Emerging Data Cleaning and Fusion for Traffic Model Calibration—Data Fusion for Microsimulation Model Calibration

PUBLICATION NO. FHWA-HRT-24-142

DECEMBER 2024

U.S. Department of Transportation
Federal Highway Administration

Research, Development, and Technology
Turner-Fairbank Highway Research Center
6300 Georgetown Pike
McLean, VA 22101-2296

FOREWORD

Traffic analysis and associated traffic analysis tools are critical to help State and local transportation agencies make decisions related to transportation investments. These tools, particularly traffic simulation tools, require extensive datasets so that the underlying models within these tools can be calibrated to real-world conditions. New data sources such as probe data, trajectory data, and connected vehicle data provide both opportunities and challenges for use in traffic simulation tools. These new data sources can be combined with existing, more traditional datasets to support the calibration of traffic simulation models.

This report focuses on how these new sources of data can be used in traffic simulation analyses. The report will be of interest to researchers, traffic analysts, and State and local transportation agencies who are interested in incorporating new sources of data into their transportation decisionmaking processes.

Carl Andersen
Acting Director, Office of Safety and
Operations Research and Development

Notice

This document is disseminated under the sponsorship of the U.S. Department of Transportation (USDOT) in the interest of information exchange. The U.S. Government assumes no liability for the use of the information contained in this document.

Non-Binding Contents

Except for the statutes and regulations cited, the contents of this document do not have the force and effect of law and are not meant to bind the States or the public in any way. This document is intended only to provide information regarding existing requirements under the law or agency policies.

Quality Assurance Statement

The Federal Highway Administration (FHWA) provides high-quality information to serve Government, industry, and the public in a manner that promotes public understanding. Standards and policies are used to ensure and maximize the quality, objectivity, utility, and integrity of its information. FHWA periodically reviews quality issues and adjusts its programs and processes to ensure continuous quality improvement.

Disclaimer for Product Names and Manufacturers

The U.S. Government does not endorse products or manufacturers. Trademarks or manufacturers' names appear in this document only because they are considered essential to the objective of the document. They are included for informational purposes only and are not intended to reflect a preference, approval, or endorsement of any one product or entity.

Recommended citation: Federal Highway Administration, *Emerging Data Cleaning and Fusion for Traffic Model Calibration–Data Fusion for Microsimulation Model Calibration* (Washington, DC: 2024) https://doi.org/10.21949/1521587

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. FHWA-HRT-24-142	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle Emerging Data Cleaning and Fusion for Traffic Model Calibration–Data Fusion for Microsimulation Model Calibration		5. Report Date December 2024	
		6. Performing Organization Code:	
7. Author(s) Xuesong (Simon) Zhou (0000-0002-9963-5369), Xiangyong (Roy) Luo, Mohammad Abbasi, Zhitong Huang (0000-0003-2871-6302), and Ankur Tyagi		8. Performing Organization Report No.	
9. Performing Organization Name and Address Leidos Inc. 1750 Presidents St. Reston, VA 20190		10. Work Unit No.	
		11. Contract or Grant No. 693JJ321D000010-693JJ322F00274N	
12. Sponsoring Agency Name and Address U.S. Department of Transportation Federal Highway Administration 1200 New Jersey Avenue, SE Washington, DC 20590		13. Type of Report and Period Covered Final Report; July 2022–August 2023	
		14. Sponsoring Agency Code HRSO	
15. Supplementary Notes The contracting officer's representative was Jennifer Foley, HOFM: (0000-0002-2695-2339).			
16. Abstract This report examines emerging data sources applicable to traffic simulation model calibration. The project team categorized these data sources based on publicly accessible datasets and emphasized their potential use for enhancing traffic simulation model calibration. The study evaluates the strengths and limitations of these emerging data sources and recommends methods for optimal data preparation and calibration. The report adapts the Federal Highway Administration's data fusion framework and *Traffic Analysis Toolbox Volume 3: Guidelines for Applying Traffic Microsimulation Modeling Software (2019 Update)* to cater to traffic microsimulation model calibration (Hale et al. 2022; Wunderlich et al. 2019). This approach enables practitioners to use emerging data sources to address the unique needs of traffic microsimulation calibration. This report aims to equip users with an understanding of these potential data sources and their effective integration in traffic microsimulation and analysis. The findings represent a key point for future advancement of research, development, and application of data-driven microsimulation calibration in transportation engineering.			
17. Key Words Data fusion, emerging data, traffic simulation, calibration, and microsimulation		18. Distribution Statement No restrictions. This document is available to the public through the National Technical Information Service, Springfield, VA 22161. https://www.ntis.gov	
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 72	22. Price N/A

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

SI* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
		AREA		
in^2	square inches	645.2	square millimeters	mm^2
ft^2	square feet	0.093	square meters	m^2
yd^2	square yard	0.836	square meters	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	square kilometers	km^2
		VOLUME		
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
	NOTE: volumes greater than 1,000 L shall be shown in m^3			
		MASS		
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2,000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
		TEMPERATURE (exact degrees)		
°F	Fahrenheit	5 (F-32)/9 or (F-32)/1.8	Celsius	°C
		ILLUMINATION		
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	candela/m^2	cd/m^2
		FORCE and PRESSURE or STRESS		
lbf	poundforce	4.45	newtons	N
lbf/in^2	poundforce per square inch	6.89	kilopascals	kPa

APPROXIMATE CONVERSIONS FROM SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
		AREA		
mm^2	square millimeters	0.0016	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.195	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	square kilometers	0.386	square miles	mi^2
		VOLUME		
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	cubic meters	35.314	cubic feet	ft^3
m^3	cubic meters	1.307	cubic yards	yd^3
		MASS		
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2,000 lb)	T
		TEMPERATURE (exact degrees)		
°C	Celsius	1.8C+32	Fahrenheit	°F
		ILLUMINATION		
lx	lux	0.0929	foot-candles	fc
cd/m^2	candela/m2	0.2919	foot-Lamberts	fl
		FORCE and PRESSURE or STRESS		
N	newtons	2.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in^2

*SI is the symbol for International System of Units. Appropriate rounding should be made to comply with Section 4 of ASTM E380. (Revised March 2003)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

3D	three dimensional
AADT	annual average daily traffic
AVI	automatic vehicle identification
BSM	basic safety message
Caltrans	California Department of Transportation
CAV	connected and automated vehicle
CBI	congestion bottleneck identification
DOT	department of transportation
DTALite	Dynamic Traffic Assignment Lite
FD	fundamental diagram
FHWA	Federal Highway Administration
GMNS	General Modeling Network Specification
GPS	Global Positioning System
HCM	*Highway Capacity Manual*
HighD	highway drone
Hz	hertz
ID	identifier
ITS	intelligent transportation system
LBS	location-based service
LiDAR	Light Detection and Ranging
LOS	level of service
MAG	Maricopa Association of Governments
MPE	model parameter estimation
NGSIM	Next Generation Simulation
OD	origin-destination
PeMS	Performance Measurement System
QPE	queue profile estimation
SUMO	Simulation of Urban MObility
TMC	Traffic Message Channel
TSE	traffic state estimation
UAV	unmanned aerial vehicle
veh/h	vehicles per hour

CHAPTER 1. BACKGROUND AND STUDY SCOPE

BACKGROUND AND STUDY SCOPE

In the rapidly evolving transportation landscape, recognizing the unique features of various emerging data sources and applying data cleaning and fusion strategies have become important areas of focus for researchers and practitioners. Located at the intersection of advanced communication technologies, data management tools, and contemporary data analytics, the traffic simulation model calibration field is enriched by an influx of data from connected and automated vehicles (CAV) and advanced intelligent transportation system (ITS) infrastructure capabilities.

This report systematically reviews the state of the art and state of practice in emerging data cleaning and fusion methodologies and their potential application in traffic simulation model calibration. Given the diverse nature of data sources and the spectrum of cleaning and fusion methods, this report summarizes the key characteristics of these data sources, the challenges involved in cleaning and fusing such data, and recommended best practices for traffic simulation model calibration.

Every data source—from legacy infrastructure-based sensors, such as loop detectors and radars, to emerging data sources, including probe data, connected vehicle data, high-resolution trajectory data from aerial sources, or video data—possesses strengths and weaknesses. This report addresses the barriers that hinder widespread adoption of emerging data in traffic analysis applications and traffic simulation model calibration and provides insight into effectively and efficiently leveraging emerging data sources for traffic simulation model calibration.

The report focuses on the following key objectives:

- **Clarification of terminologies and concepts:** This report explains the terminologies and concepts associated with emerging data cleaning and fusion. It delves into the relationships among emerging data cleaning and fusion and various sources of transportation data.

- **Insights from literature review and applications:** The report synthesizes insights from an extensive literature review and consultation with users and vendors of data cleaning and fusion tools. It examines real-world applications of emerging data cleaning and fusion in recent transportation studies, sharing valuable experiences and lessons learned from these endeavors.

- **Customization of the data fusion framework:** This section focuses on tailoring the data fusion framework, originally developed by the Federal Highway Administration (FHWA) Office of Operations, for microsimulation model calibration (Hale et al. 2022). The framework encompasses five steps: data acquisition and storage, data cleaning and fusion, data analysis, decision implementation, and evaluation and iteration. The framework's design enables transportation agencies, including nontechnical personnel, to effectively harness emerging data sources and enhance their decisionmaking capabilities.

By addressing these objectives, the report serves as a practical resource, providing guidance and insights to facilitate the seamless integration of emerging data in traffic simulation model calibration.

TERMINOLOGY USED IN THIS REPORT

The following key terms related to traffic simulation and modeling are used in this report:

- **Calibration:** The process in which the analyst selects the model parameters that cause the model to best reproduce field-measured local traffic operation conditions.
- **Microsimulation:** The process of modeling individual vehicle movements on a second or subsecond basis to assess the traffic performance of highway and street systems.
- **Model:** The specific combination of modeling software and input parameters developed by analysts for a specific application.
- **Project:** This term is limited to the physical road improvement being studied to reduce confusing the analysis of a project with the project itself.
- **Software:** Several models can be developed using a single software program. These models will share the same basic computational algorithms embedded in the software.
- **Validation:** The process in which the analyst checks the overall model-produced traffic performance for a street or road system against field measurements of traffic performance, such as traffic volumes, travel times, average speeds, and average delays.

CHAPTER 2. IDENTIFYING AND CHARACTERIZING EMERGING DATA SOURCES FOR TRAFFIC SIMULATION CALIBRATION

This chapter identifies and characterizes potential emerging data sources for traffic simulation model calibration. The first section focuses on identifying potential emerging data sources for traffic simulation model calibration and suggests a list of emerging data sources from State department of transportation (DOT) stakeholders and recent vendor contacts, based on their interests. This list includes questions designed to illuminate potential misunderstandings and complexities associated with the traffic simulation model at hand. Categorizing the identified data sources and datasets aims to provide an organized perspective on potential data that could be used. The second section in this chapter focuses on characterizing the identified emerging data sources from an analytical data fusion and data quality perspective. Data sources are categorized based on their performance metrics, such as reliability, accuracy, granularity, timeliness, and cost. This categorization is key for understanding the applicability and effectiveness of these data sources.

IDENTIFYING POTENTIAL EMERGING DATA SOURCES FOR TRAFFIC SIMULATION MODEL CALIBRATION

Categories of Emerging Data Sources

Different State DOTs have expressed joint interest in exploring emerging data sources, the potential data integration steps for *Traffic Analysis Toolbox Volume 3: Guidelines for Applying Traffic Microsimulation Modeling Software (2019 Update)* (Wunderlich et al. 2019). State agencies have also recognized the significance of the Regional Integrated Transportation Information System, which integrates real-time data from location-based service (LBS) and probe data to effectively monitor work zone performance measures.

Based on the information provided from the stakeholders and literature review, the project team has classified emerging data sources into the following categories:

- **Probe vehicle data** encompass data directly obtained from vehicles and connected devices, ensuring high reliability, accuracy, and granularity. The data are updated, although the cost may be a consideration due to data acquisition and processing. Probe vehicle data are used by many State DOTs for traffic simulation model calibration, and are used for analyzing various traffic parameters, such as hard braking, queues, speeds, and secondary crashes. Another type of probe vehicle data and related dashboards are used by traffic engineers for tracking work zone delays, queue lengths, and speeds (INRIX, 2004; Airsage, 2000; Wejo, 2013, Gupta, 2007).

- **High-resolution vehicle trajectory data** offer reliability and accuracy as they are specifically collected for focused analysis. They provide detailed per-vehicle trajectory data, which are valuable for model calibration. However, the cost can be high due to data acquisition, processing, and storage. Examples of sample data include the Highway Drone (HighD) vehicle datasets, which describe position, speed, acceleration, and lane changes, as well as classical datasets, such as Next Generation Simulation (NGSIM), and

recent FHWA projects, such as Trajectory Investigation for Enhanced Microsimulation Calibration Guidance (Krajewski 2018; NGSIM 2006; Hale et al. 2021).

- **CAV data** encompass data from connected vehicle pilot projects, the Waymo™ Open Dataset (Tancik et al. 2022), and the OpenACC dataset (European Commission 2020). CAV data are highly reliable and accurate, providing insights into the behavior of advanced vehicles. These data offer a high level of granularity and real-time updates, but the cost can be significant due to the advanced technology involved.

- **Crowdsourced platform data** exhibit varying levels of reliability and accuracy, as they rely on user-generated input. These data provide moderate granularity with detailed incident reports and benefit from high timeliness due to real-time updates. The cost involved is generally low to moderate and is mainly associated with data processing.

- **Emerging sensor technologies and traffic management systems data** (such as Light Detection and Ranging (LiDAR), and three-dimensional (3D) point clouds and their application) can vary in their reliability, accuracy, granularity, timeliness, and cost, depending on the specific technology or platform employed. Examples of sample datasets include KITTI videos and 3D point clouds captured by cameras and LiDAR mounted on vehicles driving in urban and rural areas (Geiger et al. 2013).

Traffic Trajectory Datasets

Since the early 2000s, researchers have been gathering high-resolution trajectory data to study the fundamental laws and traffic flow behaviors in microscopic traffic simulators. Among the available open-source vehicle trajectory datasets, the NGSIM trajectory dataset (NGSIM 2006) is the most extensively used. Its release has attracted enthusiasm among traffic flow researchers for the past decade. Using the NGSIM dataset as a foundation, numerous studies exploring traffic flow have been conducted worldwide, encompassing both microscopic and macroscopic levels, leading to significant new discoveries. Visualization results of the NGSIM dataset can be seen in NGSIM-I–80-Trajectory-Animation (2017) and NEXTA (2018).

Several publicly accessible datasets are available for traffic flow model calibration and studies. The HighD Dataset is a publicly accessible dataset that offers high-quality, naturalistic vehicle trajectories recorded on German highways using drones (Krajewski et al. 2018). HighD is one of the largest datasets of its kind, both in the number of vehicles and the total recording duration. The dataset features diverse traffic scenarios that provide robust modeling and analysis capabilities. It includes vehicle trajectories with a frequency of 25 Hz, offering detailed insights into vehicle behavior. The dataset also provides rich metadata, including vehicle class, length, velocity, acceleration, and lane positioning. Recorded in high definition, the dataset consists of more than 50 recordings capturing more than 110,000 vehicles and 45 h of traffic. The dataset can be used in applications such as driver behavior analysis, traffic flow analysis, autonomous vehicle algorithm development, and traffic simulation model validation.

The Waymo Open Dataset is a high-quality, multimodal sensor dataset obtained from Waymo's self-driving cars (Tancik et al. 2022). Due to its variety and quality of data, the Waymo dataset serves as a resource for the development and validation of self-driving algorithms. The dataset

includes LiDAR data collected by Waymo's proprietary sensors, offering a 360-degree field of view and high-resolution depth information. Additionally, the dataset encompasses high-resolution camera data from multiple angles, which provide visual context and help with object detection and classification tasks. The dataset is accompanied by labels and annotations for different objects in both LiDAR and camera data, facilitating machine learning model training for object detection and segmentation. With a diverse range of driving scenarios captured under different environmental conditions and locations, the dataset allows algorithms to be tested and trained in various scenarios, enhancing their robustness. The dataset also provides synchronized and calibrated sensor data, enabling integration and correlation between different modalities.

The pNEUMA experiment deployed 10 drones over multiple days in the central business district of Athens, Greece (Paipuri et al. 2021). The drones captured traffic streams within a congested area spanning 1.3 square km that covered more than 100 km-lanes of the road network, approximately 100 busy intersections (signalized and non-signalized), and numerous bus stops, and generated nearly half a million trajectories. Using unmanned aerial vehicles (UAVs), the experiment aimed to systematically investigate key traffic phenomena in a multimodal congested environment.

Other publicly accessible datasets include the Berkeley DeepDrive dataset, which provides video data and annotations for object detection and tracking, and the CityFlow dataset, which offers vehicle position, speed, acceleration, and lane-change data, along with traffic signal states and pedestrian movements (Wu et al. 2022; Tang et al. 2019).

Emerging Data Vendors and Datasets for Traffic Flow Analysis and Model Calibration

Several options exist of commercially available emerging data vendors, each offering distinct data attributes for characterization. Certain signal system manufacturers are incorporating advanced equipment into signal systems, communications, and logging systems, which allows collecting and distributing precise signal phase and timing data, as well as detector data and other relevant information. Data from different sources enable multiresolution analyses and model calibrations, specifically including:

- **Speed and traffic signal analytics data:** Automated Traffic Signal Performance Measures (ATSPMs) provide data such as movement based turning counts, vehicle delay, and estimated queue length during peak hours, which can be used for multiresolution traffic simulation models. These products facilitate the assessment of roadway performance, congestion levels, and travel time reliability. For example, leveraging movement counts and other performance measures, transportation demand models can be calibrated to account for real-world conditions, ensuring that traffic planning and simulation models accurately represent actual travel patterns such as path-level volumes.

- **Origin-destination (OD) trips/waypoint data:** Vendors offer comprehensive trip data, including OD information and travel times. These datasets can be helpful for calibrating transportation demand matrix as the key input elements of the simulator and gaining detailed insights into travel patterns.

- **Path and link volume data:** Vendors offer estimated traffic volumes derived from integrated historical data from various sources. This information plays a key role in calibrating transportation demand models and accurately forecasting future traffic patterns. By integrating traditional annual average daily traffic (AADT) and traffic volume profiles with comprehensive network-wide path data, the OD demand and traffic state estimation (TSE) modules could improve the accuracy and precision of traffic simulation models.

