World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Defining accuracy benchmarks for freeway traffic simulations in support of highway operations and planning

Ayodeji Ajidahun [1,*] and Mujeeb Abiola Abdulrazaq [2]

[1] *Department of Civil Engineering, University of New Haven, West Haven, CT, USA.*
[2] *Department of Civil and Environmental Engineering, University of North Carolina at Charlotte.*

World Journal of Advanced Research and Reviews, 2025, 27(02), 790-797

Publication history: Received on 01 July 2025; revised on 09 August 2025; accepted on 11 August 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.27.2.2900

Abstract

Accurate calibration of traffic simulation models is essential for replicating observed traffic conditions, and subsequent optimization of decision-making processes and targeted investments in transportation infrastructure. This study applies a genetic algorithm (GA) to optimize key parameters of the car-following model for a basic freeway segment in California, aiming to minimize the error between simulated and observed traffic data. Outputs generated during GA iterations were analyzed using paired T-tests and Wilcoxon signed-rank tests to compare simulated speed and flow against ground truth data. Accuracy for each sample was matched to its corresponding P-value, revealing a clear trend: when accuracy levels exceeded 80%, P-values for both speed and flow consistently rose above 0.05. This indicates that the simulated outputs became statistically indistinguishable from the observed field data after 80% accuracy. These findings demonstrate that combining statistical significance with accuracy metrics can effectively guide calibration processes and establish thresholds for acceptable simulation accuracy, contributing to a robust framework for traffic simulation studies.

Keywords: Civil engineering; Highway engineering; Traffic simulation; Traffic flow modeling; Genetic algorithm optimization; Transportation infrastructure planning

1. Introduction

Traffic simulation has become a cost-effective and indispensable tool for transportation planning, aiding engineers and planners in designing and managing efficient road systems [1, 2]. By modeling various scenarios, simulation provides insights into potential traffic conditions, operational performance, and safety outcomes, ultimately informing critical decisions for infrastructure investment and policy-making [3, 4]. However, the accuracy of these simulations in reflecting real-world conditions is paramount. A poorly calibrated model risks either underestimating or overestimating traffic performance/behavior, leading to unreliable future predictions and suboptimal outcomes for simulation-based studies.

Accurate replication of real-world traffic behavior is particularly crucial in unique or complex scenarios, where reliable predictions can have significant implications. For instance, simulations may be employed to study traffic behavior under extreme weather conditions, evaluate the relative benefits of innovative road designs, or analyze the impact of emerging technologies such as connected and autonomous vehicles (CAVs) [5] on prevailing traffic conditions. Each of these cases demands a model that aligns closely with real-world data to ensure valid and actionable results.

Over the years, researchers have developed a variety of techniques to calibrate simulation models and adjust driving behavior parameters. Most calibration efforts have relied on metaheuristic optimization algorithms such as genetic

* Corresponding author: Ayodeji Ajidahun

algorithms, particle swarm optimization, and simulated annealing [6, 7], which systematically minimize the differences between simulated and observed traffic conditions. These methods automate the calibration process, achieving high accuracy and efficiency. On the other hand, some researchers have taken a more manual approach, employing grid search methods to iteratively test and adjust parameters [8, 9]. While these approaches have advanced the field of traffic simulation calibration, their focus has been primarily on minimizing errors, without defining an acceptable threshold for accuracy in simulation studies.

Despite these advancements, no study has yet established a standard for what constitutes an acceptable level of accuracy in traffic simulation. This knowledge gap has significant implications, as simulations often serve as the foundation for policy decisions and infrastructure investments. Without a clear standard, there is a risk of over-reliance on models that may not meet the rigor required for reliable predictions. To address this gap, this study investigates the use of both parametric and non-parametric statistical methods to evaluate calibration accuracy. A basic freeway segment in California is used as a case study, providing a controlled environment for testing and analysis.

