Trajectory Investigation for Enhanced Calibration of Microsimulation Models

PUBLICATION NO. FHWA-HRT-21-071

AUGUST 2021

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

Microsimulation modeling is a commonly used approach for analyzing transportation alternatives, designing traffic control strategies, predicting future congestion problems, and evaluating advanced vehicle technology impacts. But adequate model calibration and validation continue to be among the greatest challenges to the proper application of microsimulation in the decisionmaking process. Thanks to recent advancements in data collection and processing technologies, there is significant interest in collecting trajectory-level data and using these data to calibrate driver behavior within microsimulation models.

The purpose of this report is to document the development of a novel methodology for calibrating microsimulation models using vehicle trajectory data. Given that vehicle trajectory data are not readily available, the research team collected large trajectory level datasets by mining video data collected via drones and helicopters at four sites around the United States. The team conducted four case studies comparing state-of-the-practice (traditional) calibration to the trajectory calibration method, to demonstrate the value added by using trajectory-level data for model development and calibration. This final report will be of interest to State and local departments of transportation interested in improving the state of practice for microsimulation model calibration and validation applied within their jurisdictions.

Brian P. Cronin, P.E.
Director, Office of Safety and Operations
Research and Development

Recommended citation: Federal Highway Administration, *Trajectory Investigation for Enhanced Calibration of Microsimulation Models.* (Washington, DC: 2021)
https://doi.org/10.21949/1521658

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. FHWA-HRT-21-071	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Trajectory Investigation for Enhanced Calibration of Microsimulation Models		5. Report Date August 2021 6. Performing Organization Code:
7. Author(s) David K. Hale (0000-0001-5486-9367), Xiaopeng Li (0000-0002-5264-3775), Amir Ghiasi (0000-0002-0986-9840), Dongfang Zhao (0000-0002-5424-9915), Farnoush Khalighi (0000-0003-3353-5194), Murat Aycin (0000-0002-5798-3381), Rachel James (0000-0001-9138-510X)		8. Performing Organization Report No.
9. Performing Organization Name and Address Leidos, Inc. 11251 Roger Bacon Drive Reston, VA 20190		10. Work Unit No. 11. Contract or Grant No. DTFH6116D00030
12. Sponsoring Agency Name and Address Office of Operations Research and Development Federal Highway Administration 6300 Georgetown Pike McLean, VA 22101		13. Type of Report and Period Covered Final Report; September 2018–June 2021 14. Sponsoring Agency Code HRDO-20
15. Supplementary Notes The Federal Task Manager was Rachel James (HRDO-20; ORCID: 0000-0001-9138-510X).		

16. Abstract

Traffic engineers and researchers calibrate microsimulation models using macroscopic inputs—such as aggregated traffic throughput—instead of microscopic inputs, such as intervehicle spacing and acceleration. This has led to concerns that these models have been capturing the microscopic driver behaviors inaccurately, despite the macroscopic performance measures' apparent goodness of fit. Given the recent improvements to data collection and data processing technologies, particularly concerning drone or unmanned aerial vehicle technologies and cost reductions, there is renewed interest in trajectory-based calibration for microsimulation models. Researchers behind this project developed a new methodology for trajectory-based calibration, and they tested this methodology against traditional calibration at real-world urban freeway locations: I–270 in Maryland; I−15 in California; I−75 in Florida; and I−95 in Virginia. The results provided evidence that traditional calibration indeed cannot be trusted to produce realistic vehicle trajectories. Moreover, explicit integration of trajectories into the calibration process can remedy this. Calibrated model results were most impressive at I–75, which is the only site where trajectories were collected by a helicopter (instead of by drones), producing 1.2-mi-long trajectories. This report and accompanying software scripts provide instructions and lessons learned for collecting, cleaning, post-processing, correcting, and validating trajectory data. The report and scripts also provide instructions and lessons learned for trajectory-based calibration and validation. Future applications of the proposed methodology may involve studying the importance of car-following versus lane-changing, calibrating separate driver models for different congestion regimes, and calibrating the trajectories of connected and automated vehicles.

17. Key Words	18. Distribution Statement
Vehicle trajectory, NGSIM, microsimulation, calibration, traffic simulation	No restrictions. This document is available to the public through the National Technical Information Service, Springfield, VA 22161. http://www.ntis.gov

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	109	N/A

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

2D	two-dimensional
3D	three-dimensional
API	application programming interface
B.P.R.	Binning, Pairing, RMSE
CAV	connected and automated vehicle
DBF	directed brute force
DOT	Department of Transportation
FAA	Federal Aviation Administration
FHWA	Federal Highway Administration
GEH	Geoffrey E. Havers formula
GPS	global positioning system
Hz	hertz (a unit of frequency)
I.H.O.P.	Inputs, Heuristic, Outputs, Points
ID	identification
IDM	intelligent driver model
MDOT SHA	Maryland Department of Transportation State Highway Administration
NGSIM	Next Generation Simulation
NPMRDS	National Performance Management Research Data Set
PFS	pooled fund study
RMSE	root-mean-square error
SHA	State Highway Administration
TAS	traffic analysis and simulation
TSSM	Transportation Systems Simulation Manual
UCSD	University of California at San Diego
VDOT	Virginia Department of Transportation
VIRTUAL	Video-Based Intelligent Road Traffic Universal Analysis Tool
WT	wavelet transform

EXECUTIVE SUMMARY

Traffic simulation is a commonly used approach for analyzing transportation alternatives, designing traffic control strategies, predicting future congestion problems, and evaluating advanced vehicle technology impacts. Traffic analysts value microscopic traffic simulation, also known as microsimulation, for its ability to replicate driver behavior in great detail (e.g., car-following and lane-changing). Indeed, complex facilities (e.g., advanced signal timings, freeway interchanges) cannot be evaluated carefully without this microscopic level of detail.

Traffic simulation results are generally unreliable unless the model is calibrated for local conditions. Traffic engineers and analysts calibrate microsimulation models based on macroscopic inputs (e.g., aggregated travel speed, traffic throughput) instead of microscopic inputs (e.g., intervehicle spacing, acceleration rate, time gap) because of challenges associated with collecting trajectory-level data. This has led to concerns that the microscopic driver behaviors simulated by these models are inaccurate, despite the apparent macroscopic performance measures goodness of fit.

Given the recent improvements to data collection and data processing technologies, particularly with unmanned aerial vehicle technologies and cost reductions, there is renewed interest in trajectory-based calibration for microsimulation models. Researchers behind this project developed a new methodology for trajectory-based calibration, and they tested this methodology against traditional calibration at real-world urban freeway locations. The proposed calibration methodology is highly customizable and allows users to find the right balance between practicality and calibration robustness.

This project had three primary goals: collect and process data (both traditional traffic data and trajectory data), develop a trajectory-based calibration methodology, and test this methodology by calibration microsimulation models to demonstrate the value of calibrating models with trajectories.

First, this project collected both trajectory data and more traditional traffic data, such as speed and throughput from infrastructure-based radar data, at four, real-world congested freeway sites. Data were collected using a helicopter above I–75 in Florida, and other data were collected using drones above I–270 in Maryland, I–15 in California, and I–95 in Virginia during the spring of 2019. The research team considered the data collection process successful because of the volume of obtained trajectories and the small number of mechanical failures. During the remainder of 2019 the team post-processed the trajectory data. Data were cleaned, converted from videos to a specified numeric format, validated, and corrected. Although the team initially identified many errors in the data, these errors were systematically corrected, and the resulting datasets passed validation tests by wide margins. Lessons learned from the data collection and data processing phases, as well as instructions for identifying and fixing post-processed trajectory errors, are discussed in chapter 3 of this report.

Next, the trajectory data were used to inform the development of a new trajectory-based calibration methodology. This trajectory-based method is model and simulation-software agnostic and consists of seven steps. The first four steps—inputs, heuristic, outputs, points—

consist of choices made by the analyst, while the last three steps—binning, pairing, and root-mean-squared error (RMSE)—are iterative processes that can be automated through scripting. The scripts developed as part of this project are discussed in the appendices of this document and are available for download (Github, n.d.-b).

Motivated by initial research results, the team also explored the idea of a hybrid-calibration methodology, for which both traditional traffic data and vehicle trajectories are used in the calibration process. This methodology is discussed in chapter 4 of this report.

Finally, the team tested the trajectory-based, hybrid, and traditional calibration methodologies using two separate microsimulation software platforms. The test results from the experiments provided evidence that traditional calibration methods are unreliable in producing realistic vehicle trajectories. But the explicit integration of trajectories into the calibration process can remedy this shortcoming. The test results also demonstrated the importance of validation. The initial set of validation results revealed errors in the calibration process, but after the research team made some adjustments, the calibrated models passed their validation tests.

Trajectory-based calibration achieved the best results at the I–75 site in Florida. This is likely because it was the only site where full-length trajectories approximately 1.2 m in length were collected by helicopter. By contrast, the research team deployed drones at the other three sites, where only 800-ft-long trajectories could be collected because of deployment-height limitations. To compensate for this limitation, the research team deployed multiple drones at different key network locations to obtain multiple 800-ft trajectory snippets for the same site. Given the results, it is possible the 800-ft trajectories may be too short to support robust model calibration.

As mentioned above, this seven-step process did require external scripting to enable calibration with trajectories. Thus, there is room for improvement in the efficiency and user-friendliness of trajectory-based calibration. Given the promising results obtained by applying the developed calibration method, the authors hope vendors will develop improved data collection and data processing technologies, which can collect longer vehicle trajectories at reasonable prices. The authors also hope software developers will provide user-friendly apps to streamline the proposed method, facilitating widespread improvements in microsimulation model accuracy and robustness.

CHAPTER 1. INTRODUCTION

Traffic simulation is an indispensable tool for transportation professionals. It provides a cost-effective method for predicting the impact of various changes to the transportation system on traffic flow and performance, such as travel time, speed, and capacity. These variables include land use changes, geometric design alternatives, cooperative automated vehicle (CAV) penetration rates, active traffic management strategies, nonrecurring events, and growing traffic demands.

Microscopic simulation, or microsimulation, provides detailed representations of car-following and lane-changing behaviors for analyses that focus on urban freeway interchanges and corridors. When conducted properly, the outputs of microsimulation, which include a combination of lane-specific results, static graphics, moving vehicle animation, and statistical outputs, supply valuable information to decisionmakers. In comparison with macroscopic and mesoscopic simulation, microsimulation offers the most fine-grained and detailed understanding of congestion formation, propagation, and dissipation on freeways. To ensure analyses are conducted properly, analysts must perform proper model calibration, which is the process of estimating model parameters to represent local conditions more realistically.

Conventional calibration practices, which include calibrating to traditional aggregate performance measures, such as speed and volume, may not produce the robust outcomes and accurate models that engineers expect. In one example, a study of six microsimulation models that were well calibrated to aggregate measures produced wildly divergent predictions for future conditions (Bloomberg, Swenson, and Haldors 2003). There are several possible explanations for this phenomenon. First, although these models were developed with best practices in mind, the iterative changes made to the driver behavior parameters to best match current conditions' aggregate measures may have resulted in unintended impacts to models with different underlying assumptions, such as future demand, making those estimates of future conditions unreliable. Conversely, this may also suggest that the calibrated driver behavior models were overfit to current traffic conditions and were not generalizable for other conditions, such as future demand of the same modeled area. Regardless of the explanation, the results of Bloomberg, Swenson, and Haldors (2003) imply that practitioners may have been focusing on traditional measures at the expense of underlying driver behaviors and vehicle dynamics, possibly leading to less reliable predictions of traffic flow performance and driver behavior.

Moreover, the authors are unaware of research conducted to evaluate the accuracy of individual trajectories based on models calibrated using segment-level aggregate performance measures. Thus, it may be possible to demonstrate that simulated vehicle trajectories from microsimulation models calibrated by traditional methods are quite unrealistic, even for current conditions.

Despite these legitimate concerns and uncertainties, the modeling status quo has remained relatively unchanged because of challenges with collecting data. Rarely, if ever, do projects collect before and after data to allow modelers to understand and investigate how well their future condition models captured the "after" conditions. Additionally, capturing individual vehicle trajectories is expensive, computationally intensive, and creates challenges related to personally identifiable information.

Recent advancements in computing power and drone data collection capabilities have made the collection of full-length vehicle trajectories more appealing to State agencies and their consulting companies. Indeed, interest from State transportation agencies in using drones has rapidly increased in recent years (Banks et al. 2018). These advancements, together with longstanding concerns about calibrating microsimulation models via macroscopic measures, motivated this project. The project, titled *Trajectory Investigation for Enhanced Calibration of Microsimulation Models,* was sponsored by the Traffic Analysis and Simulation Pooled Fund Study (TAS PFS) and included the following objectives:

- Collecting and processing multiple sources of full-set, trajectory-level data at four congested freeway sites.
- Developing a methodology to calibrate driver behavior components (i.e., car-following and lane-changing) of traffic simulation software using the trajectory-level data.
- Applying this methodology to fully calibrate a traffic simulation model (e.g., driver behavior, demand, route choice).
- Validating the developed calibration procedure.
- Comparing the accuracy of models, as well as the level of effort required to calibrate those models, using the new methodology against more traditional methods.
- Demonstrating the new methodology using two microsimulation software tools.

The objective of this report is to describe all the procedural and substantive components of the trajectory investigation project. The research team sought to facilitate practical implementation of a calibration and validation methodology that exploits collected trajectory data. The team began the project by synthesizing relevant literature (chapter 2). The team then selected data collection sites and data collection mechanisms (chapter 3). In the next phase, the team collected data from drones, helicopters, and through traditional methods (chapter 3). After data collection, the team employed a data processing procedure to convert traffic video footage into numeric trajectory data and validate the trajectories (chapter 3). Next, the team used calibration and validation experiments for microsimulation models. The team developed a trajectory-based calibration procedure (chapter 4) and then used it to calibrate four microsimulation models representing congested freeways across four States (chapter 5). The team also conducted traditional calibration for the same four microsimulation networks (chapter 5). The report concludes with a discussion of additional research questions and suggested next steps to move this calibration procedure into practice.

CHAPTER 2. SYNTHESIS OF LITERATURE

This critical review of literature examines existing studies related to trajectory-level data collection, data cleaning and processing, commonly encountered errors within trajectory data, and applications of trajectory-level data in the calibration of traffic microscopic simulation models. It attempts to demonstrate the significance and relevance of each one to the project at hand. The synthesis also includes detailed summaries of the challenges of microscopic simulation calibration, such as overfitting, limited computer speeds, unreliable heuristics, as well as validation or reporting metrics efforts. This chapter concludes by discussing the implications of the literature on this research project.

TRAJECTORY-LEVEL DATA COLLECTION

The review of trajectory-level data collection methods provided information for the team to develop its data collection plan. It helped to inform the team both in terms of providing insight into the data types that this project would need to collect and in determining an appropriate spatial coverage of full-set trajectories.

Ossen and Hoogendoorn (2008) wrote that errors in collected data, such as vehicle trajectories, could affect outputs derived from a simulation tool as well as parameter values estimated from a calibration process. This implies a benefit to double-checking and verifying model inputs before the calibration of model parameters.

Daamen, Buisson, and Hoogendoorn (2014) classified the data used in traffic simulation studies into the following six groups:

- Local detector data.
- Section data (vehicle reidentification).
- Vehicle-based trajectory data.
- Video-based trajectory data.
- Behavior and driving simulation.
- Stated and revealed preferences.

The first and second—local detector data and section data—reveal primarily macroscopic properties of the system. Behavior and driving simulation as well as state and revealed preferences address individual behavior, typically at the microscopic level. Vehicle-based and video-based trajectory data can be used for both microscopic and macroscopic studies. But vehicle-based trajectory data are limited by coverage—vehicles must be equipped with tracking devices. Also, video-based trajectory data are limited by spatial coverage—the cameras can see only so far. Emerging technologies in the vehicle connectivity area, both in the vehicle-to-vehicle and vehicle-to-infrastructure communication, could address these limitations.

The research team drew multiple conclusions from this literature. The first was that the aforementioned data categories could inform the team's data collection plan. For example, to define the limitations of the proposed calibration method, the plan could specify data categories

to be collected. This was useful information during discussions with the data collection company because it helped to ensure the right types of data would be collected.

Second, the limited spatial coverage motivated the team to employ some data collection plan decisions to balance practicality and robustness of calibration. On the practical level, although practitioners and researchers may prefer as much spatial coverage as possible, the team recognized the calibration methodology should incorporate existing drone data collection technologies. It should also be fairly easy to use and thus more appealing to transportation agencies for their own projects. Also, the preference for robust data led the team to seek longer vehicle trajectories because they help to capture the true nature of driver behavior, including phenomena such as intradriver heterogeneity (Taylor et al. 2015). To balance practicality and robustness, the team decided to deploy a small number of drones to sample different stages of space-time congestion propagation. The team also adopted high-definition video collection technology from a helicopter.

TRAJECTORY DATA FORMATS

The research team identified four types of real-world trajectory datasets:

- Global positioning system (GPS) based.
- LiDAR based.
- Radar based.
- Video based.

The team employed video-based trajectory data formats given the interest in developing a trajectory-based calibration methodology to exploit drone data collection technologies. These formats included Next Generation Simulation (NGSIM), developed by the U.S. Department of Transportation Intelligent Transportation Systems Joint Program Office (ITS-JPO), and the Trajectory Clustering Dataset (University of California at San Diego 2014; Morris and Trivedi 2009).

The University of California at San Diego (UCSD) Trajectory Clustering Dataset provides trajectory data obtained by a simple visual tracker. The dataset format includes truck identification (ID), location, and time. NGSIM data provide the precise location of each vehicle within the study area every tenth of a second, resulting in detailed lane positions and locations relative to other vehicles. The NGSIM dataset format is listed in table 1.

Table 1. Next generation simulation data format (USDOT 2019).

Column Name	Description	Type
Vehicle_ID	Vehicle identification number (ascending by time of entry into section)	Number
Frame_ID	Frame identification number (ascending by start time)	Number
Total_Frames	Total number of frames in which the vehicle appears in this dataset	Number
Global_Time	Elapsed time in milliseconds since January 1, 1970	Number
Local_X	Lateral (*X*) coordinate of the front-center of the vehicle, in feet, with respect to the left-most edge of the section in the direction of travel	Number
Local_Y	Longitudinal (*Y*) coordinate of the front-center of the vehicle, in feet, with respect to the entry edge of the section in the direction of travel	Number
Global_X	*X* coordinate of the front-center of the vehicle in feet	Number
Global_Y	*Y* coordinate of the front-center of the vehicle in feet	Number
v_Length	Length of the vehicle in feet	Number
v_Width	Width of the vehicle in feet	Number
v_Class	Vehicle type: 1, motorcycle; 2, auto; 3, truck	Number
v_Vel	Instantaneous velocity of the vehicle in feet per second	Number
v_Acc	Instantaneous acceleration of the vehicle in feet per second squared	Number
Lane_ID	Current lane position of the vehicle. Lane 1 is the farthest-left lane; lane 5 is the farthest-right lane. Lane 6 is an auxiliary lane between an on-ramp and an off-ramp. Lane 7 is an on-ramp; lane 8 is an off-ramp	Number
O_Zone	Origin zones of the vehicles (i.e., the place where vehicles enter the tracking system)	Plain text
D_Zone	Destination zones of the vehicles (i.e., the place where vehicles exit the tracking system)	Plain text
Int_ID	Intersection in which the vehicle is traveling. Value of 0 means the vehicle was not in the immediate vicinity of an intersection	Plain text
Section_ID	Section in which the vehicle is traveling. Value of 0 means the vehicle was in the immediate vicinity of an intersection (Int_ID)	Plain text
Direction	Moving direction of the vehicle: 1, eastbound; 2, northbound; 3, westbound; 4, southbound	Plain text
Movement	Movement of the vehicle: 1, through; 2, left turn; 3, right turn	Plain text

Column Name	Description	Type
Preceding	Vehicle ID of the lead vehicle in the same lane. A value of 0 represents no preceding vehicle	Number
Following	Vehicle ID of the vehicle following the subject vehicle in the same lane. A value of 0 represents no following vehicle	Number
Space_Headway	Space Headway in feet. Spacing provides the distance between the front-center of a vehicle to the front-center of the preceding vehicle	Number
Time_Headway	Time Headway in s. Time Headway provides the time to travel from the front-center of a vehicle (at the speed of the vehicle) to the front-center of the preceding vehicle	Number
Location	Name of street or freeway	Plain text

DATA CLEANING AND PROCESSING

Daamen, Buisson, and Hoogendoorn (2014) suggested several methods of filtering, aggregating, and correcting data before calibration and validation (e.g., particle filters, Bayesian methods, inference methods). They also noted two types of errors (i.e., random and systematic) that might occur in raw trajectory data. The research team planned to mitigate these error risks by collecting radar data that could corroborate the accuracy of the trajectory data. The team also planned to perform manual spot-check calculations of speed, acceleration, and travel distance (figure 3 and figure 4). This check would confirm that the data preserved fundamentally valid kinematic relationships (Daamen, Buisson, and Hoogendoorn 2014).

To extract trajectory data for future use, Wei et al. (2005) developed a computer-based tool to extract trajectories from videos; but, to produce a trajectory, users were required to manually click on points traversed by the same vehicle. The tool lacked post-processing and data cleaning logic. Xu and Sun (2013) proposed another video-based, vehicle-trajectory processing approach. In the model, a model-based background subtraction algorithm extracted the vehicle trajectories. A numerical study showed that 97 percent of trajectories could be accurately detected. The model lacked a data cleaning module, however, which allowed for the possibility of speed and acceleration errors. Muthurajan, Amrutsamanvar, and Vanajakshi (2017) proposed a model to extract trajectory data by using a discriminative correlation filter. The object of the model was to map coordinates of a two-dimensional (2D) digital image onto three-dimensional (3D) real-world coordinates to reduce errors.

Many researchers have proposed algorithms to process and clean trajectory data. Michalopoulos (2008) proposed a post-processing algorithm to eliminate measurement and inconsistency errors from the extracted trajectories. First, Michalopoulos (2008) extracted trajectories by using Next Generation Vehicle Interaction and Detection Environment for Operations (NG-VIDEO). Then Michalopoulos (2008) applied a two-step optimization to clean the data. The upper-level optimization seeks to minimize the inconsistency errors by looking for the optimal number of polynomial pieces. The lower level seeks to minimize suitable measures of roughness subject to interpolation constraints. A case study in Minnesota (i.e., Michalopoulos 2008) showed that the

method generates more accurate and reliable trajectories than traditional approaches (e.g., locally weighted regression).

Montanino and Punzo (2013) proposed a multistep filtering procedure that removes outliers and cuts off residual random disturbances from the signal with low-pass or average moving filters. The driving dynamics are clear and unbiased. The procedure proposed here, however, does not guarantee the platoon consistency of the trajectories. Platoon consistency refers to the physical consistency of intervehicle spacing resulting from the individual trajectories of two following vehicles. This consistency is important because it can help to identify and fix errors in the data, such as a following vehicle overtaking a leader vehicle in the same lane (Punzo, Borzacchiello, and Ciuffo 2009).

To improve the platoon consistency, Montanino and Punzo (2015) proposed a "traffic-informed" methodology to restore physical and platoon integrity of trajectories in a finite, time-space domain. Montanino built a simulation-based validation framework to verify the efficacy of the reconstruction methodology. The procedure operates on positional data and requires four steps. The first step aims to remove extreme positional errors. In this way, the method further processes the resulting signal without allowing the results to be biased by the presence of such outliers. In the second step, the method smooths the random noise in the trajectory via a traditional low-pass digital filter. The first two steps preprocess the raw measurement signal; this is necessary to guarantee that the presence of extreme errors does not bias successive filtering steps. In the third step, the method exploits the information on vehicle kinematics and traffic dynamics to restore a physically consistent trajectory by performing a local reconstruction of trajectories. Through local reconstruction, the method substitutes the physically infeasible vehicle positional data with synthetic points that meet the following criteria:

- Are physically compatible (i.e., returning physical speeds and accelerations).
- Are consistent with the space traveled in the reconstruction window (i.e., internal consistent).
- Preserve physical intervehicle spacing (i.e., platoon consistent).

In the fourth step, the method removes residual noise through an additional application of a digital low-pass filter. The research applies the method on both aggregate and disaggregate data to obtain results, which are indirect confirmation that data filtering was necessary and that the proposed reconstruction methodology is effective.

Fard, Mohaymany, and Shahri (2017) proposed a simple, two-step technique based on wavelet analysis for filtering errors and reconstructing trajectory data. The main process identifies and modifies outliers using wavelet transform (WT) and eliminates noise by applying the wavelet-based filter. In the first step, the method identifies and modifies outliers using WT. After transforming the data with WT, the method modifies outliers through locally fitting an appropriate curve. The detected outliers are replaced by values resulting from applying Gaussian kernel-based weighted regression on the vehicle trajectory. This operation extensively improves the frequency response of the acceleration profile. In the second step, the method removes noise in the dataset. The main idea in this step is to classify wavelet coefficients with thresholding function at each level as soft or hard. While under hard conditions, the method preserves the coefficients with amplitudes higher than the selected threshold; under soft conditions, the

coefficients are decreased by the value of the selected threshold. The proposed method shows better performance in a case study with NGSIM data compared with the multistep method (Fard, Mohaymany, and Shahri 2017).

Many methods have been proposed to remove trajectory noise, including moving average algorithms (Duret, Buisson, and Chiabaut 2008; Hamdar and Mahmassani 2008; Thiemann, Treiber, and Kesting 2008); smoothing algorithms (Punzo, Formisano, and Torrieri 2005; Lu and Skabardonis 2007; Toledo, Koutsopoulos, and Ahmed 2007); and Kalman filtering (Ervin et al. 1991; Ma and Andréasson 2005; Punzo, Formisano, and Torrieri 2005). Ossen and Hoogendoorn (2008) investigated the effect of measurement errors on calibration results. These researchers concluded that smoothing the data by the moving average method could be the best way to alleviate these issues.

Marczak and Buisson (2012) proposed an I-spline method to reduce noises and smooth trajectories. In the model, instead of using one single polynomial for the whole trajectory, it filtered the positions by dividing the total time interval into smaller intervals and used lower degree polynomials in each of these subintervals. This research suggests that a basis of nonnegative and monotone splines, termed I-splines, be used for good continuity characteristics. Once the smoothed trajectories are calculated, they can be represented as the linear combination of a set of the I-splines basis. Smoothed positions are differentiated to calculate smoothed velocities and accelerations. The results showed that the method can reduce the spikes in the velocity distribution and percentage of jerk values and increase the acceleration variability of smoothed trajectories.

Pal and Chunchu (2018) proposed another new smoothing method based on complete ensemble empirical mode decomposition with adaptive noise. The smoothed trajectory data were further differentiated using WT to estimate the instantaneous speed of the vehicle. The WT technique led to more accurate speeds. Internal consistency analysis of the position and speed also supported the suitability of the proposed method for speed correction. Results showed that trajectory data corrections have a significant effect on the flow-occupancy relationship, specifically at higher flow levels. The researchers did not apply this method to real-world data to test the efficiency of the model, however.

In conclusion, Fard, Mohaymany, and Shahri (2017) give the best model to remove noise and maintain platoon consistency in the dataset. Because of the unknown values of actual vehicle trajectories, however, it is unfeasible to determine directly whether the reconstructed data are near to the actual ones. Therefore, the research team cleaned data according to the proposed methods in Fard, Mohaymany, and Shahri (2017) and Pal and Chunchu (2018).

COMMON ERRORS WITHIN VIDEO-BASED TRAJECTORY DATA

Some errors may still exist in the data even after a data cleaning process. To inspect trajectory accuracy, Punzo, Borzacchiello, and Ciuffo (2011) proposed a method to analyze trajectory data. The first step is to examine the jerk values from trajectory data to check acceleration feasibility. Then, platoons and internal data consistency are verified. The last step is to examine spectral frequencies on speed, acceleration, and jerk. The research team applied the model to NGSIM

data. Results showed the method is useful for analyzing trajectory data to confirm that the data are physically valid before carrying out a study using those data.

Challenges with the identification of vehicles in each frame of raw-trajectory videos cause most of the errors in video-based trajectory datasets. A common method to locate vehicles is to extract backgrounds to find contours of vehicles, but contour errors will cause vehicle position errors. Inexact vehicle contours will cause inaccurate speed and acceleration readings. The research team desired improved contour extraction algorithms in the trajectory extraction process to reduce these errors. The research team proposed a video-based trajectory post-processing tool called Video-Based Intelligent Road Traffic Universal Analysis Tool (VIRTUAL®) (Zhao and Li 2019), which contains such algorithms.

According to Coifman and Li (2017), video image processing brings challenges in detecting projection errors, occlusions, shadows, non-rectilinear shapes of real vehicles, and vehicles with colors similar to the pavement. The 2D image plane seen by the camera is projected into the 2D ground plane of the roadway, with the implicit assumption that all objects are in the ground plane. But 3D vehicles violate the ground plane assumption and thus generate projection errors. These errors are like shadows; the higher off the ground a feature is, the farther away it projects from its true ground coordinates. Projection errors also increase as the vehicle moves farther away from the camera. Thus, the projection of the top-front of the vehicle should seemingly move faster over the ground than the actual vehicle.

Also, if the video is of low resolution, the projection will be cast onto the ground plane, leading to further distortion. The projection expands the small number of pixels occupied by the distant vehicles in the raw video, making them comparable in size to the same vehicles in the near field; this gives false confidence and masks the discretization errors of the original downsampling. It will cause small, instantaneous positioning errors because of the relatively low resolution of the video. Errors will be amplified when taking the position difference in two successive frames and will degrade the quality of conventional speed calculations if neglected. Further, it will cause accelerations to exhibit unrealistically large magnitudes. Using high-resolution videos to extract trajectories may reduce these errors.

Montanino and Punzo (2015) mentioned that motorcycles often “split lanes,” or drove between the standard lanes of travel. The legality of motorcycle lane splitting varies between States. When extracting vehicle trajectories from video, the model assigns all vehicles to a discrete lane. Thus, motorcycles that overtake other vehicles by splitting the lane appear to coexist with other vehicles in a given lane and longitudinal location.

APPLICATIONS OF TRAJECTORY-LEVEL DATA IN THE CALIBRATION OF MICROSCOPIC SIMULATION MODELS

Vehicle trajectory data, which potentially reveal the exact positions of all vehicles at all times, provide an excellent degree of specificity with which to calibrate driver behaviors. Research results have shown that trajectory data can be quite effective for calibrating both car-following behavior (e.g., Hamdar and Mahmassani 2009) and lane-changing behavior (e.g., Talebpour Talebpour, Mahmassani, and Hamdar 2015). Trajectory data are increasingly available to State

transportation agencies through the use of drones and possibly through probe data providers (Banks et al. 2018).

Kim et al. (2013) proposed a calibration framework in which exogenous and endogenous sources of travel time variation could be examined. Under this framework, numerous simulation scenarios involving different combinations of traffic demand, weather, incidents, and special events were generated through random sampling techniques (e.g., Monte Carlo sampling). Then, representative travel time distributions could be obtained by post-processing trajectory data from the various scenarios. Kim and Mahmassani (2011) also studied the effect of ignoring correlations among parameters when calibrating a car-following model. Results of the models that preserved the correlation effect were closer to the field data.