- **Micromobility data (e-scooter and bike sharing):** Various data vendors are available that specialize in micromobility data, specifically e-scooter and bike sharing information. These datasets provide insights into the use and movement patterns of micromobility vehicles, enabling a deeper understanding of this emerging mode of transportation.

Mapping from Sources of Data to Traffic Flow Analysis and Model Calibration

The following key mappings can be summarized from the original sources of data to applications such as TSE, ITS, and work zones that are of interest to stakeholders involved in traffic analysis pooled funds:

- **Roadside basic safety message (BSM) data:** Gathering BSM data from properly equipped vehicles provides insights into driver behavior, aiding in the identification of opportunities for roadway safety improvements.

- **High-resolution vehicle trajectory data:** Augmenting or replacing traditional trip and OD studies used by planners and modelers by providing detailed waypoint information.

- **Turning movement counts:** Using real-time turning movement counts during incident response or lane closures to understand alternate route usage and evaluate the effectiveness of communication strategies in influencing travel behavior.

- **Crowdsourced mapping data:** Supplementing and improving State centerline files with publicly generated navigable mapping data.

- **High-resolution map data and other asset management systems:** Using high-resolution mapping data with centimeter-level accuracy, including detailed information on curbs, road markings, and signage for enhanced asset management.

- **Probe-based speed data:** Studying congestion trends, identifying problematic locations, conducting before-and-after evaluations, and prioritizing transportation projects using speed and travel time data from vehicles with navigation systems.

- **Wireless technologies-based reidentification:** Deploying wireless technologies equipment at intersections or decision points to collect travel times and potential route choice patterns on key corridors and arterials.

- **Connected vehicle data from telematics providers:** Extracting vehicle performance measures and warnings directly from connected vehicles, including events such as heavy breaking, traction-control engagement, emissions data, and seatbelt usage in commercial vehicles.

Table 1 provides a summary of the measurement types, data quality, associated costs, and concerns for each surveillance technique used in traffic simulation model calibration.

Table 1. Comparison of data collection techniques for traffic simulation model calibration.

Surveillance Type	Measurement Type	Data Quality	Costs and Concerns
Point detectors, turning movement data.	Vehicle counts and point speed.	High accuracy, effective for incident response and lane closures and relatively low reliability.	Low installation cost and high maintenance cost.
Automatic vehicle identification.	Point-to-point OD, path flow information for tagged or probe vehicles, such as travel time and volume.	Accuracy depends on market penetration level of tagged vehicles.	Relatively high installation costs for automated vehicle identifier (ID) readers, such as Wi-Fi/ Bluetooth®.
Mobile Global Positioning System (GPS) location sensors.	Semicontinuous path trajectory and probe speed for individual equipped vehicles.	Accuracy depends on market penetration level of probe vehicles.	Public privacy concerns.
Trajectory data from video image processing.	Continuous path trajectory for vehicles on different links or lanes.	Accuracy depends on machine vision algorithms.	Relatively high installation cost for overhead video camera and communication wires.
Connected vehicle data and roadside BSM collection.	Detailed vehicle measures and BSMs from equipped vehicles.	Highly reliable and accurate.	Significant cost due to advanced technology involved.

In summary, State DOTs can explore a variety of potential emerging data sources and analytics products for traffic simulation calibration. By leveraging the data offerings from data vendors, State DOTs can fine-tune their models, calibrate simulations to real-world conditions, and make informed decisions regarding transportation infrastructure planning. Evaluating the suitability of these data sources based on specific needs and requirements can help agencies unlock the full potential of their traffic analysis and simulation capabilities.

CHARACTERIZING EMERGING DATA FOR TRAFFIC SIMULATION MODEL CALIBRATION

Key Aspects in Understanding Emerging Data Sources

Understanding emerging data sources is a key aspect of contemporary traffic analysis and simulation. A few references suggested by State DOTs provide additional insights into recognizing the key aspects of understanding emerging data sources. The study conducted by Hale et al. (2022) on data fusion and analysis explores aspects such as data trustworthiness, latency, trip tracking, different vehicle classes, and data types. These considerations are key for effectively using emerging traffic analysis and simulation data sources. Additionally, the analysis by Desai et al. (2022) provides insights into using connected vehicle data for conducting mobility analysis, work zone analytics, and assessing the presence of electric and hybrid vehicles on interstates.

Based on Hale et al (2022) and Desai et al (2022), the following information offers key insights and questions to enhance the understanding of these components of emerging data sources:

- Data aspects:
    - Product and vendor differentiation: Understanding the differences between products and vendors is key for having a comprehensive overview of available options.
    - Data trustworthiness: Evaluating the reliability of data to help guarantee valid analysis and informed decisionmaking.
    - Latency: Considering the time delay in data availability ensures timely extraction of insights.
    - Emerging trends: Staying updated on current advancements and trends in data sources for informed decisions.
- Trip aspects:
    - OD patterns: Recognizing OD patterns and routes enhances the accuracy of traffic analysis.
    - Trip tracking: Depending on the objectives of the analysis, determines whether long-term or short-term tracking of trips is required.
    - Vehicle classes: Considering the unique characteristics of different vehicle classes, such as electric or hybrid vehicles.
    - Data type decision: Aligning the decision to work with raw data or aggregated data with the specific needs and objectives of the analysis.
    - Sampling strategy: Prioritizing obtaining sufficient and representative samples rather than striving for a specific percentage of total traffic.

- Other considerations:
    - Vendor selection: Being mindful of potential vendor lock-on and associated risks when selecting data providers.
    - Provider differentiation: Understanding there is a difference between data providers and service providers. Some primarily offer raw data, while others specialize in providing summary statistics and analytics.
    - Data ownership: Enabling secure access to the lowest level of data as needed. Although visuals and analytics tools can be useful, owning and being able to access the actual data are key.
    - Data types: Distinguishing between actual, factored-up, and synthetic data for a better understanding of their reliability and trustworthiness. Actual data provide a direct representation of observed trips, while factored-up data attempt to portray the entire population's characteristics. Synthetic data are generated through modeling techniques to resemble actual data, and they are intended to provide a balance between data utility and privacy concerns.
    - Vendor longevity and adaptability: Ensuring the chosen vendor's capacity for long-term service and adaptability to changes rather than focusing on vendor lock-in.

Key Traffic State and Model Variables in Emerging Source-Based Data Fusion for Traffic Simulation Applications

Understanding how to effectively use emerging data sources is key to calibrating and improving traffic simulation applications. These data sources hold the potential to help analysts enhance ITS strategies, work zone applications, and identifying traffic system states. This enhancement provided by source-based fusion could lead to more precise models, improving traffic flow, reducing congestion, and advancing safety. It could systematically address key areas to meet the needs of pooled fund study members:

- ITS strategies: Emerging data sources can improve the calibration of traffic simulations for ITS strategies.
- Work zone applications: Work zone projects present challenges for traffic analysis and simulation due to their dynamic nature. Traffic data and other emerging data sources offer insights for calibrating simulations in work zone settings. These insights enable the development of a better work zone management plan.

Accurately identifying traffic system states creates the foundation to effectively design and execute traffic control strategies. Ubiquitous sensing techniques, which enable different types of emerging mobile sensors, LBS, and participatory sensing, can provide more reliable and richer traffic observations. Consequently, there is a need to design a system state identification framework to improve the observability of traffic systems.

The establishment of traffic system state identification framework presents a series of theoretically challenging and important modeling issues when using heterogeneous sensor data with different degrees of uncertainty sources. Specifically, the TSE problem for traffic simulation model calibration should simultaneously estimate three sets of system state variables:

- Traffic stream states such as flow rate, density, and speed on road segments of interest.
- Fundamental diagram (FD) parameters such as free-flow speed and jam density of road links.
- Congestion states represented by the queue profile and delays at traffic bottlenecks.
- Microsimulation model parameters, including route choice, headway, standstill distance, volume, speed, travel time, and car following among others.

The literature tends to categorize the traffic system states into three main ideas:

- The TSE problem is devoted to inferring time-varying traffic state variables.
- The model parameter estimation (MPE) problem is dedicated to calibrating or adjusting system parameters in traffic flow models.
- Queue profile estimation (QPE) or congestion bottleneck identification (CBI) is performed to identify congestion duration and the resulting queue profile at signalized intersections or freeway bottlenecks (FHWA 2019).

Key Driver Characteristics in Emerging Source-Based Data Fusion for Traffic Simulation Applications

Understanding driver characteristics and behavior parameters plays a key role in traffic simulation and modeling. These traits include driver reaction time, desired speeds, and acceptable critical gaps for lane changing, merging, and crossing. The recent emergence of new data sources has introduced the capability to calibrate these parameters and even to specify additional driver attributes such as cooperation, awareness, and compliance with speed limits and traffic signs.

Diverse Aspects of Driver Characteristics

Key aspects of driver characteristics include aggression, cooperation, awareness, and compliance. Aggressiveness signifies how drivers respond to traffic-flow conditions, whereas cooperation indicates the extent to which drivers prioritize collective benefit and adjust their driving behavior. Driver awareness reflects the level of traffic condition information that drivers possess, such as queues, congestion, incidents, and available alternatives. Compliance is a measure of how often drivers adhere to traffic control signs, messages on variable message signs, and other regulatory instructions.

Analysts can specify certain behavioral data to get a more detailed understanding of drivers. Observable data examples include minimum headway in car following, gap acceptance for lane

changing, response to yellow change interval, availability of real-time information, and driver's response to this information.

Observable Versus Difficult-to-Observe Data

Emerging data sources provide ample observable driver behavior data such as queue discharge and car-following headways, gap acceptance, and startup lost time. Instead of solely relying on default values, analysts can benefit from collecting observable data whenever possible since a significant portion of driver behavioral data is challenging to observe directly. Examples of difficult-to-observe data include free-flow speed and acceleration or deceleration rates, lane-change courtesy factors, and the distribution of driver types. This latter category affects aggressiveness and becomes a key data point in driver behavior analysis.

Refining Driver Behavior Parameters

If valid observed data are available, these can be used to override default values for driver behavior parameters, including free-flow speed, discharge headway, and startup lost time at intersections. While deviations from defaults are permissible, they should be documented.

Emerging data sources offer an opportunity to refine and calibrate driver characteristics. These data sources can provide insights about the percentage of habitual or commuter drivers compared to tourists, the level of driver awareness of traffic flow conditions and available alternatives, and the overall familiarity of drivers with the transportation network.

Harnessing these data sources can help analysts obtain a more indepth understanding of driver behavior. A more comprehensive understanding of driver behavior helps analysts create more accurate traffic simulation and modeling.

Studies by Shahrbabaki et al. (2018) and Bachmann et al. (2013) outline emerging data cleaning and fusion techniques. These techniques include a range of algorithms designed for noise reduction, outlier detection, and imputation of missing data. A suite of advanced data fusion methods, such as Ambühl and Menendez (2016) and Wu et al. (2018), can also be referred to, encompassing statistical matching, machine learning-based fusion, and multimodal data integration.

Probe vehicles, GPS traces, traffic cameras, and social media data are discussed to unravel the relationship among emerging data cleaning, data fusion, and various transportation data sources. Insights are gathered and summarized from consultation with model users, vendors, researchers, and industry experts to highlight the application and challenges of these data cleaning and fusion technologies. These consultations provide an understanding of experiences and challenges encountered during different stages of model applications.

Subsequent chapters in this report share real-world experiences, lessons learned, and effective practices drawn from implementing data cleaning and fusion approaches. The goal is to improve the quality of transportation data and enhance the effectiveness of simulation-based analysis.

CHAPTER 3. EVALUATING IDENTIFIED EMERGING DATA SOURCES

Chapter 3 examines the strengths and weaknesses of emerging data sources, as well as suggests techniques for preparing data for calibration. The first section highlights the importance of legacy data sources, particularly when synthesized with emerging data sources. The second section presents an indepth account of various datasets derived from data vendors. Each dataset exhibits strengths and specific applications, as shown in the following descriptions:

- Data source A: Serves as a representative example of trip and volume data vendors.
- Data source B: Encompasses probe vehicle data collected from a diverse array of sources.
- Data source C: Provides publicly available, high-resolution vehicle trajectory data as highlighted in the third section.
- Data source D: Focuses on loop detector data and GPS data derived from mobile phones.

The usage of these data sources for systemwide macroscopic observation and TSE on a freeway segment is explained in detail, including an examination of the relative value of information under varying GPS and automatic vehicle identification (AVI) market penetration rates and a demonstration of an integrated framework for simultaneous TSE, MPE, and QPE.

POTENTIAL LEGACY DATA SOURCES FOR TRAFFIC ANALYSIS WITH IDENTIFIED EMERGING DATA

Review of Traditional Data Sources and Traffic Analysis Process

This section examines the value of traditional or legacy data sources especially when integrated with newer, emerging data sources. Table 2 provides a list of these traditional data sources that are typically accessible and widely used.

Table 2. Category of legacy sensor data collection technologies.

Sensor	Description	Variables of System Performance	Use of Data Sources in Traffic Analysis and Simulation Model Calibration
Inductive loop detector	Detects vehicle movement, presence, count, and occupancy; reliable under various weather conditions.	Vehicle count, vehicle presence	FD and capacity parameters
Magnetic sensor	Detects vehicle presence; identifies stopped and moving vehicles.	Vehicle presence	Time-dependent queue profiles and congestion
Camera	Detects vehicles across several lanes, vehicle classification, flow rate, occupancy, and speed. Cameras are linked to a computer with an intelligent algorithm to retrieve traffic parameters. It is low cost and easy to install and maintain.	Flow rate, occupancy, speed, density, queue length	Car-following model and subareas OD information
Radar	Uses radio waves to detect vehicles, measure speed, and detect movement direction. It is high cost, difficult to install, and difficult to maintain.	Vehicle count, speed, direction	FD and capacity parameters
Infrared	Detects infrared radiation through sender and receiver parts. It can measure speed, vehicle volume, and lane occupancy. It is low cost but difficult to maintain.	Speed, vehicle count, occupancy	FD and capacity parameters
Ultrasonic	Uses ultrasonic waves to detect vehicle presence and occupancy. It is low cost but difficult to maintain.	Vehicle count, vehicle presence, occupancy	FD, and capacity parameters
Remote traffic microwave sensor	Uses radar technology for vehicle detection.	Average vehicle length, speed	FD, congestion bottleneck estimation
GPS	Uses satellite-based sensing to provide information on vehicle location. It is relatively expensive and difficult to install and maintain.	Coordinate, count, speed, direction	OD, route choice

Multiresolution Traffic Simulation Tools for Modeling Driver Behavior and Traffic Flow Dynamics

This research project employs a multitiered approach in assessing traffic simulation modeling tools and using them for realistic and accurate representations of traffic demand and supply scenarios. The selected tools fall under three distinct classifications, each aimed at specific resolutions of traffic flow modeling, from vehicle to aggregated flow level. These tools strive to recreate and simulate real-world traffic conditions and assist in devising efficient traffic management and control strategies.

Open-source tools used in different studies include traffic simulation packages and DTA packages (Behrisch et al. 2011; Horni et al. 2016; Auld et al. 2016; Zhou and Taylor 2014). A typical traffic simulation package, such as package I, is a space-continuous, microsimulation tool offering multimodal simulation, including road vehicles, public transport, and pedestrians (Behrisch et al. 2011). It supports various application programming interfaces to enable customization of modeling. MATSim is a multiagent simulation framework, based on activity, that is also extendable. A typical DTA package, such as package II (Zhou and Taylor 2014), is a compact dynamic network loading simulator incorporating Newell's simplified kinematic wave model (Newell, 1993).

A main objective of the model calibration is to encapsulate naturalistic driver behavior under diverse freeway and arterial street driving conditions. These behaviors include desired speed selection, acceleration changes, lane changing, car-following distances, response times, and emergency stopping. The focus of this study is not to update existing models or develop new ones, but to devise techniques for creating traffic simulation models that are more accurate, sensitive, and compatible with future simulation models.

The simulation modeling task workflow adheres to several principles to meet the objectives of model calibration. First, it seeks to statistically differentiate and characterize heterogeneous driver behaviors across different populations under different travel conditions. Second, it emphasizes the need for a comprehensive analysis of how existing microsimulation models can be enhanced using available vehicle position, speed, and acceleration data obtained from a wide range of data sources, such as GPS sensors, accelerometers, video images, and measurements collected from a driving simulator.

Data from various past and ongoing studies, as well as databases developed by transportation research institutes, can be used for analysis. These databases provide longitudinal measurements of fundamental driver behavior characteristics and assist in modeling naturalistic driving behavior.

EVALUATING IDENTIFIED EMERGING DATA FROM VENDORS

Data Source A as an Example in Trip and Volume Data Vendor Category

This section aims to examine the different datasets obtained from various data vendors. Each dataset brings strengths and specific applications. For instance, data source A falls under the trip and volume data vendor category. Data source B contains probe vehicle data from several providers. Numerous companies provide a comprehensive range of OD trips and waypoint data.

These data, rich with information such as the OD of trips and travel times, play a key role in calibrating transportation demand models and understanding travel patterns. Harnessing these indepth trip data and performing rigorous analysis can help analysts make informed decisions regarding transportation infrastructure and planning.

Ensuring the accuracy and reliability of the data is paramount, which involves verifying its quality, especially when obtained from external vendors, to prevent any inconsistencies or biased outcomes. Additionally, in instances where certain data, like turning counts or link counts are missing, techniques such as OD matrix analysis or path flow assignment are used to estimate the required information for microsimulation studies. Specifically, by utilizing trip and volume data, the analyst can perform the tasks related to traffic simulation model calibration as follows:

- Traffic volume and flow analysis:
  - Estimate of AADT analysis.
  - Estimate of traffic link count analysis.
  - Estimate of classified turn counts at intersections.

- Travel behavior and demand analysis:
  - Estimate of OD trips analysis.
  - Route choice analysis.
  - Travel mode analysis.
  - Mode choice analysis.
  - Attraction analysis.

- Network performance analysis:
  - Travel time analysis.
  - Travel speed analysis.
  - Network analysis.

Data source A, for instance, uses a wide array of geospatial information from mobile devices to offer estimates of trip OD volume, trip purpose, and travel times using data from smartphone applications and GPS devices in commercial vehicles. These data allow users to create, execute, and visualize custom queries such as OD and link flow analysis. Such queries can be broken down further by time of day and trip purpose, giving transportation planners a detailed understanding of travel patterns.