This paper makes a novel contribution by not only applying metaheuristic optimization methods for calibration but also analyzing the statistical significance of accuracy levels achieved. By matching accuracy metrics with statistical significance, this study establishes a framework for determining acceptable thresholds for simulation accuracy. These findings aim to contribute to the standardization of accuracy metrics in traffic simulation, ensuring that future models achieve the reliability necessary for critical transportation planning decisions.

2. Methods

2.1. Study Area

The study area is a segment of Interstate I-80, located in Los Angeles, California, within Yolo County, as illustrated in Figure 1. This section of the freeway is a four-lane basic segment with a total length of 5,280 feet. The study focused on the evening peak period, specifically from 4:00 PM to 5:00 PM on August 21, 2018. Field traffic data, including flow and speed, were collected and aggregated into 5-minute intervals.

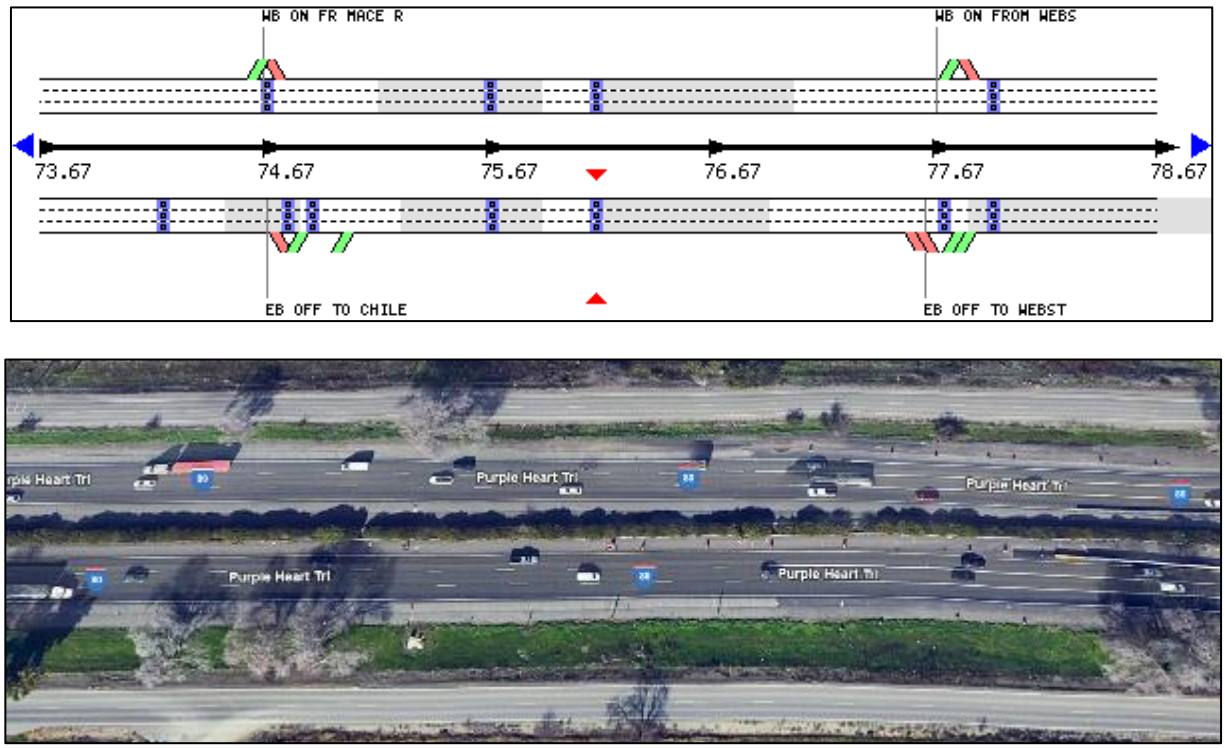


Figure 1 I-80 Freeway Segment in Yolo County (Google Earth)

Table 1 summarizes the traffic flow observed in each lane during these intervals. The table also presents the total traffic flow across all four lanes and the average traffic speed for the segment.

