MARYLAND DEPARTMENT OF TRANSPORTATION

STATE HIGHWAY ADMINISTRATION

FINAL REPORT: A DATA-DRIVEN SAFETY DASHBOARD ASSESSING MARYLAND STATEWIDE DENSITY EXPOSURE OF PEDESTRIANS, BICYCLES, AND E-SCOOTERS

MARYLAND DEPARTMENT OF TRANSPORTATION

MOTOR VEHICLE ADMINISTRATION

UNIVERSITY OF MARYLAND 1856

UNIVERSITY of MARYLAND BALTIMORE

TECHNICAL REPORT DOCUMENTATION PAGE

General instructions: To add text, click inside the form field below (will appear as a blue highlighted or outlined box) and begin typing. The instructions will be replaced by the new text. If no text needs to be added, remove the form field and its instructions by clicking inside the field, then pressing the Delete key twice. Please remove this field before completing form.

1. Report No.	**2. Government Accession No.**	**3. Recipient's Catalog No.**
4. Title and Subtitle A Data-Driven Safety Dashboard Assessing Maryland Statewide Density Exposure of Pedestrians, Bicycles, and E-Scooters		**5. Report Date** August 31, 2021 **6. Performing Organization Code**
7. Author(s) Chenfeng Xiong, Jina Mahmoudi, Weiyu Luo, Mofeng Yang, Jianyang Zheng, Carole Delion		**8. Performing Organization Report No.**
9. Performing Organization Name and Address Maryland Department of Transportation State Highway Administration 707 N Calvert St, Baltimore, MD 21202		**10. Work Unit No.** **11. Contract or Grant No.** **69A34520501050620**
12. Sponsoring Agency Name and Address United States Department of Transportation 1200 New Jersey Ave, SE Washington, DC 20590		**13. Type of Report and Period Covered** Final Report (September 2020 to August 2021) **14. Sponsoring Agency Code** **OST Policy**

15. Supplementary Notes

Conducted in cooperation with the U.S. Department of Transportation, Office of the Secretary of Transportation.

The project was supported by OST Policy funding under Notice of Funding Opportunity DOT-OSTP-SDT-2019-002

16. Abstract

This project's ultimate deliverable is a functional Vulnerable User Density Dashboard (https://mti.umd.edu/sdi) for the state of Maryland. The dashboard uses mobile device location data and electric scooter volume data to reflect pedestrian, bicycle, and electric scooter travel volumes and their exposure to roadway safety risk across all roadways in the State. The team develops advanced statistical models to study and predict pedestrian and bicycle involved crashes in a first of its kind effort to generate vulnerable roadway user risk at the link level. The results indicate high correlation between estimated volumes and observed frequency of crashes. This estimated volume and exposure data fills an important gap in understanding the spatial and temporal distributions of pedestrian and bicycle activities. Through this dashboard engineers, planners, and stakeholders can quickly identify safety risk hotspots for vulnerable road users across Maryland and start parsing out why certain locations with high volumes have less crashes. This data-driven dashboard uses emerging data sources and cutting-edge big-data analytics to derive volume estimates and crash risk predictions. It can be used by different stakeholders for situational awareness and traffic analyses of vulnerable users, i.e., pedestrians and bicycles. The predicted risk exposures and hotspots will support relevant decisions such as identifying locations for improvements and implementing pedestrian and bicycle safety countermeasures.

17. Key Words	**18. Distribution Statement**
Pedestrian; Bicycle; Exposure; Safety; Mobility; Location-Based Service Data; Visualization	No restrictions. This document is available through the National Transportation Library's Repository & Open Science Access Portal

19. Security Classif. (of this report)	**20. Security Classif. (of this page)**	**21. No. of Pages**	**22. Price**
Unclassified	Unclassified	10	

1. PROBLEM STATEMENT AND RESEARCH QUESTIONS

Maryland has applied considerable effort to realize a newly articulated vision: **Maryland will be a great place for biking and walking that safely connects people of all ages and abilities to life's opportunities**. From Mountain Maryland to the Eastern Shore and from expressways to walkable communities, safety on Maryland roadways is essential to connecting all Marylanders to recreational, economic, and social opportunities. To realize this vision pedestrian/bicycle deaths and severe injuries must be minimized and potentially eliminated. A crucial step in fostering a safer environment for pedestrians, bicycles, and scooters is to better understand current pedestrian/bicycle/scooter volumes and their exposure to roadway safety risks. Proper data, analytics, and visualization tools are imperative to improve situational awareness.

The safety exposure for road users outside of the vehicle is pivotal for appropriate risk characterization and for safety-related decision-making. However, quality pedestrian/bicycle exposure data is often regarded as a missing piece of information in safety analyses. The current state of the practice uses manual and automated count data, census data, travel surveys, and land use information from a single point in time to understand and estimate pedestrian and bicycle volumes (see next section for state of the practice). Studies have demonstrated that exposure to risk plays an important role in helping agencies better understand the cause of crashes and incident severity; however, it is difficult to accurately identify pedestrian/bicycle movements and then prioritize high-risk locations due to the lack of exposure data (FHWA, 2017). In fact, without pedestrian volumes as a measure of exposure, crash models are undermined with reduced goodness-of-fit, biased parameter estimates, and incorrect inferences (Xie et al., 2018). The focus of this project is to address the research need and data gap in a reliable, comprehensive, and up-to-date information source on pedestrian/bicycle exposure, and to convey such information in a useable platform. The project team aimed to answer the following questions:

- Can probe data be dependably converted to the density of vulnerable user movement in a dynamic platform across time **and** at the link level?
- Can crash data be parsed enough to separate modes of transportation and mapped over the density of roadway use?

2. STATE OF THE PRACTICE GAPS

States and jurisdictions are trying to address vulnerable user risk exposure challenges by adopting count data and supplementary information to analyze pedestrian/bicycle volumes at critical locations. However, there are explicit drawbacks to this approach: 1) the data sources are limited and expensive to collect; 2) the spatial and temporal coverage of the data is limited; 3) running and maintaining such models require technology transfer and staff support.

Table 1. Examples of Pedestrian/Bicycle Safety Exposure Studies

State	Coverage	Data Source(s)	Measure of Exposure
Connecticut (Qin and Ivan, 2001)	Rural areas in Connecticut	Manual counts; population and land use data	Weekly pedestrian crossing volume
California (Raford and Ragland, 2004)	Oakland	Manual counts; census data	Average annual pedestrian volume
California (Schneider et al., 2012; Schneider et al., 2013)	San Francisco	Manual counts; automated counts	Number of pedestrian crossings

District of Columbia (Molino et al., 2009; Molino et al., 2012)	Washington, D.C.	Manual counts; crossing distances	Pedestrian miles traveled
Minnesota (Hankey and Lindsey, 2016)	Minneapolis	Manual counts; census and land use data	Bicycle and pedestrian volumes
Florida (Radwan et al., 2016)	FL Statewide	Counts; population; vehicle ADT	Pedestrian miles crossed per entering vehicle
Michigan (Cai et al., 2018)	MI Statewide	Census; statewide travel survey; land-use data	Vehicle traffic and pedestrian volumes
California (Griswold et al., 2019)	CA Statewide	Counts; network and land use variables; census data	Annual pedestrian crossing volume

Most of the above studies applied statistical models to estimate facility-level volumes at intersections and pedestrian network segments based on a limited spatial and temporal scope of pedestrian and bicycle counts. The typical coverage of such data and models is at the city level. Until recently, statewide analyses of pedestrian exposure have not been seen in the literature (e.g., Cai et al., 2018; Griswold et al., 2019). A safety data tool that comprehensively covers a large-scale geographical area with detailed temporal representation is rarely seen in literature, and only a handful of commercial products are available on the market.

3. HOW THE DASHBOARD ADDRESSES THE KNOWLEDGE GAPS

To reduce several of the previously mentioned limitations, this project leverages mobile device location data and integrates it with other data sources into a data-driven analytical and visualization dashboard for transportation safety. Compared to the current state of the practice that relies on surveys and manual/automatic counts, this emerging big-data source has significant advantages:

- it no longer requires costly and time-consuming surveys or counts collections;
- mobile device location data draws evidence from a much larger sample, covering over 40% of the entire U.S. population;
- unlike cross-sectional data (i.e., data taken only on one or a few days of the year), this data provides a continuous time series of pedestrian and bicycle movements (we proposed in this project to analyze a one-year time period) and covers the entire Maryland roadway network, including rural areas where traditional pedestrian/bicycle data is lacking;
- the availability of the data source is in the entire U.S., not just Maryland, which would enable a fast and cost-effective transfer of this project to other states and jurisdictions throughout the nation.

The final product of this project is a dashboard (https://mti.umd.edu/sdi) that reports safety exposures and crash information for pedestrians, bicycles, and e-scooters at a microscopic level and is available to all practitioners across various levels of government in Maryland and nationally. Details of the technical process to achieve this deliverable are provided in the following sections.

4. DATA SOURCES AND METHODOLOGICAL APPROACH

DATA SOURCES AND DATA ASSEMBLY

The data assembly process forges a comprehensive and secure warehouse of vehicle, pedestrian, bicycle, and available e-scooter activity data based on location-based services, other transportation sector data available in Maryland, and pedestrian/bicycle-involved crash data records. This section summarizes these data sources. Note that the dashboard (https://mti.umd.edu/SDI) includes an in-depth white paper on methodology under the "Methodologies" section.

Mobile Device Location Data

This is the principal data source leveraged in the project. Collected directly from individuals' smartphones, this source of data collects location points whenever the mobile device users are using the location-based service of the devices. This data is collected continuously for each sample. Thus, individual-level mobility of the data sample could be inferred, including trips, travel times, departure times, speed profile of each trip, and adjacency to different transportation system infrastructure such as roadway network, transit stations, etc. With this level of detail, critical information concerning this project could be inferred, including travel modes (walking, biking, driving, etc.), travel routes/paths where walk/bike activities are observed, and the timing of each trip, etc. In full transparency, the Maryland Department of Transportation (MDOT) and the US DOT were never in contact with this data, nor will they have access to the data after this project. It was clearly outlined prior to the project start that this data would remain with the Maryland Transportation Institute (MTI) and not be shared due to privacy concerns.

MTI maintains access to several major mobile device data sources that have wide and comprehensive coverage of mobile device location data internationally. It has developed big-data analytical algorithms that this project can leverage through ongoing parallel collaborations with the US DOT Federal Highway Administration (FHWA). MTI also has developed a series of validated data mining algorithms to turn sample data into individual trip analytics. This project utilizes mobile device location data collected in the year 2019 (Jan 1, 2019 – Dec 31, 2019). It should be noted that raw data and trip-level results are analyzed via the separate project effort and beyond the scope of this project. Within this study, only the aggregated mobility data at the intersection and link segment level is used and analyzed.

Pedestrian and Bicycle Crash Data

Crash records visualized and analyzed in the study are collected from MDOT State Highway Administration (SHA) and cover pedestrian and bicycle crashes that occurred in Maryland from 2016-2019. Information has been aggregated at the link level and intersection level. For privacy protection, the project dashboard only highlights the centroids of links where the crashes occurred. For links that have reported crashes, the following statistics are provided:

- Involved: including the total number of crashes that occurred on that particular roadway segment in 2016-2019, as well as number of bicyclists, number of pedestrians, number of e-scooters, and number of vehicles involved in those crashes.
- Severity: aggregated number of people by injury severity levels (i.e., property damage only; injury; and fatality).
- More details: situational awareness for the most frequent movements of pedestrians, bicyclists, e-scooters, and vehicles involved in the crashes.

E-Scooter Volume Data

The project team has collected e-scooter data from the Baltimore City Department of Transportation. The project dashboard currently has one data layer displaying e-scooter volume records for Baltimore. The currently implemented data includes the average daily e-scooter volume on each roadway link in Baltimore City for October 2019. This project has demonstrated the feasibility to ingest and manage e-scooter volume data in the designed dashboard. Additional data for other time periods or other localities could be collected and hosted; however, for the purposes of this project access to that data was not available.

Additional Data for the Transportation System of Maryland

To deliver a comprehensive data-driven tool, additional information was collected and integrated into the database to train, test, and validate the data analysis algorithms and crash prediction models, including:

- Hourly vehicle and pedestrian/bicycle count data covering the entire year of 2019.
- A multimodal transportation network and its associated network-level details are collected from *OpenStreetMap* (OSM). Moreover, its embedded point-of-interest (POI) information will help the team identify rich location and infrastructure information that are essential to supplement the identification of pedestrian and bicycle activities as well as the prediction of safety hotspots.

Socio-demographics, Land Use Data, and other Supplementary Information

To support the analyses of pedestrian/bicycle activities, exposures, and predictions of crash risks, the team also relies on additional information about socio-demographics and land use variables, including:

- ***American Community Survey (ACS)***, an annual survey program conducted by the United States Census Bureau. Data collected through this survey provide information about the population (e.g., socioeconomic and sociodemographic characteristics, means of commuting to work) as well as housing (e.g., financial and physical characteristics for housing units) at many geographical scales. The 2019 five-year ACS estimates at the census block group have been used in this study as the source of socioeconomic and sociodemographic data in the development of pedestrian/bicyclist crash frequency models.
- ***Smart Location Database (SLD)*** is a nationwide spatial dataset for the U.S. The SLD is a product of the U.S. Environmental Protection Agency (EPA)'s Smart Growth Program and is publicly available on the EPA web site. The latest version of this dataset is the SLD Version 2.0, which was released in 2013 (EPA 2021). The SLD provides information on land use and built environment characteristics such as population and employment density, housing density, land use diversity, urban design attributes, destination accessibility, transit accessibility, transit service frequency, as well as socioeconomic and sociodemographic characteristics at the census block group level. These characteristics of SLD were employed in the project in examining the role of the land use and built environment attributes in modeling the frequency of pedestrian/bicyclist crashes.
- ***Level of Traffic Stress (LTS)*** was calculated for each roadway segment and intersection, to quantify the traffic stress a road segment or intersection imposes on the bicyclist as a surrogate measure of pedestrian bicycle safety. The LTS for each link was calculated using the speed limit, number of lanes, bike lanes, authorized travel modes (vehicle, bicyclist, and pedestrian), and other road geometry information from the OSM data. The intersection LTS is calculated by averaging the LTS value for all approaches. LTS information has been provided via the project dashboard at the intersection and segment level. The LTS score also contributes to the crash frequency models.

ANALYTICAL APPROACH OF USING MOBILE DEVICE LOCATION DATA

Figure 1 shows the methodology flowchart of the dashboard. The project team estimated the vehicle and pedestrian bicycle volume at the link-level (i.e., roadway segment level) and intersection-level, leveraging a set of previously developed and validated algorithms made available via the USDOT Federal Highway Administration's Exploratory Advanced Research

Program project entitled "Data analytics and modeling methods for tracking and predicting origin-destination travel trends based on mobile device data."

