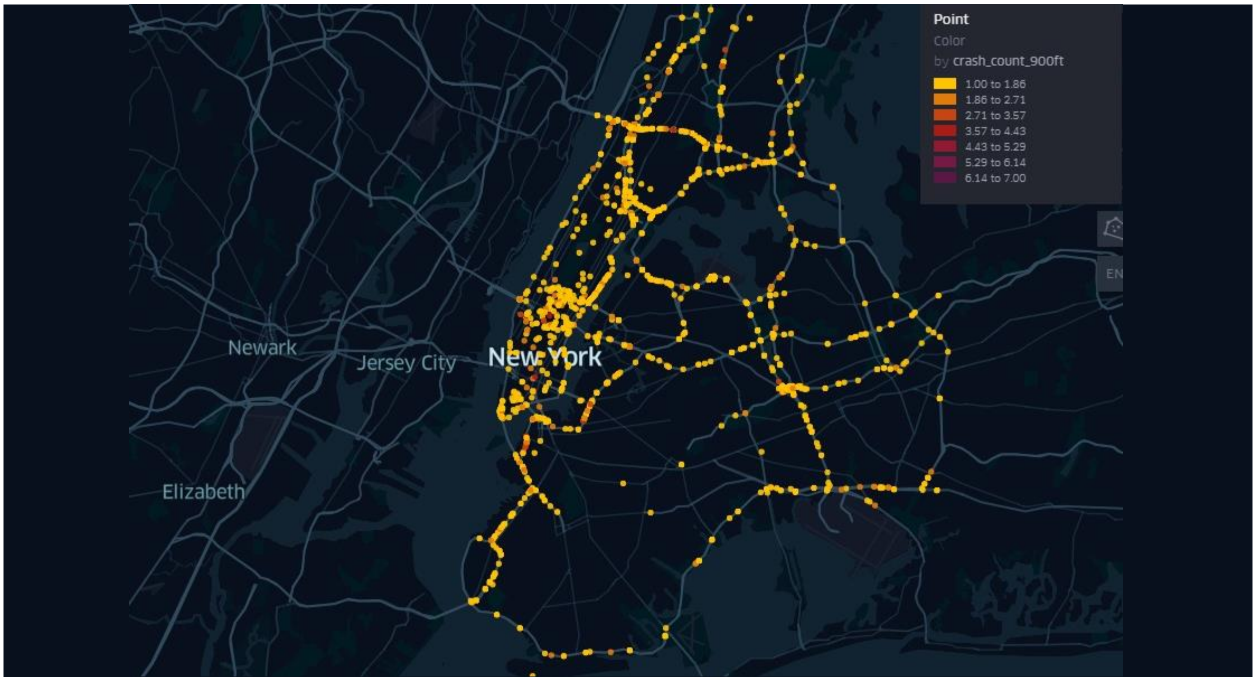


Figure 5: Map showing location of WZs with collisions. We notice the prevalence of the yellow color representing a low crash count (<2). Very few locations show in orange and red indicating a low number of work zones with more than two collisions.

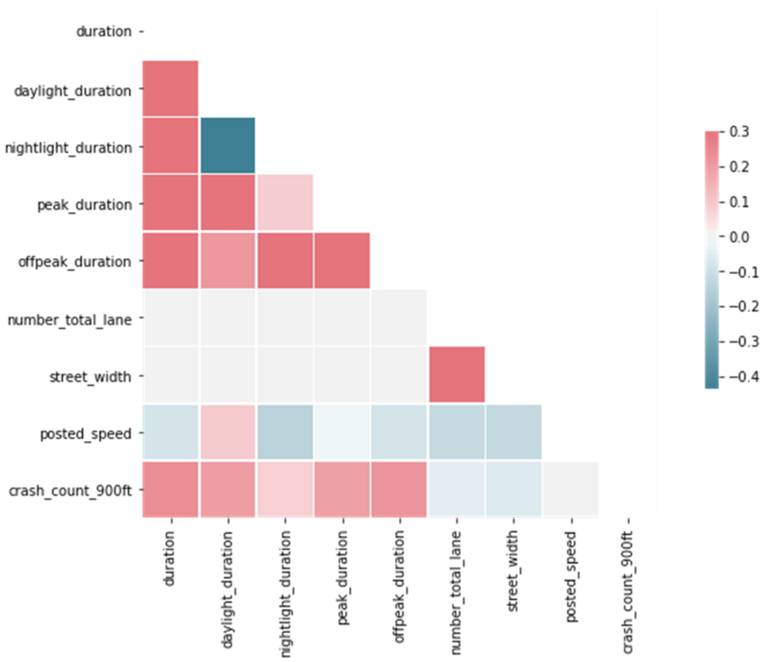


Figure 6: Correlation between input variables considered for clustering (table: 1) and number of WZ collisions. The darker red and blue colors denote the more positively and negatively the attributes are correlated, respectively. White color indicates no correlation.

Methodology

Spatial Data Wrangling

SharedStreets Referencing System

Spatial data produced by various entities use different geometric representations (Figure 11). The LION data provided a detailed database of the city's streets, but the structure of the streets was not ideal for joining disparate data sets since segment lengths did not correspond with real world block lengths (Figure 12). Additionally, the locations of crashes and WZs did not align perfectly with the street network (Figure 13). SharedStreets – a tool that creates a shared reference system for disparate street networks – provided a solution to connect the data (Figure 14). Each record in the LION, collisions and WZ data was assigned a 'SharedStreets geometry id'. Based on this id, characteristics of streets, collisions, and work zones were linked.

In some cases where one SharedStreets segment was matched to multiple LION segments, the latter were aggregated by taking the most frequent value of an attribute (Figure 15). Intersections between two adjacent roadbeds were represented as a short segment line instead of an intersection point (Figure 16). Consequently, crashes happening at those intersections were erroneously mapped as segment crashes. To distinguish between intersection and segment crashes, segments less than 100ft separating roadbeds were identified and replaced by node points from the LION dataset. This separation of segment and intersection crashes was important for better future WZ analysis because these two types of crashes represented different safety scenarios.