Table 1 Collected Traffic from Sensors**.**

Time	**Speed (mph)**				**Flow (Veh/ 5 Minutes)**			
	Lane 1	Lane 2	Lane 3	Speed	Lane 1	Lane 2	Lane 3	Flow
4:05 PM	58.30	56.30	55.60	56.90	172	150	112	434
4:10 PM	57.70	57.70	54.00	56.80	161	160	107	428
4:15 PM	57.50	58.90	53.00	56.80	172	159	119	450
4:20 PM	58.10	58.20	58.70	58.30	171	154	122	447
4:25 PM	56.70	58.30	54.20	56.60	168	138	108	414
4:30 PM	41.00	43.60	43.90	42.80	139	134	128	401
4:35 PM	46.20	46.70	46.00	46.30	160	148	127	435
4:40 PM	53.10	54.30	54.50	53.90	159	145	116	420
4:45 PM	55.70	56.30	56.60	56.20	182	164	130	476
4:50 PM	54.90	57.40	56.80	56.30	167	160	131	458
4:55 PM	36.30	44.40	44.00	41.70	117	132	122	371
5:00 PM	33.30	38.70	37.40	36.40	136	131	126	393

2.2. Car Following Model

The Wiedemann 99 (W99) model is a psycho-physical car-following model develop in 1999, derived from the original Wiedemann model proposed in 1974 (W74) [10]. It consists of 10 parameters (CC0, CC1, ..., CC9), which can be calibrated (or adjusted) to represent driving behaviors of human driven vehicles (HDVs) on freeways. Among these, CC0, CC8, and CC9 are particularly crucial in determining the model's performance. The equation governing the model is given by:

$$v_n(t+\Delta t) = min\left\{v_n(t) + 3.6 \times \left(CC8 + \frac{CC8-CC9}{80} \times v_n(t)\right)\Delta t; u_f, 3.6 \times \frac{S_{n(t)} - CC0 - L_{n-1}}{v_n(t)}; u_f\right\} \quad (1)$$

Where $v_n(t+\Delta t)$ represents the speed of the subject vehicle after Δt seconds relative to time step t, $S_{n(t)}$ is the distance between the subject and leading vehicle; L_{n-1} denotes the length of the leading vehicle; and u_f is the free-flow speed. The explanations for 10 parameters are described in Table 2.

Table 2 Traffic Parameters

W99		
Parameters	Interpretation	Default
CC0	Average standstill distance (m)	1.4
CC1	Headway (s)	1.2
CC2	Longitudinal oscillation (m)	8
CC3	Start of deceleration process (s)	-12
CC4	Minimal closing Δv (m/s)	-1.5
CC5	Minimal opening Δv (m/s)	2.1
CC6	Speed dependency of oscillation (10^{-4} rad/s)	6
CC7	Oscillation acceleration – m/s^2	0.25
CC8	Acceleration rate when starting (m/s^2)	2
CC9	Acceleration behavior at 80 km/h (m/s^2)	1.5

2.3. Calibration

A genetic algorithm (GA) is used to enhance the calibration of a microscopic traffic simulation model (the W99 car following model) by approaching near-global optimal solutions [11, 12]. The GA simulates biological evolution through selection, crossover, and mutation mechanisms. Initially, the algorithm begins with a randomly generated population of solutions, and in each iteration, higher-quality solutions have a greater chance of being selected for reproduction, producing new populations through crossover and mutation. This study employs two different GA configurations: the first uses a population size of 20, with a 20% mutation probability over 20 generations, while the second uses a population size of 30, with a 30% mutation probability over 30 generations. The objective is to assess individual GA members to obtain a sufficient sample size for generalization in the study.