Researchers have noted the strong correlation between driver behavior and traffic congestion (Geng et al. 2016; Ye and Zhang 2009)[1]. This implies that fundamental relationships between driver behavior and congestion regimes (e.g., below, near, at, or over capacity) could be developed under existing conditions and then applied to future models on a link-specific basis (i.e., based on the new congestion regime for each link under new traffic demands or traffic control strategies).

In the literature, many researchers formulated calibration approaches as optimization problems consisting of two fundamental components: a searching algorithm, and an objective function. Generally, if the objective function is analytically or numerically differentiable and unimodal, then the optimal solution can be obtained with several deterministic approaches, such as Newton's method, the Gauss-Newton algorithm, gradient descent, and the Levenberg-Marquardt algorithm (Treiber and Kesting 2013b). But many calibration/validation problems have no differentiable or unimodal objectives. In such cases, one may consider a heuristic or metaheuristic approach as a solution method. Note that these approaches do not necessarily obtain the global minima. Thus, many analysts choose to apply the heuristic multiple times (i.e., multistart), where each time the heuristic starts from a different set of initial conditions.

Downhill simplex is commonly applied in microscopic calibration and validation problems (Brockfeld, Kühne, and Wagner 2004; Kim and Mahmassani 2011). This method can be applied to problems for which the derivatives are unknown. But it requires the objective function to be unimodal. Indeed, many of the calibration/validation problems—particularly those with stop-and-go traffic—have no unique minimum or are multimodal (Treiber and Kesting 2013b). Researchers may apply stochastic solution methods to avoid getting stuck in local minima. Treiber and Kesting and Treiber (2013a) further found optimization-based estimation to be more effective than maximum likelihood estimation for calibrating car-following models.

A popular stochastic metaheuristic method applied by researchers in calibration/validation problems is the genetic algorithm (Kesting and Treiber 2008, 2009; Ranjitkar, Nakatsuji, and Asano 2004; Ranjitkar, Nakatsuji, and Kawamura 2005; Vasconcelos et al. 2014; James 2019). Inspired by the natural selection process in evolutionary biology, the algorithm typically includes three operators: mutation, crossover, and selection. A model with a set of parameters represents an "individual," and a "population" refers to several individuals. The algorithm begins with a

[1] Mahmassani et al. forthcoming.

randomly generated population of individuals, and then it calculates the objective or fitness function for all individuals of the population. Based on these values, a portion of the population is selected in pairs of individuals, called "parents," to breed a new generation by crossover and mutation operators. The method repeats this population generation until it reaches a termination condition.

Researchers have implemented several other nonlinear optimization solution methods to solve microscopic calibration/validation problems. Some of these methods include sequential quadratic programming (Wang et al. 2010), cross-entropy (Zhong et al. 2016), artificial neural networks (Colombaroni and Fusco 2014), and a dynamic time warping algorithm (Taylor, Zhou, and Rouphail 2012; Taylor et al. 2015).

Punzo, Formisano, and Torrieri (2005) provided a robust method to simplify car-following models, namely, to reduce the number of calibration parameters without sensibly affecting the capability of reproducing reality. First, calibration input parameter values are drawn by quasi-random sampling, and the traffic model is executed using those values. Then the agreement between observed and simulated trajectories is calculated in terms of the root-mean-square error (RMSE) of the instantaneous speeds or spacings. The method iterates this process until the number of evaluations is sufficient for the calculated indices to be stable. The method was applied to the intelligent driver model (IDM) using reconstructed NGSIM dataset. The analysis unveiled that the leader's trajectory is considerably more important than the parameters in affecting the variability of model performances. Sensitivity analysis also returned the importance ranking of the IDM parameters. As a result, this paper proposes a simplified model version with three (out of six) parameters. After calibrations, the full model and the simplified model show comparable performances; however, the simplified model converges to a solution more quickly.

In other objective function research, Punzo, Formisano, and Torrieri (2005) and Treiber and Kesting (2013a) concluded that following distance is a more robust measure of performance than vehicle speed because speed errors do not adequately propagate through the trajectory. Yu and Fan (2017) cited five goodness-of-fit formulas previously used by researchers to calibrate simulation models. James (2019) similarly cited four formulas previously used to calibrate simulation models, only one of which—the Geoffrey E. Havers (GEH) formula—was listed on Yu's list. Ciuffo, Punzo, and Montanino (2012) found that RMSE, also on James' list, worked best for calibrating the Gipps' car-following model. Montanino, Ciuffo, and Punzo (2012) and Punzo, Formisano, and Torrieri (2005) also selected RMSE as their goodness-of-fit formula.

The team also selected RMSE as the goodness-of-fit measure for use on this project for a few reasons. First, although the team knew that older informational documents, such as *Traffic Analysis Toolbox Volume III*, and out-of-print State departments of transportation simulation guidelines demonstrate and suggest the use of GEH, the newer "Methods and tools for supporting the Use, caLibration and validaTIon of Traffic simUlations moDEls" (MULTITUDE) report specifically discouraged using GEH to calibrate traffic simulations (Antoniou et al. 2014). Second, RMSE is an intuitive calculation that penalizes outliers in the simulated results.

For example, consider the following hypothetical, predicted travel speeds from two microsimulation models. An analyst executes each of the two models five times with five different random number seeds and examines the predicted average vehicle speeds on a key

freeway segment (table 2). The field-measured average vehicle speed is 55 mph for the same freeway segment.

Table 2. Comparison of average segment speeds from five replications of microsimulation.

Model Number	Run 1 Average Speed (mph)	Run 2 Average Speed (mph)	Run 3 Average Speed (mph)	Run 4 Average Speed (mph)	Run 5 Average Speed (mph)
Model 1	56	56	56	56	56
Model 2	55	55	55	55	60

In this example, if a simple percentage were used to quantify goodness of fit, both models would be viewed as equally good. In both cases their margin of error is 1 mph, on average. According to RMSE, however, Model 1 would have the lower RMSE and be recommended as the better calibrated model because its results are more consistently reasonable.

SUMMARY

Table 3 illustrates the primary takeaways from the literature for the research team.

Table 3. Primary literature review outcomes.

Category	Source	Finding	Impact
Data Collection	Daamen, Buisson, and Hoogendoorn (2014)	The data used in traffic simulation studies can be divided into six groups.	The team had awareness to ensure the right types of data would be collected.
Data Collection	Daamen, Buisson, and Hoogendoorn (2014)	The main limitation of video-based trajectory data is limited spatial coverage.	The team planned to deploy a small number of drones with overlapping spatial coverage to emulate longer trajectories.
Data Formats	UCSD (2014), USDOT (2019)	These were the only video-based trajectory data formats located.	The team used similar concepts when developing its own format for this project.
Data Cleaning	Daamen, Buisson, and Hoogendoorn (2014)	Daamen recommended specific data cleaning methods to avoid specific data errors.	The team planned to collect corroborative radar data and check kinematic relationships.
Data Cleaning	Fard, Mohaymany, and Shahri (2017), Pal and Chunchu (2018)	Numerous methods have been proposed to reduce noise and errors within trajectory data.	The team identified Fard and Pal as the most effective for this study.

Category	Source	Finding	Impact
Data Errors	Punzo, Borzacchiello, and Ciuffo (2011), Coifman and Li (2017)	Researchers noted errors related to projection, and kinematic validity.	The team implemented error prevention algorithms into their postprocessing tool.
Calibration	Treiber and Kesting (2013a)	Optimization-based estimation was effective for calibrating car-following models.	The team planned to develop an optimization-based approach.
Calibration	(Numerous papers)	Many heuristic methods have been successfully used to calibrate simulation models.	The team planned to develop a flexible approach that could employ any heuristic method.
Calibration	Punzo, Formisano, and Torrieri (2005), Treiber and Kesting (2013a)	Following distance is a more robust measure of performance than vehicle speed.	The team planned to incorporate headways within the overall calibration method.
Calibration	Ciuffo, Punzo, and Montanino (2012), Punzo, Formisano, and Torrieri (2005), James (2019), Montanino, Ciuffo, and Punzo (2012)	RMSE has been an effective goodness-of-fit measure for objective functions.	The team planned to incorporate RMSE within the overall calibration method.

UCSD = University of California at San Diego; USDOT = U.S. Department of Transportation.

CHAPTER 3. DATA COLLECTION AND PROCESSING

To perform the trajectory-based calibration research, the research team needed real-world data from four sites. This chapter describes the four-stage process of selecting the sites, selecting a data collection method, collecting the data, and processing the data.

SITE SELECTION AND TOOL SELECTION

This section provides a short list of candidate sites experiencing recurring congestion for further study. The research team prioritized sites having calibrated and documented microscopic models available, in addition to readily available traditional data sources (e.g., travel times, throughput counts, traffic density).

Site Selection Considerations

The research team decided to hire an external team to conduct the data collection rather than collect the data in house. The team also considered the following factors associated with the external data collection company during site selection:

- Proximity to the data collection company's offices—relocating the company's inventory and personnel would add significant costs in shipping alone. As such, the Washington, DC, metropolitan area, the Carolinas, and Florida were considered the most accessible.

- Airspace restrictions—once corridors were selected, the team prioritized unrestricted airspace for consideration. Drone data collection companies have tools at their disposal to investigate airspace restrictions. Changing the Federal Aviation Administration (FAA) license to an FAA part 107 certification revealed areas where drones cannot fly legally and which airspace requires low-altitude authorization and notification capability. Even around airports where drone surveys could be approved, drones might be permitted to fly at altitudes of only approximately 100 ft. This would limit the detection range for full-set vehicle trajectories.

- Corridor geometry—the data collection plan prioritized straight corridors because of less occlusion from trees and buildings.

- Deployment options—the data collection company desired deployment options, which could include frontage roads and fields clear of trees and power lines.

Tools Selection Consideration

The project scope required the team to apply two microsimulation tools to demonstrate the newly developed calibration procedure, which needed to be software agnostic. The team used a series of factors to inform the selection of microsimulation software packages, including license fees, perceived ease of use, availability of simulation datasets and networks, team experience with each tool, and feedback from pooled fund study (PFS) members.

Stakeholder Input

Project stakeholders were continually updated and consulted at bimonthly intervals, which allowed them to influence key project decisions. For this project, stakeholders were divided into two groups: PFS partners and subject matter experts. At this stage of the project, it was the PFS partners who provided some key opinions to influence the final decisions. The stakeholders ultimately expressed their opinion on two of the team's decisions, although they were invited to comment on other decisions as well.

The first decision influenced by stakeholders involved the choice of microsimulation platforms. The stakeholders expressed clear interest in three candidate microsimulation tools. However, the project was scoped and budgeted for only two microsimulation tools.

The second decision influenced by stakeholders involved the spatio-temporal analysis limits. The research team had originally proposed to obtain analysis limits with GPS mapping apps, which graphically display historical congestion patterns. Stakeholders were skeptical about the precision of congestion endpoints reported by these apps and suggested independent probe data analysis as a supplement. The team thus decided to analyze National Performance Management Research Data Set (NPMRDS) data (FHWA 2019) with GPS app data to corroborate the results and to ensure that the chosen analysis limits would fully capture the congestion.

Final Selections

After receiving these comments, the research team considered three microsimulation tools across seven sites. A significant obstacle perceived was the need to develop data-mapping tools to compare simulated trajectories against field-measured trajectories. The research team confirmed such tools were already available for two common microsimulation tools. This led to the selection of PTV Vissim® and Aimsun® to conduct the case studies. Moreover, the team selected I–270 in Maryland, I–15 in California, I–75 in Florida, and I–95 in Virginia as the data collection sites.

DATA COLLECTION

This section describes the selection of a primary mechanism for collecting and processing full-set trajectory level data at all sites during the various stages of congestion (e.g., bottleneck formation, full breakdown, bottleneck dissipation). It also describes the corroborative data collection mechanisms performed at all four sites.

Generalized Plan

The general data collection plan uses videos as the primary data source, probe vehicles with high-resolution GPS devices as corroborative data, and infrastructure-based radar to obtain traditional data (e.g., throughput and speed). Both drones and helicopters were used to collect the aerial videos. The drones can capture approximately 800 ft of continuous trajectories, while the helicopter can capture 6,336 ft (i.e., 1.2 mi) of continuous trajectories. The differences in data collection ability stem from the heights at which the drones and helicopters can be operated. For example, the flight altitude of the helicopter is significantly higher than the drone, enabling it to capture longer trajectories.

If project resources were unconstrained, the research team would have preferred to collect data using helicopters at all four sites. Unfortunately, project resources limited the team to helicopter data collection at only one site, as the helicopter data collection was significantly more expensive than the drone data collection. The research team ultimately decided to use helicopter data collection at the I–75 site and drone data collection at the other three sites. The team decided to use helicopters to collect aerial data at the I–75 site because it presented a unique opportunity. That location experienced less traffic congestion, with no more than two miles of queueing. A single helicopter could fully capture that queue both spatially and temporally. The team could deploy the helicopter to capture these flow dynamics in the greatest detail while applying the more economical drone approach at the other sites. The team procured the services of a second data collection company to manage the helicopter survey.

The research team developed a list of launch parameters for the data collection companies (table 5). The team developed congestion maps from NPMRDS data and used GPS mapping apps to corroborate the spatial and temporal limits of the bottlenecks. These parameters served as a starting point; during project execution, the data collection company reserved the right to tune some of the more detailed settings, such as coverage duration and locations, and lengths and device configurations. Specifications were also provided for collecting the traditional corroborative data (e.g., GPS locations for each radar collection device, on-ramp and off-ramp locations for floating car studies).

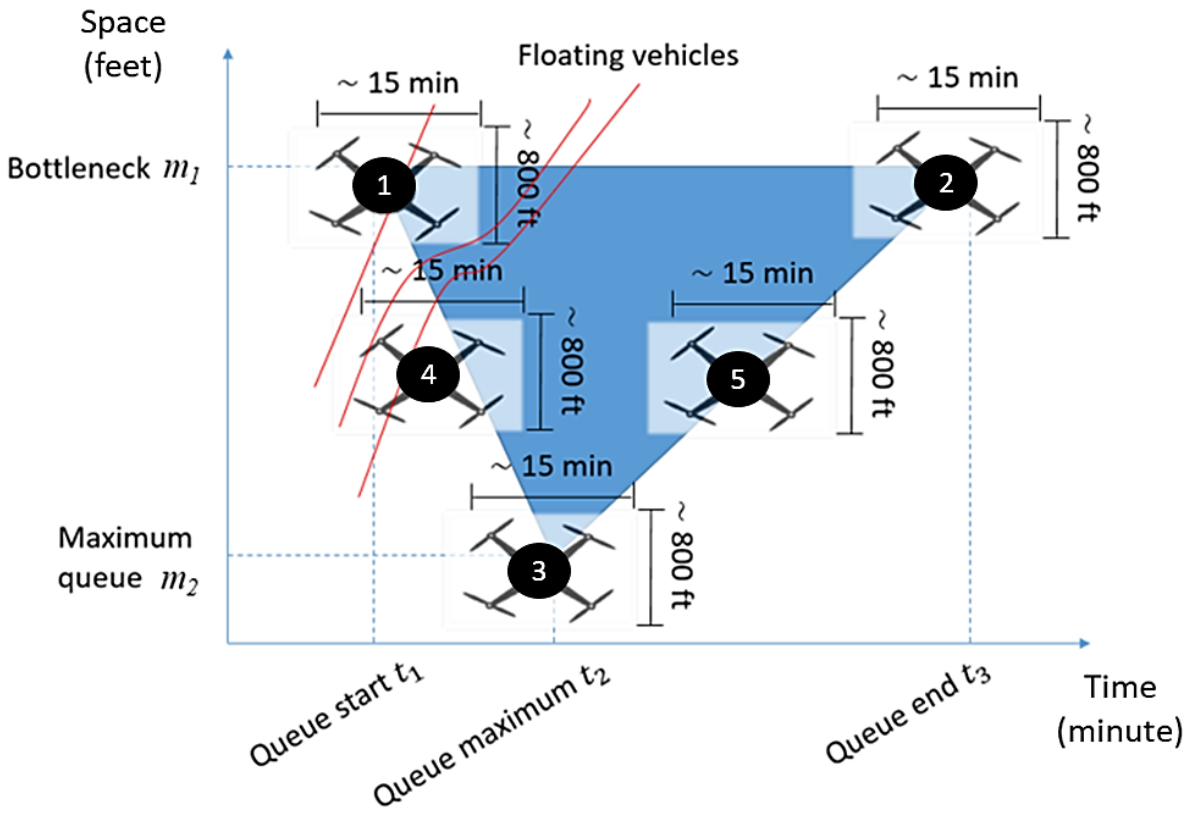
There are several data processing companies in the marketplace. The research team decided to use the university-maintained VIRTUAL tool to extract trajectory data from the videos (Zhao and Li 2019). Technical details of the video data collection (e.g., drones or helicopters, camera specs, coverage time periods, and locations) are based on site characteristics including geometries, flight restrictions, and congestion patterns (e.g., maximum queue length). The team sought to capture different phases of congestion (e.g., formation, development, and dissipation) and different queue locations (e.g., at the bottleneck where the queue starts, in the middle of the queue, and at the end of the queue where free-flow traffic transitions to queued traffic).

This project required the collection of corroborative data to validate numerical trajectories extracted from the aerially collected videos. The research team decided to use GPS-equipped probe vehicles to collect corroborative trajectory data synchronized with the video data. These probe vehicle data can cover longer distance ranges (e.g., possibly going through the entire queue). The team deployed probe vehicles simultaneously with the video data collection equipment such that the probe vehicle trajectories would overlap with the aerially collected trajectories.

In addition to collecting probe vehicle data to validate the trajectories, the research team used radar to collect count and speed data. These traditional datasets were necessary for the team to calibrate microsimulation models using standard, state-of-practice methods. This allows the team to compare model calibration results using proposed trajectory-based methods against traditional fixed detector data-based methods.

The team used traffic congestion patterns to determine where, when, and how to collect data. The main goal was to capture trajectories around and upstream of the bottleneck to learn how queues formed and propagated.

Figure 1 is a time-space queuing diagram that illustrates the relevant information and the type of data collected at typical freeway bottlenecks.



Source: FHWA.

Figure 1. Illustration. Drone data collection considerations and parameters.

The critical time-space points include bottleneck location m_1, maximum queue location m_2; and queue starting, queue maximization, and queue ending time t_1, t_2, t_3. Daily variations of critical space and time points m_1, m_2, t_1, t_2, and t_3 were considered to make sure the drone deployments did not miss the queuing process of interest to this project.

The research team identified the critical time-space points using both the NPMRDS data and corroborative GPS mapping apps. From the apps, the team identified traffic congestion states for any segment and time within a typical day. From probe data analytics sites, the team generated heat maps either directly on the sites or offline after downloading the data. Analysts may use heat maps to visually assess the congestion and determine critical analysis limits. They could alternatively use downloaded data to obtain the triangular congestion map around a bottleneck, or the starting and ending times of a bottleneck (if the ending state is unclear, or blends with another bottleneck) (figure 1). If only point measurements are available, an analyst could use interpolation methods to construct the triangular congestion map (Treiber, Kesting, and Wilson 2011).

The research team viewed NPMRDS heat maps within the Regional Integrated Transportation Information System using a cutoff speed of 45 mph to illustrate congested locations. From these

heat maps, the team visually identified the critical space and time points shown in figure 1 for each of the four study sites. The resulting analysis limits were a bit more conservative (i.e., more time-space coverage) than those obtained from the GPS mapping apps.

Agencies interested in collecting aerial data can determine drone coverage priorities with this time-space queuing diagram based on their available resources and budgets. For example, drones deployed around the bottleneck location m_1 around starting time t_1 and ending time t_3 (e.g., drones 1 and 2) should be prioritized for understanding queue formation and propagation. Next, agencies may prioritize drone deployments at the end of the queue, around location m_2 and time t_2, to capture queue dissipation (e.g., drone 3). If resources allow, agencies may consider deploying drones to collect data in the middle of the queue, with time coverage preferably around the start and end of the queue (e.g., drones 4 and 5). If the agency desires to collect corroborative data sources, such as probe vehicles, the data collection can be synchronized. For example, probe vehicles can circulate on the same segment of roadway while the drones collect video information before, during, and after the congestion (e.g., as the curves show for drone 1's coverage range).

Example Detailed Plan

This section provides some of the detailed data collection plans and specifications for one of the four sites (i.e., I–95 in Virginia). The other three sites were analyzed through a similar process.

Data Analysis

The research team identified two bottlenecks in the study area. According to the NPMRDS data and GPS apps, there was no significant difference in typical traffic between Tuesday and Thursday. For bottleneck 1, the worst congestion happens at 9:50 a.m. (location 156.0), with a maximum queue length of 4.5 mi. The maximum queue lasts until 9:55 a.m. Regarding bottleneck 2, the worst congestion happens at 7:45 a.m. (location 161.9), with a maximum queue length of 2.9 mi. The maximum queue lasts until 8:10 a.m. Bottleneck 1 was prioritized over bottleneck 2 because the bottleneck 2 back-of-queue reaches an interchange, which may complicate the extraction of accurate trajectories. The critical queuing pattern parameters are summarized in table 4.

Table 4. Queueing pattern for the example data collection plan.

Bottleneck	Queue start time t_1	Bottleneck location $m1$ (mile marker)	Maximum queue length time t_2	Back-of-bottleneck location $m2$ (mile marker)	Maximum queue length (m_1–m_2)	Maximum queue end time t_3
Bottleneck 1 Wednesday	5:35 a.m.	Mile 160.5	9:50 a.m.	Mile 156.0	4.5 mi	9:55 a.m.
Bottleneck 1 Thursday	5:30 a.m.	Mile 160.5	9:55 a.m.	Mile 156.0	4.5 mi	10:10 a.m.
Bottleneck 2 Thursday	6:15 a.m.	Mile 164.8	7:45 a.m.	Mile 161.9	2.9 mi	8:10 a.m.

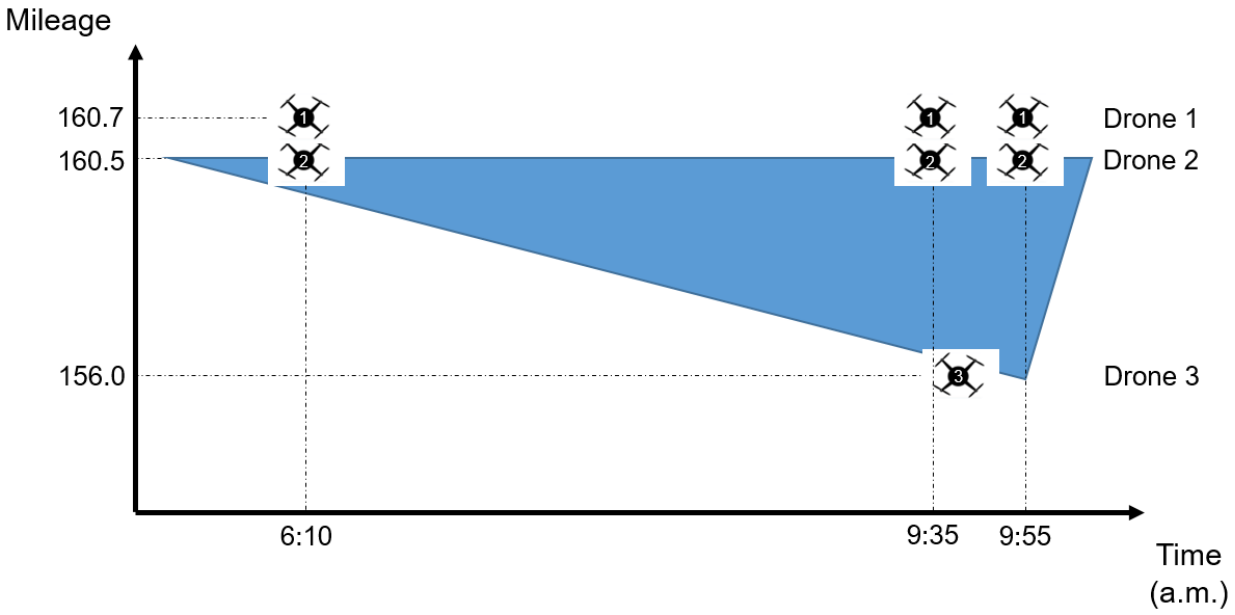
Final Plan

For each bottleneck, the team planned to deploy drones around the starting, ending, and maximum space-time patches of the queue. These parameters are specified in table 4. Bottleneck 1 had relatively simple geometry at both the bottleneck and maximum queue locations, which facilitated trajectory data analysis and simulation calibration.

The data collection plan parameters are detailed in table 5. It was impossible to extract trajectories from videos taken during darkness, so it was futile to collect data before sunrise. Thus, although congestion began around 5:30 a.m., the plan did not call for deploying drones until 6:10 a.m., because sunrise in Virginia is around 6:00 a.m. in May. To capture more trajectories at the beginning of the bottleneck, drone 1 was deployed at mile marker 160.7, which was slightly downstream of the beginning point of the bottleneck (160.5). From this analysis, the team developed detailed deployment parameters for the data collection company, as shown in table 5. Figure 2 illustrates the corresponding drone coverage.

Table 5. Parameter values for the example detailed plan.

Parameter	Drone 1	Drone 2	Drone 3
Location	38°40'12.1"N 77°15'15.6"W (Location 160.7)	38°40'10.5"N 77°15'27.0"W (Location 160.5)	38°37'06.2"N 77°17'51.0"W (Location 156.0)
Time	Wednesday, Thursday 6:10–6:30 a.m., 9:35–9:55 a.m., 9:55–10:15 a.m.	Wednesday, Thursday 6:10–6:30 a.m., 9:35–9:55 a.m., 9:55–10:15 a.m.	Wednesday, Thursday 9:40–10:00 a.m.
Coverage	400 ft, covering 850 ft	400 ft, covering 850 ft	400 ft, covering 850 ft
Camera	Pitch angle: −49.1, bottom of camera perpendicular with the road.	Pitch angle: −49.1, bottom of camera perpendicular with the road.	Pitch angle: −49.1, bottom of camera perpendicular with the road.



Source: FHWA.

Figure 2. Illustration. Drone coverage for the example data collection plan.

Outcomes and Lessons Learned

The team successfully completed data collection for this project in accordance with the aforementioned plan during the summer of 2019. The third-party data collection vendors viewed the data collection effort as successful because of the minimal number of device failures and the good weather on the scheduled days of data collection. The data collection vendors delivered nearly 3 TB of helicopter video footage data and approximately 75 GB of collected drone data. The traditional corroborative data were much smaller in size compared with the drone video footage data.

Both data collection companies documented their lessons learned for future reference. The drone data collection company provided the following insights.

First, pilot activities may benefit from clear communication before collection and during recording periods. Attempting to sync recording start times among multiple drone pilots along a corridor is a demanding task. It is helpful for each pilot to receive clear instructions for each operation from takeoff, through recording, and during landing. As each task is accompanied by some reduction in the battery life of the drone, pilots flying out of sync will reduce the amount of usable trajectory data. Furthermore, collection segments may be provided in the form of precise landmarks the pilots are to observe at the top and bottom of their respective frames. Sample snapshots of the observation area may be collected ahead of deployment for verification with the project manager.

Next, traffic flow observations may take place ahead of collection. When reviewing drone footage, some researchers noted that segments might fail to display the anticipated levels of queuing. As such, it may help to verify bottleneck propagation (including start and end of queues) through casual aerial observation, from the start to the end of the morning or the

afternoon peak periods. Finally, consider marking GPS-probe vehicles with strobe lights or other identifiers so they are easily observable from aerial footage.

The helicopter data collection company neglected to account for perspective distortion with varying focal lengths of the lenses when looking straight down. Additionally, as the focal lengths were different and the cameras were on a constant plane, it was challenging to keep a curved section of the highway properly framed across all three cameras to ensure adequate data capture. Wider focal length lenses should be used when focusing on roads with curves.

Two other insights were related to the weather. As summer weather in Florida varies considerably in the afternoon, it was nearly impossible to accurately predict the cloud ceilings and precise thunderstorm locations more than 24 h in advance. In the future, a larger time window will help alleviate weather concerns. This will increase costs, however, as the aircraft may have to wait on the ground for multiple days. Second, wind at the proper altitude made it challenging to keep the aircraft over a precise latitude-longitude location. Pilots can manually operate filming and production aircraft to hover over the ground if they are not equipped with autopilot. At altitudes of 4,000 ft or more, it is difficult to visually see aircraft drift. Multiple GPS receivers and displays minimize that issue, but human factors are still involved. The data collection vendor and agencies can explore potential solutions or define an acceptable tolerance range of location variability.

DATA PROCESSING

This section describes how the aerial videos were processed to obtain numerical vehicle trajectories. A transportation agency may lack the expertise necessary to process video data into numerical trajectories themselves. They may decide to have this process completed by a third-party data processing company or university. It may be helpful, however, for the transportation agency and its simulation analysts to understand the data processing procedure at a high level, especially regarding identifying errors in the processed trajectory data. This section provides a high-level overview of the process used to extract numerical trajectories from video and how errors were identified and corrected in the processed data.

The first step of data processing is to determine GPS coordinates for the roadway to be modeled. The second step is to apply an automated process that will generate the numeric trajectory data. The automated process will generate data that correspond with the GPS roadway coordinates. The third step is to check the data for errors and to fix any errors identified. Step three is unique to that process used to complete steps one and two, and it will vary across data collection sites and data extraction solutions. This section, however, describes some errors believed to be common, which were found and fixed by the research team during their experiments.

Creation of Numeric Trajectory Data

The research team developed scripts with programming code to determine GPS coordinates for the roadway to be modeled. These scripts will be made available to practitioners and researchers through various avenues (GitHub, n.d.-b). The research team used the university-maintained VIRTUAL tool as their automated process to generate numeric trajectory data. This tool contains several built-in data cleaning and validation functions. However, the process might include

specific knowledge and terminology in the video processing field that are unfamiliar to transportation stakeholders. Thus, to help transportation stakeholders communicate with video processing third parties, this section briefly summarizes the steps within VIRTUAL at a high level.

General Terminology

A video is composed of a series of consecutive frames. Each frame is a picture that consists of a matrix of pixels shot at the corresponding specific time. Each frame also contains a global time stamp. The foreground is the vehicles located in the area of interest (e.g., the segment of the road where vehicle trajectories can be extracted). The rest of the picture is the background that can be removed from the analysis. Certain background feature information that helps determine vehicle location and lane number can be preserved. Features are certain objects of particular shapes (e.g., rectangle, circle) on the frame that are easily detected by general video processing methods.

Background Normalization

The original background of each frame may vary because of vibrations and rotations of the video shooting devices (e.g., drones or helicopters). To match different frames, VIRTUAL identified certain distinctive features (e.g., lane markings, road lamp bases) that appear in consecutive frames. With these features, VIRTUAL determines real-world physical coordinates for each pixel in each frame, and thus it normalizes the area of interest in each frame to a static area in the physical coordinate system. With the GPS locations of these features (e.g., obtained with GPS mapping apps), the GPS coordinates of each pixel in the area of interest can be calculated. A Matlab® script with detailed explanations of the GPS coordinate conversion is available for reference.

