Salerno

DTA 2025

# 10TH INTERNATIONAL SYMPOSIUM ON DYNAMIC TRAFFIC ASSIGNMENT

16th- 18th September 2025

# ABSTRACT BOOK

# Equilibrium Analysis and Fare-Reward Scheme for Transit Morning Commute Considering the Random Delay

Zhaorui Li[1], Jianan Cao[2], Xiao Han[1], Rui Jiang[1]

[1]*School of Systems Science, Beijing Jiaotong University, Beijing, China;*

[2]*Beijing Transport Institute, Beijing, China*

## 1. Introduction

Congestion in morning commute brings inconvenience to people's daily life. Recently, numerous studies have focused on the management policies of public transits during the peak hour (Kraus & Yoshida, 2002; Lan et al., 2010; Tang et al., 2020). Also, various fare schemes have been considered, such as peak fares and off-peak discounts (Currie, 2010; Peer et al., 2016; Tang et al., 2020; Yang & Tang, 2018). Yang and Tang (2018) put forward a fare-reward scheme (FRS) to minimize the system cost on the premise of ensuring a constant revenue of the transit operator. This scheme stipulates that commuters can get the reward for a free rail transit travel at the beginning or the end in the morning commute, after a certain number of paid travels.

However, these studies excluded the random congestion, which is often experienced by commuters when they transfer to the bus after getting off the rail transit. That is because the time of bus travel highly depends on the traffic conditions. Thus, it may have an impact on the commuters' travel behavior, and then affect the implementation of fare schemes. In this study, build upon the work of Yang and Tang (2018), we investigate the impact of random delay. We first analyze the commuters' travel behavior with all uniform fares, and then classify them into three equilibrium cases according to the extent of random delay. Further, we derive the optimal FRS, and measure the system efficiency promoted by this scheme. This study sheds light on discussing the influence of random delay on the effectiveness of FRS, and provides a basis for transportation managers to formulate more realistic fare schemes.

## 2. The Model

### 2.1. Equilibrium Analysis

We consider the commute includes two parts, i.e., $N$ commuters take the rail transit first and then transfer to the bus to go to the workplace. In the transit travel, commuters experience a bottleneck with capacity $s$ caused by excessive transit queues. It is equivalent to a classic bottleneck model (Vickrey, 1969). In the bus travel, there is a random delay $T$, and $T \sim \mathrm{U}(0, b)$.

Without loss of generality, we assume that the work start time $t^* = 0$, and the time of other parts, including the transfer and the free flow, is zero. Thus, we can formulate the expected travel cost of commuters who pass through the bottleneck at time $t$,

$$C(t) = \alpha q(t) + \beta e(t) + \gamma l(t) + \varepsilon T + p_0, \quad (1)$$

where $\alpha$, $\beta$, $\gamma$, and $\varepsilon$ denotes the shadow value of queuing, early arrival, late arrival, and the congestion, $\beta < \alpha < \varepsilon < \gamma$ (Small, 1982); $q(t)$, $e(t)$, and $l(t)$ denote the time; $p_0$ is the uniform fare without FRS. The three user equilibrium cases solved are shown in Table 1.

**Table 1.** Equilibrium cases of commuters' arrival with different random delay.

| | Maximum random delay $b$ | Commuters' arrival | | | The earliest time of passing through the bottleneck |
|---|---|---|---|---|---|
| | | Always early | Probably early or late | Always late | |
| Case 1 | $\left[0, \frac{2\beta N}{(\beta+\gamma)s}\right)$ | √ | √ | √ | $-\frac{\gamma}{\beta+\gamma}\frac{N}{s} - \frac{b}{2}$ |
| Case 2 | $\left[\frac{2\beta N}{(\beta+\gamma)s}, \frac{(\beta+\gamma)N}{2\beta s}\right)$ | √ | √ | | $-b + \sqrt{\frac{2\beta N b}{(\beta+\gamma)s}} - \frac{N}{s}$ |
| Case 3 | $\left[\frac{(\beta+\gamma)N}{2\beta s}, +\infty\right)$ | | √ | | $-\frac{\gamma b}{\beta+\gamma} - \frac{N}{2s}$ |

### 2.2. Fare Reward Scheme

When introducing FRS to the morning commute, this period is divided into three intervals: the central one is a uniform fare interval (UFI), and the two shoulders are the free fare intervals (FFIs). The UFI is defined as $[t_i, t_j]$, where $\lambda N$ commuters are charged a uniform fare $p$ each. In FFIs, commuters can take a free rail transit. FRS must meet the following conditions:

$$t_j - t_i = \frac{\lambda N}{s}; \qquad \lambda N p = N p_0; \qquad AEC(\lambda) \geq AEC; \quad (2)$$

where the first formula represents the travel time of $\lambda N$ commuters; the second ensures a constant revenue for the operator; and the third shows commuters in UFI cost no less than those in FFIs. Our purpose of implementing FRS is to minimize the total system cost $TTC$, which is the expected time cost of all commuters in the whole system (i.e., excluding the fares). Further, when deriving the optimal FRS, we need to select: (1) the UFI $[t_i, t_j]$; and (2) the fare ratio $\lambda$.

***Proposition 1.*** Given the fare ratio $\lambda$, the total system cost $TTC$ is minimized if and only if the commuters pass through the bottleneck at time $t_i$ and $t_j$ don't queue before the bottleneck.

***Proposition 2.*** Given the UFI $[t_i, t_j]$, the conditions of FRS (i.e., Eq (2)) can be simplified into an inequality about the fare ratio, i.e., $f(\lambda) \leq p_0$. The optimal fare ratio $\lambda^* = \text{argmin} TTC(\lambda)$,

and $\mathrm{d}TTC(\lambda)/\mathrm{d}\lambda = -N(\mathrm{d}f(\lambda)/\mathrm{d}\lambda)$.

Proposition 1 means that when deriving the optimal FRS, one only need to solve the scenario described, and choose the best fare ratio from these scenarios. Proposition 2 simplifies the process of selecting the fare ratio $\lambda$, and suggests us that $p_0$ will affect the setting of $\lambda^*$. Combing Propositions 1 and 2, we can derive the optimal FRS. Corollary 1 indicates that the implementation effect of FRS may be affected by a low initial fare $p_0$.

***Corollary 1.*** In each case, given other parameters, we can always find a reference fare $p^*$. If $p_0 \geq p^*$, the optimal FRS is implemented with only one choice of $\lambda^*$; if $p_0 < p^*$, the sub-optimal FRS is implemented with two choices of $\lambda^*$.

## 3. Numerical Studies

In this section, we use the data of three metro rail lines in Hong Kong (*Annual Report 2015*) (see Table 2). According to previous researches (Small et al., 2005; Ubbels et al., 2005) and the assumption in this study, we set $N = s = 75000$, $\alpha = 70\mathrm{HKD/h}$, $\beta = 50\mathrm{HKD/h}$, $\gamma = 100\mathrm{HKD/h}$, and $\varepsilon = 90\mathrm{HKD/h}$. Three scenarios are designed, i.e., $b = 0.1\mathrm{h}$, $0.7\mathrm{h}$, and $1.5\mathrm{h}$. The results of the optimal FRS are listed in Table 2, where $\phi$ and $\varphi$ are the percentages of total system cost reduction (system efficiency) and personal travel cost reduction (personal efficiency), respectively; $\Delta TTC$ and $\Delta AEC$ are the values of cost reduction. Comparing the results of each rail line across scenarios, the efficiencies decrease as $b$ increases. When $b$ is large (in S3), the effect of FRS is limited. However, when $b$ is moderate or small (in S1 and S2), the effect is considerable, and the system efficiency ranges from 11% to nearly 20%.

**Table 2.** The implementation of the optimal FRS.

| | Metro rail line | $p_0$ | $\lambda^*$ | $\phi$ (%) | $\varphi$ (%) | $\Delta TTC$ ($\times 10^5$ HKD) | $\Delta AEC$ (HKD) |
|---|---|---|---|---|---|---|---|
| S1 | O | 7.5 | 0.34 or 0.66 | 19.82 | 15.85 | 5.625 | 7.50 |
| | I | 11.0 | 0.50 | 22.03 | 16.39 | 6.250 | 8.33 |
| | W | 23.0 | 0.50 | 22.03 | 13.26 | 6.250 | 8.33 |
| S2 | O | 7.5 | 0.50 or 0.56 | 11.57 | 10.09 | 5.625 | 7.50 |
| | I | 11.0 | 0.53 | 11.61 | 9.68 | 5.649 | 7.53 |
| | W | 23.0 | 0.53 | 11.61 | 8.38 | 5.649 | 7.53 |
| S3 | O | 7.5 | 0.58 | 4.6 | 4.2 | 3.608 | 4.8 |
| | I | 11.0 | 0.58 | 4.6 | 4.1 | 3.608 | 4.8 |
| | W | 23.0 | 0.58 | 4.6 | 3.7 | 3.608 | 4.8 |

## 4. Conclusion and Discussion

In this study, we consider a combined rail transit and bus commuting model with random delays and solve the equilibrium cases. On this basis, we discuss the implementation effects of FRS. We use a numerical example to illustrate the results. The formulas, derivation processes, and other examples will be shown in the full paper. Considering random delays, we demonstrate that FRS is still applicable when the random delay is not so large, and present a method for finding the optimal FRS (including the uniform fare interval and the fare ratio). This study expands the application scope of FRS, and provides insights on randomness consideration in traffic management.

## References


*Annual Report 2015*. (2015). https://www.mtr.com.hk/en/corporate/investor/2015frpt.html

Currie, G. (2010). Quick and Effective Solution to Rail Overcrowding: Free Early Bird Ticket Experience in Melbourne, Australia. *Transportation Research Record*, *2146*, 35 - 42.

Kraus, M., & Yoshida, Y. (2002). The Commuter's Time-of-Use Decision and Optimal Pricing and Service in Urban Mass Transit. *Journal of Urban Economics*, *51*, 170-195. https://doi.org/10.1006/juec.2001.2242

Lan, L. W., Lee, H.-Y., & Wen, C. (2010). Effects of Temporally Differential Fares on Taipei Metro Riders' Mode and Time-of-Day Choices. *International Journal of Transport Economics*, *37*, 1000-1022.

Peer, S., Knockaert, J., & Verhoef, E. T. (2016). Train Commuters' Scheduling Preferences: Evidence from a Large-Scale Peak Avoidance Experiment. *Transportation Research Part B: Methodological*, *83*, 314-333. https://doi.org/10.1016/j.trb.2015.11.017

Small, K. A. (1982). The Scheduling of Consumer Activities: Work Trips. *The American Economic Review*, *72*(3), 467-479. http://www.jstor.org/stable/1831545

Small, K. A., Winston, C., & Yan, J. (2005). Uncovering the Distribution of Motorists' Preferences for Travel Time and Reliability. *Econometrica*, *73*(4), 1367-1382. https://doi.org/10.1111/j.1468-0262.2005.00619.x

Tang, Y., Yang, H., Wang, B., Huang, J., & Bai, Y. (2020). A Pareto-Improving and Revenue-Neutral Scheme to Manage Mass Transit Congestion with Heterogeneous Commuters. *Transportation Research Part C: Emerging Technologies*, *113*, 245-259. https://doi.org/10.1016/j.trc.2019.05.016

Ubbels, B., Tseng, Y.-Y., & Verhoef, E. (2005). Value of Time, Schedule Delay and Reliability - Estimates Based on Choice Behaviour of Dutch Commuters Facing Congestion.

Vickrey, W. S. (1969). Congestion Theory and Transport Investment. *The American Economic Review*, *59*, 251-260.

Yang, H., & Tang, Y. (2018). Managing Rail Transit Peak-Hour Congestion with a Fare-Reward Scheme. *Transportation Research Part B: Methodological*, *110*, 122-136. https://doi.org/10.1016/j.trb.2018.02.005

# Intelligent Video Analytics for I2V communication: an experimental approach based on dynamic traffic assignment

Marisdea Castiglione*[1] and Marialisa Nigro[1]

[1]Department of Civil, Computer Science and Aeronautical Technologies Engineering, Roma Tre University, Italy

**Keywords**: I2V communication, control room, mixed traffic flow, road safety, computer vision.

## 1 Introduction

The Intelligent Video Analytics (IVA) project, sponsored by Lazio Region, aims to redefine traffic safety by integrating advanced Artificial Intelligence (AI) with 5G connectivity. Its primary goal is to develop a real-time infrastructure-to-vehicle (I2V) communication system to enhance road safety. Through real-time video processing, IVA detects anomalies like accidents, queue formations, and dangerous driving behaviors, featuring an innovative hardware module equipped with an AI-accelerated processing board and a 5G modem which enables localized data analysis and rapid transmission to Mobile Edge Computing (MEC) infrastructure.

The algorithms developed address scenarios such as accident detection, wrong-way driving, queue formation, and fire detection, while also analyzing vehicle trajectories to identify risky behaviors. A deep learning-based trajectory tracking model ensures precise tracking and handles complex scenes with high accuracy (Aharon et al., 2022). A convolutional neural network extracts features and uses matching algorithms to associate detections with existing tracks (Zhang et al., 2021), while advanced optimization techniques (Wojke et al., 2017) ensure consistent tracking even in challenging conditions involving occlusions or overlaps.

IVA's event detection and trajectory tracking capabilities can be an essential tool for traffic control rooms and for implementing I2V communication, enabling real-time communication to road users. In fact, immediate alerts reduce risks by helping drivers make informed decisions, avoid hazardous areas, and adjust routes in real-time. Moreover, the rapid dissemination of information minimizes delays and enhances overall traffic flow, contributing to a safer and more efficient transportation network.

This paper investigates the impact of communicating detected events to road users, focusing on the definition of optimal information configuration in terms of spatial distribution, temporal reactivity, and the type of information provided. The study develops simulation scenarios using dynamic traffic assignment models that incorporate mixed traffic flows, encompassing both human-driven and autonomous vehicles, to evaluate how different communication strategies affect traffic safety and congestion. The paper is structured as follows: the Methodology section defines the communication scenarios; the Results section presents simulations conducted in the EUR district of Rome, Italy; and the Conclusion section summarizes key findings and proposes directions for future research.

## 2 Methodology

The objective of this analysis is to compare a baseline scenario, without communication interventions, to project scenarios that leverage advanced technologies, including I2V communication systems, for detecting and communicating traffic conditions (such as the IVA module). Simulations are conducted to assess the impact of various information solutions on traffic flow and safety, modeling both human-drivers and autonomous vehicles' response in a dynamic traffic environment. The goal is to identify the conditions and parameters that maximize the effectiveness of the proposed interventions.

Several parameters are considered for scenario creation, including cameras distribution, congestion levels, event duration, and user communication coverage. A baseline Scenario 0 models normal traffic conditions using Dynamic User Equilibrium Assignment, representing the natural distribution of users on the network under standard conditions, serving as a benchmark to compare subsequent scenarios. High-risk hotspots, identified through historical accident data, are analyzed under scenarios that simulate incidents and lane reductions with varying severities, including full road closures.