Roadway Construction Events / Work Zones

Without access to a database of historical work zones events, two publicly available data sets on roadway construction events were investigated, the "511NY Traffic Events" and the

"Street Closures due to Construction". Each record in the 511 data provided information on the location (latitude / longitude) and active work duration of a roadway construction event. The final dataset had missing values for posted speed, street width, and number of lanes. Instead of dropping these records, missing values were imputed using the mode value for segments of the same roadway type. To determine which crashes are considered in proximity to a WZ, a buffer was created around this segment (Figure 17). The average WZ length (~900ft) was determined as a suitable distance for the buffer radius. Therefore, collisions that happened within a radius of 900ft from the segment were counted as WZ collisions.

Previous research has shown the importance of work zone length in predicting collisions (Ozturk, et al., 2015). However, the 511 data did not provide any information on the length of the WZs, so we had to analyze the Street Closures data for that. Each record in the Street Closures data provided information on a LION segment closure event due to different construction purposes, yet it lacked information on the exact active working hours on this segment. Instead, it only provided data on the overall duration of a work permit. Processing the Street Closures data led to the grouping of multiple segment closures spatially and temporally to create WZs. We then attempted to join the Street Closures WZs we created which provided length information to the 511 data which provided accurate active work durations. The process was able to successfully join 263[2] WZs only.

Work Zone Clustering & Collision Probability Prediction

This objective of the clustering process was to create clusters based on the engineered historical work zones data set to predict the collision probability for future ones (Figure 7). To improve the performance of the clustering algorithm, categorical variables were modeled using their binary counterparts, and the peak and day duration of a WZ were divided by its overall duration to get the proportion of the WZ that took place during peak and daytime. The off-peak and night durations were removed from the model for redundancy to reach a final list of input variables for clustering (Table 2). Data was normalized and split into train and test sets. The train set was used to fit a K-means clustering algorithm, first, on the larger data set that did not include the length attribute, and second, on a smaller data set that included length attribute.

The collision probability was then calculated by dividing the number of WZs witnessing a crash by the total number of WZs for each cluster. Accordingly, each WZ in the test set was assigned to the closest cluster (from the train set) based on its features to compute its predicted probability of a collision. To evaluate the model, the actual collision probabilities for each cluster in the test set were then calculated. The symmetric mean absolute percentage error (SMAPE) was used to quantify the error between the actual and predicted probabilities across clusters. A smaller SMPAE value would mean a better prediction power. To measure the prediction power a base line SMAPE value was calculated for clustering using the roadway type attribute only. The assumption was that clustering using only one attribute would have the least prediction power, and any increase in the prediction power (i.e. decrease in SMAPE) later would be considered an enhancement to the model.

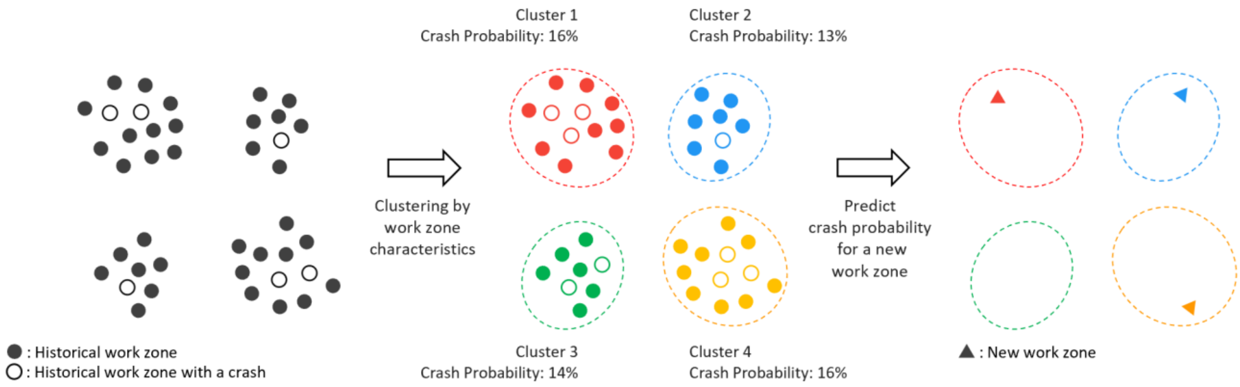


Figure 7: Graphical description of the clustering methodology

Results

Clustering & Prediction

The silhouette analysis was used to choose an optimal value for the number of clusters. The silhouette plot (Figure 8) shows that the number of clusters of value of 3, 5 , 6 and 7 are a bad pick due to the presence of clusters with below average silhouette scores. Silhouette analysis is more ambivalent in deciding between 2 and 4. Four clusters were eventually chosen since the average silhouette score was higher. While some clusters had similar mode values of an attribute, most of them show variant values (Table 3). The spatial distribution of the different clusters shows that cluster 2 of work zones seem to be occurring on secondary and tertiary streets while cluster 1, 3, 4 occurred on primary streets (Figure 9).

The predicted probability (using the train set) of a crash happening in each of the four clusters is 16%, 13%, 14%, and 16% respectively. These probabilities were compared to the actual probabilities of each cluster in the test set and SMAPE was measured at 3.8%. If compared to the base error calculated using roadway type only it showed a 2.6 % decrease. On the other hand, clustering using the smaller data set (i.e. with length) resulted in an almost zero percent error compared to a 29% base error using the roadway type only.

Dashboard & Web Application

An interactive dashboard[3] was designed to serve as a tool for transportation authorities to explore the public safety risks associated with historical work events as well as the historical crash rates on all New York City roads and intersections (Figure 18). A predictive web application[4] was also built off the results of the clustering methodology to inform the planning of multiple new construction events simultaneously. It would aid decision making to avoid situations with high crash risk.

