Data Fusion for Non-Motorized Safety Analysis

August 2021 | Final Report

Source: United States Department of Transportation. Preparing for the Future of Transportation: Automated Vehicles 3.0. October 4, 2018. Washington, DC. https://www.transportation.gov/av/3. Accessed February 3, 2020.

SAFE-D
SAFETY THROUGH DISRUPTION

Texas A&M Transportation Institute

San Diego State University
Leadership Starts Here

VT | Transportation Institute
Virginia Tech

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. 03-049	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Data Fusion for Nonmotorized Safety Analysis		5. Report Date August 2021 6. Performing Organization Code:
7. Author(s) Ipek N. Sener (TTI/TAMU) Silvy Munira (TTI/TAMU) Yunlong Zhang (TAMU)		8. Performing Organization Report No. Report 03-049
9. Performing Organization Name and Address: Texas A&M University Texas A&M Transportation Institute 3135 TAMU College Station, Texas 77843-3135 USA		10. Work Unit No. 11. Contract or Grant No. 69A3551747115/Project 03-049
12. Sponsoring Agency Name and Address Office of the Assistant Secretary for Research and Technology University Transportation Centers Program Department of Transportation Washington, DC 20590 United States		13. Type of Report and Period Final Research Report 14. Sponsoring Agency Code

15. Supplementary Notes
This project was funded by the Safety through Disruption (Safe-D) National University Transportation Center, a grant from the U.S. Department of Transportation—Office of the Assistant Secretary for Research and Technology, University Transportation Centers Program.

16. Abstract
This project explored an emerging research territory, the fusion of nonmotorized traffic data for estimating reliable and robust exposure measures. Fusion mechanisms were developed to combine five bike demand data sources in Austin, Texas, and the fused estimate was applied in two crash analyses. The research was divided into three sequential stages. The first stage involved developing and applying a guideline to process and homogenize available data sources to estimate annual average daily bike volume at intersections. The second stage was focused on developing and applying the fusion framework—demonstrating the efficacy of multiple fusion algorithms, including two novel mechanisms, suited to the data characteristics and based on the availability of actual counts. The analysis of actual and simulated data illustrated that the fusion methods outperformed the individual estimates in most cases. In the third stage, the fused data were applied in both macro (hot-spot analysis in block group level) and micro (individual safety-related perception) models in Austin to ascertain the significance of incorporating exposure in safety analysis. While the fusion framework contributes to the research in the field of decision fusion, the demand and crash models provide insights to help stakeholders formulate policies to encourage bike activity and reduce crashes.

17. Key Words Fusion, exposure, nonmotorized activity, demand models, safety analysis, crowdsourced data, Dempster Shafer	18. Distribution Statement No restrictions. This document is available to the public through the Safe-D National UTC website, as well as the following repositories: VTechWorks, The National Transportation Library, The Transportation Library, Volpe National Transportation Systems Center, Federal Highway Administration Research Library, and the National Technical Reports Library.

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 59	22. Price $0

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Abstract

This project explored an emerging research territory, the fusion of nonmotorized traffic data for estimating reliable and robust exposure measures. Fusion mechanisms were developed to combine five bike demand data sources in Austin, Texas, and the fused estimate was applied in two crash analyses. The research was divided into three sequential stages. The first stage involved developing and applying a guideline to process and homogenize available data sources to estimate annual average daily bike volume at intersections. The second stage was focused on developing and applying the fusion framework—demonstrating the efficacy of multiple fusion algorithms, including two novel mechanisms, suited to the data characteristics and based on the availability of actual counts. The analysis of actual and simulated data illustrated that the fusion methods outperformed the individual estimates in most cases. In the third stage, the fused data were applied in both macro (hot-spot analysis in block group level) and micro (individual safety-related perception) models in Austin to ascertain the significance of incorporating exposure in safety analysis. While the fusion framework contributes to the research in the field of decision fusion, the demand and crash models provide insights to help stakeholders formulate policies to encourage bike activity and reduce crashes.

Acknowledgements

This project was funded by the Safety through Disruption (Safe-D) National University Transportation Center, a grant from the U.S. Department of Transportation—Office of the Assistant Secretary for Research and Technology, University Transportation Centers Program.

We are thankful to the following agencies for their assistance in furnishing data and information required for this research: City of Austin Transportation Department, Texas Department of Transportation, and StreetLight Data, Inc.

We also thank the SAFE-D Center, particularly Dr. Sue Chrysler and Elisa Cuellar of the Texas A&M Transportation Institute (TTI) and Eric Glenn of the Virginia Tech Transportation Institute (VTTI), for their guidance and support in the management of project activities.

We would also like to acknowledge TTI researchers Mark Ojah and Hao Pang for preparing the trip generation and distribution steps of the four-step model, Boya Dai for processing the Strava data, Kyuhyun Lee (formerly with TTI) for conducting a review and evaluation of the crowdsourced data in nonmotorized travel, and Shawn Turner for his support and discussions during the execution of this project.

Finally, we would like to thank Dawn Herring of TTI and Michael Buckley of VTTI for editorial review and Frank Proulx of UrbanFoot for technical review of this report.

Table of Contents

List of Figures

List of Tables

Introduction and Research Approach

Background and Motivation

Despite many efforts to promote biking and walking in the United States in order to create healthy, sustainable, and equitable communities, the quest to reach the envisioned nonmotorized mode share is still an uphill battle in many cities. More alarmingly, although contributing to a comparatively low percentage of total trips, 13% in 2017 [1], bike and walk traffic accounted for a disproportionate share (around 19% in 2019) of the total fatal and serious injury crashes [2]. Acknowledging the exigency of the issue, safety advocates are pursuing efforts to develop evidence-based, data-driven strategies to reduce nonmotorized crashes. Although the literature is replete with studies evaluating various aspects of nonmotorized traffic safety, the majority suffer from a major limitation—the absence of nonmotorized demand or exposure data [3]. The incorporation of robust and reliable exposure estimates is instrumental to the orchestration of an efficacious crash analysis for nonmotorized traffic [4, 5]. However, the existing approaches of estimating exposure, be it through observed counts, models, or crowdsourced data, exhibit limitations in terms of spatial, temporal, or population representation, or often overall reliability. The fact that no individual data source or model is sufficiently adequate drives the research question of whether combining or integrating multiple data sources, or in simple terms creating a data fusion, can produce a better estimate of nonmotorized demand or exposure. Researchers in the motorized traffic domain, recognizing that no single data source can provide sufficient information to develop good transport models [6], are rapidly progressing their utilization of the fusion method; however, the method is yet to be explored for nonmotorized traffic research.

In light of this, this project endeavored to develop a fusion-based technique to combine multiple nonmotorized demand or exposure data. The objective was to explore, select, and customize fusion mechanisms that can accommodate the distinctive nature of the nonmotorized traffic activity data and/or model output, with an end goal of generating a better quality exposure estimate that can add value to nonmotorized safety analysis.

Overview and Research Approach

Overview of Fusion

The concept of data fusion is well established, and researchers across myriad disciplines, including transportation, have acknowledged its advantages to obtain comprehensive and rich information [7–11]. Due to numerous emerging applications of fusion approaches in both research and commercial environments, the concept is yet to reach an equilibrium of consistent terminology and standard tools [12]. For instance, given its interdisciplinarity in various application domains, the terms sensor, multi-sensor, data, and information fusion have been used in numerous research articles without much discrimination [13]. A generalized explanation of fusion is as follows: “The overall goal of data fusion is to combine data from multiple sources into information that has greater benefit than what would have been derived from each of the contributing parts” [14].

While the core notion is to generate improved information that cannot be achieved with a single source, the characterizations, framework, and applications of fusion vary and are often customized based on the field of application [15]. In the context of transportation planning, a fusion framework should be designed to facilitate automated or semi-automated decisions, rather than to be the goal or end result [16].

The key benefits of fusion include increased completeness and confidence, reduced ambiguity, and enhanced spatial and temporal coverage [17–19]. The process is also associated with challenges and limitations, both technical and institutional [20, 21]. Fusion often adds cost and complexity, generating the risk of actually producing a worse result than the most reliable single source, especially when combined with inaccurate sources [19].

Nonmotorized Traffic Data, Exposure, and Fusion

While exposure can theoretically be defined as a "measure of the number of potential opportunities for a crash to occur," in practice a wide array of facility-specific or areawide exposure measures have been used for nonmotorized safety analysis, such as bicycle or pedestrian crossing volumes, total number of bicyclists and pedestrians entering the intersection, and distance or time traveled [5]. However, data collection efforts (both household survey and sensor-based location count data) require resources (budget and time), and it is only feasible to collect data from limited locations and time periods. Bicyclists and pedestrian demand or exposure to crash risk can also be quantified through various modeling-based methodologies utilizing available short/continuous counts and household survey data, such as direct (facility) demand models, regional travel demand models, geographical information system (GIS)–based models that heavily use GIS tools and GIS-based measurements in determining activity levels, and so forth [5]. The selection of a modeling approach often warrants a trade-off consideration in terms of complexity or resource requirements (time, budget, staff, data, etc.) and accuracy or reliability of estimation [3].

In the last few years, researchers and transport planners have steered their attention toward emerging technology-based methods, such as Global Positioning System (GPS)–enabled smartphone apps, wearable tech, interactive websites, and bike-share systems, as a source of nonmotorized activity data. These crowdsourced or big data sources, characterized by their volume, velocity, and variety [22], offer great potential to understand the detailed spatiotemporal travel patterns of nonmotorized traffic at an unprecedented level of detail [23]. However, they are often blamed for lacking quality assurance and representativeness [24, 25] and considered inadequate without validation from an actual location count [26]. Therefore, rather than recognizing these activity data as a reliable exposure measure, researchers have sidelined the sources as a potential proxy solution [27] to provide complementary information regarding nonmotorized activity in an area. The above discussion of existing approaches of estimating exposure, be it the observed count, models, or crowdsourced data, underscores the promising potential of the fusion approach, which, when rightly adapted, can combine strengths and decrease uncertainties associated with individual sources in order to provide a better understanding of the trend and pattern of nonmotorized activity.

There is an abundance of research investigating numerous fusion frameworks. As discussed in various studies in the literature [e.g., 28–30], the selection, customization, and application of fusion methods for combining nonmotorized activity data sources require a thorough understanding of the inherent characteristics of the data sources and situations. The process also necessitates comprehending the characterization and application of the existing fusion mechanisms. Nonmotorized traffic activity data exhibit unique characteristics, in contrast to motorized traffic data, and the methods and steps for computing motorized traffic states cannot be directly adopted in nonmotorized traffic areas. The choice of fusion mechanisms for nonmotorized traffic data is likely to be dictated by two key factors: (a) varying temporal, spatial, and population representation and data structure of the sources; and (b) limited sample size. Regarding the first factor, nonmotorized traffic data sources are seldom at the same resolution (spatial, temporal, or population) and are most often generated in different data formats. This warrants processing of individual sources to reach a homogenized representation before any fusion effort. The processing steps and complexity depend on the format of the raw data and the scope of analysis. This issue, coupled with the second key factor (limited sample size), calls for eliciting appropriate fusion methods that can adapt and accommodate the characteristics of the nonmotorized traffic data sources considering the constraints and add value to practical application.

Although an in-depth literature review on nonmotorized traffic data and fusion approaches was conducted to provide a rationale for the selection and design of the fusion mechanisms adopted for this project, that information is not included herein to maintain brevity. The in-depth literature review will be available in an upcoming manuscript entitled "Understanding Nonmotorized Data and Fusion Mechanisms for a Better Demand/Exposure Estimate."

Research Objectives and Steps

The main objective of this research was to generate a robust exposure estimate—fusing multiple nonmotorized activity data sources and demonstrating the efficacy of the developed fusion mechanisms—to incorporate in crash analysis. The research was performed in three sequential stages:

1. Process nonmotorized traffic data sources at a given scope for the study area.
2. Develop and apply the fusion framework.
3. Apply fused estimate for crash analysis.

The tasks within these stages are described in detail in the following three sections of the report. The first section describes the design of a guideline for homogenizing multiple data sources under various scopes of estimation and then discusses the mechanism of processing or modeling multiple bike-related data to estimate demand in the study area. The second section presents the selection, proposition, and demonstration of the applicability and efficiency of multiple fusion algorithms utilizing both actual and simulated data. Two novel fusion methods accommodating the practical constraints of data availability and context are proposed. The third section presents the application of the fused estimate in both macro and micro crash analyses to the study area.

Process and Homogenize Nonmotorized Activity Data

Introduction

Because nonmotorized traffic data sources are generated in different formats and structures (as reported in Appendix A), it is imperative to process those data to obtain a homogeneous representation. In recognition of the fact that there is no clear guidance on how different data sources can be brought together to compute volume or exposure within a prespecified scale, a conceptual framework was designed to help outline the steps and aspects of processing and homogenizing multiple nonmotorized traffic sources of different structures. The key steps of the framework are:

Step 1: Select the study area.
Step 2: Gather available data sources and identify the data structure.
Step 3: Determine the scope of estimation.
Step 4: Analyze/model data sources to obtain homogenous representation.

The steps, as well as the premise for selecting the scope/unit of estimation for the case study, are explained in detail in Appendix A. The following sections summarize the methods and findings.

Methodology

For this project, the city of Austin was selected as the study area given its recent endeavors to adopt a holistic approach to increase safety and mobility for pedestrians and bicyclists and the Vision Zero initiatives [31]. With an area of 326 square miles, Austin accommodates a population of over 996,369 [32]. Downtown Austin, which is located on the north bank of the Colorado River, is the central business district of the city. The University of Texas (UT) at Austin, accommodating over 50,000 students, is located north of the downtown area. Although the eastern part of the city is flat, the western part contains some hilly terrain. The city is also home to several natural and man-made lakes. By its very nature, the city is diverse in terms of age, culture, income, and built-environment characteristics, and it has experienced a steep rise in the degree of socioeconomic spatial separation over the last few decades. Despite being heavily car dependent, especially in suburban neighborhoods, the city has observed a significant increase in bicycle commuters in the last few years. The city's strong commitment to its Vision Zero goals and long-term planning to improve nonmotorized infrastructure make Austin an excellent case study for this research. Although the data analysis and fusion framework outlined in this report are apt to accommodate both pedestrian and bicycle traffic, the case study concentrated on bike data only.

From the study area, five primary data sources relevant to bike activity were gathered: (a) actual bicycle volume counts (permanent and short count), (b) bicycle-sharing data, (c) National Household Travel Survey (NHTS) add-on data, (d) Strava data, and (e) StreetLight data. The sociodemographic and land-use data for building models were obtained from the American Community Survey (ACS), Austin Transportation Department, and other public data domains. The

selected scope of estimation was the annual average daily bike (AADB) volume at intersections (however, the framework developed in this study is scalable and can be applied at other geographic scales such as midblock locations). The rationale behind the selection was the fact that the metric is a widely accepted policy-relevant unit of demand representation for both planning and safety analysis [33]. Moreover, a large proportion of crashes occur at intersections in the study area [34]. Therefore, by providing insights into the intersection-level bicycle volume, this study could contribute to the efforts of city officials to design focused policies and effective safety implementation plans.

The five key datasets represented bike activity at various spatial, temporal, and subpopulation scales. Each of the raw data sources went through exhaustive processing to compute AADB at the intersections in the study area, which consisted of 2,518 intersections, both signalized and unsignalized. Actual observations from both short and continuous counts were processed to estimate AADB at 44 intersections, which served as the ground truth data to process and validate the models. Five demand models—whose outputs served as the inputs of the fusion framework—were developed, including (a) direct demand model (DDM), (b) four-step model, (c) bike-sharing model, (d) Strava model, and (e) StreetLight model. Appendix A provides details of the data processing and model development together with references to additional resources.

Findings

Because the output from the five models was to be used as the input for fusion, it was imperative to discern the distribution, skewness, and coverage of the AADB estimates. Overall, the AADB estimates from the models were found to be in the range from 0 to around 1,400. The intersections with high bike volume were generally concentrated in the downtown area. The estimates from the five demand models were consistent with the actual count data—illustrating a high concentration (maximum of 1,282 riders) of bike ridership in downtown Austin. As noted earlier, details of the model results as well as the spatial distributions of the AADB estimates from each of the five demand models are available in Appendix A. It is also important to note that since the locations of counts were identified based on the city's bicycle route map, the counts are drawn from locations within the limits of the map and do not include areas farther into the suburban regions.

The DDM generated estimates for the maximum number of intersections (2,518), where the minimum volume was 15 and the maximum was 1,398. The bicycle-sharing model estimated bike activity in 793 intersections located near the bike-sharing stations in the central regions. The four-step model provided estimates for 2,397 intersections. Strava and StreetLight data were available for 2,303 and 950 intersections, respectively.

Figure 1 presents the distribution of the estimated AADB from the five models. The figure shows that the volume distribution was right skewed. The majority of the estimates were found to be under 300 AADB. The mode of distribution for the bike-sharing model estimates was between 0 and 100 AADB. For the other four models, the peak was at the 100 to 200 AADB bins. Only a few observations were reported for AADB exceeding 800.