The calibration process is executed in Python, where binary chromosomes are randomly generated to represent feasible solutions. These chromosomes are then decoded into model parameters, which are fed into the SUMO simulation software. The objective function is evaluated by comparing the simulated traffic flow and speed data with observed real-world values. The calibration continues until the maximum number of generations is reached or a predefined stopping condition is satisfied. This process is depicted in Figure 2 as adopted from. In this regard, the optimization framework is formulated as follows:

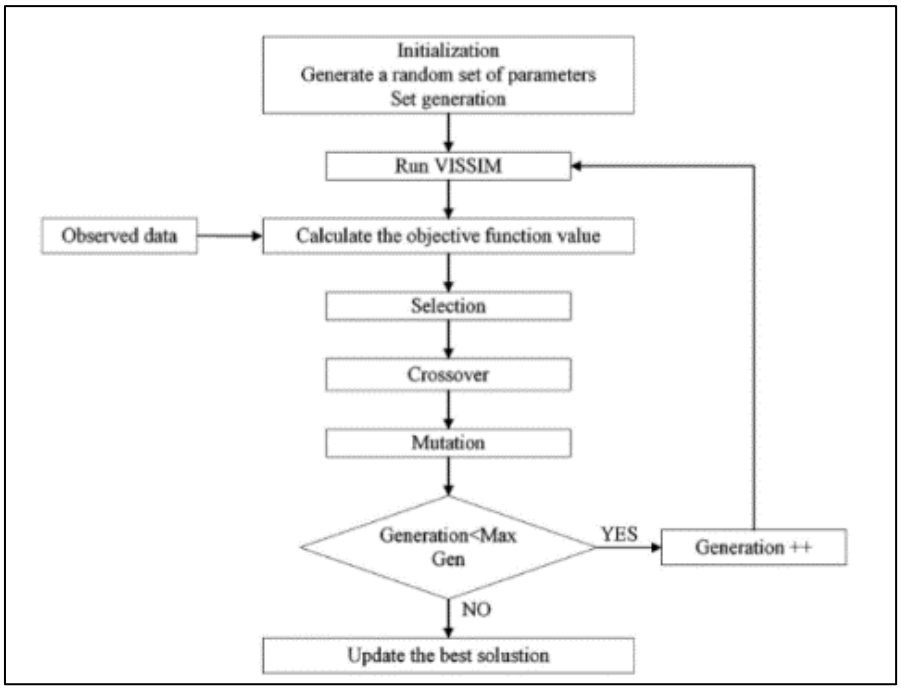


Figure 2 GA Calibration process

$$f(V^{obs}, V^{sim})$$

Subject to the constraints:

$$l_{x_i} \leq x_i \leq u_{x_i}, i = 1 \dots n,$$

Where x_i= the model parameters to be calibrated, $f(obj)$= objective function, V^{obs}, V^{sim}= observed and simulated value of model parameters, l_{x_i}, u_{x_i}= the respective lower and upper bounds of model parameter , n = number of variables. The objective function uses the Mean Absolute Normalized Error (MANE), which is provided by the following equation. The calibration using the flow and speed data as performance measures is formulated as follows:

$$Min\ MANE(q, v) = \frac{1}{N}\sum_{i=1}^{N} \left(\frac{|q_{obs,i} - q_{sim,i}|}{q_{obs,i}} + \frac{|v_{obs,i} - v_{sim,i}|}{v_{obs,i}}\right) \quad \dots\dots\dots \quad (2)$$

Where $q_{obs,i}, q_{sim,i}$ = observed and simulated traffic volume for a given time period i, $v_{obs,i}, v_{sim,i}$ = observed and simulated traffic speed for a given time period i, N = total number of observations.

2.4. Statistical Testing

During the calibration process, each candidate solution produced by the GA iterations was evaluated for statistical significance using both parametric and non-parametric methods to compare the simulated and observed traffic data. The two tests employed are described below:

2.4.1. The Paired T-test

The paired T-test is a parametric test used to compare the means of two related groups [13, 14], in this case, the simulated traffic data and the observed traffic data. This test evaluates whether there is a statistically significant difference between the two sets of data [15 ,16]. It assumes that the differences between the paired values are normally distributed. The null hypothesis for this test is that there is no significant difference between the simulated and observed traffic flow and speed data, and a p-value less than 0.05 indicates a significant difference.