Travel mode analysis examines the distribution of various transportation modes in use, which aids infrastructure planning and promoting sustainable transit options. In contrast, mode choice analysis focuses on predicting individual preferences and factors influencing transportation choices, requiring detailed personal and trip-specific information. When evaluating vendor-provided data, such as from data source A, analysts must ensure data completeness, assess its quality, and be mindful of potential limitations to derive accurate and meaningful insights applicable to both types of analysis.

Recognizing the potential of OD trips and waypoint data, analysts should approach the data with a mindful perspective that considers its integration into various State DOT activities. The data sources, namely smartphone applications and commercial vehicle GPS devices, may

inadvertently bring biases into the analysis given the variability in accessibility of cellular data plans across different income groups. Additionally, the disparities in market penetration among the trips analyzed need to be accounted for.

A thorough evaluation has been carried out to ensure the effective use of data source A. By addressing key questions and filling potential gaps, this research project enables DOTs to fully use the potential of data source A for informed planning and decisionmaking, as shown in table 3.

Table 3. Summary of data source A applications.

Reference	Application	Location Selection	Data of Year	Data Quality Evaluation
Kothuri et al. (2022)	Exploring data fusion techniques to derive bicycle volumes on a network	Boulder, CO; Charlotte, NC; Dallas, TX; Portland, OR; Bend, OR; Eugene, OR	2017–2019	Six cities selected, corridor-level OD volumes
Claros et al. (2022)	Validating and estimating AADT data	State of Wisconsin	2019 with 785,479 OD trips and 150,000 vehicles collected	100 traffic count stations selected for data validation with R2 of 0.941
Schewel et al. (2021)	Collecting AADT	Stations across 48 States	2011–2019 from FHWA; 2018–2019 from MS2	Using machine learning method to estimate AADT
Turner, Tsapakis, and Koeneman (2020)	Evaluating traffic count estimation	442 permanent count locations in Minnesota	2019	Implementing probe-based counts for approximately 90 percent of the moderate- to high-volume roadways, specifically those with an AADT of 20,000 or more
Yang, Cetin, and Ma (2020)	Information for using probe data for planning tasks	Virginia	2017–2018	Estimating AADT, estimating OD trips, estimating traffic counts, estimating turn counts and truck volumes at intersection

Reference	Application	Location Selection	Data of Year	Data Quality Evaluation
Roll (2019)	Evaluating probe data in AADT data inventory in Oregon	Oregon	2017	Compared short-term-based AADT and automatic traffic recorder between probe data and ODOT data percentage error, absolute percent error
Avner (2018)	Travel demand modeling	Frederick, MD; Fredericksburg, VA; US 322, Centre County, PA	Not applicable	Interval versus external flow, external trip distribution, percentage error
Shay (2017)	Identifying freight patterns and access to major routes via OD trip analysis	Rickenbacker area, Ohio	Not applicable	OD travel analysis, freight pattern analysis, gate entrance/exit analysis

Given the detailed information available in trip OD and link volume data provided by vendors, there is an opportunity for analysis and understanding of transportation demand models, but more critically, for the calibration of microsimulation models. The information encompasses the OD of trips, travel times, and more. Table 4 provides an overview of the general features and specifics of trip and volume data from data source A.

Table 4. General features and specifics of trip and volume data from data source A.

General Features	Specifics
Detailed trip information	OD of trips, travel times
Analysis categories	Traffic volume and flow analysis (estimate of AADT, traffic counts, turn counts, truck volumes at intersections), travel behavior analysis (estimate of OD trips, route choice, travel mode, mode choice), network performance analysis (travel time, road speed, and network analysis), demand analysis (attraction analysis)
Sample meta information	Analysis ID 884450 focusing on north segment analysis, LBS trip data, "all vehicles volume" output type, specific date range (June 27, 2021; July 11, 2021; and July 25, 2021), and metrics version R115-M116
Key performance measures and considerations	Reliability, accuracy, granularity, timeline, cost
Example data	See table 5 for detailed data and table 17 for data dictionaries of data source A
Sample location and visualization	Wisconsin I–90/I–94 OD zone selection, network topology and geometry (OpenStreetMap, osm2gmns, QGIS (a free, open-source cross-platform desktop geographic information system))

As shown in the general features in table 4 and specific samples from table 5, the provided dataset, with meta information for Analysis ID 884450, focuses on the north segment analysis, leveraging LBS trip data for OD analysis. This rich data source, coupled with the pass-through movement detection method, allows for extraction of key insights into travel patterns.
This data compiling and analysis yield an "all vehicles volume" output type, estimating the volume of all vehicles involved in the analyzed trips during three specific periods: June 27, 2021; July 11, 2021; and July 25, 2021.

The provided data demonstrates several key features:

- Reliability: The use of LBS trip data with pass-through ensures extraction of valuable insights, enhancing reliability of the information.

- Accuracy: The analysis focuses on the north segment, providing a specific geographic area for examination. Given that the provided data represent the overall volume of OD traffic across different trip purposes and vehicle types, the accuracy may not be exceptionally high.

- Timeline: The data include a specific date range (June 27, 2021; July 11, 2021; and July 25, 2021), enabling focused examination of travel patterns during a specific period. This timeline aspect enhances the understanding of the data.

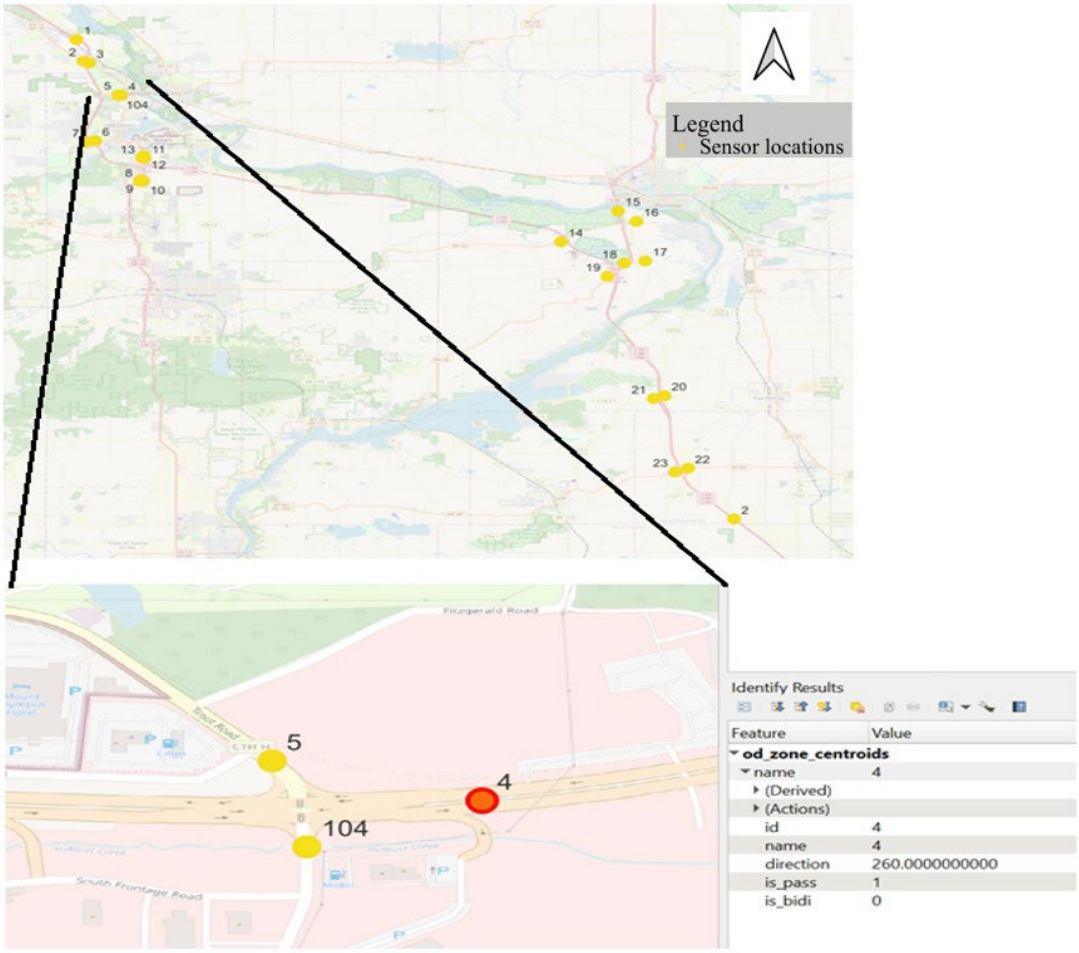
- Cost: The cost is determined based on the size of the data required. Upon a thorough examination of various factors, it appears the costs associated with this data source are reasonable and align with industry standards.

Table 5. Sample detailed data from data source A.

Mode of Travel	Zone Type	Zone ID	Zone Name	Pass-through Zone	Zone Direction (degrees)	Zone Bidirectional	Day Type	Day Part	Average Daily Zone Traffic Volume
All vehicles volume	Origin	1	1	Yes	137	No	0: all days (Monday–Sunday)	0: all day (12 a.m.–12 a.m.)	31,259
All vehicles volume	Origin	1	1	Yes	137	No	0: all days (Monday–Sunday)	1: peak p.m. (12 p.m.–6 p.m.)	16,328
All vehicles volume	Origin	1	1	Yes	137	No	1: Sunday (Sunday–Sunday)	0: all day (12 a.m.–12 a.m.)	31,259
All vehicles volume	Origin	1	1	Yes	137	No	1: Sunday (Sunday–Sunday)	1: peak p.m. (12 p.m.–6 p.m.)	16,328
All vehicles volume	Origin	10	10	Yes	27	No	0: all days (Monday–Sunday)	0: all day (12 a.m.–12 a.m.)	4,228
All vehicles volume	Origin	10	10	Yes	27	No	0: all days (Monday–Sunday)	1: peak p.m. (12 p.m.–6 p.m.)	1,831
All vehicles volume	Origin	10	10	Yes	27	No	1: Sunday (Sunday–Sunday)	0: all day (12 a.m.–12 a.m.)	4,228
All vehicles volume	Origin	10	10	Yes	27	No	1: Sunday (Sunday–Sunday)	1: peak p.m. (12 p.m.–6 p.m.)	1,831
All vehicles volume	Origin	104	104	Yes	358	No	0: all days (Monday–Sunday)	0: all day (12 a.m.–12 a.m.)	3,249

The focus of the following analysis is the Wisconsin I–90/I–94 OD zone selection with 25 sensor locations strategically selected within this area. This selection also incorporates 13 ramps equipped with these sensors. The conversion of OpenStreetMap data to the General Modeling Network Specification (GMNS) (GMNS 2022) is used for this purpose. For a comprehensive understanding of the traffic flows in the selected area, a substantial number of routable nodes and links are identified through the osm2gmns conversion process (Lu and Zhou 2022). Precisely, 3,405 nodes and 2,259 links are identified for a traffic assignment and simulation model network.

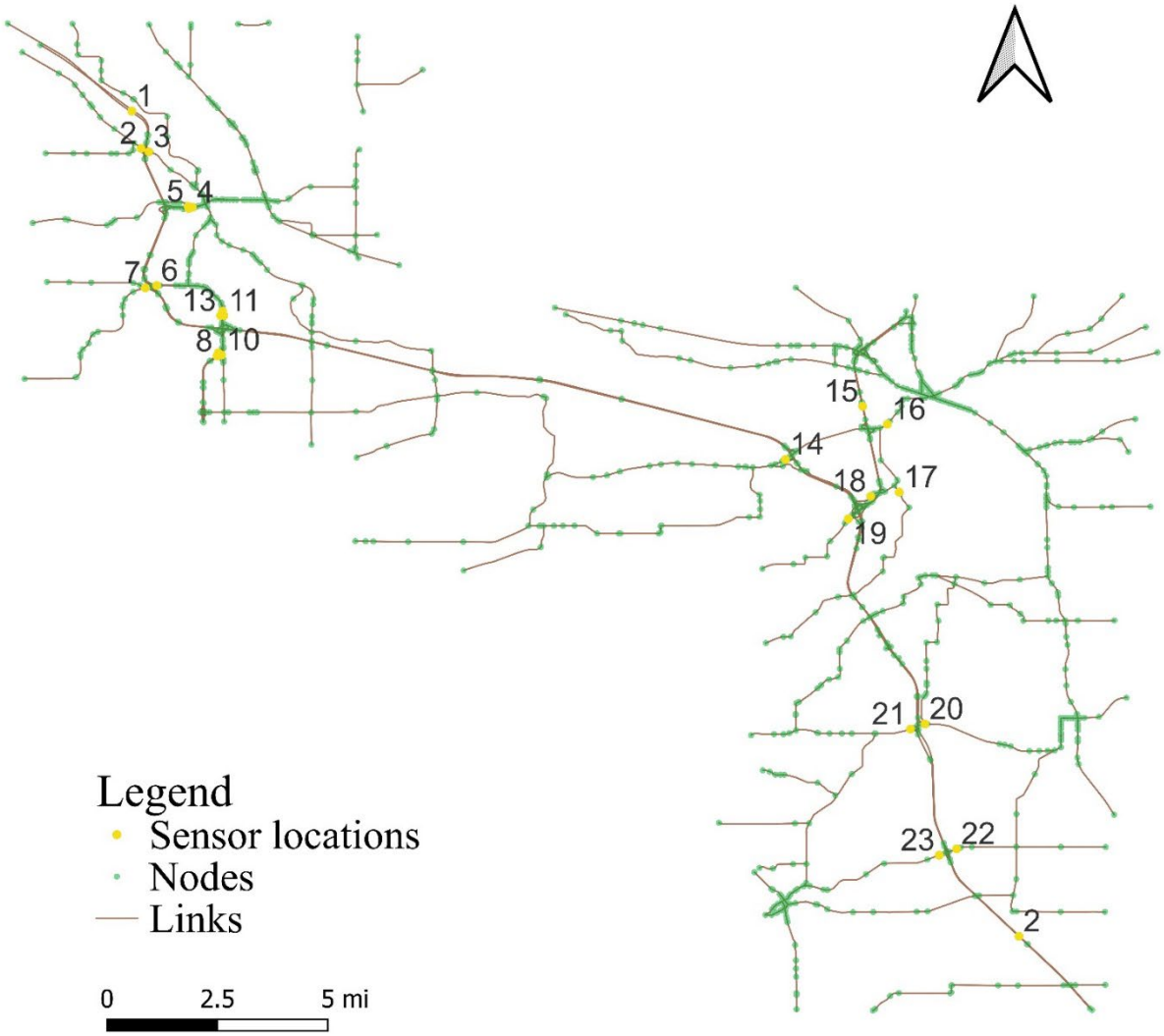
As shown in figure 1, the key elements of the simulation network's topology and geometry revolve around choosing specific road types for inclusion in the analysis and streamlining the node consolidation process. Essentially, the goal is to differentiate and decide upon various link categories, ranging from primary roads to residential streets, determining which are most pertinent for the study network. For further visualization, QGIS is used for manual fixes in the cases of missing nodes, lanes, or turning movement lanes, as well as dealing with duplicates and ensuring consolidation.



Original map: © 2024 OpenStreetMap contributors. Lines and numerical overlays added by FHWA.

Figure 1. Map. Wisconsin I–90/I–94 OD zone with selected sensor locations.

Figure 2 presents the types of roads incorporated into the analysis, with small dots representing nodes within the selected area and large yellow dots representing sensors. The OSM link types included in the analysis comprise motorways, primary, secondary, and tertiary roads, and need to be mapped and converted to standard highway functional classification codes. This approach consolidates each complex intersection into a single node, thereby facilitating traffic signal modeling in later stages of the analysis.



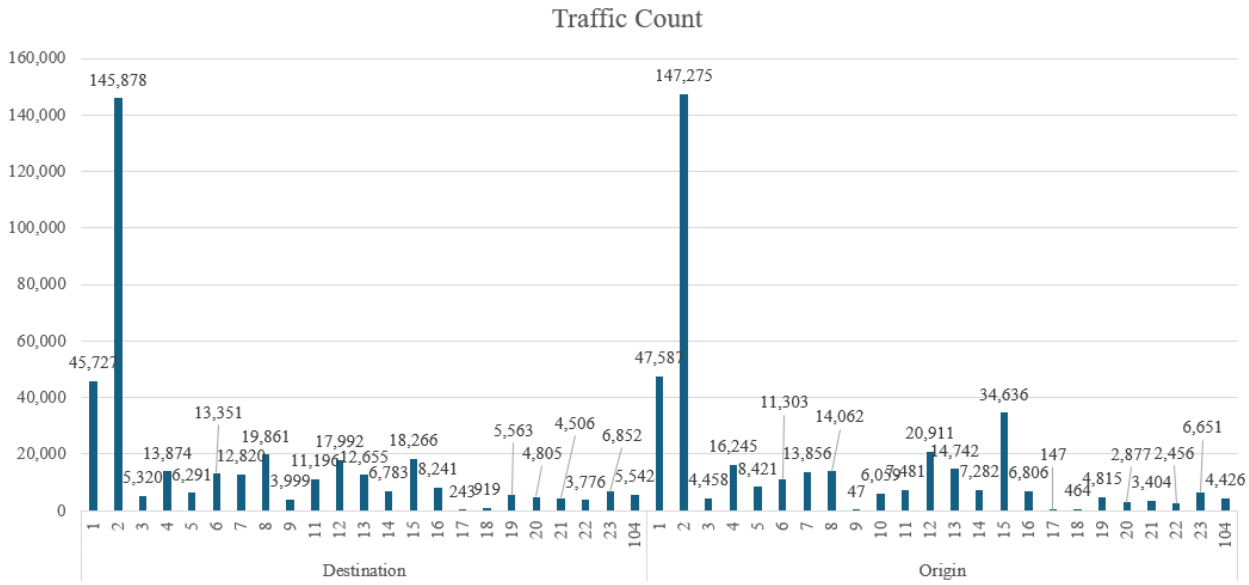
Original map: © 2024 OpenStreetMap contributors. Lines and numerical overlays added by FHWA.

Figure 2. Map. General modeling network of nodes, links, and numbered sensor locations.