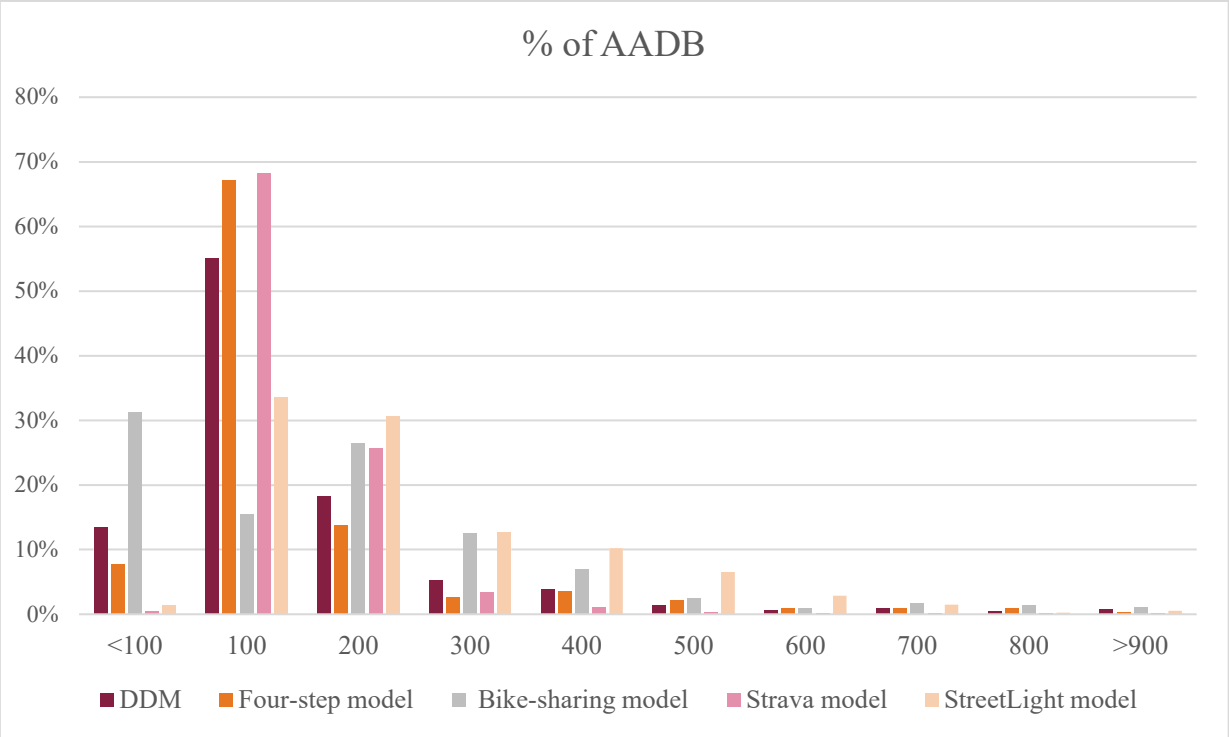

Figure 1. Chart. Frequency distribution of AADB estimates from five sources/models.

Develop and Apply Fusion Framework

Introduction

The application of the fusion concept in myriad areas has proliferated a wide range of terminologies, such as data fusion, sensor fusion, information fusion, evidence fusion, and decision fusion, which, although often used interchangeably due to the lack of standardized categorization and/or consensus, illustrate the distinct differences within the approach (e.g., theory, systems, frameworks, methods) [35]. Because there is no universal framework for fusion mechanisms, over the years researchers have proposed several generic and customized platforms accommodating the inherent attributes of the system. The task of designing and implementing a fusion framework is complex and warrants the comprehension of several aspects, including fusion architecture and algorithm selection, software implementation, and validation [36].

The literature review completed for this project (presented in detail in an upcoming manuscript "Understanding Nonmotorized Data and Fusion Mechanisms for a Better Demand/Exposure Estimate"), examined the characterization and inference process of the existing fusion practices, elucidating multiple aspects (e.g., data generation process, output type, sample size) that entail the choice and formulation of the plausible fusion mechanisms for nonmotorized demand. After discernment of the characteristics and dynamics of nonmotorized activity data and demand output, *decision fusion* was adopted in this project. Decision fusion is a mechanism of combining information wherein a decision from each source has been classified individually to obtain a

unified decision [37, 38]. It is a high-level fusion mechanism, often taking symbolic representations of events and activities, and used to obtain a more accurate decision accounting for the uncertainties and constraints [39]. The approach has been utilized for fusing motorized traffic data to estimate travel time, traffic state, origin-destination (OD) matrices, and more [10, 40, 41].

Fusion Algorithms for Nonmotorized Activity Data

The ultimate goal of any fusion is to produce estimates that are superior to the individual source estimates. The key operative word here is "superior," which implies completeness, accuracy, or a mix of both. While the completeness can be evaluated by comparing the coverage (for example, spatial coverage), the superiority of the accuracy can be demonstrated by showing that the accuracy of the fused estimate is higher than the best individual source. The domain of decision fusion is vast, and the algorithms vary across strength, weakness, and application [37, 38]. While there is no doubt a need for developing a robust state-of-the-art statistical approach for consolidating knowledge from various nonmotorized traffic data sources, it is also imperative to bring attention to the practical constraints and limitations of nonmotorized information. Therefore, during the development of the decision fusion framework for nonmotorized activity, two aspects were deemed crucial: (a) the type of output provided by the sources/models, and (b) the use of ground truth data.

Nonmotorized demand models are likely to generate knowledge or information as the abstract (crisp) level output, meaning ranking or confidence measures are not associated with the estimates [42]. Moreover, given the resource requirements of gathering on-site nonmotorized activity data, which should serve as ground truth or benchmark information for generating some fusion models, the information may not always be available. Thus, some practical scenarios may call for the option of fusion algorithms that do not demand ground truth information. In addition to these two issues, other features of nonmotorized information, such as incompleteness and conflicting cases, were also contemplated.

After the aforementioned issues were ascertained, multiple decision fusion mechanisms were adopted, which were then categorized into two types: (1) fusion without benchmark data, and (2) fusion with benchmark data. In the first category, three traditional fusion algorithms, under the rationale of the voting method, were illustrated. Then, a novel approach was proposed that uses the rationale behind the traditional weighted majority approach but generates decision weights without using the ground truth data. In the second category, the fusion framework applies a state-of-the-art statistical approach—the Dempster Shafer Theory (DST) method—which requires benchmark data but is adequate to accommodate the often incomplete and conflicting nonmotorized demand output with a mechanism of handling uncertainty involving ambiguity (or ignorance) and conflict [43]. The DST framework utilizes belief measures to allocate degrees of support to one or multiple hypotheses instead of one mutually exclusive outcome, as in the probability theory [43]. Unlike Bayesian methods, DST can handle decisions of different

granularities of classification at the same time and allow the modeling of ignorance and missing information [13, 44]. In this research, the basic DST was reinforced, incorporating a robust belief assignment method based on both a precision and recall matrix [45] and a conflict discounting mechanism based on Jousselme distance [46, 47].

A novel DST with context credibility is proposed to incorporate discounting factors based on contextual situations (such as spatial location, temporal scale, etc.) of an object or activity. The incorporation of "context" can be used to improve fusion outcomes and is deemed particularly important for the fusion of nonmotorized traffic data, which exhibit significant variation based on spatial location, as discussed in Munira and Sener [26]. Dey noted that "while most people tacitly understand what context is, they find it hard to elucidate" [48]. Dey and Abowd defined context as "any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and application themselves" [49]. The process of discerning context for traditional context-aware research can be based on automatically acquired information or be done manually. In real-world applications, context detection or sensing mostly relies on manual input (rather than an automated process) and requires profound understanding of the data dynamics and inherent situations [49]. The DST fusion with context credibility thus provides a valuable method for combining information in multiple data fusion application areas, as in the case of the current study, due to its ability to incorporate human subjectivity with mathematical probability [13].

Given the spatial variability of nonmotorized traffic data across the sources, the spatial feature of the estimates from each source was determined to serve as the context in the fusion algorithm developed in this study. Temporal variability is also essential in nonmotorized demand data but was beyond the scope of the current research due to the unavailability of relevant data to investigate the influence of temporal features as context. To the authors' knowledge, this research provides a first study within the decision fusion domain that incorporates the situation or context of sources, derived from subjective judgment, to refine the fusion algorithm. The contextual discount, based on Grey theory [50, 51], is also expected to add value to other areas of research in addition to nonmotorized traffic. This is especially the case when the sample size is a constraint and the potential for variability in reliability of the sources exists across one or multiple contexts of an entity or object.

While theoretically both the traditional and novel approaches are deemed well founded and promising, it is imperative to observe the performance of the approaches when they are applied to real data. It is also necessary to elucidate the application of the fusion framework to address the critical question of when or under what scenario the fusion methods are suitable for the application. Due to the unavailability of adequate ground truth data, the actual volume information (from 44 locations) was used to assess the volume estimates (from the five models explained in the previous section) for explaining the application and interpretation of the voting algorithms. In order to examine the performance and efficaciousness of the DST algorithms, simulated datasets conforming to real data characteristics and context were generated.

Details on the mathematical formalism and validation of the algorithms will be available in an upcoming manuscript entitled "Decision Fusion for Nonmotorized Traffic Data and Dempster Shafer with Context Credibility: Framework and Validation." A brief summary of the outcomes is presented in the following subsections.

Fusion Without Benchmark Data

The fusion mechanisms without benchmark data included four voting algorithms:

- Unanimity voting (a decision is made when all sources agree on the decision/class label).
- Simple majority voting (a decision is made when at least a majority of the sources agree on the decision/class label).
- Plurality voting (a decision is made based on the most voted label).
- Novel weighted voting (the proposed approach considers the pairwise interaction of the data sources to quantify the dissimilarity of sources from each other and assign weightage on each decision before fusion. This is basically the framework of weighted majority voting but without ground truth information. To measure the pairwise dissimilarity and compute the weight of each source, the concept of Euclidian distance was utilized).

Outputs from five sources (bike demand models) were fused using the voting algorithms. Since the decision fusion framework was developed within the classification problem domain, the first task in the process was to categorize or assign class labels to the estimates. In other words, the bike volumes were categorized in multiple categories denoted as class labels. Four class labels, reflecting the distribution of the datasets (Figure 1) and ensuring adequate observation in each of the categories, were assigned to the model outputs: < 100 AADB, 100 to 250 AADB, 251 to 400 AADB, and > 400 AADB.

Although the actual bicycle volume count data (from 44 locations) were utilized as the ground truth data, it would not be meaningful to consider this process as a validation exercise with reasonable confidence due to the limited sample size. The City of Austin was contacted in March 2021 to confirm the unavailability (as of that time) of updated counts for nonmotorized traffic. Moreover, in the ideal case (when the analysts have the option to separate the validation dataset from the training dataset), a separate validation dataset is desirable to conduct a true evaluation. Thus, this process was instead used to draw insights and gain understanding of how each of the models and the voting algorithms were performing. Table 1 presents the deviation of the individual model outputs from the actual bicycle volume count data (i.e., accuracy) along with their coverage. The DDM exhibited the highest accuracy, while the bike-sharing model had the lowest accuracy for the study area. Intuitively, a fusion algorithm can be considered effective when the accuracy is higher than that of the DDM (best individual source). In addition to the accuracy, it is also essential to examine the detection rate (or coverage) of each fusion method because the goal is to obtain high-accuracy estimates for the maximum number of intersections. Thus, it is essential to evaluate both the detection accuracy and overall accuracy of the algorithms.

Table 1. Accuracy and Spatial Coverage of the Sources

Model	Number of Intersections	Overall Accuracy (%)[a]
DDM	2,518	59
StreetLight Model	950	52
Strava Model	2,303	50
Four-Step Model	2,397	41
Bike-Sharing Model	793	27

[a] The accuracy is calculated as the percentage of correctly predicted samples in all categories.

Figure 2 presents the detection rate and accuracy among the detected AADB volumes of each voting fusion approach.

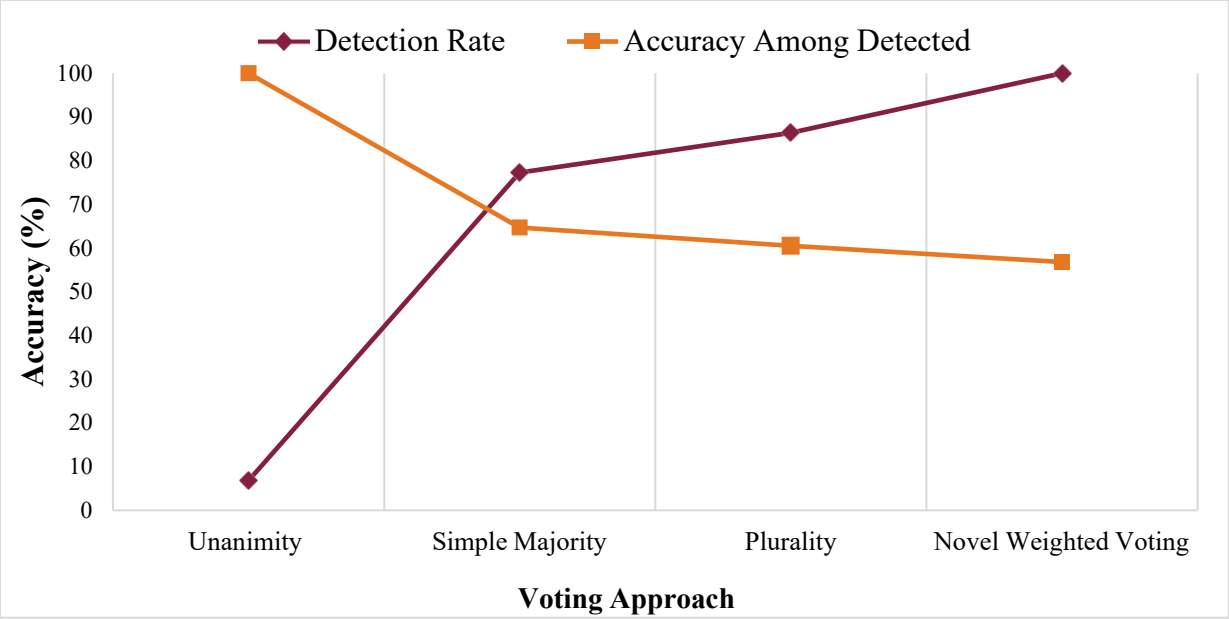

Figure 2. Chart. Detection rate and accuracy (among detected AADB) of the voting approaches.

The results indicated that the unanimity voting approach had the lowest detection rate (for the entire study area) yet the highest accuracy among the detected estimates. This finding is intuitive because to reach a decision for unanimity, each of the sources has to agree. The trade-off between detection rate and accuracy for the first four voting approaches is also made apparent in the figure. The novel weighted voting approach allocated AADB at all the intersections of the study area with a detection accuracy of 57%. Therefore, the evaluation suggests that the novel weighted approach exhibits decent performance considering the detection rate and the detected accuracy. However, if the analyst prefers accuracy over coverage, the unanimity and simple majority voting approaches may be regarded as potential options.

Based on evaluation of the overall performance of the approaches and comparison of each with the best individual source (as shown in Figure 3), the novel weighted voting approach exhibited slightly lower accuracy (57%) compared to the best individual model, the DDM (59%).

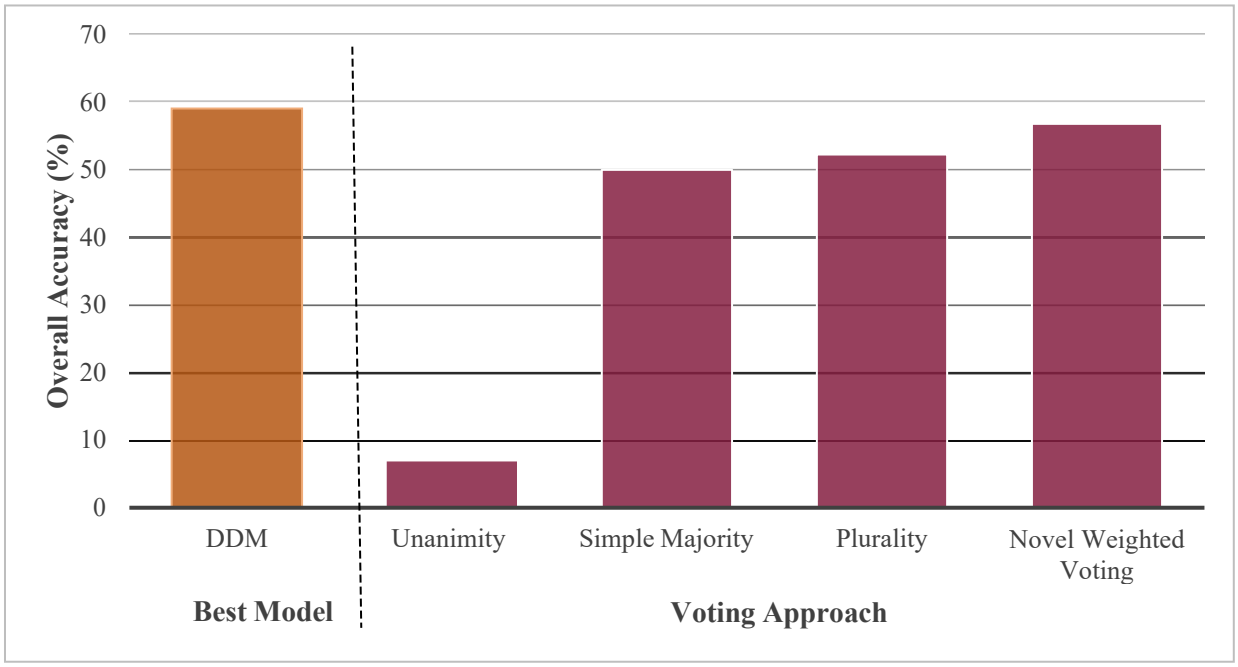

Figure 3. Chart. Overall accuracy of the voting fusion approaches.

In light of the limited sample of the validation data for the current case study, although the finding may not be intriguing from an accuracy point of view, the value of this approach can be highlighted, underscoring the fact that the method can be applied when the analyst does not have any knowledge of the accuracy of the individual sources. The application of the novel weighted method, although with no significant increase in accuracy, may instill confidence in the estimate, which is another objective of the fusion endeavor. Also, when applied to the simulated scenarios, as described in the next section, the novel weighted method outperformed the individual estimates for the first scenario and produced similar accuracy estimates, compared to the best individual source, for the third scenario.

The results of this application agree with the literature, which suggests that a fusion endeavor may not always outperform the individual estimates. Nevertheless, given the limited availability of the actual count data (or ground truth data), fusion for nonmotorized traffic data is more likely to depend on and align with the subjective discernment of the analyst. When an individual model or crowdsourced data fail to convey adequate confidence, the analyst can choose to select two or more sources to perform a fusion and select the estimate deemed the most reasonable. Thus, the main contribution of the approach is imparted from the fact that it provides the analysts an option to utilize their knowledge of the local community and the data to either use the individual estimates or opt for the fusion approach to obtain a better estimate with increased confidence.

Fusion with Benchmark Data

The DST approach was validated using simulated data since the size of the ground truth data were not deemed adequate to evaluate its performance. The purpose was to empirically test the performance of the DST approaches, both traditional and novel, by applying them on a set of artificially generated datasets with different intrinsic characteristics. The goal was to gain an understanding of which strategy, under what condition, performs the best. The simulated datasets also allowed for creation of multiple scenarios to examine the range of outcomes and understand the potential factors that contribute to the performance of the algorithm. Four distinct questions were formulated in the evaluation process. These questions were relevant to nonmotorized activity data characteristics and local situations to which the simulated scenarios were expected to respond:

1. Does the fusion algorithm always provide a better estimate?
2. Is more (number of sources) always better (for fusion estimates)?
3. Is the fusion estimate sensitive to the degree of context credibility discount?
4. Is the fusion estimate sensitive to categorization or class labels?