The "project" scenarios incorporate combinations of different congestion levels (standard conditions and 30% increased demand), event durations (30 and 60 minutes), and incident severities (from 0% to 100% capacity reduction). Vehicle information penetration rates are varied from 0% to 100%, assuming full network-wide information rather than localized communication. Key performance indicators are collected both system-wide and for vehicles directly impacted, providing a comprehensive understanding of network performance under extraordinary conditions.
Then, simulation scenarios with targeted communication on specific incidents are designed to evaluate the effectiveness of direct user notifications. These scenarios are analyzed across four key aspects:

- **Spatial Distribution**: Evaluates how the geographical coverage of notifications impacts the vehicles response;
- **Temporal Reactivity**: Assesses the timeliness of communication by analyzing user responses based on how quickly the information reaches them. Two primary scenarios are considered: 1) Real-time communication via infrastructure-to-vehicle (I2V) systems, where vehicles are notified at the exact moment the incident occurs, simulating the immediate response of the IVA system; 2) Delayed communication, where I2V systems are integrated with centralized processes, simulating scenarios where incident detection, alert transmission to a control room, and decision-making processes introduce a range of different delays;
- **Type of Information Provided**: Explores the effectiveness of three types of communication strategies in diverting traffic and reducing congestion: 1) Congestion warning, allowing users to voluntary deviations to better routes, if available; 2)Optional detour, suggesting alternative routes but leaving users free to choose whether to follow the detour or stay on the original route; 3) Mandatory detour, enforcing compulsory deviations.
- **Vehicle Reaction**: Analyzes the behavior of human-driven and autonomous vehicles in response to the provided information. Differences in compliance, route adjustment, and overall network impact are considered to highlight the implications for mixed traffic environments.

The modeling of Temporal Reactivity, Types of Information, and User Reactions is based on consolidated methodologies implemented in the Dynasmart dynamic traffic assignment model, which serves as the foundation for these scenario developments. Instead, the phase of Spatial Distribution of incident communication involves the following steps: 1) Identifying the incident location; 2) Defining the attention radius; 3) Identifying affected road links; 4) Selecting strategic road links for communication; and 5) Iteratively optimizing communication configurations.

Using GIS tools, the incident location is mapped on the OpenStreetMap (OSM) network. A buffer-based approach with adjustable radii is then used to define the area of attention, representing potential traffic impact zones where users may be influenced. The radius is iteratively refined based on simulation results to maximize the effectiveness of communication while minimizing disruptions. Affected road links are identified through spatial and directional analysis: spatial queries determine which road links intersect the buffer, and directional analysis ensures only links directing traffic toward the incident are selected. A connectivity propagation method is applied to build a tree of connected segments leading to the incident, systematically capturing directly and indirectly affected links:

- **Initialization**: Given the road network $G$, the set of affected links $L_k$ is initialized at iteration $k = 0$ with the link where the incident occurs, $l_{ij}$, where $i$ and $j$ represent the upstream and downstream nodes of the link, respectively:

$$L_{k=0} = \{l_{i,j}\} \tag{1}$$

- **Propagation**: For subsequent iterations ($k \geq 1$), the set of affected links $L_{k+1}$ is expanded as follows:

$$L_{k+1} = L_k \cup \{l_{m,i} : \forall i \in l_{i,j} \in L_k\} \tag{2}$$

  Here: - $L_k$ is the set of affected links at iteration $k$. - $l_{m,i}$ represents any link in the network where $m$ is the upstream node, and $i$ is the downstream node of any link $l_{i,j} \in L_k$.

- **Termination**: The process continues iteratively untilno new links are added to the set.

$$L_{k+1} = L_k \tag{3}$$

The final set $L_f$ includes all links leading toward the incident site. Lastly, strategic links for communication are identified to enhance the efficient dissemination of information. For each attention radius, links in the network with an hourly traffic flow exceeding a defined threshold $S_l$ are prioritized. If the number of qualifying links falls below a specified minimum $S_{min}$, additional links with traffic flows slightly below are included to meet the required minimum number of links per buffer. As the radius expands, links selected in smaller buffers are retained to ensure continuity and consistency in the communication strategy.

## 3 Results

The selected study area for simulation scenario is the EUR district of Rome. This area covers a total of 51 km$^2$ and is divided into 54 traffic zones, as depicted in Figure 1(a), which illustrates the adopted zoning system. The road network in the study area consists of 400 nodes and 812 links, providing an interconnected structure suitable for testing various urban mobility scenarios. For the simulations, an origin-destination (OD) matrix containing 2,916 OD pairs is used, representing the travel demand within the area. The simulation scenarios span 4 hours of the morning peak period, from 08:00 to 12:00. This includes a pre-load phase of 30 minutes and a network unloading phase of 1 hour to ensure a realistic representation of traffic dynamics.

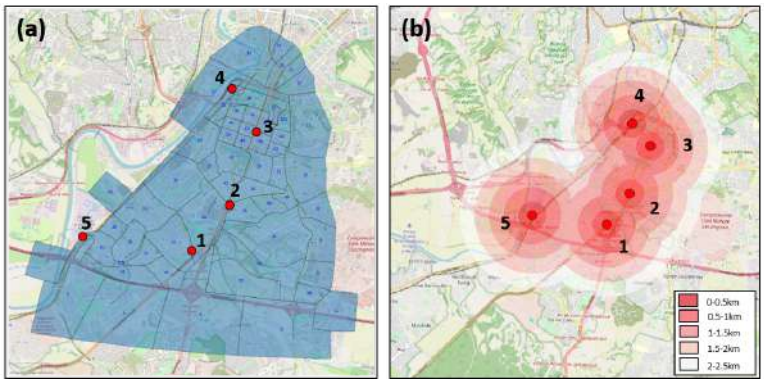



Figure 1: The EUR case study: (a) Zoning system (b) Incident Hotspots

Localized analyses were conducted based on identified incident hotspots, shown in Figure 1(b). These hotspots were selected due to their high potential for traffic disruptions, analysing the frequency of traffic incidents and their severity. Surrounding each hotspot, spatial buffers of varying radii were applied to capture the extent of the incident's impact on the network and to assess the effectiveness of different communication strategies in mitigating congestion and delays. Preliminary results analyze the impact of real-time traffic information on the road network in the presence of an incident. The objective is to compare network performance, both overall and for vehicles directly affected by the incident, by varying the percentage of vehicles receiving real-time information from 0% to 100%.
To ensure consistency across scenarios, the following baseline conditions are maintained: 1) Regular traffic congestion, representing a non-overloaded network; and 2) A fixed incident duration of 60

minutes, simulating a significant disruption. Figure 2 illustrates how the average travel times decrease and speeds improve as a higher percentage of vehicles receive real-time information. These results demonstrate the benefits of real-time communication in enhancing overall network performance. The study will also present the outcomes of a localized analysis, focusing on scenarios where punctual communication of the event is provided in real-time to vehicles potentially impacted by the incident, offering insights into the effects of targeted communication strategies on traffic flow and delays.

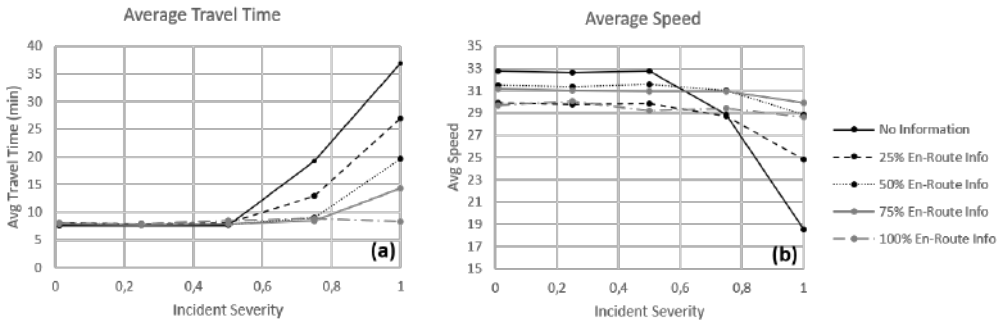



Figure 2: (a) Average Travel Times and (b) Speeds with varying levels of en-route information.

## 4 Conclusion

This study demonstrates the significant benefits of real-time traffic information, facilitated by direct I2V communication, in mitigating the impact of incidents or hazards on urban road networks. Using dynamic traffic assignment models, the analysis of the developed simulation scenarios highlights that increasing the percentage of users receiving real-time information reduces travel times and improves traffic flow. Localized analyses further reveal the effectiveness of targeted communication strategies in minimizing delays and optimizing traffic distribution in affected areas. These findings underscore the importance of timely and accurate I2V communication and information dissemination, paving the way for more efficient and resilient transportation systems. Future research will focus on refining specific communication strategies for each localized incident hotspot.

## References

Aharon, N., Orfaig, R., & Bobrovsky, B.-Z. (2022). Bot-sort: Robust associations multi-object tracking by global optimization. *arXiv*, *abs/2206.14651*. Retrieved from https://arxiv.org/abs/2206.14651 doi: 10.48550/arXiv.2206.14651

Wojke, N., Bewley, A., & Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *2017 ieee international conference on image processing (icip)* (pp. 3645–3649). doi: 10.1109/ICIP.2017.8296962

Zhang, Y., Wang, C., Wang, X., Zeng, W., & Liu, W. (2021, September). Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, *129*(11), 3069–3087. Retrieved from http://dx.doi.org/10.1007/s11263-021-01513-4 doi: 10.1007/s11263-021-01513-4

# DTA 2025

# A dynamic demand assignment in an underground transportation system for time-varying train loading calculation

Lorenzo Mussone[1*], Roberto Notari[2]

[1] *Department ABC, Politecnico di Milano, P.za L. da Vinci ,32, 20133 Milano*
[2] *Department of Mathematics, Politecnico di Milano, Via Bonardi, 9, 20133 Milano*
* *corresponding author, e-mail: lorenzo.mussone@polimi.it*

**Long Abstract**

Short abstract
The aim of this study is to develop a tool that can compute the train load throughout its route while accounting for time-dependent demand and potential failure or cancellation. We achieve this through a few stages of activity. First, we suggest a dynamic study of the underground transit system of Milan, Italy. With four lines, 110 links, and 107 stations—including 21 junctions, sometimes referred to as transfer stations—this transport system serves as the city's main hub for mobility. The flows on trains and links are calculated with a one-minute resolution using two data sets concerning passenger movements (both entering and leaving stations). The data sets pertain to a week in 2018, when the pandemic scenario had no effect on demand. After that, an ad hoc written assignment procedure is used to process the data. To examine how the exposure or significance of each station varies over time and how the train load varies from station to station, the results of those computations are displayed at various aggregate intervals. Changes become increasingly smoother as the observation interval is increased from one minute to thirty minutes, resembling the use of a low-pass filter. This suggests that aggregating data may result in inaccurate findings for some applications (such as security-related ones).

***Keywords**:* Underground networks evaluation, flow assignment, dynamic analysis, train load, segment load, station flows

## 1. Introduction

Undoubtedly, the analysis of the topology of underground networks is an intriguing and useful first step in comprehending their behavior and operation. The relationship between supply and demand actually determines the operational state, whether it is dynamic or in equilibrium, as it does in all other transportation systems. Thus, demand and supply (including topology) can be combined into a special framework that enables a thorough explanation of the network's operation. This work wants to answer to the following questions: "How can the train scheduling be changed to avoid saturation of train capacity?" or, in a similar vein: "How many additional trains must be scheduled to achieve a certain level of service?". Or alternatively, "How many train trips can be cancelled before experiencing saturation phenomena?".

Finding stations, links, and trains that are more vulnerable to malfunctions or operating decreases due to the heavy load on them, is then one of our goals. However, in the event of an interruption, we would like to know how well a network is performing. The fact that demand varies throughout the day is another factor. Therefore, a time-dependent analysis must be performed in order to gather all the data needed for managing and controlling the transportation network. Determining an appropriate time frame is definitely challenging since it must guarantee comprehensive findings while minimizing the computational load.

In this research, in describing an underground network, we devise a general approach that takes supply and demand into account. We require the network graph, the hourly passenger flow entering and leaving the network, and the time-dependent Origin-Destination (OD) matrix. We utilise this methodology to examine the Milan underground network as an example. Actually, this approach can be used in any kind of transport network as long as the scheduled service and all passenger movements (entering and exiting the stops) are known. Naturally, the

conclusions reported here are only relevant to the Milan case study because they rely on the particular OD distribution, time, and quantity demands.

This study's demand is based on data from a week in 2018 and includes the flow of people coming into and going out of Milan's underground stations. In order to assign each passenger (accessing the station) to a certain route that will get them to their ultimate destination (the leaving station), a trip assignment must be completed. The automated fared data (tickets that show the real trip data for each passenger) that is gathered from the service provider is used to build the OD matrix.

After assigning each passenger to a route, we calculate the load on each edge of the graph based on time in order to obtain a dynamical image of the operational network, as well as the load of each train over time and edges.

## 2. The overall methodology

The steps in the research approach are as follows:

- Preparation of basic tools: graph and datasets.
- Data on passenger flows are examined and reviewed for consistency and missing data,
- Hourly demand (the turnstiled data) is separated into intervals of one to ten minutes,
- Demand is assigned to the network,
- Passenger flow is computed by station, segment, and train.

### *2.1. The graph*

Due to data collected in 2018, Milan's underground network is comprised of four lines (red M1, green M2, yellow M3, and lilac M5). The blue line M4 was opened in 2022 but eventually completed in 2024. The four lines have 121 stations in all. In topological terms, several stations are the same transfer station since they have direct links between crossing lines. Therefore, there are 107 stations and 110 links on the graph at the conclusion. With the code for each station, Fig. 1 shows the entire Milan underground network graph.

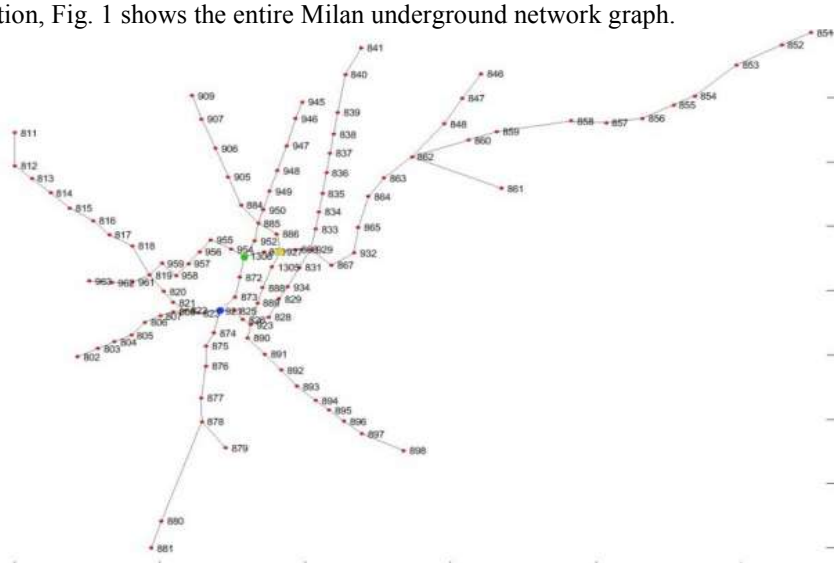



Fig. 1. The graph of the underground network of Milan, Italy (the three primary transfer stations are denoted by nodes of different colours: blue for Cadorna, green for Porta Garibaldi, and yellow for Centrale station).

### *2.2. The data sets*

Two data sets concerning passenger flows (for both access and egress - Entry and Exit - from stations) are used and pertain to the week of April 9-15, 2018 (a scenario before the pandemic), together with one concerning the service. Specifically:

1. Data from turnstiles (“Turnstiled”), aggregated by one hour,
2. Data from tickets (“Fared”), giving the time (in minutes) the ticket was detected entering and exiting;
3. The timetable for every scheduled train.

### *2.3. The assignment procedure*

An assignment procedure is used to determine the flow on edges and passengers on trains. A station's schematic

layout is depicted in Fig. 2. We made the assumption that no journeys lasted more than an hour. The study of fared data and the several trials conducted revealed that this assumption is generally true, up to some scattered case. As shown in Fig. 3, we must thus assign two hours of demand in order to create a full simulation of one hour. Additionally, it's important to remember that every line is processed simultaneously.