2.4.2. The Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test, a non-parametric test, is used when the assumption of normality is not met and the data are paired [17, 18, 19]. This test compares the distributions of two related samples (simulated and observed data), ranking the absolute differences between the pairs and then testing whether the ranks of these differences significantly deviate from zero. It does not require the data to follow a normal distribution, making it suitable for cases where the data may be skewed or contain outliers. A p-value less than 0.05 indicates that the difference between the simulated and observed data is statistically significant.

Both tests were utilized to assess the robustness and accuracy of the traffic simulation model in replicating real-world conditions. The paired T-test provides a direct comparison of means, while the Wilcoxon signed-rank test offers a more flexible approach when the data distribution does not meet parametric assumptions.

3. Results

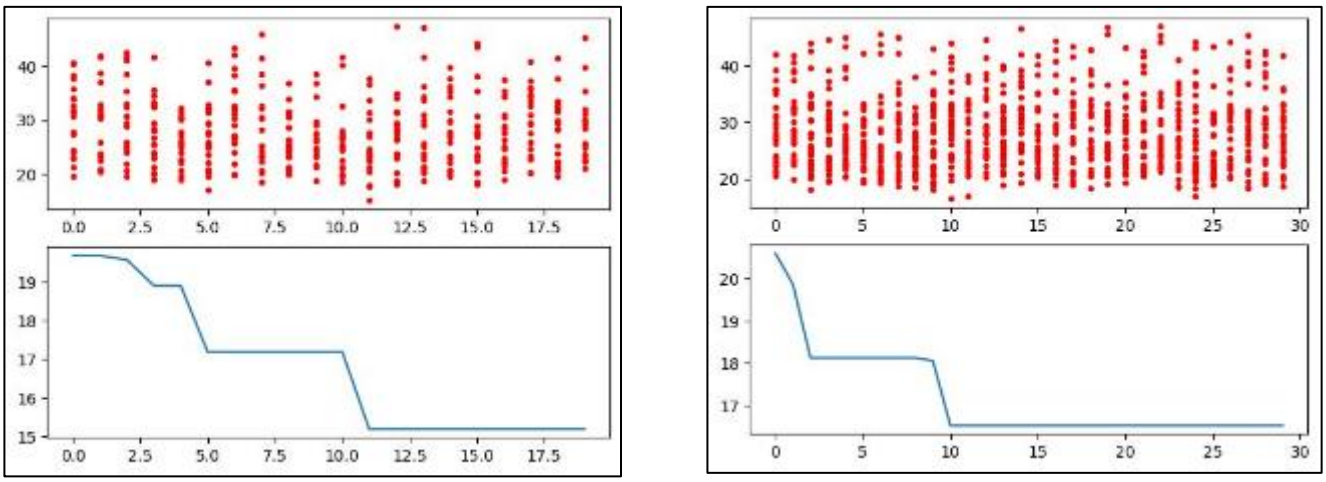
3.1. Calibration Results

As previously discussed, this study involved two distinct calibrations to analyze the impact of different parameter configurations on model performance. The first scenario employed a population size of 20, a mutation rate of 20%, and 20 generations, while the second scenario used a population size of 30, a mutation rate of 30%, and 30 generations. Together, these two scenarios produced a total of 1,300 samples (400 from the first configuration and 900 from the second). The primary aim of these analyses was to generate a sufficiently large sample size to ensure that the findings are robust and generalizable.