To address these questions, three main scenarios, each containing two sub-scenarios, were created. Each scenario was developed with three simulated datasets of varying accuracy and missing case situations. As discussed earlier, the context credibility proposed in this study depends on the subjective discernment of the analyst. During creation of the datasets, careful attention was given to instill contextual accuracy differences in the sources—with context captured through the spatial variability introduced by the location of the intersections. The reliability of the actual model estimates was reviewed with respect to their accuracy in and outside the downtown area, which showed notable variations based on the locations and an improvement in the downtown area across all sources. In an effort to replicate the actual data-based models, each of the simulated datasets was then divided into two sets, context 1 and 2, where the accuracy varied.

The three key scenarios were:

- Scenario 1: No missing observations, fairly similar accuracy across the sources
- Scenario 2: No missing observations, a varying level of accuracy across the sources
- Scenario 3: Each source has missing observations, a varying level of accuracy across the sources

For each sub-scenario, models were built for two categorizations or class labels, mainly to examine if the algorithms were sensitive to categorization type. Four class labels were assigned to the first category: < 100, 100–250, 251–400, and > 400. Twelve class labels were assigned to the second category: < 50, 50–100, 101–200, 201–300, 301–400, 401–500, 501–600, 601–800, 801–1,000, 1,001–1,200, 1,201–1,400, and > 1,400.

To evaluate and compare the performance of the DST algorithms, this research adopted a five-fold cross-validation method, maintaining a balanced trade-off between bias and variance while minimizing computation effort. For comparison, the partitioning was kept the same across all

individual sources and the fusion estimates. The final accuracy was computed by averaging the estimated accuracy of each fold. The scenario also assessed the sensitivity of the DST fusion-based context credibility, β, where β is the degree of discount due to context credibility estimated using Grey theory. Three values of β (0, 0.5, and 1) were tested, where the optimal value of $\beta > 0$ indicates that the proposed discount due to context credibility is effective and superior to the basic DST. Figure 4 presents the result of the scenario analysis using the DST method, reporting the accuracy gain/loss for optimal values of β.

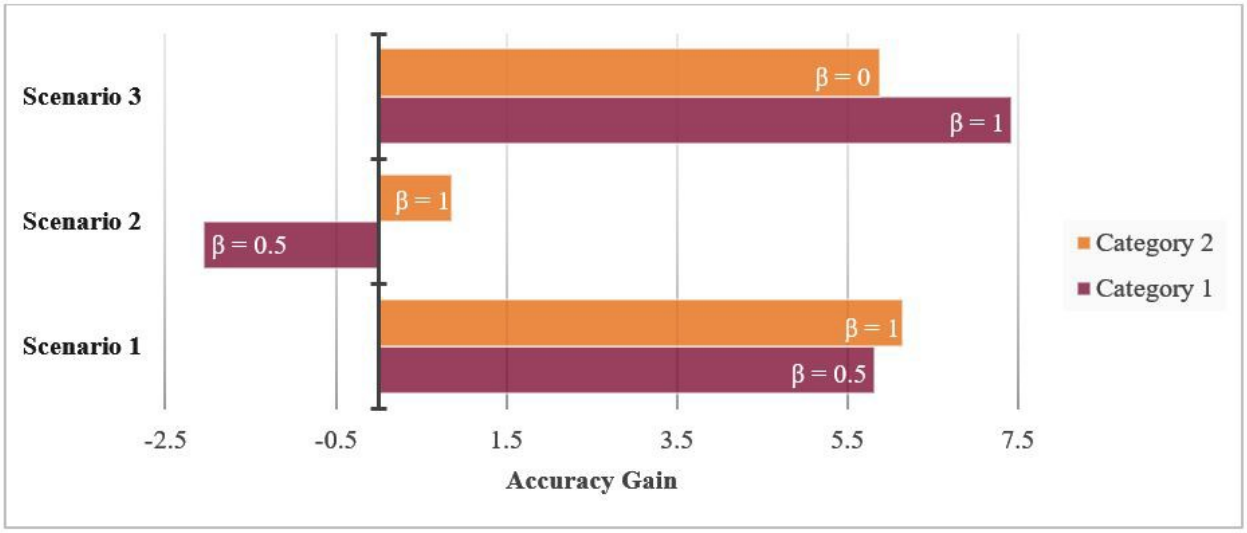

Figure 4. Chart. Accuracy gain or loss for three scenarios and two categorizations.

To examine if more is always better, the six scenarios were evaluated again after removing the lowest accuracy sources. The results showed that for Scenario 2, removal of the source that had considerably lower accuracy than the other two sources improved the accuracy of the fusion, meaning two-source fusion is better than both individual estimates and three-source fusion. Overall, the scenario analysis results suggest the following key takeaways:

- Fusion accuracy is sensitive to categorization type, meaning for a dataset, the fusion effort may yield an accuracy gain for one label but not for the others.
- The accuracy gain varies with the degree of conflict (λ) and context credibility (β).
- Even for the same dataset, the optimal values of λ and β may vary with the categorization label.
- Adding more sources may not necessarily yield improvements in the fusion. However, if there are missing values in the individual sources, even a comparatively lower accuracy source may add value to the fusion process.

Summary

The main objective of the research described in this section was to demonstrate the effectiveness of the fusion framework formulated for consolidating nonmotorized activity data. The novel weighted voting fusion algorithm proved to be an effective method of fusion, especially when the

analyst has no prior knowledge about the reliability of individual sources. In the scenario where local agencies do not have adequate actual count data but have access to various model outputs and crowdsourced datasets with enhanced coverage, voting approaches, both traditional and novel weighted, may be advantageous to obtain better coverage, thus instilling higher confidence in the estimates. This section also demonstrated the application of the DST mechanisms on the simulated data that conformed to real-world nonmotorized activity data characteristics. The most important finding is that in the majority of cases, the use of fusion methods outperforms the maximum individual source performance. The optimal value of $\beta > 0$ asserts the superiority of incorporating a contextual discount in the DST, as proposed by the study. In addition to nonmotorized fusion, the concept of context credibility can be used in other areas of fusion when deemed necessary.

Overall, the results of this numerical experiment lead to the conclusion that the performance of fusion firmly depends on the fusion method coupled with the data and situation characteristics. The findings presented in this section address the critical question of which fusion method, under what condition, can outperform individual estimates. There is no optimal categorization or β value that can accommodate all data situations. The future application of the fusion approach requires a deep understanding of the data, situation, and sensitivity of the fusion models for varying β to obtain the optimal combination for a given categorization. Thus, the findings of this research are expected to help analysts derive the best course of action regarding nonmotorized activity data fusion based on the available knowledge and local context.

Apply Fused Estimate for Crash Analysis

Introduction

This section presents the application of the fused exposure estimate (through a novel weighted voting approach) via the development of two separate crash models. The macro-level model aggregated the intersection AADB volumes into block groups to compute average zonal exposure with a goal of facilitating crash hot-spot analysis in the study area. The micro-level model examined individual safety perceptions of deterrents to biking and relating it to intersection bike activity (exposure) in the immediate neighborhoods.

Macro Crash Model

Background

Area- or macro-level safety analysis is pivotal to recognizing safety problems in a larger area to facilitate long-term policy planning to reduce crashes [52]. This research concentrated on examining traffic crash patterns at the block group level in Austin, recognizing statistically significant hot spots or high-risk locations for bicycle crashes. Among various crash evaluation measures [53–55], this research adopted the weighted crash rate by injury severity metric for two reasons. First, the metric complies with the crash score measure proposed by the City of Austin for identifying and prioritizing locations warranting safety treatments for pedestrians [56]. Second, it allows incorporating exposure related to bike volume. Although a handful of studies have

conducted hot-spot analysis for nonmotorized traffic [e.g., 57, 58], most have relied on associated vehicular volume or population measures, noting the unavailability of nonmotorized volume-related exposure data (see [55] for an example study incorporating crowdsourced Strava data as an exposure measure in a macro-level examination of nonmotorized crashes).

Data and Methodology

The traffic crash data for this study were taken from the Texas Department of Transportation's (TxDOT) Crash Records Information System (CRIS) [59] and included the disaggregated crash data, with coordinates, for all modes within the study area. CRIS also reports the severity of crashes (not injured, possible injury, non-incapacitating injury, incapacitating injury or suspected serious injury, and killed). For this study, the crashes involving bicycle traffic for an analysis period of 5 years (2014–2018) for each severity type were identified and aggregated into the block groups of Austin.

The weighted crash rate by injury severity (WCRIS) measure [60], as denoted in Equation 1, for each block group was estimated, allocating higher weight to serious injury and fatal crashes. The weights used in the WCRIS measure were adapted from the crash score measure proposed by the City of Austin Pedestrian Safety Action Plan [56]. The zonal (block group level) exposure was estimated by averaging the fused AADB for all intersections within the block group. To do so, the mid-value of each AADB category was taken and divided by the number of intersections in the zone.

$$WWWWWWWW = \frac{NNNN\ iiiiiiiiiii\ NNii\ ppNNpppppiipppppp\ iiiiiiiiiii*1+NNNNiiiiiiNNNNppNNNNiiNNNNiiiiiNN\ iiiiiiiiiii*1.5+SSiipppppNNNNppSS\ pppppiiiiNNiipp\ iiiiiiiiiii*1.75+FFNNNNNNppiiNNii*2}{\text{Zonal Exposure}} \quad (1)$$

The WCRISs in the block group were then fed into the optimized hot-spot analysis algorithm in ArcGIS, a spatial tool that uses the Getis-Ord Gi* algorithm [61] to identify statistically significant clusters of hot spots (high-concentration sites surrounded by other high-concentration sites) and cold spots (low-concentration sites near other low-concentration sites).

Findings

The analysis reported the hot spots and cold spots that were significant at the 99%, 95%, and 90% confidence levels. Appendix B exhibits where the high or low crash-risk locations were spatially clustered. The findings confirmed that hot spots generally appeared in the central downtown region of the city. There were also pockets of block group cold spots in the relatively outer area at the northern region (Cedar Park and Wells Branch) and southwest region (Barton Creek). Notably, much of the Austin region outside the downtown area lacked any statistically significant hot or cold spots.

Conclusions

Hot-spot analysis, albeit a simple and well-established practice, is still considered crucial given the fact that accurate identification of high-risk areas is essential for formulating engineering improvements to reduce crashes. Errors in hot-spot identification may lead to false negatives, meaning truly hazardous areas incorrectly designated as safe, as well as false positives, meaning

safe sites incorrectly designated as hazardous [54, 62]. The bicycle crash hot-spot identification method in this study, considering the local practice of allocating severity weightage and incorporating bicycle exposure measures, is expected to facilitate efficient resource management strategies to attain Austin's Vision Zero goals through long-term planning.

Micro Crash Model

Background

While the conventional approach to nonmotorized safety planning emphasizes the analysis of crashes at intersections or zonal areas, planners and policy makers are increasingly acknowledging the importance of proactively recognizing the factors affecting an individual's safety-related perceptions regarding biking and walking. Previous studies have already indicated that people's choice to walk or bike is likely influenced by both objective and perceived built-environment and traffic-safety-related aspects of the neighborhood [63–65], although perception often does not coincide with objective quantity [66, 67]. Perception is also likely to be affected by a number of personal attributes, such as income, gender, culture, norms, and experiences [68, 69]. An understanding of personal and environmental attributes influencing people's perceptions is crucial for formulating need-based intervention for long-term impact.

To investigate people's perceptions regarding safety-related impedances to biking, this research took advantage of the 2017 NHTS survey, which sought to obtain information about issues that influence the frequency of walking and biking [70]. The individual issues were associated with the built-environment features of the neighborhoods around individuals' home locations in addition to their demographic characteristics. Furthermore, the study specifically focused on the effects of bicycle exposure in terms of maximum AADB at a buffer zone around the household location. The rationale of the study stemmed from the hypothesis that an individual's subjective perception of the safety-related deterrent of biking may be formulated by the biking activity or exposure of their neighborhood.

Data and Methodology

The primary data source for this research was derived from the 2017 NHTS TxDOT add-on data. In the 2017 NHTS data, respondents who had at least one bike trip in the past week were asked which of the following keeps them from biking more: (a) No nearby paths or trails, (b) No nearby parks, (c) No sidewalks or sidewalks are in poor condition, (d) Street crossings are unsafe, (e) Heavy traffic with too many cars, (f) Not enough lighting at night, (g) None of the above, (h) I don't know, and (i) I prefer not to answer [71]. Among these, four issues (c, d, e, and f) were identified as safety-related reasons for not biking more and constituted the focus of the models.

To achieve the objective of the study, binary logistic regression models were developed, where the response of the individual (if a particular issue was the reason for not biking more) served as the dependent variable. To ensure adequate sample size, the four issues were aggregated into two groups for developing models: (1) street crossings unsafe, or heavy traffic with too many cars; and (2) no sidewalks or sidewalks in poor condition, or not enough lighting at night. In the Austin area,

a total of 1,095 households (2,185 persons) participated in this survey, among which 225 respondents (of age greater than 16) had at least one bike trip in the past week of the survey. While 95 of these respondents identified the first group of issues keeping them from biking more, 45 indicated not biking more because of the second group of issues.

In addition to various individual, household, and trip-related variables, the NHTS add-on also offers data with respondents' geocoded home address, which was used to generate buffer areas around each respondent's home location. The explanatory variables were developed from the built-environment characteristics for three buffer zones (0.1, 0.5, and 1 mi) around the household location and sociodemographic characteristics of each respondent. The exposure variable was generated from the fused AADB, indicating the maximum bike volume around the buffer scale of each individual's home location. The AADB was divided into three categories: low (less than 250), medium (251 to 400), and high (above 400).

Findings

Appendix C presents the results of the two binary logistic regression models. The results of both models indicate the importance of incorporating the bike exposure variables from both statistical and practical interpretation perspectives. The likelihood ratio test values for comparing the final models with the corresponding intercept-only models were much higher than the critical chi-squared value with 7 degrees of freedom at any reasonable level of significance. Also note that the log-likelihood value at convergence of the models using the DDM estimates as the bike exposure measures was lower than the log-likelihood value at convergence of the final models with the fused estimate. The log-likelihood ratio test values for comparing these latter models also indicate the superiority in fitness of the final models with the fused estimate.

The first model results suggest that people living in an area with high traffic volume and mixed land use were more likely to choose unsafe streets as the reason for not biking more. For the demographic variable, individuals who biked frequently (three times or more in a week) and used public transport (at least once in a month) identified the deterrent of unsafe streets significantly more than the other options. Also, the findings indicated the higher tendency of women toward not biking more due to unsafe street crossings and heavy traffic with too many cars. Interestingly, individuals living in lower bike traffic neighborhoods had significantly more complaints about unsafe streets than others. The second model had both similarities and differences with the first model, emphasizing the importance of better understanding individuals' perceptions in bicycling decisions. Regarding sidewalk and lighting conditions, reluctance to bike more was associated with a lower number of streetlights and higher occurrence of bike crashes. Individuals who walked for exercise more frequently or who used public transport (at least once in a month) were more likely to identify the issues related to neighborhood infrastructure as a reason for not biking more. Higher-income individuals ($100,000 and above) had significantly fewer complaints about sidewalk and lighting conditions as a deterrent for biking more frequently than others. Finally, the bike exposure variable highlighted significantly more complaints among individuals living in high bike traffic neighborhoods than individuals in medium- and low-traffic neighborhoods.

Conclusions

Overall, the best model was obtained with variables of different buffer levels. Both models suggest the significant influence of sociodemographic and built-environment characteristics on people's reasons for not biking more. Individuals who bike more frequently or who are more physically active are more likely to acknowledge the impedances to nonmotorized activity. Interestingly, respondents' subjective perceptions were consistent with the objective characteristics of their home neighborhoods, probably because the responses only included individuals who biked more or less frequently and had a better understanding of their surroundings. The notable conclusion of the model is that individuals' perceptions are associated with the location where they live. The direction of influence of the bike exposure variables provided important insights. The finding from the first model can be associated with the theory of safety in numbers [72, 73], suggesting that individuals feel safer when they are exposed to greater numbers of other bicyclists on the road in their neighborhoods. The finding from the second model can be related to individuals' increased sensitivity and demand toward infrastructure—in terms of better sidewalk conditions and adequate lightning—with increased level of cycling in the area.

While we have demonstrated example applications of crash models at the micro and macro levels, several other applications are also possible, such as development of micro-level crash models to identify risk factors at intersections [e.g., 74, 75].

Summary and Conclusions

The study presented in this report contributes to both research and professional practice by tapping the emerging research territory of fusion for nonmotorized traffic. A fastidious approach was undertaken to seek a generalizable fusion solution for nonmotorized demand computation. A major effort in this project was expended to understand the intrinsic characteristics of nonmotorized activity data and models and explore relevant and accommodative fusion mechanisms to facilitate safety-focused decision-making and infrastructure planning.

Because there was no clear guidance of how different data sources—including both traditional and crowdsourced—can be processed and brought together to compute nonmotorized exposure within a facility or area, a conceptual framework was developed for analysis. The bike demand models developed for this project not only illustrated the use of different datasets of varying forms and resolutions to bring into a homogeneous estimate at the micro level (intersection), they also shed light on the characteristics and aspects of bicycle activity. For example, the DDM developed for this research took the traditional approach to the next level by incorporating a "bikeability" index that can facilitate the modeling approach even when only limited count data are available. The model itself was reported as a reliable source of volume estimate for the entire study area, exhibiting high performance in terms of accuracy and goodness of fit. Moreover, some of the variables provided unique insights into bike travel behavior within the city, such as the significant and positive influence of the presence of bike signals and bike-accessible bridges. Additionally,

Strava and StreetLight data were examined, providing insights into the potential use of crowdsourced data in transportation studies, especially when resources are limited. In summary, the findings benefit stakeholders by explaining the determinants of bicycle activity within the region, thus providing guidance to formulate effective strategies, training, and educational programs geared toward creating a friendlier environment for bicyclists.