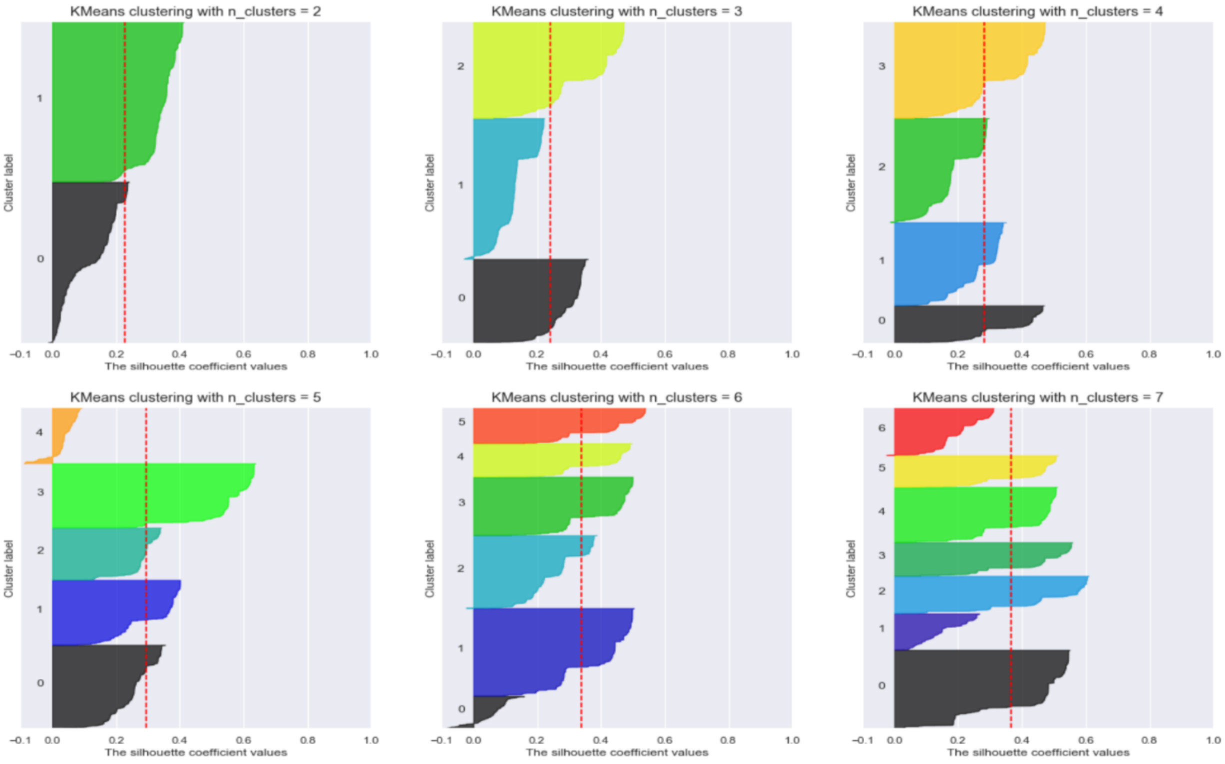


Figure 8: A silhouette plot for K-means clustering used to choose an optimal value for the number of clusters. A higher score means that the sample is well matched to its own cluster.

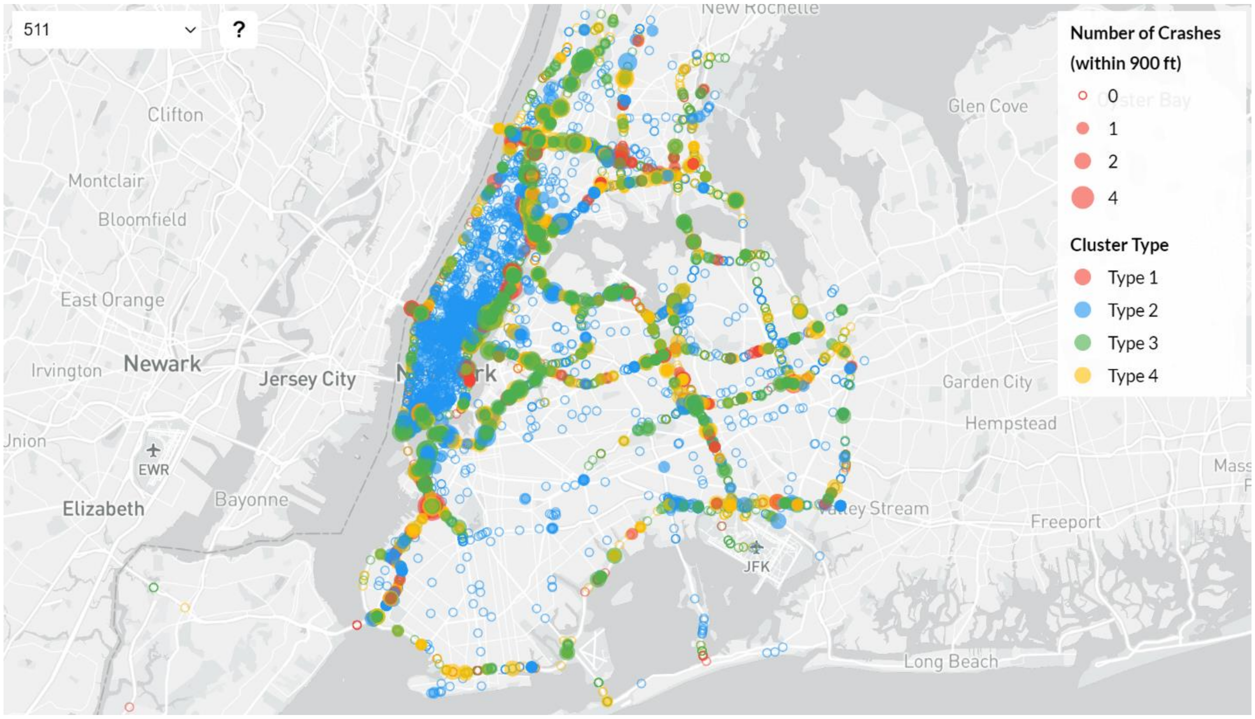


Figure 9:Spatial Distribution of the four clusters.

Conclusion

This research intended to provide transportation agencies with an understanding of work zones crash risk by proposing a clustering approach to predict the probability of a collision occurring in the proximity of a roadway construction event. This research adds to similar existing body of literature that used clustering to predict work zones' effects on public safety. One major improvement to the methodology developed by (Tay, et al., 2018) was deriving work zone length by using available permit data. Additionally, unlike past research that focused on only highways, this research was expanded to include both highway and local roads throughout NYC.

The results of the k-means clustering algorithm provided 4 distinct clusters of work zones with varying probabilities of a crash occurring. The final clustering model was built into a predictive web tool that allows for the application of the findings to future or hypothetical work zones and serves as a proof of concept for this methodology. The predictive power of these cluster assignments could be improved with further refinement of parameter weights that were not implemented in the scope of this project.

The process of this research created a pipeline of work zones' characteristics and collisions data resulted in several improvements to wrangling disparate datasets together for future use by transportation authorities. The results of this data wrangling are also available for exploration through an interactive dashboard. It also highlighted some pitfalls in the data collection process around construction work zones in New York City. To gain better insights, it would be useful to have the length of a work zone as well as the hours in which work was taking place publicly available. With access to improved datasets, the methodology used in this research could provide more accurate results.

