NITC Doctoral Dissertation Research Fellowship

Progress Report

**Project: Unsupervised Approach to Investigate Urban Traffic Crashes Based on Crash Unit, Crash Severity, and Manner of Collision**

PI: **Farzin Maniei**

Co-PI: **Dr. Stephen P. Mattingly**

**Project Introduction:**

The study objective is to examine how the fragment size affects the aggregation of traffic crash data and proposes a methodology to determine the recommended fragment size (RFS) for highways mainlane for both directions of Texas Loop 12, IH-20, IH-30, IH-35E, IH-45, IH-635, and US-75 within Dallas County limit. The research utilizes three major characteristics of traffic crashes (crash units, manner of collision, and crash severity) to calculate featured crash rates (FCRs) and cluster highway fragments using the K-means algorithm. The study evaluates the clustering results using silhouette coefficients and aims to reveal patterns in traffic crashes across highways. Additionally, the study will develop traffic crash count models for identifying crash hotspots (HSID) along highways and investigate if RFS improves the performance of these models. Hotspots will be identified using the potential for safety improvement (PSI). Overall, the study aims to provide a data-driven approach to comparing different fragment sizes and improving traffic safety analysis.

**Project Progress Updates:**

This study is divided into three stages: (1) Investigating urban traffic crashes by applying an unsupervised machine learning approach to three traffic characteristics (crash units, manner of collision, and crash severity), (2) Performing crash frequency analysis and hotspots identification (HSID) using three traffic characteristics simultaneously, and (3) Examining the impact of the fragment size on statistical models, HSID, and the potential improvement of statistical models using the recommended fragment size (RFS). As of this update, the study is progressed by taking the following steps:

* Developed and investigated an unsupervised machine learning method to cluster traffic crashes across the uninterrupted highway mainlanes using three traffic crash characteristics.
* Drafting the first paper presenting the unsupervised approach and the associated findings. Submitted the first paper to the American Society of Civil Engineering (ASCE), Journal of Transportation Engineering: Part A. The paper is currently under peer review.
* Performed literature review on crash count statistical models, hotspot identification (HSID), and traffic safety measures.
* Developed a library of Python codes to implement count regression models and compare models.
* Drafted the second paper focused on crash count models for various groups of crashes formed by three traffic crash characteristics. Submitted the second paper to the Accident Analysis and Prevention Journal.
* Investigated the impact of various fragment size on traffic crash prediction models. Evaluated the impact of recommended fragment size (RFS), identified the scenarios that using RFS for data aggregation were beneficial to the crash prediction models.
* Drafted the third paper titled: “Investigating the Impact of Recommended Fragment Size to Improve Crash Count Prediction Models”.
* Drafted the final draft of the dissertation.

**Accomplishments & Results:**

The finding from stage (1) of study confirms the results expected. In this stage, a new data-driven method for aggregating crash data is developed, which uses three dimensions of traffic crashes (number of vehicles involved, manner of collision, and crash severity) and clustering techniques to determine the recommended fragment size (RFS). The proposed approach offers a standard method for future studies to aggregate crashes and addresses concerns about the arbitrary selection of segment length in previous research. The method employs the LSDBEM and K-means clustering algorithm, and the study defines featured crash rates (FCRs) to cluster highway segments. Results show that the RFS varies for each highway travel direction and that a single "best" segment length does not exist. The method is also applied to total crash rates (TCRs) and the results are compared with FCR-based results. The FCR-based clustering results also identify significant features for each highway travel direction, which is not practical with TCR-based clustering. The proposed methodology offers a more cohesive and distinct cluster of crash data, making it more valuable for future studies. Overall, the paper proposes a data-driven approach to determine the RFS that overcomes the arbitrary selection of fragment size (segment length), providing a more standardized and practical method for aggregating crash data.