From a theoretical perspective, the DST fusion approach with credibility context, as proposed by this study, offers a unique way to incorporate the subjective judgment of experts in mathematical fusion formulation. The experiment on the simulated data, where the proposed approach outperformed the traditional approach in many scenarios, underscores the merit of the mechanism, not only for nonmotorized activity data analysis but also for application in other areas where an analyst's subjective judgment calls for considering context for belief refinement in the DST algorithm. The novel weighted approach is also expected to add value in fusion endeavors when no ground truth data are available. Nevertheless, the proposed fusion framework promotes data-driven safety analysis and informed planning while enhancing the strategic use of available information. Future research may explore other fusion algorithms, such as ensemble learning-based fusion, when more crowdsourced data are available. Also, this study considered the location of the intersections as context; based on a subjective judgment of local conditions, incorporation of additional context in the fusion, such as at the temporal scale, would be beneficial to investigate in future studies. Finally, the research demonstrated the applicability of a fused AADB estimate, both as a categorical variable and as a numeric estimate, taking the mid-value of each label. While the macro model set the stage for expanded analysis within the identified high-risk regions, the micro model provided insights into potential strategies to raise awareness through education and encouragement and to implement engineering measures to ascertain whether all residents feel safe and confident to bike more frequently.

Additional Products

The Education and Workforce Development and Technology Transfer products created as part of this project have been or will be located on the project page of the Safe-D website: Here.

The available datasets resulted from the final project have been or will be located in the Safe-D Collection of the VTTI Dataverse.

Education and Workforce Development Products

This project provided support to three students.

- Silvy Sirajum Munira served as the student researcher of the project, which constituted the core of her dissertation, entitled *Fusion with Context Credibility: Exploring and Fusing Nonmotorized Traffic Data*. Silvy graduated from the Civil & Environmental Engineering Department of Texas A&M University (TAMU) in summer 2021.

- Kyuhyun Lee is a former student researcher/research associate at the Texas A&M Transportation Institute (TTI). As part of the project, she examined the use and application of crowdsourced data. Kyuhyun has been accepted to a Ph.D. degree in the Department of Urban and Regional Planning at the University of Illinois at Urbana-Champaign. She is expected to start in fall 2021.

- Atom Arce is a former student intern at TTI and attended a five-week summer internship program designed/implemented in summer 2018 as part of this project. The main goal of the internship program was to provide an undergraduate student with expanded opportunities for guided learning. A detailed report was developed to describe the internship of Atom— (as of that time, i.e., summer 2018) a recent high school graduate (High School for Math, Science and Engineering at the City College of New York—Class of 2018) and newly admitted first-year undergraduate student (Fall 2018) at the University of Toronto.

The outputs of this research are expected to be helpful to researchers, academicians, and practitioners who are looking for a methodology to bring their data together and develop analysis/models using more reliable exposure estimates. In a similar vein, through the project's case study findings, the team expects to continue its activities to support the City of Austin in its efforts to improve nonmotorized safety and encourage safe walking and bicycling in Austin.

Technology Transfer Products

Four journal papers and two conference presentations resulted from this project. The research team is in the process of developing additional manuscripts to be submitted to peer-reviewed journals. A full list is available in Appendix D. In addition, the research team developed a slide deck (as a PowerPoint presentation) to incorporate materials and knowledge gained from this project into graduate courses/seminars, such as the graduate courses taught at TAMU, including Traffic Engineering, Engineering and Urban Transportation Systems, and Seminar. The team also developed a two-page project brief summarizing the project and presenting the key outcomes. The slide deck and the project brief will be available on the project page of the SAFE-D website.

Data Products

Appendix E provides relevant information on the data products of this project. The dataset can be found here DOI: 10.15787/VTT1/ZSJK4Z

References

1. Buehler, R. (2019). *A first look: Trends in walking and cycling in the United States 2001–2017*. Presented at the National Household Travel Survey (NHTS) Data for Transportation Applications Workshop. http://onlinepubs.trb.org/onlinepubs/Conferences/2018/NHTS/BuehlerTrendsinWalkingandCycling.pdf

2. National Highway Traffic Safety Administration. (2020). *Preview of motor vehicle traffic fatalities in 2019*. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813021

3. Turner, S., Sener, I. N., Martin, M. E., Das, S., Hampshire, R. C., Fitzpatrick, K., Molnar, L., Wijesundera, R. K., Colety, M., & Robinson, S. (2017). *Synthesis of methods for estimating pedestrian and bicyclist exposure to risk at areawide levels and on specific transportation facilities* (Report No. FHWA-SA-17-041). Federal Highway Administration.

4. Fitzpatrick, K., Avelar, R., & Turner, S. M. (2018). *Guidebook on identification of high pedestrian crash locations* (No. FHWA-HRT-17-106). Federal Highway Administration.

5. Turner, S. M., Sener, I. N., Martin, M. E., White, L. D., Das, S., Hampshire, R. C., Colety, M., Fitzpatrick, K., & Wijesundera, R. K. (2018). *Guide for scalable risk assessment methods for pedestrians and bicyclists* (No. FHWA-SA-18-032). Federal Highway Administration.

6. Picornell, M., & Willumsen, L. (2016). *Transport models and big data fusion: Lessons from experience*. Presented at the European Transport Conference 2016, Association for European Transport.

7. Wu, C., Thai, J., Yadlowsky, S., Pozdnoukhov, A., & Bayen, A. (2015). Cellpath: Fusion of cellular and traffic sensor data for route flow estimation via convex optimization. *Transportation Research Procedia, 7*, 212–232.

8. Lu, Z., Rao, W., Wu, Y. J., Guo, L., & Xia, J. (2015). A Kalman filter approach to dynamic OD flow estimation for urban road networks using multi-sensor data. *Journal of Advanced Transportation, 49*(2), 210–227.

9. Meng, C., Yi, X., Su, L., Gao, J., & Zheng, Y. (2017). City-wide traffic volume inference with loop detector data and taxi trajectories. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (p. 1). ACM.

10. Zhu, S., Guo, Y., Zheng, H., Peeta, S., Ramadurai, G., & Wang, J. (2016). *Field data based data fusion methodologies to estimate dynamic origin-destination demand matrices from multiple sensing and tracking technologies.* Nextrans.

11. Dailey, D. J., Harn, P., & Lin, P.-J. (1996). *ITS data fusion*. Washington State Transportation Commission.

12. Koks, D., & Challa, S. (2003). *An introduction to Bayesian and Dempster-Shafer data fusion*. Defence Science and Technology Organisation Salisbury (Australia) Systems Sciences Lab.

13. Wu, H. (2003). *Sensor data fusion for context-aware computing using Dempster-Shafer theory* (Doctoral dissertation, Carnegie Mellon University).

14. Gonsalves, P., Cunningham, R., Ton, N., & Okon, D. (2000, July). Intelligent threat assessment processor (ITAP) using genetic algorithms and fuzzy logic. In *Proceedings of the Third International Conference on Information Fusion* (Vol. 2, pp. THB1-18). IEEE.

15. Boström, H., Andler, S. F., Brohede, M., Johansson, R., Karlsson, A., Van Laere, J., Niklasson, L., Nilsson, M., Persson, A., & Ziemke, T. (2007). *On the definition of information fusion as a field of research* (No. HS-IKI-TR-07-006). Informatics Research Centre, University of Skövde. http://urn.kb.se/resolve?urn=urn:nbn:se:his:diva-1256

16. El Faouzi, N. E., & Klein, L. A. (2016). Data fusion for ITS: Techniques and research needs. *Transportation Research Procedia, 15*, 495–512.

17. Bachmann, C., Abdulhai, B., Roorda, M. J., & Moshiri, B. (2013). A comparative assessment of multi-sensor data fusion techniques for freeway traffic speed estimation using microsimulation modeling. *Transportation Research Part C: Emerging Technologies*, *26*, 33–48.

18. Brooks, R. R., & Iyengar, S. S. (1998). *Multi-sensor fusion: Fundamentals and applications with software*. Prentice-Hall.

19. Ince, A. N., Topuz, E., Panayirci, E., & Isik, C. (2012). *Principles of integrated maritime surveillance systems* (Vol. 527). Springer Science & Business Media.

20. Amey, A., Liu, L., Pereira, F., Zegras, C., Veloso, M., Bento, C., & Biderman, A. (2009). *State of the practice overview of transportation data fusion: Technical and institutional considerations* (Working paper). Transportation Systems.

21. Hall, D. L., & Steinberg, A. (2001). *Dirty secrets in multisensor data fusion*. Pennsylvania State Univ University Park Applied Research Lab.

22. Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note, 6*, 70.

23. Misra, A., Gooze, A., Watkins, K., Asad, M., & Le Dantec, C. (2014). Crowdsourcing and its application to transportation data collection and management. *Transportation Research Record: Journal of the Transportation Research Board, 2414*, 1–8.

24. Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics, 1*, 110–120. https://doi.org/10.1016/j.spasta.2012.03.002

25. Jestico, B., Nelson, T., & Winters, M. (2016). Mapping ridership using crowdsourced cycling data. *Journal of Transport Geography, 52*, 90–97. https://doi.org/10.1016/j.jtrangeo.2016.03.006

26. Munira, S., & Sener, I. N. (2020). A geographically weighted regression model to examine the spatial variation of the socioeconomic and land-use factors associated with Strava bike activity in Austin, Texas. *Journal of Transport Geography, 88*, 102865.

27. Saha, D., Alluri, P., Gan, A., & Wu, W. (2018). Spatial analysis of macro-level bicycle crashes using the class of conditional autoregressive models. *Accident Analysis & Prevention, 118*, 166–177. https://doi.org/10.1016/j.aap.2018.02.014

28. El Faouzi, N. E., Leung, H., & Kurian, A. (2011). Data fusion in intelligent transportation systems: Progress and challenges–A survey. *Information Fusion*, 12(1), 4-10.

29. Klein L. A. (2012). *Sensor and data fusion: A tool for information assessment and decision making*. 2nd Edition. SPIE press.

30. Proulx, F. R. (2016). *Bicyclist exposure estimation using heterogeneous demand data sources*. University of California, Berkeley, Ph.D. Dissertation. https://digitalassets.lib.berkeley.edu/etd/ucb/text/Proulx_berkeley_0028E_16522.pdf

31. City of Austin. (2015). *2014 Austin bicycle plan*. https://www.austintexas.gov/sites/default/files/files/2014_Austin_Bicycle_Master_Plan__Reduced_Size_.pdf

32. City of Austin Planning and Zoning. (2020). *Demographics*. http://www.austintexas.gov/demographics

33. Nordback, K., Marshall, W. E., Janson, B. N., & Stolz, E. (2013). Estimating annual average daily bicyclists: Error and accuracy. *Transportation Research Record: Journal of the Transportation Research Board, 2339*, 90–97.

34. Texas Department of Transportation (TxDOT). (2016). *Texas intersection safety implementation plan: Preliminary findings for Texas's Capital Area Metropolitan Planning Organization*. TxDOT.

35. Nakamura, E. F., Loureiro, A. A., & Frery, A. C. (2007). Information fusion for wireless sensor networks: Methods, models, and classifications. *ACM Computing Surveys (CSUR), 39*(3), 9-es.

36. Hall, D., & Llinas, J. (Eds.). (2001). *Multisensor data fusion*. CRC Press.

37. Kuncheva, L. I. (2014). *Combining pattern classifiers: Methods and algorithms*. John Wiley & Sons.

38. Ren, Y., Zhang, L., & Suganthan, P. N. (2016). Ensemble classification and regression-recent developments, applications and future directions. *IEEE Computational Intelligence Magazine, 11*(1), 41–53.

39. Castanedo, F. (2013). A review of data fusion techniques. *The Scientific World Journal*. https://doi.org/10.1155/2013/704504

40. Liu, K., Cui, M. Y., Cao, P., & Wang, J. B. (2016). Iterative Bayesian estimation of travel times on urban arterials: Fusing loop detector and probe vehicle data. *PLoS ONE*, *11*(6), e0158123.

41. Shi, C., Chen, B. Y., Lam, W. H., & Li, Q. (2017). Heterogeneous data fusion method to estimate travel time distributions in congested road networks. *Sensors, 17*(12), 2822.

42. Xu, L., Krzyzak, A., & Suen, C. Y. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics, 22*(3), 418–435.

43. Kronprasert, N. (2012). *Reasoning for public transportation systems planning: Use of Dempster-Shafer theory of evidence* (Doctoral dissertation, Virginia Tech).

44. Chust, G., Ducrot, D., & Pretus, J. L. (2004). Land cover discrimination potential of radar multitemporal series and optical multispectral images in a Mediterranean cultural landscape. *International Journal of Remote Sensing, 25*(17), 3513–3528.

45. Deng, X., Liu, Q., Deng, Y., & Mahadevan, S. (2016). An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Information Sciences, 340*, 250–261.

46. Jousselme, A. L., Grenier, D., & Bossé, É. (2001). A new distance between two bodies of evidence. *Information Fusion, 2*(2), 91–101.

47. Martin, A., Jousselme, A. L., & Osswald, C. (2008). Conflict measure for the discounting operation on belief functions. In *Proceedings 2008 11th International Conference on Information Fusion* (pp. 1–8). IEEE.

48. Dey, A. K. (2001). Understanding and using context. *Personal and ubiquitous computing*, *5*(1), 4-7.

49. Dey, A. K., & Abowd, G. D. (2000). *Towards a better understanding of context and context awareness*. Presented at the 2000 ACM Conference on Human Factors in Computer Systems, The Hague, Netherlands, April 1–6.

50. Chang, C. L., Tsai, C. H., & Chen, L. (2003). Applying grey relational analysis to the decathlon evaluation model. *International Journal of the Computer, the Internet and Management, 11*(3), 54–62.

51. Liu, S., & Forrest, J. Y. L. (2010). *Grey systems: Theory and applications*. Springer Science & Business Media.

52. Wang, X., Yang, J., Lee, C., Ji, Z., & You, S. (2016). Macro-level safety analysis of pedestrian crashes in Shanghai, China. *Accident Analysis & Prevention, 96*, 12–21.

53. Pulugurtha, S. S., Krishnakumar, V. K., & Nambisan, S. S. (2007). New methods to identify and rank high pedestrian crash zones: An illustration. *Accident Analysis & Prevention, 39*(4), 800–811.

54. Montella, A. (2010). A comparative analysis of hotspot identification methods. *Accident Analysis & Prevention, 42*(2), 571–581.

55. Sener, I. N., Lee, K., Hudson, J. G., Martin, M., & Dai, B. (2019). The challenge of safe and active transportation: Macrolevel examination of pedestrian and bicycle crashes in the Austin District. *Journal of Transportation Safety & Security*, 1-27.

56. Austin Transportation Department. (2018). *City of Austin pedestrian safety action plan 2018: Vision Zero*. https://austintexas.gov/sites/default/files/files/Transportation /Pedestrian_Safety_Action_Plan_1-11-18.pdf

57. Xie, K., Ozbay, K., Yang, D., Xu, C., & Yang, H. (2021). Modeling bicycle crash costs using big data: A grid-cell-based Tobit model with random parameters. *Journal of Transport Geography, 91*, 102953.

58. Apardian, R. E., & Monwar Alam, B. (2020). Pedestrian fatal crash location analysis in Ohio using exploratory spatial data analysis techniques. *Transportation Research Record: Journal of the Transportation Research Board, 2674*, 888–900.

59. TxDOT. (2017). *Texas strategic highway safety plan 2017-2022*. https://ftp.dot.state.tx.us/pub/txdot-info/library/pubs/gov/shsp.pdf

60. Zhang, C., Yan, X., Ma, L., & An, M. (2014). Crash prediction and risk evaluation based on traffic analysis zones. *Mathematical Problems in Engineering*, 2014.

61. Ord, J. K., & Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis, 27*(4).

62. Soltani, A., & Askari, S. (2017). Exploring spatial autocorrelation of traffic crashes based on severity. *Injury, 48*(3), 637–647.

63. Leslie, E., Saelens, B., Frank, L., Owen, N., Bauman, A., Coffee, N., & Hugo, G. (2005). Residents' perceptions of walkability attributes in objectively different neighbourhoods: A pilot study. *Health & Place, 11*(3), 227–236.

64. Van Dyck, D., Cerin, E., Conway, T. L., De Bourdeaudhuij, I., Owen, N., Kerr, J., Cardon, G., Frank, L. D., Saelens, B. E., & Sallis, J. F. (2012). Perceived neighborhood environmental attributes associated with adults' transport-related walking and cycling: Findings from the USA, Australia and Belgium. *International Journal of Behavioral Nutrition and Physical Activity, 9*(1), 1–14.

65. Sener, I. N., & Lee, R. J. (2017). Active travel behavior in a border region of Texas and New Mexico: Motivators, deterrents, and characteristics. *Journal of Physical Activity and Health*, *14*(8), 636-645.

66. Ma, L., & Cao, J. (2019). How perceptions mediate the effects of the built environment on travel behavior? *Transportation, 46*(1), 175–197.

67. Winters, M., Babul, S., Becker, H. J., Brubacher, J. R., Chipman, M., Cripton, P., Cusimano, M. D., Friedman, S. M., Harris, M. A., Hunte, G., Monro, M., Reynolds, C. C. O., Shen, H., & Teschke, K. (2012). Safe cycling: How do risk perceptions compare with observed risk? *Canadian Journal of Public Health, 103*(9), eS42–eS47.

68. Lawson, A. R., Pakrashi, V., Ghosh, B., & Szeto, W. Y. (2013). Perception of safety of cyclists in Dublin City. *Accident Analysis & Prevention, 50*, 499–511.

69. Chataway, E. S., Kaplan, S., Nielsen, T. A. S., & Prato, C. G. (2014). Safety perceptions and reported behavior related to cycling in mixed traffic: A comparison between Brisbane and Copenhagen. *Transportation Research Part F: Traffic Psychology and Behaviour, 23*, 32–43.

70. FHWA. (2019). *FHWA NHTS report: Changing attitudes and transportation choices*. https://nhts.ornl.gov/assets/FHWA_NHTS_Report_3E_Final_021119.pdf

71. NHTS. (2018). *Main study retrieval questionnaire*. https://nhts.ornl.gov/assets/2016/NHTS_Retrieval_Instrument_20180228.pdf

72. Jacobsen, P. L. (2015). Safety in numbers: More walkers and bicyclists, safer walking and bicycling. *Injury Prevention, 21*(4), 271–275.