As shown in table 6 and figure 3, the transportation analysis dataset is based on the demand estimation models and plays a key role in estimating traffic volume across specific OD zone pairs, as well as for determining the associated flow patterns of vehicular movement along the routes. This sample data input enables transportation planners to analyze and calibrate microsimulation models effectively.

Table 6. Sample Origin-Destination (OD) demand matrix.

O/D	1	2	3	4	5	6	7	8	9	11	12	13	14	15	16	17	18	19	20	21	22	23	104	Grand Total
1		8,366	1,587	1,219	322	1,352	550	3,436	389	664	281	776	334	72	856	8	10	581	357	447	396	406	324	1,352
2	8,467	31,960	778	746	304	569	630	414	163	809	1,460	612	608	8,152	284	39	7	477	378	343	94	1,512	173	3,260
3	1,330	1		1	6	3	50	44	28	5		10			15						19	7		203
4	1,583	1	35		2,518	432	1,094	1,194	222	161	61	211	37	15	75			96	77	23	57	27	2166	583
5	260	195	22	3,455		75	163	147	13	105	12	106			28					3			473	350
6	1,209	923	15	333	90		4,332	172	33	39	8	59	5	37	83	7		23	17	13	27		20	398
7	517	1	33	855	143	5,006		697	134	366	15	254	18	57	320			38	11	23	41	6	58	483
8	2,200	1	33	860	126	116	367		35	703	2,135	1,096		61	79			22	14	17	29	12	94	437
9	7	5									16													9
10	239	1	28	205	9	116	167	263	13	315	872	444	20	143	94			12	30	49	61		70	180
11	783	1	5	126	59	21	190	627	65		1,536	879	11	77	119			47	33	6	13	15	20	271
12	368	1	4	102	15	53	103	2,871	603	2,791		3,969	48	80	369			63	28	68	57	61	21	701
13	1,216	1	3	351	47	167	246	1,417	304	1,457	3,190		22	145	252		4	73	46	45	75	18	57	500
14	41	1		28			15		7		4			729	1,372	8		37	176	54	120	62		326
15	96	7,758	5	43	3	83	167	91	71	32	167	125	913		136	29	427	783	620	681	224	1,140	11	929
16	497	363	4	95	11	90	288	69	68	122	372	261	1,199	108		1	0	248	42	32	8	84	22	190
17														7	53				7					22
18		44												190		7			7					62
19	316	1,621		57		13	5			14	14		23	363	237	8			75	50	14	57		191
20	165	532		23		13	22	6	9	15	39	14	1	198	35	21		28		626	8	7	3	100
21	164	997		4		5	32	6			13	16	466	267	31			31	595		9	38		150
22	306	172	13	50		7	16	24	6	16	10	13	72	92				15	19	21		723	8	88
23	136	2,060	6	19			22				23	44	22	616	85	6		33	31	36	988			275
104	356	131	4	1,688	345	23	34	73	37	43			38		7					11	9			195
Grand Total	1,306	3,875	197	524	287	459	434	662	124	470	584	502	243	599	229	15	90	171	147	145	117	261	231	698



Source: FHWA.

Figure 3. Chart. Production/attraction-based zone average daily demand volume.

Key traffic modeling components of this dataset include the following:

- **OD traffic volumes observation and visualization:** Table 6 illustrates the OD traffic volume and average daily traffic volume over the study period for the different zones, respectively. They allow visualization of travel patterns within the analysis area.

- **Key role of OD matrix and estimation models:** These tools provide information about travel patterns between different traffic analysis zones in a region. These data help transportation planners assess trip patterns based on sensor count and speed data on links.

- **OD travel time analysis:** The travel time analysis is key for traffic simulation calibration. By comparing travel times measured from simulation runs and real-world data, the simulation model can be calibrated to reflect actual traffic conditions more accurately, thereby optimizing transportation system performance.

- **Traffic assignment and estimation process:** This includes allocating the demand on available routes connecting each zone pair (traffic assignment) and adjusting an OD matrix that reproduces the observed traffic counts. Both processes are key for understanding and managing traffic patterns, especially during congestion.

- **Integration with other simulation models:** The model's capabilities are expanded by integrating with mesoscopic and microscopic models, adding depth to the analysis.

- **Calibration for complex behavior responses:** Significant effort is put into calibrating transportation models to produce accurate results. Accuracy is key for complex networks and understanding the effects of demand management strategies on traveler's choices.

- **Overemphasis on supply side:** Many analyses may fail to reflect the true impact of operating conditions due to a disproportionate focus on the supply side. This overemphasis is particularly apparent when assessing multiday performance to estimate reliability.

- **Importance of model verification and calibration:** The comparison of observed and assigned volumes is a key part of model verification and calibration.

Data Source B as Probe Vehicle Data

This section details the use of probe vehicle data. These data are obtained directly from vehicles and connected devices. Updated in real-time, these data can incur significant costs due to data acquisition and processing. Data source B is a provider of data from connected and electric vehicles, offering real-time insights key for transportation monitoring, safety, and planning. Data source B's extensive dataset can be used for various transportation-related applications, including performance measurement, project evaluation, and bottleneck analysis.

The Adventures in Crowdsourcing workshop, hosted by the National Operations Center of Excellence, showcased the practicality of crowdsourced data from Nexar® (Zhang 2021). The focus was on using these data for real-time work zone management and safety in transportation systems. The workshop illustrated the advantages of crowdsourced data for work zone and lane-closure detection, along with its use for monitoring, safety, and planning purposes, such as congestion studies, performance measures, and model validation.

Zhang (2021) offers a practical case study on data source B's application for transportation monitoring and planning. The data source B sample data offers an indepth view of transportation dynamics in the Phoenix metropolitan area managed by the Maricopa Association of Governments (MAG). As illustrated in the sample data in table 7, data source B provides insights into mobility, traffic patterns, and transportation behavior, which assists researchers, urban planners, and policymakers in decisionmaking processes aimed at optimizing transportation systems and improving urban mobility.

Table 7. Sample of probe vehicle data.

Randomized Journey/Trip ID	Timestamp	Latitude	Longitude	Speed (mph)	Heading (degree)
1234554321	10/2/2019 0:59	33.21915	-111.772998	21.88	89
1234554321	10/2/2019 0:59	33.21915	-111.772656	38.01	90
1234554321	10/2/2019 0:59	33.219155	-111.772158	55.29	89
1234554321	10/2/2019 1:00	33.219158	-111.771555	66.81	89
1234554321	10/2/2019 1:00	33.219157	-111.770876	74.88	90
1234554321	10/2/2019 1:00	33.219157	-111.770144	81.79	90
1234554321	10/2/2019 1:00	33.219161	-111.769371	86.39	89
1234554321	10/2/2019 1:00	33.21917	-111.76688	94.46	89
1234554321	10/2/2019 1:00	33.219165	-111.766026	95.61	90

Note: "Heading" represents the direction of travel relative to true north, measured in degrees.

The collected data, composed of detailed attributes of a vehicle's journey in a 3-s trajectory format, includes timestamp, position, speed, heading direction, and vehicle type—limited to passenger vehicles. The vehicle movement data provide 3 s resolution with coverage 24 h per day and 7 d per week, with a penetration rate of 4–6 percent. The vehicle body class distribution data show a dominance of pickup trucks (39.9 percent) and sport utility vehicles (multipurpose vehicles) at 27.6 percent, with sedans constituting 15.1 percent as reported by Zhang (2021).

Several challenges and gaps need to be addressed to maximize the data source B's utility for traffic simulation model calibration or related planning applications. First, data interpretation is a challenge. There is a gap between raw data and information meaningful to stakeholders that requires effort to convert into digestible insights. Given that passenger cars only account for 28.1 percent of the data, careful interpretation of OD patterns is key to avoid representational biases.

Second, data management poses a challenge due to the required resources for data storage, processing, analysis, and visualization, which underscores the requirement for capable staff to manage the workload and a powerful platform to handle big data. With large data files (more than 20 gigabytes per day in the MAG region), a significant amount of processing and querying effort is needed.

The probe vehicle sample data offer an indepth view of transportation dynamics in the Phoenix metropolitan area, as shown in the data source B study. The usage of probe vehicle data provides insights into mobility, traffic patterns, and driving behavior, which assists researchers, urban planners, and policymakers in decisionmaking processes aimed at optimizing transportation systems and improving urban mobility.

The probe vehicle data can provide a rich resource for traffic simulation model calibration. Some potential applications are as follows:

- **Performance measurement:** Hot spots, congested segments, or critical zones can be identified by observing travel time, speed in corridors, subareas, or jurisdictions by time of day and day of week. The data can also facilitate project evaluations providing before-and-after studies for transportation improvement projects.

- **TSE:** Data can be converted into intersection measurements such as the turning movement count, and turning movement ratio, travel delay (divided into control delay and stop delay), level of service (LOS), queue length, and percentage of arrivals on green. The data can also yield intersection congestion profiles by date and time of day.

- **Model validation/calibration:** The probe vehicle data can assist in calibrating travel demand models by providing data on travel time, speed, and free-flow speed. Microsimulation models can benefit from queue position and timing, queue length, delay, and turning movement counts. Cross-referencing can also be applied for data validation.

- **Congestion study:** The data facilitates freeway bottleneck studies, using OD data, travel time, delay, and harsh brake occurrences. For intersection analyses, turning movement ratio, percent arrivals on green, control delay, stop delay, and LOS can be evaluated. Corridor studies can assess travel time reliability and conduct before-and-after studies.

As shown in table 8, probe data offer insights into and applications for model validation, travel studies, continuous travel monitoring, and congestion studies. The data's extensive coverage of travel behavior and key performance indicators allows for comprehensive analysis and decisionmaking in transportation planning and operations.

Table 8. Summary of data source B's traffic management and planning applications.

Reference	Purpose and Applications	Location	Year of Data	Data Quality Evaluation	Data Fusion with Other Data Sources
Islam and Abdel-Aty (2023)	Short-term conflict prediction using previous trajectory data	Orange County, Orlando, FL	2019	Long short-term memory (LSTM)-based conflict prediction framework uses connected vehicle probe data with a market penetration rate of 3 percent and achieved a 72 percent accuracy, 81 percent recall, and 28 percent false alarm rate in predicting conflict cases	None
Khadka et al. (2023)	Estimate the regional link volumes using deep neural network	1,200 locations on freeways in Dallas-Fort Worth, TX	20 workdays in September 2021	Total trajectory data collected with 15 min of time interval	None
Khadka, Li, and Wang (2022)	Queue length and propagation at freeway bottlenecks, traffic delay, time-space visualization and combined with signal performance	Dallas-Fort Worth, TX	June 1–7, 2020; December 29, 2020	A total of 36,345 vehicle trips, with selected 0.5-mi segment length on highway	High-resolution signal data
Saldivar-Carranza et al. (2022)	Performance measurements at continuous flow intersection	West Valley City, UT	August 2021 weekday	4,500 trajectories and 105,000 GPS points from August 2021 weekday data	Not reported

Reference	Purpose and Applications	Location	Year of Data	Data Quality Evaluation	Data Fusion with Other Data Sources
Sakhare et al. (2022)	Truck and passenger trajectory data penetration analysis	State of Indiana	May 9–15, 2022	10.8 million vehicles and more than 13 million trips over a 1-w period from May 9–15, 2022. Average truck penetration is 3.4 percent; overall connected vehicle penetration on interstates is 6.32 percent and 5.3 percent on non-interstate roadways.	Traffic count data, CAV data
Saldivar-Carranza et al. (2021)	Traffic signal performance: split failure, downstream blockage, and quality of progression, as well as traditional *Highway Capacity Manual* (HCM) LOS	South of Indianapolis, IN (Thompson Road, Harding Street, Epler Avenue, Southport Road, Wicker Road, County Line Road, Fairview Road, and Smith Valley Road)	July 2019 weekdays	Signalized intersections with 160,000 trajectories and 1.4 million GPS data of 3 s of time interval. Data features include GPS location, measured-on-vehicle speed, heading, timestamp, and an anonymous trajectory identification number.	None
Li et al. (2020)	Roadway hazards identification	State of Indiana	August 2019 for 1 w	Conflict analysis using 1.5 million hard breaking data. Delay and conflict analysis.	Network, intersection

Gaps Between Probe Data Collection and Demand Pattern Interpretation

The planners and simulation model users need to recognize another significant gap existing in probe data collection and demand pattern interpretation. If only a subset of vehicles equipped with transponder tags are identified by AVI counts or probe vehicles, then the estimation of market penetration rates and identification rates becomes a considerable challenge. This limitation needs explicit consideration when inferring population trip desires from the available data. Several models have been proposed to estimate population demand using probe vehicle counts, but they all acknowledge the difficulty associated with low identification rates often seen with license plate-based AVI data. For instance, one can use a three-stage procedure to estimate population OD demand from probe vehicle data, which includes estimating the tagged OD demand matrix from probe vehicle data, calculating market penetration rates using probe vehicle data and link counts, and then scaling the estimated probe vehicle demand to the total population demand using estimated market penetration rates.

To account for possible identification and representative errors, Zhou and Mahmassani (2006) further developed a joint estimation formulation and a one-sided linear penalty formulation, thereby resulting in complex optimization problems. In summary, these models need to estimate either market penetration rates or identification rates to connect probe vehicle samples to population demand using a multiplicative function structure. Estimating these rates is problematic because they are essentially time-dependent and location-dependent random variables. Their inclusion in the demand estimation problem could complicate matters.

DATA SOURCES C AND D FROM VEHICLE TRAJECTORY DATA COLLECTED AND MOBILE CENTURY EXPERIMENT

This section introduces data source C and data source D. Data source C comprises publicly accessible, high-resolution vehicle trajectory data—a high-quality dataset that records naturalistic vehicle trajectories on German highways using drone technology. Data source C has been used for studies such as driver behavior analysis, traffic flow patterns, autonomous vehicle algorithm development, and traffic simulation model validation. This dataset exhibits a wide variety of traffic scenarios, high recording frequency, comprehensive metadata, and high-definition quality. Some key characteristics of the data source C include:

- **Variety of traffic scenarios**: Data source C includes a wide range of traffic scenarios, such as lane changes, free-flowing traffic, and traffic jams. The usage of data source C allows for more robust and comprehensive modeling and analyses.

- **Vehicle trajectories**: The dataset contains trajectories for all vehicles visible in the drone's field of view. Each vehicle's position is recorded at 25 Hz (25 times per second), providing detailed insights into vehicle behavior.

- **Rich metadata**: Data source C alongside the trajectories provides metadata about each recorded vehicle, including static information, such as the vehicle class and length and dynamic information, such as the velocity, acceleration, and lane positioning.

- **High-quality recording**: The dataset was recorded in high definition with a 1920 x 1080-pixel resolution, providing clear footage for analysis.

- **Extensive data**: The dataset consisted of more than 50 recordings, capturing more than 110,000 vehicles, and 45 h of traffic.

Data source D focuses on loop detector data from the California DOT (Caltrans) Performance Measurement System (PeMS) and mobile phone-based GPS data from the Herrera et al. (2010) experiment. The data from these sources are instrumental in the systemwide macroscopic observation and TSE on a freeway segment.

This section focuses on data fusion and calibration, which delves into how the developed framework harnesses different data sources for TSE and model calibration. The integration and efficient formulation of heterogeneous sensor data sources play a key role. The main problems considered in this section are integrating diverse data sources through data fusion.

Specifically, one key analytical aspect in the data fusion process is the TSE model, which aims at estimating time-varying traffic stream states, such as flow rate, density, and speed on road segments. These states provide information for identifying traffic incidents in unobservable areas. Another analytical aspect is queue profile estimation using different data sources, which focuses on estimating the time-dependent queue length on freeway corridors or at signalized intersections. Queue profile estimation offers an intuitive representation of queue evolutions at oversaturated traffic bottlenecks, thereby supporting effective traffic simulation-based decision support and management.

The potential of combining analytical data fusion models for a comprehensive understanding of traffic conditions still needs to be explored. This approach aims to leverage the high-level queue profiles for stabilizing local estimations using local estimations, contributing to aggregated traffic modeling and hierarchical control. A comparison of related studies is also provided, examining their different modeling approaches, solution methods, benefits, and challenges. The comparison helps identify the most promising strategies. This section considers four major sensors that provide observations for the joint estimation problem: loop detectors, GPS sensors, Bluetooth sensors, and video detectors. Each sensor type contributes unique data that enrich the observations, enabling more robust and accurate traffic state and queue profile estimation.

Data Source C as Test Benchmark for Data Fusing with Emerging Multisource Heterogeneous Data

Leveraging data source C, this report aims to establish comprehensive test datasets tailored for enhanced data fusion and traffic simulation model calibration. Data source C provides key insights into traffic flow analysis, driver behavior study, autonomous vehicle algorithm development, and traffic simulation model validation, thereby enhancing the understanding of highway driving dynamics. This initiative focuses on formulating a TSE and data fusion problem within a unified modeling framework, targeting freeway segments within a defined timeframe.

The goal is to systematically estimate the traffic states and queue profiles using observations from various traffic detectors. At the same time, addressing potential inconsistencies between different components and adhering to established modeling principles for accurate system dynamics representation is key.

As represented in figure 4, the methodology involves four main types of emerging sensors that generate the necessary observations:

- Loop detectors: These detectors provide aggregated vehicle volumes at predetermined locations at specified time intervals (e.g., every 5 or 15 min).
- GPS sensors: These sensors deliver semicontinuous trajectory data of probe vehicles, including timestamps.
- Bluetooth sensors: These sensors determine the travel time of vehicles equipped with Bluetooth devices from one sensor to the next.
- Video detectors: These sensors offer high-accuracy vehicle trajectories within their monitoring range.