Deep Learning Training

With the normalized background in each frame, the VIRTUAL team trained a deep-learning neural network to track the vehicles in each video frame by cropping numerous individual vehicle images from a set of representative frames. The team tried to crop at least one image for each vehicle that appeared in the I–75 video dataset. They experimented using different numbers of cropped vehicle images for training and found that the trained neural network performance improved as the number of covered vehicles increased. They fed cropped vehicle images into the standard training process to calibrate the neural network to identify vehicles effectively in the foreground for each study video. More details on this process are available through a repository on GitHub (Redmon and Farhadi 2018; GitHub, n.d.-a).

Vehicle Identification

The VIRTUAL team then applied the trained neural network to each video to extract the contours (e.g., in a rectangle shape) of vehicles in the foreground. The contour of a vehicle in each frame provided a vehicle reference location and size information in the real-world physical coordinate system. The team used vehicle size information to classify the vehicles.

Trajectory Extraction

Because each video was available at a high frequency (e.g., 24 Hz), the contour of a vehicle, if properly detected, moved little and largely overlapped in two consecutive frames. Therefore, the reference locations of the same vehicle across consecutive frames could be easily connected based on the contour overlapping relationship to form a vehicle trajectory (provided that the contour was properly detected).

Trajectory Processing

With physical coordinates of the features obtained from Background Normalization, the VIRTUAL team developed an algorithm to identify a lane number and corresponding longitudinal location for each point along each trajectory. The team calculated speed and acceleration by differentiating the longitudinal locations of a trajectory.

Trajectory Data Format

Table 6 presents the trajectory-level data variables extracted for each vehicle at each timestep. This format is similar to the NGSIM format presented in chapter 2. This data format will allow the research team to compare directly full-set vehicle trajectories from simulation to those obtained from the field.

Table 6. Data format for full-set vehicle trajectories.

Column Name	Explanation	Note
Vehicle ID	ID number for each vehicle.	ID may not be continuous, but it is unique for each vehicle in each dataset.
Global Time (seconds)	Time in s referenced to 12:00:00 a.m.	—
Frame ID	Frame number in the corresponding video.	The beginning portion of the video with severe vibrations might be cut until the frame becomes stable.
Local X (feet)	Position in the cross-section direction.	The reference point is this vehicle's front bumper location.
Local Y (feet)	Position in the direction along the road.	The reference point is this vehicle's front bumper location.
Global *X* (Longitude)	Vehicle's GPS longitude location.	—
Global *Y* (Latitude)	Vehicle's GPS latitude location.	—
Width (feet)	Vehicle width.	—
Length (feet)	Vehicle length.	—

Column Name	Explanation	Note
Class	Vehicle class.	Can take three values: one motorcycle; two light-duty vehicle; three heavy-duty vehicle.
Speed (ft/s)	Vehicle speed.	Calculated by the moving average method (2-s interval).
Acceleration (ft/s^2)	Vehicle acceleration.	Calculated by the moving average method (2-s interval).
Lane Number	Lane number.	From right to left; 1, 2, 3, 4…, *n*
Space Headway (ft)	Distance between this vehicle's front bumper to its following vehicle's front bumper.	−1 when there is no leading vehicle or the leading vehicle is out of scope.

— = No data; *n* = number of lanes.

Identifying Data Errors

After the video was converted into the data variables in Table 6, the data were checked for errors. The researchers noticed a few issues with the trajectories produced by the VIRTUAL tool. This section describes the issues observed and the resolutions used to fix the data. This process will be unique to each individual data collect effort. With this project, there were five issues identified: misalignment of trajectories because of roadway curvature, broken trajectories, infeasible kinematic relationships, timestep errors, and lane number ID. This section discusses how these errors were identified and resolved.

Roadway Curvature

One way to check the trajectory data for errors is to plot these data on a map. Once the numeric trajectory data are available for review, part of that data will include the global latitude and longitude (X and Y) positions of vehicles at each timestep. The research team used an online GPS visualizer tool to import these X and Y positions from a comma-delimited text file, also known as a comma-separated values (CSV) file. The team saved the GPS data into a CSV file with two columns of latitude and longitude (in this order). The file-size limit was 10 MB, so each file was only a subset of the trajectory data. The team thus performed spot checks from the top, middle, and bottom portions of the dataset. In doing so, they noticed that some GPS coordinates were slowly straying off the road while moving further downstream, when viewing the trajectories superimposed on a map. The data processing team ultimately extracted more GPS coordinates so that the trajectories would better fit the roadway curvature.

Crossing Trajectories and Short Trajectories

A second way to check the trajectory data for errors is to plot vehicles' road positions (i.e., Local Y) over time (i.e., Global Time). For this, any software, such as Excel®, can be used. The research team wrote code in Matlab to clean the data.

One problem the researchers observed was crossing trajectories. This implied either a crash or one vehicle passing another without changing lanes. However, the researchers observed no crashes in the videos and passing without lane-changing is physically impossible. The issue in this case was caused by a camera rotation problem, which made identifying the current vehicle lane number more challenging. The data processing team developed a dynamic curve-fitting algorithm to resolve this issue.

A second problem researchers identified was that tracking was lost for a small portion of detected vehicles, which caused broken trajectories. This occurred when the automated tools temporarily lost track of one or more vehicles because of issues with video quality (e.g., shaking camera). This resulted in trajectories that appeared to have occasional gaps (or were broken) when plotted as described earlier in this section. The team developed a trajectory connection algorithm to overcome this problem.

Kinematic Relationships

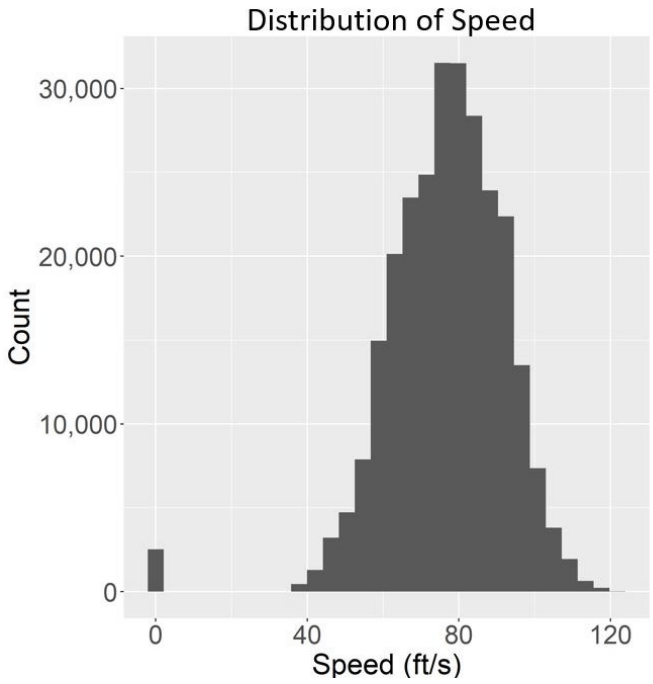
The research team noticed that some speed and acceleration values were unrealistically high. The developers of VIRTUAL further observed that the integral of acceleration over time was sometimes inconsistent with the change of speed over time, and the integral of speed was sometimes inconsistent with the change of location over time. These kinematic anomalies were caused by the nature of the video processing method used in vehicle trajectory extraction. The developers of VIRTUAL ultimately developed an algorithm to reconcile the connected trajectories extracted from videos with their kinematic characteristics.

Timestamps

For one of the time periods at one of the data-collection sites, the team noticed the timestamps associated with vehicle trajectories were clearly several hours different from the time when data were collected. The data processing team fixed this problem and regenerated that set of numeric trajectories.

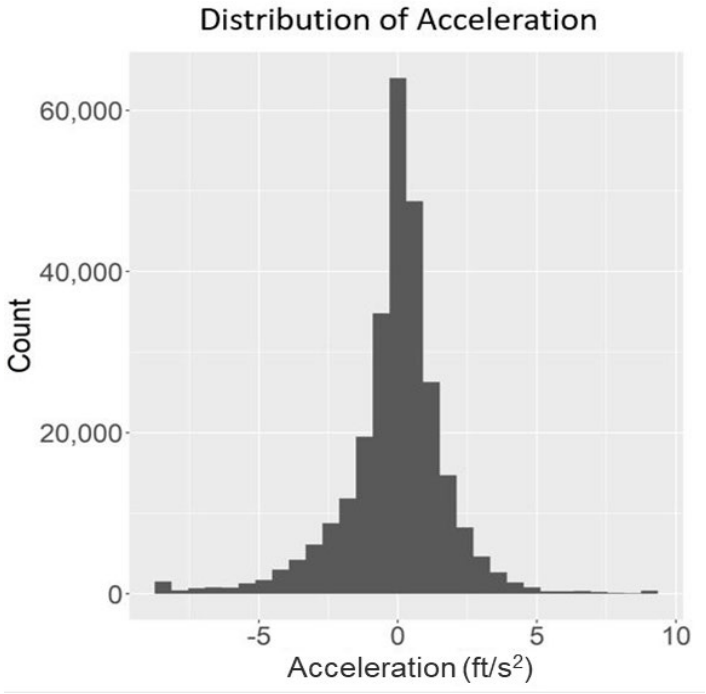
Verification of Data Accuracy

The team verified the accuracy of extracted trajectory data before using them for the calibration experiments. One method involved plotting the speeds and accelerations, as proposed by Coifman and Li (2017), to visually observe whether the speeds and accelerations looked reasonable. These results are illustrated in figure 3 and figure 4. The research team viewed accelerations as being unrealistic whenever their absolute value exceeded 10 ft/s^2. In the NGSIM dataset, approximately 10 percent of the vehicle trajectories indicated such unrealistic acceleration values. In the data collected for this project, the portion of unrealistic acceleration values was close to 0 percent. In figure 3, a non-negligible number of stopped vehicles were extracted by VIRTUAL. The team used the video data to confirm that these vehicles were stopped on the shoulder waiting for a tow truck while data was collected.



Source: FHWA.

Figure 3. Graph. Distribution of speeds from trajectory data at all four sites.



Source: FHWA.

Figure 4. Graph. Distribution of accelerations from trajectory data at all four sites.

A second verification method involved the corroborative probe vehicle data. The data processing team compared the GPS location of probe vehicles and the extracted trajectories by calculating a

correlation coefficient between these two datasets. Based on the calculation, the team successfully identified probe vehicles within the I–95, I–270, and I–15 networks. However, the team could not identify probe vehicles within the I–75 network, because the probe vehicles were deployed outside the time window of the aerial video shooting period.

Lessons Learned

The research team documented lessons learned. These may prove useful for agencies when planning their own trajectory-data post-processing efforts.

The team recommends careful consideration and selection of data collection locations and times. For constructing a congestion map, it may help to analyze multiple days of data to ensure traffic conditions are close to what was expected. This will also help to understand congestion pattern variabilities. Because of uncertainties in the congestion pattern across different days, teams may wish to keep separation between data collection operations (i.e., keeping small time intervals when flying drones and helicopters to collect data). Teams may also consider increasing their data coverage range and time periods (e.g., flying more drones and helicopters) to ensure the formation and disappearance of traffic congestion is captured successfully.

When the video data are being collected via drones or helicopters, severe vibrations tend to produce errors in the trajectory data. Although some data collection companies might be aware of this, it might help to discuss ways to improve stability of the collected aerial video in advance.

It is difficult to identify vehicles during nighttime because light conditions at night significantly differ from those during the day. Thus, it is a possibility that only partial vehicle trajectories can be extracted from nighttime videos unless otherwise specified by one's data extraction software vendor. Procedures to identify and track vehicles during nighttime hours may be discussed before the data collection process.

Signage may block vehicles within the areas of interest. Vehicles fully occluded by signs are impossible to identify and track automatically. When developing a data collection plan, analysts may wish to identify launch parameters (e.g., location, angle) that avoid or minimize vehicle occlusion by signs, overpasses, or other objects.

It may help to deploy some easy-to-detect artificial landmarks in the video, such as construction cones or colored signs, to help mark coordinates of the study area.

The time and effort needed to detect vehicles and process trajectories may depend on the length and quality of the videos. If videos are shot in similar roadway segments in a stable position and from a consistent angle, the time and effort to process videos and extract trajectories will be less. Video processing is significantly more challenging in cases where video quality is not as expected (e.g., unstable or shaky video footage). In such cases, extensive coding and cross-validation may be needed.

CHAPTER 4. CALIBRATION AND VALIDATION METHODOLOGY

This chapter describes vehicle-trajectory-based, traditional, and hybrid calibration methodologies for microsimulation models developed during this project. This chapter also describes corresponding validation methodologies that could help to ensure robustness of the calibrated models.

Regardless of the driver behavior calibration method used (e.g., trajectory, traditional, or hybrid), each of the methodologies makes two assumptions:

- The analyst previously developed and debugged a microsimulation model using available input data.
- The analyst calibrated the model demands using available count data.

With respect to assumption 1, the research team worked with State agencies and explored repositories of previously developed models to locate functional microsimulation models of the four data collection sites. The development of the base model is outside of the scope of this project, and the authors of this paper refer the reader to *Traffic Analysis Toolbox Volume III* (Wunderlich, Vasudevan, and Wang 2019), *Transportation Systems Simulation Manual* (TSSM)[2], State departments of transportation (DOT) guidelines, or various software user guides for additional resources on base model development.

With respect to assumption 2, the authors adopted a sequential form of calibration, calibrating local segment phenomena before proceeding to system-scale issues (e.g., Chu et al. 2004, Toledo et al. 2003). For this project, the team calibrated traffic demands (step zero) before adjusting any driver behavior parameters according to the trajectory-based, traditional, or hybrid calibration methodologies.

STEP ZERO: OBTAINING A BENCHMARK MODEL

The research team believed that driver behaviors are more sensitive to traffic congestion levels than vice-versa. This motivated a calibration sequence that could attain more accurate congestion levels prior to driver behavior calibration. Among the different variables, the number of roadway lanes and traffic demand volumes may have the greatest effects on congestion levels. Therefore, the team quickly gravitated toward a sequence where the traffic demands are calibrated manually first to achieve the best possible matching of simulated and field-measured throughputs (i.e., vehicle trips or discharge rates) for key locations in the network, prior to any trajectory-based calibrations.

Demands are difficult to measure and are thus often calibrated by traffic modelers (Creasey and Sampson 2020; Barceló 2010). The team's approach to calibrating demands was traditional: input demands were modified manually, in an ad-hoc fashion, without software assistance. Following each iteration of modifying input demands, the researcher inspected the simulated throughput at a few key network locations. Although the researcher attempted to achieve better agreement of simulated and field-measured throughput at the key locations over numerous

[2] List et al. forthcoming.

iterations, there was no formal goodness-of-fit measure or acceptable level of error. This process was much more art than science.

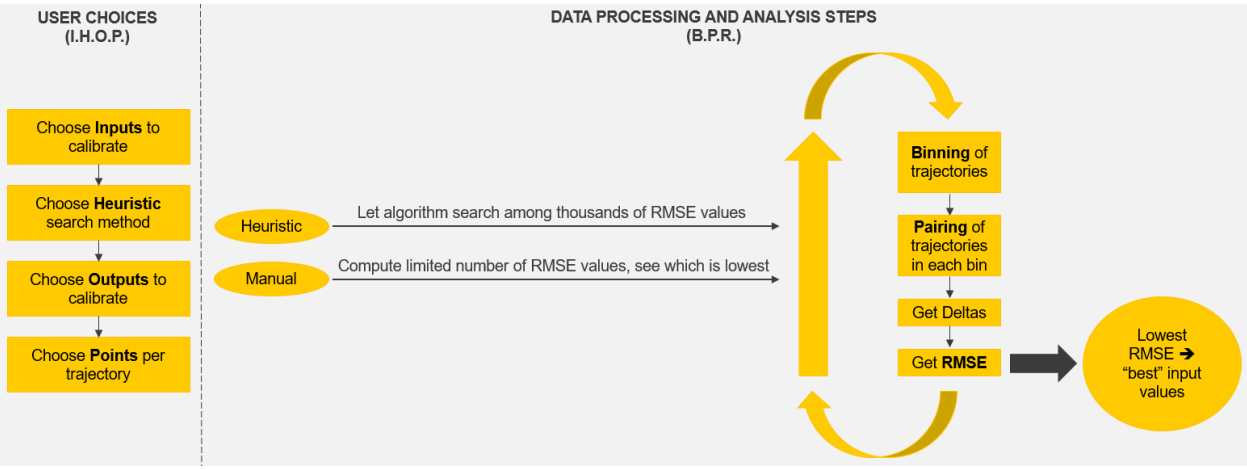
Eventually, upon deciding that little additional improvement was possible or likely, the researcher identified and preserved the set of input demands achieving the best possible agreement between simulated and field-measured throughputs. The researcher entered this best set of input demands into one simulation dataset as the benchmark starting point for subsequent trajectory-based calibrations of driver behavior. The benchmark dataset had car-following and lane-changing parameters set to their default values. This was the starting point for both the trajectory-based, traditional, and hybrid calibration methodologies.

TRAJECTORY-BASED CALIBRATION METHODOLOGY

Researchers designed the trajectory-based calibration methodology to provide a balance between robustness and practicality for State agencies. The developed methodology has seven discrete steps. Please note that the preliminary step of calibrating demands to improve the agreement of simulated and measured throughput (previous section) should be completed and finalized before starting this seven-step calibration procedure.

The first four steps are preliminary choices made by the user, while the last three steps involve automated data processing. The overall methodology assumes that trajectory data have been collected and archived in the format previously described in the "Trajectory Data Format" section. The overall methodology is illustrated in figure 5.

Prior to beginning the calibration process, the analyst should set assign sufficient data to validate the calibrated driver behaviors. This process is discussed in the "Trajectory-Based Validation Method" section later in this chapter.



Source: FHWA.
B.P.R. = Binning, Pairing, RMSE; I.H.O.P. = Inputs, Heuristic, Outputs, Points.

Figure 5. Flowchart. Proposed seven-step trajectory-based calibration method.

Step 1: Inputs

As with many calibration methodologies for traffic microsimulation, the first step is to choose which input parameters to calibrate and which candidate values to consider for each parameter. Car-following and lane-changing parameters are the most natural choices for trajectory-based calibration on urban freeways. Every microsimulation tool contains a car-following model and a lane-changing model with inputs that can be calibrated. For example, the Wiedemann car-following model has 10 input parameters that can be calibrated. The Gipps car-following model has maximum acceleration, maximum deceleration, reaction time, and minimum following distance, all of which can be calibrated. Calibration of lane-changing input parameters may also be considered.

There are tradeoffs between practicality and robustness when choosing which input parameters to calibrate. If a user chooses to calibrate only one car-following parameter, the task is relatively easier than calibrating all of the car-following parameters. However, the resultant calibrated model might not be robust enough to be trustworthy under various conditions. By contrast, if a user chose to calibrate five car-following and five lane-changing parameters for a total of 10 input parameters, this would likely produce a more robust model, but it might be impractical to evaluate the resulting billions of combinations of values. Indeed, if all 10 parameters have 10 possible numeric values, this would produce 10 billion possible combinations of parameter values. Although automated parameter search space algorithms exist, evaluating 10 billion possible parameter sets is likely to be a time consuming and computationally expensive endeavor.

The developed methodology places no restriction on the inputs that can be chosen. As such, the analyst may compromise between practicality and robustness when choosing which input parameters to calibrate. This choice may become easier over time as the analyst gains experience with more microsimulation projects. Over time, the analyst might notice which simulation models lack sufficient accuracy and which projects produce excessive levels of effort.

Moreover, some agency manuals have recommendations and requirements as to what inputs should be calibrated and what acceptable parameter ranges should be considered (e.g., VDOT Traffic Engineering Division 2020, Colorado Department of Transportation 2018). Additionally, some simulation researchers have published product-specific information on which inputs have more impact than others, making them higher priorities for calibration. Armed with this information, an analyst could choose to leave less impactful inputs at default values, possibly making the calibration process more efficient.

This procedure is software and model agnostic, leaving the choice of which model to calibrate to the analyst. For the purposes of the case study, the research team selected a subset of the available car-following and lane-changing parameters in both Aimsun and Vissim for its calibration experiments. The team also selected a few candidate values for each parameter to limit the number of overall candidate solutions. The team used its experience with the microsimulation tools, along with available tool-specific guidance in the literature, to determine which input parameters and candidate values to use in the experiments. These selections led to 162 candidate solutions for the Vissim case study and 156 candidate solutions for the Aimsun

case study. Chapter 5 provides more specificity and details on the car-following and lane-changing parameters selected and the parameter search space.

Step 2: Heuristics

The second step in the proposed method is to choose which search method to use for identifying the best set of calibration parameter coefficients. The search methods considered may be some form of an exhaustive search algorithm or heuristic method. Heuristic methods contain special logic to automatically eliminate many combinations of values unlikely to be optimal (e.g., genetic algorithm). Heuristics may be valuable when the number of calibration parameters and the size of search spaces are larger, resulting in a higher number of possible solutions. Without heuristics, it is necessary to explicitly evaluate all input parameter value combinations through exhaustive enumeration or brute-force searching (Hale et al. 2015). Like step 1, the proposed trajectory calibration method places no restriction on the search method, as no one-size-fits-all search method exists.

As discussed in step 1, the research team decided to limit the number of calibrated parameters and the search spaces such that there were only 162 candidate solutions for VISSIM and 156 candidate solutions for Aimsun. Given the limited number of possible solutions for the case studies presented in this paper, the authors chose to use directed brute force (DBF) searching.

The above discussion illustrates a potential interdependence between step 1 and step 2 of the proposed calibration process. In other words, a user's choice of inputs to calibrate (step 1) might be influenced by which search methods the user wishes to implement (step 2). Conversely, a user's choice of search method (step 2) might be influenced by what input parameters the user needs or wants to calibrate (step 1) because many inputs usually cannot be calibrated by exhaustive enumeration. These choices might be affected by other considerations such as the size of the traffic network, the number of time periods to simulate, the speed of the computer, and the speed of the simulation product. These factors affect the amount of time needed per simulation run, which in turn affects the amount of time needed for calibration. As an example, suppose the user has selected a car-following and lane-changing model with six total parameters; assume that the user has identified five candidate values for each of the six parameters. In this case, the number of possible solutions is $5^6 = 15{,}625$. As such, a heuristic search method will be necessary to limit the search space to the most likely optimal values, unless the analyst can produce 15,625 simulation runs in a reasonable period of time. By contrast, suppose the user is willing to restrict calibration to the two most sensitive parameters, with five candidate values for each parameter. In this case, the search space is $5^2 = 25$ possible parameter sets. In this scenario, the analyst may consider an exhaustive search method or a heuristic.

Although the proposed calibration methodology places no restriction on the search method that can be used, the research team selected the DBF search method in step 2; this was motivated by the limited number of input parameters and candidate values chosen in step 1.

Step 3: Outputs

The third step of the proposed method is to choose output performance measures on which the analyst will evaluate effectiveness of the calibration procedure by comparing the simulated

output performance measures against observed output performance measures. The analyst may select one or more performance measures for comparison. The output performance measure selection could be motivated by literature, State simulation modeling guidance, sensitivity analysis, or engineering judgment.

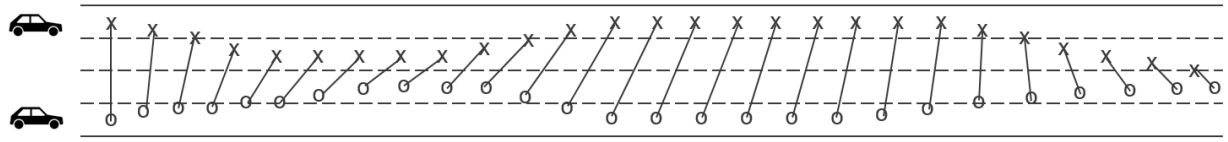
For a traditional calibration of a freeway microsimulation model, these outputs are typically aggregate performance measures such as segment or lane-by-lane speed, throughput, density, and bottleneck duration (Wunderlich, Vasudevan, and Wang 2019).

For a purely trajectory-based calibration, the authors suggest using headways for car-following dynamics, and lane identification (ID) numbers for lane-changing dynamics, throughout the entire trajectory. This decision was motivated by the literature review, where several papers noted that the use of headways more effectively captures car-following dynamics throughout a trajectory (Punzo, Formisano, and Torrieri 2005; Treiber and Kesting 2013a). In the absence of similar literature on how to capture lane-changing effects, the team hypothesized that lane ID numbers could be effective. Additionally, both headways and lane IDs were also readily available within the proposed trajectory data format

If multiple performance measures of interest are identified, the user can specify the relative weighting of these performance measures to the optimization problem (e.g., are car-following and lane-changing equally important or is one performance measure more important than the other?). Step 7 of the method will demonstrate how these output measures and relative weightings affect calculations within the calibration process.

Step 4: Points

The fourth step in the proposed method enables the analyst to decide how many times the procedure compares the predicted performance measure to the performance measure observed in the field. The could be based on a desired number of points (e.g., figure 6 has 27 points), time (e.g., every 2 s) or space (e.g., every 50 ft). There is a tradeoff with this decision. More frequently comparing the trajectories (i.e., a higher number of points) will likely improve model robustness, but at the expense of increased run times for analyses. Conversely, if trajectories are compared less frequently, this will reduce the required run time, but may also reduce the effectiveness of the calibration process. The analyst is free to choose the number of times they wish for observed and simulated performance measures to be compared. This decision may be informed by the literature, early sensitivity analysis, or engineering judgment.



Source: FHWA.

Figure 6. Diagram. Comparison of full-set trajectories.

The spatial and temporal scope of the available data may affect the number of points. For example, if vehicles were traveling at approximately 55 mph (81 ft/s), they would traverse an

800-ft drone coverage area within 10 s. If the analyst selects one comparison point every 2 s, then there would be only five comparison points per full trajectory.

As with steps 1 and 2, this choice can be targeted to achieve the best possible compromise between robustness and practicality. The methodology is flexible and leaves this decision up to the analyst. The proposed methodology neither adopts nor requires a specific number of points. Moreover, the analyst may choose to use time-based or space-based points. To demonstrate this flexibility, the research team used 2-s (for Vissim) and 164-ft (for Aimsun) intervals between points for its own case study experiments. In the absence of guidance in the literature, the researchers hypothesized that interval durations near average driver reaction times might appropriately balance the tradeoff decision in step 4.

Summary of Decisions

The previous four steps allow the analyst to make decisions that impact the practicality and robustness of the adopted trajectory calibration method. These decisions and their tradeoffs are summarized below:

- Inputs—what car-following and lane-changing calibration parameters does the analyst wish to calibrate? Choosing more parameters will likely result in a model more representative of the drivers in the sample but significantly increase the time and resources required to complete calibration.

- Heuristic—which heuristic method should the analyst choose to solve their calibration problem? Exact solution methods will guarantee that they are using the best parameter set, but heuristics have a much more reasonable run time.

- Outputs—which performance measures does the analyst wish to use to determine the accuracy of their calibration procedure? The analyst will compare their simulated and observed performance measures to determine how close their model is matching field results. Traditional performance measures include throughput, speed, and density, while the recommended trajectory performance measures include headway and lane ID.

- Points—how many times does the analyst want to compare the predicted performance measure to the performance measure observed in the field? The more frequently the two values are compared, the better their results are likely to be. This significantly increases the complexity of the problem and will likely require more resources to complete calibration.

Next, this section will discuss the data processing steps. Steps 5–7 may require automation through software or scripts because these steps involve iterations, once for each candidate combination of input parameter values. The researchers developed scripts to assist with this process as described in the appendices (Github, n.d.-b).

Step 5: Binning

There is significant evidence in the literature of inter- and intra-driver heterogeneity in trajectory level data as a function of driver attributes, driver aggression, level of congestion, operational

condition, weather conditions, lane type, and leading vehicle type. Thus, one of the challenges with the trajectory calibration process is ensuring that sufficiently similar trajectories are compared (e.g., it would be inappropriate to compare an aggressive driver's trajectory to a defensive driver's trajectory, just as it would be inappropriate to compare a trajectory collected in congested conditions to a trajectory collected in uncongested conditions).

Thus, the fifth step in the proposed method involves the binning of trajectories (both simulated and field-observed) into specific groups to enable the method to compare sufficiently similar trajectories. The binning process seeks to identify groups of drivers that are likely to behave similarly, minimizing the heterogeneity within the group (James and Hammit 2019).

This step may be the most important and sensitive of the proposed seven-step calibration procedure. Step 5 is critical because within any given bin, the analyst needs to have a reasonable expectation that the comparison between simulated and field-measured trajectories is appropriate.

Step 5 is another decision with tradeoffs for the analyst to consider: increasing the number of bins for the data will likely reduce the heterogeneity within the binned data, but will increase the computational time of the calibration procedure. The analyst is free to choose the types of bins for their data, but must ensure that at the end of the binning process there are enough trajectories for sampling during the pairing step (step 6). The analyst is encouraged to conduct a literature review and an early sensitivity analysis to determine the best bins for their unique dataset. The following subsections describe some suggested bins, but an analyst must be sure to select bins that make the most sense for their unique data sample.

Origin and Destination Lanes

Proper treatment of origin and destination bins, as shown in figure 7, may improve both the efficiency and the robustness of calibration. For example, suppose two trajectories are being compared as shown in figure 6. If a simulated vehicle enters on a different origin lane than the field-observed vehicle, any subsequent differences in their trajectories would not imply an inaccuracy in the simulation. Therefore, on a segment such as the one shown in figure 7, the analyst could separate trajectories into four separate bins according to their origin lane.

Source: FHWA.

Figure 7. Illustration. Origin and destination bins.

However, the research team handled the destination bins differently. The motivation behind this decision was that if pairs of vehicles (one simulated, one field-observed) entering on origin lane 1 frequently exit on different downstream mainline lanes, this could help to expose more deficiencies in the car-following and lane-changing models. As such, the analyst may choose to include all downstream mainline lane numbers within the same destination bin to capture (and reconcile) these driver behavior discrepancies. However, vehicles exiting at the off-ramp may appropriately follow different trajectories than vehicles exiting on the mainline. For this reason, the analyst may define the off-ramp as a second destination bin.

Vehicle Type

The vehicle type bin could be relatively simple (e.g., passenger car, heavy vehicle) or more complex (one bin for each of the 15 vehicle types defined by FHWA (Hallenbeck, Selezneva, and Quinley, 2014). The proposed trajectory data format (shown earlier in the "Trajectory Data Format" section) explicitly provides a numeric code to indicate the vehicle classification. In this data sample, only passenger cars and heavy vehicles were observed. However, the sample of heavy vehicles collected was too small to inform statistically significant analyses of driver behavior. Thus, heavy vehicles were removed from the data sample.