Table 3 Calibration Results

W99		**Parameters**		
Parameters	Interpretation	Default	Calibration 1	Calibration 2
CC0	Average standstill distance (m)	1.40	0.50	1.54
CC1	Headway (s)	1.20	1.18	1.03
CC2	Longitudinal oscillation (m)	8.00	7.14	8.93
CC3	Start of deceleration process (s)	-12.00	10.72	12.60
CC4	Minimal closing Δv (m/s)	-1.50	-0.23	-0.35
CC5	Minimal opening Δv (m/s)	2.10	0.25	0.44
CC6	Speed dependency of oscillation (10^{-4} rad/s)	6.00	5.06	5.69
CC7	Oscillation acceleration – m/s^2	0.25	0.21	0.31
CC8	Acceleration rate when starting (m/s^2)	2.00	2.30	2.69
CC9	Acceleration behavior at 80 km/h (m/s^2)	1.50	2.98	3.04
Metric	Accuracy	72%	85%	83%

The results reveal a trend of diminishing returns with increased computational effort. Specifically, the first configuration (20 generations) achieved an accuracy of 85%, while the second configuration (30 generations) resulted in a slightly lower accuracy of 83%.



Case One Case Two

Figure 3 Optimization Results.

These findings suggest that increasing the population size, mutation rate, and number of generations beyond a certain point may not yield proportional improvements in model accuracy. Despite this, both configurations significantly outperformed the default parameter settings, which produced an accuracy of 72%. These results highlight the value of parameter optimization while emphasizing the need to balance computational resources with expected gains. The next section delves deeper into the statistical tests conducted to evaluate these outcomes and examines the trends observed.

3.2. Statistical Testing

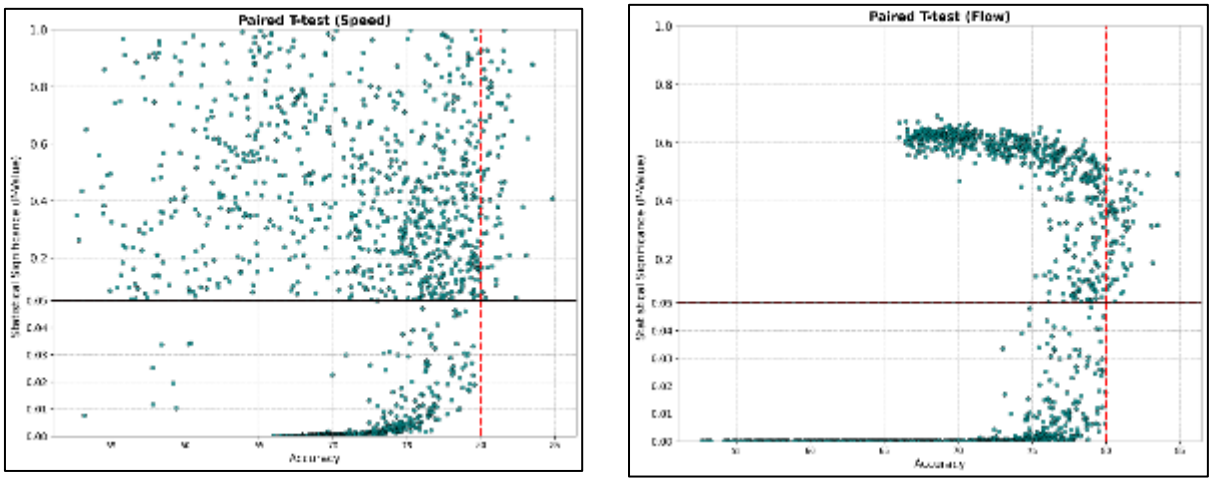


Figure 4 Paired T-test for both Speed and Flow

During the calibration process, candidate solutions from both simulations underwent rigorous statistical testing to assess their significance for speed and flow. Non-parametric and parametric methods, including the paired t-test and Wilcoxon signed-rank test, were applied to ensure consistency across the results. Following these tests, a scatter plot was generated to explore the relationship between accuracy levels and statistical significance (p-values).