73. Leden, L. (2002). Pedestrian risk decrease with pedestrian flow: A case study based on data from signalized intersections in Hamilton, Ontario. *Accident Analysis & Prevention, 34*(4), 457–464.

74. Munira, S., Sener, I. N. & Dai, B. (2020). A Bayesian spatial Poisson-lognormal model to examine pedestrian crash severity at signalized intersections. *Accident Analysis and Prevention*. 144, 105679.

75. Strauss, J., Miranda-Moreno, L. F., & Morency, P. (2013). Cyclist activity and injury risk analysis at signalized intersections: A Bayesian modelling approach. *Accident Analysis & Prevention*, 59, 9–17.

76. Lee, K., & Sener, I. N. (2020). Emerging data for pedestrian and bicycle monitoring: Sources and applications. *Transportation Research Interdisciplinary Perspectives*, *4*, 100095.

77. Munira, S., Sener, I. N., & Zhang, Y. (2021). Estimating bicycle demand in the Austin, Texas Area: Role of a bikeability index. *Journal of Urban Planning and Development*, 147(3), 04021036.

78. Swanson, K. (2012). *Bicycling and walking in the United States benchmarking report*. Alliance for Biking and Walking.

79. City of Austin. (2018). *The City of Austin Transportation Department 2018 annual report*. https://1d0d7cb9-bfde-4667-8eeb-20a5ee919bd6.filesusr.com/ugd/956239_455271f084ea4a4d9caf7bb098fb18ab.pdf

80. American Community Survey (ACS). (2019). *Commuting characteristics by sex, 2019 ACS 1-year estimates*. https://data.census.gov/cedsci/table?q=Commuting%20Characteristics%20by%20Sex&tid=ACSST1Y2019.S0801&hidePreview=false

81. Geller, R. (2009). *Four types of cyclists*. Portland Office of Transportation. https://www.portlandoregon.gov/transportation/44597?a=237507

82. Krizek, K. J., Iacono, M., El-Geneidy, A. M., Liao, C. F., & Johns, R. (2009). *Access to destinations: Application of accessibility measures for non-auto travel modes* (No. MN/RC 2009-24). Minnesota Department of Transportation.

83. Dill, J. (2009). Bicycling for transportation and health: The role of infrastructure. *Journal of Public Health Policy, 30*(1), S95–S110.

84. Griswold, J., Medury, A., & Schneider, R. (2011). Pilot models for estimating bicycle intersection volumes. *Transportation Research Record: Journal of the Transportation Research Board, 2247*, 1–7.

85. Hankey, S., Lu, T., Mondschein, A., & Buehler, R. (2017). *Merging traffic monitoring and direct-demand modeling to assess spatial patterns of annual average daily bicycle and pedestrian traffic*. Presented at the Transportation Research Board 2017 Annual Meeting.

86. Hasani, M., Jahangiri, A., Sener, I. N., Munira, S., Owens, J. M., Appleyard, B., Ryan, S., Turner, S. M., & Ghanipoor Machiani, S. (2019). Identifying high-risk intersections for walking and bicycling using multiple data sources in the city of San Diego. *Journal of Advanced Transportation*. https://doi.org/10.1155/2019/9072358

87. Munira, S., & Sener, I. N. (2017). *Use of the direct-demand modeling in estimating nonmotorized activity: A meta-analysis*. Safety through Disruption (Safe-D) National University Transportation Center (UTC) Program.

88. City of Austin. (2017). *Austin Texas bike map*. https://austintexas.gov/sites/default/files/files/Transportation/2017_Austin_Bike_Map_-_Side_2.pdf

89. Chan-Lau, M. J. A. (2017). *Lasso regressions and forecasting models in applied stress testing* (Working paper). Institute for Capacity and Development. https://doi.org/10.5089/9781475599022.001

90. Chen, P., Zhou, J., & Sun, F. (2017). Built environment determinants of bicycle volume: A longitudinal analysis. *Journal of Transport and Land Use, 10*(1), 655–674.

91. Parsons Brinckerhoff Quade & Douglas. (1999). *Development and calibration of the statewide land use-transport model*. Oregon Department of Transportation.

92. Feudo, F. L., Morton, B., Capelle, T., & Prados, E. (2017). Modeling urban dynamics using LUTI models: Calibration methodology for the Tranus-based model of the Grenoble urban region. In *HKSTS 2017—22nd International Conference of Hong Kong Society for Transportation Studies*.

93. Wegener, M. (2004). Overview of land-use transport models. In D. Hensher, K. Button, K. Haynes, & P. Stopher (Eds.), *Handbook of transport geography and spatial systems* (Vol. 5, pp. 127–146). Emerald Group. https://doi.org/10.1108/9781615832538-009

94. Collins, T. W., Grineski, S. E., & Aguilar, M. D. L. R. (2009). Vulnerability to environmental hazards in the Ciudad Juárez (Mexico)–El Paso (USA) metropolis: A model for spatial risk assessment in transnational context. *Applied Geography, 29*(3), 448–461.

95. Eco-Counter. (2019). *Products*. https://www.eco-compteur.com/en/produits/multi-range/urban-multi/

96. BCycle. (2018). *Austin B-cycle expands service to UT campus and west campus neighborhoods*. https://www.bcycle.com/news/2018/02/14/austin-b-cycle-expands-service-to-ut-campus-and-west-campus-neighborhoods

97. Kuzmyak, J. R., Walters, J., Bradley, M., & Kockelman, K. M. (2014). *Estimating bicycling and walking for planning and project development: A guidebook* (No. 08-78). NCHRP.

98. Lee, K., & Sener, I. N. (2021). Strava Metro data for bicycle monitoring: A literature review. *Transport Reviews*, 41(1), 27-47.

99. StreetLight. (2018). *StreetLight active mode methodology, data sources and validation* (Version 1.0). https://www.streetlightdata.com/wp-content/uploads/StreetLight_Active-Mode_Methodology-and-Validation_190108.pdf

100. City of Austin. (2021). *Open data inventory*. https://data.mobility.austin.gov/open-data/

Appendix A

Framework for Processing Nonmotorized Traffic Data

Nonmotorized traffic data, both traditional and crowdsourced, are generated in different structures and formats. See Lee and Sener [76] for an overview of data for pedestrian and bicycle monitoring, with a focus on sources and applications of emerging data.

Table 2 tabulates the key characteristics of the commonly available nonmotorized traffic data sources, including short-duration and permanent counts, travel surveys including ACS and NHTS, and crowdsourced data. The breakdown of the characterization of the individual source is vital in interpreting and further processing the data to extract volume-related information. Table 2 depicts that none of the nonmotorized traffic data sources can furnish activity data for full spatial, temporal, or population-level resolution. For example, permanent sensor data can provide data for all hours and seasons across the year, but only for the location for which they have been installed. Data sources from wearable tech only provide data for the people using the apps. Therefore, to compute activity-related information, each of the datasets has to be adjusted, scaled, or modeled. The application of the data sources, either as direct input of demand models or as parameters or aids of the model-building process, also varies widely. Moreover, while some of the sources can be processed to estimate activity, some, such as Web 2.0 tech, are not meant for demand representation. In addition, some of the data sources need minimal processing, while others have to go through an array of steps to compute demand given the scope of estimation.

Nonetheless, there is no clear guidance on how different data sources can be processed and brought together to compute nonmotorized volume or exposure within a facility or area. Motivated by the ardent need for a comprehensive guideline, this research developed a conceptual framework that outlines the steps and aspects of processing and homogenizing different sources. The framework, which consists of processing data and applying fusion mechanisms, can be applied to all nonmotorized traffic data, such as bicycle and pedestrian data. However, for demonstration purposes, as a case study, only bicycle-related data were collected and processed.

The key steps for gathering and processing nonmotorized traffic data are:

Step 1: Select the study area.
Step 2: Gather available data sources and identify the data structure.
Step 3: Determine the scope of estimation.
Step 4: Analyze/model data sources to obtain homogenized representation.

Table 2. Characterization of Nonmotorized Traffic Data Sources*

Data Sources	Temporal Coverage or Frequency of Data Collection	Spatial Coverage	Population Resolution	Data Output	Role in Demand Estimation
Short-Duration Count (i.e., manual, video image, etc.)	15 minutes to 24 hours for multiple days or week	Selected locations (intersection or mid-block)	Total population	Aggregated volume in a road segment or intersection for a specific time period	• Estimating flow patterns and annual average or peak/off-peak hour volume • Computing adjustment factor for scaling short-duration count • Developing DDM
Permanent Count (i.e., inductance loop, magnetometer, etc.)	Constant count for at least 1 year	A few locations (intersection, trail, or mid-block)	Total population	Aggregated volume in a road segment or intersection	
American Community Survey https://www.census.gov/programs-surveys/acs	1-year, 3-year, and 5-year estimates	Area-level estimate for all geographies down to the block group level	Total population	Commute, social, economic, demographic, and housing characteristics of the U.S. population	• Estimating parameters for demand models • Developing mode choice model
National Household Travel Survey https://nhts.ornl.gov/	Every 5 to 7 years	Down to core-based statistical area level in the United States	Total population	Individual travel data for all trips, modes, purposes, trip lengths, etc.	
NHTS Add-On Data https://nhts.ornl.gov/addOn.shtml	Every 5 to 7 years for add-on partners	Smaller and more precise level of geography compared to NHTS data	Total population	Additional data include origin and destination geocoordinates of each trip	
Fitness App Data (e.g., Strava: https://metro.strava.com/)	Continuous data collection	Entire United States (based on users)	Only Strava users	Origin-destination and node/street-level volume	• Estimating system user volume • Estimating total volume
Bike-Sharing Data	Continuous data collection	Depends on the coverage of the stations	Only system users	Origin-destination of the trips	
StreetLight Data https://www.streetlightdata.com/	Continuous data collection	Entire United States (based on users)	A sample of the total trips	Node/street-level volume	

* See also Appendix E for additional information about the data sources as well as their resources as relevant to this research study.

The following sections explain the steps in detail and present a case study as an example of the application.

Select Study Area

The first step of the framework is to select a study area for gathering, processing, and analyzing nonmotorized traffic data. The study area can be a city, county, multiple census tracts, or traffic analysis zones (TAZs), etc. The selection of a study area depends on the ultimate objectives of the researchers (such as to analyze pedestrian crashes in a city or design bike lanes in an urbanized area), data availability, and so forth. The size of the study area should also be a consideration since some nonmotorized models (such as the DDM) exhibit scalability and transferability issues [77]. For example, while building a bike DDM for Austin, Munira et al. [77] noted that the model prediction was drawn for locations within the limits of the city's bicycle route map and did not include areas farther into the suburban regions.

For this study, the city of Austin was selected as a study area. Austin makes an excellent case study for this research for multiple reasons. Austin ranked 27th among U.S. cities in terms of high bicycling and walking levels in 2012 [78]. The same report indicated that Austin stood 15th and 37th for bicycle and walk commute share, respectively, compared to other U.S. cities. As a bike friendly city, Austin is endeavoring to adopt a holistic approach to increase safety and mobility for pedestrians and bicyclists of all ages. The Austin City Council adopted the 2014 Austin Bicycle Master Plan to develop a connected and protected walking and biking network. The city accommodates a total of 267.5 miles of bicycle facilities, including protected and buffered bicycle lanes and urban trails [79]. Approximately 36% of the region's arterial streets have traditional painted bicycle lanes [31]. The planning and implementation of various projects since 2009 has resulted in a significant increase in bicycle commuters in the city. The citywide mode share of bicycles doubled in 2011 (around 2%) compared to 2009 [31]. This was particularly the case in some census tracts in central Austin, which had seen a considerably higher mode share than the suburban regions [30]. On the other hand, according to the 2019 ACS data, the bike mode share across the counties located within the city varied from 0.2% to 0.9% [80].

The city also adopted the Vision Zero initiative to reduce traffic-related deaths and injuries to zero by the year 2025. To promote safe walking and biking among the residents, the city has focused on various programs and activities, including nonmotorized friendly street design and pedestrian safety action plans. Given the strong commitment to its Vision Zero goals and long-term planning to improve nonmotorized infrastructure, Austin needs reliable and robust nonmotorized activity data and tools to facilitate strategic data-informed decisions. Therefore, a comprehensive data fusion framework for reliable nonmotorized demand/exposure estimation, as applied in this study, would be of great value to policy makers and practitioners, along with scholars.

Gather Available Data Sources and Identify Data Structure

The second step is to identify available data sources relevant to nonmotorized volume estimates in the region. Table 3 depicts the characterization of the nonmotorized traffic data sources. It is

important to gather all available data sources and recognize the attributes associated with each source, such as frequency of data collection; geographic unit; and spatial, temporal, and population-level resolution.

For Austin, five primary data sources relevant to bike activity were gathered:

- Actual bicycle volume counts (permanent and short counts)
- Bicycle-sharing data
- NHTS add-on data
- Strava data
- StreetLight data

The ACS and other land-use-related data were also extracted for the study area and served as auxiliary data sources for modeling purposes.

The five key datasets represented bike activity at various spatial, temporal, and subpopulation scales. Table 3 presents the structure of the datasets. Table 3 confirms that the data sources exhibit heterogeneous structures. In order to obtain a homogeneous estimate that is meaningful and relevant to policy planning, each of the data sources needs to be cleaned, processed, and modeled within the scope of estimation, as explained in the next section.

Table 3. Structure of the Bicycle Data Sources in Austin

Data Sources	Temporal Coverage or Year of Data Collection	Spatial Coverage	Population Resolution	Source
Video-Based Short Count	24-hour count in 2017	44 intersections	Yes	City of Austin Transportation Department
Inductive Loop-Based Permanent Count	Continuous count from 2012 to 2017	11 locations	Yes	Eco-Counter (through City of Austin)
NHTS Add-On Survey	Gathered in 2017	1,095 households in Austin	Yes	TxDOT
Strava Metro Data	Trips in 2017	2,303 intersections in Austin	No	TxDOT (through an internal agreement with Strava)
Bike-Sharing Data	Trips in 2017	63 stations in Downtown Austin	No	Public website
StreetLight Data	Trips in 2018	950 zones for intersections in Austin	No	StreetLight Inc.
ACS Data	2017	Census tract or block group level in Austin	N/A	Public website
Land-Use-Related Data	2014 to 2017	Point or area-based data in Austin	N/A	Internal communication with the City of Austin and public website

Determine Scope of Estimation

The processing of individual data sources inevitably entails determining the scope for demand analysis, which can be temporal scope, spatial scope, and population-level scope. The following section explains the premise and reasoning behind the selection of scope/unit of estimation for the case study.

Temporal Unit of Volume/Demand Estimation

It is imperative to select a temporal unit of analysis for representing nonmotorized traffic demand from all sources. Intuitively, estimates at a finer unit (such as hourly volume) require data at a finer temporal resolution compared to estimates at a larger unit (such as annual average volume). However, data at a finer resolution may not be available from all sources. For example, the four-step demand modeling approach generally estimates traffic as annual average daily traffic (AADT) volume. Transforming estimations at a finer unit requires the involvement of multiple assumptions and an understanding of the local conditions.

The temporal unit of analysis for this study was selected as AADB volume. The selection was also motivated by the fact that the particular metric is a widely accepted policy-relevant unit of demand representation for policy planning and safety analysis.

Spatial Unit of Volume/Demand Estimation

The task of estimating nonmotorized demand requires the analyst to select a geographic or spatial unit for which the volume will be estimated. The plausible spatial units are facility level (such as intersection or mid-block) or zone level (such as TAZ or block) [3].

For this research, intersections were selected as the spatial unit of volume analysis. The selection stemmed from the fact that a large proportion of crashes occur at intersections in the Texas region [34]. A microscopic level (intersection) analysis was expected to facilitate focused policy efforts and effective safety implementation plans. Moreover, the actual count data for the study area were available at the intersection level and were expected to serve as the ground truth and benchmark data for models.

To identify the intersections for the study area, a bicycle network developed and managed by the City of Austin Transportation Department was obtained. The bike route network is generally different from the regular street maps because it identifies various bicycle-related facility segments (off-street and on-street bicycle facilities, special facilities, etc.) in addition to regular roadways that allow bicyclists. The process of extracting intersections from the bike network is explained in an upcoming section.

Population Scale

Since some of the data sources represent a different subpopulation of the actual bike activity, it is logical to expand or scale such activity from individual sources to the total population level. While the analyst may not have to choose a particular population scale for homogenization, it is imperative to understand the representativeness and skewness of the sources.

Analyze/Model Data Sources to Obtain Homogenized Representation

Given the selected scope of bike activity estimation, AADB at intersections, the data sources of different structures had to go through multiple steps of analysis. Using the data sources tabulated in Table 3, the researchers developed five models:

- DDM
- Four-step model
- Bike-sharing model
- Strava model
- StreetLight model

A tabulated summary of the key features of the modeling steps is presented in Table 4, followed by an in-depth discussion of individual processes.

Table 4. Key Steps of Volume Estimation Models

Model	Main Input	Use of Explanatory Variables	Model Type or Steps	Adjust Population Scale?
DDM	Short and permanent count data	Yes	Regression	No
Four-Step Model	Demographics and employment data	Yes	Trip generation, distribution, mode choice, assignment	No
Bike-Sharing Model	Bike-sharing volume	No	Trip assignment	Yes
Strava Model	Node/road-level app volume	Yes	Regression	Yes
StreetLight Model	Zone-level volume	Yes	Regression	Yes

The description of the modeling process is divided into two parts. Because each of the data sources had to undergo some processing and modeling tasks, the common steps that were utilized multiple times, such as gathering auxiliary data, regression model building, traffic assignment, and actual volume processing, are explained in the first part. Then, the models developed from the induvial sources are described.

Common Modeling Steps

In this section, some of the steps that were relevant to multiple or all models developed in this study are explained to avoid repetition when later discussing the individual models. The discussion includes the processing of bike network data to obtain intersections for the study area and the processing of explanatory variables for models. The section also explains the generalized variable and model building process and the traffic assignment task required for the bike-sharing and four-step models.

Processing Bike Network

The bike network of the study area is one of the crucial inputs required for activity or demand estimation of nonmotorized traffic. It is required to locate the intersections or street segment locations for which the demand has to be computed. The network map is also necessary for route choice analysis in the traffic assignment task. A bike or pedestrian route network is generally different from regular street maps because it identifies various bicycle or pedestrian-specific facilities and excludes the access-controlled routes on which nonmotorized activities are not allowed.