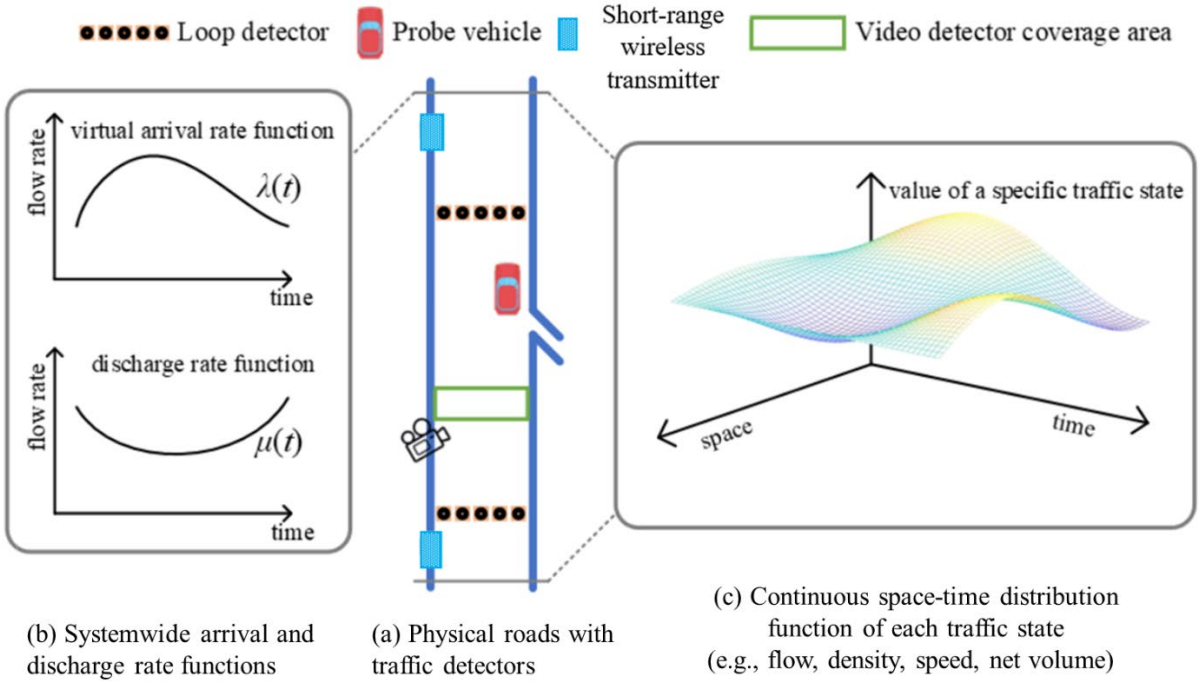


Figure 4. Data fusion modeling on freeway segments with different types of traffic detectors (Lu 2022).

To comprehensively evaluate the model's performance under complex traffic conditions, six segments are selected from data source C. These segments encompass a range of traffic conditions, including both light and heavy traffic scenarios and transitions between these states. Detailed specifications regarding the dataset employed in this study can be found in table 9.

Table 9. Summary of data source C used in this research (Lu 2022).

Dataset	ID	Direction	Month	Weekday	Start Time	End Time
1	12	1	201709	Thursday	17:21	17:36
2	25	1	201710	Monday	8:55	9:14
3	26	1	201710	Monday	9:20	9:38
4	25	2	201710	Monday	8:55	9:14
5	26	2	201710	Monday	9:20	9:38
6	46	2	201711	Wednesday	8:47	9:06

Note: The publicly accessible source code and dataset of Lu's research are available at https://github.com/jiawlu/Traffic_State_Estimation-Computational_Graph.

To reduce the potential impact of data inaccuracies, a preprocessing stage is conducted to discard any observations that do not comply with quality control standards. Considerations at this stage may include GPS drift, Bluetooth mismatches, or other sensor-specific issues. The surviving observations indicate accurate data distribution and are suitable for the proposed model.

The computational graphs used in this study are developed using an open-source machine learning framework. The freeway segments selected for Lu's study (2022) are approximately 420 m with no ramps. To mitigate the impact of vehicle identification errors on segment boundaries, the range of data processing and subsequent estimates is limited to between 30 and 410 m for all datasets. The duration of each dataset fluctuates around 1,000 s, mainly determined by drone battery consumption. Table 10 presents the configurations of virtual traffic detectors used in the study. The loop detector is situated at locations 120 and 320 m from the segment upstream, with an aggregation time interval of 1 min. GPS data are collected at a sampling rate of 10 percent and reported every 5 s. The Bluetooth detector is positioned at locations 40 m and 400 m from the segment upstream, with a sampling rate of 5 percent. Finally, the video detector is located between 220 and 230 m from the segment upstream. These configurations provide information for understanding data collection setup and methodology.

Table 10. Configurations of virtual traffic detectors (Lu 2022).

Detector Name	Configurations
Loop detector	Location: 120 m and 320 m from the segment upstream Aggregation time interval: 1 min
GPS	Sampling rate: 10 percent Reporting frequency: 5 s
Bluetooth detector	Location: 40 m and 400 m from the segment upstream Sampling rate: 5 percent
Video detector	Location: 220–230 m from the segment upstream

Fixing Traffic Flow Model Parameters to Show the Value of Data Fusion with Different Data Sources

Lu's study (2022) evaluates the benefits of a joint estimation framework over precalibrated traffic flow models to enhance the accuracy of state estimations. First, data source C is used to calibrate the parameters of FDs offline and treat these parameters as constants in the state estimation model. Using the virtual detector setting in table 10, table 11 compares speed estimations with fixed traffic flow model parameters. The results show an increased estimation error across all six datasets, indicating that employing precalibrated traffic flow models can significantly compromise state estimation accuracy.

Further investigation was performed for how much a joint estimation framework can improve the accuracy of state estimations compared to using precalibrated traffic flow models in TSE. As illustrated in table 11, precalibrated traffic flow models within the data fusion framework for queue profile estimation significantly affect state estimation accuracy, particularly in datasets with heavy congestion. This discrepancy is primarily due to potential differences between traffic environments (e.g., weather, traffic incidents) where traffic flow models are calibrated, and the traffic states are estimated. Significantly, the joint calibration approach results in an approximate 40-percent improvement in both mean absolute error and mean absolute percentage error. The result in table 11 highlights the substantial value of the joint estimation framework in enhancing the accuracy of state estimations.

Table 11. Comparison between speed estimations with calibratable and fixed traffic flow models (Lu 2022).

Dataset	Fixed Traffic Flow Model Parameters (mph)	Joint Calibration of TSE/QPE and Traffic Flow Models (mph)	Improvement Due to Joint Calibration (percent)
1	1.73/9.00	1.20/5.88	30.6/34.7
2	2.49/25.18	1.26/14.25	49.4/43.4
3	3.84/29.23	2.31/18.65	39.8/36.2
4	1.82/7.85	1.22/5.03	33.0/35.9
5	2.11/10.18	0.98/4.60	53.6/54.8
6	1.97/9.88	1.24/5.90	37.1/40.3

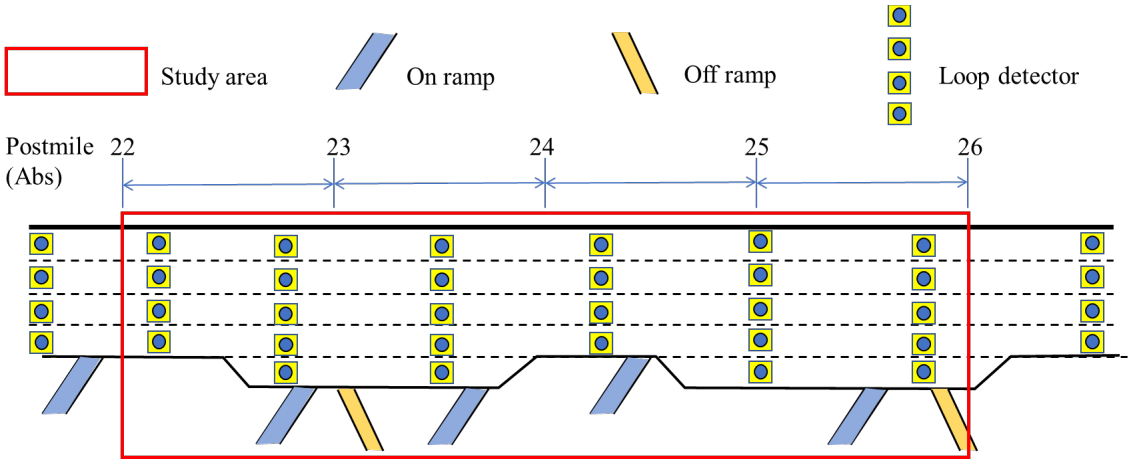
Note: Values in each cell denote mean absolute error and mean absolute percentage error.

Data Source D with Legacy and Emerging Data Sources on a Real-World Freeway Corridor with a Downstream Bottleneck

This section emphasizes the application of the developed framework to estimate traffic states by leveraging diverse data sources. This approach contributes to efficient integration of heterogeneous sensor data, which underpins the calibration of traffic dynamics. This section investigates the cross-resolution TSE framework applied to a 3-mi freeway corridor characterized by a downstream bottleneck. As depicted in figure 5, the focus is the absolute post mi 22–25 on I–880 N in Alameda County, CA. This section was analyzed between 10 a.m.–12 p.m. on February 8, 2008. The data for this analysis, presented in table 12, stem from two types collected during the Herrera et al. (2010) experiment. The travel time of the probe vehicles along

the entire corridor was extracted from the GPS data, providing macroscopic observation on a systemwide level.

The estimation model settings in this section parallel those used in the preceding section, except for the integration of macroscopic modeling. This adjustment aligns with the central theme of this report.



Source: FHWA.

Figure 5. Diagram. Freeway corridor on I–880 N (post mi 22–25).

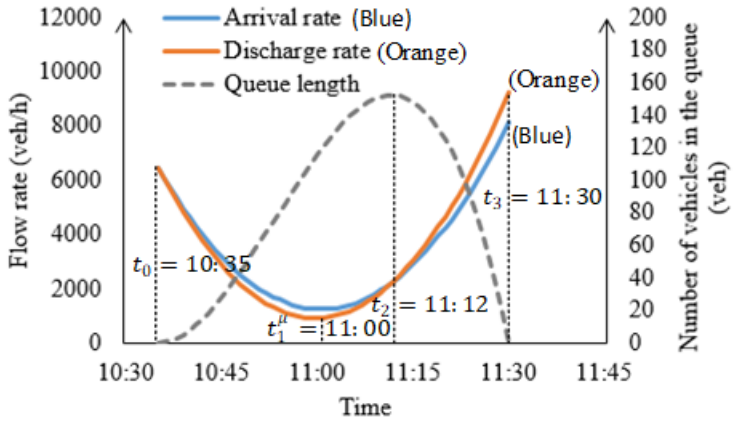
Table 12. Configurations of traffic detectors.

Detector Name	Configurations
Loop detector	Location (post mile): 22.23, 22.53, 22.78, 23.37, 24.01, 24.48, 24.92 Aggregation time interval: 5 min
GPS	Sampling rate: 1.74 percent (192 probe vehicles) Reporting frequency: 3.5 s on average

Figure 6-A presents the estimation results at a broad scale, which highlights the calibrated arrival rate, discharge rate, and queue length curves, focusing on key time points that show changes in traffic conditions. At the first critical point, the traffic-flow arrival rate matches the discharge rate, and queueing begins. Then, the discharge rate reaches its lowest point. After that, the arrival rate matches the discharge rate again, corresponding with the maximum queue length and the start of queue dissipation. Finally, the queue clears, marking the end of the congestion period.

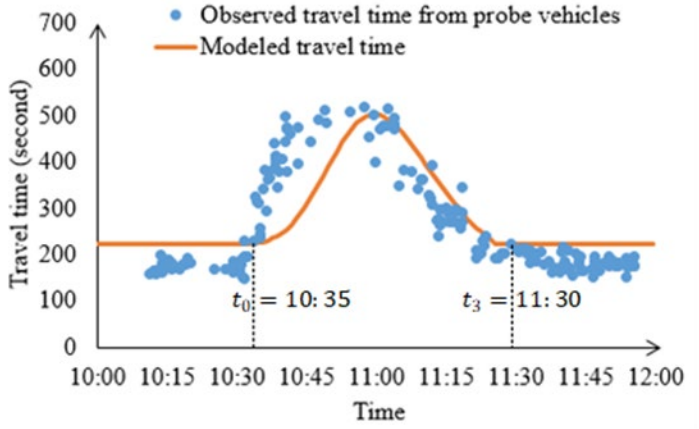
Figure 6-A shows the modeled time-dependent travel time curve alongside observed travel time data from probe vehicles. While the project team's method has generally resulted in a good alignment between the modeled and observed travel times, indicating its effectiveness in capturing macroscopic traffic dynamics, addressing the observed inconsistencies is important.

Specifically, as highlighted in figure 6-B, there is a consistent overestimation of travel time outside of congested periods, which warrants further investigation and potential adjustment of the data fusion approach to improve the model's accuracy across different traffic conditions.



veh/h = vehicle per hour.

A. Estimated arrival rate, discharge rate, and queue length curves.



B. Estimated travel time curve and observed travel time from probe vehicles.

Figure 6. Charts. Estimation results of systemwide measures at the macroscopic level (Lu 2022).

Performance Evaluation for Using Emerging Data Sources in Traffic Simulation Model Calibration

This section assesses the efficiency of typical simulation model calibration using a broad range of metrics. Table 13 provides a comparative overview of related papers and reports on traffic analysis using emerging data sources.

Table 13. Comparative overview of related papers and reports for traffic analysis using emerging data sources.

Reference	Applications	Data Source/ Category	Data Features	Reliability	Accuracy	Granularity	Strengths	Weaknesses
Wejo Group (2023)	Traffic alerts, drive time, parking service, location analysis, roadway analysis, trip analysis, trip trend analysis, speed analysis, traffic volume count	LBS/probe vehicle data	Timestamp, position, speed (kilometer per hour), heading direction, parcels, infrastructure	15 min, (refers to the specific time window used for data aggregation. Within this interval, all collected data points are pooled together to provide a cohesive snapshot of traffic conditions.)	N/A	N/A	Offer a broader utility of data for diverse transportation analysis applications	N/A
Sharma, Ahsani, and Rawat (2017)	Speed bias, incident detection, congestion detection, latency, volume count, buffer time index	LBS/probe vehicle data	Traffic Message Channel (TMC) ID, direction, country, road, time segment, speed, referenced speed, average speed, travel time, confidence	85–100 percent	Realtime with more than 78.8-percent accuracy	1 s	Reliable data with accuracy, incident, and congestion detection capabilities	N/A
Adu-Gyamfi et al. (2017)	Speed, latency, similarity index	LBS/probe vehicle data	TMC ID, direction, country, road, time segment, speed, referenced speed, average speed, travel time, confidence	85–100 percent	N/A	1 s	Reliable for monitoring transportation infrastructure performance over time	Various levels of amplitude bias between LBS and benchmarked data

Reference	Applications	Data Source/ Category	Data Features	Reliability	Accuracy	Granularity	Strengths	Weaknesses
Gong and Fan (2017)	Travel time, reliability	LBS/probe vehicle data	TMC ID, direction, country, road, time segment, speed, referenced speed, average speed, travel time, confidence	85–100 percent	N/A	N/A	Capable of identifying and ranking recurrent freeway bottleneck	Frequency of congestion or planning time index may not identify traffic congestion
Saldivar-Carranza et al. (2021)	Video analysis for UAV/CAV	LiDAR/high-resolution vehicle trajectory data	Automated traffic signal performance measures, signal operations, signal retiming, volume counts, turning movement analysis, asset management, precision navigation, identification of maintenance issues	N/A	N/A	N/A	Valuable for traffic signal analysis, asset management, and precise navigation	N/A

N/A = not applicable.

Based on the analysis in Chapter 3, the following characteristics of each data source are summarized:

- Probe vehicle data:
  - Derived directly from vehicles and connected devices, offering high reliability, accuracy, and granularity.
  - Provides real-time updates, making monitoring and analyzing traffic conditions across a variety of operational planning applications.
  - Includes cost considerations associated with data acquisition and processing.
- High-resolution vehicle trajectory data:
  - Exhibits high reliability and accuracy.
  - Allows for granular analysis with per-vehicle trajectory data, providing detailed insights into individual vehicle behavior.
  - Includes cost considerations of data acquisition, processing, and storage.
- CAV data:
  - Provides highly reliable and accurate data, offering insights into the behavior of advanced vehicles.
  - Provides high granularity with real-time updates, enabling a detailed understanding of CAV operations.
  - Includes cost considerations associated with advanced technology.
- Crowdsourced platforms data:
  - Provides data with which reliability and accuracy may vary.
  - Offers moderate granularity, with detailed incident reports a key strength.
  - Ensures high timeliness with real-time updates, making monitoring dynamic traffic conditions useful.
  - Includes cost considerations that are low to moderate, mainly involving data processing.
- Emerging sensor technologies, traffic management systems, and data platforms provide data with which reliability, accuracy, granularity, timeliness, and costs vary based on the specific technology or platform used.

CHAPTER 4. DATA FUSION FRAMWORK FOR CALIBRATION OF MICROSIMULATION MODELS

This chapter customizes the data fusion framework developed under a previous FHWA Office of Operations project for use in microsimulation model calibration (Hale et al. 2022).

OBJECTIVES OF CALIBRATION AND POTENTIAL CALIBRATION MEASURES

This section describes the objective of calibration, which involves determining the optimal set of parameter values for the model to accurately replicate observed system performance measures. This section also identifies appropriate calibration measures for datasets fused from emerging and legacy data sources.

The most recent FHWA *Traffic Analysis Toolbox Volume 3: Guidelines for Applying Traffic Microsimulation Modeling Software (2019 Update)* established standardized practices for effectively using microsimulation tools (Wunderlich et al. 2019). The key modeling steps and principles of this analysis will be outlined in the report, with a focus on their relevance in facilitating fusion and calibration of traffic simulation data using emerging data sources. The project team's proposed approach uses multiple data sources, which can enhance the granularity of data and improve the fidelity of simulation models. Recent advancements in sensor technology and in-vehicle monitoring systems have enabled the collection of extensive driver behavior data through real-world driving datasets.

Data Collection Recommendations

Existing agency-specific techniques and documents focusing on data collection can be tailored to project-specific requirements. These sources should be employed with consideration for suitable data collection methods. General resources on traffic data collection include:

- *Introduction to Traffic Engineering: A Manual for Data Collection and Analysis* (Currin 2012).
- *Manual of Transportation Engineering Studies* (Robertson 1994).
- HCM *2010* (TRB 2010).
- *Traffic Analysis Toolbox Volume VI* (Dowling 2007).
- *Traffic Monitoring Guide* (FHWA 2022).

Given that data collection can be costly, researchers should clearly define the required data for the study and allocate the budget accordingly. In situations with funding constraints, resources should be used in a way that guarantees the availability of sufficient, high-quality data to examine the potential impact of transportation investments.

Data For Base Model Development

Microsimulation primarily requires three categories of input data:

- Road geometry (lengths, lanes, curvature).
- Traffic controls (signal timing, signs).
- Demand (entry volumes, turning volumes, OD table).