The entire process is predicated on the idea that every subject works as efficiently as possible: trains are consistently on schedule, and every passenger gets to the tracks and back in the given time. Because of this, it's possible that passenger arrival times will be earlier than actual ones. A description of the functions developed for these tasks is shown in Fig. 3.

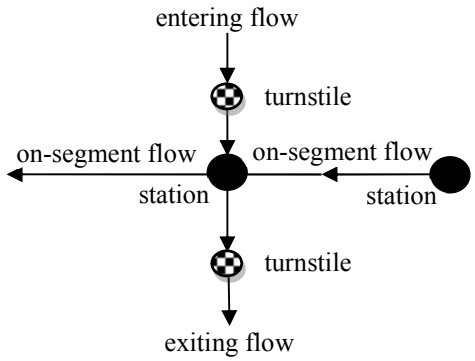



Fig. 2: Schematic representation of passenger flows at a station node.

|----------1h----------|---------2h---------|----------3h-------|
|-----warm up-------|**----simulated-----**|---for balance---|
|-----assignment----|----assignment----|

Fig. 3: Warm-up, assignment, and balance intervals in the assignment procedure.

```
for each station
    calculate access time to track
    extract entry data from Turnstiled dataset (one hour aggregated)
    apply REBO6 programme to get 1 minute data
    select OD matrices for the two hours of assignment
    for each minute (1:120 [warms up and simulated intervals])
        distribute 1minute demand to destinations through a Monte Carlo routine
        for each destination
            calculate egress time
            extract the best route and the list of stations belonging to it
            for each route find the segments between stations
                calculate the earliest departure time
                find the first useful train
                calculate the travel time (and transfer times, if any)
                update the destination matrix (it is function of time)
                load each segment with the assigned demand (according to the travel time)
                update the train's load
            end
        end
    end
```

Fig. 4 Meta-language program of the assignment procedure.

*2.4. The REBO6 programme*

For each station and day of the week, the turnstiled data consists of hourly data of entries or egresses. The idea behind REBO6 programme is that entries every 10 minutes can be obtained from entries every 60 minutes as the result of three different processes: (a) 67.5% of the entries are uniformly distributed every 10 minutes; (b) 22.5% of the entries are allocated to the 10 minute slots through a random probability modelled according to the total number of entries during the hour before and next the current hour; (c) 10% of the entries are allocated to the 10 minute slots according to a completely random probability vector. The percentages have been chosen to be in good alignment with fared entries.

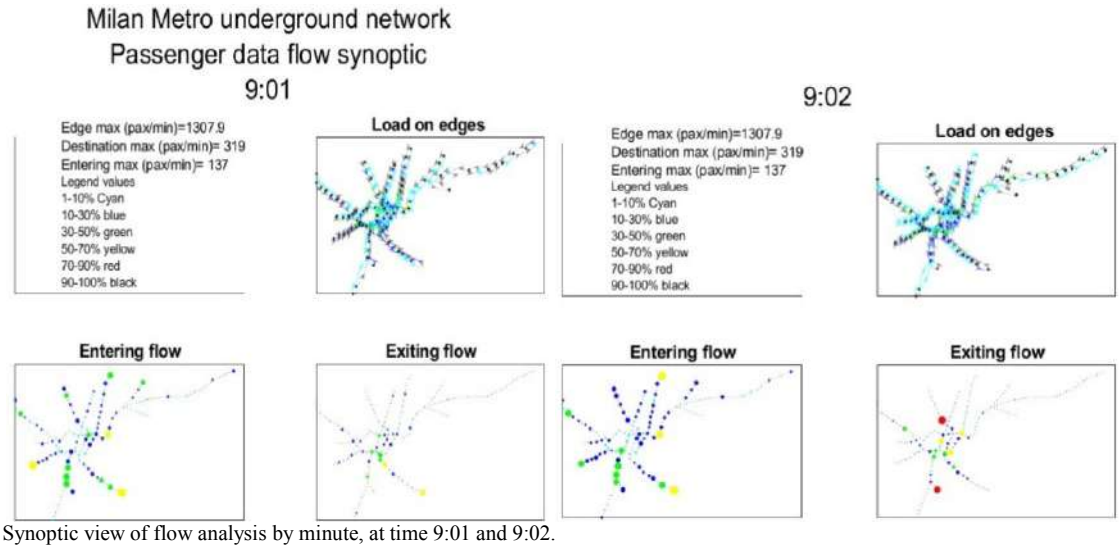



Fig. 5: Synoptic view of flow analysis by minute, at time 9:01 and 9:02.

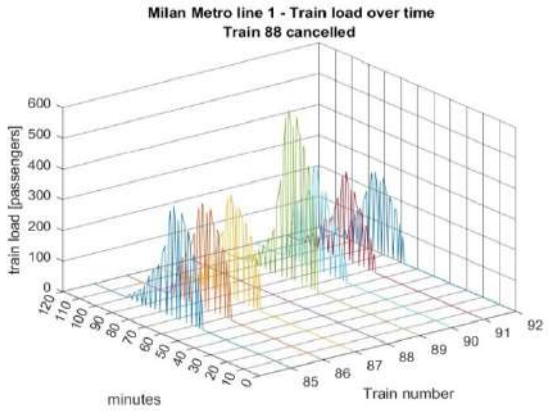



Fig. 6: Train load over time: effect of train (number 88). Cancellation

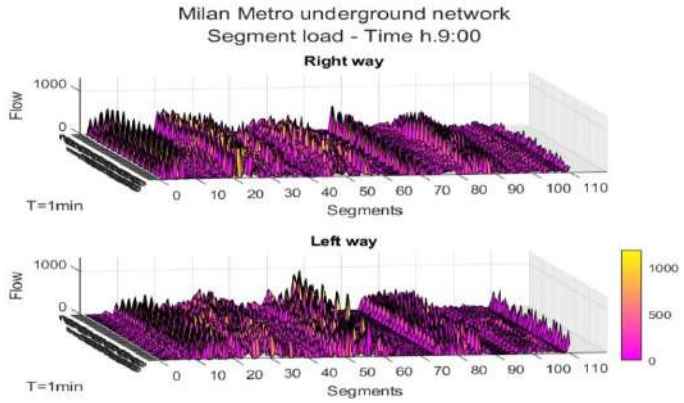



Fig. 8: Flows over one hour by steps of one minute on the two ways of the segments from 9:00 to 10:00 a.m.

## 3. Results

The assignment process enables us to analyse every simulation result. Results are shown in two different ways: graphically and through tables and statistics.

Initially, a synoptic graphic shows three subplots, each of which has the Milan graph that shows the incoming flow, the outgoing flow by node, and the flow on the edges. The synoptic charts are updated based on the results, which can be generated at intervals of one to sixty minutes (Fig. 5 shows an example of two successive frames at one minute). Second, the train load is shown graphically along with the segment load variation over time, departure flows, and entering flows (Fig. 6).

Thirdly, during the course of an hour, a 3D figure (Fig. 7) shows the load for each part for the left and right way of circulation.

## 4. Conclusions

This study proposes a system for gradually selecting stations and trains in an underground network for which the effects of a disruption would be more significant. The main focus of contributions to transportation planning is the understanding of the train load factor for each segment per time interval with a minimum of one minute, followed by the train schedule improvements. We concentrated on underground networks since they provide us with all the information available regarding passenger flows, but their application to other transportation systems is unrestricted. The same justification, however, pertaining to data availability applies to the Milan instance. Of course, the format of the data may affect how they are pre-processed.

With a time resolution of 1 to 30 minutes, the methodology is based on a dynamic approach that considers the change in demand (ridership) over time. To replicate the assignment procedure and recover the four demand components, entry and exit from the stations, segment load, and train load, specific programs are put into place. It is important to note that required evaluation approaches, e.g. using statistical indicators like mean or variance, are not particularly appropriate for this type of analysis because they tend to filter out singular phenomena, like a peak in passengers, which are essential for our goals and must be disclosed.

A variety of aggregation times were used to test the effects of passenger flows in an underground network. Although they are connected, each of them provides distinct information. So, it is crucial to understand what it means to assume a specific period for data aggregation. This problem is also addressed in the study by examining the composition and temporal development of passenger flows, including those that indicate the load on segments and on-train, and those that arrive and exit the stations. In terms of their dynamics, these four passenger flows represent unique processes, and the suggested method demonstrated the ability to evaluate them.

Implications for planning concern mainly the use of the overall methodology to simulate, and then to study how, by changing the train's scheduled service, the train passenger load changes. This has two fallouts: first, the passenger flow at stations and platforms can be controlled and dimensioned opportunely; second, the relevance of stations can be distributed homogeneously among stations avoiding or mitigating critical concentrations only in a few stations.

Future research will address:

- Improving the assignment procedure with other models to create the percentage OD matrix;
- Considering the train schedule times and the passenger times (from turnstile to the platform and vice versa, and transfer times) stochastic;
- Taking into account saturation of train vehicles (though no cases were observed in the Milan underground system), this features would make the procedure still more general.

# Introducing the Flex-GLS Framework: Harnessing Crowd-Sourced Flexibility Data to Enhance Accuracy in Dynamic Origin-Destination Matrices Estimation

Marisdea Castiglione*[1], Guido Cantelmo[2], Ernesto Cipriani[1], and Marialisa Nigro[1]

[1]Department of Civil, Computer Science and Aeronautical Technologies Engineering, Roma Tre University, Italy
[2]Department of Technology, Management and Economics, Technical University of Denmark, Denmark

**Keywords**: Dynamic OD Matrices Estimation; Spatio-Temporal Flexibility; Crowd-Sourced Data; Floating Car Data; Trip Purpose.

## 1 Introduction

In response to the rapidly evolving challenges of urban environments, there is a growing necessity to enhance traditional Origin-Destination Matrices Estimation (ODME) models with data sources that provide broader and more comprehensive insights (Cantelmo et al., 2014). Traditional fixed-location data collection tools, foundational in developing reliable traffic metrics, often fall short of capturing the complex nature of travel demand (Carrese et al., 2017). Crowdsourced data, which includes mobile phone data, GPS-based data, and social media analytics, offers valuable avenues for obtaining high-resolution information that accurately reflects urban travel patterns. Particularly, the integration of location-based crowdsourced data into ODME models can significantly enhance our understanding of the activities and purposes of travelers' trips (Timokhin et al., 2020).

Previous research (Castiglione et al., 2024) delved into the motivations behind individual trips, assessing how travel flexibility is influenced by the type of activities performed at destinations, utilizing sources such as Floating Car Data (FCD) and Google Popular Times (GPT). These studies revealed that flexibility parameters vary across different activities and over time, providing detailed insights into spatio-temporal flexibility for various components of travel demand, such as rigid or flexible demand components. Specifically, six demand components $C$ were identified, each characterized by distinct levels of temporal and spatial flexibility, and represented through a series of sample OD matrices derived from the aggregation of FCD trips based on their spatio-temporal flexibility. This body of evidence underscores the critical importance of integrating crowd-sourced data into the framework of established dynamic ODME models like the Generalised Least Squares (GLS) model (Cascetta et al., 1993), enhancing their accuracy in modern urban settings.

This paper introduces the Flex-GLS approach, a novel adaptation of the GLS model designed to incorporate multiple demand components characterized by detailed spatio-temporal flexibility metrics derived from real-world, crowdsourced data. The Flex-GLS model utilizes a dynamic traffic assignment to efficiently allocate these identified demand components across the network, reflecting realistic traffic conditions and optimizing route choices in real time. This approach aims to more accurately depict travel demand by blending temporal and spatial flexibility measures, thereby more effectively capturing the complex dynamics of urban travel. The practical implementation of the Flex-GLS model is exemplified through a case study in the EUR district of Rome, Italy, utilizing an extensive FCD dataset that recorded over 1.5 million trips between September and December 2020.

## 2 Methodology

Temporal Flexibility (TF) and Spatial Flexibility (SF) define an individual's ability to adjust the timing and locations of their activities, respectively. In this context, the traditional GLS model is extended to

encompass multiple demand components $C$, each characterized by its unique spatio-temporal flexibility distribution $\sigma_C$. Given $n_{t,C}$ sample $OD$ matrices, where each cell represents trips from an origin $O$ to a destination $D$ within a time interval $t$ for a travel demand component $C$, derived from the aggregation of crowd-sourced data based on shared spatio-temporal flexibility, as detailed in Castiglione et al. (2024), the modified GLS objective function is provided as follows:

$$d^* = \arg\min_d \left( \sum_t \left( \sum_l w_l \cdot (v_l(d) - \hat{v}_l)^2 + \sum_{od} \sum_C w_C \cdot (d_{od,C} - \hat{d}_{od,C})^2 \right) \right) \tag{1}$$

Here, $v_l(d)$ denotes simulated traffic flows from a certain demand matrix $d$, against observed traffic counts $\hat{v}_l$. $d^*$ represents the estimated matrix that minimises the discrepancy between simulated and observed flows. Any generic demand matrix $d$ can then be segmented into $C$ components where $d_{od,C}$ indicates the demand for each $OD$ pair per component, while $\hat{d}_{od,C}$ is the seed matrix for each demand component obtained from the classified FCD in Castiglione et al. (2024). The weights $w_l$ and $w_C$ are assigned based on the inverse of traffic counts and demand component variances, respectively. Incorporating multiple demand components, while straightforward conceptually, significantly complicates the estimation process, especially for large urban networks. The Flex-GLS model, however, addresses this complexity by using conditional probabilities to treat demand components as one composite $OD$ variable, ensuring computational efficiency. Demand components are thus defined as:

$$\begin{cases} d_{C,t,od} & = p_C(t) \times d_{t,od} \\ p_C(t) & = \frac{d_{C,t}}{\sum_t d_{C,t}} \end{cases} \tag{2}$$

Where $p_C(t)$ is obtained from the classified FCD sample OD matrices. The Flex-GLS model utilizes a gradient descent algorithm to estimate the demand for each $OD$ pair and time interval $t$. At every gradient descent step, the model dynamically assigns the overall demand onto the network, and then refines the individual demand components through a constrained Maximum Likelihood Estimation (MLE) problem, leveraging prior probabilities $P_C(t)$ and variances $\sigma_C^2$ based on seed FCD data.

The MLE constraints in the model aim to ensure data consistency, with Temporal and Spatial Flexibility treated as complementary. Temporal constraints allow adjustments in demand component proportions within a time interval $t$, while ensuring overall consistency across a temporal window $T$. This is critical for accurately capturing variations in travel behavior, considering narrower time windows for commuters versus broader windows for other, more flexible activities (e.g. shopping). Similarly, Spatial Flexibility enables the redistribution of demand from one origin $O$ to various destinations within the same time interval, maintaining, however, the proportionality in demand components. The spatio-temporal MLE problem constraints are thus formalized as:

$$\begin{cases} \sum_C P_C(t) = 1 & \forall t \in T, \forall od \\ \frac{\sum_{t \in T} \sum_d d_{C,t,od}}{\sum_{t \in T} \sum_d d_{t,od}} \approx \frac{\sum_{t \in T} \sum_d d_{C,t,od}}{\sum_{t \in T} \sum_d d_{t,od}} & \forall C, \forall t \in T, \forall d \end{cases} \tag{3}$$

## 3 Results

For benchmarking purposes, initial tests of the Flex-GLS model were conducted on a toy network with one origin and two destinations to evaluate its performance across different scenarios of data reliability. These preliminary tests explored various alignments and deviations of seed data from real conditions, preparing for a more robust application. The findings inform strategic choices of modeling approaches, advocating for the Flex-GLS when detailed component data is available and for standard GLS in scenarios where data reliability may be compromised. This section presents the results from applying the Flex-GLS model to a comprehensive case study of the EUR district in Rome, Italy. The EUR district, encompassing 51 km$^2$ with 54 traffic zones (Figure 1[a]), provides a complex real-world environment for testing the model's effectiveness. Figure 1[b] shows the road network along with the locations of 8 traffic count detectors, highlighting strategic points for data collection and traffic monitoring.