Driver Type

Drivers can be categorized in many different ways, and it is up to the analyst to select the segmentation that makes the most sense for the data available. The simplest approach could simply be two bins: aggressive drivers and cautious drivers. The proposed trajectory data format does not explicitly provide driver aggressiveness. Thus, the research team suggests a simple calculation to infer driver aggressiveness: Time headway is equal to space headway divided by speed, both of which are included in the trajectory data format. The analyst may use a simple rule to divide the drivers into two halves: above average time headways and below average time headways. These are then assumed to be the cautious and aggressive driver bins, respectively. This approach is demonstrated in chapter 5.

For datasets in which attributes are known about the driver, as in the second Strategic Highway Research Program (SHRP2) Naturalistic Driving Study (NDS) dataset, it may make more sense to divide the trajectories by driver attributes such as age, gender, or driving experience (James and Hammit 2019).

Weather

Research has demonstrated that weather affects drivers' car-following and lane-changing behaviors (Hammit, Ghasemzadeh, James, Ahmed, and Young 2018; Wunderlich, Vasudevan, and Wang 2019). However, it is sometimes difficult to collect accurate trajectory data during periods of poor weather. For example, drones may be flown only under ideal conditions, or automated video processing algorithms may struggle with extracting trajectories in adverse weather conditions such as rain, snow, and fog, in which the views of vehicles are limited. If the analyst has access to trajectories in poor weather conditions, they will want to separate trajectories by weather condition. These trajectories may be collected by instrumented vehicle, as

was done in the SHRP2 NDS (Calida et. al 2016). For this project, however, trajectories were collected only during clear weather conditions because of limitations with aerial data collection.

Operational Conditions

The operational condition of a facility has been shown to impact traffic flow (Berthaume, James, Hammit, Foreman, and Melson 2018; Kondyli et. al 2019). Examples of operational conditions include levels of congestion, lane width, and work zones, among others. Previous research has demonstrated that in addition to impacting traffic flow, operational conditions are a source of intradriver heterogeneity that influence driver behavior. In the team's dataset, diverse operational conditions were not collected. However, depending on the nature of one's data, an analyst may want to consider creating bins for different operational conditions.

Summary

Given the datasets available for this research, the selected bins included origin and destination lanes, vehicle type, and driver type. The team divided each origin lane and on-ramp into separate bins, producing four separate bins (three general purpose lanes and one on-ramp). Within each of these bins, the team further divided the data by destination lane type: either general purpose lane or off-ramp. This produced eight total bins. The team next binned the data by driver type: aggressive or conservative. Within each of the eight bins (separated by origin and destination lane), drivers maintaining a below median time headways were classified into an aggressive driver bin, while above the median time headways were classified into a conservative driver bin. Finally, given the low number of heavy vehicles and motorcycles in the underlying data, those vehicle types were filtered out of each of the bins such that only passenger cars were included. This resulted in 16 total bins for the case studies documented in this paper. It should be noted that some freeway segments were assigned a different number of origin bins because they had different numbers of mainline lanes.

Finally, once the bins are finalized, the data should be separated into calibration and validation data. The team then set aside 20 percent of the trajectories in each bin for subsequent validation experiments. This allowed the team to validate the robustness of the calibration process on similar data, but on data that were not used for calibration.

The project team developed scripts to accomplish steps 5, 6, and 7. These details and source codes are provided in the appendices and are available online (GitHub, n.d.-b).

Step 6: Pairing

After the analyst bins the trajectories, the calibration method pairs a simulated trajectory to an observed trajectory within the same data bin from step 5 (e.g., origin lane 1, general purpose lane destination, aggressive driver, passenger car). Pairing vehicles that entered at similar times may help calibration effectiveness, because driver behaviors are sensitive to traffic congestion levels, which change over time (Geng et al. 2016; Ye and Zhang 2009)[3].

[3] Mahmassani et al. forthcoming.

The research team used timestamp-based pairing for the case study experiments featured in chapter 5. The team paired a field-observed vehicle with any simulated vehicle entering the study area within 4 s of one another. In the case that multiple trajectories qualified for pairing, one trajectory was randomly selected. It is important to note that this threshold may be dataset specific. The 4 s threshold was found to work well on the data available to the team, but may need to be identified through sensitivity analysis or engineering judgment for different datasets.

To limit the amount of time required for calibration, the analyst may choose a maximum number of paired trajectories for each bin, keeping in mind that each full trajectory pair may have several points to compare (see step 4). The research team applied a maximum of 25 paired trajectories per bin for its own experiments. It should be noted that the selection of 25 paired trajectories for these case studies was arbitrary and would benefit from sensitivity analyses in future studies.

Pseudocode for trajectory pairing logic is given below:

Repeat for all 32 bins:

> *Repeat for all 25 paired trajectories:*
>
> > *Determine which of the 25 simulated vehicles entered the study area at the earliest time.*
> >
> > *If a field-observed vehicle entered the study area within 4 s of the simulated vehicle's entry time, pair these two trajectories and remove them from the pool of unpaired trajectories.*

For the team's case studies, at the end of this process there could be 800 trajectory pairs (32 bins multiplied by 25 maximum trajectories pairs within each bin) and 4,000 comparison points (8,000 trajectory pairs multiplied by 5 comparison points per trajectory) per trial simulation run. This is highly unique to the team's specific case study and will be different for other studies.

Step 7: RMSE

After the analyst pairs the trajectories, step 7 quantifies the similarity between the observed and simulated trajectories. The objective of the calibration process is to minimize the difference between the observed and simulated trajectories based on the performance measure identified in step 3.

An objective function value of 0 indicates a perfectly calibrated model, whereas an objective function value of infinity indicates an infinitely worthless model. As discussed in chapter 2, RMSE was found to work well as the goodness-of-fit function for driver behavior calibration (Ciuffo, Punzo, and Montanino 2012; Punzo, Formisano, and Torrieri 2005; James 2019; Montanino, Ciuffo, and Punzo 2012) and was adopted as part of this methodology.

For the vehicle trajectory-based calibration, the RMSE needed to reflect the similarity of the trajectories for both car-following (i.e., headways) and lane-changing (i.e., lane ID). Thus, the team needed to normalize both performance measures such that they could be combined into a single dimensionless value, even though they possess fundamentally different units of

measurement. Normalizing is a multistep calculation. The calculations are relatively simple, especially if they are implemented within macros, scripts, or software.

During the normalization process, the analyst first chooses the relative importance of car-following versus lane-changing. For example, a relative weighting of 50–50 would mean the analyst wants car-following and lane-changing to be equally influential within the calibration process. Similarly, a weighting of 75–25 would mean the analyst wants car-following to be three times as influential as lane-changing. To demonstrate the normalization calculation later in this section, a 67–33 weighting will be assumed, indicating that car-following is twice as important as lane-changing. This may be expressed as a relative weighting for headways (i.e., $rw_H = 0.67$), and a relative weighting for lanes (i.e., $rw_L = 0.33$). These relative weights should always sum to 1.0.

Second, the analyst may define the acceptable minimum and maximum range limits for headways and lane numbers. The team did not explicitly research the most appropriate range limits for headways. Within the drone-collected trajectory data, it was intuitive that headway values far below 1.0 s could indicate either an invalid headway, an unusual headway, or an unsafe headway. Similarly, headway values far above 5.0 s could indicate that the "following" vehicle is actually in a free-flow mode, as opposed to engaging in any significant following behavior. Therefore, to demonstrate the normalization calculation, a headway range of 0.5 to 5.0 s is assumed (i.e., $H_{min} = 0.5$; $H_{max} = 5.0$). By contrast, the lane number range has a physical limit. The minimum lane number value is 1, and the maximum value can be equal to the number of lanes on the roadway whose trajectories are being calibrated (e.g., $L_{min} = 1.0$, $L_{max} = 4.0$ for a 4-lane roadway).

Third, the analyst may define the normalized maximum value of a variable called delta (Δ_{nmax}). The word or variable "delta" is used in some models and contexts to indicate the difference between two values. In the trajectory-based calibration method, Δ_H represents the difference between simulated and field-observed headway, and Δ_L represents the difference between the simulated and field-observed lane number. Although any value can be used here, a Δ_{nmax} of 10 will be assumed for the sake of demonstration. In other words, a Δ_{nmax} value of 0 would indicate a perfectly calibrated paired trajectory point, whereas a Δ_{nmax} of 10 would indicate a maximally unrealistic trajectory point.

Next, total delta (Δ_{tot}) for a single paired trajectory point is computed as the sum of Δ_H and Δ_L. The formula to compute Δ_H is given in figure 8. The formula for Δ_L takes the same functional form, but with lane-based variables substituted in for headway-based variables. Note that headways outside the allowable range are not allowed in the calculation of total delta or RMSE.

$$\Delta_H = \Delta_{nmax} \times rw_H \times \frac{|H_{obs} - H_{sim}|}{(H_{max} - H_{min})}$$

Figure 8. Equation. Normalized delta headway calculation formula.

Where:

Δ_H = delta headway.
Δ_{nmax} = normalized maximum delta value.

rw_H = relative weighting for headways.
H_{obs} = field-observed headway (s).
H_{sim} = headway in simulation (s).
H_{max} = acceptable maximum headway (s).
H_{min} = acceptable minimum headway (s).

In this hypothetical example, let us assume that at one of the comparison points (step 4) the observed headway is 1.8 s and the simulated headway is 2.8 s. Additionally, assume the vehicle is observed to be in lane 4 in the field and lane 1 in the simulation. The sample calculation is as follows:

$\Delta_H = (10 \times 0.67) \times \left|1.8 - 2.8\right| \div (5.0 - 0.5) = 1.49$
$\Delta_L = (10 \times 0.33) \times \left|4.0 - 1.0\right| \div (4.0 - 1.0) = 3.30$
$\Delta_{tot} = 1.49 + 3.30 = 4.79$

The net effect is a total delta value between 0 and 10 for each paired trajectory point where delta headway ranges between 0 and 6.7 and delta lanes ranges between 0 and 3.3. A total delta value of 0 implies that the paired trajectory point exhibits the same headway and same lane number in the field as in simulation. A total delta value of 10 indicates the maximum possible discrepancy between a trajectory point in the field and in simulation.

Fifth and finally, after obtaining the normalized total delta values for all points and trajectory pairs in each bin, the normalized RMSE calculation is finally possible, as shown in figure 9.

$$RMSE = \sqrt{\frac{\sum_T \Delta_{tot}{}^2}{T}}$$

Figure 9. Equation. Final RMSE for a candidate simulation run.

Where:
T = total number of trajectory pairs across all bins.
Δtot = delta total for one paired trajectory point.

After this step, there is a singular RMSE value for each candidate combination of input parameter values simulated. The lowest RMSE value then indicates which input parameter values allow the simulation to best replicate field conditions. Because Δ_{nmax} was set equal to 10, an RMSE value of 0 would indicate a perfectly calibrated model, whereas an RMSE of 10 would indicate a completely unrealistic model. Ideally, multiple random number seed replications could be executed to obtain a more statistically reliable RMSE for each combination of input values.

It should be noted that the case studies in chapter 5 assumed a 50–50 relative weighting between headways and lane IDs, indicating that car-following and lane-changing behavior were equally important.

Summary of Trajectory-Based Calibration Method

In the proposed seven-step trajectory calibration procedure, a single RMSE is computed for each combination of input parameter values. The combination of values producing the lowest RMSE is then considered the best solution. The process is divided into four user choices and three data processing steps. The seven steps are illustrated previously in figure 5. The preliminary user choice steps (inputs, heuristic, outputs, points) are abbreviated as I.H.O.P. The subsequent data processing steps (binning, pairing, RMSE) are abbreviated as B.P.R. Below is a summary of the questions that an analyst wishing to use this methodology should ask themselves at each step:

- Inputs—what car-following and lane-changing calibration parameters do I wish to calibrate? A higher number of parameters will likely result in a model more representative of the drivers in my sample, but it would significantly increase the time and resources required to complete calibration.

- Heuristic—what heuristic method should I choose to solve my calibration problem? Exact solution methods will guarantee that I am using the best parameter set I can find, but heuristics have a much more reasonable run time.

- Outputs—which performance measures do I wish to use to determine the accuracy of my calibration procedure? I will compare my simulated and observed performance measures to determine how close my model is matching field results. Traditional performance measures include throughput, speed, and density, while the recommended trajectory performance measures include headway and lane ID.

- Points—how many times do I want to compare my predicted performance measure to the performance measure I observed in the field? The more frequently I compare, the better my results are likely to be; however, this significantly increases the complexity of the problem and will likely require more resources to complete calibration.

- Binning—how should I divide my data to allow me to make comparisons between simulated and observed trajectories? These bins should have minimal interdriver heterogeneity within the clustered trajectories. If I create more bins, I am likely to have smaller clusters of drivers that are more similar. The more bins I use, however, the more computationally taxing my calibration procedure becomes.

- Pairing—within each bin, how do I identify which simulated trajectory to compare to which observed trajectory?

- RMSE—which set of parameter values should I use to calibrate the model? The one with the minimum RMSE. I have multiple performance measures, however. How do all of my performance measures contribute to my overall assessment of the accuracy of the model? What do I consider to be the relative importance of all of my selected performance measures, comparatively?

TRAJECTORY-BASED VALIDATION METHOD

In the proposed method, the calibrated simulation model produces the lowest RMSE value. This value conveys the degree of agreement between the simulated trajectories and the field-collected trajectories. It further conveys a best-case scenario in terms of how much improvement may be possible in the modeling of trajectories. However, an unbiased data source is necessary to assess the expected predictive ability of the calibrated model. As such, one fundamental approach to validating the calibrated model is to set aside a portion of the collected data as holdout data for validation, which is excluded from the calibration process. The process of converting aerial video clips into post-processed numeric trajectory datasets produces thousands of candidate vehicle trajectories for possible inclusion in the calibration process. As such, the analyst can set aside one portion of the field-collected trajectories for validation purposes prior to performing trajectory-based calibration.

Dividing the Data

For the validation to be as unbiased as possible, the analyst should ensure the validation and calibration datasets have the same fundamental composition. For example, the validation and calibration datasets could have the same proportion of trajectories from each bin defined in step 5. The validation and calibration datasets could also have similar profiles or distribution of departure times as discussed in step 6. To achieve the similar and unbiased compositions described in this paragraph, the analyst could randomly sample trajectories from the available bins and time periods and then assign a portion of them to each bucket (e.g., 80 percent of trajectories to the calibration dataset, and 20 percent of trajectories to the calibration dataset). For a rigorous approach, the analyst could apply statistical tests such as Kolmogorov–Smirnov (KS) (Massey 1951) to confirm that the validation and calibration datasets have similar compositions.

There are many different schools of thought for how to split a data sample into calibration and validation data. The size of the underlying dataset will inform this decision. Ideally, even after splitting the data into two groups (one for calibration and one for validation), each bin established in step 5 must have enough trajectories (e.g., the research team wanted 25 trajectories per bin in step 6). One common method for splitting the data into calibration-validation datasets is the "holdout" method, whereby a predetermined amount of the data are used for calibration and the remaining data are used exclusively for validation. This predetermined amount of holdout data could be determined arbitrarily or by using sensitivity analysis. A more rigorous approach known as k-fold validation allows 100 percent of the data to be applied toward calibration (Anguita et al. 2012; James, Hammit, and Boyles 2019).

The research team used sensitivity analysis to inform the selection of calibration-validation data split. The team observed that more realistic trajectories are obtained using an 80–20 split of the data sample. In other words, the team calibrated a model using 80 percent of the real-world trajectories, and then compared simulated trajectories from that model to the remaining 20 percent of the real-world trajectories.

Assessing the Results

Suppose an analyst has successfully applied the proposed trajectory-based calibration method and now has a calibrated model producing the lowest possible trajectory RMSE value. The analyst can now take their calibrated model using the parameter set that produced the lowest trajectory RMSE value and compare the simulated trajectories against the observed trajectories in the validation dataset. A well-fit, predictive model specification will produce a trajectory RMSE value that is acceptably low, defined by the analyst, even using data that was not used to fit the model.

If the trajectory RMSE calculated by comparing the simulated trajectories against the observed trajectories in the validation dataset is too high, this could imply some sort of inconsistency or bias in the original calibration process. It could also indicate overfitting to the data that were sampled for calibration. Because trajectory-based calibration is relatively new, there is no specific procedure for resolving such discrepancies. However, some options could include:

- Ensuring that the calibration and validation datasets (e.g., 80 and 20 percent of trajectories collected from the field, respectively) have a similar proportion of vehicles in each bin compared with typical traffic (i.e., 100 percent of trajectories from the field).
- Ensuring that the trajectories in the calibration and validation datasets are sufficiently similar (e.g., using KS statistical tests).
- Executing more random number replications for each candidate solution defined in step 1 of I.H.O.P. B.P.R.
- Reviewing optimal calibrated model parameters as judged by the calibration and validation datasets to better understand differences and biases between those datasets.
- Fixing more errors in the trajectory data (described in chapter 3).
- Changing the bins defined in step 5 of I.H.O.P. B.P.R.
- Applying k-fold validation to allow more trajectories to be applied toward calibration.

TRADITIONAL CALIBRATION METHOD

The project scope of work required calibrating four microsimulation models according to both traditional and trajectory-based methods to assess advantages and disadvantages of the newly developed trajectory method. For this purpose, the researchers endeavored to perform traditional calibration in a similar manner as trajectory-based calibration. Thus, the research team followed a procedure analogous to the trajectory-based procedure, updating the seven-step procedure as needed. The motivation for this was to facilitate comparisons of driver behavior calibration, which was a key project objective. Moreover, applying certain aspects of the trajectory methodology toward traditional calibration could produce better outcomes than can be achieved in practice, during which input parameters may be modified in a trial-and-error fashion.

The first step was to complete the same “step zero” as required by the trajectory-based calibration methodology. Specifically, step zero for both traditional and trajectory-based methods involve calibration of input volume demands to improve the agreement of simulated and field-measured throughputs at key network locations. After step zero, researchers followed the seven-step I.H.O.P. B.P.R methodology as closely as possible for traditional calibration to avoid introducing unnecessary bias into the comparison experiments. Although steps 4 through 6 are

irrelevant to traditional calibration, the researchers applied steps 1, 2, 3, and 7 in much the same manner for both traditional and trajectory-based calibration.

Specifically, step 1 (inputs) of the traditional calibration methodology was identical to the trajectory-based calibration methodology. That is, the traditional calibration method selected the same input parameters and parameter search spaces as those selected for the trajectory-based methodology. The motivation behind this decision was that the car-following and lane-changing models that are calibrated are the same, whether one is using traditional data (and calibration methods) to calibrate the model or trajectory data (and calibration methods). The parameters and the parameter search spaces used in calibration are unique to each case study and described in chapter 5.

Step 2 (heuristic) of the traditional calibration methodology was also identical to the trajectory-based calibration method. The traditional calibration methodology used the DBF search method to simulate exhaustively each possible combination of input values. The decision to keep step 1 and step 2 of the traditional calibration methodology consistent with the trajectory calibration methodology significantly reduced the computation time for the case studies because the same 162 candidate solutions for Vissim and 156 candidate solutions for Aimsun were used for traditional calibration without the need for additional simulations or datasets.

The methodologies were most different during step 3 (outputs) of the traditional calibration. The team used traditional performance measures of average segment speeds and throughput as the output variables to compare between the simulation and observed data (instead of headways and lane numbers, which were used for trajectory calibration).

As mentioned above, steps 4, 5, and 6 are unique to the trajectory-based method and were not used as part of the macroscopic calibration method.

For step 7, the RMSE goodness-of-fit calculation for traditional calibration was similar in application. For normalization, the team used the highest and lowest observed speeds and throughputs as the ranges. Additionally, the team assumed a 50–50 weighting for throughputs and speeds, meaning that throughput and speed were considered equally important by the calibration objective function.

This approach to steps 1, 2, 3, and 7 meant that for every unique combination of input parameter values both a traditional RMSE and a trajectory-based RMSE could be calculated. The team identified the parameter set corresponding with the lowest traditional RMSE value as the optimal set of calibration parameters for the traditionally calibrated model; analogously, the parameter set corresponding with the lowest trajectory-based RMSE value was identified as the optimal set of calibration parameters for the trajectory-based calibrated model.

This approach allowed the research team to assess the impact of traditional calibration on the realism of both traditional performance measures and vehicle trajectories. Similarly, it allowed the team to assess the impact of trajectory-based calibration on the realism both trajectories and traditional performance measures.

Calculation of traditional RMSEs involved extensive, automated comparisons of simulated throughputs to field-observed throughputs and of simulated speeds to field-observed speeds. The

data collection company obtained these field-observed throughputs and speeds through traditional methods, such as radar and floating car runs, at the same times and locations as the aerial drone data collections. Thus, the field-observed throughputs, speeds, headways, and lane numbers represent the same traffic conditions, facilitating direct comparison of traditional and trajectory-based calibration.

The team acknowledged that the adopted traditional methodology is inconsistent with the methodologies used in practice, such as those described in *Traffic Analysis Toolbox Volume III* (Wunderlich, Vasudevan, and Wang 2019). The adoption of this specific calibration process enabled the team to compare more directly the new methodology with the traditional methodology. Future research is encouraged to compare the new trajectory-based calibration methodology against traditional calibration methodologies that are considered state-of-practice.

TRADITIONAL VALIDATION METHOD

For traditional validation, researchers compared speed-flow scatterplots from simulation to field-observed values at radar sensor locations in the network. Radar data from the data collection company included 5-min traffic counts for each lane, broken down into 10 mph speed intervals. Table 7 shows two records of the radar data spreadsheet.

Table 7. Speed-flow readings from radar data.

Date	Start time	<10 mph	10 to <20 mph	20 to <30 mph	30 to <40 mph	40 to <50 mph	50 to <60 mph	60 to <70 mph	70 to <80 mph	80 to <90 mph	Total Count (veh)
5/2/19	05:30	0	0	2	31	97	44	4	1	0	179
5/2/19	05:35	0	1	1	14	118	47	3	0	0	184

veh = vehicle.

Each point in the speed-flow scatterplots represents a throughput and speed value in a 5-min interval of a specific highway segment. To obtain segment-based throughput values, the team summed up the lane-based throughput numbers for each time-interval and converted it to an hourly based value (see figure 10 for calculation). Next, the team aggregated speed values through all lanes for each time interval. This aggregation was a weighted average of the mean speed of each column, with the corresponding vehicle counts used as weights. For example, the segment-based throughput and speed for the first data record in table 7 is calculated as an example and shown in figure 10 and figure 11.

$$throughput = \frac{179\ veh}{5\ min} \times \frac{60\ min}{1\ h} = 2{,}148\ veh/h$$

Figure 10. Equation. Calculation of throughput for traditional validation.

Where:

throughput = throughput in units of vehicles per hour

$$speed = \frac{2 \times 25 + 31 \times 35 + 97 \times 45 + 44 \times 55 + 4 \times 65 + 1 \times 75}{179} = 46.1 \text{ mi/h}$$

Figure 11. Equation. Calculation of speed for traditional validation.

Where:

speed = vehicle speed in units of miles per hour

This validation effort considered results only from the model parameter settings associated with purely traditional calibration. The researchers calculated segment-based throughput and speed values for all 5-min intervals, for each highway segment having radar data, and for both simulation and field-observed results. Finally, the team plotted speed-flow scatterplots for the highway segments and checked for reasonable correlation between observed and simulated traffic flow conditions.

HYBRID CALIBRATION METHOD

A hybrid calibration objective function that involves both trajectories and traditional measures is also possible, and was explored as part of this project. The method enables an analyst to calibrate a model with trajectories and aggregate traffic data (e.g., throughput and speed) considering both traditional and trajectory performance measures.

To conduct a hybrid calibration, the user may either choose another relative weighting (i.e., the relative importance of trajectories versus traditional measures) or perform multiple sequential calibrations (i.e., trajectory calibration after traditional calibration). For the case studies detailed in chapter 5, the authors chose the former option. To conduct hybrid calibration using the relative weighting method, the analyst must have previously completed a fully trajectory-based and traditional calibration procedure. The RMSEs obtained using both of those methods will be used in the hybrid calibration process.

To complete a hybrid calibration, the analyst should first independently calculate the normalized trajectory and traditional RMSE using the same defined normalized maximum value of a variable (Δ_{nmax}). Next, the analyst chooses the relative importance of trajectory versus traditional performance measures. For example, a relative weighting of 67-33 would mean the analyst wants trajectory measures to be twice as influential as traditional (macroscopic) measures within the calibration process. This may be expressed as a relative weighting for trajectory measures (i.e., $rw_{TRAJ} = 0.67$) and a relative weighting for traditional measures (i.e., $rw_{TRAD} = 0.33$). These relative weights should sum to 1.0. In the final step, the analyst obtains a hybrid RMSE. The hybrid RMSE is the sum of the traditional RMSE and trajectory RMSE, which were previously obtained by completing the traditional and trajectory-based calibration procedures, adjusted by the relative weights. To obtain the adjusted trajectory RMSE for a set of calibration parameters, the analyst multiplies the trajectory RMSE by rw_{TRAJ}. Similarly, to obtain the adjusted traditional RMSE, the analyst multiplies the traditional RMSE by rw_{TRAD}. After this step, there is a singular RMSE value for each candidate combination of input parameter values simulated. The lowest hybrid RMSE value then indicates the input parameter value set that allows the simulation to best replicate field conditions in a way that reflects the chosen relative weightings.

By implementing this approach, the team hoped to identify which relative weighting of traditional and trajectory measures might produce the best overall solution, according to both traditional and trajectory measures.

CONCLUSIONS

This chapter described a vehicle trajectory-based calibration methodology for microsimulation models developed during this project. This chapter also described the corresponding validation method that ensures robustness of the calibrated models. This chapter also described the traditional calibration and validation methodology adopted by this research project. The traditional results serve as a baseline from which to assess the valued added by the trajectory calibration methodology. Chapter 5 details four case studies that allowed the team to compare the results of the developed methodology against results from a traditional calibration procedure.

CHAPTER 5. CALIBRATION AND VALIDATION EXPERIMENTS

This chapter describes the calibration and validation experimental results for the four chosen highway sites: I–270, I–15, I–75, and I–95. In these experiments, calibration and validation were performed using both traditional and trajectory-based methodologies, as described in chapter 4, to compare and assess the newly developed methodology. To develop and validate the proposed microsimulation procedure, the team conducted extensive field data collection efforts at the four sites as described in chapter 3. Video data were collected by drones at all sites except I–75, where video data were obtained by helicopter. At each of the sites using drones for data collection, 3 or 4 point-locations were selected to capture traffic conditions for about 1 hour during rush hours on weekdays; the temporal and spatial section of these points were informed by the queue accumulation polygon discussed in chapter 3. Next, the video data were processed, and vehicle trajectories were extracted by the post-processing tool described in chapter 3. The extracted trajectory data format was described previously in table 6. Post-processed trajectory data were then verified and corrected as described in chapter 3.

I–95, I–75, AND I–270 CASE STUDIES

The research team used Vissim to test the proposed methodology at the I–95, I–75, and I–270 sites. The team coded the I–75 and I–95 models from scratch. The Maryland Department of Transportation State Highway Administration (MDOT SHA) provided the I–270 model, but the team reset driver behavior input parameters to their default values. To identify the most important space and time sections of roadway to capture, the team analyzed three 800-ft sections—the maximum drone coverage length—of I–95 and I–270 using the methods described in chapter 3 (figure 1). For I–75, the team captured a continuous 1.2-mi section by relying on helicopter coverage instead. Warm-up periods, also known as simulation fill time or initialization time, were 15 min for all calibration and validation runs[4]. Before calibrating driver behavior, the team first calibrated input demands to achieve better matching of field-measured and simulated segment throughput, described as “step zero” in chapter 4. The team then executed the I.H.O.P. B.P.R. procedure as follows.

Step 1: Inputs

The research team selected a subset of the available car-following and lane-changing parameters for calibration and selected a small number of candidate values for each parameter to limit the number of overall candidate solutions. The team used prior experience working with the software along with available guidance in the literature (Habtemichael and Picado-Santos 2013; Lownes and Machemehl 2006; MDOT SHA 2017) to determine which input parameters and candidate values to use in the experiments. The team decided to calibrate three car-following and four lane-changing parameters: CC1 (spacing time), CC4 (negative following threshold), CC5 (positive following threshold), deceleration reduction distance (own), deceleration reduction distance (trailing), accepted deceleration (trailing vehicle), and safety distance reduction factor. The literature recommended calibrating CC0 (jam spacing). However, the dataset collected for this project lacked sufficient stop-and-go traffic conditions to allow the team to calibrate this

[4]Analysts may consider using a simulation warm-up time of at least twice the estimated travel time at free-flow conditions to traverse the length of the network (Dowling, Skabardonis, and Alexiadis 2004).

parameter. To make the problem more practical, the team identified ranges of values to consider for each calibration parameter. The values considered for the Wiedemann 99 car-following model are as follows:

- CC1: 0.7, 0.8, and 0.9 s.
- CC4: −0.25 and −0.35.
- CC5 was set equal to −CC4 (e.g., if CC4 = −0.25, then CC5 = 0.25).

The following lane-change model parameters and candidate values were selected for calibration:

- Deceleration reduction distance (own): 50, 100, and 200 ft.
- Deceleration reduction distance (trailing) was set equal to the value above.
- Accepted deceleration (trailing vehicle): −1.64, −3.28, and −6.27 ft/s^2.
- Safety distance reduction factor: 0.2, 0.4, and 0.6.

The remaining parameter values were kept at their default values. The number of combinations was determined as follows:

- Number of combinations = 3 (CC1) × 2 (CC4 and CC5) × 3 (deceleration reduction distance) × 3 (accepted deceleration of trailing vehicle) × 3 (deceleration reduction distance (own and trailing)) = 162

These selections led to 162 candidate solutions for the I–95, I–75, and I–270 case studies.

Step 2: Heuristics

Chapter 4 recommended several viable options for the heuristic step. For this case study, the research team selected the directed brute force (DBF) search method in step 2 to limit the amount of time needed for its own calibration experiments. The team hypothesized that DBF search would perform faster than heuristics given the limited number of calibration parameters and the small search spaces for each parameter. However, for case studies with more parameters and larger search spaces (e.g., hundreds or thousands of possible combinations) heuristic methods will likely be much faster and should be considered.

Experienced modelers typically know that the results of a microsimulation run based on only one random number seed replication tend to be less reliable (Hale 1997). Ideally, multiple random number seed replications will be executed to obtain more statistically reliable output for each combination of input values. Therefore, the team performed 10 random number seed replications for each of the 162 combinations, which resulted in 1,620 microsimulation runs. For real projects, analysts could consider saving time and resources by avoiding multiple random number seed replications for combinations that lack promise.

Step 3: Outputs

Chapter 4 addressed the rationale for selecting output measures. For trajectory-based calibration, the team chose headways and lane numbers and selected a 50–50 relative importance weighting. For traditional calibration, the team chose average segment speed and throughput and selected a

50–50 relative importance weighting. The selections of 50–50 relative importance weightings were arbitrary for these case studies; future research could include sensitivity analysis of relative importance weighting to determine the best selection for traditional and trajectory-based calibration.