Microsimulation models also necessitate data on vehicle and driver characteristics such as vehicle length, maximum acceleration rate, and driver aggressiveness.

Data for Determining Travel Conditions

In addition to being necessary for base model development, travel demand data are also key for identifying travel conditions such as entry volumes, turning movements at intersections within the study area, and vehicular OD tables. Other data that can influence travel conditions include weather data, incident data, transit data, freight data, bottleneck throughput data, and travel time data.

Data for Model Calibration

Model calibration involves adjusting the model's parameters so that its outputs match observed data as closely as possible. Calibration data should represent the range of conditions the model is expected to replicate. For this reason, the report also provides instructions and methods for gathering data required for model calibration.

Summary of the 2019 Update to *Traffic Analysis Toolbox Volume III*

The 2019 update to *Traffic Analysis Toolbox Volume III* (Wunderlich, Vasudevan, and Wang 2019) includes detailed technical information for data collection and analysis, model calibration, and analysis of alternatives. The update incorporates a complex corridor-based example problem to illustrate the application of the updated information.

The following are highlights of this update:

- Focus calibration and alternatives analysis on the representation of time-dynamic system performance measures, including bottleneck formation and dissipation.

- Encourage comprehensive experimental design based on various travel conditions instead of solely relying on normative average days.

- Replace subjective criteria with statistically valid and data-derived criteria, eliminating subjectivity from the process.

- Develop a data-driven, repeatable, and potentially automatable calibration process.

- Integrate time-dynamic representation of congestion to improve the realism of simulation analyses.

- Ensure better representation of recurrent and nonrecurrent conditions such as incidents, weather, and variations in travel demand, and help integrating various data sources to create meaningful and realistic models.

- Emphasize accurate modeling and calibration of bottleneck dynamics.

Following this information, the transportation community can adopt a data-driven and statistically valid approach for conducting objective analyses. Analysts are recommended to engage stakeholders and partners throughout the application of microsimulation models to ensure the credibility of results, recommendations, and conclusions while minimizing any potential unforeseen tasks.

Performance measures should be carefully selected to differentiate the alternatives identified in the objectives statement. These measures can be observed using field data or generated from simulation outputs and should effectively distinguish between the alternatives.

Microsimulation models offer analytical strengths in various scenarios, such as impacting lane-level capacities and throughput; analyzing time-dynamic congestion patterns; and evaluating multiple intersections, interchanges, and facilities over time. The microsimulation models are valuable for informing decisionmakers in areas such as signalized network systems, freeway operations, managed lane deployments, incident management, corridor management, work zone planning, and ITS technologies and applications.

Adhering to the following principles is important for conducting a practical analytical study using a microsimulation model:

- Use measurable field data.
- Recognize that the quality and quantity of data influence the analysis.
- Collect an appropriate quantity of data based on the required analytical accuracy.
- Use relatively recent and time-variant data.
- Use contemporaneous data to ensure relevance and accuracy.

Potential Calibration Measures

This section illustrates the objective of calibration (e.g., find the set of parameter values for the model that best reproduces observed system performance measures) and identifies suitable calibration measures for datasets fused by emerging and legacy data. The following are general considerations for an effective analytical study using a microsimulation model:

- Establish the objective, hypotheses, and well-defined performance measures before developing the microsimulation model. Analysts should consider the unique definitions and interpretations of terms at the microscopic level.

- Select the appropriate tool based on its limitations ensuring it accurately represents traffic operations theory and aligns with the study's purpose, needs, and scope.

- Acquire sufficient and reliable data, recognizing that the quality and quantity of data directly influence the accuracy of the microsimulation model results.

- Calibrate the model specifically to local conditions and prevailing travel conditions to ensure an accurate representation and effective management of transportation systems.

Establishing appropriate calibration measures to calibrate traffic simulation models using new and emerging data sources is important. Transportation agencies often employ various methods to estimate travel times and validate their models. These methods include data collection from sensors and cameras, traffic flow modeling, historical data analysis, and real-time traffic monitoring. Model validation typically involves comparing predicted travel times or traffic patterns with observed data to ensure accuracy and reliability.

The CBI tool has facilitated the use of traffic simulation calibration with emerging data sources. This tool adopts approaches that consider various factors such as visibility, weather effects, and the exclusion of congestion caused by traffic signals. Its advanced features allow for a more accurate identification of congestion patterns, distinguishing between recurring and nonrecurring congestion durations. Additionally, the CBI tool can detail the spatial extent of queues over several days, which arise from interactions within vehicular flow. The CBI tool also incorporates new performance measures, offering numerical and graphical representations to assess and rank traffic bottlenecks with a higher level of detail and accuracy than existing methods.

The development of the CBI tool is a breakthrough in traffic analysis, enabling a comprehensive evaluation of congestion patterns and the identification of bottleneck locations. Leveraging sensor speed profiles, this tool enhances the ability to pinpoint areas that require intervention or improvements to alleviate congestion and improve traffic flow. While these calibration measures serve as key indicators, they can be customized to align with specific project requirements and local conditions. With the abundance of emerging data sources, integrating the CBI tool and other data-driven techniques can further enhance the effectiveness and efficiency of traffic simulation calibration.

Data Cleaning, Data Fusion, and Model Calibration Challenges in Traffic Simulation Applications

To investigate the characteristics of identified emerging datasets and suggest potential uses of each for various aspects of traffic simulation calibration, a summary of methodologies for data fusion in general TSE and model calibration is presented in table 14.

Table 14. Methodologies for data fusion in TSE and simulation model calibration.

Fusion Level	Data Fusion Methods	Traffic Flow Analysis and Model Variables
Data-level fusion	Joint probabilistic data association (Fortmann, Bar-Shalom, and Scheffe 1983), K-nearest neighbor (Keller, Gray, and Givens 1985), probabilistic data association (Bar-Shalom and Tse 1975)	Fusion motion data, fuse numerical data, remove outlier and noise, missing value estimation
Feature-level fusion	Kalman Filter (Welch and Bishop 1995), Extended Kalman Filter (Houtekamer and Mitchell 1998), neural network (Lawrence 1993), fuzzy logic (Hájek 2013), Bayesian (Bernardo and Smith 2009), Gaussian Mixture Model (Reynolds 2009)	Travel time estimation, traffic states estimation/prediction, turning ration estimation, traffic flow prediction, traffic passenger prediction, traffic speed estimation/prediction, traffic incident/accident, traffic congestion prediction, pedestrian candidate identification
Decision-level fusion and traffic flow model-based fusion	Dempster-Shafer (Sentz and Ferson 2002), fuzzy logic, software agent, hybrid, and convolutional neural network (O'Shea et al. 2015)	Traffic management transportation, traffic control decision, traffic control signal operation, misbehave vehicle detection, TSE, lane-changing detection

CUSTOMIZING DATA FUSION FRAMEWORK FOR TRAFFIC SIMULATION CALIBRATION

This section customizes the data fusion framework developed under a previous FHWA project for use in microsimulation model calibration.

Overview of the Five-Step Framework to Support Data Fusion, Analysis, and Decisionmaking

The five-step data fusion framework offers a procedure that enhances the decisionmaking capabilities of transportation agencies. This framework is beneficial for personnel lacking technical expertise in data science or analytics and is designed to leverage the opportunities presented by an array of emerging data sources.

Step 1: Data Acquisition and Storage

The framework's foundation rests on data, which inform all subsequent steps. This stage involves identifying, obtaining, and securely storing relevant data, which could range from traditional traffic datasets to more complex, high-frequency big data. Techniques for managing data with varying temporal and spatial granularity are also discussed.

Step 2: Data Cleaning and Fusion

Recognizing the data limitations inherent in the dataset, this step encompasses the process of data cleaning and tagging to address any flaws, uncertainties, or incomplete aspects, thereby improving the accuracy and reliability of the data. Techniques for identifying missing records, duplications, logical inconsistencies, outliers, and stale data are incorporated. Once data are cleaned, the fusion process begins, allowing the amalgamation of multiple datasets to provide rich, unique insights. Five fusion techniques—spatial, temporal, complementary, redundant, and cooperative fusion—are highly recommended for utilization, each offering unique benefits and applications.

Step 3: Data Analysis

Building on cleaned, tagged, and fused data, this stage involves employing analytical techniques to transform data into actionable information such as data mining, multivariate cluster analysis, supervised and unsupervised learning, and reinforcement learning. The aim is to extract insights from complex, multisourced fused data to inform decisionmaking.

Step 4: Decision Implementation

This step uses the actionable information derived from the previous step to formulate and execute decisions. Depending on the operational level of the decision (strategic, tactical, or operational), the decisionmaking process might be automated, manual, or semiautomated with a human-in-the-loop approach. The selection of the decisionmaking approach can depend on factors such as complexity, impact, and time horizon.

Step 5: Evaluation and Iteration

Finally, the effectiveness of the decision is evaluated based on key performance indicators suitable for the decision level. This stage also includes measuring baseline performance, evaluating postdecision performance, visualizing, and communicating performance, and providing feedback to the framework. The goal is to learn from the evaluation process and repeat the framework with improvements, making it a continuous cycle of learning and refinement.

Key Principles of Customization

In customizing the five-step data fusion framework to the datasets discussed in chapter 3, the following principles are proposed with a summary in table 15:

- **Nature of the data:** Recognize each dataset's unique attributes and potential applications. For instance, data source A offers detailed OD trip data, providing insights into travel patterns and geographic-specific data. Data source B probe vehicle data offers semicontinuous trajectory data, allowing for real-time route and detour analysis. High-resolution vehicle trajectory data provide a variety of traffic scenarios with high recording frequency. Loop detectors and mobile phone-based GPS data are key for real-time TSE and historical analysis. Assessing the data's reliability, accuracy, granularity, and timeline is key, and stakeholders should qualitatively assess the strengths and weaknesses of each data source.

- **Domain knowledge:** Apply domain knowledge about traffic microsimulation models to inform the data fusion process. Analysts should understand the specifics of traffic microsimulation models and how they relate to the data. This understanding of models should include systemwide travel patterns; FD parameters including traffic volume, density, and speed distribution function; and microscopic model parameters such as car-following model parameters.

- **Data compatibility:** Ensure the datasets can be meaningfully combined during the framework's fusion and analysis stages. The data should be compatible in terms of both temporal and spatial granularity. Emphasis should be placed on map matching needs for aligning GPS points to underlying networks and the fusion of complementary data sources such as speed from probe data, volume from loop detectors, and weather data. For example, GMNS offers a common format to organize and structure diverse datasets, thereby facilitating a better understanding of the data and enabling interoperability in traffic simulation model calibration.

- **Data cleaning and preparation:** Check large, diverse datasets for accuracy, consistency, and completeness and prepare the data for subsequent framework steps. In traffic simulation model calibration, data cleaning and preparation involve ensuring all required data points are present and complete, checking for consistency in data formatting, and eliminating duplicates. Range constraints are applied to maintain plausibility, while error checking identifies anomalies, and standard validation ensures data align with accepted benchmarks. The age of data is also considered to ensure relevance to current traffic conditions. Missing or inconsistent data may be handled using data imputation techniques while documenting all modifications for transparency.

- **Algorithm selection:** Choose algorithms that are the most suitable depending on the data and specific objectives. For instance, multivariate cluster analysis might be useful for identifying patterns in the OD trip data, while supervised learning (Zhu 2005) could predict traffic patterns based on historical data. A need to fully integrate the time-dynamic representation of congestion was identified, highlighting the importance of evolving away from static or average demand patterns toward a more realistic reflection of congestion dynamics.

- **Model calibration and validation:** Use emerging data sources to calibrate and validate microsimulation models, which involves using the data to estimate the model's parameters, then using separate data to test the model's performance. The framework should provide information on how to perform these steps effectively. Emphasize accurate bottleneck modeling by correctly representing the bottleneck location, onset time, and duration. Also important is aiding in the integration of various data sources to create meaningful and realistic models for recurrent and nonrecurrent conditions.

Table 15. Analysis of data fusion steps, strategies, and customization for traffic simulation calibration using emerging data sources.

Data Fusion Step	Strategies	Customization for Traffic Simulation Calibration Using Emerging Data Sources
1	1.1 Identify and obtain different emerging data sources	Stakeholders and model users should understand the unique attributes and applications of each dataset. Assess the reliability, accuracy, granularity, and timeline of each dataset.
	1.2 Ingest/store data	Apply domain knowledge about traffic microsimulation models to inform the data fusion process. Ensure compatibility of the datasets in terms of both temporal and spatial granularity.
2	2.1 Data quality and cleaning	Stakeholders and model users should check the data for accuracy, consistency, and completeness. Apply range constraints to maintain plausibility, error checking to identify anomalies, and standard validation to ensure alignment with accepted benchmarks.
	2.2 Data fusion	Understand how to meaningfully merge datasets, considering temporal and spatial compatibility. Enhance the calibration of the traffic simulation model by fusing complementary datasets.
3	3.1 Data analysis methods	Select appropriate algorithms based on the nature of the data and specific objectives. Consider using multivariate cluster analysis for identifying patterns, supervised learning for predicting traffic patterns, and physics-informed neural networks for embedding physical laws governing traffic flow into the learning process.
	3.2 Strategies implementation	Identify different traffic problems such as temporal (time of day, week, year) and spatial (corridor, segment, station, work zones). Consider evolving from static or average demand patterns toward a more realistic reflection of congestion dynamics.

Data Fusion Step	Strategies	Customization for Traffic Simulation Calibration Using Emerging Data Sources
4	4.1 Manual decisionmaking	Use emerging data sources for calibrating and validating microsimulation models. This involves estimating the model's parameters using the data and testing the model's performance using separate data. Focus on the correct representation of the bottleneck location, onset time, and duration.
	4.2 Automated decisionmaking	Automated decisionmaking processes can save time and labor for microsimulation model calibration. Accurate methodologies can improve the reliability and accuracy of results.
5	Evaluation for results	A cross-validation method should be applied for parameters in different locations. This ensures the inputs for microsimulation model calibration accurately represent the real-world situation. Provide information on integrating various data sources to create realistic models for both recurrent and nonrecurrent conditions.

In step 2 of table 15, data quality and cleaning considerations when using emerging data sources are presented:

- **Device issues:** Adopting a multidevice configuration requires additional indexing, synchronization, and mapping skills and methods to ensure the device's reliability as a data collection tool. There is a cost involved in ensuring device stability and maintenance.

- **Data preprocessing issue:** Managing huge and unstructured data require complex preprocessing methods. Particularly in a multisensor and heterogeneous sources environment, dealing with such data becomes a challenge. Data quality and missing data values must be handled to avoid inaccurate results.

- **Research issue:** Identifying key data to fit the project purpose can be a challenge. The diversity of algorithms and methods requires intensive study to formulate a suitable model to solve domain problems. Finding the complete, suitable, and relevant dataset is another challenge when evaluating data fusion proposed models or solutions, especially for a data-dependent model.

- **System architecture issue:** Using hardware specifications requires compatible software to be integrated to ensure system performance and stability. Integrating different devices, sensors, algorithms, and methods has complexities. Collecting data from various sources and inputting them into the data fusion model requires integrating several systems equipped with stable communication networks. In cloud computing as a data center, security and privacy leakage are other key areas that must be considered.

- **Processing complexity issue:** Collecting and processing data from various sources in real time requires complex, distributed, and dynamic systems. Heterogeneous data sources increase data completeness and reliability. However, dealing with heterogeneous data with different characteristics may require a model combination as a solution that increases the complexity.

Underlying Components for Cross-Layer Consistency in Traffic Simulation Model Calibration

The following specifics are key, interconnected components of traffic simulation model calibration. These components play a key role in maintaining cross-layer consistency by integrating and leveraging different resolutions of datasets. Trajectory-based modeling provides a detailed understanding of individual vehicle behavior and traffic dynamics, allowing for an accurate representation of real-world traffic flow. Joint estimation of traffic state and queue profiles enhances the observability of the traffic system, enabling effective management of congestion and queue evolution. Microscopic traffic flow modeling focuses on capturing fine-grained details of traffic patterns and interactions, ensuring the fidelity of simulation models. By incorporating these components into the calibration process, the traffic simulation models can achieve higher accuracy and reliability, facilitating better decisionmaking and optimization of transportation systems.

In the calibration process, capacity parameters play an important role in defining a given roadway's maximum sustainable flow rate. The process begins with calibrating these capacity parameters, which includes determining factors such as the number of vehicles that can pass through a roadway segment in a given period. The process of calibrating the capacity parameters involves collecting field measurements of capacity, which can be obtained from traffic counts and sensor data. These real-world measurements serve as benchmarks against which model estimates of capacity are compared.

Next, the process involves obtaining model estimates of capacity. These predictions are generated from traffic simulation models, which consider factors such as roadway geometry, traffic volume, and vehicle speed. The accuracy of these estimates is key, as they directly impact the fidelity of the simulation models.