The observed trends, as depicted in the figures below, highlight a key distinction: while lower accuracy levels achieved acceptable statistical significance for speed, they failed to meet the statistical significance requirements for flow in both tests. This indicates that lower accuracy may satisfy the speed criterion but falls short for flow requirements.

Conversely, at higher accuracy levels, particularly those exceeding the 80% threshold, the p-values consistently surpassed the rejection region for both speed and flow across all tests. These findings suggest that an accuracy level of 80% or higher is generally sufficient to produce results that are statistically indistinguishable from the ground truth data. This underscores the importance of achieving higher accuracy levels to ensure robust and reliable outcomes.

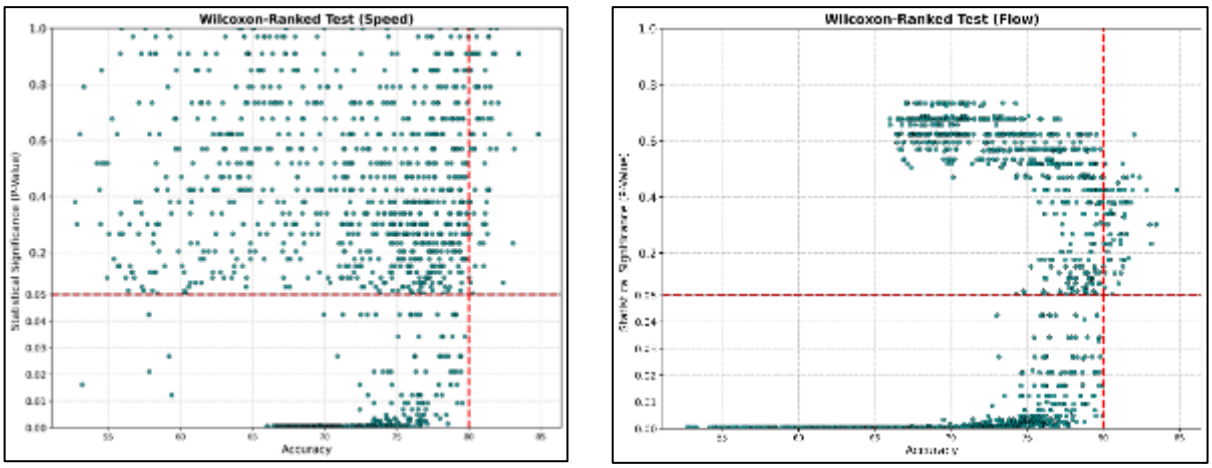


Figure 5 Wilcoxon Signed Ranked Test for both Speed and Flow

4. Conclusion

This study demonstrates the effectiveness of using a genetic algorithm (GA) to calibrate a car-following model for simulating traffic behavior. By optimizing key parameters in the Wiedemann 99 (W99) model, we significantly improved its accuracy in replicating observed traffic conditions on a California freeway segment. The key takeaway is the establishment of a benchmark for simulation accuracy. The results show that achieving an accuracy level of 80% or higher ensures that simulated traffic speeds and flows are statistically indistinguishable from real-world data, validated through paired T-tests and Wilcoxon signed-rank tests. This finding provides a clear threshold for model reliability, essential for making sound decisions in transportation planning.

Additionally, the study highlights the importance of balancing optimization efforts with computational efficiency, as further increases in accuracy yield diminishing returns. These results contribute to the growing body of knowledge on traffic simulation calibration and set a foundation for future studies to refine and apply these methods in broader contexts, such as the integration of connected and autonomous vehicles.

Ultimately, this research provides a framework for establishing acceptable levels of accuracy in traffic simulations, ensuring their reliability for policy-making and infrastructure planning.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Barceló, J. (2010). *Fundamentals of traffic simulation* (Vol. 145). Springer.

[2] Kotusevski, G., & Hawick, K. A. (2009). A review of traffic simulation software. *Research Letters in the Information and Mathematical Sciences*, *13*, 35-54.