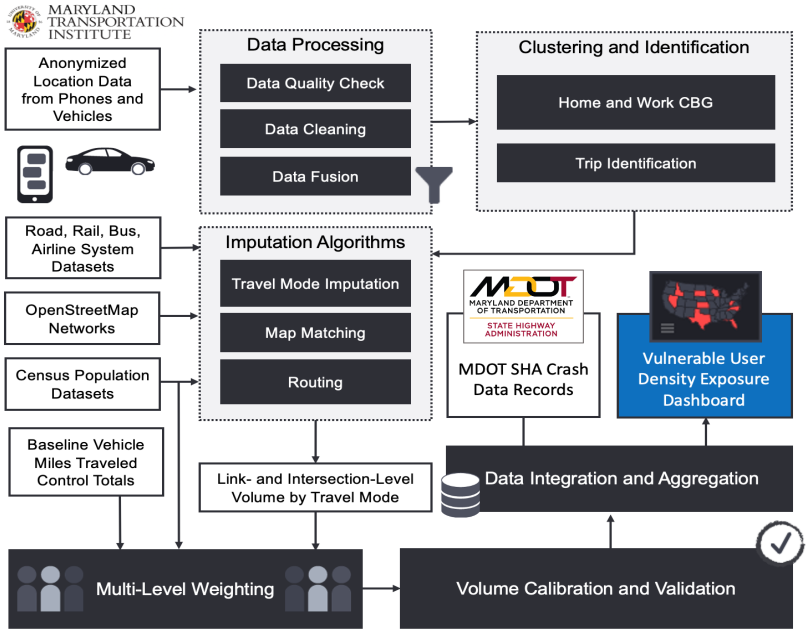

Figure 1. Methodology flowchart

Several key steps were taken in order to derive pedestrian and bicycle mobility data at each intersection and roadway segment:

- First, we employed a state-of-the-practice data preprocessing to integrate and clean the mobile device location data;
- The team then clustered location points into activity locations and identified home and work locations at the census block group (CBG) level to protect privacy;
- Next, we applied previously developed and validated imputation algorithms to identify all trips from the cleaned data panel, including trip origin, destination, departure time, and arrival time. Most importantly, travel modes and routes were imputed so that pedestrian and bicycle activities can be inferred and aggregated to each intersection and roadway segment;
- Lastly, postprocessing steps were taken to expand our sample to the entire population, validate our estimated volumes, and visualize the mobility data on a large-scale and interactive map-based visualization platform embedded in the project dashboard.

Data Processing

Some common issues, such as unordered and duplicated records, need careful treatment before extracting any information from mobile device location data. The state-of-the-practice methods for raw data cleaning and quality control often include identifying and merging duplicate device observations, removing outliers, and checking on the obvious data consistency issues (e.g., devices with unreasonably high-speed readings). The data processing procedure taken by the research team is based on the four dimensions of data quality assessment: consistency, accuracy, completeness, and timeliness. More details are offered in Appendix A.

Trip Identification

Trips are not initially included in any mobile device location data sources. Instead, location sightings are continuously generated while the sample device moves, stops, stays static, or starts a new trip. As a result, MTI developed a trip identification algorithm, which can detect which location sightings form a trip together. The first step is to sort device observations by time. The algorithm assigns a random ID to each trip it identifies. Many location points in the dataset may belong to no trips. The algorithm assigns "0" to the trip ID of these locations to tag them as static points. Then, a recursive algorithm checks every point to identify if they belong to the same trip as their previous point. Results and computational algorithms were validated based on a variety of independent datasets such as the National Household Travel Survey (NHTS) and the American Community Survey (ACS), and peer-reviewed by an external expert panel in a US DOT Federal Highway Administration's Exploratory Advanced Research Program project (Zhang and Ghader, 2020). By running the algorithm on our data and comparing the trip lengths and travel times with the reported travel distances and travel times from the 2017 National Household Travel Survey, a satisfactory match was observed. Moreover, we compared mobility trends calculated by our methods and by other data sources including Apple, Google, and SafeGraph and found high consistency. All data and trip identification remained securely within MTI's systems.

Imputations of Home and Work Locations, Travel Modes, Routes, and Matching to Map

Home and Work Locations: A typical methodology for identifying home and work clusters is to identify the most frequently visited clusters during the night and during the day. The algorithm first applies a HDBSCAN clustering algorithm to cluster all device observations into activity locations. This step takes the cleaned multi-day location data as input and applies an iterative algorithm until no cluster has a radius larger than two miles. The iterative algorithm consists of two parts: HDBSCAN, based on a minimum number of point parameters and filtering non-static clusters based on time and speed checks. After finalizing the potential stay clusters, the algorithm combines nearby clusters to avoid splitting a single activity. Here, instead of setting a fixed time period for each type, e.g., 8pm to 8am as the study period for home CBG identification and the other half day for work CBG identification, the framework examines both temporal and spatial features for the entire activity location list. The benefits are two-fold: the results for workers with flexible or opposite work schedules is more accurate and the employment type for each device can be detected simultaneously.

Travel Mode Imputation: A machine learning model was developed and applied to impute the ground transportation travel modes for non-air trips, including vehicle (car and bus), rail, and walk/bike. Feature engineering directly affects the model performances. Three types of features are incorporated for transportation travel mode imputation, including location recording intervals, trip-level features, and features about multimodal transportation network (a total of 32 features). More details are presented in Appendix A.

Map Matching and Routing: The team developed and implemented a computationally efficient method to map crash and vehicle/pedestrian/bicycle/e-scooter movements data to an all-street roadway. This allows project engineers to evaluate the weighted vehicle/pedestrian/bicycle/e-scooter volumes at each intersection and each roadway segment. In addition, the team has worked with MDOT SHA to incorporate other available safety data sources from the state of Maryland into the map matching process.

Postprocessing

Weighting: The sample data must be expanded to produce population-level statistics of safety exposure for those vulnerable road users. The LBS represents a sample of the population, so device-level weights were needed to expand the device sample to the broader population. Also, for an observed device, only a sample of all trips may be recorded, so trip-level weights were needed. In order to obtain device-level weights, MTI used the home locations, obtained from the imputed home CBG information. The weight for each device is equal to the number of devices observed in the device's imputed home location divided by the population of the zone where home is located. So, all devices residing in that CBG would have the same device-level weight. These weighting processes expand the data analytics to cover the entire population, such that the mobility of the individuals who are not observed in the data or that do not own a smartphone with location-based service is also captured by our analytics.

Validation The estimated vehicle and pedestrian/bicycle volumes are validated before visualization and being used for the project dashboard. For vehicle volume validation, the project team employed hourly traffic count data from MDOT SHA, including 1,933 portable traffic count stations (189,456 records) and 62 permanent traffic count stations (968,880 records). For pedestrian bicycle volume validation, the project team collected 15-minute interval pedestrian/bicycle count data from the above dataset that records the number of pedestrians and bicycles coming from each approach of an intersection. A total number of 845 count locations (89,500 records) were included in the pedestrian/bicycle count data. The pedestrian bicycle counts are further aggregated into 1-hour interval for validation.

The team validated data using two standard metrics that are widely adopted in research and practice: (1) R-Squared (R^2), a statistical measure that quantifies how much the dependent variable can be explained by the independent variables, and, (2) Normalized Mean Absolute Error (NMAE), defined as Mean Absolute Error (MAE) divided by the mean value of all observations.

The validation of vehicle volumes has a 0.98 R^2 and 10% NMAE. The validation of pedestrian/bicycle volumes has a 0.80 R^2 and 32% NMAE. The accuracy of the project estimated vehicle and pedestrian/bicycle volumes meets the Federal Highway Administration validation target (Cambridge Systematics, 2010) and is in line with the accuracy of other commercial products (e.g., StreetLight, 2020). The accuracy of validation for pedestrian/bicycle volumes is naturally lower than vehicle volumes, partly due to the data discrepancy. The current data collection method aggregates pedestrians and bicycles on each approach of an intersection. Unlike vehicle trajectories, which follow the roadway network directions, pedestrian/bicycle trajectories at an intersection are widely unknown under existing collection methods (manual counts). In other words, the mobile device location data covers more pedestrian/bicycle activities than volumes that are crossing the intersection. On the other hand, for intersections/segments where bus stops are located, certain pedestrian/bicycle trajectories could be categorized as bus trajectories and lead to underestimation. This is a parsing that should be further explored.

Visualization After extensive exploration, traditional vectorized visualization, which is typically used in GIS, Tableau, or Mapbox applications, was found to not unsuitable for this project. The data being visualized in this project is dramatically larger in size and covers all roadway segments and intersections in Maryland, which leads to noticeable time lag when zooming in and out using the vectorized visualization. Instead, the team developed a rasterization method to effectively visualize the mobility data generated from the analyses process. The team first defined multiple

zoom levels for the maneuver of the visualization tool (in total, sixteen zoom levels have been defined). For each zoom level, a different scale is used to represent different level of visualization fidelity. Then, the visualization generates multiple tiles of figure files in the PNG format. This rasterization method is superior in visualizing "big" data. It brings all the heavy computation offline, while traditional vectorization method has to perform rendering and vectorizing on the go. This finding was critical to stakeholders reviewing the tool. Visualization being critical in the dashboard development led the team to spend significant time refining the exponentially large datasets.

ANALYTICAL APPROACH OF ESTIMATING RISK

As part of this project, crash frequency models were developed for pedestrian and bicyclist involved crashes at Maryland intersections and road segments in order to predict and identify safety risk hotspots. This section summarizes and discusses the results of those crash frequency models at the higher level. More details about the modeling study are offered in the project white paper (see Appendix B).

Four statistical models were developed and estimated in this project to examine the role of various key contributing factors including safety risk exposure factors in pedestrian and bicyclist crashes that have occurred in the state of Maryland. These models are the Poisson model, the negative binomial model, the zero-inflated Poisson model, and the zero-inflated negative binomial (ZINB) model. The results of these models have been compared to identify the most suitable model that best fits the data. For the purposes of this study the ZINB model provided the best results.

The model incorporates vehicle and pedestrian bicycle volume estimates and other geometric design characteristics, socio-demographics, built environment features, as well as traffic-related information. The results indicate that key contributing factors to pedestrian bicycle-related crashes include number of intersections legs and level of traffic stress (LTS) ratings, commute mode shares, road network density and activity density, percentage of low-income workers, among other factors. The results also highlight that the inclusion of estimated vehicle and pedestrian bicycle volume significantly improves the performance of the model. The model was then used to predict the number of pedestrian and bicycle crashes, which in turn generated the crash risk for each intersection and road segment within the state of Maryland.

5. DESCRIPTIONS AND INSIGHTS OF THE DASHBOARD TOOL

The final deliverable of this project addresses the initial research questions. First, the 2019 safety exposure of pedestrians, bicyclists, and e-scooters in Maryland was successfully estimated at the road segment and intersection level based on emerging data-driven methods and statistical and predictive modeling. Second, such exposure information was processed, visualized, and packaged on an online interactive platform (https://mti.umd.edu/sdi) for researchers and practitioners to use for a complete situational awareness of vulnerable roadway user safety. These collectively address the research need and data gap through a reliable, comprehensive, and up-to-date information source on pedestrian/bicycle exposure, and convey such information through useable visualization tools. The following information is provided via the dashboard:

- Base map visualizations of vehicle volume, pedestrian/bicycle volume, and e-scooter volume by different season, day of week, and time of day.

- An informational panel offering statistics for each intersection/link segment. Statistical measurements including vehicle volume, pedestrian/bicycle volume, e-scooter volume, predicted safety risks, and level of traffic stress (LTS) by different season, day of week, and time of day. Measurement estimates are provided for the year 2019.
- Ranking analytics highlighting the top ranked locations in terms of those abovementioned measurements in Maryland and in each jurisdiction of Maryland.
- Pedestrian- and bicycle-involved crash visualization and informatics including units involved, crash injury severity, and vehicle/person movements. Information is aggregated to each link segment. Crash data employed covers the time period of 2016-2019.

The development of this safety dashboard is leading edge, though we note that the US DOT is involved in multiple efforts nationally within Pooled Fund Studies to generate greater data-driven situational awareness of the transportation system through location-based data (e.g. TPF-5(384)). Several of these national efforts use mobile device location-based data, which is collected via the mobile device's downloaded application location pinging through GPS-enabled smartphones.

We have limited information on crash risk rates, in part due to the lack of exposure information. This tool fills the information gap and supports users from throughout MDOT and local jurisdictions to identify safety risk hot spots and improve the situational awareness of existing pedestrian and bicycle activities. Users of this dashboard can analyze vulnerable user risk and further investigate through engineering analyses. Moreover, the improved understanding of vulnerable user density also advances the modeling and predictions of crashes, as highlighted by our modeling practice. This is also fully leveraged in the dashboard, where users may rank roadway facilities at the link level by volume and predicted safety risks, and by a specific jurisdiction. This would assist local agencies with their respective priority decisions, such as developing safety countermeasures and pedestrian/bicycle safety improvement plans and decisions. The dashboard is not a substitute for engineering analyses, rather it is a starting point for increased understanding of the transportation ecosystem. It can be used in planning to identify the limits of a project with similar safety risks, identification of locations where government agencies need to pay closer attention, and also provides insights on locations that have low risk but high volumes of vulnerable roadway users, which may indicate that a roadway treatment is successful in reducing crashes.

6. LESSONS LEARNED, CHALLENGES, AND OPPORTUNITIES

This project has successfully integrated mobile device location data into a data-driven analytical and visualization dashboard for assessing safety exposure of pedestrians, bicycles, and e-scooters. This study demonstrated crash prediction models empowered by data-driven volume data at intersections and link segments. The pedestrian and bicycle volume estimates reliably demonstrated an increased goodness of fit of predictions, which in research had not been recorded in full, and enables a spectrum of future studies to be pursued in the transportation safety domain.

The project has also learned additional lessons that could potentially benefit other data-driven projects within and beyond the scope of transportation safety.

- Zero-inflated negative binomial models were found to be superior compared to other statistical approaches at the intersection level and segment level. The models are also successfully employed in this project for predictions of safety risks.

- A novel rasterization visualization was developed to facilitate large-scale visualization using the data. Compared to vectorization, this method significantly improves the front-end performance by moving most of the computation to the backend offline environment.