For the study area, a bicycle network map developed and managed by the City of Austin Transportation Department was obtained. The map included information on both bicycle facility type (on-street/off-street, etc.) and bicycle comfort level for each segment. The comfort level was computed based on a study by Geller [81] that considered several factors, including traffic speeds and volumes, roadway widths, bicycle facility type, and other readily available metrics, to determine how comfortable a segment is for people of all ages and abilities. The comfort level was categorized into four types: high comfort sections, medium comfort sections, low comfort sections, and extremely low comfort sections. The bike network excluded road segments that were without any bike facility or of low comfort but with an alternative route option.

The residents of Austin are allowed to suggest new routes or update the comfort levels of the existing routes. Therefore, the mapping follows a robust process that takes continuous advantage of public feedback to avoid the routes where bicyclists never or very seldom ride. The network was thus expected to provide a good representation of the actual route choices by the bicyclists of the study area.

Because the network was processed to be used in intersection identification and route choice analysis, it was imperative to perform a comprehensive quality check to examine its completeness and accuracy. As expected, the route map did not include access-controlled roads (i.e., interstate highways) where bicyclists are not allowed. Moreover, acknowledging the fact that even the most carefully developed networks are bound to have errors, the researchers checked the obtained network file for two common digitizing errors: overshoots and undershoots. Overshoot and undershoot errors happen when a line is not connected with the neighboring line with which it should intersect [82]. Following an exhaustive manual investigation of the network, the researchers performed a network correction task (in ArcGIS) to avoid error in the traffic assignment process.

The final task was to identify intersections from the route network. Both three- and four-legged intersections were identified in the process. The study area consisted of 2,518 intersections.

Processing Explanatory Variables

In an effort to assemble explanatory variables for the study, prior studies were reviewed with a focus on DDMs and bike determinants [83–87]. This study sought to build a rich set of explanatory variables using insights from the earlier studies as well as the data available for the study area. To develop the explanatory variables for the model, first researchers performed a comprehensive search to see which data were publicly available. Data were gathered from the City of Austin data portal, 2017 ACS, City of Austin Planning and Development Review Department, Texas Education Agency, Austin Transportation Department Arterial Management Division, Capital Metro, and BCycle (Austin bike-sharing agency) data portal. After identifying the gap between the required and available datasets, the researchers contacted relevant authorities and asked them to provide additional data for research purposes. Additionally, a review of factors included in past studies identified a need to explore new variables related to the features that have the potential to drive the bicycle demand at an intersection. Therefore, the researchers examined some additional

variables—including the presence of bike-sharing stations or bike signals around an intersection, bike-accessible bridges, and so forth—anticipating their possible impacts on bicycle volume in the area.

Since all of the raw datasets obtained from different sources were at different spatial scales, the datasets were cleaned and processed to bring them to homogenous spatial scales (buffer level). Over 400 variables for three buffer zones—0.1 mi, 0.5 mi, and 1 mi—were created. The variables were categorized into seven groups following the categorization suggested by Munira and Sener [87]: demographics, socioeconomics, network/interaction with vehicle traffic, pedestrian- or bicycle-specific infrastructure, transit facilities, major generators, and land use.

The demographic and socioeconomic variables included age, gender, education, race, household size and occupancy status, income, and commute mode and time of the surrounding population. The network and bicycle-specific infrastructure-related variables included different types of bicycle infrastructure developed by the City of Austin [88] based on the conditions and comfort level, as well as bike signal, intersection density, and bike-sharing stations. Various transit-facility-related variables were compiled, including frequency of transit stops, transit route length, and distance from hub locations. Major generators and land-use variables, such as the number of schools, offices, industries, open areas, mixed-use developments, water areas, and bicycle-accessible bridges, were also gathered based on available data.

Creating a Bikeability Index

A bikeability index, a composite measure to quantify the bike friendliness of the network, was created to develop a DDM addressing the issue of small sample size (limited actual count data) from the study area. The five attributes of the bikeability index were bicycle route length, high comfort bicycle route length, connectivity of bicycle-friendly streets, destination density, and transit coverage. Details of the procedure can be found in Munira et al. [77].

Variable Selection and Model Building

The variable selection was completed through an extensive three-stage procedure. First, a simple ordinary least squares model was developed to analyze the relative strengths of relationships between each of the explanatory variables and the dependent variable. This included identification of the variables that had significant association at a 90% confidence level with the dependent variable. Second, Pearson's correlation coefficients were examined for each pair of explanatory variables to investigate the correlation between variables. This process yielded a large number of significantly correlated variables, from which highly correlated (at 0.7) variable pairs were recognized. Finally, by iterating several combinations of variables that are not highly correlated, the researchers selected a final model based on its predictive accuracy, such as mean absolute error (MAE), root mean square error (RMSE), misclassification error, and fitness (adjusted R^2). The statistical significance of individual variables and intuitive interpretation, based on insights from the literature, were also considered while selecting the variables for the best model.

The performance of the models was evaluated using cross-validation, which is a resampling technique that helps identify a parameter value, ensuring a proper balance between bias and variance [89]. For cross-validation, a subset of the data, known as the training set, was used to train the model, and the remaining data points served as a test set or validation set. The model on the training set seeks a minimum mean squared error. A 10-fold, cross-validation method was used to evaluate and compare the performance of the developed models. This method split the feature vector sets into 10 approximately equal-sized distinct partitions. While one set was used for testing, the other set was used for training. Then, the procedure was repeated 10 times, and all accuracy rates over these 10 runs were averaged to provide a more reliable estimate. The performance evaluation criterion was the average accuracy.

The process of variable selection could have also been performed by utilizing various state-of-the-art machine learning approaches, including Lasso or Random Forest. However, this study opted for the manual approach over those machine learning processes because they are often referred to as black box approaches with limited interpretability [90]. The approach followed in this study promotes a deep understanding of the influence of individual variables and the dynamics of their relationship, given the local condition, in order to make informed decisions.

Traffic Assignment

A traffic assignment task is needed to allocate traffic to the facilities of a transport network. This is the last step in the four-step modeling process. The step is also required in bike-sharing traffic allocation because the raw data generally come in a zone-to-zone (station-to-station) trip format.

For conducting the traffic assignment task, several software packages are available, such as PTV VISUM from PTV Group, EMME by INRO, CUBE Voyager by Citilabs, and TransCAD by Caliper, all of which require a commercial license. In this study, open-source software was preferred. Tranus is open-source software that can be used in developing land-use and transportation models at an urban or regional scale. First developed in 1982 by Modelistica, the software has gone through several versions, incorporating theoretical developments and practical requirements [91]. This project used Tranus version 12.10.1.

Tranus has been utilized in a large number of studies and for regions of varying socioeconomic and cultural contexts, such as Latin America, Europe, the United States, and Japan. The Oregon Department of Transportation used Tranus to develop an integrated land-use and transport model at the statewide level [91]. A research project promoted by the French National Research Agency utilized Tranus to develop an integrated land-use and transport model for the urban region of Grenoble, France, and the researchers reported that they were able to achieve their goal of building a meaningful and policy-relevant model using the Tranus system [92]. Wegener [93] reviewed 20 contemporary urban land-use transport models and noted that Tranus stands out as a particularly advanced and well-documented demand modeling platform with an attractive user interface. The Center for International Intelligent Transportation Research of the Texas A&M Transportation Institute (TTI) and Modelistica developed a binational travel demand model for El Paso, Texas,

and Ciudad Juarez, Mexico, using Tranus to help transportation agencies of both regions anticipate traffic flow and needs [94].

In this study, the traffic assignment task in the Tranus platform went through an exhaustive process utilizing an array of assumptions and hypotheses. The models were customized based on the characteristics and requirements of the study area.

Processing Actual Bicycle Volume Counts

Two types of bicycle count data were obtained for the study area. The short-count (24-hour) data were obtained from the City of Austin Transportation Department, and continuous-count data were obtained from Eco-Counter, which is a company that assists with continuous data collection for pedestrians and bicyclists in specific locations across cities around the world [95]. The continuous-count data were needed to calculate adjustment factors that could be incorporated with the short count to estimate the AADB volume for the specific locations [33].

The 24-hour bicycle count data were available for 44 locations. According to city officials, the sites were selected using the City of Austin bicycle route map [88] and based on the professional judgment of local planners. Following standard procedures, the data were collected by the city using a video recorder in each of the intersections on typical weekdays distributed over 5 months (April, May, June, August, and October) in 2017. The permanent location counts were obtained from Eco-Counter for 11 locations in the Austin area; the company provided continuous counts for the locations since 2012. The count data from the permanent counters were used to estimate the daily and monthly factors, which were then applied to calculate the AADB volume for each location where the short-count data were available.

Figure 5 shows the location of the 44 intersections with short-count data and presents the estimated AADB for each location. The intersections exhibited notable variation in terms of AADB volume, with a minimum of 43 and a maximum of 1,282 riders.

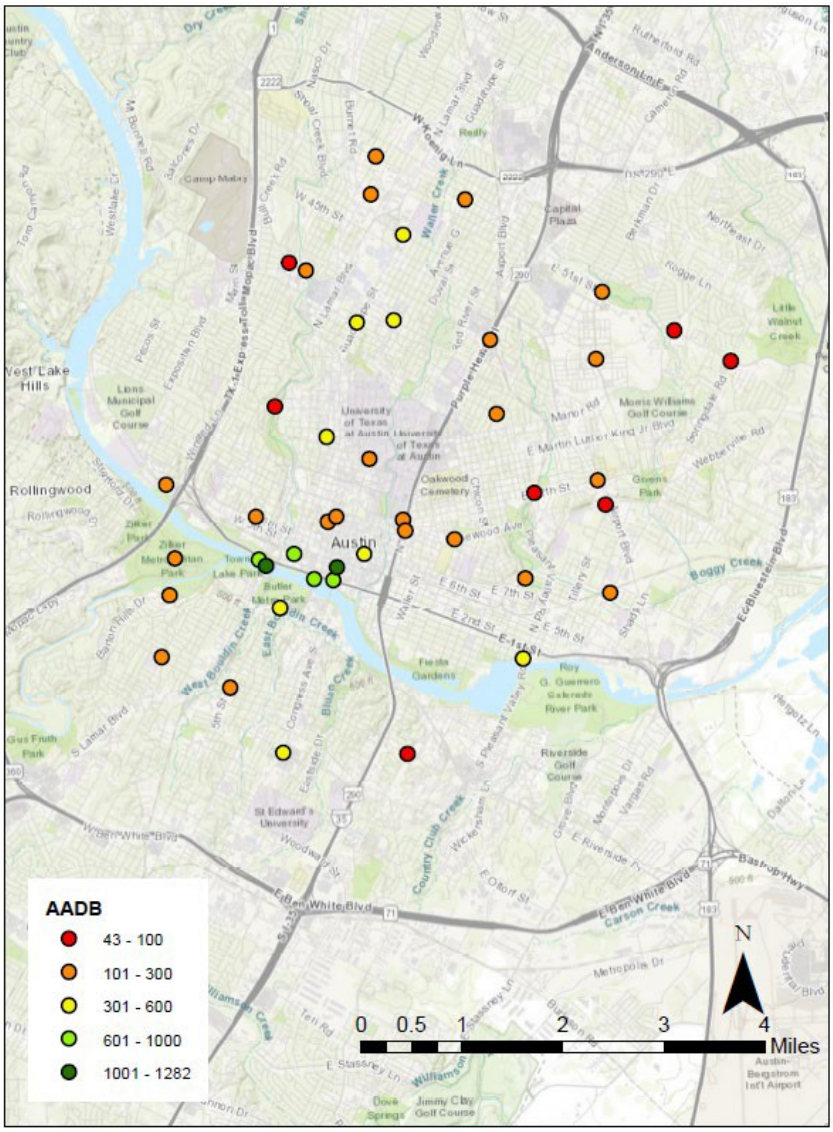

Figure 5. Map. Actual AADB volume in Austin.

Model Building from Individual Sources

This section describes the demand models developed utilizing multiple sources. The characteristics and features of different datasets are explained briefly. The steps mentioned in the previous section are referred to when necessary.

Direct Demand Model

Among several approaches to estimate and predict the demand of pedestrian and bicycle travel, the direct (facility) demand model is the most frequently used modeling approach in the area of pedestrian/bicyclist safety [3]. This research utilized the actual AADB estimate, bikeability index, and explanatory variables (as discussed in the previous section) to develop a DDM (Table 5) that can estimate AADB for all intersections of the study area. More details on the procedure are available in Munira and Sener [77].

Table 5. Negative Binomial Regression Model of Bicycle Travel

Variable (buffer width)	Estimates	T-Stat
(Intercept)	4.96	8.82
Bikeability index (1 mi)	0.02	1.65
Black or African American population (1 mi) (in 100 s)	−0.02	−3.23
Population with no or some academic degree (0.1 mi)	0.03	2.98
Total population of age under 14 (0.5 mi) (in 100 s)	−0.12	−2.99
Bike signal (0.1 mi)	0.30	2.46
Presence of bicycle-accessible bridge (0.1 mi)	0.54	2.25
Model Statistics: N (sample size): 44 Adjusted R^2: 0.7 RMSE: 171 MAE: 132		

The adjusted R^2 for the prediction model was 0.7, and RMSE was 171. The model also provided an estimate of AADB for all intersections of the study area, where the predicted AADB varied from a minimum of 15 (mainly at the areas away from the downtown core) to a maximum of 1,398 (at the downtown core).

Figure 6 presents the spatial distribution of the AADB estimates resulting from the DDM.

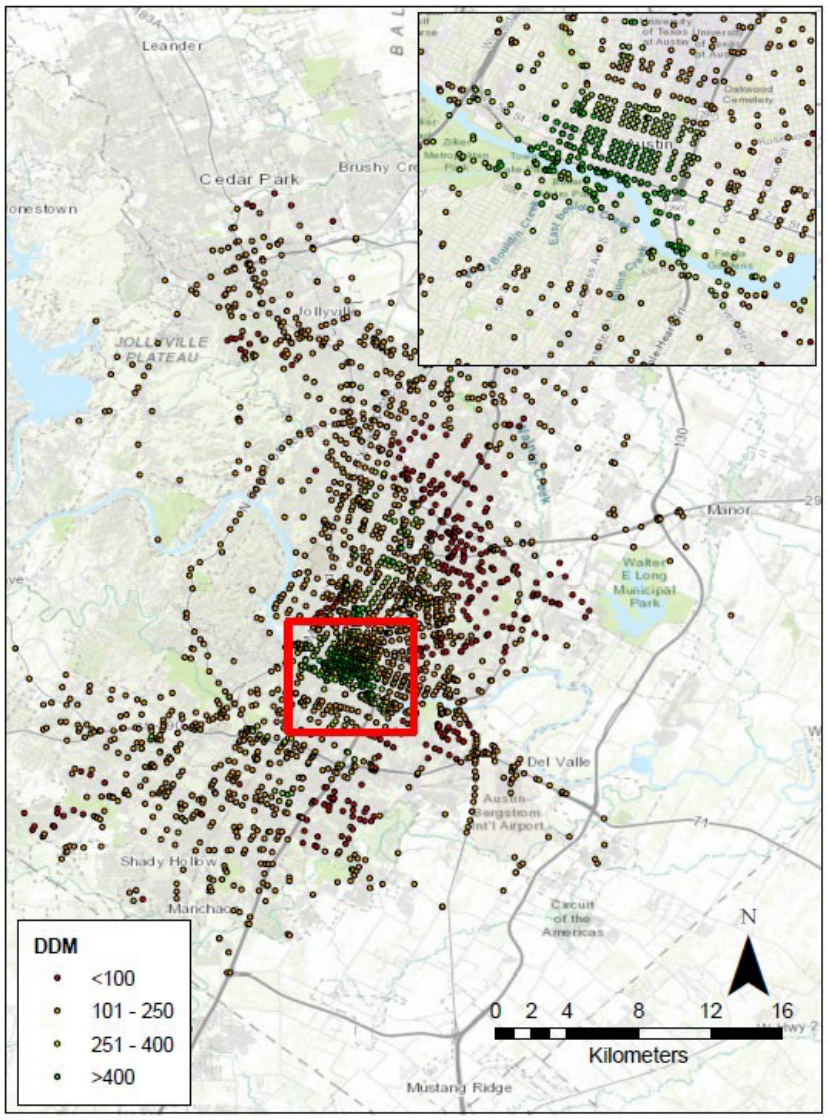

Figure 6. Map. AADB estimates from the DDM.

Bicycle-Sharing Model

Bicycle-sharing data were obtained from the data portal of the BCycle-Austin bike-sharing agency. The bike-sharing system in Austin mainly covers the downtown region, with only 63 stations as of 2018 [96]. The research extracted bicycle trip data from all stations for 2017, when a total of 193,492 trips were logged. The dataset contained limited information regarding each trip, including trip start and end time and station.

Initial investigation on the temporal distribution of the trips revealed that the highest numbers of trips were observed during March through April, followed by September. This distribution might be attributed to the pleasant and mild weather during these months. The Texan summer, when temperatures climb into the mid to high 90s with high humidity, might be the reason for low bike-

sharing trip activity from June through August. Intuitively, Friday, Saturday, and Sunday observed a higher number of trips compared to other days of the week. Moreover, the majority of the trips were made by users with a 24-hour pass (compared to annual pass, monthly pass, and 3-day pass users). This finding explains the high number of plausible recreational purpose trips that started and ended at the same station. Despite having a wide temporal coverage, the spatial coverage of the data was limited because the sharing stations mainly cover the downtown region. Trip starting and ending volumes were concentrated near the Lady Bird Lake (downtown) area. Moreover, the bike-sharing activity only accounted for a proportion of the total bike activity in that area, and thus needed population-level scaling.

The process of estimating annual average bicycle volume from this specific data went through three key stages: (a) examination and cleaning to obtain data in a meaningful format, (b) trip assignment, and (c) population-level scaling. In the first step, the distribution of the data, in terms of spatial unit, temporal unit, membership level, etc., was investigated. The step also included data cleaning; for example, trips starting and ending at the same station were removed because they would not add value to the assignment process (i.e., no trips would be assigned to the routes). Then, an OD matrix was developed using the station-level data. For this purpose, trips for the months April, May, and June, which represented the typical months of bike-sharing activity, were extracted. Then an OD table was built using the average trips for each station (OD) pair for the above-mentioned three months. For example, if there were six trips in three days (during the three months) for an OD pair, the average trip for that particular OD pair was taken as 2. Trips were rounded up for the decimal numbers. In the second step, a trip assignment process was conducted (as explained previously) to assign the trips at the intersection level. The output of this step was intersection-level bike-sharing volume. Finally, the estimate was scaled to population level using actual AADB and utilizing a negative binomial model.