The primary data source for this analysis has been a FCD dataset, which includes 1.5 million car trips recorded between September and December 2020 in the Metropolitan City of Rome. A subset of 180,000 trips with destinations within the study area has been classified into 'Home', 'Work', and 'Other' categories using rule-based, spatial clustering techniques to identify regular trip patterns. Additionally, activity data was obtained from Google Popular Times for 752 Points of Interest (POI) located within the study area, collected in December 2020. From this data, six demand components were identified, each with its respective spatio-temporal flexibility distributions: Home, Work, Services (MA1), Sustenance (MA2), Shopping (MA3), and Drop-Off/Pick-Up (DO-PU). These components were analysed according to the procedure detailed in (Castiglione et al., 2024).

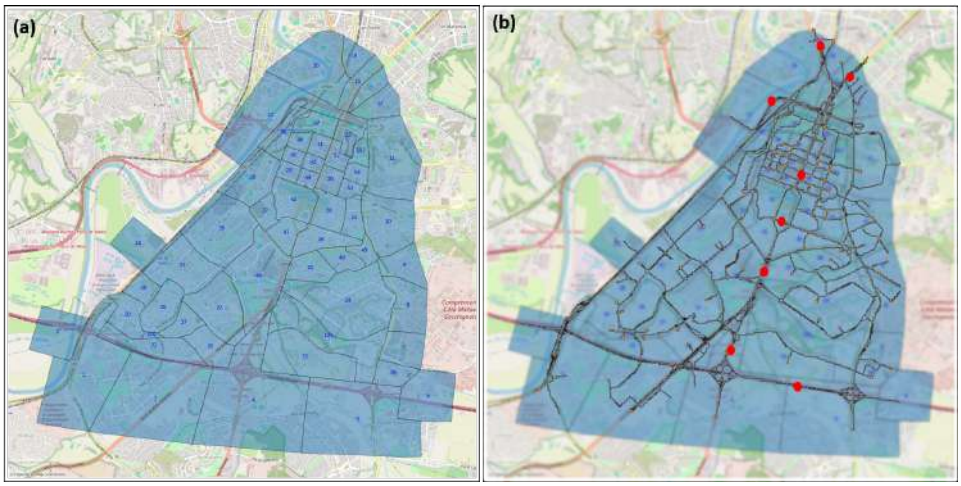



Figure 1: Case study: (a) Traffic Zones (b) Road Network and Detectors of the Eur district of Rome

For the real network tests, the dynamic traffic assignment simulator Dynasmart was employed to obtain the traffic flows for each time interval, allowing for a detailed analysis of network dynamics under various demand scenarios. Both the Flex-GLS and the standard GLS models were tested across 16 time intervals from 08:00 AM to 12:00 PM, with each interval lasting 15 minutes. Each of the six components was evaluated over specific time windows to reflect varying temporal dynamics: a time window $T = 4$ time intervals for 'Home' and 'Work'; $T = 8$ time intervals for 'Services' (MA1) and 'Drop-Off/Pick-Up' (DO-PU); and $T = 16$ time intervals for 'Sustenance' (MA2) and 'Shopping' (MA3).

The performance of both models in terms of RMSE is highlighted in two subsequent tables. Table 1 below presents the comparative results of both the Flex-GLS and standard GLS models from the initial to the last iteration in terms of RMSE of demand estimation accuracy and link flow reproduction. Following the initial comparison, Table 2 provides detailed performance results discretized into the six demand components, highlighting their respective abilities to handle the spatio-temporal flexibility inherent in each category.

Table 1: RMSE comparisons between Flex-GLS and GLS models

| **RMSE** | **GLS** | **Flex-GLS** |
|---|---|---|
| Detected Counts vs. Simulated Flows (Initial) | 84.8 | 84.8 |
| Detected Counts vs. Simulated Flows (Final) | 33.0 | 25.3 |
| Real vs. Seed Demand | 19.1 | 19.1 |
| Real vs. Estimated Demand | 13.2 | 7.6 |

Table 2: Performance comparison across demand components

| **RMSE** | **GLS - Real** | **Flex-GLS - Real** |
|---|---|---|
| Home | 4.3 | 3.5 |
| Work | 1.4 | 0.6 |
| DO-PU | 3.6 | 1.2 |
| MA1 | 5.8 | 4.2 |
| MA2 | 5.1 | 2.7 |
| MA3 | 7.2 | 3.3 |

## 4 Conclusions

This paper presents the Flex-GLS model, an enhancement to the traditional GLS framework for ODME, demonstrating the significant potential of integrating crowd-sourced data. This study validates the model in a real-world setting within the EUR district of Rome, Italy, employing crowd-sourced flexibility data obtained from Floating Car Data and Google Popular times in previous research. The Flex-GLS model successfully utilizes this data to estimate six distinct demand components — Home, Work, Drop-Off/Pick-Up, Services (MA1), Sustenance (MA2), and Shopping (MA3), each characterized by unique spatio-temporal flexibility patterns. The results highlight the model's capability to outperform traditional GLS models by leveraging nuanced insights on the demand structure. This approach not only enhances accuracy in estimating traffic flows but also adapts to the variability inherent in urban travel. Particularly in the EUR district, where traffic data coverage spans less than 1% of links, the Flex-GLS model demonstrates robustness in its estimates, even with minimal traffic data. This feature is invaluable for extensive urban networks where comprehensive data collection poses significant challenges.

## References


Cantelmo, G., Cipriani, E., Gemma, A., & Nigro, M. (2014, June). An Adaptive Bi-Level Gradient Procedure for the Estimation of Dynamic Traffic Demand. *IEEE Transactions on Intelligent Transportation Systems*, *15*(3), 1348–1361. Retrieved 2024-01-17, from https://ieeexplore.ieee.org/abstract/document/6740860 (Conference Name: IEEE Transactions on Intelligent Transportation Systems) doi: 10.1109/TITS.2014.2299734

Carrese, S., Cipriani, E., Mannini, L., & Nigro, M. (2017, August). Dynamic demand estimation and prediction for traffic urban networks adopting new data sources. *Transportation Research Part C: Emerging Technologies*, *81*, 83–98. Retrieved 2024-01-17, from https://linkinghub.elsevier.com/retrieve/pii/S0968090X17301432 doi: 10.1016/j.trc.2017.05.013

Cascetta, E., Inaudi, D., & Marquis, G. (1993, November). Dynamic Estimators of Origin-Destination Matrices Using Traffic Counts. *Transportation Science*, *27*(4), 363–373. Retrieved 2024-01-17, from https://pubsonline.informs.org/doi/10.1287/trsc.27.4.363 doi: 10.1287/trsc.27.4.363

Castiglione, M., Cantelmo, G., Cipriani, E., & Nigro, M. (2024). From trip purpose to space-time flexibility: A study using floating car data and google popular times. *Transportmetrica B: Transport Dynamics*, *13*(1). Retrieved from https://doi.org/10.1080/21680566.2024.2440596 doi: 10.1080/21680566.2024.2440596

Timokhin, S., Sadrani, M., & Antoniou, C. (2020, August). Predicting Venue Popularity Using Crowd-Sourced and Passive Sensor Data. *Smart Cities*, *3*, 818–841. doi: 10.3390/smartcities3030042

# Marginal Congestion Costs, Implications in a Dynamic Context

Jeroen Verstraete – KU Leuven

Chris Tampère - KU Leuven

## 1. Introduction

This paper focuses on understanding marginal congestion costs in more depth and the implications it has in practice. In transportation systems, marginal costs represent the additional impact caused by introducing one more traveler into the network. These costs can take various forms, such as increased travel times or higher emissions, and are often converted into monetary terms for consistency and comparison. Marginal congestion costs specifically refer to the additional travel time costs imposed on the system by an extra traveler, typically expressed in units of time. The term "social" reflects the impact on the entire system, including all affected travelers (Mayeres et al., 2005).

Focusing on marginal costs, rather than solely on absolute costs, is gaining increasing attention in transportation research. While individual travelers typically consider only their own costs, marginal costs provide valuable insights into the broader traffic system. They highlight which routes impose the greatest additional cost on society per extra traveler, offering a clearer understanding of the current traffic state. This perspective also enables the identification of socially optimal routes—those with the lowest total marginal cost—and facilitates analysis of strategies to encourage travelers to adopt these routes for the benefit of the entire system (Gavanas et al., 2017; Krichene et al., 2018).

First in section 2, the basic theoretical framework is discussed. In section 3, some more complexity is added before theoretical conclusions and implications to practice are discussed in section 4. In the full paper, a reflection is made on how marginal social congestions costs are calculated in practice (including static and dynamic traffic models) and how these calculations differs from theory.

## 2. Theoretical Framework on Basic Cases

In this framework, small extra delays in free flow conditions are neglected. Only when a queue appears, adding an extra traveler further increases the travel costs for all others passing through the bottleneck. The theoretical framework uses traffic flow theory, first order LWR models (Lighthill and Whitham, 1955; Richards, 1956). Let start by analyzing the most basic case, a single bottleneck with no other on-or offramps.

### Single Bottleneck

As a simple example, assume the following scenario: a two lane road changes to a one lane road at location $x_0$. The used fundamental diagram is given in Figure 1. The capacity for one lane is $q_1$, which means that the capacity of the two lane segment is $q_2 = 2 * q_1$. The demand intensity is $q_a$,

with $q_1 < q_a < q_2$ for the first hour and $q_b$, $q_b < q_1 < q_2$ for the second hour. The scenario starts from an empty road. The XT-plot of this scenario is given in Figure 2.

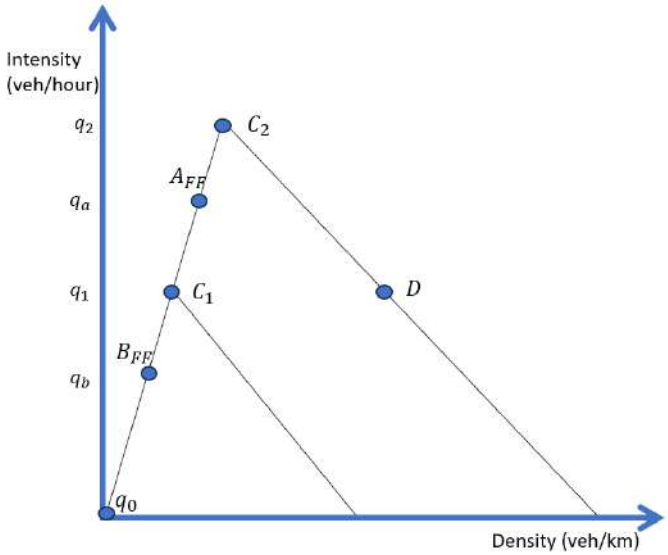



*Figure 1 Fundamental Diagram for one and two lane road, with all relevant traffic states indicated*

Now an extra vehicle, illustrated as the dotted green line in the XT-plot, is added compared to the normal case. This means that for a very short time, the demand is $q_a + 1$. This results in a short shockwave that is slightly sharper. The new boundary between the congestion and free flow state is illustrated in red. We denote $T$ as the remaining time of the bottleneck once the extra vehicle has passed through the bottleneck.

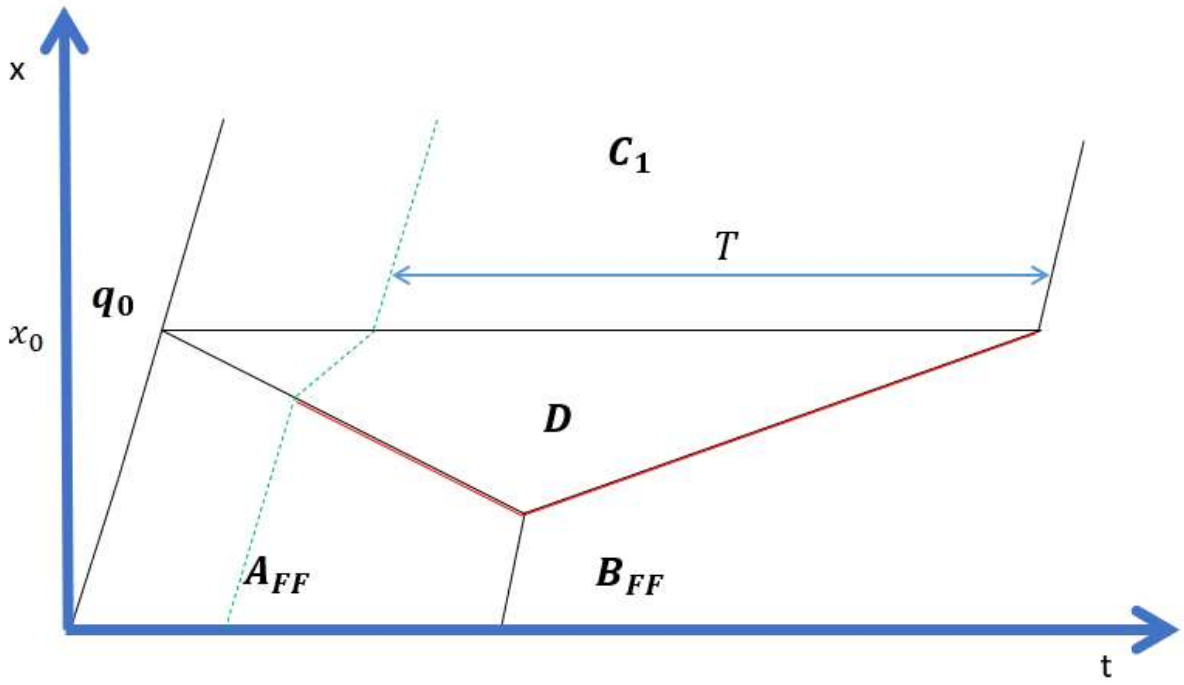



*Figure 2 XT-plot of the single bottleneck case, traffic states are indicated in bold, while shockwaves are shown in black*

Every person during this period $T$ is delayed with $\frac{1}{q_1}$. The total amount of people still going through the bottleneck after the extra traveler is $T * q_1$, leading to a total increased delay of $\frac{1}{q_1} * (T * q_1) = T$. An analogue derivation can be made if the extra traveler would depart in the second hour.

**Dissolving Queue**

Imagine now a bottleneck as a result from an incident. For some time, one lane is blocked and thus the capacity is limited to $q_1$. The road can use both lanes again at time $t_0$. The demand now stays constant at $q_a$. The resulting XT-plot of this scenario can be seen in Figure 3. Again, an extra traveler is added, and the new boundaries between the different states are illustrated in red. We denote $T_1$ and $T_2$ as the time after the extra traveler passed the bottleneck that there is a queue with one lane or two lanes, respectively. This means that $T = T_1 + T_2$. During a period of $T_1$, $T_1 * q_1$ cars are being delayed with $\frac{1}{q_1}$. During the period $T_2$, $T_2 * q_2$ cars are being delayed with $\frac{1}{q_2}$. The total extra delay equals thus: $\frac{1}{q_1} * T_1 q_1 + \frac{1}{q_2} * T_2 q_2 = T_1 + T_2 = T$. If the extra traveler would pass the bottleneck at the time the two lanes are opened again, the calculations become trivial and the marginal cost equals $T_2 = T$.