Several challenges of the project are also noted and worth further exploration:

- It was found challenging to identify pedestrian/bicycle traces given the existing data evidence. More data must be collected to improve the algorithms as well as validate the data-driven approach. Evidence could be drawn from dedicated pedestrian/bicycle GPS surveys, video records, and other supplementary data sources based on advanced learning and artificial intelligence. Additional empirical data would further improve the methodology.
- The needs of differentiating pedestrian and bicycle volumes and exposures are noted from internal and external feedback. This refinement would support mode-specific decision making. However, these two modes are currently combined as one due to the close similarity. The team expects to explore ways to distinguish them in the near future. Possible directions include adding data and features such as exclusive facility information (bike lanes, walking trails etc.), further exploring the unique characteristics of speed distributions, and so forth.
- E-scooter analyses was also found challenging. As an emerging mobility option, e-scooter was relatively less studied and thus limited references and data are available for the team to construct a machine learning model and extract the associated mobility from the mobile device data. In addition, e-scooter safety data is largely lacking. Most crash reporting procedures do not record e-scooter crashes as a standalone crash category. The recommendation from this research is that it is imperative e-scooters be recorded as their own category in some fashion given the abundant safety risks faced by e-scooter users to support research and safety mitigation strategies.

In summary, this project successfully analyzed and integrated emerging mobile device location data to fill a critical gap in the field of transportation safety, i.e., measuring pedestrian and bicycle volumes and safety exposure across all links within the State of Maryland. The novel approach to use big-data on a data-driven dashboard that visualizes volume estimates for vulnerable road users and predicts and identifies safety hotspots significantly improve the state of the practice for planners and engineers. Compared to the current use of surveys and manual/automatic counts, leveraging this emerging big-data source has significant advantages. With the improved understanding of pedestrian and bicycle volumes, safety risk exposure of those vulnerable users can also be analyzed at more disaggregated level. The tool can immediately help stakeholders gain more insights into pedestrian and bicycle traffic that previously was rarely available in practice. Stakeholders who reviewed this dashboard immediately identified with the ability to see “the bigger picture”, the ability to rank the top locations within the state or specific counties, or knowing that the riskiest locations or not single intersections but the trajectory of multiple “clusters” of intersections. The regions around Washington D.C. especially are already using this dashboard to provide more situational awareness as the narrative shifts from single point improvements to a system of improvements for increased vulnerable roadway user mobility. Ultimately this tool has already accomplished its principal goal to accelerate the use of reasonably accurate big-data to support data-driven decision making to improve the safety of vulnerable roadway users in Maryland. As of Summer 2021, the MDOT SHA has funded the expansion of the tool to support new requests from practitioners across the state.

REFERENCES

Alluri, P., Raihan, M.A., Saha, D., Wu, W., Huq, A., Nafis, S. and Gan, A., 2017. Statewide analysis of bicycle crashes.

Arévalo-Támara, A., Orozco-Fontalvo, M. and Cantillo, V., 2020. Factors Influencing Crash Frequency on Colombian Rural Roads. Promet-Traffic & Transportation, 32(4), pp. 449-460.

Bao, J., Liu, P., Yu, H. and Xu, C., 2017. Incorporating twitter-based human activity information in spatial analysis of crashes in urban areas. Accident analysis & prevention, 106, pp.358-369.

Cambridge Systematics (2010). Travel Model Validation and Reasonableness Checking Manual—Second Edition (No. FHWA-HEP-10-042).

Fabozzi, F.J., Focardi, S.M., Rachev, S.T. and Arshanapalli, B.G., 2014. The basics of financial econometrics: Tools, concepts, and asset management applications. John Wiley & Sons.

Fitzpatrick, K., Avelar, R.E. and Turner, S.M., 2018. Guidebook on Identification of High Pedestrian Crash Locations. Federal Highway Administration. Office of Safety Research and Development. URL: https://www.fhwa.dot.gov/publications/research/safety/17106/17106.pdf

Furth, P.G., 2020. Level of Traffic Stress (LTS) Criteria. Available at: http://www.northeastern.edu/peter.furth/research/level-of-traffic-stress/.

Hosseinpour, M.H., Prasetijo, J., Yahaya, A.S. and Ghadiri, S.M.R., 2012. Modeling vehicle-pedestrian crashes with excess zero along Malaysia federal roads. Procedia-social and behavioral sciences, 53, pp.1216-1225.

Jiang, X., Abdel-Aty, M., Hu, J. and Lee, J., 2016. Investigating macro-level hotzone identification and variable importance using big data: A random forest models approach. Neurocomputing, 181, pp.53-63.

Lee, J., Abdel-Aty, M., Choi, K. and Huang, H., 2015. Multi-level hot zone identification for pedestrian safety. Accident Analysis & Prevention, 76, pp.64-73.

Lee, J., Abdel-Aty, M. and Shah, I., 2019. Evaluation of surrogate measures for pedestrian trips at intersections and crash modeling. Accident Analysis & Prevention, 130, pp.91-98.

Lord, D. and Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation research part A: policy and practice, 44(5), pp.291-305.

Maryland State Government. 2021a. "Maryland Statewide Vehicle Crashes". Available at: https://opendata.maryland.gov/Public-Safety/Maryland-Statewide-Vehicle-Crashes/65du-s3qu.

Maryland State Government. 2021b. "Maryland Statewide Vehicle Crashes - Person Details (Anonymized)". Available at: https://opendata.maryland.gov/Public-Safety/Maryland-Statewide-Vehicle-Crashes-Person-Details-/py4c-dicf.

Mekuria, M.C., Furth, P.G. and Nixon, H., 2012. Low-stress bicycling and network connectivity. Mineta Transportation Institute Publications. Available at: https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?referer=https://scholar.google.com/&httpsredir=1&article=1073&context=mti_publications.

Raihan, M.A., Alluri, P., Wu, W. and Gan, A., 2019. Estimation of bicycle crash modification factors (CMFs) on urban facilities using zero inflated negative binomial models. Accident Analysis & Prevention, 123, pp.303-313.

Ramsey and Bell, 2014. Smart Location Database Version 2.0 User Guide. Available at: https://www.epa.gov/sites/production/files/2014-03/documents/sld_userguide.pdf.

Saad, M., Abdel-Aty, M., Lee, J. and Cai, Q., 2019. Bicycle safety analysis at intersections from crowdsourced data. Transportation research record, 2673(4), pp.1-14.

Sanders, R.L., Frackelton, A., Gardner, S., Schneider, R. and Hintze, M., 2017. Ballpark method for estimating pedestrian and bicyclist exposure in Seattle, Washington: Potential option for resource-constrained cities in an age of big data. Transportation Research Record, 2605(1), pp.32-44.

Strauss, J., Miranda-Moreno, L.F. and Morency, P., 2015. Mapping cyclist activity and injury risk in a network combining smartphone GPS data and bicycle counts. Accident Analysis & Prevention, 83, pp.132-142.

StreetLight (2020). StreetLight Volume Methodology & Validation White Paper. Version 2.0. October 2019.

Tiwari, G., Bangdiwala, S., Saraswat, A. and Gaurav, S., 2007. Survival analysis: Pedestrian risk exposure at signalized intersections. Transportation research part F: traffic psychology and behaviour, 10(2), pp.77-89.

Ukkusuri, S., Miranda-Moreno, L.F., Ramadurai, G. and Isa-Tavarez, J., 2012. The role of built environment on pedestrian crash frequency. Safety science, 50(4), pp.1141-1151.

United States Environmental Protection Agency (EPA). 2021. "Smart Location Mapping". Available at: https://www.epa.gov/smartgrowth/smart-location-mapping#SLD.

Xie, S.Q., Dong, N., Wong, S.C., Huang, H. and Xu, P., 2018. Bayesian approach to model pedestrian crashes at signalized intersections with measurement errors in exposure. Accident Analysis & Prevention, 121, pp.285-294.

Xie, K., Ozbay, K., Kurkcu, A. and Yang, H., 2017. Analysis of traffic crashes involving pedestrians using big data: Investigation of contributing factors and identification of hotspots. Risk analysis, 37(8), pp.1459-1476.

Zhang, L., Ghader, S., Darzi, A., Pan, Y., Yang, M., Sun, Q., Kabiri, A. and Zhao, G., 2020. Data Analytics and Modeling Methods for Tracking and Predicting Origin-Destination Travel Trends Based on Mobile Device Data. Federal Highway Administration Exploratory Advanced Research Program.

APPENDIX A – DETAILS OF DATA ANALYTICS

To reduce several of the previously mentioned limitations, this project leverages and integrates mobile device location data into a data-driven analytical and visualization dashboard (https://mti.umd.edu/sdi). Compared to the current state of the practice that relies on surveys and manual/automatic counts, leveraging this emerging big-data source has significant advantages:

1. it no longer requires costly and time-consuming surveys or counts collections;
2. mobile device location data draws evidence from a much larger sample, covering over 40% of the entire U.S. population;
3. unlike cross-sectional data (i.e. data taken only on one or a few days of the year), this data provides a continuous time series of pedestrian and bicycle movements (we propose to analyze a one-year time period) and covers the entire Maryland roadway network, including rural areas where traditional pedestrian/bicycle data is lacking;
4. the availability of the data source is in the entire U.S., not just Maryland, which would enable a fast and cost-effective transfer of this project to other states and jurisdictions throughout the nation.

This dashboard reports exposure and crash information for pedestrians, bicycles, and e-scooters at a microscopic level, as illustrated in Figure 1, below.

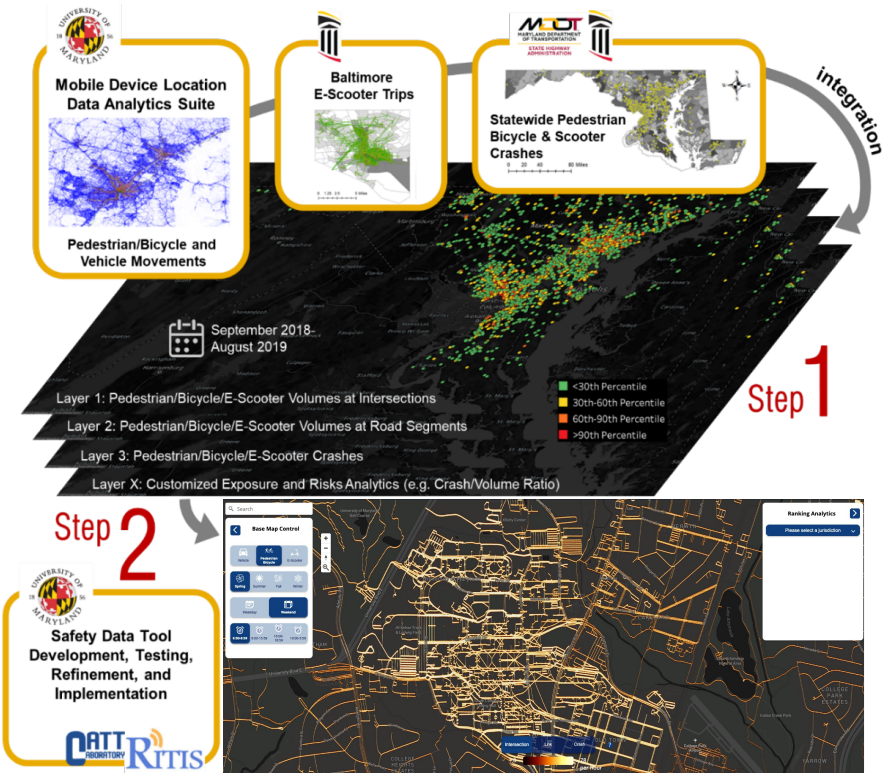

Figure 1. Data-Driven Safety Dashboard Deployment Framework for Maryland

The development of this safety dashboard is leading edge, though we note that the U.S. Department of Transportation (U.S. DOT) is involved in multiple efforts nationally within Pooled Fund Studies to generate greater data-driven situational awareness of the transportation system through location-based data (e.g. TPF-5(384)). Several of these national efforts use mobile device location-based data, which is collected via the mobile device's downloaded application (app) location pinging through GPS-enabled smartphones. This data source is capable of revealing multimodal mobility, including pedestrian and bicycle movements (Zhang et al., 2019). The tool aims to help agency users from various teams throughout all MDOT Offices identify safety risk hot spots and improve the situational awareness of existing pedestrian and bicycle activities as they tie to historical crashes. Users may also rank roadway facilities at the link level by volume, and predicted safety risks, or by a specific county, which would assist local county agencies with their respective priority rankings.

Analytical functions (e.g., subareas, sorting, filtering, etc.) are customized according to agency needs. During the project, the following activities and steps have fostered the iterative process in support tool development, refinement, and implementation:

1) Monthly team meeting series that brings together all team partners and US DOT OST. This has synchronized all team partners and collected feedback iteratively about the project.
2) Close engagement of key stakeholders across MDOT and local jurisdictions of the safety tool. When deploying the tool, stakeholders within the state were requested to review and comment on the design and analytical functions (e.g., subareas, sorting, filtering, etc.). The stakeholder comments were collected and addressed to develop the final tool.
3) Quarterly Pedestrian-Bicycle Emphasis Area Team (P-BEAT) meeting and quarterly CODES (crash outcomes data evaluation system) board meetings. These meetings have a broad audience who are stakeholders and potential end-users of the safety tool, working in the area of pedestrian/bicycle safety within Maryland.
4) Iterative feedback was also collected via two peer exchange events with other Safety Data Initiative project teams.

The following sections elaborate on the technical steps of data sources and assembly, data analytics, algorithms, visualization, as well as crash prediction models. Section 3 describes data sources being used in the development of the project tool. Section 4 discusses the methodologies employed in analyzing mobile device data and generating the final data products, i.e., measurements of pedestrian and bicycle volumes. Section 5 presents the models of crash frequency at intersections and roadway segments. These models are then used in the dashboard for predictions of safety risks.

Figure 2 shows the methodology flowchart of the dashboard. The project team estimated the vehicle and pedestrian bicycle volume at the link- and intersection-level leveraging a set of previously developed and validated algorithms made available via the USDOT Federal Highway Administration's Exploratory Advanced Research Program project entitled "Data analytics and modeling methods for tracking and predicting origin-destination travel trends based on mobile device data".

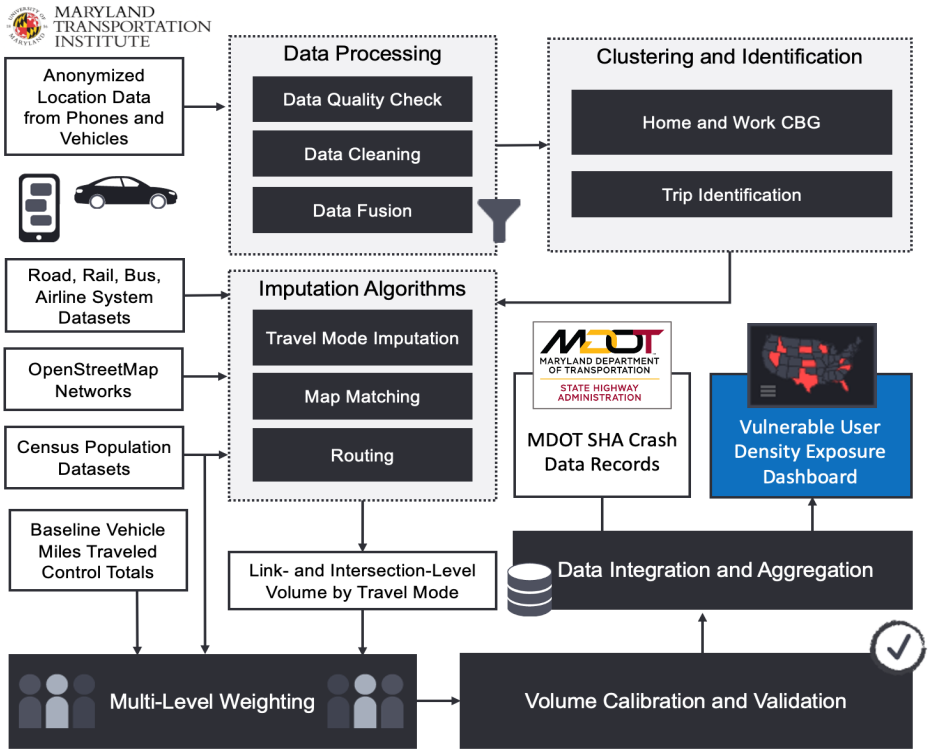

Figure 2. Methodology flowchart of the project.