Figure 7 presents the spatial distribution of the AADB estimates resulting from the bike-sharing model.

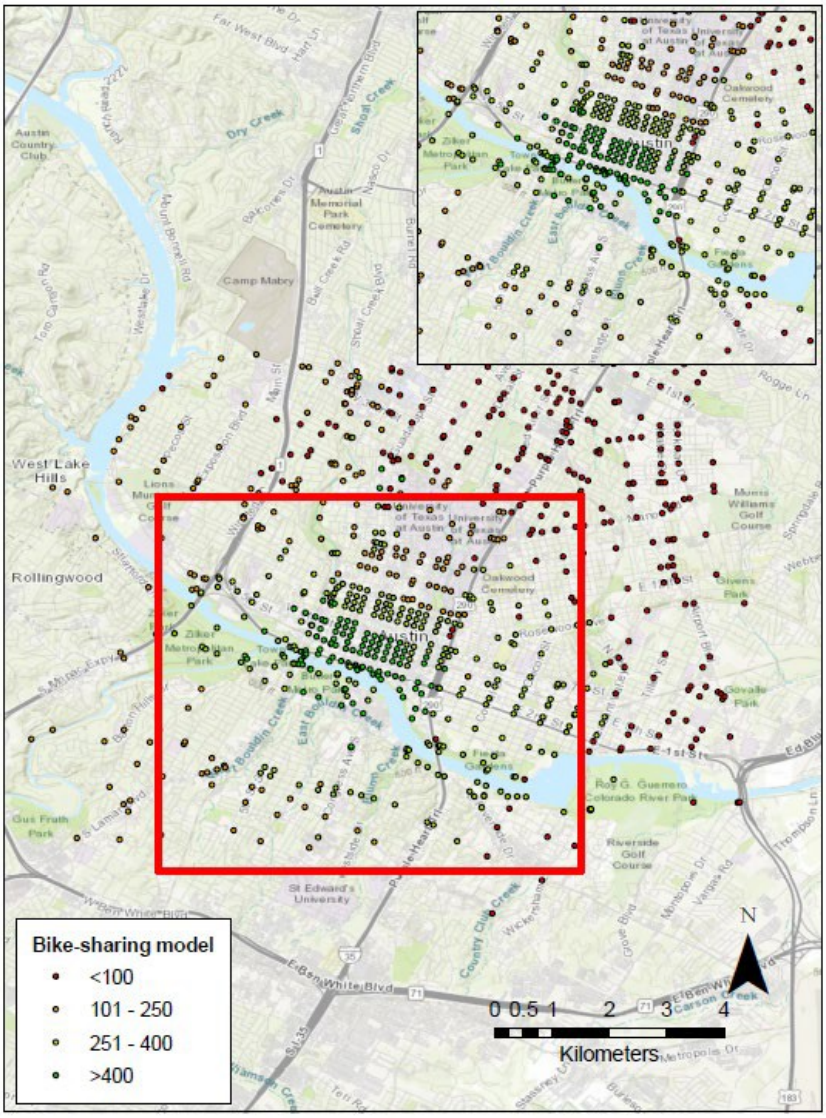

Figure 7. Map. AADB estimates from the bike-sharing model.

Four-Step Model

This study developed a four-step demand model to estimate intersection-level bicycle volume in the Austin area. The general steps of the model were trip generation, trip distribution, modal split, and trip assignment.

The first two steps, trip generation and trip distribution at the TAZ level, were performed using the traffic forecasting model tools from the local transportation planning agency, known as the Capital Area Metropolitan Planning Organization (CAMPO). The trip generation step computed the number of trip ends (daily) produced in and/or attracted to each TAZ of the study area, based on sociodemographic and land-use information, for multiple trip purposes (i.e., home-based work, non-home-based work, and non-home-based other). The step utilized TripCAL6 in TexPACK v3.0

beta version, which was created by TTI travel forecasting program staff (https://travel-forecasting.tti.tamu.edu/) for the demographic preparation and trip generation process. For the trip distribution step, the outputs from TripCAL6 in TexPACK v3.0 were converted to conform to the format of the CAMPO 2010 model. The CAMPO model uses a standard gravity model equation and applies friction factors to represent the effects of impedance (i.e., travel time, spatial separation) between zones. The output of the trip distribution step is an OD matrix that, for each trip purpose, indicates the travel flow between each pair of TAZs.

As the trip distribution output for multiple trip purposes, including all modes of transportation, it is necessary to develop a mode choice model to obtain an OD matrix for bicycle traffic. To do so, the general framework of the nonmotorized demand model proposed by the National Cooperative Highway Research Program [97] was followed. The main dataset utilized for this analysis was 2017 NHTS add-on data for the Austin region. To facilitate the model-building process, TAZ-level socioeconomic and land-use variables were also used.

For the study area, 27,950 trips were extracted, of which 353 were bicycle trips. Given that the bicycle trips only occurred for a limited number of purposes (i.e., no bike trips for pick-up/drop-off purposes), only trips of purpose that had at least one bike trip were selected. Then, the TAZ location of each trip's origin and destination (based on coordinates) was identified. As the final step of data generation for the model building process, the land-use variables for each TAZ were matched to each trip's origin and destination TAZ. Thus, one trip was associated with two features of each land-use variable, denoted as origin land use and destination land use.

Three mode choice models were developed: home-based non-work trip model, home-based work trip model, and non-home-based work trip model. For all three trip purposes, the binary logit model was used to estimate an OD score (based on utility) based on the variables of distance between origin and destination (skim), origin land use, and destination land use. The main rationale was that the decision to bike depends on the characteristics of both origin and destination. Table 6 presents the mode choice model results for the three trip purposes.

Table 6. Binary Logit Model for Three Trip Purposes

Variable	Home-Based Non-Work Model		Non-Home-Based Work Model		Home-Based Work Model	
	Estimate	T-Stat	Estimate	T-Stat	Estimate	T-Stat
(Intercept)	0.77	1.90	2.11	2.53	2.08	4.03
Skim (distance)	−0.30	−4.76	−0.18	−1.58	−0.19	−3.35
Origin—Frequency of transit stop	0.06	1.28	—	—	—	—
Origin—Mixed land use	0.31	1.44	—	—	−1.65	−1.38
Destination—Frequency of transit stop	0.06	1.34	—	—	—	—
Destination—Mixed land use	0.50	1.51	—	—	—	—
Destination—High comfort bike facility	0.0003	1.47	0.0004	1.18		
Origin—Low comfort bike facility	—	—	−0.0005	−1.28	−0.0002	−1.81
Destination—Low comfort bike facility	—	—	—	—	−0.0001	−1.04
Origin—Commercial land use	—	—	−0.06	−0.93	—	—
Destination—Commercial land use	—	—	−0.09	−1.66	—	—
Misclassification error	0.24		0.19		0.25	
Receiver operating characteristic (ROC)	0.76		0.82		0.74	

The final model variables were then used to calculate the OD score for each TAZ pair and for each trip purpose. The computed OD score was categorized into several bins to estimate the rates of mode split for each bin by trip purpose, as outlined by Kuzmyak et al. [97]. The graph and equation developed from this step exhibited a clear pattern of higher rates of bike mode share for TAZ pairs of higher OD scores. The OD score and mode split relationship were then used to estimate bicycle trips for the study area. The output of the process was a trip distribution table, at the TAZ level, for bicycle traffic. Since bike trips are generally short, TAZs that were more than 20 car minutes away from each other were removed from the trip distribution table.

Finally, Tranus was used for the trip assignment step to allocate the trips at the intersections of the study area. The model outcome was found to underestimate intersection-level volume when compared with the actual AADB. This underestimation was probably due to the model being developed using the demographic and trip characteristics of the 2010 CAMPO model. Thus, to

scale the volume to 2017, actual AADB was used in a negative binomial model to estimate AADB from the four-step model.

Figure 8 presents the spatial distribution of the AADB estimates resulting from the four-step model.

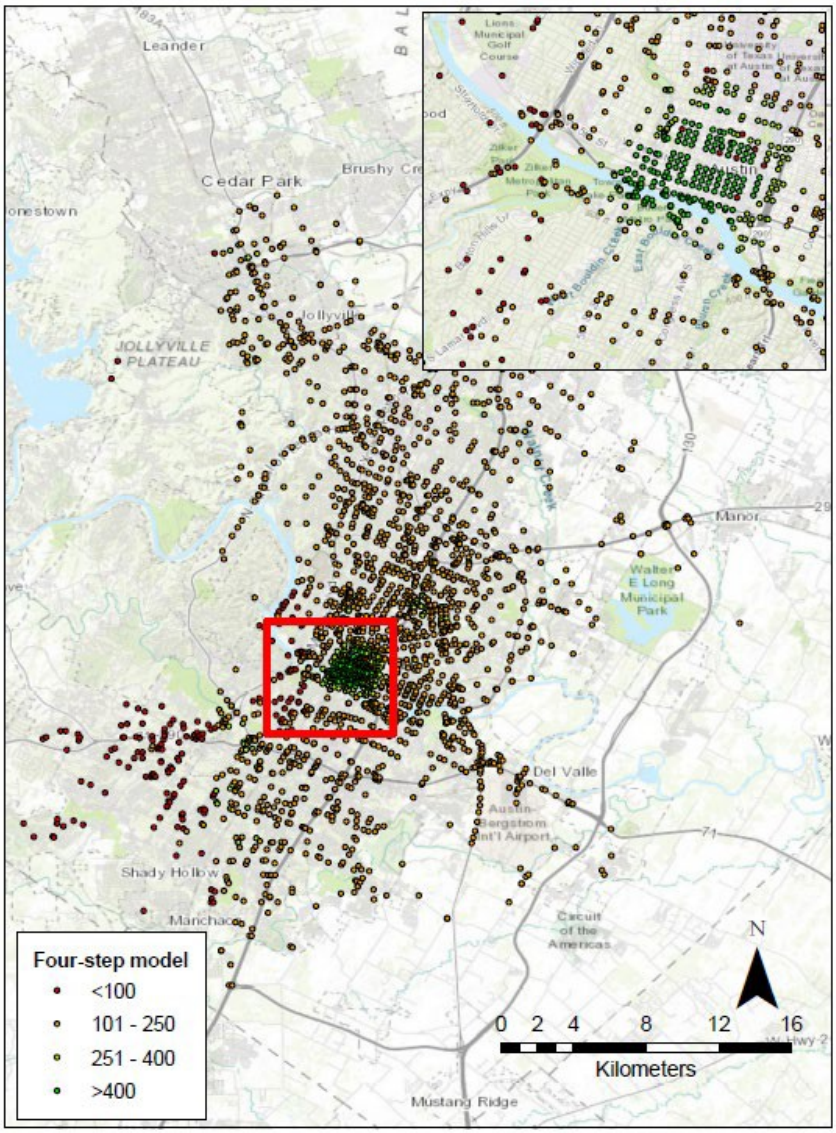

Figure 8. Map. AADB estimates from the four-step model.

Strava Model

Strava Metro is a data service that produces anonymized and aggregated activity data from users of the Strava app, which allows cyclists and runners to track their activities (such as rides, runs, and walks) on a smartphone or other GPS device. Strava allows transportation agencies, city governments, and corporations to access the data in a subscription-based format. Lee and Sener

[98] provide an extensive review of Strava Metro data for bicycling monitoring. For this study, bike activity data were obtained from Strava Metro through TxDOT.

The obtained dataset contains three subsets in three formats: streets, origin-destination, and nodes. Because the spatial unit for this study was the intersection, node-level data (street intersections) were extracted. The researchers processed the total bicycle volume count for all nodes for the year 2017 to obtain the daily average estimate. In order to overlay the Strava nodes with the street intersections, the bicycle street network for the study area was used. The process extracted the Strava activity count for 2,303 intersections. Although Strava provides a large sample size with enhanced temporal and spatial resolution, it only represents a subpopulation. In order to scale the volume to the population level, the relationship between the actual AADB and Strava volume was built utilizing a negative binomial model, with the intersection density as an explanatory variable. Given that the intention here was to conduct a population-level scaling for Strava data, the Strava model was intentionally kept simple with one explanatory (or independent) variable based on its significance (at the 95% confidence level) as well as the performance of the model.

Figure 9 presents the spatial distribution of the AADB estimates resulting from the Strava model.

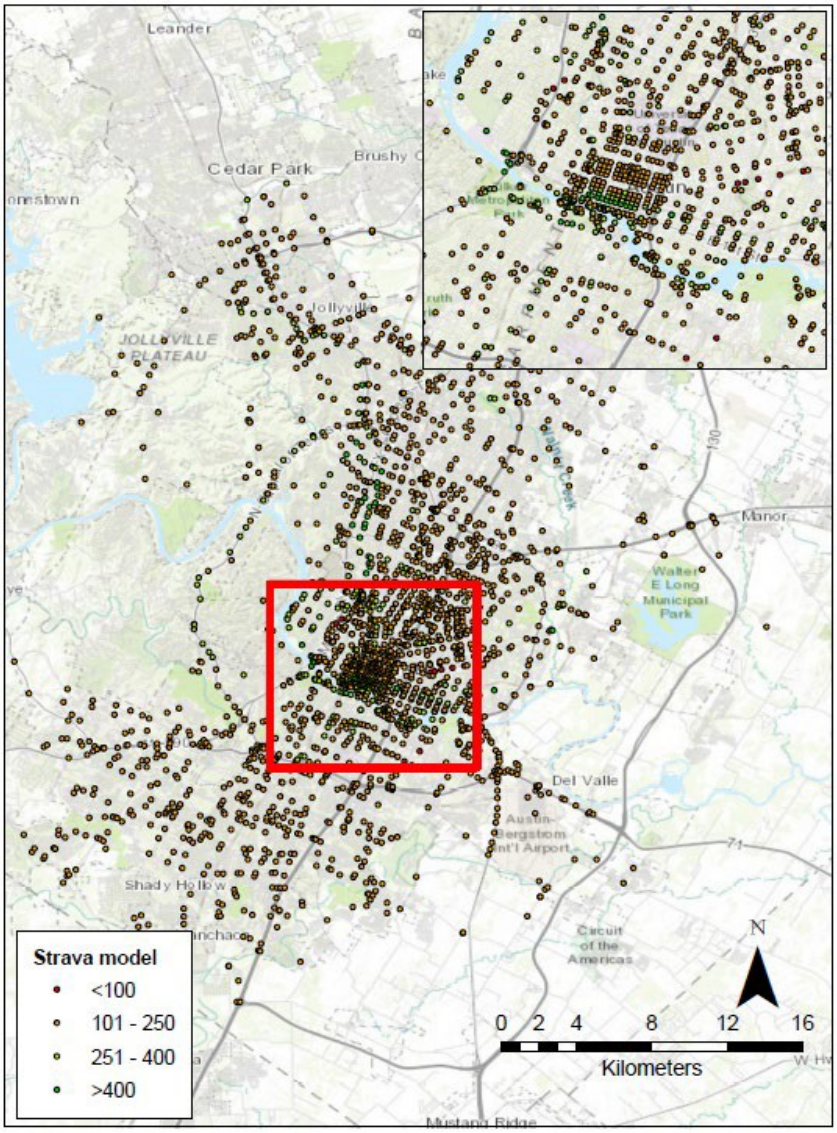

Figure 9. Map. AADB estimates from the Strava model.

StreetLight Model

StreetLight generates data representing walking and biking activity metrics that are derived from three main sources: general location-based services data, mode-tagged location-based services data, and validated bicycle and pedestrian counts [99]. Additional sources, such as GPS-enabled travel diaries and traditional surveys about active mode behavior, are also used during the algorithmic development of the metrics.

The raw datasets go through a series of data processing steps to measure the active mode trips for an area. The platform utilizes a probabilistic approach for mode inference (car, bike, walk) based on machine learning models and using multiple trip-related features. StreetLight [99] has noted that due to the relatively low sample, pedestrian and bike activity are not adjusted for population

biases. To represent activity metrics, StreetLight generates index values that reflect a sample of bicycle trips starting in, passing through, or ending in defined zones.

This research gathered StreetLight Index data for bike traffic in terms of annual average daily volume in 2018 for 950 intersections of the study area. In order to scale the volume to the population level, the relationship between the actual AADB and StreetLight Index was built utilizing a negative binomial model with the population under age 14 (within a 1-mi buffer) as an explanatory variable. Given that the intention here was to conduct a population-level scaling for StreetLight data, the StreetLight model was intentionally kept simple with one explanatory (or independent) variable based on its significance at the 95% confidence level as well as the performance of the model.

Figure 10 presents the spatial distribution of the AADB estimates resulting from the StreetLight Model.