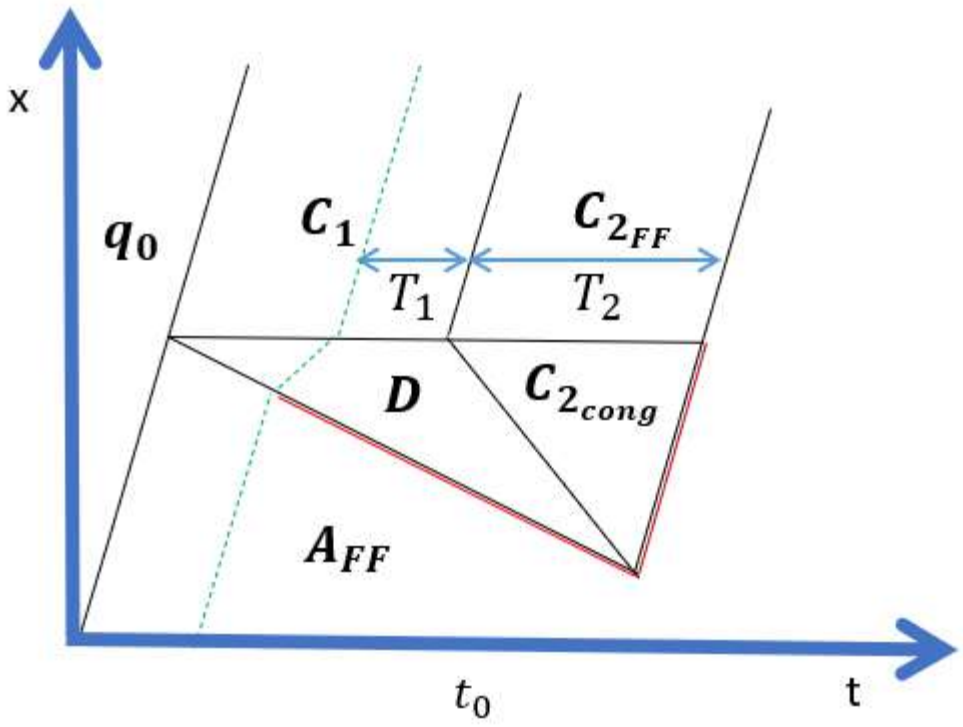



*Figure 3 XT-plot of the dissolving queue case, traffic states are indicated in bold, while shockwaves are shown in black*

## 3. Increasing Complexity

For simplicity, the extra complexity is only discussed on the single bottleneck case, caused by fluctuating demand. The same reasonings can be made on the dissolving queue case.

**An off-ramp before the bottleneck**

Let's start by adding an off-ramp before the bottleneck. Now assume the extra traveler takes this exit, located at $x_1$. The responding XT-plot is visualized in Figure 4. As can be seen by the figure, no extra delay is caused to other travelers. Intuitively, the following process is happening: first an extra traveler takes some place in the queue, pushing others backwards. Second, the extra traveler takes the exit, making room again. All cars after the extra traveler will reach the bottleneck at the same time compared to when there was no extra traveler. Therefore, there is no extra delay and no marginal congestion cost.

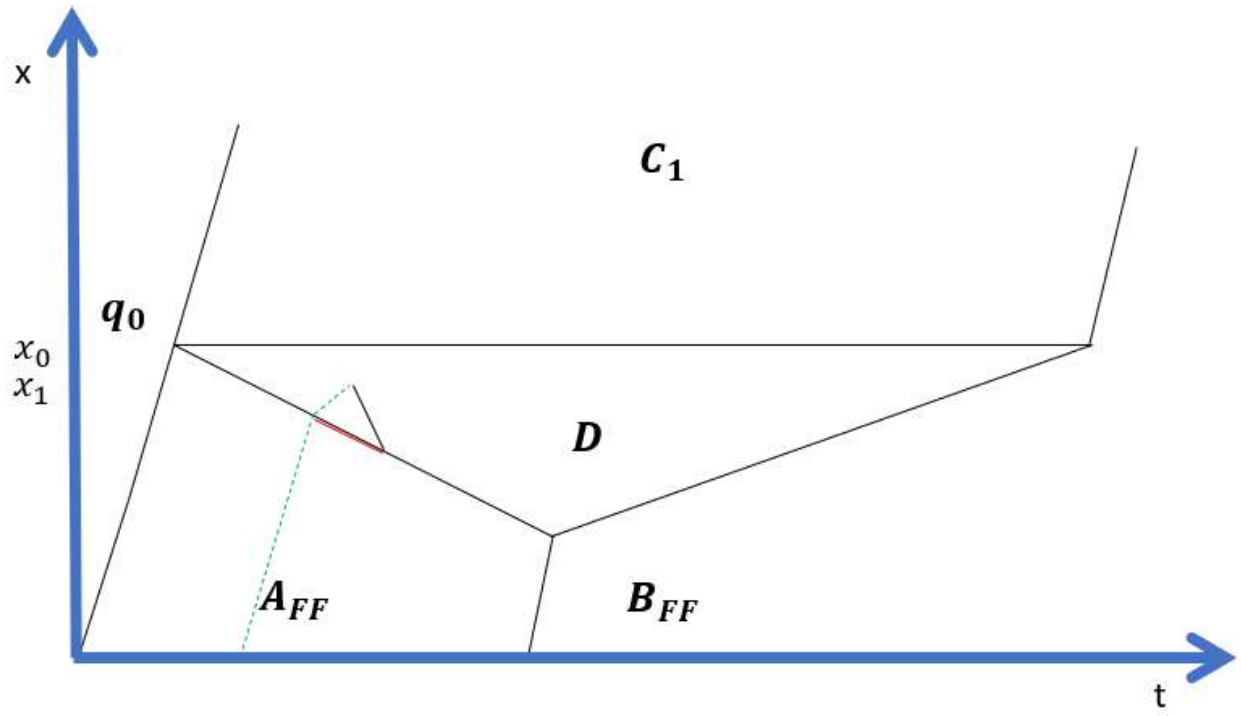



*Figure 4 XT-plot with an extra off-ramp before the bottleneck*

**Multiple off-ramps before the bottleneck**

Now assume that there would have been multiple off-ramps before the bottleneck. For travelers going through the bottleneck, no extra delay is caused. However, some travelers that either also take that exit or an exit before can experience delay. To be precise, all cars in the little triangle, defined by the trajectory of the extra car (green dotted line), the new queue border (red line) and the shockwave when the car disappears (black line), are $\frac{1}{q_1}$ later on the same x-position. If their exit would lie within this triangle, they experience extra delay. Once they passed the new shockwave, the delay is gone and they are at exactly the same time on each x-position. This triangle can be bigger or smaller, depending on the time the extra vehicle enters the system and also on the location of the off-ramp. However, the total extra delay is much lower than the extra delay caused by an extra traveler going through the bottleneck.

**Merging point in queue**

Assume now that the queue of the single bottleneck spills back to a merging point of two main roads. At the merging point, each road takes some priority. An extra traveler originating from the left road does not cause a delay on the travelers origination from the right road. The travelers from the right road will still take the same position in the queue once merged. However, the extra traveler originating from the left road causes a proportional larger delay to all other travelers originating from the left. This extra delay can be quantified as $\frac{1}{q_c * P_{left}}$, with $q_c$ the bottleneck capacity and $P_{left}$ the proportion of vehicles going through the bottleneck origination from the left road. If $T$ again represents the time until the end of the bottleneck since the extra traveler went through the bottleneck, then the total amount of travelers going through the bottleneck originating

from the left road after the extra traveler is $T * q_c * P_{left}$. Resulting in a total extra delay of $\frac{1}{q_c * P_{left}} * T * q_c * P_{left} = T$. Again, the total extra delay or the marginal social congestion cost is $T$. In this calculation, it is assumed that $P_{left}$ is constant. If this would not be the case, more complex calculations can be made, resulting in the same total extra delay of $T$.

## 4. Theoretical Conclusion & Implications in Practice

If the extra traveler goes through the bottleneck, the marginal social congestion cost is $T$, the remaining time the bottleneck is active once the extra traveler has gone through it. If the extra traveler goes not through the bottleneck, there can be marginal social congestion costs, but generally of a much lower magnitude. This implies that if marginal costs need to be allocated to links, the marginal social congestion cost is only non-zero at bottleneck locations.

The magnitude of the marginal social congestion cost on these few links can be large. Many bottlenecks on highways remain active for multiple hours during the rush periods, with a value of time of around 20 euros (Small, 2012), the total marginal cost can be 60-100euro. This is much larger then any other marginal effect (noise, safety, …). When optimizing routes in a social perspective, this would give a free pass to avoid the bottleneck and e.g. take the off-ramp and route through the underlying network.

The marginal social congestion costs on the underlying network is usually much lower, as the bottlenecks are typically not constant active. E.g. at signalized intersections it happens that the queue remains during multiple cycles, but if at one point the queue disappears completely, the marginal social cost is reset.

In the full paper, more theoretical cases will be discussed, including the possibility of having negative marginal social costs by having an extra traveler. These insights will be compared to how currently marginal congestion costs are calculated, with static and dynamic models. Furthermore, the discussion with other marginal social costs and the implications of the difference in magnitude values will be further elaborated.

## 6. Bibliography


Gavanas, N., Tsakalidis, A., Pitsiava-Latinopoulou, M., 2017. Assessment of the marginal social cost due to congestion using the speed flow function. Transp. Res. Procedia, 3rd Conference on Sustainable Urban Mobility, 3rd CSUM 2016, 26 – 27 May 2016, Volos, Greece 24, 250–258. https://doi.org/10.1016/j.trpro.2017.05.115

Krichene, W., Suarez Castillo, M., Bayen, A., 2018. On Social Optimal Routing Under Selfish Learning. IEEE Trans. Control Netw. Syst. 5, 479–488. https://doi.org/10.1109/TCNS.2016.2619910

Lighthill, M.J., Whitham, G.B., 1955. On kinematic waves II. A theory of traffic flow on long crowded roads. Proc. R. Soc. Lond. Ser. Math. Phys. Sci. 229, 317–345. https://doi.org/10.1098/rspa.1955.0089

Mayeres, I., Proost, S., Dender, K.V., 2005. The Impacts of Marginal Social Cost Pricing. Res. Transp. Econ., Measuring the Marginal Social Cost of Transport 14, 211–243. https://doi.org/10.1016/S0739-8859(05)14008-6
Richards, P.I., 1956. Shock Waves on the Highway. Oper. Res. 4, 42–51.
Small, K.A., 2012. Valuation of travel time. Econ. Transp. 1, 2–14. https://doi.org/10.1016/j.ecotra.2012.09.002

# Transit route design based on trip-based agglomeration of travel demand

April 15, 2025

Keywords: Transit routes, Trip-based agglomeration, Road networks, Mixed integer linear program

## 1 Introduction

The design of routes in transit networks is an important problem as it affects other aspects of transit system design such as the setting of timetables, planning of crews, etc. (Vermeir *et al.*, 2021). Traditionally, this problem has been tackled considering road network data and transit demand matrix as inputs (Chakroborty & Wivedi, 2002, Vermeir *et al.*, 2021). Note that these approaches require an origin-destination (OD) matrix representation of transit demand and thus require the trip ends to be agglomerated towards the a priori decided zone centroids (and then to the potential bus stops). However, a novel trip-based agglomeration (TBA) method proposed by Kumar *et al.* (2024) represents each trip as an ordered pair of origin and destination in the form of a line segment and these line segments are agglomerated to form corridors (note that each corridor is a line segment whose length is dependent upon the endpoints of the agglomerated trips). Kumar *et al.* (2024) showed that the proposed TBA method can identify major flow-carrying corridors in a single step as opposed to multi-step post-processing that may arise in the OD-based agglomeration (ODA) methods.

This paper proposes a novel method to design transit routes given the corridors obtained from the TBA method. Since corridors are line segments that may not be aligned with the underlying road network, it is required to determine a route on the underlying road network that closely represents the given corridor while satisfying the following desirable conditions. One of the conditions is that the determined route should pass through some of the important points on the corridor such as the endpoints, points with high-demand activity, etc. Another condition that the developed route should satisfy is that its length should be as small as possible. While ensuring the above two conditions, it may happen that there may be many turns along the determined route, and therefore the last condition is to keep the number of turns in the route as low as possible. In order to ensure that all these conditions are met as much as possible, an optimization problem is proposed to design the transit routes that maximizes the weighted sum of the number of important points on corridor that are visited in the determined route minus the length of the determined route minus the number of turns in the determined route.

A major difference between the proposed optimization problem and the studies that require OD matrix representation of transit demand is that the latter studies benchmark their results on small-scale networks like Mandl's network with 15 nodes (Vermeir *et al.*, 2021), Mumford network with 127 nodes (Ahmed *et al.*, 2019), etc., whereas in this work transit routes are designed for a large-scale network with approximately $20,000$ nodes.

Note that the proposed optimization formulation has high-level analogies to the orienteering problem (OP) (Vansteenwegen *et al.*, 2011) but there are important differences also as explained next. One of the versions of OP involves finding a route between two given points that maximizes the number of points that are visited from a pre-determined set of points under the constraint that the length of the determined route does not exceed a predetermined threshold. However, the

proposed optimization problem considers the number of turns in the objective function because the determined route is developed on a given road network where the angle of turns that are made is important unlike that in OP. Furthermore, the length of the determined route is part of the objective function in the proposed optimization formulation whereas there is a constraint on the length of the determined route in OP. In the next section, the developed optimization formulation is presented.

## 2 Problem Formulation

This section describes the proposed optimization formulation for developing transit routes. Let $c$ be a directed transit demand corridor as obtained from the TBA method of Kumar *et al.* (2024) (note that a directed corridor is a directed line segment). The source and the target endpoints of the given corridor $c$ are denoted as $o$ and $d$, respectively. Let $G = (V, E)$ be a directed graph representing the underlying road network. For all $(i, j) \in E$, let $l_{ij}$ be the distance from node $i$ to node $j$. Suppose $s \in V$ and $t \in V$ represent the nodes closest to the source endpoint and the target endpoint, respectively, of the given corridor $c$. The output of the proposed optimization problem is a route that starts from node $s$ and ends in node $t$ while satisfying the conditions described in the subsequent discussion. Let $P$ be a given set of important points on corridor $c$ such as points with a high-demand activity (note that set $P$ does not include the endpoints $o$ and $d$). Let $Q \subset V$ be the set of nodes that are closest to the points in set $P$. Let $H$ be the set of all the pair of edges $((i, j), (j, k))$ (such that $(i, j) \in E$, $(j, k) \in E$ and $i \neq k$) where a turn is said to be present. Let $w_p$ and $w_t$ be the scaling factors for different terms in the objective function. The target of the proposed optimization formulation is to determine a route connecting $s$ to $t$ so that the sum of the number of nodes from set $Q$ that lie in the determined route minus $w_p$ times the length of the determined route minus $w_t$ times the number of turns in the determined route is maximized. The proposed optimization formulation is given as follows.

$$\text{Maximize} \quad \left[\sum_{\forall k \in Q} \rho_k\right] - \left[w_p \sum_{\forall (i,j) \in E} x_{ij} l_{ij}\right] - \left[w_t \sum_{\forall ((i,j),(j,k)) \in H} f_{ijk}\right] \tag{1}$$