Several key steps were taken in order to derive pedestrian and bicycle mobility data at each intersection and roadway segment:

- First, we employed a state-of-the-practice data preprocessing to integrate and clean the mobile device location data;
- The team then clustered location points into activity locations and identified home and work locations at the census block group (CBG) level to protect privacy;
- Next, we applied previously developed and validated imputation algorithms to identify all trips from the cleaned data panel, including trip origin, destination, departure time, arrival time. Most importantly, travel modes and routes were imputed so that pedestrian and bicycle activities can be inferred and aggregated to each intersection and roadway segment;
- Lastly, postprocessing steps were taken to expand our sample to all population, validate our estimated volumes, and visualize the mobility data on a large-scale and interactive map-based visualization platform embedded in the project dashboard.

Data Processing

Some common issues, such as unordered and duplicated records, need careful treatment before extracting any information from mobile device location data. The state-of-the-practice methods for raw data cleaning and quality control often include identifying and merging duplicate device observations, removing outliers, and checking on the obvious data consistency issues (e.g., devices with unreasonably high-speed readings). Figure 3 shows a general data cleaning procedure for mobile device location data taken by the research team based on the four dimensions of data quality assessment: consistency, accuracy, completeness, and timeliness.

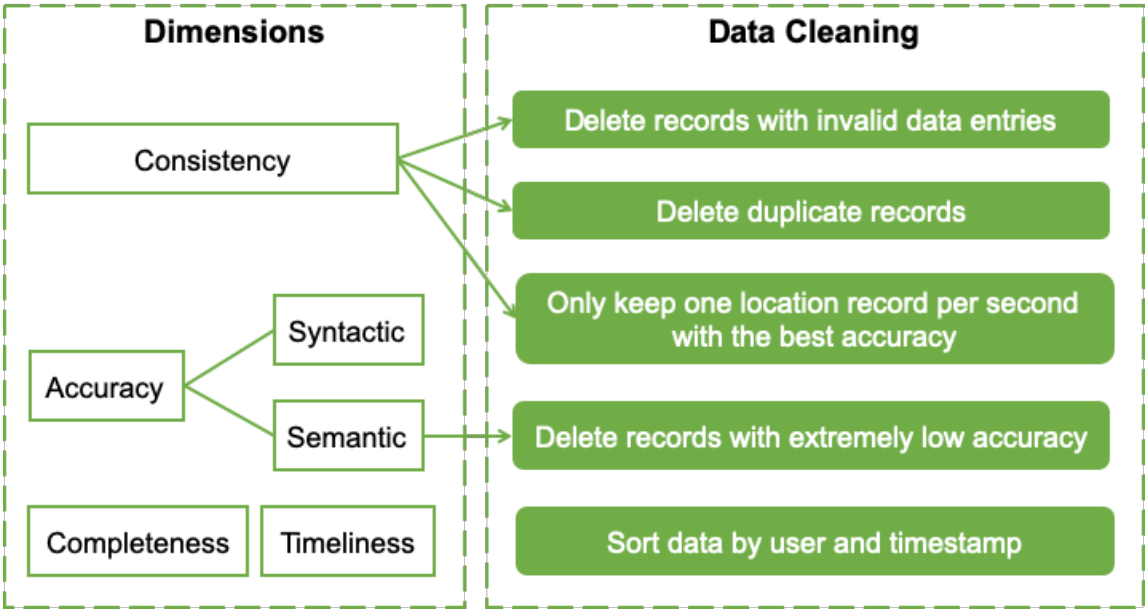

Figure 3. State-of-the-practice data cleaning procedure employed in the study.

The completeness dimension cannot be considered without prior knowledge of the actual individual movements and mobile device usage. The timeliness is addressed by using daily feeds of mobile device location data for our application. For the first two dimensions related to data cleaning, the consistency dimension defines certain semantic rules that a set of data items should obey. A common type of semantic rule is integrity constraints. For example, the latitude and longitude of a location observation should be within a reasonable range. According to the integrity constraints, the cleaning procedure first deletes records with invalid entries and duplicate records to reduce redundancy. Since one subject cannot be at more than one place at the same time, the procedure keeps only one location record per second (with the highest accuracy, if applicable). Another important dimension of data quality assessment is accuracy, including syntactic and semantic accuracy. The syntactic accuracy measures the closeness of a value to all the elements of its corresponding definition domain. The semantic accuracy measures the closeness of a value to its real-world value. For example, an accuracy of 10 meters in a location sighting indicates that the subject should be within a radius of 10 meters from the observed location with a certain confidence level, e.g., 95%. Therefore, the cleaning procedure removes the noisy records with extremely poor accuracy, e.g., two miles.

Location data providers describe their sample sizes with statistics such as daily active users (DAU) and monthly active users (MAU). MAUs are devices that are observed at least once a month and DAUs are devices that are continuously observed throughout the month. Reported data coverage by major data providers ranges between 5% to 70%, depending on whether they report MAU or DAU and how they define active users.

While the overall sample size is measured by daily and monthly active users, these measures do not take into consideration that some devices may provide many sightings every day while other devices may only provide a few sightings in a very small number of days. Table 1 presents more information about the mobile device location dataset used in this research. The following definitions describe the variables presented in the table:

- Daily population coverage: number of devices with identifiable home census block group (CBG) divided by the population of the study area.
- Temporal consistency: average number of days a device is observed in the study period.
- Frequency: the average location observations per device per day.
- Geographical representativeness: variance of population coverage among different zones of the study area, measured by a Gini coefficient between 0 and 1, with 0 indicating equal sampling rate in all zones and 1 indicating that all observed devices are from a single zone.
- Device representativeness: a measure of the variance in the location point frequency among observed devices. This measure shows if observed devices are comparable in terms of their data frequency and are also measured by a Gini coefficient falling between 0 and 1. Raw data representativeness has a lower value if all observed devices have more consistent data frequency.
- Hourly and daily temporal coverage: a measure of the variation of the number of location point observations among different hours of the day and different days of the month, respectively. Lower values between 0 and 1 indicate a more equitable distribution.

Table 1 Data quality of the mobile device location dataset

Selected Raw Data Quality Metrics	Mobile Device Location Dataset
Daily population coverage (%)	23.92
Geographical representativeness (0~1)	0.09
Frequency (observations per device per day)	190
Temporal consistency (days per device)	14.67
Device representativeness (0~1)	0.67
Hourly temporal coverage (0~1)	0.249
Daily temporal coverage (0~1)	0.03

Trip Identification

Trips are not initially included in any mobile device location data sources. Instead, location sightings are continuously generated while the sample device moves, stops, stays static, or starts a new trip. As a result, we developed a trip identification algorithm, which can detect which location

sightings form a trip together. We first sort device observations by time. The algorithm assigns a random ID to each trip it identifies. Many location points in the dataset may belong to no trips. The algorithm assigns "0" to the trip ID of these locations to tag them as static points. For every location point, we calculate distance, time, and speed between the point and its immediate previous and next points, if exist. Three hyperparameters need to be set for the algorithm: distance threshold, time threshold, and speed threshold. The speed threshold is used to identify if a location point is recorded on the move. The distance and time thresholds are used to identify stay locations and trip ends. At this step, the algorithm identifies the device's first observation with speed from≥speed threshold. This identified location point is recorded on the move, so a hashed trip ID is generated and assigned to this point. All points recorded before this point, if exist, are set to have "0" as their trip ID. Next, a recursive algorithm identifies if the next points are on the same trip and should have the same trip ID.

Then, a recursive algorithm has been developed to check every point to identify if they belong to the same trip as their previous point. If they do, they are assigned the same trip ID. If they do not, they are either assigned a new hashed trip id (when their speed from≥speed threshold) or their trip ID is set to "0" (when their speed from<speed threshold). Identifying if a point belongs to the same trip as its previous point is based on the point's "speed to", "distance to" and "time to" attributes. If a device is seen in a point with distance to≥distance threshold but is not observed to move there (speed to<speed threshold), the point does not belong to the same trip as its previous point. When the device is on the move at a point (speed to≥speed threshold), the point belongs to the same trip as its previous point; but when the device stops, the algorithm checks the radius and dwell time to identify if the previous trip has ended. If the device stays at the stop (points should be closer than the distance threshold) for a period of time shorter than the time threshold, the points still belong to the previous trip. When the dwell time reaches above the time threshold, the trip ends, and the next points no longer belong to the same trip. The algorithm does this by updating "time from" to be measured from the first observation in the stop, not the point's previous point. The algorithm may identify a local movement as a trip if the device moves within a stay location. To filter out such trips, all trips that are shorter than 300 meters are removed.

Results and computational algorithms have been validated based on a variety of independent datasets such as the National Household Travel Survey (NHTS) and the American Community Survey (ACS), and peer-reviewed by an external expert panel in a U.S. Department of Transportation Federal Highway Administration's Exploratory Advanced Research Program project (Zhang and Ghader, 2020). By running the algorithm on our data and comparing the trip lengths and travel times with the reported travel distances and travel times from the 2017 national household travel survey, a satisfactory match is observed. Moreover, we compared mobility trends calculated by our methods and by other data sources including Apple, Google, and SafeGraph and found high consistency.

Imputations of Home and Work Locations, Travel Modes, Routes, and Matching to Map

Home and Work Locations A typical methodology for identifying home and work clusters is to identify the most frequently visited clusters during the night and during the day. The algorithm first applies HDBSCAN clustering algorithm to clusters all device observations into activity locations. This step takes the cleaned multi-day location data as input and applies an iterative

algorithm until no cluster has a radius larger than two miles. The iterative algorithm consists of two parts: HDBSCAN based on a minimum number of point parameters and filtering non-static clusters based on time and speed checks. After finalizing the potential stay clusters, the algorithm combines nearby clusters to avoid splitting a single activity. Here, instead of setting a fixed time period for each type, e.g., 8pm to 8am as the study period for home CBG identification and the other half day for work CBG identification, the framework examines both temporal and spatial features for the entire activity location list. The benefits are two-fold: the results for workers with flexible or opposite work schedules would be more accurate and the employment type for each device could be detected simultaneously.

Travel Mode Imputation A machine learning model is developed and applied to impute the ground transportation travel modes for non-air trips, including vehicle (car and bus), rail, and walk/bike. More details are presented in the following sub-sections. Feature engineering directly affects the model performances. Three types of features (including a total of 32 features) are considered for ground transportation travel mode imputation, as shown in **Table 2**.

Table 2. Features for Imputing Ground Transportation Travel Mode Imputation

Features	Number of Features
Location Recording Interval Feature	
Average # of records per minute	1
Trip Features	
Origin-destination straight-line distance	1
Cumulative trip distance	1
Travel time	1
Average travel speed	1
0^{th}, 5^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th}, 100^{th} percentile travel speed	7
Multimodal Transportation Network Features	
0^{th}, 5^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th}, 100^{th} percentile distance to the nearest rail lines	7
0^{th}, 5^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th}, 100^{th} percentile distance to the nearest bus lines	7
Origin/Destination distances to the nearest rail station	2
Origin/Destination distances to the nearest bus stop	2
Percentage of records within 165-feet of all rail stations	1
Percentage of records within 165-feet of all bus stops	1

The Location Recording Interval (LRI) feature, represented by the average number of sightings per minute, indicates the location service usage during a trip. The trip features can show the characteristics of each trip, including the origin-destination straight-line distance, cumulative trip distance (network distance), travel time, average travel speed, and different percentiles of travel speed, which are all derived from our sighting data. The multimodal transportation network features are important to distinguish between different ground transportation travel modes. Here, the distance for each sighting to its nearest rail and bus lines are generated to calculate the 0^{th}, 5^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th}, and 100^{th} percentile distance to rail and bus lines; the distance for the origin/destination of each trip to its nearest rail and bus stations/stops are also calculated. Also, the percentage of records within 165 feet of all rail stations or bus stops are calculated for each trip.

The U.S. national bus and rail lines and bus stops and rail stations (including metro and Amtrak Stations) are collected from the Homeland Infrastructure Foundation-Level Data (HIFLD) and U.S. Department of Transportation Bureau of Transportation Statistics.

After comparing the performance of different machine learning models, the Random Forest (RF) machine learning model is selected as the final model to impute the ground transportation travel modes. The model is trained using over 11,000 sample data with labeled travel mode information. Synthetic Minority Over-Sampling Technique (SMOTE) is then applied to the training data to address the imbalanced sample problem, where the minority class from the existing samples is synthesized (Bohte and Maat, 2009). The randomized search approach is used to fine-tune the model. During the model training process, 10-fold cross-validation (CV) is conducted to evaluate the model performance. The training results show that the RF model can achieve 97.1% cross-validation accuracy for ground transportation travel mode imputation. The trips with the imputed four modes are further aggregated into three modes, including vehicle (car and bus), rail, and walk/bike.

Map Matching and Routing The team has developed and implemented a computationally efficient method to map crash and vehicle/pedestrian/bicycle/e-scooter movements data to an all-street roadway network (illustrated in Figure 4). This allows project engineers to evaluate the weighted vehicle/pedestrian/bicycle/e-scooter volumes at each intersection and each roadway segment. Through this project, the team has worked with MDOT SHA to incorporate other available safety data sources in the state of Maryland.

(a) **(b)**

Figure 4. Map Matching of (a) Crashes and (b) Mobility Movements at Intersection and Roadway Segment Levels

A spatial index method, KD-Tree, is first used to find all the roads within 328 ft (or 100 meters) for each sighting. The next step is to construct the complete path between all the sightings snapped to the road networks using routing algorithms. For each sighting, the team first compares its travel direction and the travel direction of its nearby roads within 328 ft. The closest candidate link with an absolute travel direction difference smaller than 30 degrees is selected as a valid match. Then, the path between the consecutive matched sightings is reconstructed by using the shortest path

algorithm based on road length. In the meantime, reasonableness checks are also conducted during the routing process. For each pair of consecutive sightings snapped to the network, the routed distance is first calculated by adding the length of all the road segments routed between the two sightings.