Step 4: Points

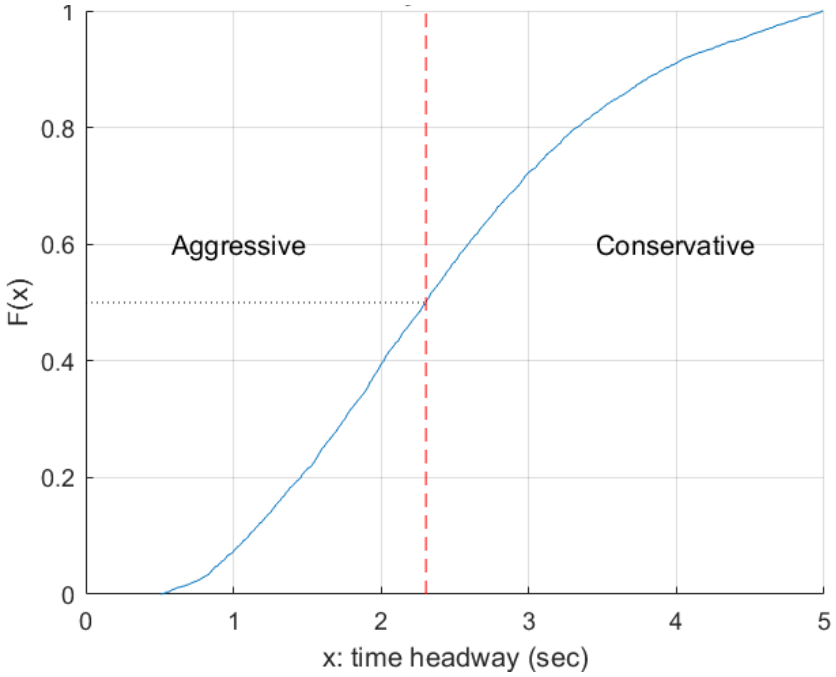
The fourth step in the proposed methodology is to choose the number of comparison points per full-set trajectory. The methodology neither adopts nor requires a specific number of points. The research team used 2-s intervals between points for these case study experiments, with no specific rationale other than estimating what might achieve the right balance between practicality (e.g., computer run time) and robustness. Again, future research could include sensitivity analysis to determine the optimal number of comparison points per trajectory.

Step 5: Binning

The fifth step in the proposed methodology involves binning trajectories into specific groups to enable point-by-point comparisons of vehicle headways and lane numbers of sufficiently similar simulated and observed trajectories. The team wished to select enough bins to allow robust calibration results, but also wanted the number of bins to be small enough to make the experiments efficient and practical. The team defined 16 bins: two driver types (aggressive and conservative), one vehicle type (passenger car), four mainline origin lanes, two destinations (off-ramp and mainline).[5] The team read vehicle type and lane number information directly from the trajectory data file format described in table 6.

To characterize driver aggressiveness, researchers used time headways as described in chapter 4. Time headway is equal to space headway divided by speed, both of which are provided by the data in table 6. The team used simple coding logic to divide the trajectories into two halves: above average (aggressive drivers) and below average (cautious drivers) time headways. The team then sorted these trajectories into the cautious and aggressive driver bins, respectively. figure 14 illustrates that half of the I–95 drivers exhibited time headways above 2.25 s.

[5] The I–75 case had only 8 bins instead of 16 bins because no off-ramps were present.



Source: FHWA.
The red dashed line denotes the mean (50th percentile) time headway (2.25 s);
x = time headway; F(x) = cumulative distribution of time headways.

Figure 12. Graph. Cumulative distribution of time headways (I–95 study area).

Step 6: Pairing

The team developed scripts to automate the trajectory pairing process. The team believed that pairing simulated and field-observed trajectories entering the study area at approximately the same time could produce robust calibration results. The team chose a 4-s and 200-ft threshold. Any simulated vehicle entering the study area within 4 s and 200 ft of a field-observed vehicle could be paired with that vehicle. To limit the amount of time required for calibration, the team limited the number of paired trajectories per bin to a maximum of 25.

The thresholds for pairing (e.g., space and time window) and the maximum number of paired trajectories per bin were decided based on engineering judgment. In the future, an analyst may want to consider conducting sensitivity analysis on these three parameters to determine the best selections for their data sample.

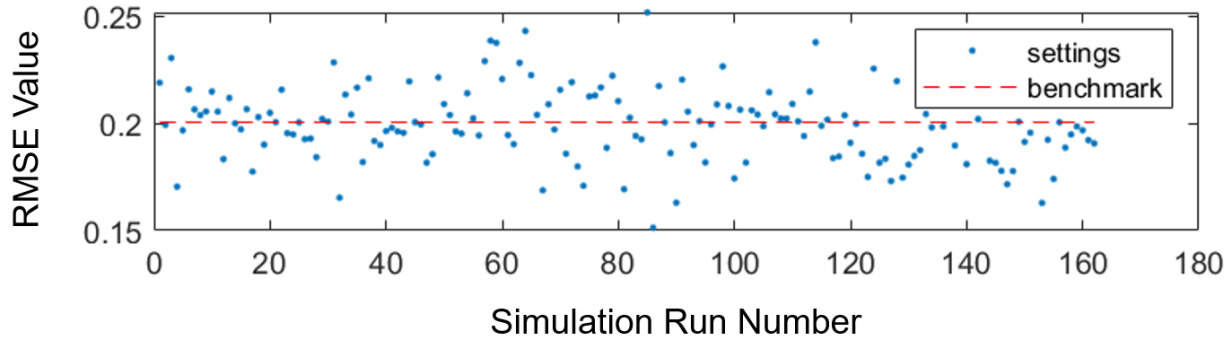
Step 7: RMSE

The team's literature review (chapter 2) indicated that RMSE is an effective goodness-of-fit measure for calibrating traffic simulation models and driver behavior models. For the trajectory-based RMSE, the team used a 50–50 relative weighting of headways and lane numbers to blend them into a single value. This allowed the RMSE value to reflect both car-following and lane-changing effects. For the traditional RMSE, the team used a 50–50 relative weighting of throughputs and speeds to blend them into a single value. The calculations to accomplish this are

described in chapter 4. The team further obtained hybrid calibration via relative weightings of traditional and trajectory RMSE, as described in chapter 4.

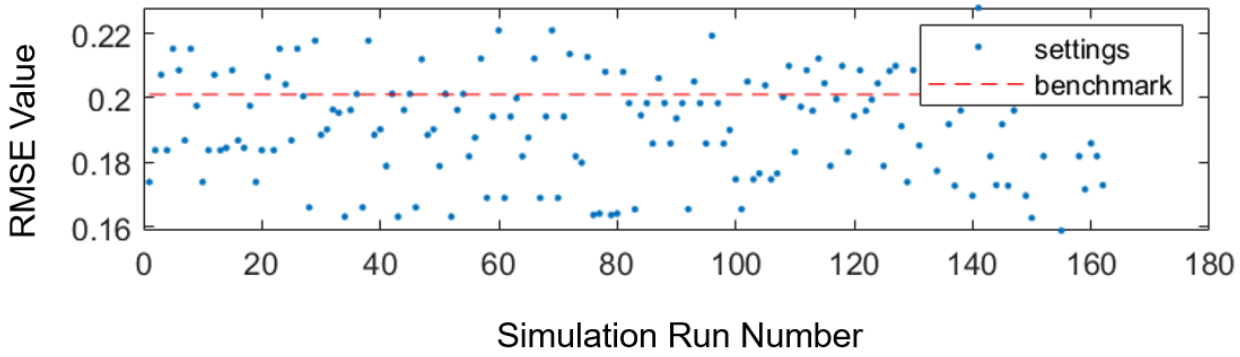
The team developed a script to compute both a trajectory-based RMSE and a traditional RMSE for all 162 simulation runs. Figure 13, figure 14, and figure 15 illustrate the trajectory-based RMSE values for I–95, I–75, and I–270, respectively.

The team selected a maximum normalized delta value of 1.0 for these experiments, such that the RMSEs in these figures were effectively constrained to a range of 0.0 to 1.0. The benchmark RMSE value in these figures represents step zero of the proposed method. In step zero, traffic throughput volumes were calibrated to achieve more accurate simulated throughputs. The driver behavior parameters were left at default values for the benchmark model. This simulation model was then used as the starting point for subsequent calibration of driver behavior.



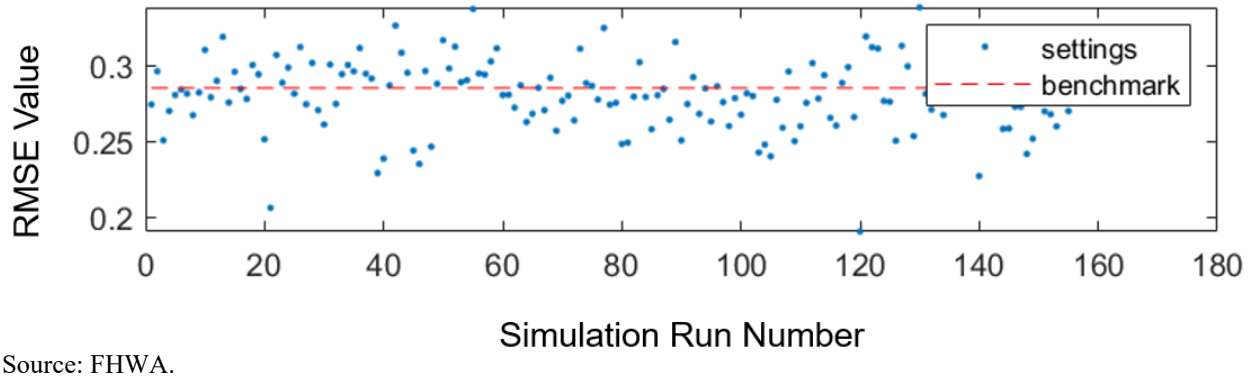
Source: FHWA.

Figure 13. Scatterplot. I–95 trajectory RMSE.



Source: FHWA.

Figure 14. Scatterplot. I–75 trajectory RMSE.



Source: FHWA.

Figure 15. Scatterplot. I–270 trajectory RMSE.

Calibration Results

This section includes both model-specific calibration results and overall model calibration implications.

Model Specific Calibration Results

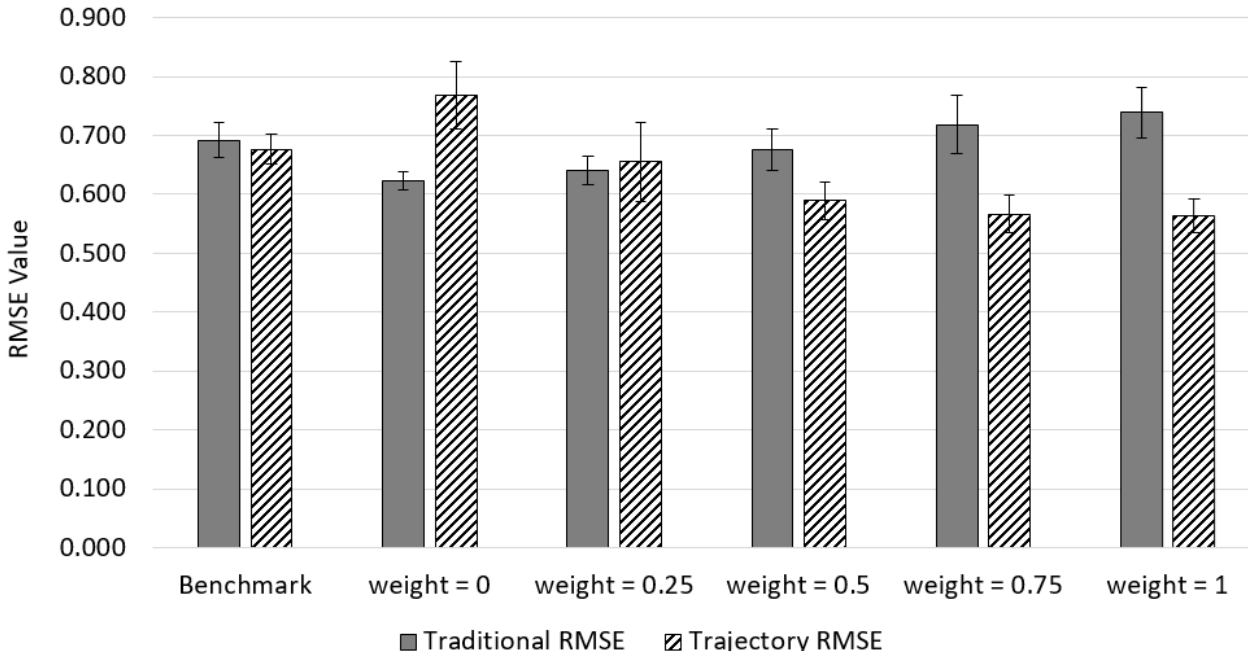
The results appeared to confirm suspicions held by the research team: if trajectories are excluded from the calibration process, simulated trajectories may be unrealistic in terms of headway and lane ID, even if aggregate measures have good agreement with those observed in the field. Figure 16, figure 17, and figure 18 provide results for the I–95, I–75, and I–270 networks, respectively.

To help with interpreting these graphs, please note that traditional calibration is indicated in these figures as "weight = 0," while a pure trajectory-based calibration of driver behavior is indicated in these figures as "weight = 1." The step zero model with calibrated throughputs and default driver behavior parameters is labeled "Benchmark." "Weight = 0.25," "weight = 0.5," and "weight = 0.75" are all hybrid calibration model results, where the relative importance of traditional performance measures (e.g., throughput and speed) and trajectory performance measures (e.g., lane ID and headway) varied. In weight = 0.5, traditional and trajectory performance measures were considered equally important. For weight = 0.25, the traditional performance measures were considered to be three times as important as the trajectory measures. Conversely, for weight = 0.75, trajectory performance measures were considered to be three times as important as the traditional performance measures. Hybrid RMSEs calculations are discussed in chapter 4.

In these bar charts, only the best solution (i.e., with the lowest RMSE out of 162 candidates) is shown for each weight, with a box-and-whisker notation to indicate the range of random number seed outcomes.

For the I–95 network shown in figure 16, the traditional calibration methodology (weight = 0) improves the estimate of traditional RMSE (i.e., average lane speed and throughput) compared with the benchmark model, in which the driver behavior parameters were not calibrated. However, the trajectories produced using the traditional calibration methodology are less

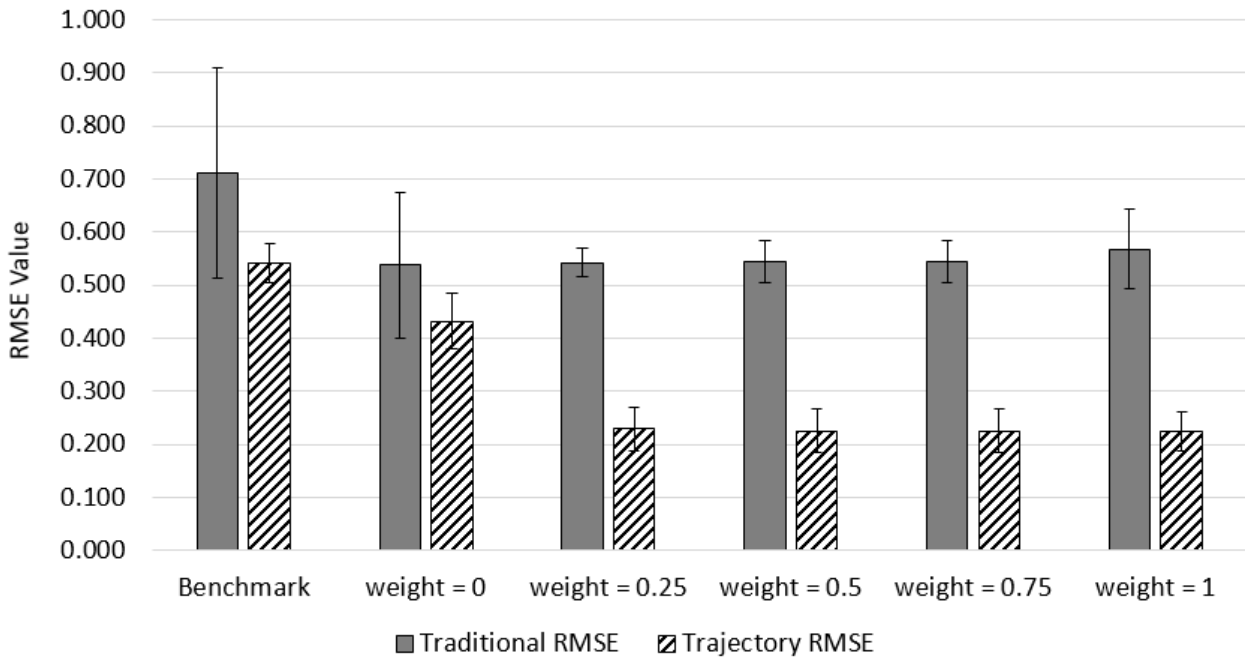
accurate than the trajectories produced using the default driver behavior parameters. The trajectory-based calibration methodology (weight = 1) improves the trajectory RMSE compared with the benchmark model and the traditional calibration model, indicating that the individual vehicle's lane assignment and headways better match what was observed in the real-world data. However, this methodology produces worse traditional RMSE than the benchmark and traditional calibration model, indicating that the lane-specific speed and throughput do not match the field data as well.



Source: FHWA.
Weight is the relative weighting of trajectory-to-traditional calibration.

Figure 16. Bar Chart. I–95 calibration results.

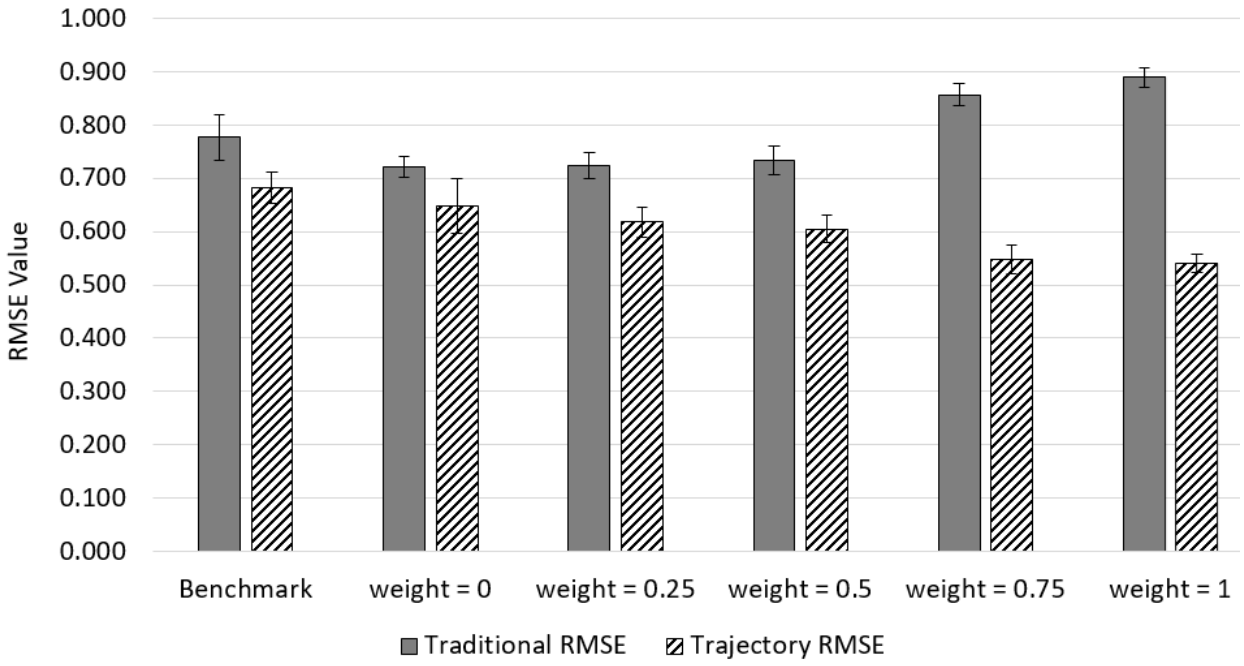
For the I–75 network shown in figure 17, the traditional calibration methodology (weight = 0) improves the traditional RMSE and the trajectory RMSE compared with the benchmark model, where the driver behavior parameters are not calibrated. This indicates that following the traditional calibration methodology produces simulated headways, lane numbers, average lane speed, and throughput that better match the observed data compared with the model where the driver behavior are held at default values. The trajectory-based calibration methodology (weight = 1) improves the trajectory RMSE significantly compared with the benchmark model and the traditionally calibrated model. Moreover, this model performed equivalently well at producing lane-specific speed and throughputs (traditional RMSE) compared with the traditionally calibrated model. The team hypothesized that the longer trajectories extracted from the helicopter data contributed to the improvement in modeling results.



Source: FHWA.
Weight is the relative weighting of trajectory-to-traditional calibration.

Figure 17. Bar Chart. I–75 calibration results.

For the I–270 network results shown in figure 18, the traditional calibration methodology (weight = 0) improves the estimate of traditional RMSE (i.e., average lane speed and throughput) and trajectory RMSE (i.e., lane number and headways) compared with the benchmark model, where driver behavior parameters were not calibrated. The trajectory-based calibration methodology (weight = 1) improves the trajectory RMSE compared with the benchmark model and the traditional calibration model, indicating that the individual vehicle's lane assignment and headways better match what was observed in the real-world data. However, this methodology produces worse traditional RMSE than the benchmark and traditional calibration model, indicating that the average lane speed and throughput do not match the field data as well. The authors hypothesize that this may have been due to I–270 network complexities that made the pairing process more difficult (e.g., managed lanes, overpasses, lots of on-ramps and off-ramps).



Source: FHWA.
Weight is the relative weighting of trajectory-to-traditional calibration.

Figure 18. Bar Chart. I–270 calibration results.

Model Calibration Implications

The effect of pure trajectory calibration on the traditional measures was inconsistent across the different sites. At the I–75 site, trajectory-based calibration (weight = 1) made the traditional measures more accurate than those obtained with the benchmark and traditionally calibrated models. At the other two sites, however, trajectory-based calibration made the traditional measures less accurate. That is, by calibrating the models only considering headways and lane IDs, the observed and measured throughputs and speeds were somewhat less accurate compared to the benchmark model and model calibrated using traditional data. These results suggest that it may be important to consider both macroscopic properties of traffic flow (e.g., throughput and speed) and vehicle trajectories (e.g., headways and lane IDs) in the calibration process, to achieve a model whose outputs are realistic. Moreover, these results suggest that future research is necessary to determine if the limited spatial and temporal scope of trajectories collected via drones (i.e., 800 ft) is sufficient to capture data for calibration, as the calibration performed with longer trajectories performed much better. If longer trajectories are indeed a requirement, future data collection efforts could explore the impact of flying multiple drones in tandem and stitching together videos, in lieu of using helicopters for data collection.

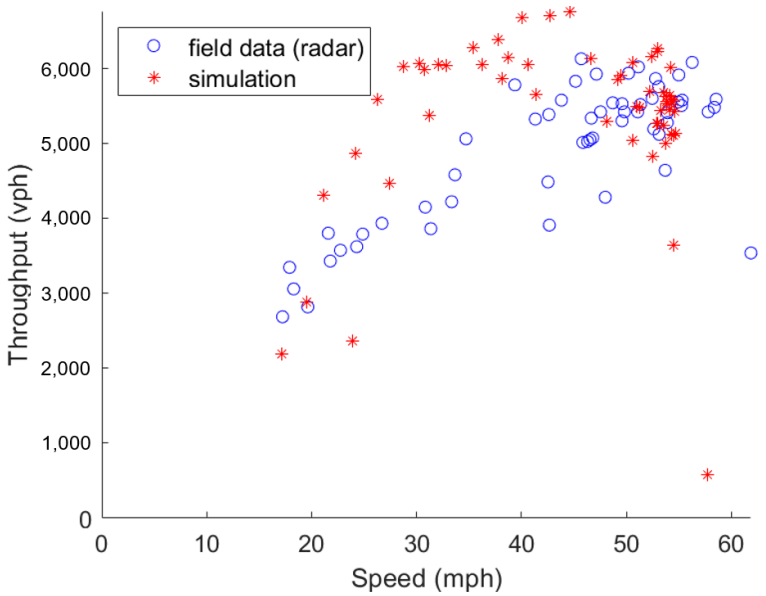
Additionally, the results suggest there exists an opportunity to calibrate a model considering both trajectory and traditional data and performance measures. This is the hybrid calibration method discussed in chapter 4 and noted on figure 16 through figure 18 by intermediate weights (e.g., 0.25, 0.5, and 0.75). Figure 16 through figure 18 demonstrate that the hybrid calibration method does not typically identify the best (i.e., lowest) trajectory or traditional RMSE. However, the hybrid calibration method does a much better job of balancing the need for accurate trajectories

(i.e., headways, lane numbers) and macroscopic traffic performance measures (i.e., average lane speed, throughput) than either methodology that excludes the other data type (i.e., purely trajectory-based or purely traditional calibration). The methodology is flexible and allows any relative weighting between 0 and 1. If time and resources exist, analysts could consider conducting sensitivity analysis on this parameter by producing bar charts such as this before choosing which calibrated model (at which relative RMSE weighting) to use for future predictions or alternatives analysis.

Traditional Validation Results

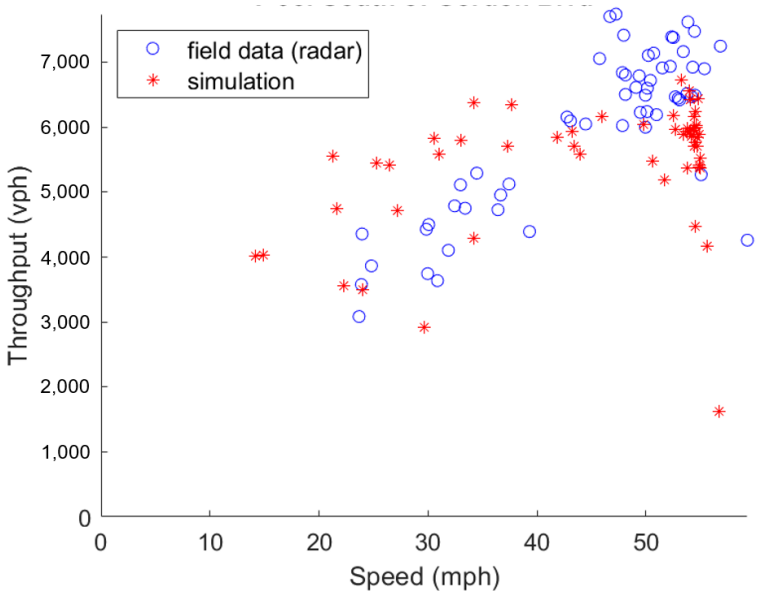
Following the calibration experiment, the research team performed validation by both traditional methods and trajectory-based methods. Chapter 4 provided details for the team's approach to both. Regarding traditional validation, the team constructed speed-flow diagrams for two radar locations. The I–95 model validation results are illustrated in figure 19 and figure 20. The model validation results for I–270 are shown in figure 21 and figure 22. For I–75, the team created speed-flow diagrams for three locations (figure 23, figure 24, and figure 25).

The similarity between simulation and field-data patterns, as shown by the points in the scatterplot, indicate a reasonable correlation between observed and simulated traffic flow conditions, albeit with clear room for improvement. In other words, these results imply that the calibrated models meet minimum validity standards but could be more robust if other samples of field data could be applied toward traditional calibration.



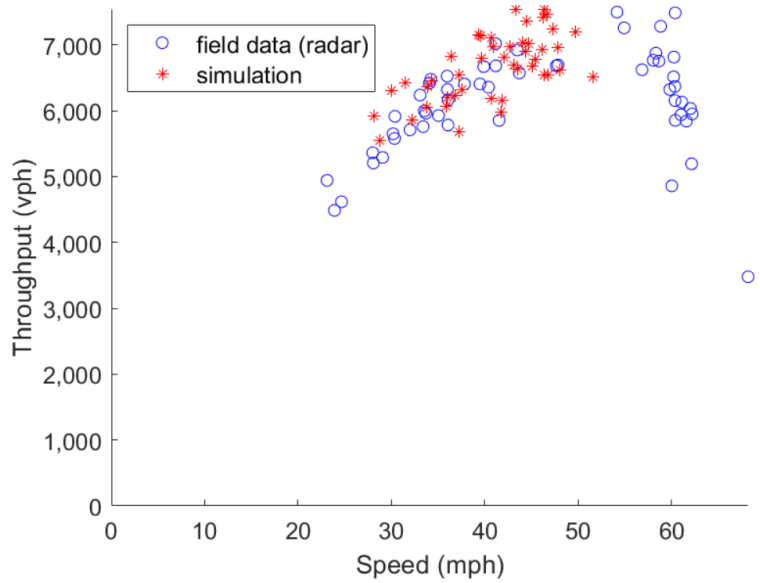
Source: FHWA.

Figure 19. Scatterplot. I–95 speed-flow diagram north of Gordon Boulevard.



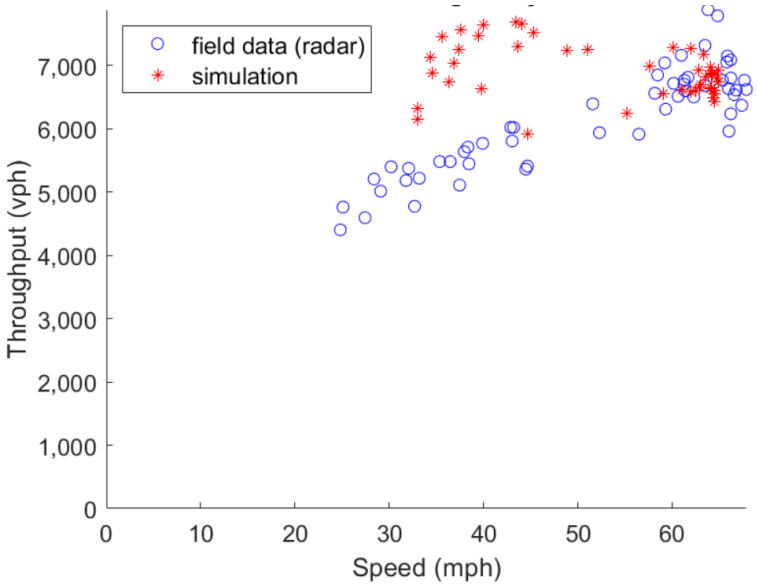
Source: FHWA.

Figure 20. Scatterplot. I–95 speed-flow diagram south of Gordon Boulevard.