Subject to:

$$\sum_{\forall j \in V \setminus \{s\}} x_{sj} = 1 \tag{2}$$

$$\sum_{\forall i \in V \setminus \{s\}} x_{is} = 0 \tag{3}$$

$$\sum_{\forall i \in V \setminus \{t\}} x_{it} = 1 \tag{4}$$

$$\sum_{\forall j \in V \setminus \{t\}} x_{tj} = 0 \tag{5}$$

$$\sum_{\forall i \in V \setminus \{t\}} x_{ik} = \sum_{\forall j \in V \setminus \{s\}} x_{kj} \qquad \forall k \in V \setminus \{s, t\} \tag{6}$$

$$\rho_k \leq \sum_{\forall i \in V \setminus \{t\}} x_{ik} \qquad \forall k \in Q \tag{7}$$

$$f_{ijk} \geq x_{ij} + x_{jk} - 1 \qquad \forall ((i, j), (j, k)) \in H \tag{8}$$

$$u_j \geq u_i + x_{ij}n - n + 1 \qquad \forall i, j \in V \text{ and } i \neq j \tag{9}$$
$$\rho_k \in \{0, 1\} \qquad \forall k \in Q \tag{10}$$
$$x_{ij} \in \{0, 1\} \qquad \forall (i, j) \in E \tag{11}$$
$$f_{ijk} \in \{0, 1\} \qquad \forall ((i, j), (j, k)) \in H \tag{12}$$

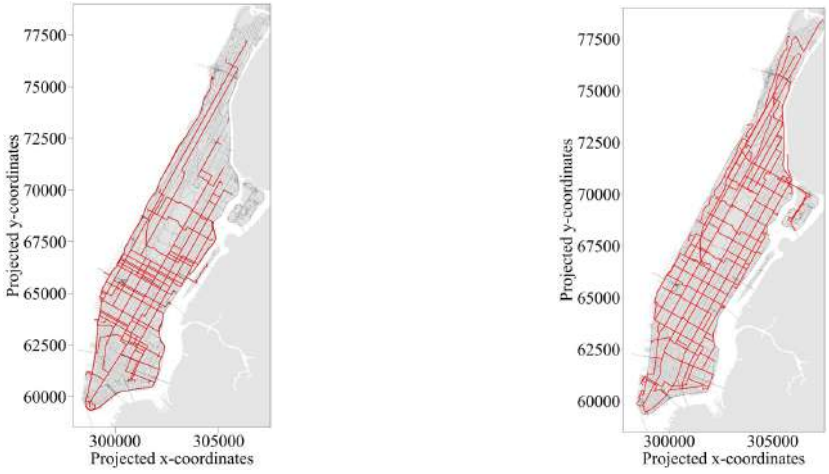
The expression in Equation 1 presents the objective function as described before. Here, for all $k \in Q$, $\rho_k$ is equal to 1 if node $k$ lies in the determined route, otherwise it is equal to 0. Also, for all $(i, j) \in E$, $x_{ij}$ is equal to 1 if the edge $(i, j)$ lies in the determined route, otherwise it is equal to 0. For all $((i, j), (j, k)) \in H$, $f_{ijk}$ is equal to 1 if the edges $(i, j)$ and $(j, k)$ lie in the determined route, otherwise it is equal to 0. Thus, the first term from the left is the number of nodes from set $Q$ that lie in the determined route, the middle term is $w_p$ times the length of the determined route and the last term is $w_t$ times the number of turns in the determined route. Equation 2 ensures that only one edge that lies in the determined route leaves the source node $s$. Equation 3 ensures that none of the edges that lie in the determined route enter the source node $s$. Equation 4 ensures that only one edge that lies in the determined route enters the source node $t$. Equation 5 ensures that none of the edges that lie in the determined route leave the source node $t$. Equation 6 ensures that for each node $k \in V \backslash \{s, t\}$, the number of incoming edges to node $k$ that lie in the determined route is equal to the number of outgoing edges from node $k$ in the determined route. The constraint in Equation 7 ensures that for all $k \in Q$, $\rho_k$ is equal to 0 if there is no incoming edge to node $k$ in the determined route. The constraint in Equation 8 ensures that for all $((i, j), (j, k)) \in H$, $f_{ijk}$ is equal to 1 if both the edges $(i, j)$ and $(j, k)$ lies in the determined route. Equation 9 is a sub-tour elimination constraint that ensures no cycles are formed in the determined route. Here, $u_i, \forall i \in V$ are unrestricted decision variables used for eliminating subtours in the determined route. Equations 10 to 12 are the definitional constraints for the decision variables. Since some of the decision variables in the above formulation are integers (except the variables $u_i, \forall i \in V$), and the objective function and the constraints are linear, the proposed optimization formulation is a Mixed Integer Linear Program (MILP).

## 3 Results and discussions

In this section, the proposed optimization formulation is applied on a given set of corridors to determine the transit routes for a large-scale road network. A set of 40 corridors obtained by the TBA method of Kumar *et al.* (2024) for Manhattan, New York City, and the underlying road network for Manhattan (consisting of approximately 20,000 nodes and 30,000 directed edges (Dataset 1, 2014)) are used as inputs for the proposed optimization formulation. Note that 40 corridors are taken as input as there are about 40 transit routes currently in Manhattan. The optimization problem is solved[1] for each corridor and this process is repeated 40 times to obtain a set of 40 transit routes. Figure 1a shows the computed transit routes in red color. Note that the average computation time required to determine each transit route is approximately 75 seconds. Figure 1b shows the existing transit routes in Manhattan (Dataset 2, 2020). It can be seen that at a high-level, the obtained routes look similar to the current routes but there are noticeable differences. That is because the determined transit routes are computed solely based on the information on transit demand and the underlying road network whereas the existing transit routes are designed based on other information as well such as site-specific constraints that may not allow transit routes to pass through some specific locations, welfare policies that may require providing routes at low demand areas, etc. Therefore, the proposed optimization framework only provides an initial estimate of the transit routes that should be suitably modified by the transit route planners to consider various practical constraints and government policies. Note that the obtained transit routes are statistically compared with the existing transit routes

[1]The optimization problem is solved using a competitive solver for MILPs (Manual, 1987) and codes are run on a 64-bit Windows operating system with 32GB RAM and an Intel i7-3770 processor.

in Manhattan using the Kolmogorov–Smirnov test to check similarity in terms of aspects such as length, orientation, etc. It is found that the obtained routes are not significantly different from the current routes statistically. These results are not presented here for brevity but interested readers can refer to Supplementary material 1 (2024). In summary, this work provides a novel way to efficiently generate transit routes based on the demand corridors obtained from the TBA method of Kumar *et al.* (2024). An important contribution of this work is that it considers the information regarding large-scale road networks while designing transit routes in contrast to existing studies that test their results on smaller networks.




(a) (b)

Figure 1 – *The left figure shows the transit routes as obtained from the proposed formulation and the right figure shows the existing transit routes in Manhattan, New York City.*

The present work can be extended as follows. In this paper, only the development of transit routes was focused on, but developing schedules for running transit on such routes would be an important step. The corridors that were taken as input for developing transit routes were assumed to be fixed but different set of corridors can be generated from the TBA method of Kumar *et al.* (2024) as the proposed algorithms in Kumar *et al.* (2024) are stochastic and thus considering stochasticity in the distribution of corridors while developing transit routes would also be very interesting. Finally, the information of other criteria such as area served by the transit routes, length of the transit routes, etc., can be incorporated in the proposed optimization formulation while developing the transit routes.

## References

Ahmed, Leena, Mumford, Christine, & Kheiri, Ahmed. 2019. Solving urban transit route design problem using selection hyper-heuristics. *European Journal of Operational Research*, **274**(2), 545–559.

Chakroborty, Partha, & Wivedi, Tathagat. 2002. Optimal route network design for transit systems using genetic algorithms. *Engineering optimization*, **34**(1), 83–100.

Dataset 1. 2014. *NYC open data*. https://data.cityofnewyork.us/City-Government/road/svwp-sbcd. Accessed on 2024-09-19.

Dataset 2. 2020. *MTA bus time*. https://bt.mta.info/. Accessed on 2024-09-19.

Kumar, Praveen, Chakroborty, Partha, & Gehlot, Hemant. 2024. Novel Trip Agglomeration Methods for Efficient Extraction of Urban Mobility Patterns. *Networks and Spatial Economics*, 1–30.

Manual, CPLEX User's. 1987. Ibm ilog cplex optimization studio. *Version*, **12**(1987-2018), 1.

Supplementary material 1. 2024. *Distribution of parameters for transit routes*. https://sites.google.com/view/praveenkumar-iitk/supplementary-material?authuser=0. Modified on 2025-04-01.

Vansteenwegen, Pieter, Souffriau, Wouter, & Van Oudheusden, Dirk. 2011. The orienteering problem: A survey. *European Journal of Operational Research*, **209**(1), 1–10.

Vermeir, Evert, Engelen, Wouter, Philips, Johan, & Vansteenwegen, Pieter. 2021. An exact solution

approach for the bus line planning problem with integrated passenger routing. *Journal of Advanced Transportation*, **2021**(1), 6684795.

# Probabilistic Inference of Dynamic Traffic Assignment Using a Generalized Bathtub Model

**Zihan Wan,**[1,*] **Wei Ma**[1]
[1] Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University
[*] Correspondence: zi-han.wan@connect.polyu.hk

## ABSTRACT

**The increasing complexity of urban traffic networks makes monitoring individual links challenging, limiting the effectiveness of traditional dynamic traffic assignment (DTA) methods. The macroscopic fundamental diagram (MFD) provides a powerful tool for managing urban traffic at a network level. However, existing MFD-based DTA frameworks do not fully explore the inseparable coupling among uncertainties in trip lengths, speed-density MFDs, and route choices, all of which jointly influence macroscopic traffic dynamics. To address this gap, we propose a new formulation of probabilistic dynamic traffic assignment using the generalized bathtub model, which explicitly accounts for stochasticity in trip length distributions, speed-density MFDs, and route choice behaviors. In the model formulation, instead of aggregate path flow data, we use a detailed hierarchy of flow data, including path-specific inflow and outflow data, and other probe vehicle-related data. We define a random utility function to assess how these uncertainties impact route choices, and develop a data-driven inference algorithm to solve stochastic dynamic user equilibrium. Under the equilibrium condition, it is able to simultaneously derive the approximate distributions of trip lengths, MFD parameters, and multinomial logit route choice coefficients. Numerical experiments are conducted on synthetic data on a simulation network and realistic taxi trajectory data in Futian district of Shenzhen. The experiment results validate the effectiveness and accuracy of our proposed inference algorithm.**

## INTRODUCTION

The aggregate capacity of a region in a network, represented by the macroscopic fundamental diagram (MFD), is initially proved by Daganzo (2007) and Geroliminis and Daganzo (2008), and later confirmed in other cities (Ambühl and Menendez, 2016; Geroliminis and Sun, 2011; Loder et al., 2017). An MFD describes the aggregate relationship between average flow $q$ (vehicles/second) and the accumulation $n$ (vehicles) inside a region, and is instrumental for downstream tasks including perimeter control (Geroliminis et al., 2012; Guo and Ban, 2020; Zhong et al., 2018) and road pricing (Simoni et al., 2015; Zheng et al., 2012). Many studies have employed loop detector data to estimate an MFD either in real or simulated settings (Geroliminis and Daganzo, 2008; Keyvan-Ekbatani et al., 2013; Leclercq et al., 2014); however, as loop detectors are fixed-point measurement devices, the estimated MFDs may be biased and can not reflect the aggregate traffic dynamics. Taking advantage of communication technologies, recent efforts use probe vehicles to overcome the limitations of loop detectors (Ambühl and Menendez, 2016; Saffari et al., 2022). Traffic flow in reality is far more complicated than deterministic and homogeneous fluid flow governed by physical laws. Therefore, it is necessary to estimate a stochastic MFD based on probe vehicle trajectories, taking into account the heterogeneity in vehicles and road geometry inside each region.

In the current literature, the studies of dynamic traffic assignment (DTA) combined with regional networks and MFD-based models are not extensive. Initial efforts focus on simple networks with few origin-destination pairs or non-overlapping paths (Laval et al., 2017; Leclercq and Geroliminis, 2013). Yildirimoglu and Geroliminis (2014) proposed the first DTA framework for multi-regional MFD-based models but did not discuss the distributions of trip lengths; Batista and Leclercq (2019) examined perceived errors in both speed and trip lengths, and solved network equilibrium by Monte Carlo simulations. Batista et al. (2019, 2018); Paipuri et al. (2020) investigated the estimation of regional trip lengths with a hierarchy of network information, including previous region, next region, path. Overall, they only provide deterministic point-valued estimates of regional trip lengths; in the case of macroscopic traffic modeling, trip length in each region is stochastic and might take a wider range of values, reflecting the truth that a region is composed of numerous links.

In dynamic traffic assignment (DTA), the path length distributions and temporal evolution of speed inside different regions influence path travel times and further influence how trips are distributed in a regional network. The generalized bathtub model (GBM)

(Jin, 2020; Vickrey, 2019, 2020), coupled with the MFD, is a suitable macroscopic framework to model the traffic dynamics resulting from travelersŕoute choice behaviors. The GBM draws an analogy between water flow in a bathtub and vehicular flow in a homogeneous region, and assumes that speed is only dependent on the vehicular density inside a region. In this paper, based on prove vehicle trajectories, we propose an MFD-based stochastic dynamic traffic assignment model that is able to:

- Estimate the posterior distributions of coefficients in a random utility model to quantify the randomness in travelers' route choices
- Derive the approximate trip length distribution in each region
- Infer the posterior distributions of parameters in the functional form of the MFD to quantify the uncertainty in speed-density relationship in each region

## METHODOLOGY

In this section, we start by presenting the hierarchy of variables whose interactions lead to path flow and speed observations based on macroscopic fundamental diagrams (MFDs) in Table 1.

| | **Description** |
|---|---|
| **Symbols** | |
| $\Omega$ | set of origin-destination (OD) pairs |
| $R_\omega$ | set of regional paths connecting OD pair $\omega$ |
| $U^t_{r_\omega}$ | utility of route $r_\omega$ at time $t$ |
| $\epsilon_{r_\omega}$ | error term in the utility of $r_\omega$ |
| $B$ | set of bathtubs (regions) |
| $\Delta_\omega$ | path-region incidence matrix of size $\lvert R_\omega\rvert \times \lvert B\rvert$ for OD pair $\omega$ |
| $G_{r_\omega}$ | $\lvert B\rvert \times \lvert B\rvert$ subgraph matrix of path $r_\omega$ indicating the order of traffic flow over the bathtubs, i.e., $G_{r_\omega}(i,j) = 1$ if bathtub $i$ immediately follows bathtub $j$ on path $r_\omega$ and 0 otherwise |
| $\mathbf{I}$ | the identity matrix |
| **Latent variables and parameters** | |
| $d_{r_\omega,j}$ | travel distance in bathtub $j$ on path $r_\omega$ |
| $\lambda_{r_\omega,j}$ | parameter for the exponentially distributed $d_{r_\omega,j}$ |
| $\boldsymbol{\lambda}_{r_\omega}$ | a length-$\lvert B\rvert$ vector of $\lambda_{r_\omega,j}$ |
| $v_0^j$ | free flow speed in bathtub $j$ |
| $d_{jam}^j$ | jam density in bathtub $j$ |
| $\boldsymbol{\theta}_\omega$ | vector of logit coefficients for OD pair $\omega$ |
| **Observations** | |
| $Q^t_\omega$ | travel demand for OD pair $\omega$ at time $t$, unit: veh(s) |
| $F^t_{r_\omega,\text{enter}}$ | number of vehicles selecting path $r_\omega$ at time $t$ |
| $F^t_{r_\omega,j,\text{enter}}$ | number of vehicles selecting path $r_\omega$ and entering bathtub $j$ at time $t$, i.e., $F^t_{r_\omega,j,\text{enter}} = F^t_{r_\omega,\text{enter}}$ if bathtub $j$ is the origin of OD pair $\omega$ |
| $\mathbf{F}^t_{r_\omega,\text{enter}}$ | a length-$\lvert B\rvert$ vector of $F^t_{r_\omega,j,\text{enter}}$ |
| $F^t_{r_\omega,j,\text{exit}}$ | number of exit vehicles from bathtub $j$ on path $r_\omega$ at time $t$ |
| $\mathbf{F}^t_{r_\omega,\text{exit}}$ | a length-$\lvert B\rvert$ vector of $F^t_{r_\omega,j,\text{exit}}$ |
| $N^t_j$ | number of total vehicles inside bathtub $j$ at time $t$ |
| $\mathbf{N}^t$ | a length-$\lvert B\rvert$ vector of $N^t_j$ |
| $N^t_{r_\omega,j}$ | number of total vehicles on path $r_\omega$ inside bathtub $j$ at time $t$ |
| $\mathbf{N}^t_{r_\omega}$ | a length-$\lvert B\rvert$ vector of $N^t_{r_\omega,j}$ |
| $\bar{v}^t_j$ | average speed over $N^t_j$ probe vehicles inside bathtub $j$ at time $t$ |

**Table 1:** Notation summary.

### Traffic dynamics in Regional DTA

A core question in dynamic traffic assignment (DTA) is how to replicate the travelers' route choice behaviors. The first principle of Wardrop (Wardrop, 1952) gives rise to the theoretical framework which describes the selfish behaviors of drivers to minimize their own travel times. The deterministic user equilibrium assumes that all drivers are rational and well-informed about possible routes and their travel times. A more realistic assumption is that drivers evaluate route choices, and trips are distributed, according to the perceived utility $U^t_{r_\omega}$:

$$U_{r_\omega}^t = \theta_{\omega,1}\text{TTD}_{r_\omega} + \theta_{\omega,2}\text{TTT}_{r_\omega}^t + \epsilon_{r_\omega}, \forall r_\omega \in R_\omega, \forall \omega \in \Omega, \tag{1}$$

where TTD denotes total travel distance, TTT denotes total travel time, and $\epsilon$ accounts for the uncertainty specifically related to route/path $r_\omega$. $\theta_{\omega,2}$ is the value of time as in Lu et al. (2008); Zhang et al. (2013). Equation 1 represents the stochastic extension of user equilibrium (SUE), where no single traveler has incentives to improve perceived utility by switching to another route (Daganzo, 1982; Daganzo and Sheffi, 1977).