References

AASHTO. (2010; 2014). Highway Safety Manual (1st Edition) with supplement 2014. American Association of State Highway and Transportation Officials (AASHTO). Retrieved from https://app.knovel.com/hotlink/toc/id:kpHSM00002/highway-safety-manual/highway-safety-manual

Gross, F., Persaud, B., & Lyon, C. (2010). A Guide to developing Quality Crash Modification Factors.

Hébert, A., Guédon, T., Glatard, T., Jaumard, B. (2019). High-Resolution Road Vehicle Collision Prediction for the City of Montreal.

Ozbay, K., Yang, H., Ozturk, O., & Yildirimoglu, M. (2013). Modeling Work Zone Crash Frequency by Quantifying Measurement Errors in Work Zone Length. Accident Analysis & Prevention, 55(192-201).

Ozturk, O., Yang, H., Ozbay, K., & Xie, K. (2015). Work Zone Safety Analysis and Modeling: A State-of-the-art Review. Traffic Injury Prevention, 16(4), 387-396.

Tay, R. S., Sekuła, P., Vander Laan, Z., Farokhi Sadabadi, K., Skibniewski, M. J., & . (2018). Predicting Work Zone Collision Probabilities via Clustering: Application in Optimal Deployment of Highway Response Teams. Journal of Advanced Transportation.

Yang, H., Ozbay, K., Xie, K, Bartin, B. (2014). Modeling Crash Risk of Highway Work Zones with Relatively Short Durations. Journal of Transportation Research.

Appendix

Footnotes

[1] In addition to this, 900 ft also represent the width of the standard Manhattan block.

[2] Many reasons could be behind this low number. Adding to the fact that some 511 events might not require a permit, some also might not require full closure and therefore will not be listed – according to the Street Closures data description the data identifies locations where a street is subject to a full closure. Besides, the process of merging and creating WZs from the Street Closures and eventually linking those to the 511 data is prone to error and may have influenced the result.

[3] https://workzone-collision-analysis.github.io/capstone/dashboard/

[4] https://workzone-collision-predict.herokuapp.com/

Tables

Table 1:Input variables considered for WZ clustering

Season of the year when the WZ was active	Roadway type
Day of the week when the WZ was opened	Street width
Duration of the WZ during the peak hours	Total number of lanes
Duration of the WZ during off-peak hours	Number of travel lanes
Duration of the WZ during daylight	Number of parking lanes
Duration of the WZ during night	Speed

Table 2: Input variables used for WZ clustering

Variable	Description	Columns
Season	Season of year	4
Roadway type	Type of road	5
Weekend	1 for Saturday or Sunday, 0 otherwise	2
Street Width	Maximum width of the street	1
Peak	Peak duration / overall duration	1
Day	Day duration / overall duration	1
Speed	Roadway posted speed	1
Length	Total length of segments in feet	1

Table 3: Mode values of attributes in each cluster

Attribute	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Season	Fall	Summer	Winter	Summer
Roadway type	Highway	Street	Highway	Highway
Weekend	Most work zones occurred on weekdays			
Street Width	30	25	40	40
Peak	Off-Peak	Off-peak	Off-peak	Off-Peak
Day	Night	Night	Night	Night
Speed	50	25	50	50
Collision Probability	16%	13%	14%	16%