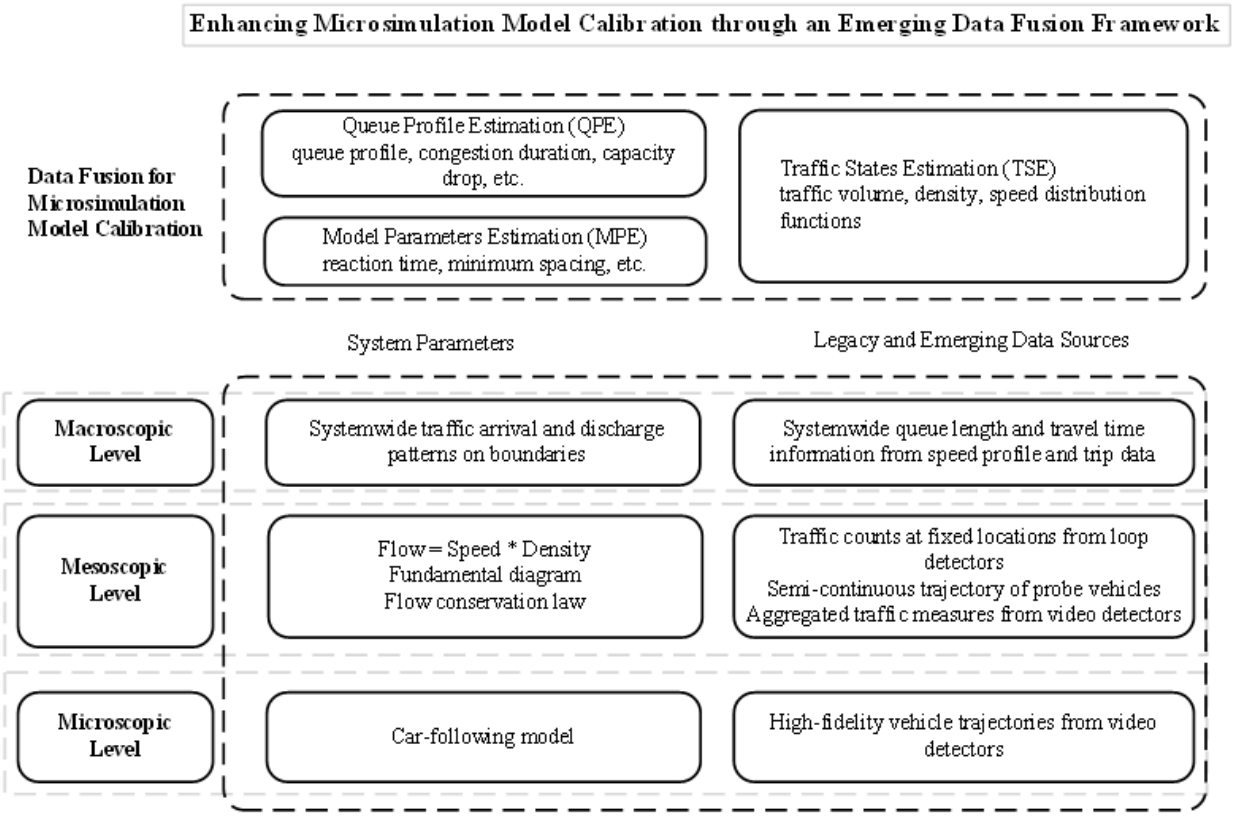
Once the field measurements and model estimations of capacity are ready the next step is selecting calibration parameters. These variables such as traffic demand and driving behavior parameters within the traffic simulation models can be adjusted to better align the model estimates with the field measurements, traffic signal timings, and vehicle speed limits.

With the calibration parameters selected, the objective calibration function is set to quantify the difference between the model estimates and the field measurements. The calibration process aims to minimize this difference, thereby maximizing the accuracy of the simulation models.

Additionally, route choice parameters are calibrated, typically proceeding in two phases: global calibration and link-specific fine-tuning. The global calibration phase involves adjusting parameters that affect the overall behavior of drivers in choosing routes, such as travel time and route length preferences. The link-specific fine-tuning phase focuses on adjusting parameters that affect the choice of specific links in the network, such as turn penalties at intersections.

Simultaneously, the analysis of traffic flow parameters is key. To bridge macroscopic system states more effectively, like flow and density, and to integrate legacy data including sensor counts, occupancy, and segment-based speed measurements, the calibration of variables is divided into microscopic, mesoscopic, and macroscopic studies, as illustrated in figure 6. The choice between these studies depends on the level of detail required. For instance, microscopic-level studies delve into car-following and lane-changing models, providing granular insights into individual vehicle behavior. Mesoscopic-level studies, however, could examine headway and spacing distributions, offering an intermediate level of detail. Finally, macroscopic-level studies may encompass the FD and traffic wave models, presenting an aggregate view of traffic flow and speed-density relationships.

The primary objective of the framework and methodology development in figure 7 is to arm users with the necessary tools to effectively use emerging data sources to develop detailed and accurate traffic simulations. By illustrating the customization of these general data fusion steps, the report intends to improve users' ability to maximize value from these data sources. This approach not only enhances the traffic simulation models' accuracy but also increases their relevance in today's traffic situations. The capacity to adapt and apply these steps further highlights the flexibility of the data fusion framework and its appropriateness for a wide range of traffic simulation calibration tasks.



Source: FHWA.

Figure 7. Diagram. Emerging data fusion framework for microsimulation model calibration within a multiresolution framework.

Macroscopic and Mesoscopic Modeling Framework for Enhanced Data Fusion Using Multisource Heterogeneous Data

Accurate traffic system state identification is key for the design and execution of control strategies. Emerging technologies, including mobile sensors, LBS, and participatory sensing, offer richer traffic observations, necessitating a system state identification framework to enhance traffic system observability. This report addresses three key challenges in identifying traffic system state, as shown in figure 6. Essentially, the proposed approach aims to enhance traffic analysis by estimating time-varying traffic states like flow rate, density, and speed on specific road segments and queue lengths at congested areas.

Combining these two approaches in a unified model can offer a more comprehensive understanding of traffic conditions. However, combining the two approaches is challenging due to the complexity of data fusion, which involves intricate correlations among components. The goal is to create an efficient and consistent model-driven framework for data fusion using various traffic data sources and advanced machine learning techniques, improving local estimates, aggregated traffic modeling, and hierarchical traffic control. Table 16 shows different computationally efficient and inherently consistent model-driven frameworks for using multisource heterogeneous traffic data and advanced computational techniques from general traffic data fusion applications.

Table 16. Comparison of related studies for data fusion and traffic flow mode calibration.

Source	Task	Modeling Approach	Solution Method	Benefits and Challenges
Sun, Jin, and Ritchie (2017)	TSE	Nonlinear optimization	Closed-form formula, Gauss-Newton method	Estimate traffic parameters and states simultaneously. Initial state estimation is sensitive to the measurement errors.
Shi, Mo, and Di (2021)	TSE	Nonlinear optimization	Gradient descent method (Amari 1993)	Incorporate traffic flow models and field observations into a machine learning framework. A single data source is used.
Wang et al. (2016)	TSE	State-space model	Particle filtering (Djuric et al. 2003)	Enable joint TSE and incident detection. High computational complexity.
Canepa and Claudel (2017)	TSE	Mixed integer linear programming	Mathematical programming solver	Leverage Hamilton-Jacobi equations to solve estimation problems exactly. Limited to small-scale applications.
Liu et al. (2009)	QPE	Lighthill-Whitham-Richards shock wave theory (Leclercq 2007)	Numerical derivations	Estimate time-dependent queue length for congested signalized intersections. Not ideal for oversaturated scenarios.
Ramezani, Haddad, and Geroliminis (2015)	QPE	Shock wave theory (Rassweiler et al. 2011)	Data mining	Applicable in oversaturated conditions. Queue development and dissipation varying among different lanes can be further considered.
Duret and Yuan (2017)	TSE	Lighthill-Whitham-Richards model in Lagrangian space (Porto, Senatore, and Zaldarriaga 2014)	Numerical solutions obtained with Godunov scheme	Propose a Lagrangian space formulation to assimilate Eulerian and Lagrangian observations. Require accurate preset traffic flow parameters.
Jabari and Liu (2013)	TSE	State-space model (Roesser 1975)	Kalman filtering (Welch and Bishop 1995)	Analytically tractable Gaussian model of (stochastic) first-order traffic flow. Discretized space-time state representation affects the investigation of wave propagation dynamics.

Microscopic Trajectory-Based Traffic Flow Modeling for Simulation Development

The domain of traffic flow studies is witnessing a range of opportunities that arise primarily from advancements in technology, the refinement of data collection methodologies, and insights derived from comprehensive data analysis.

Technological Advancements in Traffic Monitoring and Data Collection

The incorporation of modern technologies such as roadside video cameras, drones, and video-based traffic flow monitoring systems mark a leap from the early days of high-speed aerial photography and manual data reduction (Treiterer and Myers 1974; Coifman, Li, and Xiao 2018). These systems provide detailed tracking of multiple vehicles simultaneously, capturing nearly the entire scope of traffic conditions and vehicle motion information. High-resolution cameras deployed on elevated structures or aerial vehicles further enhance data quality, leading to a deeper understanding of individual driver behaviors (Knoop, Hoogendoorn, and Van Zuylen 2008; Zhang et al. 2016; Krajewski et al. 2018).

Enhanced Understanding of Driving Behavior and Traffic Phenomena

Comprehensive data analysis has shed light on several previously overlooked aspects of driving behavior and traffic phenomena. Key among these aspects are the understanding of asymmetric driving behavior and its impact on traffic flow (Yeo 2008; Yeo and Skabardonis 2009), and the identification of the driver memory effect, where a vehicle's acceleration and deceleration is influenced by historical traffic conditions (Treiber and Helbing 2003; Wang et al. 2019). These insights offer an opportunity for more accurate traffic modeling and prediction.

Indepth Analysis of Complex Traffic Behaviors

The detailed analysis of complex traffic behaviors, such as lane changing and car following, has been made possible by the availability of trajectory data. These data have highlighted variations in these behaviors based on drivers and vehicle types and have been instrumental in the calibration of behavioral parameters of different driving styles (Zheng 2014; Sharma, Zheng, and Bhaskar 2018).

Additional Challenges in Framework Customization

Emphasis is also placed on the need to procure additional trajectory data to enhance the precision of traffic flow studies. Although strides have been made with the advent of datasets such as NGSIM, gaps persist, which emphasizes the need for more comprehensive data sources and deep customization.

First, the current trajectory data are limited in spatial and temporal scope. For instance, the NGSIM dataset only covers small highway or arterial segments, and there is a need for more comprehensive tracking of vehicles and traffic flow dynamics (FHWA 2016). Second, there is a limited sampling of traffic scenarios. The NGSIM data, for example, contains very little free-flow trajectory data and limited road segment geometries, hindering the understanding of some important traffic phenomena and the calibration of traffic flow models.

Third, inaccuracies exist in video-based trajectory collection methods. Despite advancements in image processing techniques, vehicle recognition, and localization errors are still a concern because of optical constraints of cameras and potential increases in financial costs due to multiple camera usage. Finally, the high equipment costs and data preprocessing time obstruct the collection of more high-quality trajectory data. Traditional video-based collection methods can only monitor fixed road segments, leading to financial and time costs and limiting the widespread usage of these methods.

Table 17 provides an overview of four different datasets categorized based on their data features and their applications in traffic simulation calibration. The letters Y (yes) and N (no) represent the mapping of each data source to the detailed calibration needs at different resolutions.

The following explains each column in table 17:

- Data source/category: Specifies the source or category of the datasets used in traffic simulation calibration elements, including four datasets mentioned in chapter 2.

- Data features: Describes the specific features or variables captured in each dataset across different emerging and legacy data sources.

- Macroscopic systemwide travel pattern: Indicates if the dataset provides information on the systemwide travel OD pattern, traffic arrival on boundaries.

- Time-dependent congestion: Specifies if the dataset includes information on systemwide queue length and travel time derived from speed profiles and trip data.

- FD: Specifies if the dataset provides traffic flow system measures, including FD parameters (traffic volume, density, and speed distribution function).

- Bottleneck-related system measures: Indicates if the dataset includes other bottleneck-related system measures, such as queue profile, congestion duration, capacity drop, and similar parameters.

- Microscopic model parameters: Indicates if the dataset is used for microsimulation and car-following model calibration.

- Route choice: Indicates if the dataset includes information on route choice.

- Applications of related simulation model: Describes common purposes and applications of each dataset, such as transportation planning, travel monitoring, performance measurement, OD analysis, travel modeling, traffic flow analysis, autonomous vehicle algorithm development, macroscopic observation, and TSE on freeway segments.

Table 17. Mapping of four different datasets and their applications in traffic simulation calibration elements.

Data Source/ Category	Data Features	Macroscopic Systemwide Travel Pattern	Time-Dependent Congestion	FD Parameters	Bottleneck-Related System Measures	Microscopic Model Parameters	Route Choice	Applications of Related Simulation Model
OD trip data	Vehicle volume, zone ID, day type, average daily zone traffic	Y	Y	Y	N	N	Y (partially related to subarea)	Transportation planning, travel monitoring
Semicontinuous trajectory data	Timestamp, position, speed, heading	Y	Y	Y	Y	N	N	Performance measurement, OD analysis, travel monitoring and modeling
High-resolution vehicle trajectory data	Vehicle trajectories	Y	Y	Y	Y	Y	Y	Traffic flow analysis, autonomous vehicle algorithm development
Loop detector data and mobile phone-based GPS data	Vehicle volumes, timestamps of probe vehicles, travel time	Y	Y	Y	Y	N	N	Macroscopic observation and TSE on freeway segments

CHAPTER 5. CONCLUSION AND RECOMMENDATIONS

Adding new data sources in traffic simulation and modeling is important, leading to more detailed and flexible depictions of real-world traffic conditions. This process involves important steps such as data integration, calibration methods, and validation procedures. Data integration refers to incorporating new sources, such as connected vehicles, UAVs, crowdsourced data, and social media, into calibrating traffic simulation models. Adding new data sources leads to increased accuracy and realism in simulated traffic conditions, a better understanding of dynamic demand patterns and driving behaviors, and the ability to monitor real-time changes in traffic flow and network conditions.

Calibration methods involve adjusting the traffic simulation model using data from these new sources to ensure the model accurately mirrors real-world traffic characteristics. The outcomes include more precise simulations of traffic flow, travel times, and congestion patterns and more reliable predictions of network performance under different scenarios. The calibration methods also provide a better way to estimate OD demand patterns and evaluate the impact of transportation policies and strategies.

Validation procedures, the third step, involve checking the accuracy and reliability of the traffic simulation model using real-world data from new sources. These steps allow for checking the model's ability to reproduce observed traffic conditions, identifying any differences in simulation results, and confirming the model's usefulness in decisionmaking and policy evaluation. The process ends with a review of how effectively integrating new data sources has improved simulation accuracy.

To illustrate how to use emerging and legacy data for improving accuracy of traffic simulation model calibration, chapter 2 of this report provided an overview of potential emerging data sources for traffic simulation model calibration, aiming to improve traffic simulation and analysis. These data sources were categorized based on commonly available emerging datasets, offering a perspective on their applicability. This chapter can help users of traffic simulation models gain a deep understanding of the potential emerging data sources for calibration and their effective use in traffic simulation and analysis.

Chapter 3 highlighted the strategic fusion of legacy and emerging data sources, providing benefits for OD demand, routing, and car-following models. It also examined the strengths and weaknesses of emerging data sources and suggested techniques for preparing these data for calibration. Chapter 4 of this report delved into the customization process of the data fusion framework (Hale et al. 2022) and *Traffic Analysis Toolbox Volume 3: Guidelines for Applying Traffic Microsimulation Modeling Software (2019 Update)* (Wunderlich et al. 2019), tailored for traffic microsimulation model calibration. Each step of the data fusion framework for supporting data fusion, analysis, and decisionmaking can be tailored to address the needs of traffic simulation calibration using emerging data sources.

The information provided in these chapters can help transportation practitioners and researchers leverage emerging data sources effectively, enhance the accuracy of traffic simulation models, and make informed decisions to optimize transportation systems. A comprehensive understanding of emerging data sources and the customization process can help advance traffic simulation calibration practices and improve the efficiency and effectiveness of traffic management strategies.

REFERENCES

Adu-Gyamfi, Y. O., S. K. Asare, A. Sharma, and T. Titus. 2017. "Automated Vehicle Recognition with Deep Convolutional Neural Networks." *Transportation Research Record* 2645:1, 113–122.

Airsage. 2000. "AirSage for Transportation Consulting Firms" (web page). https://airsage.com/solutions/transportation-consulting-firms/, last accessed March 27, 2024.

Amari, S. I. 1993. "Backpropagation and Stochastic Gradient Descent Method." *Neurocomputing*, 5:4–5, 185–196.

Ambühl, L., and M. Menendez. 2016. "Data Fusion Algorithm for Macroscopic Fundamental Diagram Estimation." *Transportation Research Part C: Emerging Technologies* 71: 184–197.

Auld, J., M. Hope, H. Ley, V. Sokolov, B. Xu, and K. Zhang. 2016. "POLARIS: Agent-Based Modeling Framework Development and Implementation for Integrated Travel Demand and Network and Operations Simulations." *Transportation Research Part C: Emerging Technologies* 64:101–116.

Avner, J. 2018. *"Using Big Data in Small and Medium Sized Regions."* Presented at the 97th Annual Meeting of the Transportation Research Board, Washington, DC. https://onlinepubs.trb.org/onlinepubs/Conferences/2018/Tools/JAvner.pdf, Last accessed November 8, 2018.

Bachmann, C., B. Abdulhai, M. J. Roorda, and B. Moshiri. 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.

Bar-Shalom, Y., and E. Tse. 1975. "Tracking in a Cluttered Environment with Probabilistic Data Association." *Automatica* 11:5, 451–460.

Behrisch, M., L. Bieker, J. Erdmann, and D. Krajzewicz. 2011. *SUMO–Simulation of Urban Mobility: An Overview*. In *Proceedings of SIMUL 2011, The Third International Conference on Advances in System Simulation*. Barcelona, Spain: International Academy, Research, and Industry Association. https://sumo.dlr.de/pdf/simul_2011_3_40_50150.pdf, last accessed May 10, 2024.

Bernardo, J. M., and A.F. Smith. 2009. *Bayesian theory* (Vol. 405). New York, NY: John Wiley & Sons.

Canepa, E. S., and C. G. Claudel. 2017. "Networked Traffic State Estimation Involving Mixed Fixed-Mobile Sensor Data Using Hamilton-Jacobi Equations." *Transportation Research Part B: Methodological* 104:686–709.

Claros, B., G. Vorhes, M. Chitturi, A. Bill, D. A. Noyce. 2022. “Filling Traffic Count Gaps with Connected Vehicle Data.” *International Conference on Transportation and Development 2022* 192–199.

Coifman, B., L. Li, and W. Xiao. 2018. “Resurrecting the Lost Vehicle Trajectories of Treiterer and Myers with New Insights into a Controversial Hysteresis. *Transportation Research Record* 2672(20): 25–38.

Currin, T. R. 2012. *Introduction to Traffic Engineering: A Manual for Data Collection and Analysis*. Boston, MA: Cengage Learning.

Desai, J., J. K. Mathew, H. Li, R. S. Sakhare, D. Horton, and D. M. Bullock. 2022. *National Mobility Analysis for All Interstate Routes in the United States: December 2022*. West Lafayette, IN: Purdue University.

Djuric, P. M., J. H. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. F. Bugallo, and J. Miguez. 2003. “Particle filtering.” *IEEE Signal Processing Magazine*, 20:5, 19–38.

Dowling, R. 2007. *Traffic Analysis Toolbox Volume VI: Definition, Interpretation, and Calculation of Traffic Analysis Tools Measures of Effectiveness.* Report No. FHWA-HOP-08-054. Washington, DC: Federal Highway Administration.