In the stage (2), the study develops a methodology to explore the effect of including traffic crash dimensions on the crash prediction models, significance and magnitude of contributing factors, and traffic crash hotspot identification results for IH 20 within Dallas County limit. In the study, four scenarios are defined to form different crash groups using number of vehicle involved in crash (crash unit), manner of collision (crash type), and crash severity. In scenario 1, all crashes combined form a single crash group; scenario 2 creates the crash groups only based on crash units as the first dimension. For scenario 3, crash groups include crash units and manner of collision (crash type). Finally, in scenario 4, all three crash dimensions are used to form crash groups. The methodology investigates the performance of various statistical model forms by estimating crash count prediction models for each scenario crash group with sufficient observations; these models include Poisson, negative binomial, zero-inflated Poisson, zero-inflated negative binomial, generalized Poisson type 1, and type 2. The modeling results show that different crash groups' significant contributing factors vary. Also, almost half of each crash group's top ten hotspot segments are not mutually shared. The contributing factors and hotspot segments show substantial differences for each scenario, s scenario 4 shows especially considerable changes. This new approach provides a more robust understanding of crash groups when simultaneously considering all three crash dimensions.

In the stage (3), This study investigates the potential improvement in the performance of crash prediction models using the RFS obtained by LSDBEM/K-means. The study aims to explore the impact of various fragment sizes (ranging from 0.10 mile to 0.25 mile with an increment of 0.01 mile) on the crash count prediction modeling process including the multicollinearity among explanatory variables, the selection of suitable model types, and model performance and accuracy. In particular, it examines the potential benefits of determining the RFS rather than arbitrarily selecting a fragment size or testing a range of fragment sizes. The study results show that the multicollinearity among the explanatory variables varies for the different fragment sizes, and the minimum multicollinearity occurs when using the RFS. Also, the different fragment sizes cause changes to the dispersion values for the dependent variables and impact the preferred model type. The predictive model performances reached their minimum AIC at a fragment size of 0.24 and 0.20 mile under scenarios 1, 2, and 3 for IH 20 EB and IH 20 WB, respectively. The minimum AIC values for the scenario 4 models inconsistently appear at different fragment sizes . Although, the RMSE results showed that the minimum RMSE, occurred at RFS for some crash groups for scenario 1, 2, and 3. Also, there are several crash groups that the RMSE at RFS, , is within proximity (20%) of , meaning the performance of crash prediction models was as good as using the fragment size associated with . However, the RMSE values for scenario 4 did not demonstrate similar results. Finally, it is found that the GP2 and GP1 models are mostly suitable for the crash group formed by total number of crashes. The top two model candidates for SV-related and MV-related crash groups are {ZIP, GP2} and {GP1, GP2}, respectively.

**Updated Schedule for Completion of All Tasks:**

Per study progress and challenges, the project schedule is updated as follows:

|  |  |  |
| --- | --- | --- |
| Item | Description | Completion  Status/Date |
| 1 | Crash data acquisition, geovalidation, cleaning, and preparation. | Completed |
| 2 | Literature reviews on potential candidate for unsupervised feature selections | Completed |
| 3 | Develop and implement unsupervised feature selections and clustering method | Completed |
| 4 | Drafting and submitting paper 1: Unsupervised Approach to Investigate Urban Traffic Crashes Based on Crash Type, Crash Severity, and Manner of Collision | Completed |
| 5 | Performing statistical models and hotspots identifications | Completed |
| 6 | Drafting and submitting paper 2: Crash Frequency Analysis and Hotspots Identification using Crash Type, Crash Severity, and Manner of Collision | Completed |
| 7 | Examining the impact of the fragment size on statistical models and HSID | Completed |
| 8 | Drafting dissertation and submit to the committee | Oct. 23rd, 2023 |
| 9 | Defend the dissertation and submit the final draft of dissertation | Dec 6th, 2023 |
| 10 | Addressing potential comments from the committee | Dec 16th, 2023 |
| 11 | Defend the dissertation and submit the final draft of dissertation | Dec. 31st, 2023 |

**Performance Metrics:**

A graduate student, Farzin Maniei, has been working on the study as his Ph.D. dissertation.