[3] Bartin, B., Ozbay, K., Gao, J., & Kurkcu, A. (2018). Calibration and validation of large-scale traffic simulation networks: A case study. *Procedia computer science*, *130*, 844-849.

[4] Casas, J., Ferrer, J. L., Garcia, D., Perarnau, J., & Torday, A. (2010). Traffic simulation with aimsun. In *Fundamentals of traffic simulation* (pp. 173-232). Springer.

[5] Shladover, S. E., Su, D., & Lu, X.-Y. (2012). Impacts of cooperative adaptive cruise control on freeway traffic flow. *Transportation Research Record*, *2324*(1), 63-70.

[6] Kesur, K. B. (2009). Advances in genetic algorithm optimization of traffic signals. *Journal of Transportation Engineering*, *135*(4), 160-173.

[7] Lidbe, A. D., Hainen, A. M., & Jones, S. L. (2017). Comparative study of simulated annealing, tabu search, and the genetic algorithm for calibration of the microsimulation model. *Simulation*, *93*(1), 21-33.

[8] Chowdhury, M. M. H., & Chakraborty, T. (2024). Calibration of SUMO Microscopic Simulation for Heterogeneous Traffic Condition: The Case of the City of Khulna, Bangladesh. *Transportation Engineering*, *18*, 100281. https://doi.org/https://doi.org/10.1016/j.treng.2024.100281

[9] Gao, K., Zhang, Y., Su, R., Yang, F., Suganthan, P. N., & Zhou, M. (2018). Solving traffic signal scheduling problems in heterogeneous traffic network by using meta-heuristics. IEEE Transactions on Intelligent Transportation Systems, 20(9), 3272-3282.

[10] Wiedemann, R. (1974). Simulation des Straßenverkehrsflusses.

[11] Katoch, S., Chauhan, S. S., & Kumar, V. (2021). A review on genetic algorithm: past, present, and future. Multimedia tools and applications, 80, 8091-8126.

[12] Sánchez-Mangas, R., García-Ferrrer, A., de Juan, A., & Arroyo, A. M. (2010). The probability of death in road traffic accidents. How important is a quick medical response? Accident Analysis & Prevention, 42(4), 1048-1056. https://doi.org/https://doi.org/10.1016/j.aap.2009.12.012

[13] Kim, T. K. (2015). T test as a parametric statistic. Korean journal of anesthesiology, 68(6), 540-546.

[14] Manfei, X., Fralick, D., Zheng, J. Z., Wang, B., Xin, M. T., & Changyong, F. (2017). The differences and similarities between two-sample t-test and paired t-test. Shanghai archives of psychiatry, 29(3), 184.

[15] Afifah, S., Mudzakir, A., & Nandiyanto, A. B. D. (2022). How to calculate paired sample t-test using SPSS software: From step-by-step processing for users to the practical examples in the analysis of the effect of application anti-fire bamboo teaching materials on student learning outcomes. Indonesian Journal of Teaching in Science, 2(1), 81-92.

[16] Yu, Z., Guindani, M., Grieco, S. F., Chen, L., Holmes, T. C., & Xu, X. (2022). Beyond t test and ANOVA: applications of mixed-effects models for more rigorous statistical analysis in neuroscience research. *Neuron*, *110*(1), 21-35.

[17] Elkin, L. A., Kay, M., Higgins, J. J., & Wobbrock, J. O. (2021). An aligned rank transform procedure for multifactor contrast tests. The 34th annual ACM symposium on user interface software and technology,

[18] Rosner, B., Glynn, R. J., & Lee, M.-L. T. (2006). The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics*, *62*(1), 185-192.

[19] Windi, W. A., Taufiq, M., & Muhammad, T. (2021). Implementasi wilcoxon signed rank test untuk mengukur efektifitas pemberian video tutorial dan ppt untuk mengukur nilai teori. *Produktif: Jurnal Ilmiah Pendidikan Teknologi Informasi*, *5*(1), 405-410.