Duret, A., and Y. Yuan. 2017. “Traffic State Estimation Based on Eulerian and Lagrangian Observations in a Mesoscopic Modeling Framework.” *Transportation Research Part B: Methodological* 101:51–71.

European Commission. 2020. “Open ACC database” (web page). Joint Research Centre. https://data.jrc.ec.europa.eu/dataset/9702c950-c80f-4d2f-982f-44d06ea0009f, last accessed March 6, 2024.

FHWA. 2016. Next Generation Simulation (NGSIM) Vehicle Trajectories and Supporting Data. Provided by ITS DataHub through Data.transportation.gov. http://doi.org/10.21949/1504477, last accessed May 9, 2024.

FHWA. 2017. “NGSIM-I-80-Trajectory-Animation” (NGSIM animation files in GitHub repository). https://github.com/cemsaz/NGSIM-trajectory-animation, last accessed March 6, 2024.

FHWA. 2018. “About Next Generation Simulation (NGSIM)” (NEXTA visualization tool files in GitHub repository). https://github.com/xzhou99/NeXTA_4_Trajectory_Visualization, last accessed March 6, 2024.

FHWA. 2019. *Congestion Bottleneck Identification (CBI) Tool* (software). https://highways.dot.gov/research/resources/software/congestion-bottleneck-identification-cbi-tool-software-download, last accessed May 9, 2024.

FHWA. 2020. "General Modeling Network Specification (GMNS)" (GMNS files in GitHub repository). https://github.com/zephyr-data-specs/GMNS, last accessed March 6, 2024.

FHWA. 2022. “Traffic Monitoring Guide” (web page). https://www.fhwa.dot.gov/policyinformation/tmguide/, last accessed May 9, 2024.

FHWA. 2024. “The Next Generation Simulation Program (NGSIM)” (web page). https://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm, last accessed March 6, 2024.

Fortmann, T., Y. Bar-Shalom, and M. Scheffe. 1983. “Sonar Tracking of Multiple Targets Using Joint Probabilistic Data Association.” *IEEE Journal of Oceanic Engineering* 8, no. 3: 173–184.

Geiger, A., P. Lenz, C. Stiller, and R. Urtasun. 2013. “Vision Meets Robotics: The KITTI Dataset.” *The International Journal of Robotics Research* 32, no. 11: 1231–1237.

Gong, L., and W. Fan. 2017. “Applying Travel-Time Reliability Measures in Identifying and Ranking Recurrent Freeway Bottlenecks at the Network Level.” *Journal of Transportation Engineering, Part A: Systems* 143 no. 8: 04017042.

Gupta, S., J. A. Stamatoyannopoulos, T. L. Bailey, and W. S. Noble. (2007). “Quantifying Similarity Between Motifs.” *Genome Biology*, *8*(2): R24.

Hájek, P. 2013. *Metamathematics of Fuzzy Logic* (Vol. 4). Berlin, Germany: Springer Science+Business Media.

Hale, D. K., X. Li, A. Ghiasi, D. Zhao, F. Khalighi, M. Aycin, and R. M. James. 2021. *Trajectory Investigation for Enhanced Calibration of Microsimulation Models.* Report No. FHWA-HRT-21-071. Washington, DC: Federal Highway Administration.

Hale, D., M. Pack, and S. Qian. 2022. “A Framework to Support Data Fusion, Analysis, and Decisionmaking.” Presented at the *Transportation Research Board Freeway Operations Committee Midyear Meeting.* Washington, DC: Transportation Research Board Operations Committee.

Herrera, J. C., D. B. Work, R. Herring, X. J. Ban, Q. Jacobson, Q. and A. M. Bayen. 2010. “Evaluation of Traffic Data Obtained via GPS-Enabled Mobile Phones: The Mobile Century Field Experiment.” *Transportation Research Part C: Emerging Technologies* 18 no. 4: 568–583.

Horni, A., K. Nagel, and K. W. Axhausen, (Eds.) 2016. *The Multi-Agent Transport Simulation MATSim* (pp. 3–7). London, England: Ubiquity Press.

Houtekamer, P. L., and H. L. Mitchell. 1998. “Data Assimilation Using an Ensemble Kalman Filter Technique.” *Monthly Weather Review* 126, no. 3:796–811.

Islam, Z., and M. Abdel-Aty. 2023. “Traffic Conflict Prediction Using Connected Vehicle Data.” *Analytic Methods in Accident Research* 39: 100275.

INRIX. 2004. “Traffic Data” (web page). https://inrix.com/products/ai-traffic/, last accessed May 10, 2024.

Jabari, S. E., and H. X. Liu, H. X. 2013. "A Stochastic Model of Traffic Flow: Gaussian Approximation and Estimation." *Transportation Research Part B: Methodological*, 4715–41.

Keller, J. M., M. R. Gray, and J. A. Givens. 1985. "A Fuzzy K-Nearest Neighbor Algorithm." *IEEE Transactions on Systems, Man, and Cybernetics* no. 4:580–585.

Khadka, S., P. T. Li, and Q. Wang. 2022. "Developing Novel Performance Measures for Traffic Congestion Management and Operational Planning Based on Connected Vehicle Data." *Journal of Urban Planning and Development* 148, no. 2:04022016.

Khadka, S., P. S. Wang, P. T. Li, and F. J. Torres. 2023. "A New Framework for Regional Traffic Volumes Estimation with Large-Scale Connected Vehicle Data and Deep Learning Method. *Journal of Transportation Engineering, Part A: Systems* 149 no. 4:04023015.

Knoop, V. L., S. P. Hoogendoorn, and H. J. Van Zuylen. 2008. "Capacity Reduction at Incidents: Empirical Data Collected from a Helicopter. *Transportation Research Record* 2071, no. 1:19–25.

Kothuri, S., J. Broach, N. McNeil, K. Hyun, S. Mattingly, M. Miah, K. Nordback, and F. Proulx. 2022. *Exploring Data Fusion Techniques to Estimate Network-Wide Bicycle Volumes*. Portland, OR: Transportation Research and Education Center.

Krajewski, R., J. Bock, L. Kloeker, and L. Eckstein. 2018. "The HighD Dataset: A Drone Dataset of Naturalistic Vehicle Trajectories on German Highways for Validation of Highly Automated Driving Systems." *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*:2118–2125.

Lawrence, J. 1993. *Introduction to Neural Networks*. Nevada City, CA: California Scientific Software Press. https://calsci.com/IntroToNN.html, last accessed March 6, 2024.

Leclercq, L. 2007. "Hybrid Approaches to the Solutions of the "Lighthill–Whitham–Richards" Model." *Transportation Research Part B: Methodological* 41 no. 7:701–709.

Li, H., J. K. Mathew, W. Kim, and D. M. Bullock. 2020. *Using Crowdsourced Vehicle Braking Data to Identify Roadway Hazards*. West Lafayette, IN: Purdue University. https://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1033&context=jtrpaffdocs, last accessed March 6, 2024.

Lu, J. 2022. "Connected and Automated Mobility Modeling on Layered Transportation Networks: Cross-Resolution Architecture of System Estimation and Optimization." Doctoral dissertation. Arizona State University. https://keep.lib.asu.edu/items/171423, last accessed March 8, 2024.

Lu, J., and X. S. Zhou. 2022. *Modeling Partially Schedulable Connected and Automated Mobility Systems on Layered Virtual-Track Networks: Modeling Framework and Open-Source Tools.* Berlin, Germany: ResearchGate. https://www.researchgate.net/publication/364341612, osm2gmns — osm2gmns 0.7.3 documentation, last accessed March 6, 2024.

Newell, G. F. 1993. "A Simplified Theory of Kinematic Waves in Highway Traffic, Part I: General Theory." *Transportation Research Part B: Methodological* 27, no. 4: 281–287.

Ohio Department of Transportation. 2022. *Ohio Design Traffic Forecasting Manual: Volume 3: Travel Demand Modeling*. Columbus, OH: Ohio Office of Statewide Planning & Research. https://transportation.ohio.gov/static/Programs/StatewidePlanning/Modeling-Forecasting/Vol-3-TravelDemandModeling.pdf, last accessed March 4, 2024.

O'Shea, K., and R. Nash. 2015. *An Introduction to Convolutional Neural Networks*. Report No. arXiv preprint arXiv:1511.08458. Ithaca, NY: Cornell University.

Paipuri, M., E. Barmpounakis, N. Geroliminis, and L. Leclercq. 2021. "Empirical Observations of Multi-Modal Network-Level Models: Insights from the pNEUMA Experiment." *Transportation Research Part C: Emerging Technologies* 131, 103300.

Porto, R. A., L. Senatore, and M. Zaldarriaga. 2014. "The Lagrangian-Space Effective Field Theory of Large Scale Structures." *Journal of Cosmology and Astroparticle Physics* 05, 022.

Ramezani, M., J. Haddad, and N. Geroliminis. 2015. "Dynamics of Heterogeneity in Urban Networks: Aggregated Traffic Modeling and Hierarchical Control." *Transportation Research Part B: Methodological* 74:1–19.

Rassweiler, J. J., T. Knoll, K. U., Köhrmann, J. A. McAteer, J. E. Lingeman, R. O. Cleveland, M. R. Bailey, and C. Chaussy. 2011. "Shock Wave Technology and Application: An Update." *European Urology* 59(5):784–796.

Reynolds, D. A. 2009. *Gaussian mixture models*. *Encyclopedia of Biometrics*. Boston, MA: Springer. 741, 659–663.

Robertson, H. D. 1994. *Manual of Transportation Engineering Studies*. Englewood Cliffs, NJ: Prentice Hall.

Roesser, R. 1975. "A Discrete State-Space Model for Linear Image Processing." *IEEE Transactions on Automatic Control* 20(1):1–10.

Sakhare, R. S., M. Hunter, J. Mukai, H. Li, and D. M. Bullock. 2022. "Truck and Passenger Car Connected Vehicle Penetration on Indiana Roadways." *Journal of Transportation Technologies* 12(4):578–599.

Saldivar-Carranza, E., H. Li, J. Mathew, M. Hunter, J. Sturdevant, and D. M. Bullock. 2021. "Deriving Operational Traffic Signal Performance Measures from Vehicle Trajectory Data." *Transportation Research Record* 2675(9):1250–1264.

Saldivar-Carranza, E., H. Li, M. Taylor, and D. M. Bullock. 2022. "Continuous Flow Intersection Performance Measures Using Connected Vehicle Data." *Journal of Transportation Technologies* 12(4):861–875.

Schewel, L., S. Co, C. Willoughby, L. Yan, N. Clarke, and J. Wergin. 2021. *Non-Traditional Methods to Obtain Annual Average Daily Traffic (AADT).* Report No. FHWA-PL-21-030. Washington DC: Federal Highway Administration.

Sentz, K., and S. Ferson. 2002. *Combination of Evidence in Dempster-Shafer Theory*. Report No. SAND2002-0835. Albuquerque, NM: Sandia National Laboratories. https://www.stat.berkeley.edu/~aldous/Real_World/dempster_shafer.pdf, last accessed March 11, 2024.

Shahrbabaki, M. R., A. A. Safavi, M. Papageorgiou, and I. Papamichail. 2018. "A Data Fusion Approach for Real-Time Traffic State Estimation in Urban Signalized Links." *Transportation Research Part C: Emerging Technologies* 92:525–548.

Sharma, A., V. Ahsani, and S. Rawat. 2017. *Evaluation of Opportunities and Challenges of Using INRIX Data for Real-Time Performance Monitoring and Historical Trend Assessment*. Report No. SPR-P1(14) M007. Lincoln, NE: Nebraska Department of Transportation. https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1236&context=ndor, last accessed March 11, 2024.

Sharma, A., Z. Zheng, and A. Bhaskar. 2018. "A Pattern Recognition Algorithm for Assessing Trajectory Completeness." *Transportation Research Part C: Emerging Technologies* 96:432–457.

Shay, N. 2017. "Using StreetLight InSight Data for a Small Area Study: Rickenbacker Area." Columbus, OH: Mid-Ohio Regional Planning Commission. https://www.otdmug.org/wordpress/wp-content/uploads/2017/09/20170908_MORPC_Shay_OTDMUG.pdf, last accessed March 11, 2024.

Shi, R., Z, Mo, and X. Di. 2021. "Physics-Informed Deep Learning for Traffic State Estimation: A Hybrid Paradigm Informed by Second-Order Traffic Models." *Proceedings of the AAAI Conference on Artificial Intelligence* 35(1):540–547.

Sun, Z., W. L. Jin, and S. G. Ritchie. 2017. "Simultaneous Estimation of States and Parameters in Newell's Simplified Kinematic Wave Model with Eulerian and Lagrangian Traffic Data." *Transportation Research Part B: Methodological* 104:106–122.

Tancik, M., V. Casser, X. Yan, S. Pradhan, B. Mildenhall, P. P. Srinivasan, J. T. Barron, and H. Kretzschmar. 2022. "Block-Nerf: Scalable Large Scene Neural View Synthesis." In Proceedings of the *IEEE/CVF Conference on Computer Vision and Pattern Recognition* 8248–8258. New Orleans, LA: IEEE.

Tang, Z., M. Naphade, M. Y. Liu, X. Yang, S. Birchfield, S. Wang, R. Kumar, D. Anastasiu, and J. N. Hwang. 2019. "Cityflow: A City-Scale Benchmark for Multi-Target Multi-Camera Vehicle Tracking and Re-identification." In Proceedings of the *IEEE/CVF Conference on Computer Vision and Pattern Recognition* 8797–8806. New Orleans, LA: IEEE.

Treiber, M., D. Helbing, D. 2003. "Memory Effects in Microscopic Traffic Models and Wide Scattering in Flow-Density Data." *Physical Review E* 68(4):046119.

Treiterer, J., and J. Myers. 1974. "The Hysteresis Phenomenon in Traffic Flow." *Transportation and Traffic Theory* 6:13–38.

Turner, S., I. Tsapakis, I., and P. Koeneman. 2020. *Evaluation of StreetLight Data's Traffic Count Estimates from Mobile Device Data*. Report No. MN 2020-30. St. Paul, MN: Minnesota Department of Transportation. https://rosap.ntl.bts.gov/view/dot/57948, last accessed March 12, 2024.

Wang, D., F. Yang, K. L. Tsui, Q. Zhou, and S. J. Bae. 2016. "Remaining Useful Life Prediction of Lithium-Ion Batteries Based on Spherical Cubature Particle Filter." *IEEE Transactions on Instrumentation and Measurement* 65(6):1282–1291.

Wang, X., R. Jiang, L. Li, Y. L. Lin, and F. Y. Wang. 2019. "Long Memory is Important: A Test Study on Deep-Learning Based Car-Following Model." *Physica A: Statistical Mechanics and its Applications* 514:786–795.

Wejo. 2013. *Real-Time Traffic Data for Smart Mobility Intelligence* (software). https://www.wejo.com/products/real-time-traffic-intelligence.

Wejo Limited. 2023. "Powering More Informed Decision Making" (web page). https://www.wejo.com/, last accessed March 4, 2024.

Welch, G., and G. Bishop. 1995. *An Introduction to the Kalman Filter*. Report No. TR 95-041. Chapel Hill, NC: University of North Carolina. https://www.cs.unc.edu/~welch/media/pdf/kalman_intro.pdf, last accessed March 12, 2024.

Wu, F., D. Wang, M. Hwang, C. Hao, J. Lu, J. Zhang, C. Chou, T. Darrell, and A. Bayen. 2022. *Decentralized Vehicle Coordination: The Berkeley DeepDrive Drone Dataset.* Report No. arXiv:2209.08763v2. Ithaca, NY: Cornell University.

Wu, X., J. Guo, K. Xian, and X. Zhou. 2018. "Hierarchical Travel Demand Estimation Using Multiple Data Sources: A Forward and Backward Propagation Algorithmic Framework on a Layered Computational Graph." *Transportation Research Part C: Emerging Technologies* 96:321–346.

Wunderlich, K. E., M. Vasudevan, and P. Wang. 2019. *Traffic Analysis Toolbox Volume 3: Guidelines for Applying Traffic Microsimulation Modeling Software (2019 Update)*. Report No. FHWA-HOP-18-036. Washington, DC: Federal Highway Administration.

Yang, H., M. Cetin, and Q. Ma. 2020. *Guidelines for Using StreetLight Data for Planning Tasks.* Report No. FHWA/VTRC 20-R23. Charlottesville, VA: Transportation Research Council (VTRC).

Yeo, H., and A. Skabardonis. 2009. "Understanding Stop-and-Go Traffic in View of Asymmetric Traffic Theory." *Transportation and traffic theory 2009: Golden jubilee* 99–115. Boston, MA: Springer. https://link.springer.com/chapter/10.1007/978-1-4419-0820-9_6, last accessed March 12, 2024.

Zhang, W. 2021. "Using Wejo Data for Transportation Monitoring, Safety, and Planning at MAG." Presented at *Adventures in Crowdsourcing: Vehicle Video Analytics and Connected Vehicle Data for Operations and Planning webinar*. Washington, DC: National Operations Center of Excellence.

Zhang, W., G. Jordan, and V. Livshits. 2016. "Generating a Vehicle Trajectory Database from Time-Lapse Aerial Photography." *Transportation Research Record* 2594(1):148–158.

Zheng, Z. 2014. "Recent Developments and Research Needs in Modeling Lane Changing." *Transportation Research Part B: Methodological* 60:16–32.

Zhou, X., and H. Mahmassani. 2006. "Dynamic Origin–Destination Demand Estimation Using Automatic Vehicle Identification Data." *IEEE Transactions on Intelligent Transportation Systems* 7:105–114. https://doi.org/10.1109/TITS.2006.869629, last accessed March 12, 2024.

Zhou, X., and J. Taylor. 2014. "DTALite: A Queue-Based Mesoscopic Traffic Simulator for Fast Model Evaluation and Calibration. *Cogent Engineering* 1(1): 961345. https://doi.org/10.1080/23311916.2014.961345, last accessed March 13, 2024.

Recycled
Recyclable

Recommended citation: Federal Highway Administration,
Emerging Data Cleaning and Fusion for Traffic Model Calibration–Data Fusion for Microsimulation Model Calibration
(Washington, DC: 2024) https://doi.org/10.21949/1521587

HRSO-50/12-24(WEB)E