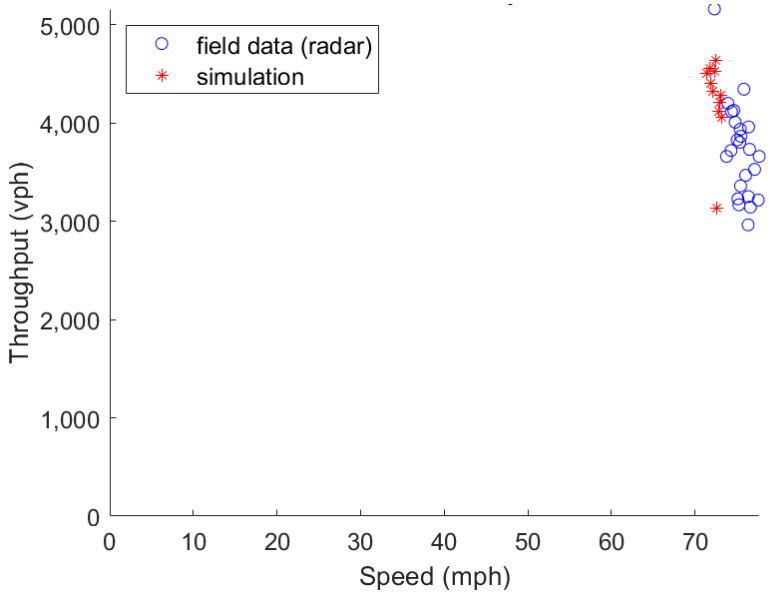
Source: FHWA.

Figure 21. Scatterplot. I–270 speed-flow diagram north of Middlebrook Road.



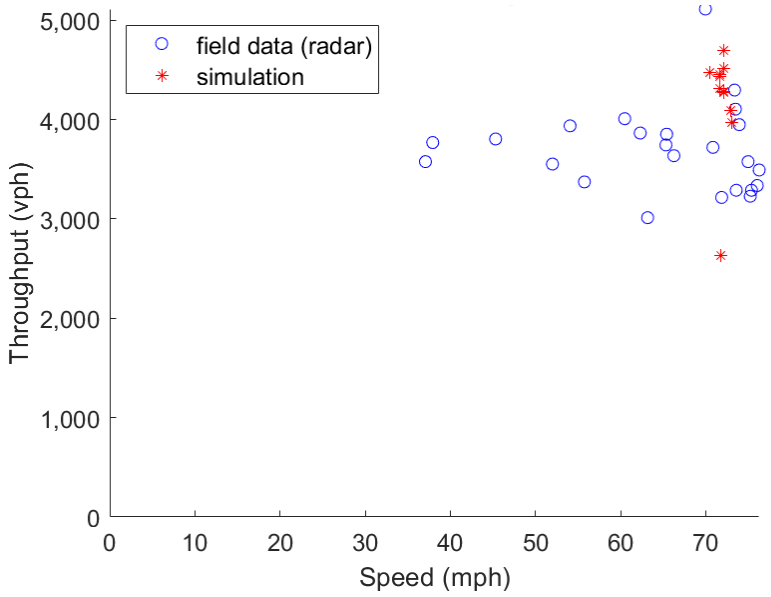
Source: FHWA.

Figure 22. Scatterplot. I–270 speed-flow diagram north of Montgomery Avenue.



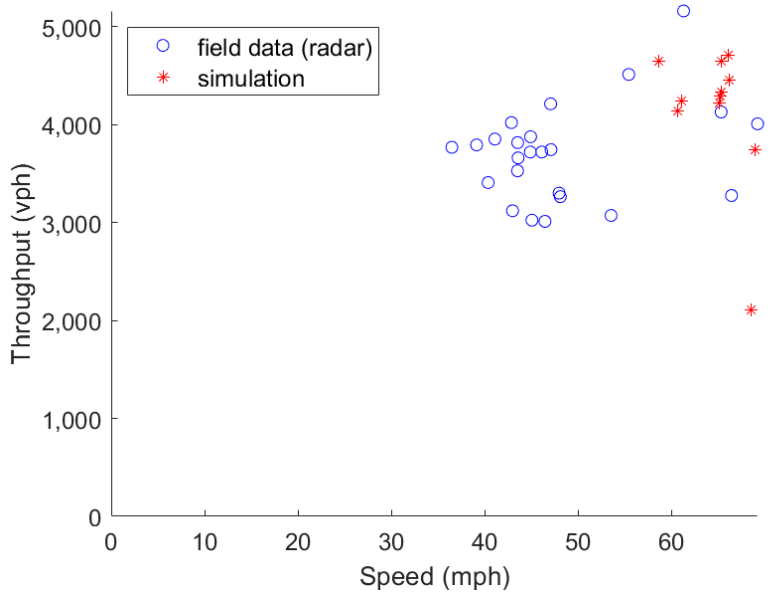
Source: FHWA.

Figure 23. Scatterplot. I–75 speed-flow diagram south of New Tampa Boulevard.



Source: FHWA.

Figure 24. Scatterplot. I–75 speed-flow diagram north of New Tampa Boulevard.



Source: FHWA.

Figure 25. Scatterplot. I–75 speed-flow diagram south of I–275 Crossover.

Trajectory Validation Results

As described in chapter 4, for trajectory-based validation, the datasets were divided into two separate groups: calibration data and validation data. Following calibration, the calibrated model was evaluated according to validation (holdout) data. Deciding how to proportion the data into calibration and validation bins was more art than science, as no hard and fast rule exists in the literature. The team decided to apply 80 percent of the trajectories toward calibration and 20 percent toward validation primarily to achieve enough trajectories in each bin for calibration.

This section discusses model specific validation results, validation challenges and solutions, and model calibration implications.

Model Specific Validation Results

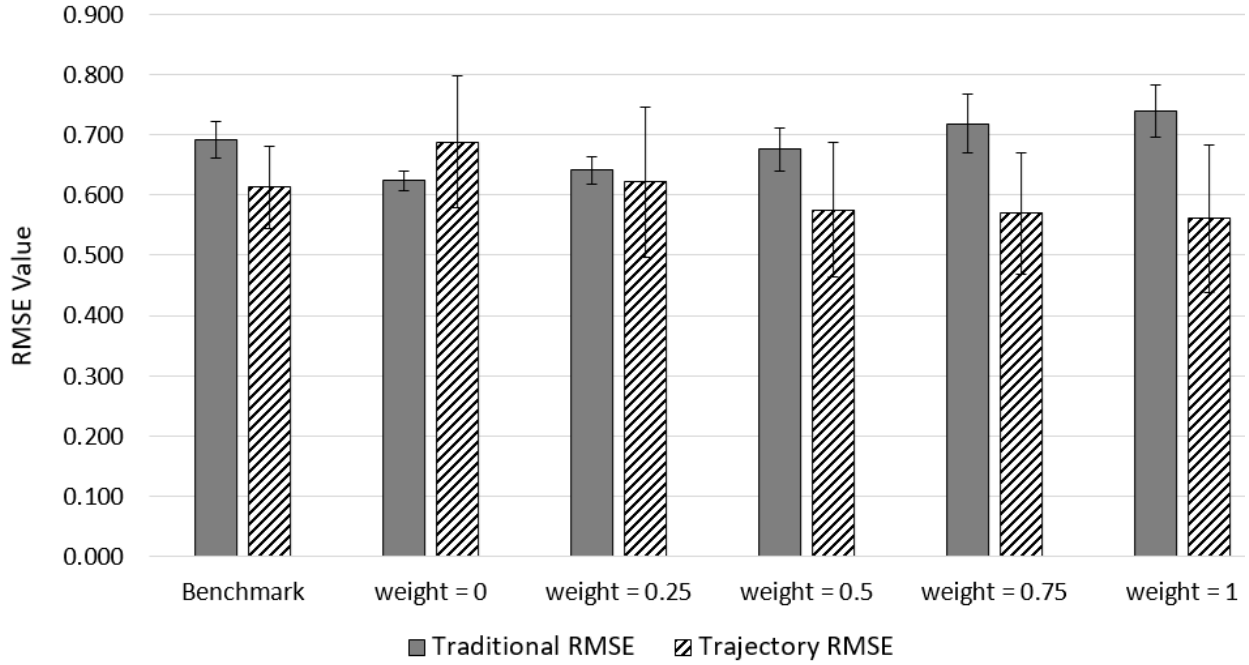
The following validation exercise tested five calibrated models. These five weights represent the hybrid calibration at three different relative weights, the model calibrated considering only traditional data, and the model calibrated considering only trajectory data. The resulting RMSE values are graphed within figure 26 for I–95, in figure 27 for I–75, and in figure 28 for I–270. Traditional RMSE values are unchanged from the calibration results because traditional validation was performed using the speed-flow diagrams instead of a separate validation dataset. These results provide further evidence that the trajectory-calibrated models provide more realistic trajectories than the benchmark or traditionally calibrated models.

For the I–95 network, the traditional calibration methodology (weight = 0) improves the estimate of traditional RMSE (average lane speed and throughput) compared with the benchmark model, in which the driver behavior parameters were not calibrated. However, the trajectories produced using the traditional calibration methodology are less accurate than the trajectories produced using the default driver behavior parameters. The trajectory-based calibration methodology (weight = 1) improves the trajectory RMSE compared with the benchmark model and the traditional calibration model, indicating that the individual vehicle's lane assignment and headways better match what was observed in the real-world data. However, this methodology produces worse traditional RMSE than the benchmark and traditional calibration model, indicating that the average lane speed and throughput do not match the field data as well.

For the I–75 network, the traditional calibration methodology (weight = 0) improves the traditional RMSE and the trajectory RMSE compared with the benchmark model, in which the driver behavior parameters were not calibrated. This indicates that following the traditional calibration methodology produces simulated headways, lane numbers, average lane speed, and throughput that better match the observed data compared with the model where the driver behavior are held at default values. The trajectory-based calibration methodology (weight = 1) improves the trajectory RMSE compared with the benchmark model and the traditionally calibrated model. Moreover, this model performed equivalently well at producing lane-specific speed and throughputs (traditional RMSE) compared with the traditionally calibrated model. The team hypothesizes that the longer trajectories extracted from the helicopter data contributed to the improvement in modeling results.

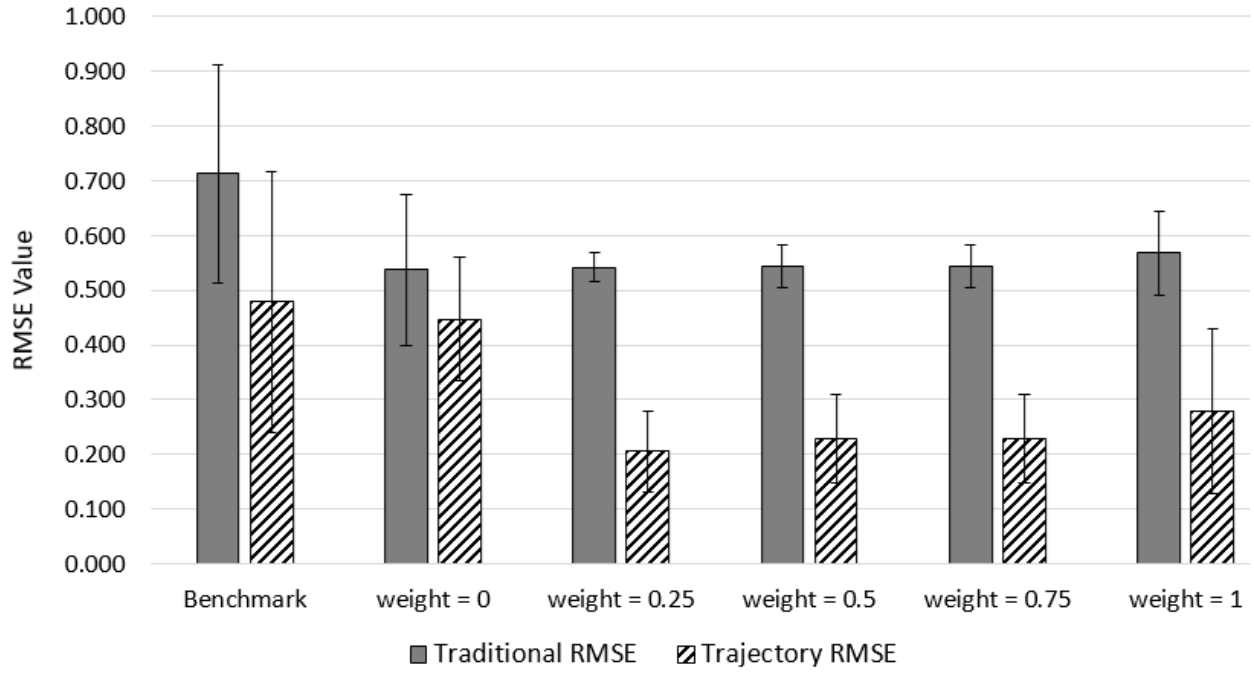
For the I–270 network, the traditional calibration methodology (weight = 0) improves the estimate of traditional RMSE (average lane speed and throughput) and trajectory RMSE (lane number and headways) compared with the benchmark model, where driver behavior parameters were not calibrated. The trajectory-based calibration methodology (weight = 1) improves the trajectory RMSE compared with the benchmark model and the traditional calibration model, indicating that the individual vehicle's lane assignment and headways better match what was observed in the real-world data. However, this methodology produces worse traditional RMSE than the benchmark and traditional calibration model, indicating that the average lane speed and throughput do not match the field data as well.

In the I–75 results, it is notable that the hybrid calibration (weight = 0.25) model predicted trajectories more accurately than the purely trajectory-based (weight = 1) model. This may be due to having only 10 random number seed replications or having only 20 percent of the trajectory data applied toward validation. Moreover, the error bar for the purely trajectory-based calibration shows that the distribution of the modeling results obtained with 10 random number seeds has a high variance; this may skew the average trajectory RMSE for the purely trajectory-based calibration on the higher side. The authors believe that if they had applied more random number seed replications or if they had used a more rigorous validation approach, such as k-fold validation, such anomalies would be less likely.



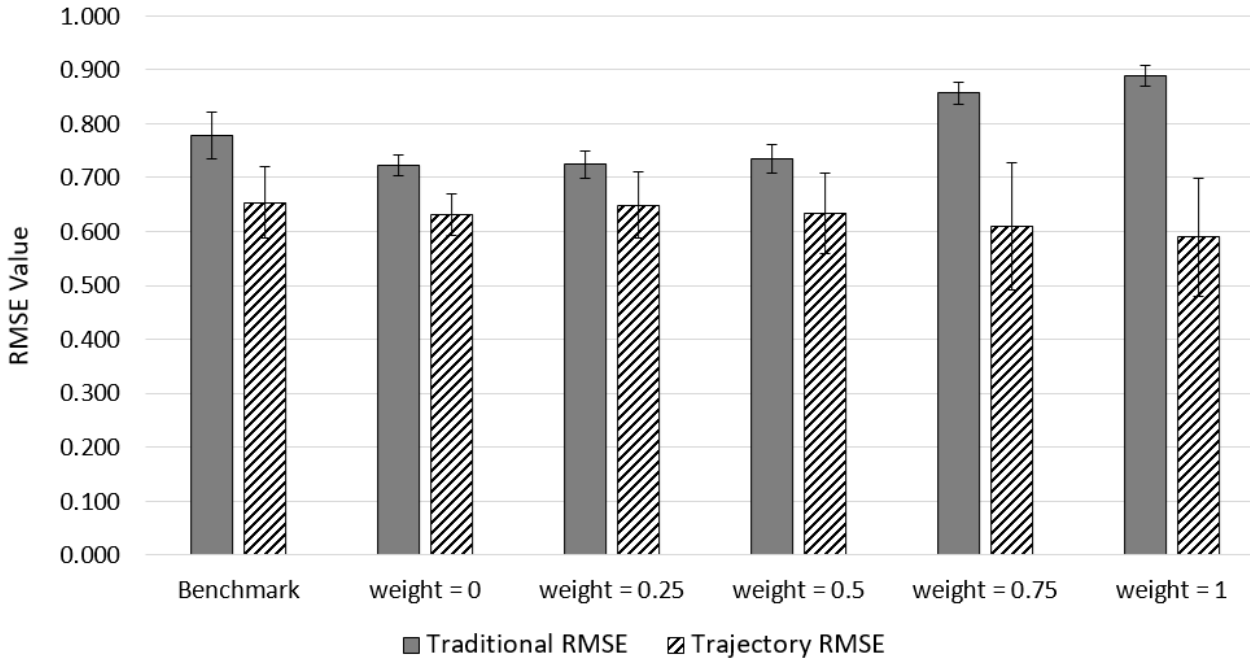
Source: FHWA.
Weight is the relative weighting of trajectory-to-traditional calibration.

Figure 26. Graph. I–95 validation results.



Source: FHWA.
Weight is the relative weighting of trajectory-to-traditional calibration.

Figure 27. Graph. I–75 validation results.



Source: FHWA.
Weight is the relative weighting of trajectory-to-traditional calibration.

Figure 28. Graph. I–270 validation results.

Validation Challenges and Solutions

Although the above validation results look fairly promising, the team's initial validation experiment was a failure. Initial results showed that the trajectory-calibrated models failed to provide more realistic trajectories than the benchmark or traditionally calibrated models. This led the team to find and fix bugs in their scripting code for processing the trajectory data. It also led the team to perform 10 random number seed replications per candidate solution and revise the proportion of vehicles in each bin.

Originally, the calibration and validation datasets were created by sorting trajectories randomly, such that the bins were not at all considered. The improved sorting logic performed random sorting of trajectories within each bin, instead of random sorting throughout the entire set of collected trajectory data. This ensured that the calibration and validation datasets would better reflect typical traffic and the distribution of the other dataset, as discussed in chapter 4.

The team believes this adjustment, to account for binning when creating the calibration and validation datasets, was the most important correction that allowed the validation experiment to be successful. Ultimately the validation exercise, which led to the three key changes in the way validation was performed (i.e., fixed bug in the scripting code, 10 random number seed replications, consistent proportion of vehicles in each bin), correspondingly led to revisions in the calibrated models, which allowed those models to become more robust and trustworthy. Other possible solutions to failed validations are discussed at the end of the "Trajectory-Based Validation Method" section of chapter 4.

Model Calibration Implications

Overall, the trends of traditional RMSE versus trajectory RMSE are consistent across the calibration data (figure 16 through figure 18) and the validation data (figure 26 through figure 28). This indicates that the trajectory calibration method can capture generalizable trends in driver behavior without overfitting to the calibration data sample.

Additionally, the results provide further evidence that the trajectory-calibrated models provide more realistic trajectories than the benchmark or traditionally-calibrated models. This suggests that trajectories should be considered during calibration to ensure simulated trajectories are realistic. Indeed, these results suggest that there is no one size fits all approach to calibration: models calibrated with macroscopic performance measures generally produce results that more accurately depict the macroscopic characteristics of traffic flow (e.g., throughput, speed), while models calibrated with trajectories more accurately capture individual vehicle movement (e.g., headways, lane ID). The exception to this trend occurred with the I–75 model, where the model calibrated purely with trajectory data could capture the macroscopic traffic flow characteristics almost as well as the model calibrated purely with traditional performance measures like throughput and speed. The team hypothesizes that with longer trajectories, reliable calibration of traffic flow (throughput and speed) and trajectories (headway, lane IDs) considering a purely trajectory-based calibration procedure is possible.

Until longer trajectories are more ubiquitously available to transportation agencies, there exists a hybrid approach to modeling, which uses both macroscopic performance measures and vehicle

trajectories in the calibration process. This approach balances the tradeoff between accurately capturing the vehicle trajectories and the characteristics of traffic flow, regardless of the length of the available trajectory data.

I–15 CASE STUDIES

The team conducted calibration and validation experiments for I–15 using Aimsun. The team performed data collection and modeling on four discrete sections of the highway. The lengths of these sections were 1,250 ft, 1,045 ft, 545 ft, and 440 ft. Warmup periods were 15 min for all calibration and validation runs, which satisfied warmup period recommendations discussed in Dowling, Skabardonis, and Alexiadis (2004). One should note that analysts may want to consider using a simulation warm-up time of at least twice the estimated travel time at free-flow conditions to traverse the length of the network. Prior to the calibration of driver behavior, the team first calibrated the simulation input demands to achieve a better matching of field-measured versus simulated throughput at key segments, described as "step zero" in chapter 4. The I.H.O.P. B.P.R. procedure was then executed as follows.

Step 1: Inputs

The team did not directly calibrate parameters in the Gipps model. Instead, they used parameters that affect headway and lane-changing directly, which were the chosen performance measures in this study. They decided to calibrate three car-following and two lane-changing parameters: reaction time, car-following aggressiveness, sensitivity factor deviation, lane-changing cooperation, and lane-changing aggressiveness. To make the problem more practical, the team identified ranges of values to consider for each calibration parameter based on their previous modeling experience. The values considered for the Gipps car-following model (Vasconcelos et al. 2014) are as follows:

- Reaction time: 0.85, 0.90, 0.95, 1.00, 1.05, and 1.10 s.
- Car-following aggressiveness: 0.0, −0.1, −0.2, −0.3, −0.4, −0.5.
- Sensitivity factor deviation: 0.00, 0.05, and 0.10.

The following lane-change model parameters and candidate values were selected for calibration:

- Lane-changing cooperation: 50, 60, 70, 80, 90, and 100.
- Lane-changing aggressiveness: 0, 10, 20, and 30.

The remaining parameter values were kept at their default values. These selections led to 156 candidate solutions.

Step 2: Heuristics

Chapter 4 recommended several viable options for the heuristic step. For this case study, the research team selected the DBF search method in step 2 to limit the amount of time needed for the calibration experiments. The team hypothesized that DBF search would perform faster than heuristics given the limited number of calibration parameters and the small parameter search spaces.

To limit the experiment to 156 candidate solutions, the team did not evaluate all possible combinations of the above values through simulation. Instead, the team first simulated all 108 possible combinations of car-following values while preserving the default lane-changing values. The team then retained the best car-following values, namely, reaction time 0.95 s, car-following aggressiveness −0.5, and sensitivity factor deviation 0, while testing all 24 combinations of lane-changing values. Because car-following aggressiveness = −0.4 also performed well during the first 108 simulations, the team again retested all 24 combinations of lane-changing values at a car-following aggressiveness of −0.4. The team performed 10 random number seed replications for each of the 156 combinations.

Step 3: Outputs

Chapter 4 addressed the rationale for selecting output measures. For trajectory-based calibration, the team used headways and lane numbers with a 50–50 relative importance weighting. For traditional calibration, the team used speed and throughput with a 50–50 weighting. The selection of 50–50 relative importance weightings was arbitrary for these case studies; future research could include sensitivity analysis of relative importance weighting to determine the best selection for traditional and trajectory-based calibration.

Step 4: Points

The fourth step in the proposed methodology is to choose the number of comparison points per full-set trajectory. The methodology does not adopt or require a specific number of points. The research team used 164 feet intervals between points for these case study experiments, with no specific rationale other than estimating what might achieve the right balance between practicality (e.g., computer run time) and robustness. Again, future research could include sensitivity analysis to determine the optimal number of comparison points per trajectory.

Step 5: Binning

The fifth step in the proposed methodology involves binning trajectories into specific groups to enable point-by-point comparisons of vehicle headways and lane numbers of sufficiently similar simulated and observed trajectories. The team wished to select enough bins to allow robust calibration results, but also wanted the number of bins to be small enough to make the experiments efficient and practical. The team defined 28 bins for the I–15 case study:

- One driver type.
- Two vehicle types (light-duty and heavy).
- Seven origins (on-ramp and six mainline lanes).
- Two destinations (off-ramp and mainline).

Step 6: Pairing

The team developed scripts to automate the trajectory pairing process. The team believed that pairing simulated and field-observed trajectories entering the study area at approximately the same time could produce robust calibration results. The team chose a 4-s and 200-ft threshold. Any simulated vehicle entering the study area within 4 s and 200 ft of a field-observed vehicle

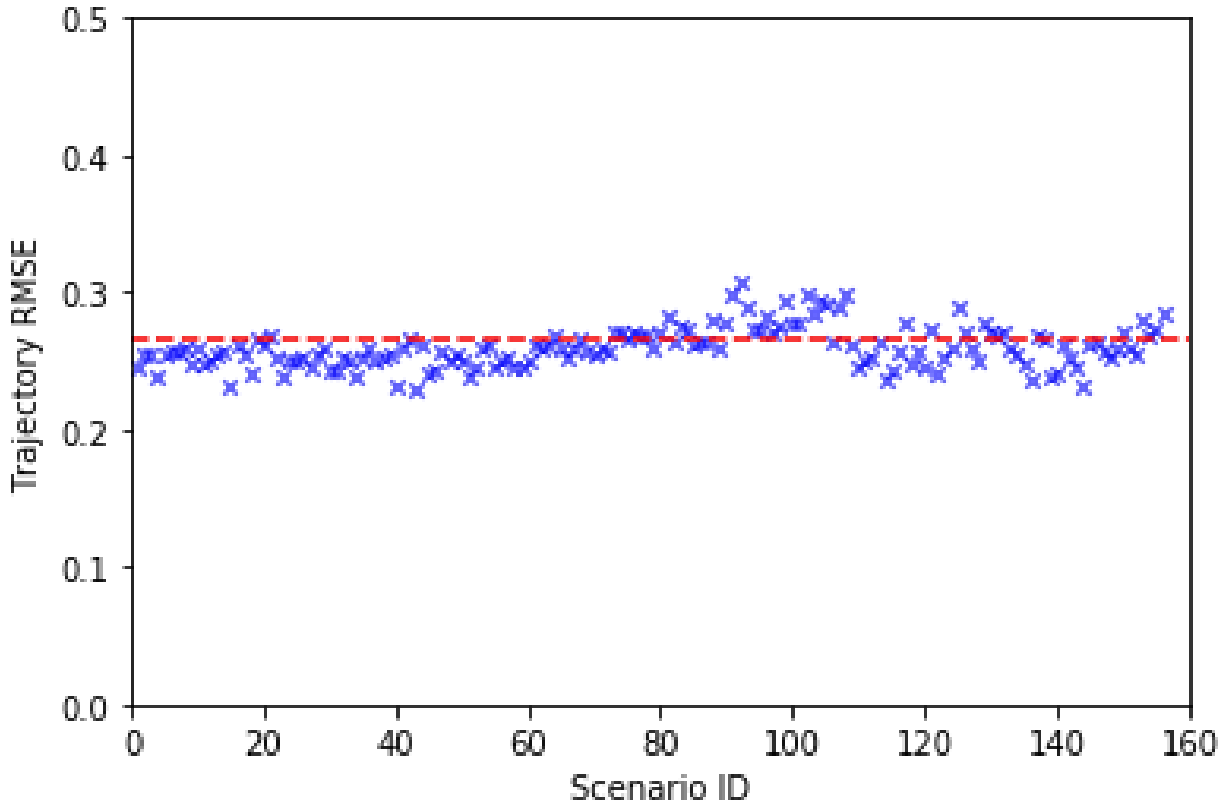
could be paired with that vehicle. To limit the amount of time required for calibration, the team limited the number of paired trajectories per bin to a maximum of 25.

The thresholds for pairing (e.g., space and time window) and the maximum number of paired trajectories per bin were decided based on engineering judgement. In the future, an analyst may want to consider conducting sensitivity analysis on these three parameters to determine the best selections for their data sample.

Step 7: RMSE

The team's literature review (chapter 2) indicated that RMSE is an effective goodness-of-fit measure for calibrating traffic simulation models and driver behavior models. For the trajectory-based RMSE, the team used a 50–50 relative weighting of headways and lane numbers to blend them into a single value. This allowed the RMSE value to reflect both car-following and lane-changing effects. For the traditional RMSE, the team used a 50–50 relative weighting of throughputs and speeds to blend them into a single value. The calculations to accomplish this are described in chapter 4. The team further obtained hybrid calibration via relative weightings of traditional and trajectory RMSE as described in chapter 4.

Figure 29 illustrates the trajectory-based RMSE values for I–15. The team selected a maximum normalized delta value of 1.0 for these experiments, such that the RMSEs in figure 29 were effectively constrained to a range of 0.0 to 1.0. The horizontal dashed line in figure 29 represents the benchmark RMSE value obtained after step zero of the proposed method. In step zero, traffic demand volumes were calibrated to achieve more accurate simulated throughputs. This simulation model was then used as the starting point for subsequent calibration of driver behavior.



Source: FHWA.

Figure 29. Scatterplot. I–15 trajectory RMSE.

Calibration Results

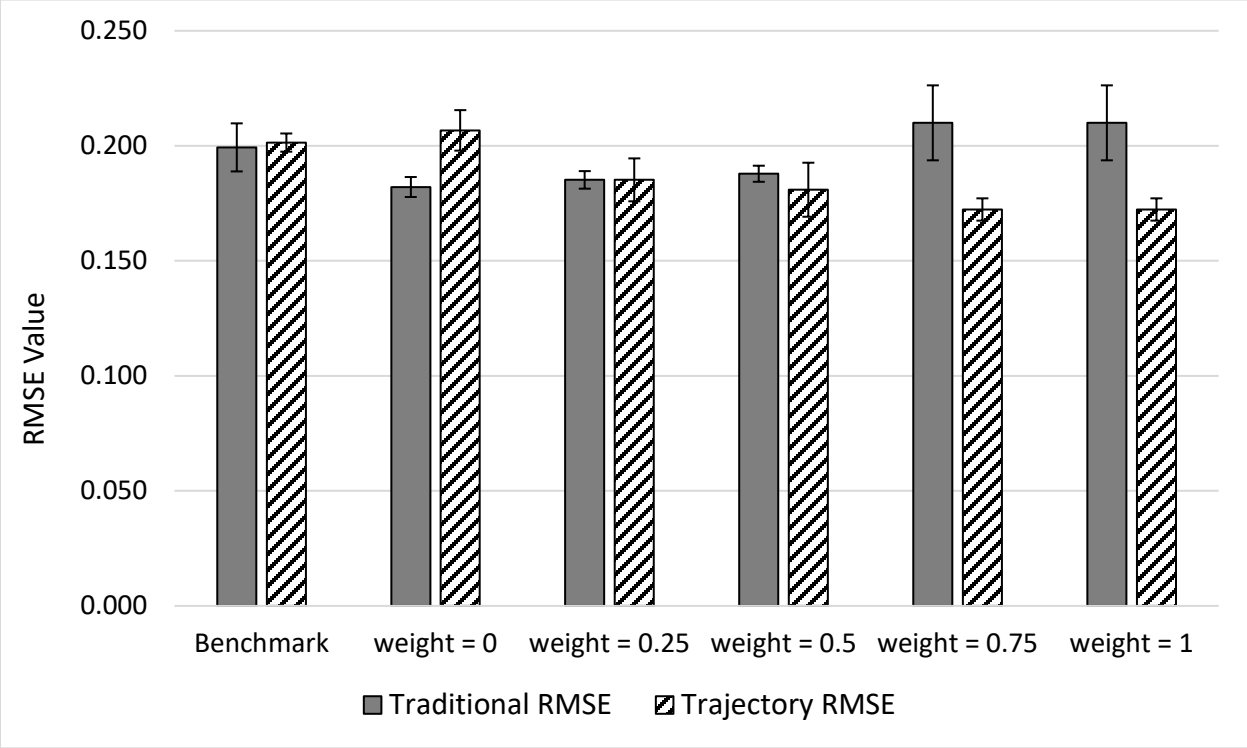
Figure 30 provides results for the I–15 network. To help with interpreting these graphs, please note that traditional calibration is indicated in these figures as "weight = 0," while a pure trajectory-based calibration of driver behavior is indicated in these figures as "weight = 1." The step zero model with calibrated throughputs and default driver behavior parameters is labeled "Benchmark." "Weight = 0.25," "weight = 0.5," and "weight = 0.75" are all hybrid calibration model results, where the relative importance of traditional performance measures (e.g., throughput and speed) and trajectory performance measures (e.g., lane ID and headway) varied. In weight = 0.5, traditional and trajectory performance measures were considered equally important. For weight = 0.25, the traditional performance measures were considered to be three times as important as the trajectory measures. Conversely, for weight = 0.75, trajectory performance measures were considered to be three times as important as the traditional performance measures. Hybrid RMSEs calculations are discussed in chapter 4. Only the best solution—that with the lowest RMSE out of 156 candidates—is shown for each weight, based on 10 random number seed replications. In this particular bar chart, the box-and-whisker notation shows the range of all 10 random number seed replications.