Then, two reasonableness checks will be conducted:

- If the routed distance is greater than the cumulative distance between the two observed snapped to the network by 1.24 miles or more, we consider the route as invalid and in need of revision.
- The travel time on these links will be calculated based on the timestamp difference of the two snapped sightings. With the routed distance and travel time, the average travel speed on these links can be calculated. If the speed exceeds 112 mph (180 km/h), we consider one of the two sightings is matched to the wrong link.

If either of these two violations is observed, we conduct an incremental approach by randomly removing one of the sightings, conduct the routing with the previous/next sighting snapped to the network, and examine the distance and travel speed until they do not violate the 1.24-mile threshold or the 112-mph threshold.

Postprocessing

Weighting The sample data needs to be expanded to produce population-level statistics of safety exposure for those vulnerable road users. The devices available in our dataset represent a sample of the population, so device-level weights are needed to expand the device sample. Also, for an observed device, only a sample of all trips may be recorded, so trip-level weights are needed as well. In order to obtain device-level weights, we have used the home locations, obtained from the imputed home CBG information. The weight for each device is equal to the number of devices observed in the device's imputed home location divided by the population of the zone where home is located. So, all devices residing in that CBG would have the same device-level weight. For instance, if our sample includes 100 devices in a CBG with a population of 2,000, each device would be assigned a weight of 20. For trip-level weights, we have calculated number of trips per person (trip rate) for each CBG during an average weekday from our sample. We have also calculated this trip rate number for each state from the 2017 National Household Travel Survey. We have used a single trip rate for all trips generated from each state, equal to the NHTS trip rate divided by our observed trip rate.

Validation The estimated vehicle and pedestrian/bicycle volumes are validated before visualization and being used for the project dashboard. For vehicle volume validation, the project team employed hourly traffic count data from the Maryland Department of Transportation State Highway Administration (MDOT SHA), including 1,933 portable traffic count stations (189,456 records) and 62 permanent traffic count stations (968,880 records). For pedestrian bicycle volume validation, the project team collected 15-minute interval pedestrian/bicycle count data from MDOT SHA that records the number of pedestrians and bicycles coming from each approach of an intersection. A total number of 845 count locations (89,500 records) were included in the pedestrian/bicycle count data. The pedestrian bicycle counts are further aggregated into 1-hour interval for validation.

The team measured the validation performance using two standard metrics that are widely adopted in the research and practice:

- R-Squared (R^2): A statistical measure that quantifies how much the dependent variable can be explained by the independent variables.
- Normalized Mean Absolute Error (NMAE): Defined as Mean Absolute Error (MAE) divided by the mean value of all observations.

The validation of vehicle volumes has a 0.98 R^2 and 10% NMAE. The validation of pedestrian/bicycle volumes has a 0.80 R^2 and 32% NMAE. The accuracy of the project estimated vehicle and pedestrian/bicycle volumes meets the Federal Highway Administration validation target (Cambridge Systematics, 2010) and is in line with the accuracy of other industrial data products (e.g., StreetLight, 2020). The accuracy of validation for pedestrian/bicycle volumes is lower than vehicle volumes. This is partly due to the data discrepancy. The best validation data we can collect is the counts of incoming pedestrians and bicycles from each approach of an intersection. Unlike vehicle trajectories which will more or less follow the roadway network directions, the fuzziness of pedestrian/bicycle trajectories near an intersection can be much higher. In other words, the mobile device location data covers more pedestrian/bicycle activities than volumes that are crossing the intersection. On the other hand, for intersections/segments where bus stops are located, certain pedestrian/bicycle trajectories could be categorized as bus trajectories and lead to underestimation. This limitation will be further explored. As an immediate next step, the team will work on additional data collection at roadway segments and intersections of all pedestrians and bicyclists for a more comprehensive validation and recalibration of the algorithms, if deemed necessary.

Visualization is an important final step for the project. The project dashboard needs an intuitive and yet effective visualization platform to engage its users and offer quick and accurate analytical insights.

With weighted and validated vehicle, pedestrian/bicycle trajectories and volume data, the team works on a proper visualization method to display the data to potential dashboard and data users. After extensive exploration, traditional vectorized visualization that are typically used in GIS, Tableau, or Mapbox applications is found not suitable. The data being visualized in this project is dramatically bigger in size and covers all roadway segments and intersections in Maryland, which leads to noticeable time lag when zooming in and out using the vectorized visualization.

Instead, the team has developed a rasterization method to effectively visualize the “big” mobility data generated from the analyses process. The team first defined multiple zoom levels for the maneuver of the visualization tool (in total, sixteen zoom levels have been defined). For each zoom level, a different scale is used to represent different level of visualization fidelity. Then, the visualization generates multiple tiles of figure files in the PNG format. For instance, at the zoom level 16, there are a total number of 2^16 * 2^16 tiles generated for visualization. Instead of doing vectorized visualization, each tile was discretized into pixels based on latitudes and longitudes (256*256 pixels are defined in each tile). We then used heat normalization based on

geoinformation of over 700 million activities (1.4 trillion location points in our vehicle and pedestrian/bicycle mobility data) to visualize the pixels. This process is summarized in Figure 5.

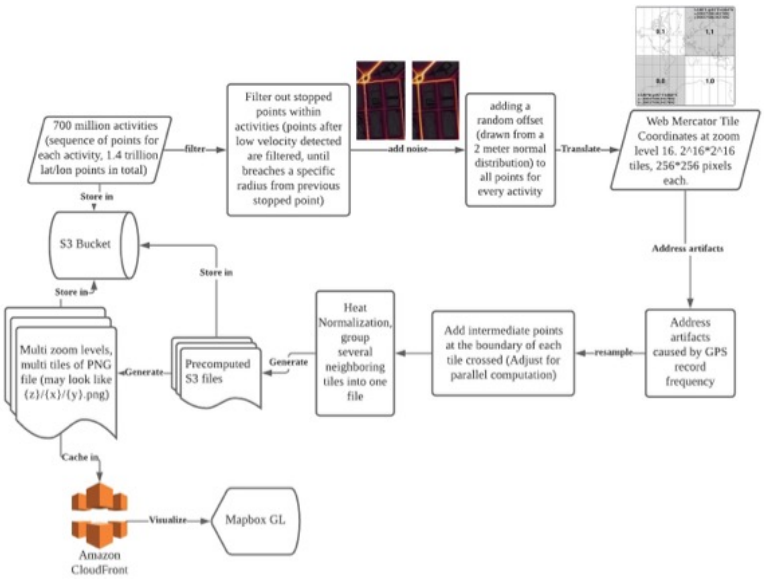

Figure 5. Rasterization Visualization Process

This rasterization method is superior in visualizing "big" data. It brings all the heavy computation offline, while traditional vectorization method has to perform rendering and vectorizing on the go. The team leveraged cloud-computing infrastructure and expertise based on Amazon CloudFront to parallelize the computation of the aforementioned large amount of data and finally visualize via Mapbox GL JS environment. Figure 6 illustrates an example of the rasterization at the zoom level 16 for College Park, Maryland. Pedestrian/bicycle data is visualized. All generated tiles are merged seamlessly to construct a high-fidelity map. The brighter color indicates higher level of pedestrian/bicycle volumes.

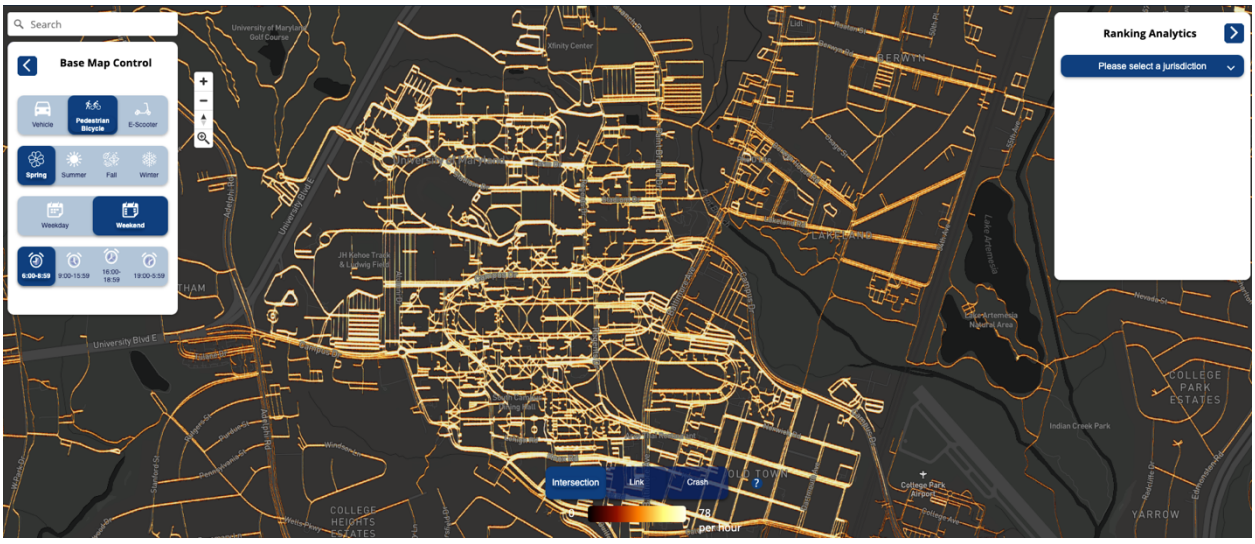

Figure 6. An illustrative example of rasterization visualization (College Park, MD is shown)

Final Data Product

The final data product of this step of mobility data analytics is two-fold. First of all, the safety exposure of pedestrians and bicyclists of Maryland in the year of 2019 has been developed and measured using the metrics of road segment level and intersection level volumes of these vulnerable road users. Second, such information is processed, visualized, and packaged on an online interactive platform (https://mti.umd.edu/sdi) for researchers and practitioners to use. These collectively address the research need and data gap in a reliable, comprehensive, and up-to-date information source on pedestrian/bicycle exposure, and convey such information via useful visualization tools.

APPENDIX B

Project White Paper:
Modeling the Frequency of Pedestrian and Bicyclist Crashes at Intersections: Big Data-driven Evidence from Maryland

Jina Mahmoudi, Ph.D., P.E.
Research Scientist
Maryland Transportation Institute (MTI)
Department of Civil and Environmental Engineering, University of Maryland College Park
1173 Glenn Martin Hall, College Park, MD 20742
Email: zhina@umd.edu

Chenfeng Xiong, Ph.D., Corresponding Author
Associate Research Professor and Assistant Director, Maryland Transportation Institute (MTI)
Department of Civil and Environmental Engineering
University of Maryland College Park
School of Medicine (SOM)
Shock Trauma Anesthesiology Research (STAR) Center
University of Maryland Baltimore
Email: cxiong@umd.edu

Mofeng Yang
Ph.D. Candidate and President of the ITS-ITE UMD Student Chapter, Maryland Transportation Institute (MTI)
Department of Civil and Environmental Engineering, University of Maryland College Park
1173 Glenn Martin Hall, College Park, MD 20742
Email: mofeng@umd.edu

Weiyu Luo
Ph.D. Student, Maryland Transportation Institute (MTI)
Department of Civil and Environmental Engineering, University of Maryland College Park
1173 Glenn Martin Hall, College Park, MD 20742
Email: wyl@umd.edu

Word count: 6,240 words text + 5 tables x 250 words (each) = 7,490 words

ABSTRACT

This study leverages big location-based service data collected from mobile devices in 2019 to conduct a pedestrian and bicyclist safety analysis. Statistical models are estimated for pedestrian and bicyclist crash frequency at Maryland intersections using the location-based service data as risk exposure data. The analysis is performed by employing prominent frequency modeling methodologies including Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial regression techniques.

The findings indicate that inclusion of big location-based service exposure data in the analysis improves the performance of the models. Further, the results suggest that key contributing factors to pedestrian and bicyclist crashes at Maryland intersections include: *i)* intersection design- and traffic-related attributes, such as number of intersection legs, presence of a traffic signal, average level of traffic stress rating, and safety risk exposure measures such as the average daily pedestrian, bicyclist, and vehicle volumes at the intersection; *ii)* travel-related attributes including public transportation and nonmotorized mode shares within the intersection's census block group; *iii)* land use and built environment attributes such as road network density, activity density, and extent of walkability within the census block group; *iv)* socioeconomic and sociodemographic attributes including the percentage of low-income workers, households with no vehicles, African American population, and senior population within the census block group.

The findings of the study show how big location-based service exposure data can be utilized to identify pedestrian and bicyclist safety risks and guide data-driven, evidence-based policy decision-making to improve the safety of vulnerable road users.

INTRODUCTION

Between 2015 and 2019, more than one hundred pedestrians and bicyclists were killed each year in the state of Maryland *(1–3)*. In 2019, 3,136 pedestrian crashes and 848 bicycle crashes occurred in Maryland, of which 2,872 pedestrian crashes and 695 bicycle crashes resulted in injuries or fatalities. This means that in 2019, over 90% of pedestrian crashes and over 80% of bicycle crashes in Maryland resulted in injury or death *(2,3)*. Further, approximately one out of every four individuals killed in traffic crashes in Maryland was a pedestrian *(2)*.

To combat these statistics and improve pedestrian/bicyclist safety throughout the state, a newly articulated vision was adopted in the 2019 Maryland Bicycle and Pedestrian Master Plan: *"Maryland will be a great place for biking and walking that safely connects people of all ages and abilities to life's opportunities"(4)*. The Master Plan lists its first safety objective as reducing the number of pedestrian/bicycle injuries and fatalities within Maryland's transportation system *(4)*. Thus, to realize the vision of the Maryland Bicycle and Pedestrian Master Plan, pedestrian- and bicyclist-involved injuries and fatalities must decrease throughout the state.

A crucial step towards that goal is to identify and gain a better understanding of the factors that pose a risk to the safety of pedestrians and bicyclists in Maryland. Many factors have been suggested to play a role in pedestrian and bicyclist crashes, including those representing pedestrian and bicyclist risk exposure *(5–10)*, land use and the built environment *(7,11–14)*, and sociodemographic/socioeconomic status *(7,11,12,14,15)*.

Among these factors, pedestrian/bicyclist exposure has been regarded as a critical factor in pedestrian and bicyclist crash analysis, the omission of which can lead to biased or overstated effects for the other factors *(16)*. Exposure pedestrian and bicycle data have traditionally been collected through surveys or count collections at sample locations *(13, 17,18)*. A few of the limitations of these conventional data collection methods include cost, time, accuracy, and subjectivity *(13)*. As a result, reliable pedestrian/bicyclist exposure data are often unavailable. For these reasons, high-quality and readily-available pedestrian and bicyclist exposure data are considered as a limitation in safety analysis *(16)*.