The number $F_{r_\omega,\text{enter}}^t$ of observed probe vehicles selecting path $r_\omega$ at time $t$ is the result of route decision making, in which route choices follow a multinomial logit (MNL) model (Dial, 1971), where $\epsilon_{r_\omega}$ is independently and identically Gumbel distributed. Some might criticize that MNL fails to capture the correlations between overlapping routes; since the number of regions is significantly smaller than that of links, we assume that the severity of this fault is reduced in macroscopic traffic dynamics. In this study, we explore how expected path total travel distance (TTD) and total travel time (TTT) would influence the likelihood of instantaneously choosing a route $r_\omega$ at time $t$:

$$\begin{aligned}
\text{TTD}_{r_\omega} &= \Delta_{r_\omega,:}\frac{1}{\boldsymbol{\lambda_{r_\omega}}} \\
\text{TTT}_{r_\omega}^t &= \Delta_{r_\omega,:}(\frac{1}{\mathbf{v}^t} \circ \frac{1}{\boldsymbol{\lambda_{r_\omega}}}) \\
\boldsymbol{\theta}_\omega &\sim \text{MVN}(\boldsymbol{\mu}_{\theta_\omega}, \boldsymbol{\Sigma}_{\theta_\omega}) \\
p_{r_\omega}^t &= \frac{\exp(\boldsymbol{\theta}_\omega^T[\text{TTD}_{r_\omega}, \text{TTT}_{r_\omega}^t])}{\sum_{k\in R_\omega}\exp(\boldsymbol{\theta}_\omega^T[\text{TTD}_k, \text{TTT}_k^t])} \\
F_{1,\text{enter}}^t, ..., F_{|R_\omega|,\text{enter}}^t &\sim \text{Multinomial}(Q_\omega^t; p_1^t, ..., p_{|R_\omega|}^t),
\end{aligned} \tag{2}$$

where $\Delta_{r_\omega,:}$ is the row of path-region incidence matrix $\Delta_\omega$ corresponding to path $r_\omega$; $\frac{1}{\boldsymbol{\lambda}_{r_\omega}}$ is the vector of inverses of $\lambda_{r_\omega,j}$'s, which is the vector of expected travel distances in bathtubs on path $r_\omega$; $\mathbf{v}^t$ is the speed vector obtained from bathtub MFDs; $p_{r_\omega}^t$ is the probability of choosing path $r_\omega$ at time $t$.

The observed speed $\bar{v}_j^t$ of probe vehicles in bathtub $j$ at time $t$ follows

$$\begin{aligned}
v_0^j, d_{jam}^j &\sim \text{MVN}(\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2) \\
v_j^t &= \text{MFD}_j(N_j^t; v_0^j, d_{jam}^j) \\
\bar{v}_j^t &\sim \text{Normal}(v_j^t, \sigma_{\text{MFD}_j}^2),
\end{aligned} \tag{3}$$

where $N_j^t$ denotes the number of total vehicles in bathtub $j$ at time $t$ and $\sigma_{\text{MFD}_j}^2$ is the measurement error for individual vehicle speed in bathtub $j$.

The probability $p_{r_\omega,j}^t$ of a probe vehicle exiting bathtub $j$ on path $r_\omega$ at time $t$ is dependent on the average speed in bathtub $j$ and exponentially distributed path-specific regional travel distance $d_{r_\omega,j}$. For an exponentially distributed random variable $d$ with parameter $\lambda$, the probability of $d$ being smaller than a threshold $t$ is given by $\text{p}(d \leq t) = 1 - e^{-\lambda t}$. We further assume that the vehicles taking path $r_\omega$ exit bathtub $j$ are independent and follow identical Bernoulli distribution. Given an exponentially distributed travel distance $d_{r_\omega,j}$ with parameter $\lambda_{r_\omega,j}$, we have:

$$
\begin{aligned}
p^t_{r_\omega,j} &= 1 - e^{-\lambda_{r_\omega,j}\Delta_t v^t_j} \\
F^t_{r_\omega,j,\text{exit}} &\sim \text{Binomial}(N^t_{r_\omega,j}, p^t_{r_\omega,j}).
\end{aligned} \tag{4}
$$

We assume that traffic dynamics abide the law of flow conservation among bathtubs (regions), represented as follows:

$$
\begin{aligned}
\mathbf{N}^t &= \sum_{\omega\in\Omega}\sum_{r_\omega\in R_\omega} \mathbf{N}^t_{r_\omega} \\
\mathbf{N}^t_{r_\omega} &= \mathbf{N}^{t-1}_{r_\omega} + \mathbf{F}^t_{r_\omega,\text{enter}} + (G_{r_\omega} - \mathbf{I})\mathbf{F}^{t-1}_{r_\omega,\text{exit}}.
\end{aligned} \tag{5}
$$

## Solution algorithm of regional DTA

Our objective is to solve the stochastic dynamic traffic equilibrium in a regional network, in terms of the parameters of trip length distributions and posterior distributions of MFD and MNL parameters. In other words, we want to find $d_{r_\omega,j} \sim \text{exponential}(\lambda_{r_\omega,j})$, $\theta_\omega \sim \text{MVN}(\boldsymbol{\mu}_{\theta_\omega}, \Sigma_{\theta_\omega})$, $v^j_0, d^j_{jam} \sim \text{MVN}(\boldsymbol{\mu}_j, \boldsymbol{\sigma}^2_j)$ so that

$$
\begin{aligned}
F^t_{1,\text{enter}}, ..., F^t_{|R_\omega|,\text{enter}} &\sim \text{Multinomial}(Q^t_\omega; p^t_1, ..., p^t_{|R_\omega|}) \\
p^t_{r_\omega} &= P(U^t_{r_\omega} \le U^t_k) \\
\forall k \ne r_\omega \in R_\omega, &\forall \omega \in \Omega.
\end{aligned} \tag{6}
$$

It is difficult to directly derive posterior distributions of latent variables in complex systems involving both spatial and temporal interactions; hence, we resort to stochastic variational inference (SVI )for scalable approximate inference (Hoffman et al., 2013; Kingma, 2013; Ranganath et al., 2014; Wingate and Weber, 2013). In SVI, the objective function to minimize the following function:

$$
\pi(\Psi) = -\frac{1}{N}\sum_{i=1}[\log(p_\Phi(\boldsymbol{X}|\boldsymbol{Z}_i)p_\Phi(\boldsymbol{Z}_i)) - \log q_\Psi(\boldsymbol{Z}_i)], \tag{7}
$$

where $\boldsymbol{X}$ denotes observed variables, $\boldsymbol{Z}$ denote latent variables, $\Psi$ represents the parameters of approximate posterior distributions. Minimizing $\pi(\Psi)$ implicitly minimizes the KL-divergence $\text{KL}(q_\Psi(\boldsymbol{Z})\|p_\Phi(\boldsymbol{Z}|\boldsymbol{X}))$. In each iteration $k$, the relative gap is defined as $\frac{\pi(\Psi^k)-\pi(\Psi^{k-1})}{\pi(\Psi^k)}$, and we consider that the network is in equilibrium if gap $\le$ tol, where tol is predefined tolerance. Algorithm 1 summarizes the SVI solution of MFD-based stochastic DTA.

# RESULTS

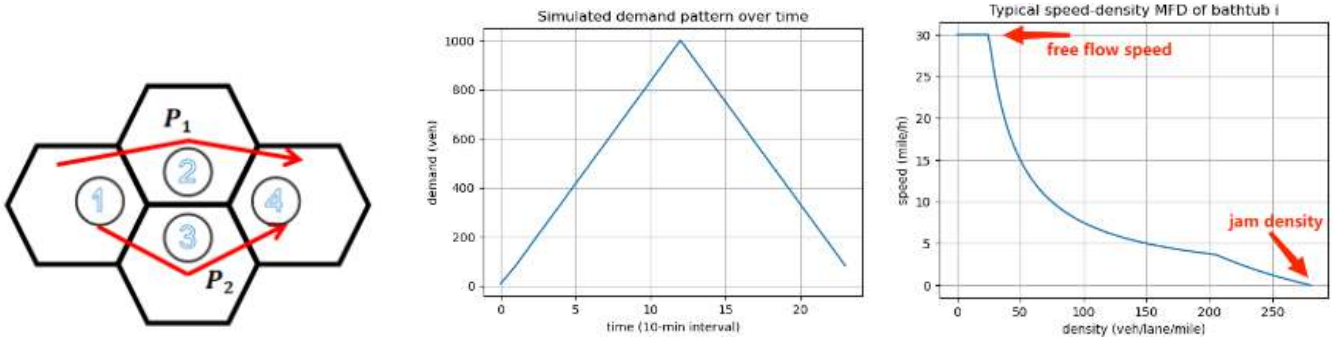
## A toy network with known MFD parameters

In this subsection, we examine the inference performance of our proposed modeling framework in a simulated bathtub network with a single OD pair and two connecting paths. The model is tested on a simplified setting where a subset of variables are fixed and known, instead of being latent, since speed information is not observable in the simulation.

Basic information of the components in the simulation is displayed in the following Figure 1. The prescribed paths connecting the designated OD pair are $\mathcal{P}_1 = (1, 2, 4)$ and $\mathcal{P}_2 = (1, 3, 4)$ in Figure 1A, respectively. The deterministic demand pattern $Q_t$ in Figure 1B is exogenously determined to mimic the traffic dynamics in morning peak hours, and the vehicles are only allowed to enter the network via bathtub 1. In the simulation setting, the ground truth average speed-density bathtub-wide MFD (Figure 1C) is assumed to be publicly accessible to help drivers' route decision making. In this hypothetical environment, we are able to record the trajectory

**Algorithm 1** Stochastic Variational Inference for MFD-Based Stochastic Dynamic Traffic Assignment

Input the OD pair set $\Omega$, route choice set $R_\omega$, path-region incidence matrix $\Delta_\omega$, directed flow matrix $G_{r_\omega}, \forall r_\omega \in R_\omega, \forall \omega \in \Omega$
Input the travel demand $Q_\omega^t$ , $\forall \omega \in \Omega$, and the duration $T$
initialize $k = 1$, tol, maximum number of iterations $N_{\max}$, learning rate $\alpha$
**while** gap $\geq$ tol **and** $k \leq N_{\max}$ **do**
  Sample MNL coefficients $\theta_\omega \sim \text{MVN}(\boldsymbol{\mu}_{\theta_\omega}, \Sigma_{\theta_\omega})$ and MFD parameters $v_0^j, d_{jam}^j \sim \text{MVN}(\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2)$ from approximate posteriors
  **for** $i = 1, 2, \ldots, T$ **do**
    **if** $i = 1$ **then**
      Set perceived speed inside bathtub $j$ equal to free flow speed $v_0^j$
    **else**
      Calculate the perceived speed inside bathtub $j$ based on $\text{MFD}_j$ and $N_j^{t-1}$
    **end if**
    Calculate MNL-based path inflow $F_{r_\omega,\text{enter}}^t$ by Eq 2 and calculate the likelihood $\text{Multinomial}(F_{1,\text{enter}}^t, ..., F_{|R_\omega|,\text{enter}}^t | Q_\omega^t; p_1^t, ..., p_{|R_\omega|}^t)$
    Calculate $\mathbf{N}^t$ and $\mathbf{N}_{r_\omega}^t$ by Eq 5 based on observations $\mathbf{N}_{r_\omega}^t$, $\mathbf{F}_{r_\omega,\text{enter}}^t$, $\mathbf{F}_{r_\omega,\text{exit}}^{t-1}$
    Calculate the likelihood of MFD speed observation $\text{Normal}(\bar{v}_j^t | \text{MFD}_j(N_j^t; v_0^j, d_{jam}^j), \sigma_{\text{MFD}_j}^2)$ by Eq 3
    Given the speed, calculate the likelihood of path region outflow $\text{Binomial}(F_{r_\omega,j,\text{exit}}^t | N_{r_\omega,j}^t, p_{r_\omega,j}^t)$ by Eq 4
    $\theta_{old} \leftarrow \theta$
  **end for**
  Collect all the likelihoods and compute $\pi^k$ and gap
  $\Psi \leftarrow \Psi - \alpha\eta$, where $\eta$ is the gradient of $\pi^k$ with respect to $\Psi$
  $k \leftarrow k + 1$
**end while**




A Toy network with a single OD pair and two paths

B Simulation demand pattern

C Bathtub-level speed-density MFD

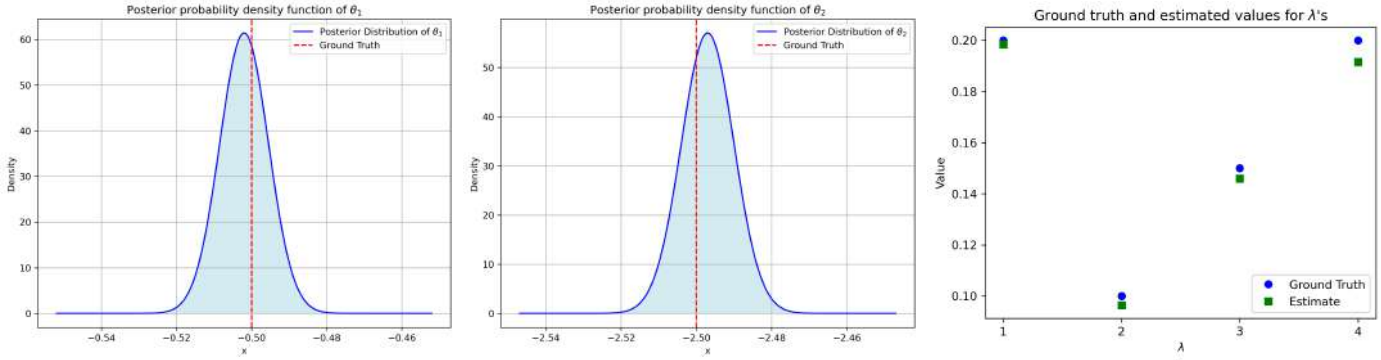
**Figure 1:** Essential components in the simulation of a single-OD toy network.

of each trip; therefore, our observed data consist of time-dependent path inflows $F_{i,\text{enter}}^t$ and path-specific time-dependent bathtub outflows $F_{i,exit}^t$, where time index $t \in \{1, ..., T\}$ and path index $i \in \{1, 2\}$.