The overall results for the I–15 network showed that if trajectories are excluded from the calibration process, simulated car-following and lane-changing behaviors may not accurately reflect trajectories observed in the field data. This holds true even if aggregate measures have

good agreement with those observed in the field. The purely traditional calibration made trajectories less realistic than those obtained with default driving behavior parameters at I–15, even though the calibration process improved the traditional RMSE.

Like the previous case studies, the trajectory-based calibration (weight = 1) improves the accuracy of the trajectories—specifically, lane number and headway—compared with the benchmark model and the traditionally calibrated model (weight = 0). The traditional RMSE performance is lower, however, indicating that the average lane speed and throughput are reflective of what was observed in the data.

The symmetry of these results on a second microsimulation software platform again forces us to question which data collection effort produces a better representation of ground truth. In addition, these results imply that a hybrid calibration might produce the most reliable results for all types of measures.



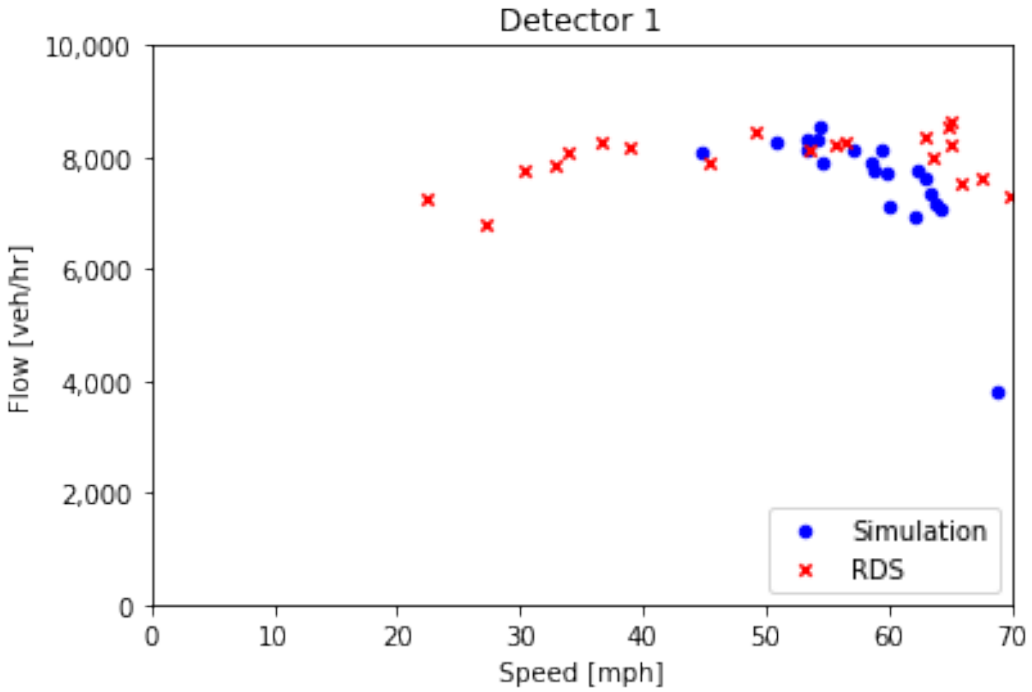
Source: FHWA.
Weight is the relative weighting of trajectory-to-traditional calibration.

Figure 30. Bar Chart. I–15 calibration results.

Traditional Validation Results

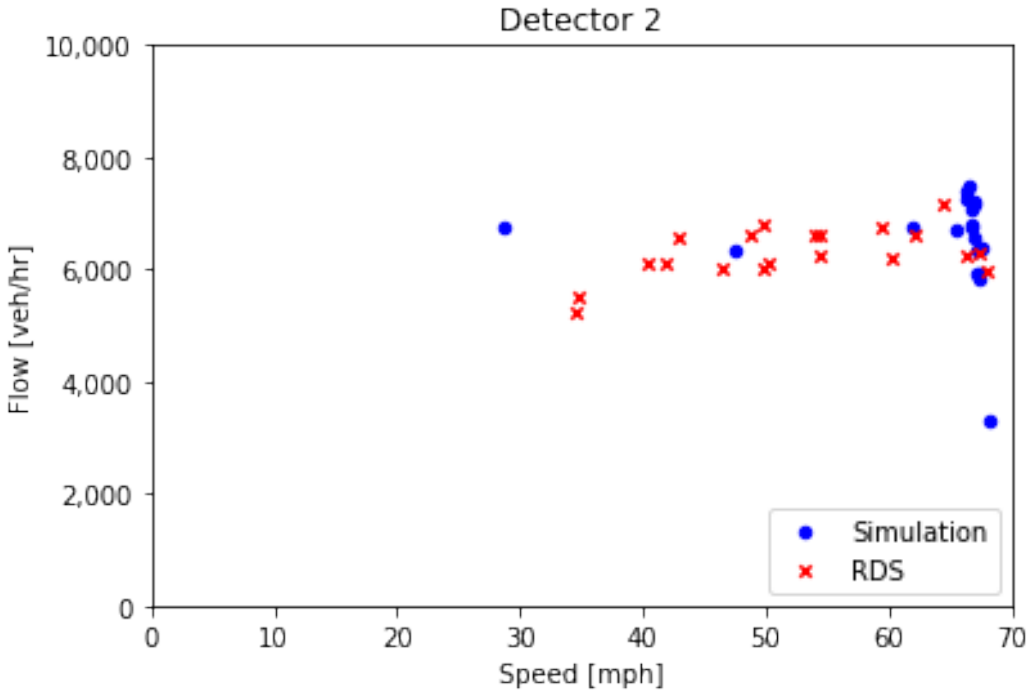
Following the calibration experiment, the research team performed validation by both traditional methods and trajectory-based methods. For traditional validation, the team constructed speed-flow diagrams for four detector locations, as seen in figure 31 through figure 34 for I–15. The

similarity of simulation and field data patterns, as shown by the points in the scatterplot, indicate a reasonable correlation between observed and simulated traffic flow conditions.



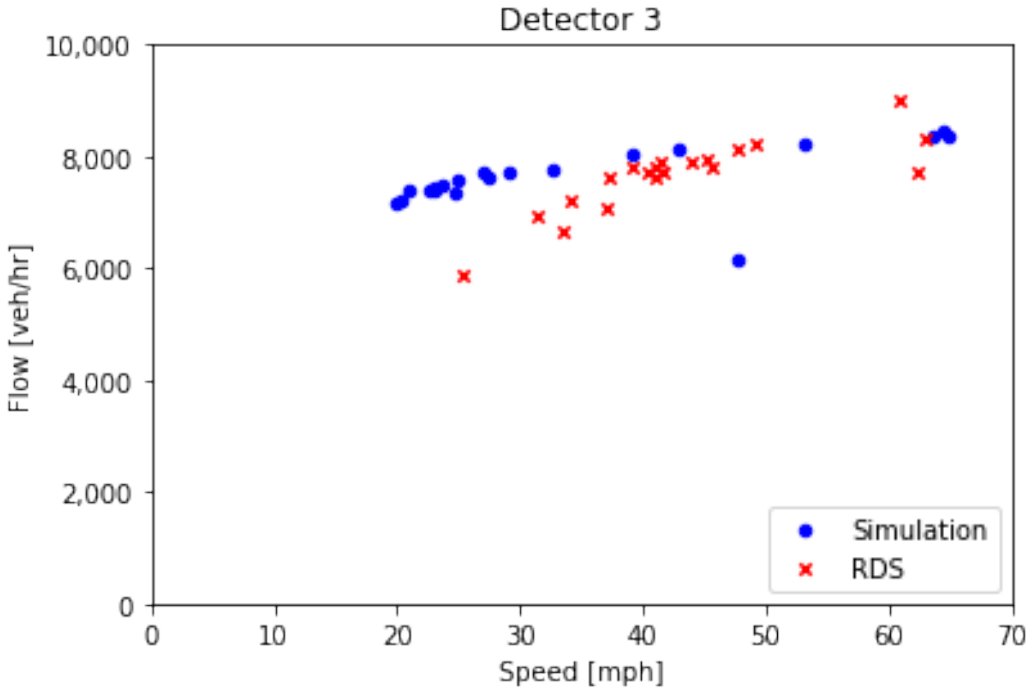
Source: FHWA.
RDS = real dataset.

Figure 31. Scatterplot. I–15 speed-flow diagram at detector 1.



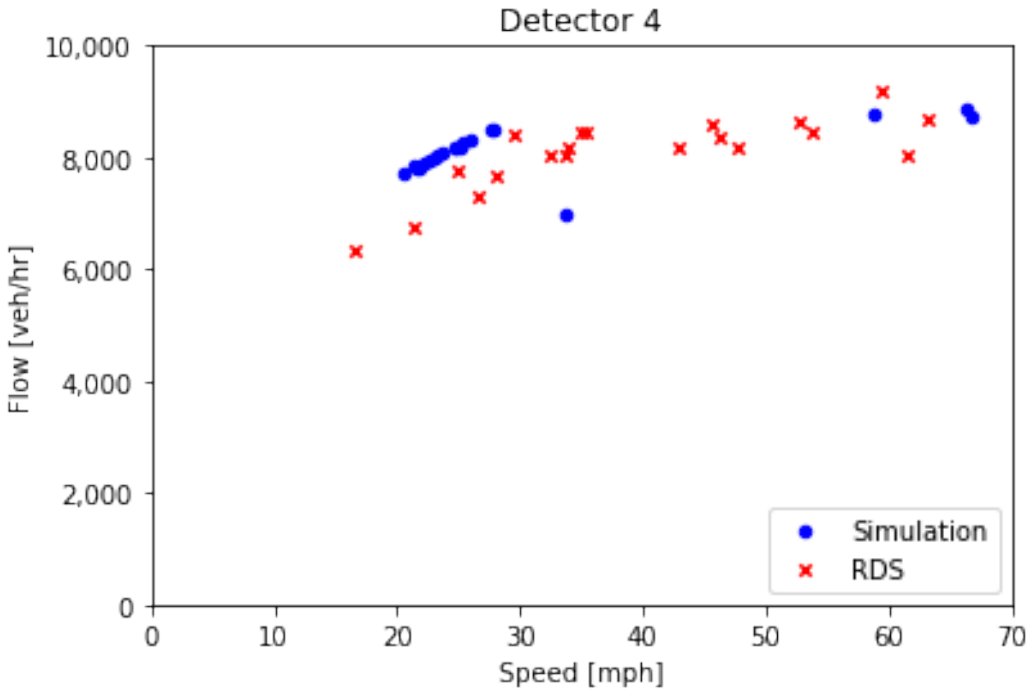
Source: FHWA.
RDS = real dataset.

Figure 32. Scatterplot. I–15 speed-flow diagram at detector 2.



Source: FHWA.
RDS = real dataset.

Figure 33. Scatterplot. I–15 speed-flow diagram at detector 3.



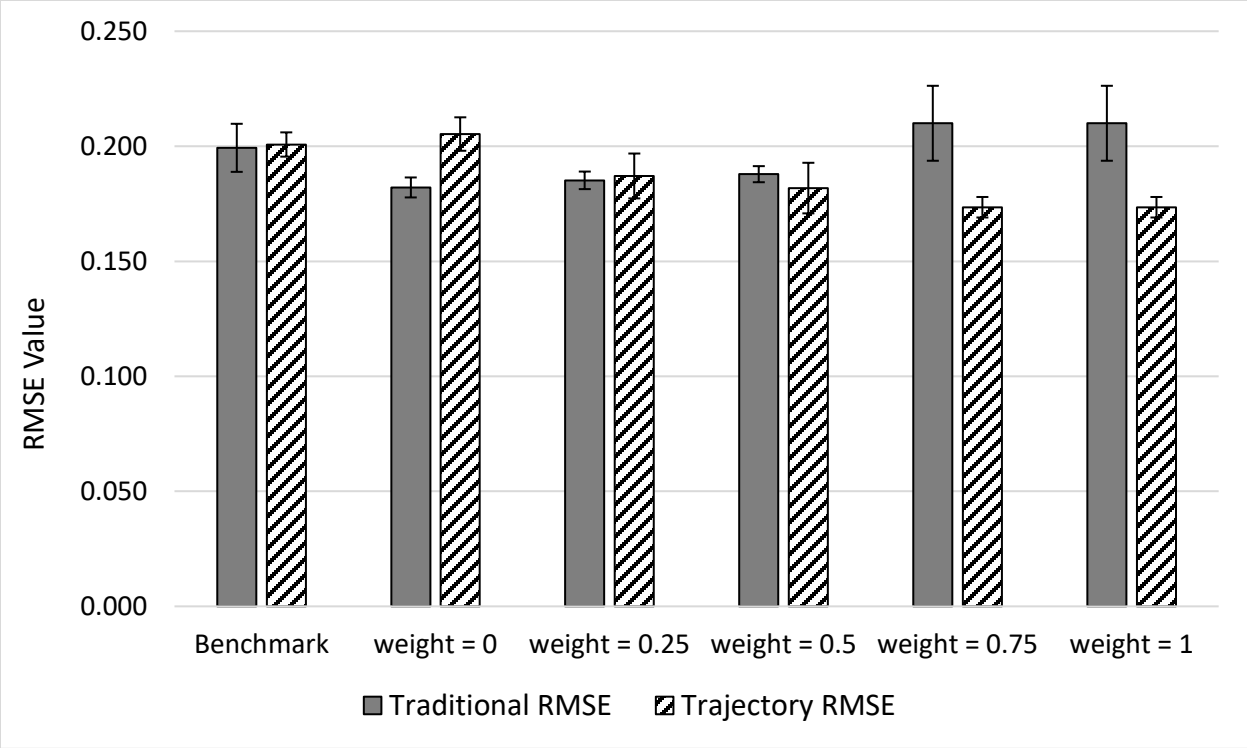
Source: FHWA.
RDS = real dataset.

Figure 34. Scatterplot. I–15 speed-flow diagram at detector 4.

Trajectory Validation Results

As described in chapter 4, for trajectory-based validation, the datasets were divided into two separate groups: calibration data and validation data. Following calibration, the calibrated model was evaluated according to validation (holdout) data. The decision of how to proportion data into calibration and validation bins was more art than science, as no hard and fast rule exists in the literature. The team decided to apply 80 percent of the trajectories toward calibration and 20 percent toward validation primarily to achieve a sufficient number of trajectories in each bin for calibration.

The following validation exercise tested five calibrated models. These five weights represent the hybrid calibration at three different relative weights, the model calibrated considering only traditional data, and the model calibrated considering only trajectory data. The resulting RMSE values. The team intended to demonstrate that the calibrated simulation models could effectively predict field-observed trajectories from the validation dataset. The resulting RMSE values are graphed within figure 35 for I–15.



Source: FHWA.
Note: weight is the relative weighting of trajectory-to-traditional calibration.

Figure 35. Graph. I–15 RMSEs from the validation data.

Figure 35 shows the same trend of traditional and trajectory RMSE exhibited earlier by figure 30, which verifies that using trajectory performance measures in the calibration process improves the accuracy of the trajectories without overfitting to the calibration data. The purely trajectory-based calibration methodology produced trajectories that were most reflective of what was observed in the data with respect to lane numbers and headways. The improvement in trajectory RMSE from the benchmark model to purely trajectory-based calibration (weight = 1) using the validation dataset is 14 percent, which is the same improvement from the calibration dataset. The trajectory calibration methodology produced the highest observed traditional RMSE, however, indicating that the model's average lane speed and throughputs were less similar to what was observed in the collected dataset.

As with the other case studies, the first validation attempt failed. In the first attempt at validation, the change in trajectory RMSE from the benchmark model to the purely trajectory-based calibrated model was −7 percent. After this initial failed validation, the team observed that the proportion of trajectories in each bin was highly unequal between the calibration and validation datasets. This highlights the importance of ensuring that during step 5 of I.H.O.P. B.P.R, the proportion of trajectories in each bin is consistent with typical traffic. It ensures that the calibrated model can be truly realistic and truly predictive of future conditions.

OVERALL RESULTS: PURE TRAJECTORY-BASED CALIBRATION

To gain further insight, the research team generated some overall summary results accounting for all four sites. Table 8 focuses on the potential benefits of a pure trajectory-based calibration (i.e., weight = 1), relative to a pure traditional calibration (i.e., weight = 0) or a benchmark model (calibrated demands, uncalibrated driver behavior), in terms of trajectory accuracy and realism ($RMSE_{trajectory}$) using the calibration data as the test data (corresponding with figure 16 through figure 18). The first column is the difference between the $RMSE_{trajectory}$ of the traditional calibration (weight = 0) relative to the $RMSE_{trajectory}$ of the benchmark model. This implies that traditional calibration (i.e., minimizing the difference between the simulated and observed lane-specific throughput and speed) does not reliably improve the realism of vehicle trajectories compared to the benchmark model. Interestingly, for two of the four models, the benchmark model using uncalibrated driving behavior parameters produced trajectories that better matched field data compared to the model calibrated using lane-specific throughput and speed measurements. This is a major cause for concern to the research team: using throughput and speed as calibration data alone may not produce the robustly calibrated models that engineers expect.

The second column represents the difference between $RMSE_{trajectory}$ of the trajectory calibration (weight = 1) relative to the $RMSE_{trajectory}$ of the benchmark model. The second column implies that pure trajectory-based calibration (i.e., minimizing the difference between the observed and simulated headways and lane ID) produces large improvements in trajectory accuracy compared to the benchmark model.

The third column is the difference between the $RMSE_{trajectory}$ of the trajectory-based calibration method (weight = 1) and the traditional calibration method (weight = 0). The third column highlights the sizeable benefit to trajectory accuracy ($RMSE_{trajectory}$) when performing trajectory-based calibration (weight = 1) instead of traditional calibration (weight = 0).

Table 8. Impacts of calibration on trajectories from the calibration dataset.

Interstate	Traditional (%)	Trajectory (%)	Benefit (%)
I–95	−13	+17	+30
I–75	+20	+59	+39
I–270	+5	+21	+16
I–15	−3	+14	+17

Table 9 reveals a corresponding set of results according to the validation dataset, which is data that were not used for calibration (set aside during step 5 of the I.H.O.P. B.P.R procedure). The first column of results implies that traditional calibration (i.e., minimizing the difference between the simulated and observed throughput and speed) has mixed impact on the realism of vehicle trajectories ($RMSE_{trajectory}$) relative to the benchmark model. This provides further evidence that traditional calibration methods cannot be trusted to produce accurately modeled vehicle trajectories.

The second column implies that the trajectory-based calibration method consistently produced improved simulated trajectory predictions relative to the benchmark model. This observation was true across four different case studies conducted in two different microsimulation platforms.

Finally, the third column highlights the potential benefits of trajectory-based calibration (weight = 1) relative to traditional calibration (weight = 0): an increase in trajectory accuracy for every case study conducted. These benefits are somewhat smaller than the table 8 benefits. However, the research team expected this, because the validation data were not used in the calibration process. Thus, table 9 illustrates the more-likely outcomes of using the calibrated models to make predictions, because they were evaluated against a separate validation dataset. Although the team does not have access to something comparable to a before-and-after (current and future) dataset, table 9 demonstrates that the trajectory-based calibration method can capture generalizable trends in driver behavior that match validation data (not used in the calibration procedure) quite well and should be explored further in future research.

Table 9. Impacts of calibration on trajectories from the validation dataset.

Interstate	Traditional (%)	Trajectory (%)	Benefit (%)
I–95	−12	+9	+21
I–75	+7	+42	+35
I–270	+3	+10	+7
I–15	+1	+14	+15

Table 10 examines the impacts of calibration on traditional measures. The first column is the difference between the $RMSE_{traditional}$ of the traditional calibration (weight = 0) method relative to the $RMSE_{traditional}$ of the benchmark model. The first column of results implies that traditional calibration (i.e., minimizing the difference between the simulated and observed lane-specific throughput and speed) improves the realism of traditional performance measures, although the magnitude of that improvement varies across networks.

The second column is the difference between the $RMSE_{traditional}$ of the trajectory calibration (weight = 1) relative to the $RMSE_{traditional}$ of the benchmark model. The second column implies that pure trajectory-based calibration does not reliably improve the simulated aggregate traffic flow performance measures. At two of the four sites, the reduction in accuracy was marginal. Moreover, at the site where helicopters were used for trajectory data collection, the accuracy of the aggregate traffic flow performance measures increased substantially (relative to the models calibrated with drone data), even though traditional measures were not explicitly used by the calibration objective function.

The third column highlights the advantage of traditional calibration (i.e., minimizing the difference between the simulated and observed throughput and speed) over trajectory-based calibration (i.e., minimizing the difference between the simulation and observed headway and lane IDs) in terms of aggregate traffic flow performance measure accuracy. As shown in table 10, the site where helicopters were used for data collection instead of drones (which produced significantly longer trajectories), the trajectory calibration method performed nearly as well as the traditional calibration method at replicating aggregate traffic flow performance measures, even though traditional measures (e.g., throughput, speed) were not considered as part of the

objective function. At 3 of the 4 sites, the traditional calibration method (where the objective was to explicitly match segment-level throughput and speed) produced considerably more accurate simulated aggregate traffic flow performance measures. However, as observed in table 8 and table 9, this came at the significant expense of accurate individual vehicle trajectories. This motivated the study of hybrid calibrated models.

Table 10. Impacts of calibration on traditional measures.

Interstate	Traditional (%)	Trajectory (%)	Benefit (%)
I–95	+10	−7	+17
I–75	+25	+20	+5
I–270	+7	−14	+21
I–15	+9	−5	+14

OVERALL RESULTS: HYBRID CALIBRATION

The research team developed the hybrid calibration method to enable analysts to include both trajectories and aggregated traffic-flow performance metrics in the calibration process. The section discusses the benefits of conducting a hybrid calibration process in microsimulation models. In the absence of guidance, this section will analyze the results of the 50-50 hybrid calibrated model, which considered trajectories and aggregated traffic-flow performance metrics as equally important in the calibration process.

Table 11 details the impact of the hybrid calibration method (weight = 0.5) on the accuracy of simulated trajectories ($RMSE_{traditional}$) versus the benchmark model and the pure traditional calibrated model (weight = 0). The second column indicates that the hybrid calibration method more accurately simulated throughput and speed compared to a model using default parameters. As shown in column 3, the site where helicopters were used for data collection instead of drones (which produced significantly longer trajectories), the trajectory calibration method performed nearly as well as the traditional calibration method at replicating aggregate traffic flow performance measures, even though traditional measures (e.g., throughput, speed) were not considered as part of the objective function. The remaining three sites suggest that the traditional calibration method was slightly more accurate in simulating throughput and speed that match field data compared to the hybrid calibration method. However, as will be shown in table 12 and table 13, this slight reduction in the traditional RMSE is a small price to pay in return for a significant increase in trajectory accuracy.

Table 11. Impacts of trusted hybrid model on traditional measures.

Interstate	Versus Benchmark	Versus Traditionally Calibrated
Interstate 95	+3%	-8%
Interstate 75	+24%	-1%
Interstate 270	+6%	-2%
Interstate 15	+6%	-3%

Table 12 details the impact of the hybrid calibration method (weight = 0.5) on the accuracy of simulated trajectories ($RMSE_{trajectory}$) versus the benchmark model and the pure trajectory

calibrated model (weight = 1) using calibration data as the observed data. As one can see, using both trajectories and aggregate traffic flow performance metrics in the calibration process through a unique hybrid calibration method produces models that are much more robust, and accurately simulate vehicle trajectories that match what was observed in field data.

Table 12. Impacts of trusted hybrid model on trajectories from the calibration dataset.

Interstate	Versus Benchmark	Versus Traditionally Calibrated
Interstate 95	+13%	+23%
Interstate 75	+58%	+48%
Interstate 270	+11%	+7%
Interstate 15	+10%	+13%

Table 13 details the impact of the hybrid calibration method (weight = 0.5) on the accuracy of simulated trajectories ($RMSE_{trajectory}$) versus the benchmark model and the pure trajectory calibrated model (weight = 1) using the validation data as the observed data. This table provides further evidence that a hybrid approach to calibration performs much better at producing accurately simulated trajectories compared to models that were not calibrated (benchmark) or models that were calibrated only with macroscopic data. Moreover, when comparing table 12 and table 13, one does not observe a significant decrease in accuracy despite using holdout data as the observed data; this strongly suggests that the hybrid calibration method produces models that are not overfit to the calibration data, and captures generalizable trends in driver behavior.

Table 13. Impacts of trusted hybrid model on trajectories from the validation dataset.

Interstate	Versus Benchmark	Versus Traditionally Calibrated
Interstate 95	+6%	+16%
Interstate 75	+52%	+49%
Interstate 270	+3%	+0%
Interstate 15	+9%	+11%

Thus, the authors of this paper believe that a hybrid calibration approach—using both trajectories and lane-specific throughput and speed—for the calibration of driver behavior models in microsimulation provides a substantial improvement over current best practices. This is because the hybrid calibration approach produces models whose simulated trajectories match field observations much more accurately, without sacrificing the accuracy of macroscopic traffic flow performance metrics.

MODELING IMPLICATIONS

Based on the research documented in this chapter, the authors make the following general calibration recommendations for calibrating microsimulation models:

- The driver behavior parameters of microsimulation models should always be calibrated based on real-world data. This research adds to a body of literature that recognizes the importance of model calibration and suggests that default driver behavior parameters are

not sufficient for capturing real-world driver behavior in microsimulation model analyses.

- Moreover, when appropriate data are unavailable, the authors of this paper caution practitioners against choosing high-resolution microsimulation models when conducting modeling analysis projects. Although the visualizations produced using microsimulation models are helpful for communication of project impacts, it is our professional responsibility to ensure the models are calibrated appropriately to reflect local conditions without being overfit to the data.

- The authors of this paper believe that a hybrid calibration approach—using both trajectories and lane-specific throughput and speed—for the calibration of driver behavior models in microsimulation provides a substantial improvement over current best practices. The hybrid calibration approach produces models whose simulated trajectories more accurately match field observations without sacrificing the accuracy of macroscopic traffic flow performance metrics.

- When calibrating driver behaviors such as car-following and lane-changing, it is preferable to incorporate lane-specific measures instead of segment-specific measures.

- The longer trajectories (collected via helicopter) are more desirable than data collected via individual drones. Calibrations completed using longer trajectories (>1.2 mi) outperformed the shorter trajectories (800 ft) in terms of both simulated trajectories and the simulated lane-specific throughput and speed. Moreover, the model calibrated using longer trajectories performed just as well at capturing lane-specific throughput and speed as the model calibrated using lane-specific throughput and speed. This may suggest that with longer trajectories, a purely trajectory-based calibration method may be sufficiently reliable, but future research is needed on the topic.

- The validation experimental results highlight the importance of validation in identifying problems in the calibrated models, fixing those problems, and making the calibrated models more robust and predictive.

CHAPTER 6. CONCLUSION

Because of recent improvements in data collection and processing technologies for vehicle trajectories, trajectory-based calibration of microsimulation models is now a more feasible and practical option for transportation agencies to consider. In response, this project produced a methodology that explicitly incorporates vehicle trajectories into the calibration process (chapter 4). This project accomplished a comprehensive data collection and data processing effort at four real-world congested freeway sites (chapter 3). The data collection vendors delivered nearly 3 TB of helicopter video footage data and approximately 75 GB of collected drone data. The traditional corroborative data were much smaller in size compared with the drone video footage data. The project team will provide online public access to the raw video footage, post-processed trajectory data, and the traditional corroborative data. The team used this inventory to test the new, trajectory-based calibration methodology in two microsimulation software programs (chapter 5). The project team developed several scripts to automate the trajectory-based calibration methodology; the scripts are available online and are described in the appendix of this document (Github, n.d.-b).

MODEL CALIBRATION IMPLICATIONS

The research team's goal was to make this methodology as practical and straightforward as possible. The resulting seven-step methodology can be remembered through the acronym I.H.O.P. B.P.R. The first four steps—inputs, heuristics, outputs, and points—are user choices, whereas the last three steps—binning, pairing, and RMSE—are iterative processes that can be automated through scripting. Chapter 4 provides an overview of the developed methodology, and chapter 5 presents four case studies applying the methodology for model calibration.

The foremost practical challenges of the new methodology appear to be in the automated post-processing, binning, and numeric comparison of trajectory data. The research team developed add-on scripts to automate these data processing steps. This project report, coupled with the available scripts, may inspire early adopters to try trajectory-based calibration for the first time. However, to achieve widespread adoption and cost-effectiveness, the same process may need to be streamlined by software developers who can provide user-friendly, interactive apps that implement the methodology more efficiently.

The experimental results from this project imply that data from traditional calibration methods (e.g., average lane speed, throughput) cannot be trusted to accurately predict vehicle trajectories (e.g., lane number, headway) even though they are replicating traditional performance measures well. This has significant implications for how practitioners think about calibrating their models.

Interestingly, a calibration method that only used vehicle trajectories for calibration (i.e., the methodology produced by this research effort) also may not be the best solution. As observed through the case studies, the trajectory calibration methodology produced models that better match vehicle trajectories, but do not always match macroscopic performance measures as well as traditionally calibrated models. The exception to this observation was at the site where trajectories were collected by helicopters, instead of drones. The model calibrated using longer trajectories performed just as well at capturing lane-specific throughput and speed as the model

calibrated using lane-specific throughput and speed. This may suggest that with longer trajectories, a purely trajectory-based calibration method may be sufficiently reliable, but future research is needed on the topic.

There exists an opportunity to calibrate a model considering both trajectory and traditional data and performance measures. This is the hybrid model discussed in chapter 4. Case studies documented in chapter 5 demonstrate that the hybrid model does not typically identify the best (i.e., lowest) trajectory or traditional RMSE. However, the hybrid calibration method does a much better job of balancing the need for accurate trajectories (i.e., headways, lane numbers) and macroscopic traffic performance measures (i.e., average lane speed, throughput) than either methodology that excludes the other data type (i.e., purely trajectory-based or purely traditional calibration).

Based on the documented research, the authors of this report make the following general recommendations:

- The driver behavior parameters of microsimulation models should always be calibrated based on real-world data. This research adds to a body of literature that recognizes the importance of model calibration and suggests that default driver behavior parameters are not sufficient for capturing real-world driver behavior in microsimulation model analyses.