Meanwhile, the potential for emerging big (i.e., crowdsourced) data to provide accurate and reliable pedestrian and bicyclist risk exposure data has been realized in recent years *(9, 13,17,18)*. Nonetheless, current safety studies using consistent big data remain limited in number and scope, especially with respect to pedestrians. This is in part due to limited sources of crowdsourced data for pedestrians. Lee and Sener *(18)* categorized existing crowdsourced pedestrian and bicyclist data sources into two categories: 1) active data sources, for which travelers' active input is required (e.g., public bike-sharing programs, Strava); and 2) passive data sources, for which travelers' active input is not required (e.g., global positioning systems (GPS), location-based services (LBS)). They concluded that while actively crowdsourced bicycling data have many potential applications in research and practice, pedestrian data have not received the same attention, as limited sources are available for actively crowdsourced pedestrian data. Further, they stated that passively crowdsourced pedestrian and bicycle data are not currently available due to a high level of uncertainty and low locational precision *(18)*.

As exposure data are needed to contextualize crash analyses and prioritize countermeasures to lower safety risks *(13)*, utilization of high-quality and consistent exposure data is imperative in conducting pedestrian and bicyclist crash analysis. In other words, more comprehensive pedestrian and bicyclist crash analyses are those that use more reliable pedestrian and bicyclist exposure data, such as those provided through emerging crowdsourced services (i.e., big data).

To fill the important gap in understanding pedestrian and bicycle safety, crash frequency models have been developed in this study for pedestrian and bicyclists crashes at Maryland intersections by using emerging mobile-device location big data. This research was conducted as part of the Vulnerable Road User Density Exposure Dashboard project—a tool that utilizes big mobile device location data to provide data and insights on volumes and safety risk exposure of vulnerable road users (e.g., pedestrians, bicycles) at intersections and roadway segments within Maryland. This study is among the first to use consistent mobile-device location big data for both bicyclists and pedestrians in crash frequency analysis. The findings can assist policy decision-makers by providing big data-driven evidence and guiding potential solutions for vulnerable road users' exposure and safety.

LITERATURE REVIEW

Vulnerable road users such as pedestrians and bicyclists are susceptible to increased safety risks compared to other road users. Studies on pedestrian and bicyclist safety issues are abundant and identify key contributing factors to pedestrian- and bicyclist-involved crashes as well as suitable methodologies for crash frequency analysis.

To address the fundamental issues typically associated with crash frequency data, previous research studies have employed various methodologies to analyze pedestrian- and bicyclist-involved crash frequency. According to Lord and Mannering *(19)*, one of the main issues characterizing crash frequency data is overdispersion, which happens when the variance of the crash counts is considerably larger than the mean. The other issue that usually affects crash frequency data is having excess zeros, which happens when crash counts contain a significant number of zero values *(10,19)*.

To gain a better understanding of the factors that affect pedestrian/bicyclist safety and the methodologies applied to crash frequency data, a few studies are discussed in this section.

To predict bicycle crash frequency at intersections, Saad et al. *(9)* used bicycle crowdsourced data from Strava and developed a negative binomial model. The study found that the frequency of bicycle crashes at intersections were positively associated with intersection size, the intersection being a signalized intersection, the number of intersection legs being four (compared to three-legged intersections), as well as the risk exposure factors (i.e., total entering volume and bicycle volume at the intersection). The study results also indicated that the frequency of bicycle crashes at intersections was negatively associated with the presence of a bike lane at the intersection.

Raihan et al. *(10)* used a zero-inflated negative binomial model to develop crash modification factors for bicycle crashes in Florida's urban areas. The study found that road design characteristics such as lane width and speed limit had positive effects on reducing bicycle crashes. Lower bicycle crash probabilities on segments were associated with increased bicycle activity. However, increased bicycle activity was associated with higher bicycle crash probabilities at intersections. Increased bicycle crash probabilities at intersections were also associated with the number of bus stops within the intersection influence area.

Ukkusuri et al. *(11)* examined the role of various built environment, land use, road network, and sociodemographic factors as well as key exposure measures including traffic volume, transit ridership, and proportion of nonmotorized trip-makers in the frequency of total, injury-causing, and fatal pedestrian crashes. The study employed negative binomial and zero-inflated negative binomial regressions to develop crash frequency models and found that increased numbers of total and/or fatal pedestrian crashes were associated with increased

proportions of industrial and commercial land use, increased transit ridership, increased numbers of subway stations, increased proportions of intersections with four and five approaches, increased proportions of primary roads without access restriction, and increased number of lanes.

Sanders et al. *(13)* employed Poisson regression to examine the role of various factors in pedestrian exposure at intersections as well as bicycle exposure at various road segments in Seattle, Washington. The study found that variables representing population and land use (i.e., number of households, number of commercial properties, and the presence of a university near the intersection) were significantly associated with pedestrian exposure at intersections. Moreover, bicycle exposure was associated with the number of bicycle lanes on the road segment and land use variables such as the presence of a university or a school near the count location. The findings of that study provided insights into the factors affecting pedestrian and bicyclist risk exposure, which is a key contributing factor to pedestrian and bicyclist crashes.

Jestico et al. *(20)* used a crowdsourced bicycling incident dataset for the Capital Regional District in British Columbia, Canada, to identify design attributes associated with unsafe intersections between multi-use trails and roads. Negative binomial regression was used to model the link between the number of bicycle crashes and near-miss incidents and the infrastructure characteristics at multi-use trail-road intersections. The results showed that factors associated with bicycle incident frequency at multi-use trail-road intersections included bicycling volumes, vehicle volumes, and trail sight distance.

Many other studies also investigated factors affecting pedestrian and bicyclist safety risk exposure and modeled pedestrian- and bicyclist-involved crash frequency. The key contributing factors to pedestrian/bicyclist safety exposure and crash frequency that emerge from the literature include: sociodemographic and socioeconomic factors such as proportion of the population by race or age group *(7,11,12,14,15)*; land use and built environment factors such as population density, employment density, activity diversity, bus stop density, and ratio of residential, industrial, and commercial uses *(7,12–14)*; and traffic- and travel-related factors such as vehicle, pedestrian, and bicycle volumes as exposure measures *(5–10)*.

Further, the literature review reveals that the most prominent methodologies that have been applied to pedestrian and bicyclist crash frequency analysis are Poisson regression, negative binomial (NB) regression, zero-inflated Poisson (ZIP) regression, and zero-inflated negative binomial (ZINB) regression *(5,14,19–21)*. The Poisson regression is usually considered the starting point in crash frequency modeling *(5)*. Moreover, while the ZIP and ZINB regression methodologies have frequently been applied in empirical research to account for the preponderance of zeros observed in crash count data, the ZINB regression is applicable for count data that exhibit both overdispersion and excess zeros issues *(10)*.

Table 1 summarizes a few of the previous pedestrian and bicyclist safety studies.

TABLE 1 Examples of Past Studies on Pedestrian and Bicyclist Safety Models

Study	Unit of Analysis	Study Area	Safety Measure	Methodology	Key Exposure Measure(s)
Ukkusuri et al. 2012 *(10)*	Census tract, zip code	New York City (NYC), NY	Total pedestrian crashes, severe crashes, and fatal crashes	Negative binomial, and zero-inflated negative binomial models	Traffic volume, pedestrian activity, operating speeds

Hosseinpour et al. 2012 *(5)*	Road segment	Federal Road Network, Malaysia	Frequency of pedestrian crashes	Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial models	Motorized traffic volume
Lee et al. 2015 *(15)*	Zip code	Various locations in FL	Pedestrian crashes per crash location zip code, crash-involved pedestrians per residence zip code	Bayesian poisson lognormal simultaneous equations spatial error model	Log of population, log of vehicle miles traveled
Sanders et al. 2017 *(13)*	Intersection, Road segment	Seattle, WA	Pedestrian and bicyclist counts	Poisson model	— [a]
Jestico et al. 2017 *(20)*	Multi-use trail intersection	Capital Regional District, British Columbia, Canada	Frequency of bicyclist crash and near miss incidents	Negative binomial model	Bicyclists, vehicles, and pedestrian volumes
Xie at al. 2017 *(7)*	Grid cell (300×300 ft^2)	Manhattan (NYC), NY	Pedestrian crash cost	Tobit model	Vehicle miles traveled, taxi trips, subway ridership
Mansfield et al. 2018 *(14)*	Census tract	United States	Frequency of pedestrian fatalities	Negative binomial model, zero-inflated negative binomial model, zero-inflated negative binomial mixed model	Vehicle miles traveled density (te thousand VMT/mi^2) by roadway functional class
Saad et al. 2019 *(9)*	Intersection	Orange County, FL	Frequency of bicycle crashes	Negative binomial model	Total entering volume, bicycle volume
Raihan et al. 2019 *(10)*	Intersection, road segment	Urban areas, FL	Bicycle crash modification factors	Zero-inflated negative binomial model	Bicycle activity (Strava volumes)
Lee et al. 2019 *(21)*	Intersection	Orange and Seminole Counties, FL	Pedestrian crashes	Negative binomial, and zero-inflated negative binomial models	Observed and predicted pedestrian trips

Notes: — [a]: This was an exposure study; therefore, the exposure measures were the response variables in the models (i.e., pedestrian and bicyclist counts).

METHODOLOGY

Data

Various data sources were utilized in this study to analyze the frequency of crashes involving vulnerable road users, such as pedestrians and bicyclists at Maryland intersections. These data sources provide data on the key factors that, based on the literature review, contribute to the

occurrence of pedestrian/bicyclist crashes. The data sources used in this study are described below.

American Community Survey (ACS)
The American Community Survey (ACS) is an annual survey program conducted by the United States Census Bureau. Data collected through this survey provide information about the population (e.g., socioeconomic/sociodemographic characteristics, means of commuting to work) as well as housing (e.g., financial and physical characteristics for housing units) at many geographical scales.

The 2019 five-year ACS estimates at the census block group were used in this study as the source of socioeconomic and sociodemographic data in development of pedestrian/bicyclist crash frequency models.

Big Location-based Service (LBS) Data
The present study leverages large-scale, location-based service data, which were collected from anonymized mobile devices belonging to Marylanders. These anonymized mobile device location data are publicly available on the University of Maryland COVID-19 Impact Analysis Platform *(22)*, which was developed by the Maryland Transportation Institute (MTI). The project was partially funded by the U.S. Department of Transportation's Bureau of Transportation Statistics and National Science Foundation's RAPID Program.

The LBS data panel first integrated and processed locations for human movements in Maryland for the entire year of 2019. Then, a cloud-based computing platform was deployed using a multilevel weighting method, a spatial-temporal algorithm, as well as machine learning models to derive multimodal travel patterns. The data panel and the computational algorithms were validated based on a variety of independent datasets, such as the National Household Travel Survey and the American Community Survey, and peer-reviewed by an external expert panel from the U.S. Department of Transportation Federal Highway Administration's Exploratory Advanced Research Program project.

A map-matching and routing algorithm was also developed to assign the multimodal trips to the transportation network to estimate vehicle, pedestrian, and bicyclist volumes at intersections and links. In the present study, the LBS data was used to characterize safety risk exposure for vulnerable road users (i.e., pedestrians and bicyclists) by providing information on pedestrian, bicyclist, and vehicle volumes at intersections throughout Maryland.

Level of Traffic Stress (LTS)
Developed in 2012 *(23)*, Level of Traffic Stress (LTS) is a scale that rates a road segment based on the traffic stress it imposes on bicyclists. LTS ranges from 1 (for the lowest level of traffic stress) to 4 (for highest level of traffic stress). The four levels of LTS are described in more detail by Furth *(24)*. As a surrogate measure of pedestrian and bicyclist safety, the LTS was included in this analysis to quantify the traffic stress imposed on vulnerable road users at Maryland intersections.

National Transit Map (NTM)
Initially released in 2016, the National Transit Map (NTM) database is a product of the U.S. Department of Transportation's Bureau of Transportation Statistics. This nationwide database provides information on fixed-guideway and fixed-route transit service across the entire U.S.

including transit agencies' stops, routes, and schedules *(25)*. In this research, the NTM has been used to obtain information about transit stop locations.

National Walkability Index (NWI)
The National Walkability Index is a nationwide spatial data resource provided by the U.S. Environmental Protection Agency's Smart Growth Program. The National Walkability Index dataset ranks each U.S. Census block group according to its relative walkability *(26)*. The present study utilized walkability scores from this national dataset in the analysis of pedestrian and bicyclist crashes that occurred at Maryland intersections.

OpenStreetMap (OSM) Network
For the purpose of this study, road geometry and transportation network information, including the location of intersections, are extracted from the OpenStreetMap network, which provides open-source maps and map data for the world.

Smart Location Database (SLD)
First released in 2011, the Smart Location Database (SLD) is a nationwide spatial dataset for the U.S available through the U.S. Environmental Protection Agency. The latest version of this dataset is the SLD Version 2.0, which was released in 2013 *(26,27)*.

The SLD provides information on land use and built environment characteristics such as population and employment densities, land use diversity, urban design attributes, destination accessibility, transit accessibility, and socioeconomic/sociodemographic characteristics at the census block group level. These characteristics of SLD make it a suitable dataset for examining the role of land use and the built environment in the frequency of pedestrian and bicyclist crashes.

Vulnerable Road User (i.e., Pedestrian and Bicyclist) Crash Data
Data from the Maryland State Government's open data portal *(28)* was utilized in this study to obtain pedestrian and bicyclist crash data.

Analytical Methods
Different statistical models were developed and estimated in the present study to examine the role of various key contributing factors including safety risk exposure factors in pedestrian and bicyclist crashes that occurred at Maryland intersections. These models are the Poisson model, the negative binomial model, the zero-inflated Poisson model, and the zero-inflated negative binomial model. The results of these models were compared to identify the most suitable model that best fits the data. Due to paper length constraints, the mathematical formulations of these models are not reviewed here but can be found in various available sources—including Washington et al. *(29)*.

Pedestrian/Bicyclist Crash Frequency Models

Model Dependent Variable
The dependent variable for the statistical models is the frequency of pedestrian and bicyclist crashes at a particular intersection within the state of Maryland. The frequency of pedestrian and bicyclist crashes is a nonnegative count; therefore, the application of the four formerly-

mentioned regression models, which address nonnegative count data as a dependent variable, is appropriate *(see 29)*. Table 2 tabulates the frequency of pedestrian and bicyclist crashes in 2019 at Maryland intersections as extracted from the Maryland State Government's open data portal *(28)*.

TABLE 2 Frequency of Pedestrian/Bicyclist Crashes at Maryland Intersections in 2019

Number of Pedestrian and Bicyclist Crashes at the Intersection	Frequency	Percent
0	190,412	98.92
1	1,937	1.01
2	120	0.06
3	20	0.01
4	5	0.00
5	3	0.00
Total	192,497	100.00

The table shows that pedestrian/bicyclist crashes did not occur at a large proportion of intersections, leading to a data distribution that is positively skewed with many observations being zero. The preponderance of zeros is a common characteristic of count data that represent occurrence of an event (in this case, occurrence of a pedestrian/bicyclist crash at an intersection).