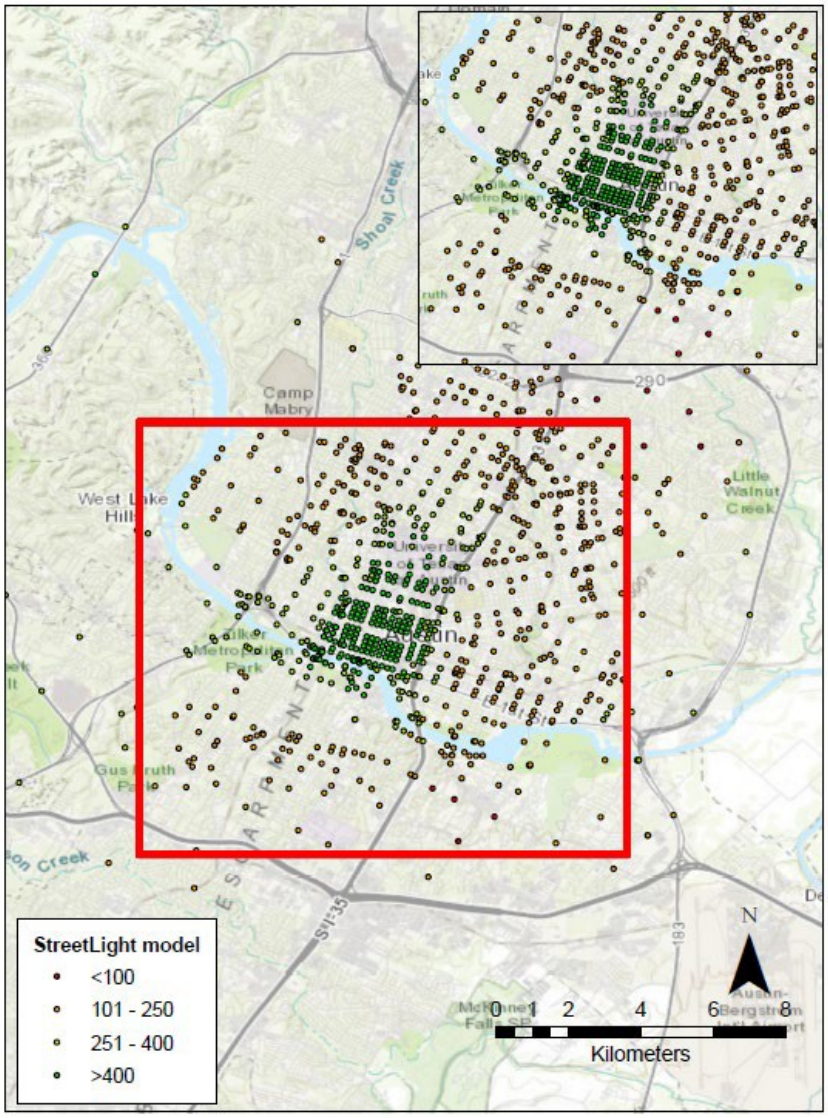

Figure 10. Map. AADB estimates from the StreetLight model.

Appendix B

Appendix B provides the spatial clustering of the high or low crash risk locations in Austin— using the fused AADB estimate obtained from the proposed fusion algorithm and based on the macro crash analysis (i.e., block group level hot-spot analysis of bicycle crashes) conducted in this study.

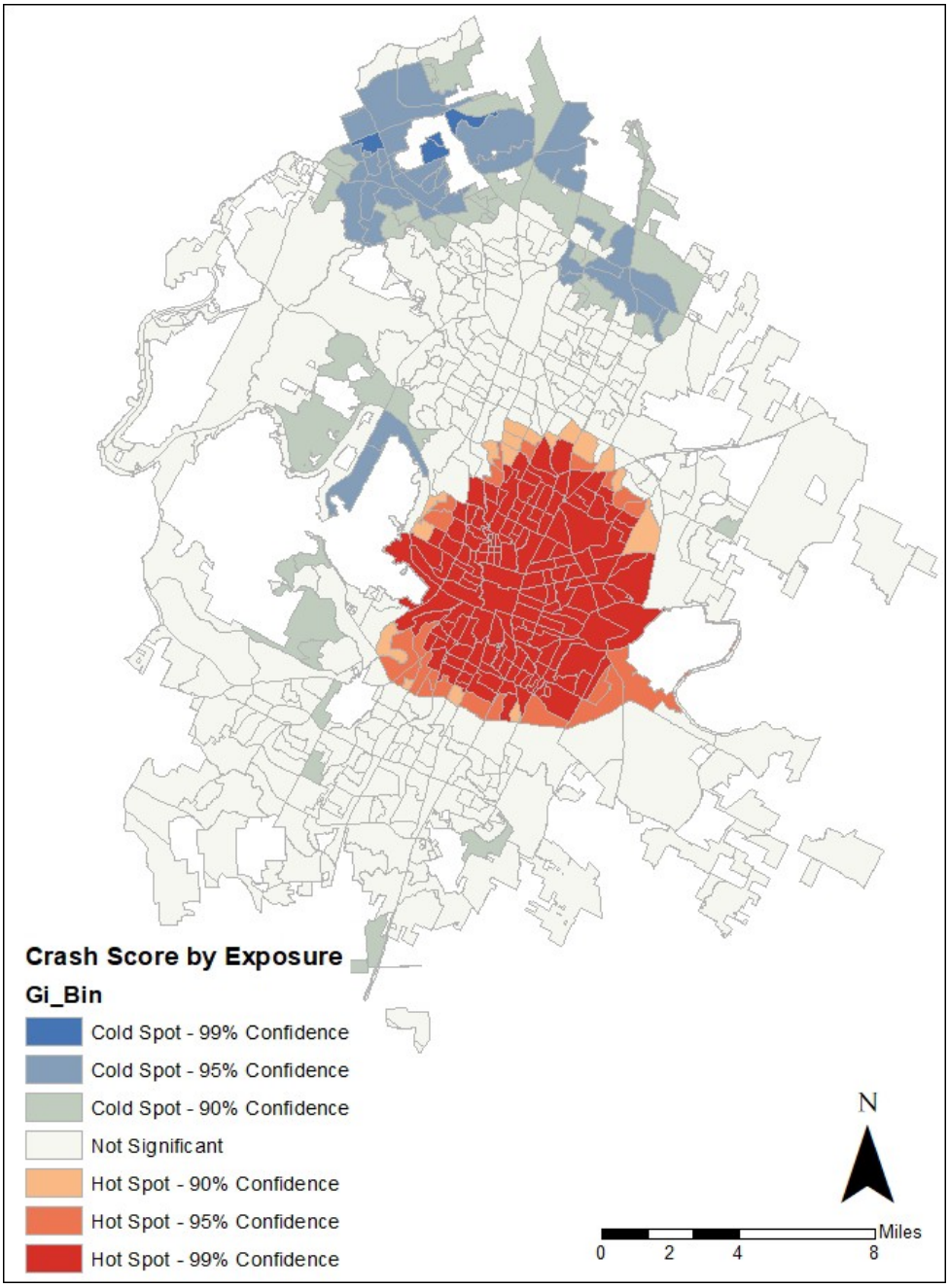

Figure 11. Map. Hotspot analysis.

Appendix C

Appendix C provides the results of the micro analysis crash models (i.e., binary logit models of the reasons for not biking more) developed in this study for the Austin case study.

Variable		Buffer	*(Model 1)* Street Crossings Unsafe, or Heavy Traffic with Too Many Cars		*(Model 2)* No or Poor Condition of Sidewalks, or Not Enough Lighting at Night	
			Estimate	T-Stat	Estimate	T-Stat
(Intercept)		None	−3.25	−1.89	0.04	0.04
Max AADB Category [ref = high (above 400)]	Low (less than 250)	1.0 mi	1.05	1.97	−1.51	−2.02
	Medium (251 to 400)	1.0 mi	1.60	2.93	−1.51	−2.03
Log of Zonal AADT		1.0 mi	0.20	1.18	-	-
Mixed Land Use		0.5 mi	0.22	2.39	-	-
Number of Street Lights		1.0 mi	-	-	−0.001	−2.47
Count of Bike Crashes		0.1 mi	-	-	0.32	1.79
Count of Walk Trips for Exercise in Past 7 days		None	-	-	0.10	2.24
Count of Bike Trips in Past 7 Days [ref = "less than three times"]	Three or More	None	0.78	2.61	-	-
Count of Public Transit Usage in Past 30 Days [ref = "never"]	At Least Once	None	0.82	2.49	0.57	1.31
Gender (ref = male)	Female	None	0.37	1.22	-	-
Household Income Category [ref = "less than 100k"]	100k and above	None	-	-	−1.58	−3.65
Sample Size			223		203	
Log-likelihood at constant			−151.81		−97.87	
Log-likelihood of the final model			−138.70		−81.66	
Log-likelihood ratio test			$\chi2(7) = 26.22$, $p<0.0005$		$\chi2(7) = 32.42$, $p<0.0005$	

Appendix D

Appendix D provides the full list of publications and presentations that have been (or will be) a result of this project.

Peer-Reviewed Journal Articles

- Munira, S., Sener, I. N., & Zhang, Y. (2021). Estimating bicycle demand in the Austin, Texas Area: Role of a bikeability index. *Journal of Urban Planning and Development*, *147*(3), 04021036.
- Munira, S., & Sener, I. N. (2020). A geographically weighted regression model to examine the spatial variation of the socioeconomic and land-use factors associated with Strava bike activity in Austin, Texas. *Journal of Transport Geography*, *88*, 102865.
- Lee, K., & Sener, I. N. (2021). Strava Metro data for bicycle monitoring: A literature review. *Transport Reviews*, *41*(1), 27–47.

The following paper was produced as a collaborative effort from this project and another SAFE-D project, entitled Data Mining to Improve Planning for Pedestrian and Bicyclists Safety.

- Lee, K., & Sener, I. N. (2020). Emerging data for pedestrian and bicycle monitoring: Sources and applications. *Transportation Research Interdisciplinary Perspectives*, *4*, 100095.

Conference Presentations

- Munira, S., & Sener, I. N. Examining the spatial variation of the socioeconomic and land-use factors associated with bike activity: A case study using crowdsourced Strava data in Austin, Texas. Presented at the *International Conference on Transport & Health*, Virtual Conference, June 14–30, 2021.
- Lee, K., & Sener, I. N. Emerging data for pedestrian and bicycle monitoring: Sources and applications. Presented at the *15th World Conference on Transport Research*, Mumbai, India, May 26–31, 2019.

Manuscripts in Preparation

- The role of crowdsourced data in understanding nonmotorized demand: A case study for the City of Austin based on Strava and StreetLight data
- Direct demand modeling in estimating nonmotorized activity: A literature review
- Understanding nonmotorized traffic data and fusion mechanisms for a better demand/exposure estimate
- Decision fusion for nonmotorized traffic data and Dempster Shafer with context credibility: Framework and validation
- Where they live matters: Exploring walk and bike perception to support policy regarding neighborhood infrastructures

Appendix E

Appendix E describes the characteristics of the data used in this research study.

Link to datasets: here on the Safe-D Dataverse Site

Project Description

This project explored an emerging research territory, the fusion of nonmotorized traffic data for estimating reliable and robust exposure measures. The researchers developed fusion mechanisms to combine five bike demand data sources in Austin and demonstrated the applicability of the fused estimate in two crash analyses. The data used in this study were gathered from five primary data sources: (a) actual bicycle volume counts, (b) bicycle-sharing data, (c) NHTS add-on data, (d) Strava data, and (e) StreetLight data. In addition, the sociodemographic and land-use data for building models were obtained from the American Community Survey, Austin Transportation Department, and other public data domains.

Data Scope

Table 7 provides a generic description of the datasets used in this project.

	Data Sources	Temporal Coverage or Year of Data Collection	Spatial Coverage	Source	Data Access
Primary bicycle data sources	Video-Based Short Count	24-hour count in 2017	44 intersections	City of Austin Transportation Department	Requested
	Inductive Loop-Based Permanent Count	Continuous count from 2012 to 2017	11 locations	Eco-Counter	Requested
	NHTS Add-On Survey	Gathered in 2017	1,095 households in Austin	TxDOT	Requested
	Strava Metro Data	Trips in 2017	2,303 intersections in Austin	TxDOT (through an internal agreement with Strava)	Requested
	Bike-Sharing Data	Trips in 2017	63 stations in Downtown Austin	https://data.austintexas.gov/Transportation-and-Mobility/Austin-MetroBike-Trips/tyfh-5r8s	Public Website
	StreetLight Data	Trips in 2018	950 Zones for intersections in Austin	StreetLight Inc.	Requested
Secondary data sources	ACS Data	2017	Census tract or block group level in Austin	https://data.census.gov/cedsci/	Public Website

	Data Sources	Temporal Coverage or Year of Data Collection	Spatial Coverage	Source	Data Access
	City of Austin's Bike Network Data	2018	City of Austin boundary	City of Austin Transportation Department	Requested
	Land-Use Data	2016	City of Austin boundary	https://data.austintexas.gov/Locations-and-Maps/Land-Database-Data-Only-2016/4nsn-uea6/data?pane=feed	Public Website
	School Data	2015	City of Austin boundary	http://schoolsdata2-tea-texas.opendata.arcgis.com/datasets/059432fd0dcb4a208974c235e837c94f_0	Public Website
	Traffic Signal Data	2019	City of Austin boundary	https://data.austintexas.gov/dataset/Traffic-Signals-and-Pedestrian-Signals/p53x-x73x	Public Website
	Transit Data	2018	City of Austin boundary	https://data.texas.gov/Transportation/CapMetro-Shapefiles-JUNE-2018/rwce-6ann	Public Website
	Traffic Volume Data	2017	City of Austin boundary	https://gis-txdot.opendata.arcgis.com/datasets/4dfba4bdbd8044c58e1ce1a1c5fdbcd2_0?geometry=-141.889%2C24.544%2C-58.261%2C37.664	Public Website
	Network Speed Data	2020	Texas	https://data.austintexas.gov/Locations-and-Maps/Street-Centerline/m5w3-uea6	Public Website
	Crash Data	2014-2018	City of Austin boundary	https://cris.dot.state.tx.us/public/Query/app/welcome	Requested
	Employment Data	2017	City of Austin boundary	TxDOT (through an internal agreement with InfoGroup)	Requested

Data Specification

Some of the datasets used in this project were obtained through the public domain, and thus are available for sharing. Some other datasets were requested from corresponding data providers or government agencies and obtained under specific data use agreements, and thus are not available directly from the authors. With these data access regulations in mind, following are details on the data specifications and/or reference information of each dataset.

Primary Data Sources

Short-Count Data: The bike short-count data were obtained from the City of Austin Transportation Department. The city holds an Open Data Inventory available at https://data.mobility.austin.gov/open-data/. The inventory "provides comprehensive list of public transportation datasets made available by the City of Austin Transportation Department. This inventory also identifies datasets which may be under active development, or datasets which are not currently available but have been identified for future development" [100].

- Requests related to the bicycle short-count data should be directed to the City of Austin Transportation Department.

Continuous-Count Data: The bike continuous-count data were obtained from Eco-Counter, which is a company that assists with continuous data collection for pedestrian and bicyclists in specific locations across cities around the world [95].

- Requests related to the continuous-count data should be directed to Eco-Counter.

NHTS TxDOT Add-On Survey: The 2017 TxDOT NHTS add-on survey data were obtained from TxDOT. NHTS data specifications are available online at https://www.txdot.gov/government/enforcement/data-access.html.

Requests related to the continuous-count data should be directed to TxDOT.

Strava Data: The Strava Metro data were obtained from TxDOT—through its internal agreement with Strava. Details on Strava Metro data can be found at https://metro.strava.com/.

Requests related to the Strava Metro data should be directed to Strava.

StreetLight Data: The StreetLight data were obtained from StreetLight. Details on StreetLight data can be found at https://www.streetlightdata.com/.

Requests related to the StreetLight data should be directed to StreetLight.

Bike-Sharing Data: The bike-sharing trip data were obtained from the BCycle (Austin bike-sharing agency) data portal, available online at https://data.austintexas.gov/Transportation-and-Mobility/Austin-MetroBike-Trips/tyfh-5r8s.

Secondary Data Sources

ACS Data: The ACS datasets were obtained from the ACS data inventory at https://data.census.gov/cedsci/.

- The processed ACS data include sociodemographic variables at three buffer zones (0.1 mi, 0.5 mi, 1 mi) around the intersections of the study area.

City of Austin's Bike Network Data: The bike network data were obtained from the City of Austin Transportation Department. Detailed specifications can be found at

https://services.arcgis.com/0L95CJ0VTaxqcmED/arcgis/rest/services/TRANSPORTATION_bicycle_facilities/FeatureServer.

Requests related to the bike network data should be directed to the City of Austin Transportation Department.

Land-Use Data: The land-use data were obtained from Austin's open data portal at https://data.austintexas.gov/Locations-and-Maps/Land-Database-Data-Only-2016/4nsn-uea6/data?pane=feed.

The processed land-use data include counts of land use across different categories at three buffer zones (0.1 mi, 0.5 mi, 1 mi) around the intersections of the study area.

School Data: The school data were obtained from the Texas Education Agency's public open data site, available online at http://schoolsdata2-tea-texas.opendata.arcgis.com/datasets/059432fd0dcb4a208974c235e837c94f_0.

The processed school data include counts of schools across three buffer zones (0.1 mi, 0.5 mi, 1 mi) around the intersections of the study area.

Traffic Signal Data: The traffic signal data were obtained from Austin's open data portal at https://data.austintexas.gov/dataset/Traffic-Signals-and-Pedestrian-Signals/p53x-x73x.

The processed signal data include counts of traffic signals across three buffer zones (0.1 mi, 0.5 mi, 1 mi) around the intersections of the study area.

Transit Data: The transit data were obtained from Capital Metro's open data site at https://data.texas.gov/Transportation/CapMetro-Shapefiles-JUNE-2018/rwce-6ann.

The processed transit data include counts of transit stops across three buffer zones (0.1 mi, 0.5 mi, 1 mi) around the intersections of the study area.

Traffic Volume Data: The traffic volume (i.e., AADT) data were obtained from TxDOT's open data portal at https://gis-txdot.opendata.arcgis.com/datasets/4dfba4bdbd8044c58e1ce1a1c5fdbcd2_0?geometry=-141.889%2C24.544%2C-58.261%2C37.664.

The processed data include average AADT across three buffer zones (0.1 mi, 0.5 mi, 1 mi) around the intersections of the study area.

Network Speed Data: The network speed data were obtained from Austin's open data portal at https://data.austintexas.gov/Locations-and-Maps/Street-Centerline/m5w3-uea6.

The processed data include average network speed across three buffer zones (0.1 mi, 0.5 mi, 1 mi) around the intersections of the study area.

Crash Data: The crash data were obtained from CRIS at https://cris.dot.state.tx.us/public/Query/app/welcome.

Requests related to the crash data should be directed to CRIS.

Employment Data: The employment data were obtained from TxDOT—through its internal agreement with Infogroup.

Requests related to the employment data should be directed to TxDOT.

Estimated Data Sources

Fused Bike Volume Data: The fused bike volume data include fused bike volume or exposure estimates at the intersections of the study area.

Citation Metadata

Author list with researcher ORCIDs (for datasets directly available from the study authors):

Ipek Nese Sener, 0000-0001-5493-8756
Sirajum Munira, 0000-0002-4953-2628

Contact information (email) for corresponding author (project PI): i-sener@tti.tamu.edu

Keywords: Fusion, exposure, nonmotorized activity, demand models, safety analysis, crowdsourced data, Dempster Shafer