- Moreover, when appropriate data are unavailable, the authors of this paper caution practitioners against choosing high-resolution microsimulation models when conducting modeling analysis projects. Although the visualizations produced using microsimulation models are helpful for communication of project impacts, it is our professional responsibility to ensure the models are calibrated appropriately to reflect local conditions without being overfit to the data.

- The authors of this paper believe that a hybrid calibration approach—using both trajectories and lane-specific throughput and speed—for the calibration of driver behavior models in microsimulation provides a substantial improvement over current best practices. The hybrid calibration approach produces models whose simulated trajectories more accurately match field observations without sacrificing the accuracy of macroscopic traffic flow performance metrics.

- When calibrating driver behaviors such as car-following and lane-changing, it is preferable to incorporate lane-specific measures instead of segment-specific measures.

- The longer trajectories (collected via helicopter) are more desirable than data collected via individual drones. Calibrations completed using longer trajectories (>1.2 mi) outperformed the shorter trajectories (800 ft) in terms of both simulated trajectories and the simulated lane-specific throughput and speed. Moreover, the model calibrated using longer trajectories performed just as well at capturing lane-specific throughput and speed as the model calibrated using lane-specific throughput and speed. This may suggest that with longer trajectories, a purely trajectory-based calibration method may be sufficiently reliable, but future research is needed on the topic.

- The validation experimental results highlight the importance of validation in identifying problems in the calibrated models, fixing those problems, and making the calibrated models more robust and predictive.

COST-EFFECTIVENESS OF THE NEW METHOD

Over the years, agencies and their consultants have learned how to manage and plan for the level of effort associated with traditional calibration methodologies. The same cannot yet be said for trajectory-based calibration. The effort to conduct trajectory-based calibration, especially without user-friendly apps within popular microsimulation packages, is larger than that required for traditional calibration simply because of inexperience with the new methodology. Moreover, only four case studies have been conducted using this methodology; thus, the model improvement benefits are somewhat uncertain. Therefore, user-friendly apps will be needed to lessen the risks of pursuing trajectory-based calibration and achieve wider adoption of trajectory-based calibration.

The approach to data collection is another factor in cost-effectiveness. Although helicopter data collection is more expensive than drone data collection, trajectory-based calibration at I–75 produced a simulation model having significantly better predictive ability compared with the sites calibrated with drone data, as demonstrated by the validation experiments. The collected trajectory data by helicopter at I–75 produced the largest sample size of trajectories of any site and the longest length of full-set trajectories (1.2 mi). By contrast, the team deployed multiple drones at the other sites to sample different stages of space-time congestion propagation; these data samples are much shorter in time and space, only collecting data for about 15 min at a time and only capturing 800 ft of each trajectory. Based on this research, the trajectory-based calibration methodology may ultimately perform better using longer trajectories than what can currently be captured by a single drone. This has a major cost implication for the methodology. Some solutions include flying multiple, simultaneous drones that are synced to enable the trajectories to be stitched together during post-processing. Another approach may involve obtaining trajectories from commercial probe data providers, although probe data may bring its own set of challenges (e.g., GPS accuracy, sample size of probe vehicles, availability of trajectory data through commercial providers).

TAKEAWAYS FOR TRANSPORTATION AGENCIES

Analysts and agencies may consider certain tradeoffs when contemplating their approach to calibration. It may be helpful for agencies to migrate toward incorporating finer-grained performance measures and gradually adopt such measures for inclusion within the calibration process for microsimulation models. This is because improving the realism of driver behavior modeling may be one of the best available ways to improve the predictive ability of microsimulation models. It stands to reason that fine-grained output performance measures (e.g., headway, lane ID), as opposed to coarse segment-based measures (e.g., segment throughput and speed), are best suited for determining the best input model parameters to control fine-grained driver behaviors (i.e., car-following and lane-changing). A hybrid blend of measures may prove to be an excellent compromise, but additional research is necessary.

In addition, agencies should strongly consider validation. In the case studies from this project, validation helped to find and fix problems in the calibrated models. The separate research teams both found problems in their calibrated models despite working independently, working with different software, and analyzing different freeway sites. Without the validation effort, these problems never would have been discovered. In real-world projects, this could mean using problematic models for important future predictions because the calibration results look favorable and validation was never completed. Even when calibration results look favorable, there could be unknown biases lurking inside the model. After finding and fixing problems revealed by validation, the revised calibrated model may be more robust and predictive than before.

Finally, the authors recognize that trajectory data are often not readily available. This report strongly suggests that it is time to start exploring methods to collect trajectory data more ubiquitously to inform microsimulation model calibration. In the interim, the authors do not want to discourage agencies from using traditional calibration methods when trajectory data are unavailable. As was documented in the case studies, the traditional calibration method significantly outperformed the benchmark model, in terms of replicating traditional performance measures from the field. However, all four case studies suggest that traditional calibration methods do not appear to produce accurate car-following behaviors, lane-changing behaviors, or simulated vehicle trajectories. As a result, the ability of traditionally calibrated microsimulation models to predict future conditions, or perform alternatives analyses, may be compromised. Additional research is recommended in this area to fully understand the implications of these observations.

In summary, there are additional costs required to collect trajectory-level data and set up the automated calibration procedure. The resulting microsimulation models, however, may provide more realistic predictions and avoid overfitting to the traditional measures.

FUTURE RESEARCH AND DEVELOPMENT

The research documented in this report represent a first step toward maturing methods that use trajectories for microsimulation model calibration. As with most exploratory research, the experiments in this project raised the following additional issues and questions that future research could potentially examine:

- The extent to which unrealistic driver behaviors are compromising the accuracy of high-profile performance measures is unclear.
- In this project's trajectory-based calibration experiments, the relative importance of car-following and lane-changing was always set to 50–50. Would the process be more robust under a different relative weighting? If so, could the optimal relative weightings be predicted based on traffic network conditions?
- Can lengthy and accurate trajectories be formed by stitching together shorter trajectories (collected by multiple drones) in a cost-effective manner?

- Alternatively, can lengthy and accurate trajectories be obtained through probe data, and can they be successfully applied toward calibrating microsimulation models?

- In this project's trajectory-based calibration experiments, the research team achieved favorable results by applying 80 percent of the trajectory data toward calibration and 20 percent of the trajectory data toward validation. Would this distribution of data work well for other sites? If not, could the appropriate distribution of data be predicted according to site characteristics or traffic characteristics?

- How reliable are the available data collection and data processing methods for trajectory data compared with traditional data?

- To what extent would it be helpful to calibrate multiple driver behavior models for multiple congestion regimes (e.g., below capacity, near capacity, at capacity, above capacity)? The benefit is uncertain, because driver behavior may have minimal impact on overall mobility at the below-capacity and above-capacity regimes.

APPENDIX A. VISSIM SCRIPTS

The following seven scripts are developed and available online (Github, n.d.-b):

1. WGS_to_cartesian—this Python® code converts the WGS84 coordinates (longitudes and latitudes) to Vissim Cartesian coordinates. This conversion code uses several reference points. The WGS84 and Vissim coordinates for these reference points are derived and entered in the code manually.

2. Enumerate—this code enumerates through all parameter settings and calls the Vissim_eval code for each parameter set. It also reads the location and time limits of the trajectory data and passes it to the Vissim_eval code, so that Vissim_eval collects simulated trajectories only within the same location and time limits. Vissim_eval returns the macroscopic measures (throughput and speed) and trajectories to Enumerate. Enumerate then saves these outcomes.

3. Vissim_eval—this code is a main simulator component. Inputs to this code are parameter set, time, and location limits of trajectory collection, and several other simulation parameters. With these inputs, Vissim_eval calls Vissim, sets the Vissim simulation parameters, runs the simulation, and collects microscopic and macroscopic measures.

4. MOE_eval—this code is used to post-process and evaluate the simulation outcomes. First, macroscopic RMSEs are calculated by comparing the field and simulation throughput and speed data. These RMSEs are then normalized to fit within a range between 0 and 1. Second, microscopic RMSEs are derived with the following procedures:

    a. Reads the field-collected trajectories.
    b. Reformats the field data into a structure similar to NGSIM dataset.
    c. Combines all field data zones into one big table and makes vehicle identifications (ID) unique.
    d. Calculates time headways for the field data.
    e. Defines a time headway cutoff and categorizes vehicles in the field data into "conservative" and "aggressive."
    f. Gets a sample of vehicles in the simulated trajectory dataset for each bin (using "sampling" code).
    g. Finds the corresponding data points in the field data (using "pairing" code).
    h. Calculates the headway and lane number errors for each pair of vehicles at each bin.
    i. Calculates microscopic RMSE for each bin and then returns the average of all bins as the final microscopic RMSE.
    j. Finally, MOE_eval saves the macroscopic and microscopic RMSE values.

5. Sampling—this script gets the sample size and headway cutoff values as inputs and returns a sample of simulated trajectories for each bin.

6. Pairing—this script gets the field and simulated trajectories, simulated trajectories sample indices (outputs of "sampling" code), and time and location tolerances parameters as

inputs and returns the corresponding field trajectories indices (similar to the format of simulated trajectories sample indices). The corresponding field trajectories shall fall within the temporal and spatial tolerance parameters.

7. Plots—this code plots the graphics.

APPENDIX B. AIMSUN SCRIPTS

- The model is developed in Aimsun version 8.4, and as vehicle headways can be obtained using an application programming interface (API), the license should include API as well.

- There is one Aimsun model (.ang) and two databases (sqlite). As mentioned in the report, the team ran two sets of trajectory scenarios and the regular Aimsun outputs of these scenarios (not vehicle trajectories) are in separate databases. The outputs of scenarios 1-108 is in "Trajectory_Base_Output_1_108.sqlite" and the outputs of scenarios 109-156 is in "Trajectory_Base_Output_109_156.sqlite". However, for the second set the scenario IDs in database starts from 1. Therefore, the IDs of scenarios 109-156 are 1-48 in the database Trajectory_Base_Output_109_156.sqlite. To retrieve Aimsun outputs from the existing databases the corresponding database should be linked in the "Outputs to Generate" tab of scenario.

- Trajectory outputs are saved as .csv files with the ID of the scenario. This is how the API (included in the material) saves the output.

- To get trajectory outputs when running any new scenario, the "Headway_Output.py" API should be linked as an external API in "Aimsun Next APIs" tab of scenario. Also, "AAPI.py" library should be saved in the same directory as "Headway_Output.py". Then after running a scenario, a .csv file of trajectory outputs will be saved in the same directory.

- There are two required files to run the model. First, the path_assignment file (included in the material), which should be linked in the "Main" tab of scenario. Second, the ramp metering input files (saved in SDRMS folder in the material), which should be linked in "Aimsun Next APIs" tab of scenario. By double clicking on the existing SDRMS API, you can update the directory of the relevant input files.

- To be able to set up trajectory scenarios easily the team used a script, "SensitivityAnalysis" (Aimsun id: 21348132), which is already included in the model. This script runs from a replication. Before running the replication, the calibration parameters and their ranges, in addition to section IDs (if any section-level calibration parameter such as lane-changing cooperation is used), should be updated. Then you can right click on the replication and select the script. Depending on the parameters and the ranges, several scenarios will be run.

- The "Trajectory_RMSE.py" script requires the coordinate of the study locations from .csv files, which are in folder "Road_Geometry".

- To get XY coordinates in vehicle trajectories, the Aimsun model's unit should be set to metric (not English).

ACKNOWLEDGMENTS

The authors thank members of the traffic analysis and simulation pooled fund study for making this study possible. The authors also thank the subject matter experts for their helpful ideas, insights, and guidance during the project.

REFERENCES

Anguita, D., L. Ghelardoni, A. Ghio, L. Oneto, and S. Ridella. 2012."The 'K' in K-fold Cross Validation." Bruges, Belgium: In *Proceedings of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. Bruges, Belgium: ESANN.

Antoniou, C., J. Barcelò, M. Brackstone, H. Celikoglu, B. Ciuffo, V. Punzo, P. Sykes, T. Toledo, P. Vortisch, and P. Wagner. 2014. *Traffic Simulation: Case for Guidelines*. Italy: European Commission, Joint Research Centre.

Banks, E., S. J. Cook, G. Fredrick, S. Gill, J. S. Gray, T. Larue, J. L. Milton et al. 2018. *Successful Approaches for the Use of Unmanned Aerial Systems by Surface Transportation Agencies*. NCHRP Project 20-68A. Washington, DC: National Cooperative Highway Research Program.

Barceló, J. 2010. *Fundamentals of Traffic Simulation (International Series in Operations Research and Management Science, 145)*. New York: Springer.

Berthaume, A. L., R.M. James, B. E. Hammit, C. Foreman, and C. L. Melson. 2018. "Variations in Driver Behavior: An Analysis of Car-Following Behavior Heterogeneity as a Function of Road Type and Traffic Condition." *Transportation Research Record* 2672, no. 37: 31–44.

Bloomberg, L., M. Swenson, and B. Haldors. 2003. "Comparison of Simulation Models and the HCM." Presented at the *82nd Annual Meeting of the Transportation Research Board*. Washington, DC: Transportation Research Board.

Brockfeld, E., R. D. Kühne, and P. Wagner. 2004. "Calibration and Validation of Microscopic Traffic Flow Models." *Transportation Research Record* 1876, no. 1: 62–70.

Calida, B., M. A. Perez, M. Ahmed, A. Ghasemzadehkhoshgroudi, J. Clapp. 2016. *IAP Project: Concept to Countermeasure- Research to Deployment Using the 2nd Strategic Highway Research Program (SHRP2) Naturalistic Driving Study (NDS) Data: Adverse Weather and Speeding.* Federal Highway Administration Implementation Assistance Program: Wyoming Department of Transportation. https://doi.org/10.15787/VTT1/CDUJU5.

Chu, L., H. Liu, J-S Oh, and W. Recker. 2003. "A Calibration Procedure for Microscopic Traffic Simulation." Presented at *Institute of Electrical and Electronics Engineers Conference on Intelligent Transportation Systems*. Shanghai, China: IEEE.

Ciuffo, B., V. Punzo, and M. Montanino. 2012. "The Calibration of Traffic Simulation Models: Report on the Assessment of Different Goodness of Fit Measures and Optimization Algorithms." *Joint Research Centre Scientific Report*.

Colombaroni, C., and G. Fusco. 2014. “Artificial Neural Network Models for Car-following: Experimental Analysis and Calibration Issues.” *Journal of Intelligent Transportation Systems* 18, no. 1: 5–16.

Coifman, B., and L. Li. 2017. “A Critical Evaluation of the Next Generation Simulation (NGSIM) Vehicle Trajectory Dataset.” *Transportation Research Part B: Methodological* 105: 362–377.

Traffic Analysis and Forecasting Guidelines. July 2018. Colorado Department of Transportation. Denver, CO.

Creasey, T., and B. Sampson. 2020. “Intersection Capacity Analysis: Are You Doing It Wrong?.” *Institute of Transportation Engineers Journal* 90, no. 1.

Daamen, W., C. Buisson, and S. P. Hoogendoorn, eds. 2014. *Traffic Simulation and Data: Validation Methods and Applications*. Boca Raton, FL: CRC Press.

Dowling, R., A. Skabardonis, and V. Alexiadis. 2004. *Traffic Analysis Toolbox Volume III: Guidelines for Applying Traffic Microsimulation Modeling Software*. Report No. FHWA-HRT-04-040. Washington, DC: Federal Highway Administration.

Duret, A., C. Buisson, and N. Chiabaut. 2008. “Estimating Individual Speed-Spacing Relationship and Assessing Ability of Newell's Car-Following Model to Reproduce Trajectories.” *Transportation Research Record* 2088, no. 1: 188–197.

Ervin, R. D., C. C. MacAdam, K. Gilbert, and P. Tchoryk. 1991. “Quantitative Characterization of the Vehicle Motion Environment (VME).” In *Vehicle Navigation and Information Systems Conference Proceedings: VNIS ’91*, Vol. 2, 1011–1029. Warrendale, Pennsylvania: Society of Automotive Engineers.

Fard, M. R., A. S. Mohaymany, and M. Shahri. 2017. “A New Methodology for Vehicle Trajectory Reconstruction Based on Wavelet Analysis.” *Transportation Research Part C: Emerging Technologies* 74: 150–167.

National Performance Management Research Data Set (NPMRDS). 2019. Washington, DC: Federal Highway Administration. https://npmrds.ritis.org/analytics/.

Geng, X., H. Liang, H. Xu, and B. Yu. 2016. “Influences of Leading-Vehicle Types and Environmental Conditions on Car-Following Behavior.” *IFAC-PapersOnLine* 49, no. 15: 151–156.

GitHub. n.d-a.“AlexeyAB/darknet.” (web page). https://github.com/AlexeyAB/darknet.

GitHub. n.d.-b. “Trajectory Investigation.” (web page). https://github.com/AMSResearchProgram/trajectoryinvestigation.

Google® Earth™. “Mountain View, California” (web page). https://www.google.com/earth, last accessed February 28, 2020.

Google®. 2020. “Google Maps Platform Documentation” (web page). https://developers.google.com/maps/documentation/, last accessed February 28, 2020..

Habtemichael, F., and L. Picado-Santos. 2013. “Sensitivity Analysis of VISSIM Driver Behavior Parameters on Safety of Simulated Vehicles and Their Interaction with Operations of Simulated Traffic.” Presented at the *92nd Annual Meeting of the Transportation Research Board*. Washington, DC: Transportation Research Board.

Hale, D. K. 1997. “How Many NETSIM Runs Are Enough?” *McTrans Newsletter* 11, no. 3: 4–5.

Hale, D. K., C. Antoniou, M. Brackstone, D. Michalaka, A. T. Moreno, and K. Parikh. 2015. “Optimization-Based Assisted Calibration of Traffic Simulation Models.” *Transportation Research Part C: Emerging Technologies* 55: 100–115.

Hallenbeck, M. E., O. I. Selezneva, and R. Quinley. 2014. *Verification, Refinement, and Applicability of Long-Term Pavement Performance Vehicle Classification Rules*. Report No. FHWA-HRT-13-091. Washington, DC: Federal Highway Administration.

Hamdar, S. H., and H. S. Mahmassani. 2008. “Driver Car-Following Behavior: From Discrete Event Process to Continuous Set of Episodes.” Presented at the *Transportation Research Board 87th Annual Meeting*. Washington, DC: Transportation Research Board.

Hamdar, S. H., and H. S. Mahmassani. 2009. “Life in the Fast Lane: Duration-Based Investigation of Driver Behavior Differences Across Freeway Lanes.” *Transportation Research Record* 2124, no. 1: 89–102.

Hammit, B. E., A. Ghasemzadeh, R. M. James, M. M. Ahmed, and R. K. Young. 2018. “Evaluation of Weather-Related Freeway Car-Following Behavior Using the SHRP2 Naturalistic Driving Study Database.” *Transportation Research Part F* 59: 244–259.

James, R. M. 2019. “The Development of a Holistic Approach to Modeling Driver Behavior: Accounting for Driver Heterogeneity in Car-Following Models.” PhD diss. University of Texas.

James, R. M., and B.E. Hammit. 2019. “Identifying Contributory Factors to Heterogeneity in Driving Behavior: Clustering and Classification Approach.” *Transportation Research Record* 2673, no. 10: 343–353.

James, R. M., B. E. Hammit, and S. D. Boyles. 2019. “Methods to Obtain Representative Car-Following Model Parameters from Trajectory-Level Data for Use in Microsimulation.” *Transportation Research Record* 2673, no. 7: 62–73.

James, R. M. *Traffic Analysis and Simulation*. Report No. TPF-5(176). Washington, DC: Federal Highway Administration. National Cooperative Highway Research Program. 2020. Transportation Pooled Fund Program. Accessed April 29, 2020. https://pooledfund.org/Details/Study/403.

Kesting, A., and M. Treiber. 2008. “Calibrating Car-Following Models by Using Trajectory Data: Methodological Study.” *Transportation Research Record* 2088, no. 1: 148–156.

Kesting, A., and M. Treiber; Appert-Rolland, C., F. Chevoir, P. Gondret, S. Lassarre, J.P. Lebacque, and M. Schreckenberg, eds. 2009. “Calibration of Car-Following Models Using Floating Car Data.” In *Traffic and Granular Flow 07:* 117–127. Berlin, Heidelberg: Springer.

Kim, J., and H. S. Mahmassani. 2011. “Correlated Parameters in Driving Behavior Models: Car-Following Example and Implications for Traffic Microsimulation.” *Transportation Research Record* 2249, no. 1: 62–77.

Kim, J., H. S. Mahmassani, P. Vovsha, Y. Stogios, and J. Dong. 2013. “Scenario-Based Approach to Analysis of Travel Time Reliability with Traffic Simulation Models.” *Transportation Research Record* 2391, no. 1: 56–68.

Kondyli, A., D. K. Hale, M. Asgharzadeh, B. Schroeder, A. Jia, and J. Bared. 2019. “Evaluating the Operational Effect of Narrow Lanes and Shoulders for the Highway Capacity Manual.” *Transportation Research Record* 2673, no. 10: 558–570.

Lownes, N. E., and R. B. Machemehl. 2006. “VISSIM: A Multi-Parameter Sensitivity Analysis.” In *Proceedings of the 2006 Winter Simulation Conference*, 1406–1413. Monterey, CA: IEEE.

Lu, X-Y, and A. Skabardonis. 2007. “Freeway Traffic Shockwave Analysis: Exploring the NGSIM Trajectory Data.” Presented at the *Transportation Research Board 86th Annual Meeting*, Washington, DC: Transportation Research Board.

Ma, X., and I. Andréasson. 2005. “Dynamic Car-following Data Collection and Noise Cancellation Based on the Kalman Smoothing.” In *Proceedings of IEEE International Conference on Vehicular Electronics and Safety*, 35–41. Shaanxi, China: IEEE.

Marczak, F., and C. Buisson. 2012. “New Filtering Method for Trajectory Measurement Errors and its Comparison with Existing Methods.” *Transportation Research Record* 2315, no. 1: 35–46.

VISSIM Modeling Guidance. August 2017. Maryland Department of Transportation State Highway Administration. Baltimore, MD.

Massey Jr., F. J. March 1951. “The Distribution of the Maximum Deviation between Two Sample Cumulative Step Functions.” *The Annals of Mathematical Statistics* 22, no. 1: 125–128.

Montanino, M., B. Ciuffo, and V. Punzo. 2012. “Calibration of Microscopic Traffic Flow Models Against Time-Series Data.” In *Proceedings of the 2012 International IEEE Conference on Intelligent Transportation Systems*, 108–114. Anchorage, AK: IEEE.

Montanino, M., and V. Punzo. 2013. "Making NGSIM Data Usable for Studies on Traffic Flow Theory: Multistep Method for Vehicle Trajectory Reconstruction." *Transportation Research Record* 2390, no. 1: 99–111.

Montanino, M., and V. Punzo. 2015. "Trajectory Data Reconstruction and Simulation-Based Validation against Macroscopic Traffic Patterns." *Transportation Research Part B: Methodological* 80: 82–106.

Morris, B., and M. Trivedi. 2009. "Learning Trajectory Patterns by Clustering: Experimental Studies and Comparative Evaluation." In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 312–319. Miami, FL: IEEE.

Muthurajan, B., R. Amrutsamanvar, and L. D.Vanajakshi. 2017. "A Semi-Automated Image Processing Solution for Extracting Microscopic Traffic Data." Presented at the *10th Urban Mobility India Conference*. Telangana, India: Government of India, Ministry of Housing and Urban Affairs.

Ossen, S., and S. P. Hoogendoorn. 2008. "Validity of Trajectory-Based Calibration Approach of Car-Following Models in Presence of Measurement Errors." *Transportation Research Record* 2088, no. 1: 117–125.

Pal, D., and M. Chunchu. 2018. "Smoothing of Vehicular Trajectories under Heterogeneous Traffic Conditions to Extract Microscopic Data." *Canadian Journal of Civil Engineering* 45, no. 6: 435–445.

Punzo, V., D. J. Formisano, and V. Torrieri. 2005. "Nonstationary Kalman Filter for Estimation of Accurate and Consistent Car-Following Data." *Transportation Research Record* 1934, no. 1: 2–12.

Punzo, V., M. T. Borzacchiello, and B. Ciuffo. 2009. "Estimation of Vehicle Trajectories from Observed Discrete Positions and Next-Generation Simulation Program (NGSIM) Data." Presented at the *Transportation Research Board 88th Annual Meeting*, Washington, DC: Transportation Research Board.

Punzo, V., M. T. Borzacchiello, and B. Ciuffo. 2011. "On the Assessment of Vehicle Trajectory Data Accuracy and Application to the Next Generation Simulation (NGSIM) Program Data." *Transportation Research Part C: Emerging Technologies* 19, no. 6: 1243–1262.

Punzo, V., M. Montanino, and B. Ciuffo. 2014. "Do We Really Need to Calibrate All the Parameters? Variance-Based Sensitivity Analysis to Simplify Microscopic Traffic Flow Models." *IEEE Transactions on Intelligent Transportation Systems* 16, no. 1: 184–193.

Ranjitkar, P., T. Nakatsuji, and M. Asano. 2004. "Performance Evaluation of Microscopic Traffic Flow Models with Test Track Data." *Transportation Research Record* 1876, no. 1: 90–100.

Ranjitkar, P., T. Nakatsuji, and A. Kawamura. 2005. "Experimental Analysis of Car-Following Dynamics and Traffic Stability." *Transportation Research Record* 1934, no. 1: 22–32.

Redmon, J., and A. Farhadi. 2018. *YOLOv3: An Incremental Improvement*. arXiv:1804.02767. Ithaca, NY: arXivLabs, Cornell University.

Talebpour, A., H. S. Mahmassani, and S. H. Hamdar. 2015. "Modeling Lane-Changing Behavior in a Connected Environment: A Game Theory Approach." *Transportation Research Part C: Emerging Technologies* 59: 216–232.

Taylor, J., X. Zhou, and N. M. Rouphail. 2012. "Calibrating Dynamic Car-Following Model Using Vehicle Trajectory Data: A Dynamic Time Warping Approach." Presented at the *Transportation Research Board 91st Annual Meeting*, Washington, DC: Transportation Research Board.

Taylor, J., X. Zhou, N. M. Rouphail, and R. J. Porter. 2015. "Method for Investigating Intradriver Heterogeneity Using Vehicle Trajectory Data: A Dynamic Time Warping Approach." *Transportation Research Part B: Methodological* 73: 59–80.

Thiemann, C., M. Treiber, and A. Kesting. 2008. "Estimating Acceleration and Lane-Changing Dynamics from Next Generation Simulation Trajectory Data." *Transportation Research Record* 2088, no. 1: 90–101.

Toledo, T., H. N. Koutsopoulos, and K. I. Ahmed. 2007. "Estimation of Vehicle Trajectories with Locally Weighted Regression." *Transportation Research Record* 1999, no. 1: 161–169.

Toledo, T., H. Koutsopolous, A. Davol, M. Ben-Akiva, W. Burghout, I. Andreasson, T. Johansson, and C. Lundin. 2003. "Calibration and Validation of Microscopic Traffic Simulation Tools: Stockholm Case Study," *Transportation Research Record* 1831: 65–75.

Highway Capacity Manual, A Guide for Multimodal Mobility Analysis. 6th ed. 2016. Transportation Research Board, Washington, DC.

Treiber, M., and A. Kesting. 2013. "Microscopic Calibration and Validation of Car-Following Models—A Systematic Approach." *Procedia—Social and Behavioral Sciences* 80: 922–939.

Treiber, M., and A. Kesting. 2013. "Traffic Flow Dynamics." In *Traffic Flow Dynamics: Data, Models and Simulation.* Berlin, Heidelberg: Springer-Verlag.

Treiber, M., A. Kesting, and R. E. Wilson. 2011. "Reconstructing the Traffic State by Fusion of Heterogeneous Data." *Computer-Aided Civil and Infrastructure Engineering* 26, no. 6: 408–419.

University of California, San Diego. 2014. "UCSD Trajectory Data." http://cvrr.ucsd.edu/bmorris/datasets/dataset_trajectory_clustering.html.

Next Generation Simulation (NGSIM) Vehicle Trajectories and Supporting Data. 2019. Washington, DC: U.S. Department of Transportation.

https://data.transportation.gov/Automobiles/Next-Generation-Simulation-NGSIM-Vehicle-Trajector/8ect-6jqj.

Vasconcelos, L., L. Neto, S. Santos, A. Bastos Silva, and Á. Seco. 2014. "Calibration of the Gipps Car-Following Model Using Trajectory Data." *Transportation Research Procedia* 3: 952–961.

VDOT VISSIM User Guide, v. 2.0. 2020. Richmond, VA: Virginia Department of Transportation, Traffic Engineering Division,.

Wang, H., W. Wang, J. Chen, and M. Jing. 2010. "Using Trajectory Data to Analyze Intradriver Heterogeneity in Car-Following." *Transportation Research Record* 2188, no. 1: 85–95.

Wei, H., C. Feng, E. Meyer, and J. Lee. 2005. "Video-Capture-Based Approach to Extract Multiple Vehicular Trajectory Data for Traffic Modeling." *Journal of Transportation Engineering* 131, no. 7: 496–505.

Wunderlich, K. E., M. Vasudevan, and P. Wang. 2019. *TAT Volume III: Guidelines for Applying Traffic Microsimulation Modeling Software 2019 Update to the 2004 Version*. Report No. FHWA-HOP-18-036. Washington, DC: Federal Highway Administration.

Xin, W., J. Hourdos, and P. Michalopoulos. 2008. "A Vehicle Trajectory Collection and Processing Methodology and its Implementation to Crash Data." Presented at the *Transportation Research Board 87th Annual Meeting*, Washington, DC: Transportation Research Board.

Xu, F., and L. Sun. 2013. "An Efficient Video-Based Vehicle Trajectory Processing Approach." In *ICTIS 2013: Improving Multimodal Transportation Systems-Information, Safety, and Integration*. Reston, VA: American Society of Civil Engineers.

Xyntarakis, M., V.Alexiadis, R. Campbell, E. Flanigan, and Cambridge Systematics. 2016. *Active Transportation and Demand Management (ATDM) Trajectory-Level Validation State of the Practice Review*. Report No. FHWA-JPO-14-193. Washington, DC: Federal Highway Administration.

Ye, F., and Y. Zhang. 2019. "Vehicle Type-Specific Headway Analysis Using Freeway Traffic Data." *Transportation Research Record* 2124, no. 1: 222–230.

Yu, M., and W. D. Fan. 2017. "Calibration of Microscopic Traffic Simulation Models Using Metaheuristic Algorithms." *International Journal of Transportation Science and Technology* 6, no. 1: 63–77.

Zhao D., L. Xiaopeng. 2019. "Video-Based Intelligent Road Traffic Universal Analysis TooL (VIRTUAL)." University of South Florida. http://www.research.usf.edu/dpl/content/data/PDF/18B141.pdf.

Zhong, R. X., K. Y. Fu, A. Sumalee, D. Ngoduy, and W. H. K. Lam. 2016. “A Cross-Entropy Method and Probabilistic Sensitivity Analysis Framework for Calibrating Microscopic Traffic Models.” *Transportation Research Part C: Emerging Technologies* 63: 147–169.

Recycled
Recyclable

HRDO-20/08-21(WEB)E