Figures

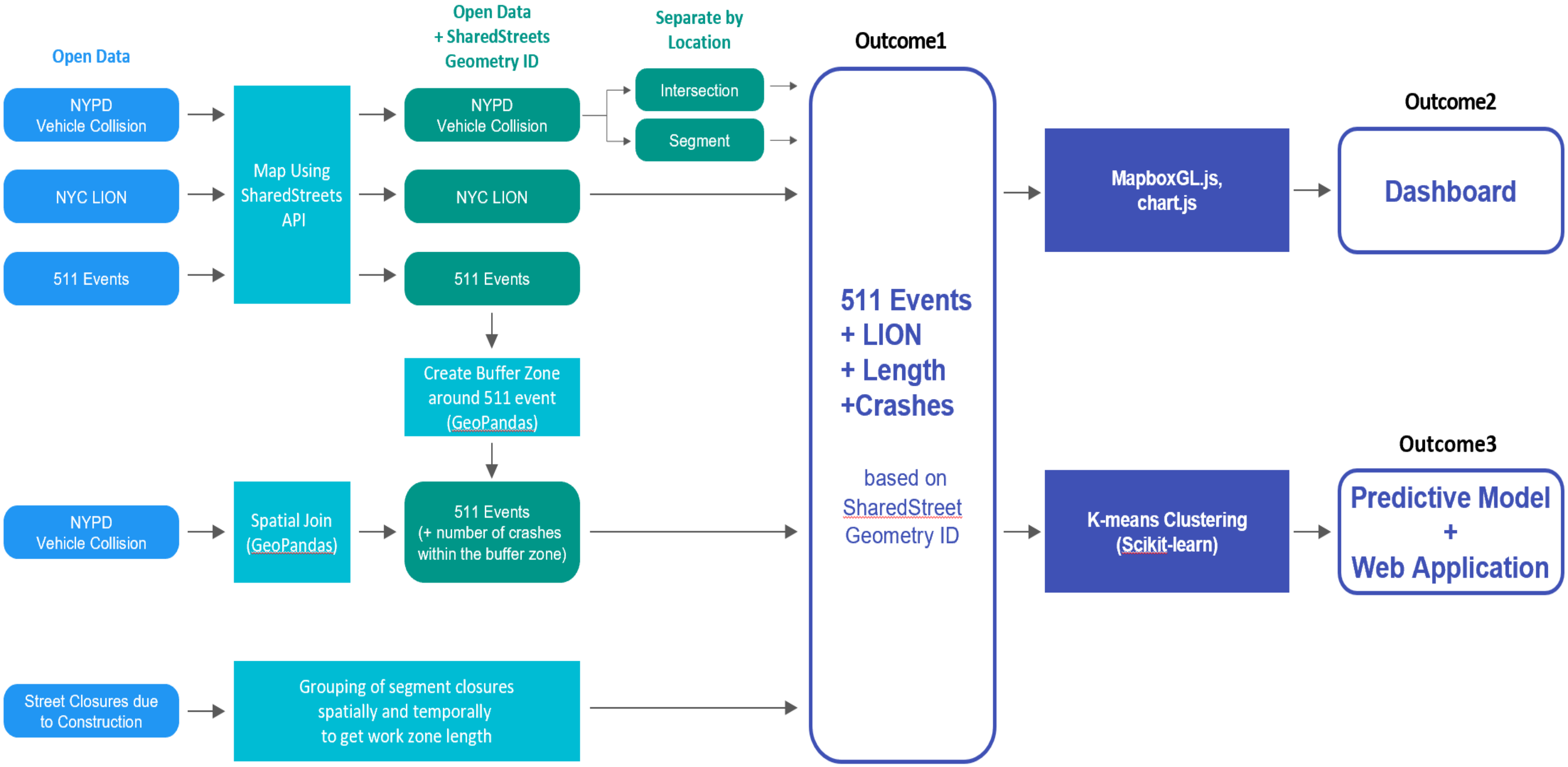

Open Data
NYPD Vehicle Collision
NYC LION
511 Events
Map Using SharedStreets API
Open Data + SharedStreets Geometry ID
NYPD Vehicle Collision
NYC LION
511 Events
Separate by Location
Intersection
Segment
Create Buffer Zone around 511 event (GeoPandas)
NYPD Vehicle Collision
Spatial Join (GeoPandas)
511 Events (+ number of crashes within the buffer zone)
Street Closures due to Construction
Grouping of segment closures spatially and temporally to get work zone length
Outcome1
511 Events + LION + Length +Crashes
based on SharedStreet Geometry ID
MapboxGL.js, chart.js
Outcome2
Dashboard
K-means Clustering (Scikit-learn)
Outcome3
Predictive Model + Web Application

Figure 10: Flow Chart of Tasks

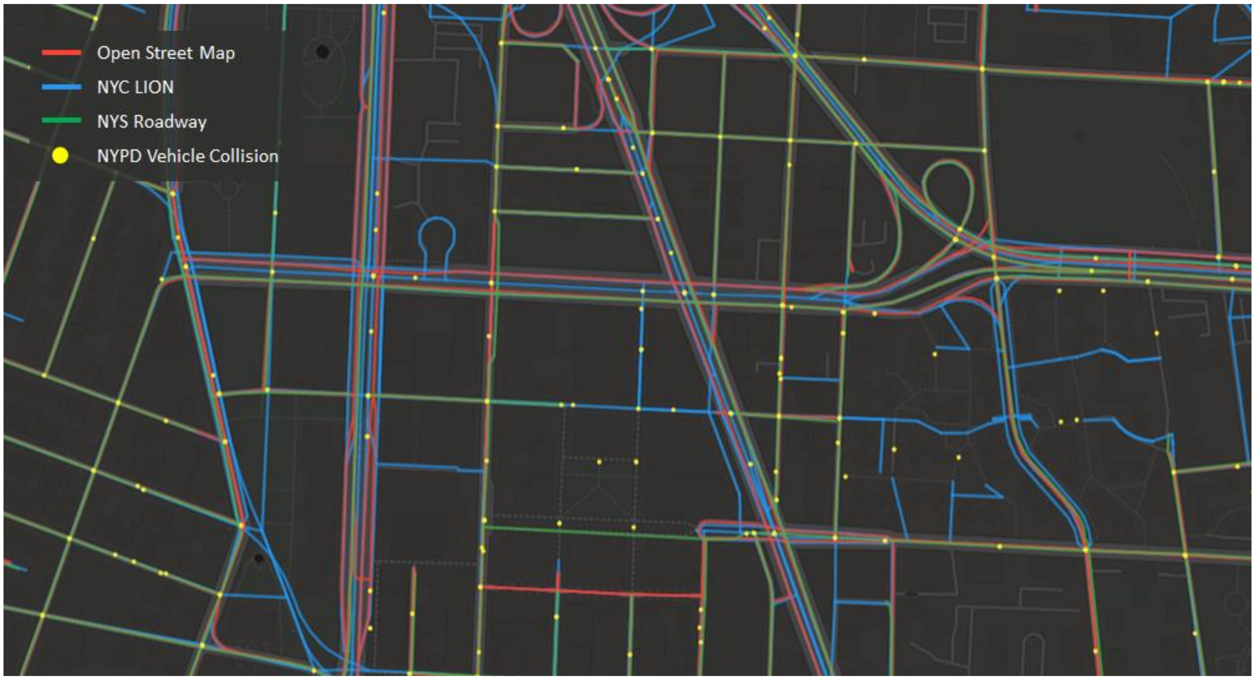
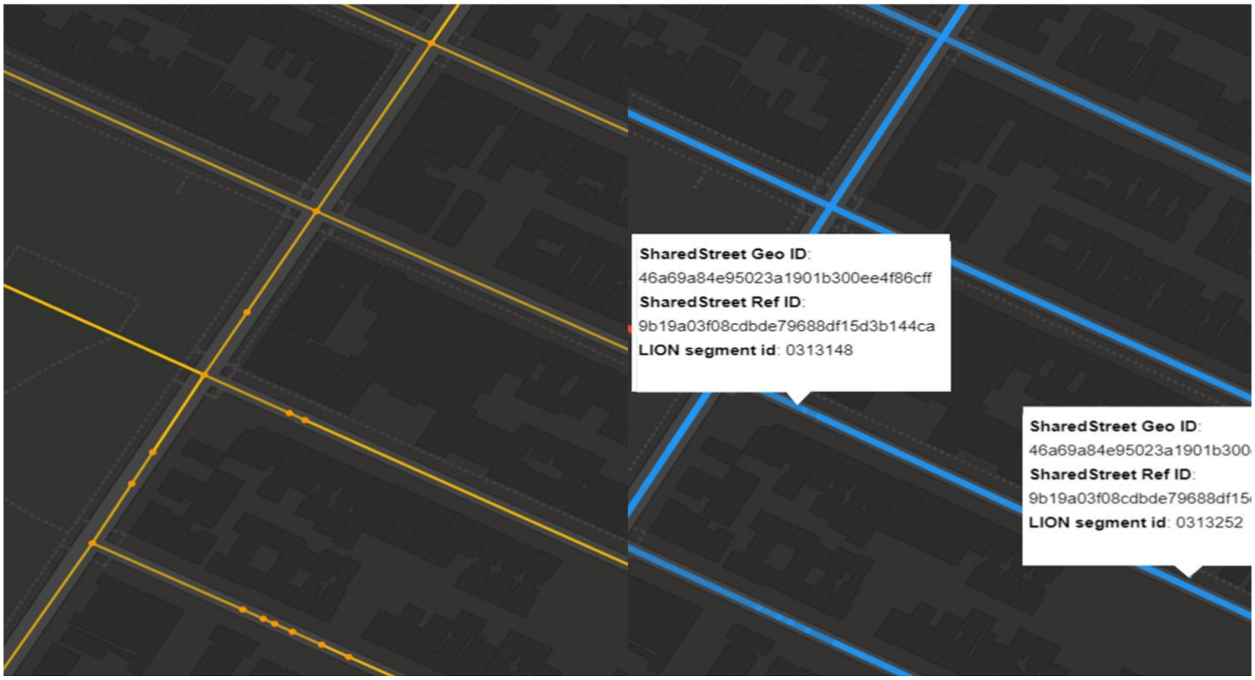