Model Independent Variables

The independent variables included in the models were selected based on the literature review and engineering judgement. These independent variables represent the key contributing factors that affect the occurrence of pedestrian and bicyclist crashes, including sociodemographic and socioeconomic factors, land use and built environment factors, and design-, traffic-, and travel-related factors—a few of which characterize the safety risk exposure for pedestrians and bicyclists at intersections. Table 3 provides information on the original independent variables considered for inclusion in the models.

TABLE 3 Independent Variables for Pedestrian/Bicyclist Crash Frequency Models

Variable	Description	Mean	SD	Data Source
Intersection Design- and Traffic-related Attributes				
Legs	Number of intersection approaches	5.61	1.26	OSM
Traffic Signal	Intersection is signalized – 1: yes, 0: no	0.04	0.19	OSM
Average Level of Traffic Stress (LTS)	Average of LTS rating for all intersection approaches	1.62	0.99	LTS
Average Daily Pedestrian/Bicyclist Volume	Average daily pedestrian/bicyclist volume passing through the intersection (January 1 to January 7, 2019)	84.93	210.40	LBS (MTI)
Average Daily Vehicle Volume	Average daily vehicle volume passing through the intersection (January 1 to January 7, 2019)	376.31	797.99	LBS (MTI)
Travel-Related Attributes				

Automobile Mode Share	Automobile commute mode share for CBG	84.61	13.22	ACS
Public Transportation Mode Share	Public transit commute mode share for CBG	6.12	8.60	ACS
Nonmotorized Mode Share	Walk/Bike commute mode share for CBG	2.40	5.44	ACS
Land Use and Built Environment Attributes				
Road Network Density	Total road network density for CBG	11.20	8.35	SLD
Pedestrian-oriented Network Density	Network density in terms of facility miles of pedestrian-oriented links per mile^2 of CBG	8.18	6.73	SLD
Multimodal Network Density	Network density in terms of facility miles of multimodal links per mile^2 of CBG	2.01	2.35	SLD
Intersection Density	Intersection density in terms of automobile-oriented intersections per mile^2 of CBG	1.21	2.83	SLD
Residential Density	Gross residential density (housing units/acres) for CBG	2.23	3.46	SLD
Employment Density	Gross employment density (jobs/acres) for CBG	2.34	10.38	SLD
Activity Density	Gross activity density [(employment + housing units)/acres] for CBG	4.58	11.64	SLD
Land Use Diversity	Employment and household entropy for CBG	0.50	0.22	SLD
National Walkability Index	Walkability index score for CBG	9.45	4.20	NWI
Number of Transit Stops	Count of bus stops within CBG	2.64	5.41	NTM
Sociodemographic and Socioeconomic Attributes				
Population Over 65	Percent of population ≥ 65 years old in CBG	16.63	9.27	ACS
Population Under 18	Percent of population < 18 years old in CBG	21.32	7.45	ACS
Male Population	Percent of the male population in CBG	48.58	6.46	ACS
African American Population	Percent of African American population in CBG	24.92	27.62	ACS
Enrolled in School	Percent of CBG population enrolled in school	25.01	8.83	ACS
Unemployed	Percent of unemployed population in CBG	3.15	2.96	ACS
Low-wage Workers	Percent of CBG workers earning ≤ $ 1250/month	21.61	4.75	SLD
Households with No Cars	Percent of zero-car households in CBG	6.46	10.45	SLD

Notes: CBG = Census block group; SD = Standard Deviation; MTI = Maryland Transportation Institute.

Pearson pairwise correlation coefficients were computed to examine the correlations between all original independent variables. To lower the risk of multicollinearity, highly correlated variables were not simultaneously included in the models. The final independent variables/models were selected by comparing the estimated models based on their Akaike's information criterion (AIC) and the Bayesian information criterion (BIC). Comparison of AIC and BIC is a common method of model selection. Various studies employed this method for model selection when analyzing pedestrian and bicyclist crashes *(7,9,32)*. The model with the smallest AICs and/or BICs is considered a more appropriate model among a set of candidate models *(30)*.

RESULTS AND DISCUSSION

Table 4 provides model estimation results for the models developed to relate the frequency of pedestrian and bicyclist crashes to various key contributing factors that affect crash occurrence at intersections. The following table shows several statistically significant associations between the dependent variable (frequency of pedestrian and bicyclist crashes at the intersection) and various independent variables representing key contributing factors that have been suggested by previous research to impact the occurrence and frequency of pedestrians and bicyclist crashes.

TABLE 4 Results of the Pedestrian/Bicyclist Crash Frequency Models

Dependent Variable: Frequency of Pedestrian/Bicyclist Crashes at the Intersection				
Independent Variable	**Type of Model**			
	Poisson	**NB**	**ZIP**	**ZINB**
Intersection Design- and Traffic-related Attributes				
Legs (*Reference: Number of Intersection Legs = 3 Legs*)	—	—	—	—
Number of Intersection Legs = 4	0.0604	0.0585	0.0093	0.0028
Number of Intersection Legs ≥ 5	0.3396***	0.3441***	0.2390***	0.2281***
Traffic Signal (*1: Signalized Intersection, 0: Otherwise*)	1.0143***	1.0241***	0.9123***	0.9351***
Average Level of Traffic Stress (LTS)	0.4517***	0.4331***	0.2239***	0.2217***
Average Daily Pedestrian/Bicyclist Volume	0.0004***	0.0006***	0.0003***	0.0004***
Average Daily Vehicle Volume	0.0002***	0.0002***	0.0001***	0.0001***
Travel-related Attributes				
Automobile Mode Share	0.0035	0.0035	0.0036	0.0036
Public Transportation Mode Share	0.0058*	0.0070**	0.0061*	0.0068**
Nonmotorized Mode Share	0.0058	0.0045	0.0078**	0.0071*
Land Use and Built Environment Attributes				
Road Network Density	0.0255***	0.0262***	0.0168***	0.0162***
Multimodal Network Density	-0.0163**	-0.0148*	-0.0091	-0.0083
Intersection Density	-0.0102	-0.0143**	-0.0072	-0.0086
Activity Density	0.0038***	0.0047***	0.0038***	0.0047***
Land Use Diversity	-0.1385	-0.0571	-0.0451	-0.0385
National Walkability Index	0.1157***	0.1080***	0.0688***	0.0687***
Number of Transit Stops	-0.0042	-0.0062	-0.0030	-0.0054
Sociodemographic and Socioeconomic Attributes				
Population Over 65 (%)	-0.0056*	-0.0060**	-0.0054**	-0.0053*
Population Under 18 (%)	-0.0033	-0.0047	-0.0046	-0.0051
Male Population (%)	-0.0044	-0.0040	-0.0034	-0.0031
African American Population (%)	0.0022**	0.0018*	-0.0002	-0.0002
Enrolled in School (%)	-0.0014	-0.0002	-0.0010	-0.0004
Unemployed (%)	0.0032	0.0050	0.0028	0.0035
Low-wage Workers (%)	0.0166***	0.0195***	0.0162***	0.0177***
Households with No Cars (%)	0.0066***	0.0076***	0.0067***	0.0072***
Model Goodness of Fit/Information Criteria				
Pseudo R^2	0.1968	0.1809	—	—
Akaike's Information Criterion (AIC)	20141.06	19952.39	19574.20	19466.10
Bayesian Information Criterion (BIC)	20395.25	20216.75	19848.73	19750.79

Dispersion Parameter (Alpha)	—	1.9938	1.3203	1.3203
Likelihood Ratio Chi^2 Test of Alpha = 0	—	chibar2(01) = 190.67***	—	—
Number of Observations (i.e., Intersections) = 192,497				

Notes: *, **, *** = Coefficient is significant at the 10%, 5% and 1% significance level, respectively; — = N/A.

Intersection Design and Traffic Attributes

The model estimation results indicate that the frequency of pedestrian and bicyclist crashes at an intersection is associated with intersection design- and traffic-related characteristics such as number of intersection legs, presence of a traffic signal at the intersection, average level of traffic stress (LTS) for the intersection, and average daily pedestrian/bicyclist and vehicle volumes at the intersection.

More specifically, frequency of pedestrian and bicyclist crashes at the intersection is positively associated with the intersection having a larger number of approaches (i.e., number of intersection legs ≥ 5), as compared to having three approaches (i.e., number of intersection legs = 3). This result is consistent with findings of previous research suggesting that a higher number of intersection approaches contributes to pedestrian and bicycle crashes *(9,11)*. One explanation for this finding could be that intersections with fewer approaches create fewer turning conflicts *(8)*. Another reason could be that intersections with more approaches may have higher vehicular and pedestrian/bicyclist volumes, and thereby are more prone to crashes.

Further, frequency of pedestrian and bicyclist crashes at the intersection is associated with the presence of a traffic signal at the intersection. This result is in line with results of past studies that found higher numbers of pedestrian and bicycle crashes as well as higher injury risk for bicyclists associated with signalized intersections *(6,9,21)*. With respect to pedestrians, findings by Tiwari et al. *(31)* can offer an explanation for the results obtained in the present study. The referenced study found that as signal waiting time increased at signalized intersections, pedestrians became impatient and violated the traffic signal *(31)*. Therefore, attempting to cross the intersection prematurely by pedestrians at signalized intersections may have contributed to the higher frequency of pedestrian-involved crashes at signalized intersections within the study area.

Average LTS for the intersection shows a positive correlation with frequency of pedestrian and bicyclist crashes at the intersection. This is a reasonable result considering that higher LTS ratings represent conditions that impose higher traffic-related stress on bicyclists including interaction with higher speed traffic, close proximity to high speed traffic, and multilane traffic *(24)*—all of which can contribute to bicyclist-involved crashes.

As expected, frequency of pedestrian and bicyclist crashes is also positively associated with the average daily pedestrian/bicyclist volume as well as the average daily vehicle volume passing through the intersection. These variables represent safety risk exposure for pedestrians and bicyclists in this study and the results corroborate past findings suggesting that increased frequencies of pedestrian- and bicyclist-involved crashes at intersections are associated with increased levels of risk exposure measures such as vehicle volumes, pedestrian volumes, and bicycle volumes *(9,18,20)*.

Travel Attributes

The travel-related characteristics of the census block group within which the intersection is located also play a role in frequency of pedestrian and bicyclist crashes. The results show that

increased frequencies of pedestrian and bicyclist crashes at the intersection are associated with the intersection being located in a census block group with higher public transportation and nonmotorized mode shares. One explanation can be that these alternative modes of travel may lead to increased risk exposure for pedestrians and bicyclists, which can in turn lead to increased crash frequency for such vulnerable road users. The results corroborate those of past research that found the proportion of commuters who travel to work by transit or nonmotorized modes were among the contributing factors to the number of pedestrian crashes *(11)*.

Land Use and Built Environment Attributes

The model results also provide evidence for the impact of land use and built environment characteristics on frequency of pedestrian and bicyclist crashes. Total road network density within the census block group has a statistically significant and positive association with the frequency of pedestrian and bicyclist crashes at intersections, whereas multimodal network density and intersection density within the census block group have negative associations with those crashes.

Road density has been previously found to be an important factor in pedestrian and bicyclist crashes *(12)*. The negative association between multimodal network density and frequency of pedestrian and bicyclist crashes (only statistically significant in Poisson and NB models) is a reasonable finding, which highlights the role of multimodal network designs in increasing safety of vulnerable road users. On the other hand, the negative association between the frequency of pedestrian and bicyclist crashes and intersection density (only statistically significant in the NB model) seems counter-intuitive, as previous research found that higher intersection density was associated with hotzones for pedestrian and bicycle crashes *(12)*. It should be noted, however, that this variable represents the density of automobile-oriented intersections within the census block group, and automobile-oriented facilities as defined in the SLD include facilities on which automobiles are allowed but pedestrians are restricted *(27)*. As these characteristics restrict pedestrian and bicyclist activities, they may lead to low or no pedestrian/bicyclist volumes at some intersections, thereby reducing the number of pedestrian and bicyclist crashes at those intersections.

Activity density within the census block group is another influential land use factor in pedestrian and bicyclist crashes. As expected, increased numbers of pedestrian and bicyclist crashes at the intersection are associated with the intersection being located in a census block group with increased activity density. This is in line with a previous study that found population density as a contributing factor to pedestrian and bicyclist crashes *(12)*.

The variable representing the National Walkability Index for the census block group also exhibits a positive and statistically significant coefficient in all four models. This indicates that increased frequencies of pedestrian and bicyclist crashes at the intersection are associated with increased walkability within the census block group. This is an expected result since higher extents of walkability could mean higher pedestrian activity, which can in turn mean higher risk exposure for pedestrians at intersections.

Sociodemographic and Socioeconomic Attributes

Socioeconomic and sociodemographic characteristics of the census block group also affect the frequency of pedestrian and bicyclist crashes at intersections. Based on the model results, higher numbers of pedestrian and bicyclist crashes at the intersection are associated with lower percentages of seniors (i.e., population over 65 years old) within the census block group. This

result can be due to decreased levels of activity such as walking trips for this age group, which can lead to lower levels of exposure and thereby lower numbers of pedestrian crashes. Other research has also found pedestrian crashes to be negatively associated with greater proportions of population over the age of 65 years *(11)*. Further, higher frequencies of pedestrian and bicyclist crashes at intersections are associated with a higher percentage of the African American population within the census block group. This result is only statistically significant in the Poisson and NB models; nonetheless, it is supported by past findings indicating that percentage of the African American population is an important crash risk factor for pedestrian and bicyclist crashes *(12)*. Further, based on the results, higher numbers of pedestrian and bicyclist crashes at the intersection are associated with increased percentages of low-wage workers and increased percentages of households with no private vehicle within the census block group. These findings can be an indication of higher usage levels of nonmotorized and public transit modes by individuals with a lower socioeconomic status. These alternative modes of travel may result in higher risk exposure, leading to higher numbers of pedestrian and bicyclist crashes at intersections. The results of the present study are consistent with those of past research that found increased income levels were associated with decreased numbers of pedestrian-involved crashes *(15,32)*. The results also corroborate previous findings indicating that increased proportions of households without a vehicle were associated with increased levels of pedestrian-involved crashes *(15)*, and higher percentage of households owning two or more vehicles were associated with higher risk for pedestrian and bicyclist crashes *(12)*.