Figure 11. A visualization of disparate spatial data sets showing how different street representations (i.e. LION, OpenStreetMap, & NYS Roadway Inventory) do not match perfectly. It also shows how the NYPD vehicle collisions do not snap accurately on the network.



SharedStreet Geo ID: 46a69a84e95023a1901b300ee4f86cff
SharedStreet Ref ID: 9b19a03f08cdbde79688df15d3b144ca
LION segment id: 0313148

SharedStreet Geo ID: 46a69a84e95023a1901b300
SharedStreet Ref ID: 9b19a03f08cdbde79688df15
LION segment id: 0313252

Figure 12: SharedStreets API: Unifying LION street network with SharedStreets/OpenStreetMap geometry. In the left visualization you can see how a block face is segmented by several nodes that do not represent reality. The image to the right shows how using SharedStreets we were able to unify a block face into one segment while keeping the old LION IDs

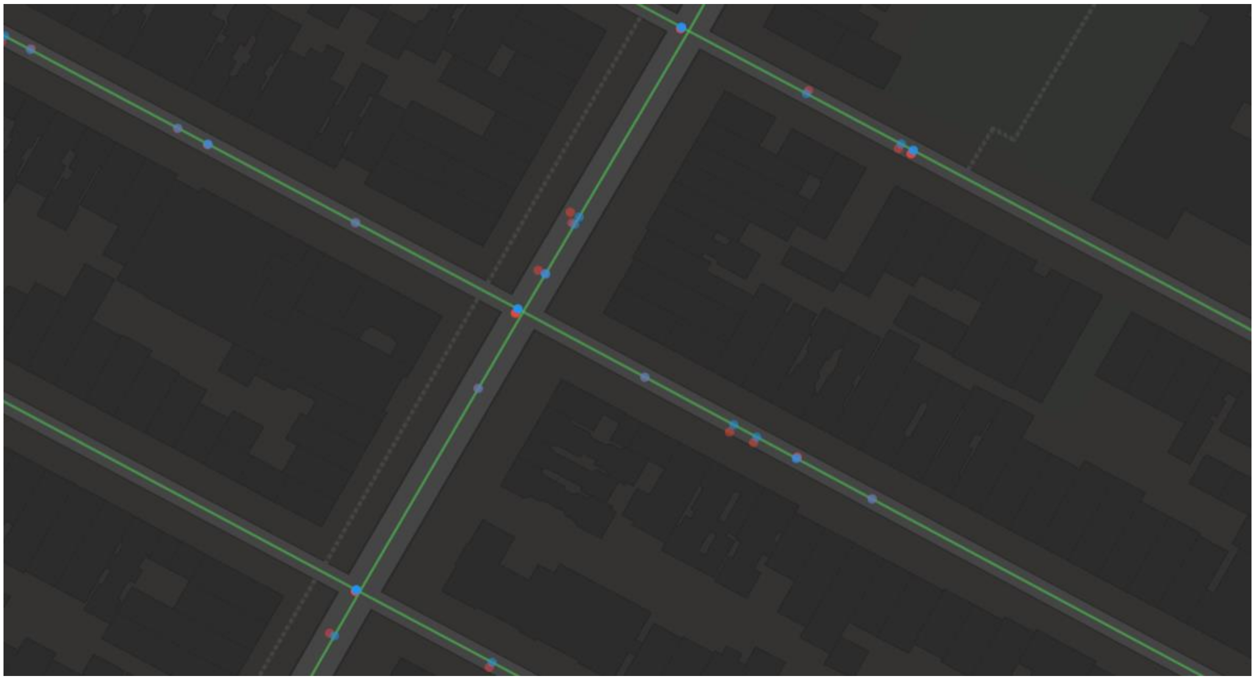


Figure 13: SharedStreets API: Mapping vehicle collisions on SharedStreets geometry. A visualization showing old crash points (red) and a new list of projected crash points (blue) that fall accurately on the network snapped using SharedStreets.