Model Selection

To arrive at the model that offers the best data fit, several methods have been employed:

- The main assumption of the Poisson model, which requires the mean of the count variable to be equal to its variance *(29)* has been checked. The mean and variance of the pedestrian and bicyclist crash frequency variable in this study are 0.0118028 and 0.0141571, respectively. Therefore, the variance is larger than the mean, indicating presence of overdispersion in data.
- The likelihood ratio Chi^2 test (a postestimation test for the NB model indicating if the dispersion parameter alpha is equal to zero) has been performed. The result of this test is statistically significant (p-value < 0.0001), which suggests that the dependent variable is overdispersed and is not adequately estimated by the Poisson model.
- The AIC and BIC for all the models are computed and compared with each other. The comparison reveals that the ZINB model has the smallest AICs and BICs among all four models.
- Consideration has been given to the dispersion parameter (alpha). The dispersion parameter has been estimated by the NB, ZIP, and ZINB models to be greater than zero (1.9938 in the NB model and 1.3203 in the ZIP and ZINB models). Thus, due to the data being overdispersed, the model developed using these data is better estimated using the ZINB modeling methodology compared to the ZIP modeling methodology.

Thus, it seems that among the four models presented in Table 4, the ZINB model is the most suitable model for estimating the dependent variable (frequency of pedestrian/bicyclist crashes at Maryland intersections).

The ZINB model is subsequently used to predict the number of pedestrian and bicycle crashes at each intersection to assess crash risk. The model is highly capable of capturing the pedestrian/bicycle high-crash-risk intersections. As depicted in Figure 1, the intersections with the highest predicted risks cover major locations where the observed crashes occurred.

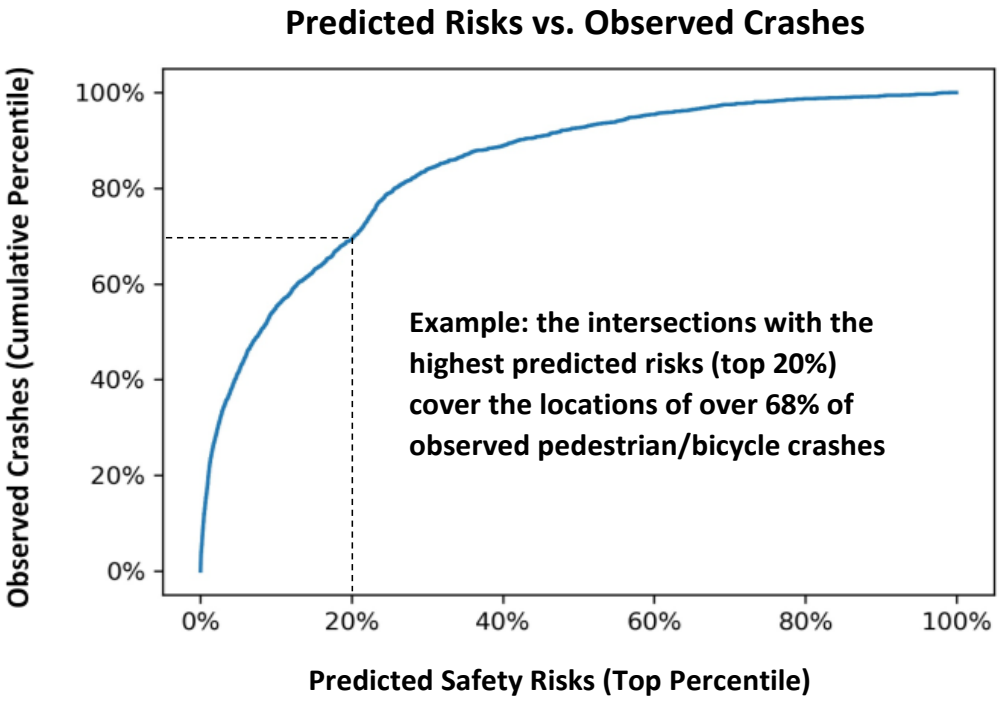

FIGURE 1 ZINB model performance.

Assessment of Contribution of the Big Location-based Service Data to Model Performance
To evaluate whether inclusion of the big Location-based Service (LBS) volume data contributes to improvement of the models, AIC and BIC of all four models have been computed for with and without LBS data scenarios. Table 5 summarizes these information criteria for all four models—with and without the variables representing the LBS volumes.

TABLE 5 Model Improvement Assessment Based on LBS Variables

Type of Model / LBS Variable (s) Included	Poisson Model	NB Model	ZIP Model	ZINB Model
	Information Criteria (AIC; BIC)			
Average Daily Pedestrian/Bicyclist Volume & Average Daily Vehicle Volume	20141.06; 20395.25	19952.39; 20216.75	19574.20; 19848.73	**19466.10; 19750.79**
No LBS Volume Variables	20407.44; 20641.30	20237.90; 20481.93	19651.22; 19905.41	19558.50; 19822.87

Comparison of the AICs and BICs (Table 5) indicates that for all four model types, the model that includes the LBS volume variables has the smallest AIC and BIC values. This means that the addition of the LBS volume variables is an improvement to the models. Moreover, the appropriateness of the ZINB model with the LBS variables is further emphasized by having the smallest AIC and BIC values among all model types.

CONCLUSIONS

This study leverages mobile device location data to conduct a data-driven pedestrian and bicyclist safety analysis by estimating the frequency of pedestrian and bicyclist crashes at Maryland intersections. The study employs the most commonly used crash frequency modeling methodologies including the Poisson, negative binomial (NB), zero-inflated Poisson (ZIP), and zero-inflated negative binomial (ZINB) regression techniques.

The ZINB outperformed the other methodologies and thus was concluded to be the most suitable methodology to model the count data in this study (i.e., the frequency of pedestrian and bicyclist crashes at a particular intersection).

Additionally, the results of the study provide evidence that increased frequencies of pedestrian and bicyclist crashes at Maryland intersections are associated with the intersection having greater than five approaches (i.e., number of intersection legs ≥ 5); the intersection being signalized; an increased average LTS for the intersection; and increased average daily pedestrian, bicyclist, and vehicle volumes passing through the intersection—which emphasizes the contributing role of safety risk exposure factors in pedestrian and bicyclist crashes. Results also indicate that increased frequencies of pedestrian and bicyclist crashes at intersections are associated with higher public transportation and nonmotorized commute mode shares, higher total road network density, higher levels of activity density, a greater extent of walkability, and higher percentages of a low-socioeconomic-status population within the census block group.

Moreover, the results show that inclusion of big location-based service (LBS) pedestrian/bicyclist and vehicle volume data in the models improves the performance of the models. As consistent and high-quality pedestrian and bicyclist exposure data is often regarded as a limitation in safety analysis *(9,13,16)*, this finding highlights the critical role of big LBS exposure data in contextualization of pedestrian/bicyclist crash analysis, where the main contribution of the present study also lies.

The present study has a few limitations, which can be addressed in future work. First, various other factors with the potential to impact the frequency of pedestrian and bicyclist crashes were not included in the analysis. Among such factors are vehicle speeds, parking availability, and pavement conditions. Further, a similar crash frequency analysis can also be performed for road segments. In addition, future research can conduct safety risk analysis for other vulnerable road users such as e-scooter users.

Nonetheless, the findings of this study contribute to the body of knowledge on safety by providing evidence on the role of big LBS exposure data in pedestrian/bicyclist safety analysis. These findings highlight the tremendous potential of this emerging source of big data, which offers many advantages, including elimination of costly and resource-intensive surveys and count collections and potential of generalization to other areas by providing data for the entire U.S.

Knowledge gained from the findings of this study can assist policy decision-makers in gaining a deeper understanding of the factors that contribute to pedestrian/bicyclist crashes and in developing more effective, data-driven safety interventions and policies to protect these vulnerable road users.

REFERENCES

1. Maryland Department of Transportation. https://zerodeathsmd.gov/wp-content/uploads/2021/06/StatewideBR-19Jan22-2021-1.pdf. Accessed July 9, 2021.
2. *2020 Maryland Pedestrian Safety Program Area Brief*. https://zerodeathsmd.gov/wp-content/uploads/2021/05/FFY21_Ped_ProgramAreaBrief_FINAL.pdf. Accessed July 9, 2021.
3. *2020 Maryland Bicycle Safety Program Area Brief*. https://zerodeathsmd.gov/wp-content/uploads/2021/05/FFY21_Bicycle_ProgramAreaBrief_FINAL.pdf. Accessed July 9, 2021.
4. *2040 Maryland Bicycle and Pedestrian Master Plan 2019 Update*. https://www.mdot.maryland.gov/OPCP/BW_2019_01_08MDOT_Final_Version_High_Res_with_Page_Borders.pdf. Accessed July 10, 2021.
5. Hosseinpour, M.H., Prasetijo, J., Yahaya, A.S., Ghadiri, S.M.R. Modeling Vehicle-Pedestrian Crashes with Excess Zero along Malaysia Federal Roads. *Procedia-Social and Behavioral Sciences*, 53, 2012, pp. 1216–1225.
6. Strauss, J., Miranda-Moreno, L.F., Morency, P. Mapping Cyclist Activity and Injury Risk in a Network Combining Smartphone GPS Data and Bicycle Counts. *Accident Analysis & Prevention*, 83, 2015, pp.132–142.
7. Xie, K., Ozbay, K., Kurkcu, A., Yang, H., 2017. Analysis of Traffic Crashes Involving Pedestrians Using Big Data: Investigation of Contributing Factors and Identification of Hotspots. *Risk Analysis*, 37(8), 2017, pp.1459–1476.
8. *Guidebook on Identification of High Pedestrian Crash Locations.* FHWA-HRT-17-106. FHWA, U.S. Department of Transportation, 2018.
9. Saad, M., Abdel-Aty, M., Lee, J., Cai, Q. Bicycle Safety Analysis at Intersections from Crowdsourced Data. *Transportation Research Record,* 2673(4), 2019, pp.1–14.
10. Raihan, M.A., Alluri, P., Wu, W., Gan, A. Estimation of Bicycle Crash Modification Factors (CMFs) on Urban Facilities Using Zero Inflated Negative Binomial Models. *Accident Analysis & Prevention*, 123, 2019, pp.303–313.
11. Ukkusuri, S., Miranda-Moreno, L.F., Ramadurai, G., Isa-Tavarez, J. The Role of Built Environment on Pedestrian Crash Frequency. *Safety Science*, 50(4), 2012, pp.1141–1151.
12. Jiang, X., Abdel-Aty, M., Hu, J., and Lee, J. Investigating Macro-level Hotzone Identification and Variable Importance Using Big Data: A Random Forest Models Approach. *Neurocomputing*, 181, 2016, pp.53–63.
13. Sanders, R.L., Frackelton, A., Gardner, S., Schneider, R., Hintze, M. Ballpark Method for Estimating Pedestrian and Bicyclist Exposure in Seattle, Washington: Potential Option for

Resource-constrained Cities in an Age of Big Data. *Transportation Research Record*, 2605(1), 2017, pp.32–44.
14. Mansfield, T.J., Peck, D., Morgan, D., McCann, B.,Teicher, P. The Effects of Roadway and Built Environment Characteristics on Pedestrian Fatality Risk: A National Assessment at the Neighborhood Scale. *Accident Analysis & Prevention*, 121, 2018, pp.166–176.
15. Lee, J., Abdel-Aty, M., Choi, K., Huang, H. Multi-level Hot Zone Identification for Pedestrian Safety. *Accident Analysis & Prevention*, 76, 2015, pp.64–73.
16. *Synthesis of Methods of Estimating Pedestrian and Bicyclist Exposure to Risk at Areawide Levels and on Specific Transportation Facilities*. FHWA-SA-17-041. FHWA, U.S. Department of Transportation, 2017.
17. Yin, L., Cheng, Q., Wang, Z., Shao, Z.'Big Data'for Pedestrian Volume: Exploring the Use of Google Street View Images for Pedestrian Counts. *Applied Geography*, 63, 2015, pp.337–345.
18. Lee, K., and Sener, I.N. *Emerging Data Mining for Pedestrian and Bicyclist Monitoring: A Literature Review Report*. Safety through Disruption, National University Transportation Center (UTC) Program. 2017. https://safed.vtti.vt.edu/wp-content/uploads/2020/07/UTC-Safe-D_Emerging-Data-Mining-for-PedBike_TTI-Report_26Sep17_final.pdf. Accessed July 10, 2021.
19. Lord, D., and Mannering, F. The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives. *Transportation Research Part A: Policy and Practice*, 44(5), 2010, pp.291–305.
20. Jestico, B., Nelson, T.A., Potter, J., Winters, M. Multiuse Trail Intersection Safety Analysis: A Crowdsourced Data Perspective. *Accident Analysis & Prevention*, 103, 2017, pp.65–71.
21. Lee, J., Abdel-Aty, M., Shah, I. Evaluation of Surrogate Measures for Pedestrian Trips at Intersections and Crash Modeling. *Accident Analysis & Prevention*, 130, 2019, pp.91–98.
22. *COVID-19 Impact Analysis Platform*. MTI. https://data.covid.umd.edu. Accessed July 7, 2021.
23. *Low-stress Bicycling and Network Connectivity*. Mineta Transportation Institute Publications. https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?referer=https://scholar.google.com/&httpsredir=1&article=1073&context=mti_publications. Accessed July 10, 2021.
24. Furth, P.G. *Level of Traffic Stress (LTS) Criteria*. http://www.northeastern.edu/peter.furth/research/level-of-traffic-stress/. Accessed July 10, 2021.
25. *National Transit Map*. Bureau of Transportation Statistics. U.S. Department of Transportation. https://www.bts.gov/content/national-transit-map. Accessed July 9, 2021.
26. *Smart Location Mapping*. United States Environmental Protection Agency. https://www.epa.gov/smartgrowth/smart-location-mapping#SLD. Accessed July 10, 2021.
27. *Smart Location Database Version 2.0 User Guide*. https://www.epa.gov/sites/production/files/2014-03/documents/sld_userguide.pdf. Accessed July 10, 2021.
28. *Maryland Statewide Vehicle Crashes*. https://opendata.maryland.gov/Public-Safety/Maryland-Statewide-Vehicle-Crashes/65du-s3qu. Accessed July 10, 2021.

29. Washington, S.P., Karlaftis, M.G., Mannering, F. *Statistical and Econometric Methods for Transportation Data Analysis*. Chapman and Hall/CRC. 2003.
30. Fabozzi, F.J., Focardi, S.M., Rachev, S.T., Arshanapalli, B.G. *The Basics of Financial Econometrics: Tools, Concepts, and Asset Management Applications*. John Wiley & Sons. 2014.
31. Tiwari, G., Bangdiwala, S., Saraswat, A., Gaurav, S. Survival Analysis: Pedestrian Risk Exposure at Signalized Intersections. *Transportation Research Part F: Traffic Psychology and Behaviour*, 10(2), 2007, pp.77–89.
32. Bao, J., Liu, P., Yu, H., Xu, C. Incorporating Twitter-based Human Activity Information in Spatial Analysis of Crashes in Urban Areas. *Accident Analysis & Prevention*, 106, 2017, pp.358–369.