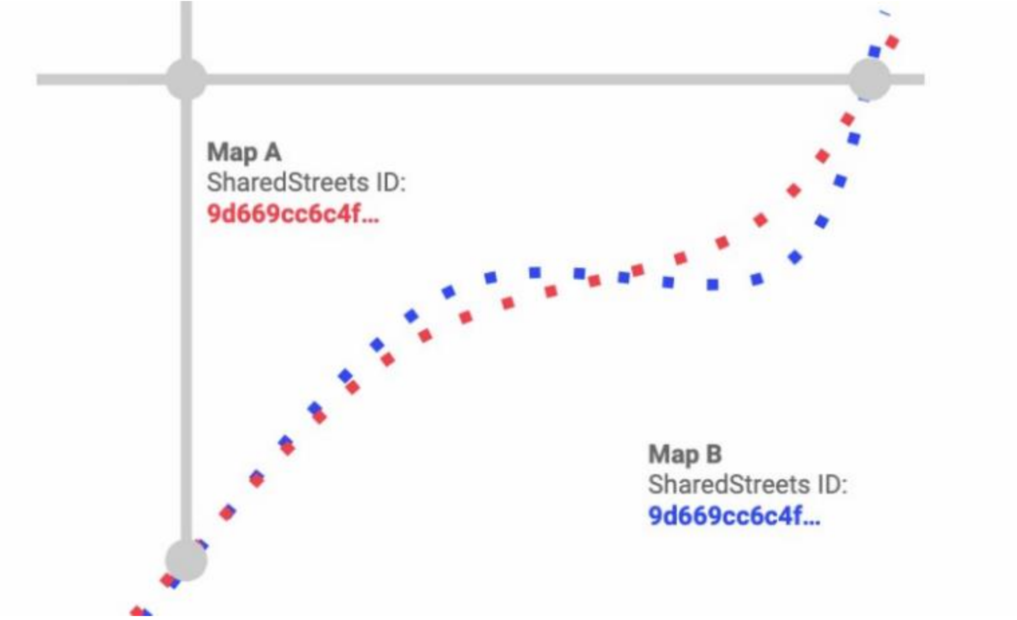
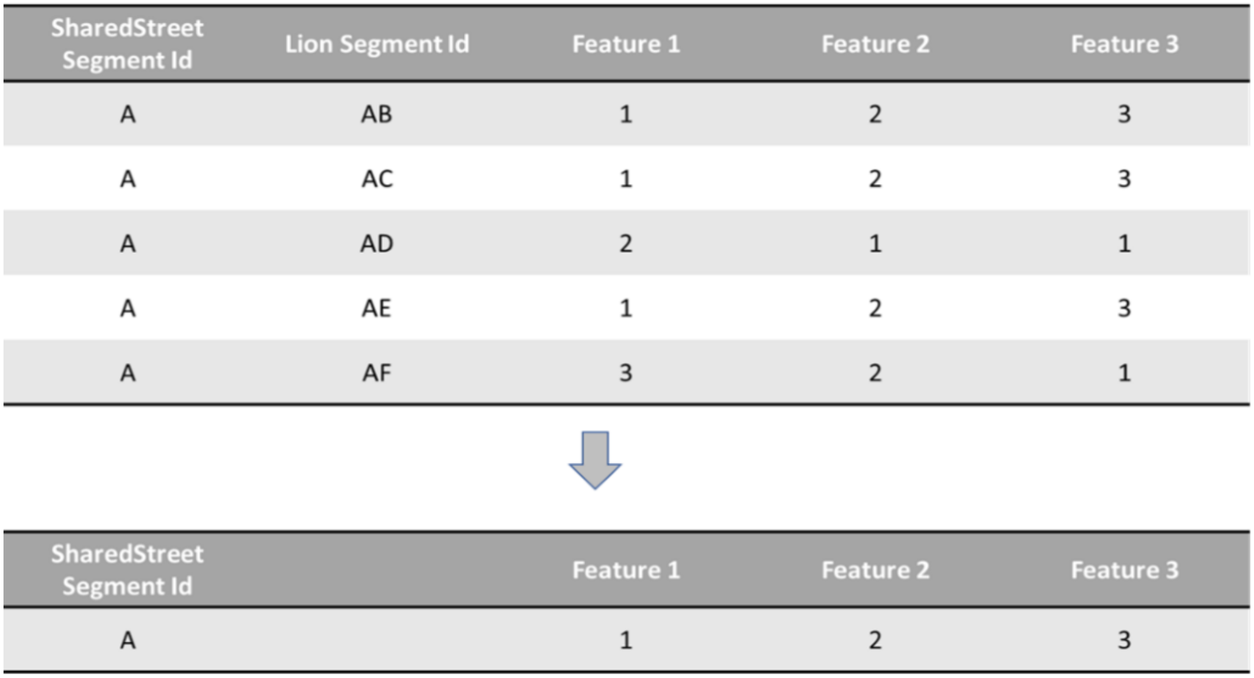


Figure 14: SharedStreets referencing system connects two street networks with a common ID.



SharedStreet Segment Id	Lion Segment Id	Feature 1	Feature 2	Feature 3
A	AB	1	2	3
A	AC	1	2	3
A	AD	2	1	1
A	AE	1	2	3
A	AF	3	2	1

SharedStreet Segment Id	Feature 1	Feature 2	Feature 3
A	1	2	3

Figure 15: A demonstration of the aggregation process used to group multiple LION segments that were assigned to the same SharedStreets id. The most frequent value of an attribute was used to describe the segment.

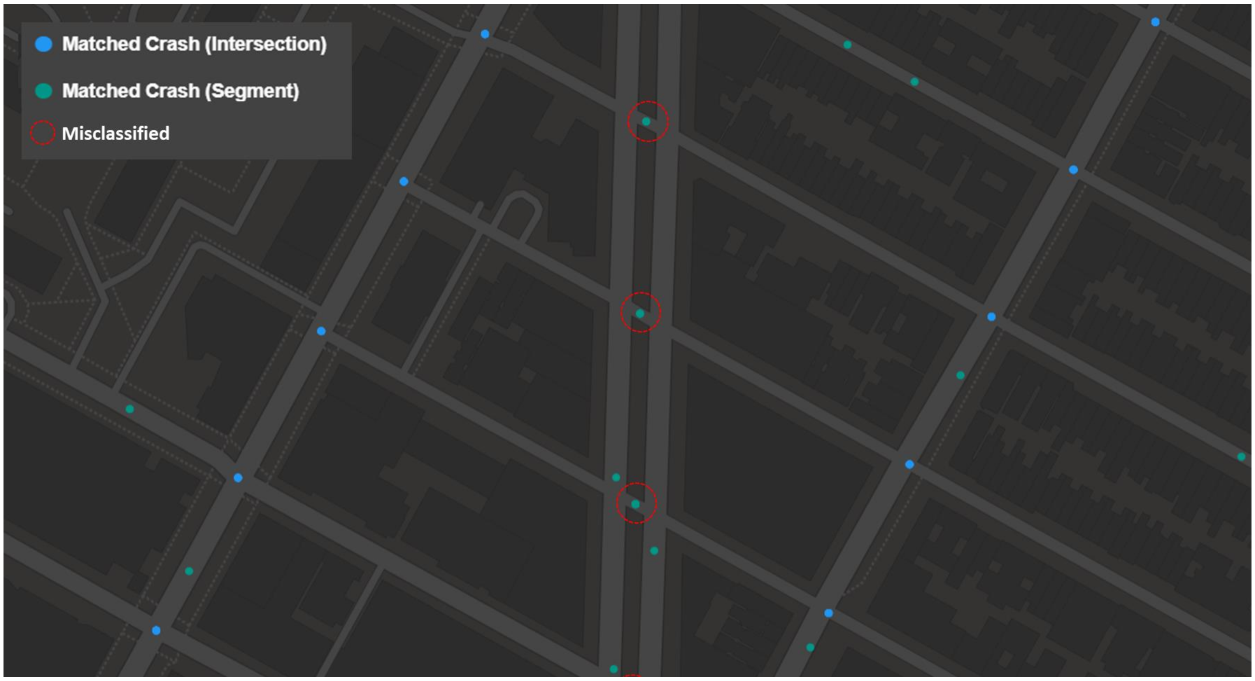
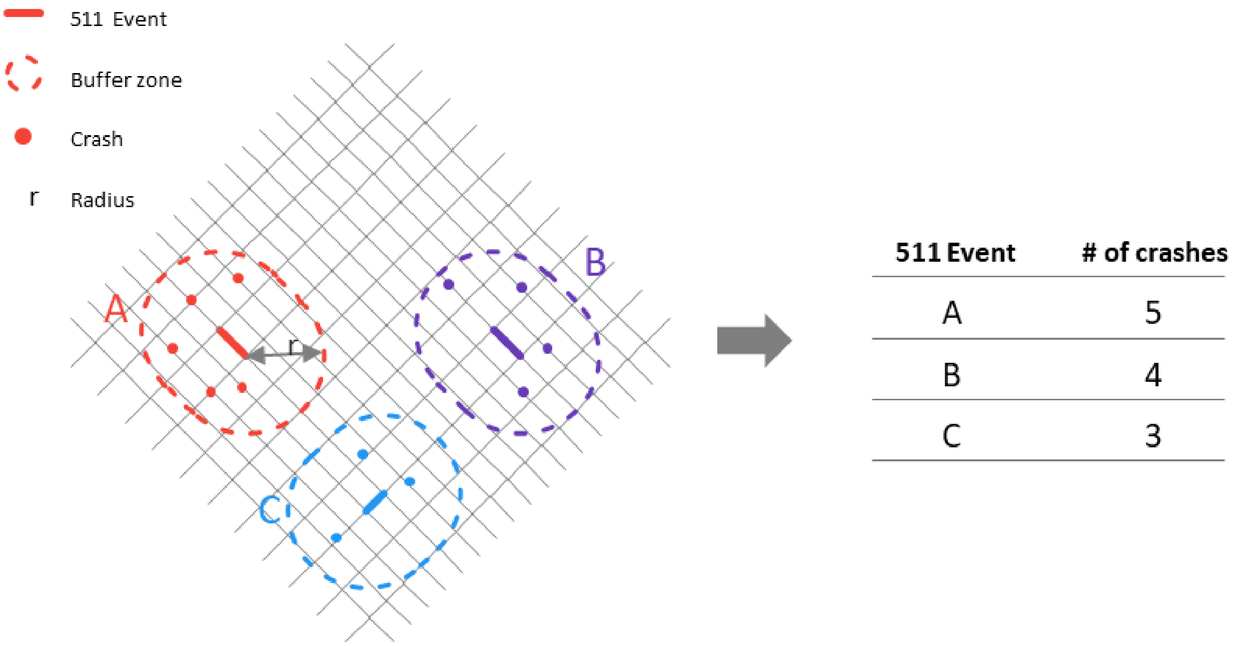
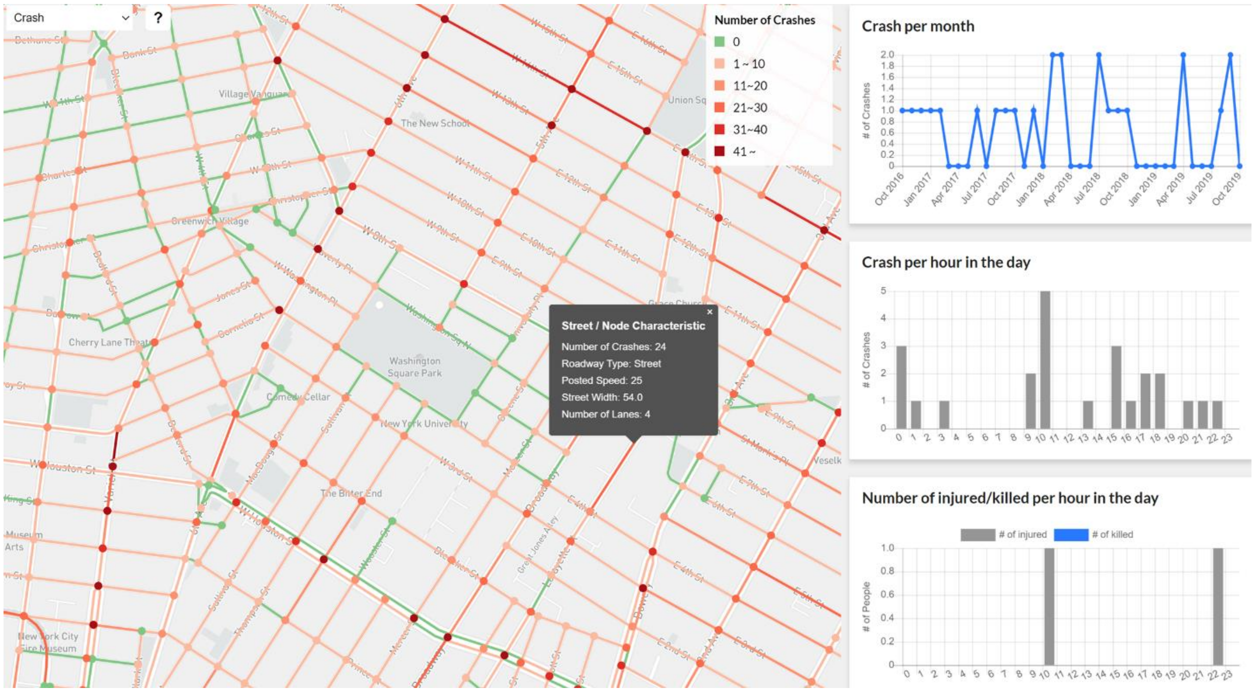


Figure 16. A map showing misclassified crashes. The circled locations represent crashes that were misclassified as segment crashes when they actually occurred at an intersection.



511 Event	# of crashes
A	5
B	4
C	3

Figure 17: An illustration of the methodology used for counting work zone related crashes. Each color represents a temporal range (i.e. active work zone duration) of a 511 event.



Street / Node Characteristic
Number of Crashes: 24
Roadway Type: Street
Posted Speed: 25
Street Width: 54.0
Number of Lanes: 4

Figure 18: A screen shot of the dashboard that can be visited here: https://workzone-collision-analysis.github.io/capstone/dashboard/.