In the simplified setting, route choice behaviors are replicated by presuming unknown fixed values $(-0.5, -2.5)$ for $\theta_1$ and $\theta_2$. Therefore, the goal is to infer the approximate posterior distributions of multinomial logit coefficients $\boldsymbol{\theta}$ and unknown exponential parameters $\lambda$'s. In Figure 2A and Figure 2B, the resulting posterior probability density functions for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are plotted. While the ground truth values for $\boldsymbol{\theta}$ are marked by red vertical dashed lines, it is easy to observe that the modes of normal posteriors match the ground truth values closely. The measurements of variance of MNL coefficients are also negligible by the spike-shaped PDFs. Figure 2C also provides the comparison between accurately estimated and ground truth values of $\lambda$'s for the exponentially distributed distances.

## A real urban network with stochastic MFDs

To evaluate the proposed stochatsic DTA inference framework in a real urban network, we divide the Futian District, Shenzhen into hexagonal regions by h3-py (Uber, 2018) in Figure 3A. To aggregate OD demand data and observed route choices, we use taxi GPS data, which contain time, longitude, latitude, and occupancy status so that we could track the movement of each OD trip as a sequence of hexagonal regions (Zhang et al., 2015). As preliminary results, the estimated regional MFDs are presented in Figure 3B, each indexed by the Uber h3 hexagon id. Each solid line represents the average speed-density relationship in the region; by repeatedly sampling from approximate posteriors of MFD parameters, we are able to quantity the uncertainty in MFD relationship. It is observed that the corresponding proportions of speed observations fall between the confidence intervals whose endpoints are empirical percentiles of the MFD samples.

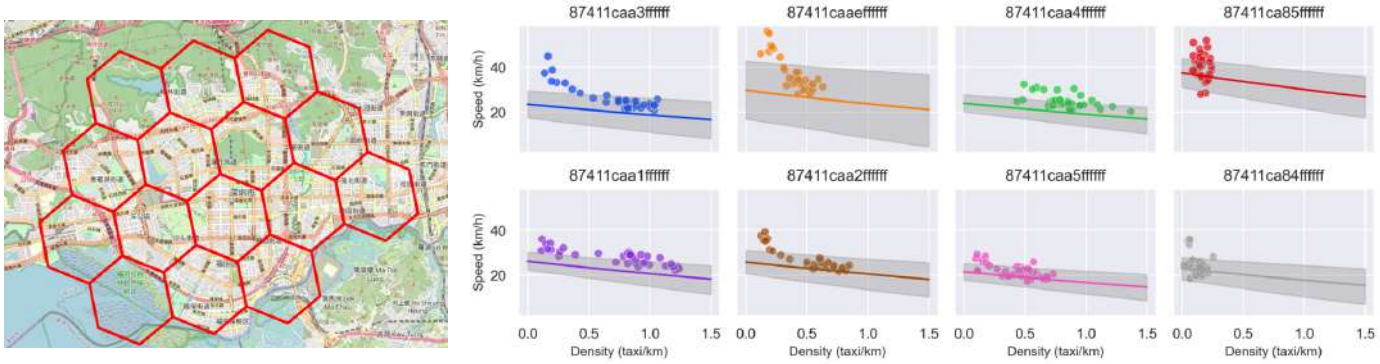



A Posterior PDF of $\theta_1$

B Posterior PDF of $\theta_2$

C Ground truth and estimated values of $\lambda$'s

**Figure 2:** Inference results for MNL coefficients and exponentially distributed distance parameters for the simulation environment.




A Decomposition of Futian District, Shenzhen into hexagonal regions

B Estimated MFDs and corresponding confidence intervals for a subset of regions indexed by Uber h3 hex id

**Figure 3:** Inferred stochastic speed-density MFDs based on taxi trajectories in Futian District, Shenzhen.

## REFERENCES


Ambühl, L. and Menendez, M. (2016). Data fusion algorithm for macroscopic fundamental diagram estimation. *Transportation Research Part C: Emerging Technologies*, 71:184–197.

Batista, S., Leclercq, L., and Geroliminis, N. (2019). Estimation of regional trip length distributions for the calibration of the aggregated network traffic models. *Transportation Research Part B: Methodological*, 122:192–217.

Batista, S., Leclercq, L., Krug, J., and Geroliminis, N. (2018). Trip length estimation for the macroscopic traffic simulation: scaling microscopic into macroscopic networks. In *TRB 2018, Transportation research board annual meeting*, page 22p.

Batista, S. F. and Leclercq, L. (2019). Regional dynamic traffic assignment framework for macroscopic fundamental diagram multi-regions models. *Transportation Science*, 53(6):1563–1590.

Daganzo, C. F. (1982). Unconstrained extremal formulation of some transportation equilibrium problems. *Transportation Science*, 16(3):332–360.

Daganzo, C. F. (2007). Urban gridlock: Macroscopic modeling and mitigation approaches. *Transportation Research Part B: Methodological*, 41(1):49–62.

Daganzo, C. F. and Sheffi, Y. (1977). On stochastic models of traffic assignment. *Transportation science*, 11(3):253–274.

Dial, R. B. (1971). A probabilistic multipath traffic assignment model which obviates path enumeration. *Transportation research*, 5(2):83–111.

Geroliminis, N. and Daganzo, C. F. (2008). Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transportation Research Part B: Methodological*, 42(9):759–770.

Geroliminis, N., Haddad, J., and Ramezani, M. (2012). Optimal perimeter control for two urban regions with macroscopic fundamental diagrams: A model predictive approach. *IEEE Transactions on Intelligent Transportation Systems*, 14(1):348–359.

Geroliminis, N. and Sun, J. (2011). Hysteresis phenomena of a macroscopic fundamental diagram in freeway networks. *Procedia-Social and Behavioral Sciences*, 17:213–228.

Guo, Q. and Ban, X. J. (2020). Macroscopic fundamental diagram based perimeter control considering dynamic user equilibrium. *Transportation Research Part B: Methodological*, 136:87–109.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*.

Jin, W.-L. (2020). Generalized bathtub model of network trip flows. *Transportation Research Part B: Methodological*, 136:138–157.

Keyvan-Ekbatani, M., Papageorgiou, M., and Papamichail, I. (2013). Urban congestion gating control based on reduced operational network fundamental diagrams. *Transportation Research Part C: Emerging Technologies*, 33:74–87.

Kingma, D. P. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Laval, J. A., Leclercq, L., and Chiabaut, N. (2017). Minimal parameter formulations of the dynamic user equilibrium using macroscopic urban models: Freeway vs city streets revisited. *Transportation research procedia*, 23:517–530.

Leclercq, L., Chiabaut, N., and Trinquier, B. (2014). Macroscopic fundamental diagrams: A cross-comparison of estimation methods. *Transportation Research Part B: Methodological*, 62:1–12.

Leclercq, L. and Geroliminis, N. (2013). Estimating mfds in simple networks with route choice. *Procedia-Social and Behavioral Sciences*, 80:99–118.

Loder, A., Ambühl, L., Menendez, M., and Axhausen, K. W. (2017). Empirics of multi-modal traffic networks–using the 3d macroscopic fundamental diagram. *Transportation Research Part C: Emerging Technologies*, 82:88–101.

Lu, C.-C., Mahmassani, H. S., and Zhou, X. (2008). A bi-criterion dynamic user equilibrium traffic assignment model and solution algorithm for evaluating dynamic road pricing strategies. *Transportation Research Part C: Emerging Technologies*, 16(4):371–389.

Paipuri, M., Xu, Y., González, M. C., and Leclercq, L. (2020). Estimating mfds, trip lengths and path flow distributions in a multi-region setting using mobile phone data. *Transportation Research Part C: Emerging Technologies*, 118:102709.

Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR.

Saffari, E., Yildirimoglu, M., and Hickman, M. (2022). Data fusion for estimating macroscopic fundamental diagram in large-scale urban networks. *Transportation Research Part C: Emerging Technologies*, 137:103555.

Simoni, M. D., Pel, A. J., Waraich, R. A., and Hoogendoorn, S. P. (2015). Marginal cost congestion pricing based on the network fundamental diagram. *Transportation Research Part C: Emerging Technologies*, 56:221–238.

Uber (2018). H3: Ubers hexagonal hierarchical spatial index. https://eng.uber.com/h3/.

Vickrey, W. (2019). Types of congestion pricing models. *Economics of Transportation*, 20:100140.

Vickrey, W. (2020). Congestion in midtown manhattan in relation to marginal cost pricing. *Economics of Transportation*, 21:100152.

Wardrop, J. G. (1952). Road paper. some theoretical aspects of road traffic research. *Proceedings of the institution of civil engineers*, 1(3):325–362.

Wingate, D. and Weber, T. (2013). Automated variational inference in probabilistic programming. *arXiv preprint arXiv:1301.1299*.
Yildirimoglu, M. and Geroliminis, N. (2014). Approximating dynamic equilibrium conditions with macroscopic fundamental diagrams. *Transportation Research Part B: Methodological*, 70:186–200.
Zhang, D., Zhao, J., Zhang, F., and He, T. (2015). Urbancps: A cyber-physical system based on multi-source big infrastructure data for heterogeneous model integration. In *Proceedings of the ACM/IEEE Sixth International Conference on Cyber-Physical Systems*, pages 238–247.
Zhang, K., Mahmassani, H. S., and Lu, C.-C. (2013). Dynamic pricing, heterogeneous users and perception error: Probit-based bi-criterion dynamic stochastic user equilibrium assignment. *Transportation Research Part C: Emerging Technologies*, 27:189–204.
Zheng, N., Waraich, R. A., Axhausen, K. W., and Geroliminis, N. (2012). A dynamic cordon pricing scheme combining the macroscopic fundamental diagram and an agent-based traffic model. *Transportation Research Part A: Policy and Practice*, 46(8):1291–1303.
Zhong, R., Chen, C., Huang, Y., Sumalee, A., Lam, W., and Xu, D. (2018). Robust perimeter control for two urban regions with macroscopic fundamental diagrams: A control-lyapunov function approach. *Transportation Research Part B: Methodological*, 117:687–707.

# Designing the Publicly-Owned Centralized Platform for Ride-Hailing Services with Shared Automated Vehicles: A System Optimal Dynamic Traffic Assignment Approach

Jinxiao Du[a], Wei Ma[a,b,*]

[a]*Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China*
[b]*The Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, Guangdong, China*

**Abstract**

As ride-hailing services continue to grow in popularity, they contribute to escalating congestion in traffic networks. This paper proposes a publicly-owned centralized platform (POCP) for shared automated vehicles (SAVs) that connects travelers, traffic network companies (TNCs), and parking spaces. It provides a novel insight for the regulation of government to save social costs by adjusting travelers' behaviors with spatial-temporal trip subsidies. In the proposed POCP, (i) government operates this centralized platform with the objective of minimizing social cost, collecting and distributing necessary information (*i.e.*, trip orders, trip fees, discounts, parking vacancies) to corresponding users; (ii) travelers request and pay the trip order in the centralized platform and their trip order choices may be associated with the trip fee and discount; (iii) TNCs receive trip orders from the platform and provide mobility services by routing SAVs to maximize the profit. We formulate the operation of POCP as a bi-level programming, where the government aims at minimizing the social cost by providing travelers with spatial-temporal discounts in the upper-level problem, and TNC aims at maximizing the profit given travel demand from the upper level as a dynamic traffic assignment (DTA) in lower level problem. We prove that the proposed bi-level programming could be equivalently decomposed into two sub-problems with linear programming to solve for the minimum social cost and minimum subsidy amount, respectively. To validate our proposed model and solution algorithms, we conducted numerical experiments in both small and real-world large networks. The results reveal substantial savings in social costs, equivalent to 110.38% of the subsidies spent in the small network and 121.96% in the large network. These findings underscore the efficacy of the proposed POCP model, emphasizing the potential of spatial-temporal subsidies in reducing traffic congestion and improving social welfare.

*Keywords:* Ride-Hailing Platform, System Optimal Dynamic Traffic Assignment (SO-DTA), Shared Automated Vehicles (SAVs), Bi-level Optimization

## 1. Introduction

This paper proposes a novel insight to improve the current ride-hailing platform and save social welfare in dynamic traffic networks from the government's perspective and designs a publicly-owned centralized platform (POCP). We formulate the operation of the POCP as a nonlinear bi-level programming where the government aims at minimizing the social cost by adjusting travelers' behaviors with the spatial-temporal discount in the upper level and TNC aims at maximizing its profit by providing mobility service of SAVs in the lower level. The bi-level problem can be decomposed by two sub-problems, where the first sub-problem could be solved by a series of LPs and the second sub-problem is a LP. Therefore, the solution process becomes tractable for large-scale networks and numerical experiments demonstrate the great potential of the design of POCP.

The contributions of this study are summarized as follows:

- This research proposes a design for a publicly-owned centralized platform (POCP) for shared automated vehicles (SAVs), which serves as an intermediary connecting travelers, and TNCs. The platform, operated by

*Corresponding author
*Email addresses:* jinxiao.du@connect.polyu.hk (Jinxiao Du), wei.w.ma@polyu.edu.hk (Wei Ma)

the government, aims to reduce social costs by subsidizing trips via spatial-temporal discounts. Conversely, TNCs focus on providing mobility services to travelers by optimally routing SAVs to maximize their profit.

- The study formulates a SO-DTA problem for the POCP as non-linear bi-level programming. This model incorporates both traveler and vehicle flows to capture network-wide traffic dynamics.
- We demonstrate that the proposed non-linear bi-level programming of POCP could be equivalently converted into two sub-problems with linear programming (LPs) to solve for the minimum social cost and minimum subsidy amount, respectively. The first LP could be obtained by the branch and bound algorithm. It provides us with a direct and efficient way to obtain the spatial-temporal subsidy to save the social cost using the minimum subsidy amount.

## 2. Problem Settings

The POCP acts as an intermediary to connect travelers and TNCs. We would illustrate the roles of each stakeholder for POCP as follows.

The government operates the platform to minimize the social cost and offers spatial-temporal subsidies to travelers by providing discounts on trip fees in the POCP. These discounts are dependent on the destination, time of orders, and the type of SAVs.

Travelers, who are informed about the trip fees set by TNC and the discount set by the government in POCP, place trip orders on the platform. They pay the actual trip fare to the platform, which equals the trip fee multiplied by the discount. Additionally, travelers exhibit the following characteristics:

- Sensitivity to subsidies varies among travelers. Travel demand can be divided into regular demand and flexible demand. Regular demand pertains to travelers who would request an order at a specific time (denoted as $r$) regardless of the discounts. On the other hand, flexible demand is sensitive to discounts, and their requesting time may change due to varying travel costs at different time
- Travelers have the option to utilize ride-sharing services and are divided into two groups: solo travelers and ride-sharing travelers. Solo travelers, who prefer not to use ride-sharing services, are the only occupants of a Shared Automated Vehicle (SAV). Ride-sharing travelers, on the other hand, opt for ride-sharing services. Accordingly, the type of SAV is defined as follows:
  - $x = 0$ represents an idle SAV.
  - $x = -1$ represents an SAV with a solo traveler (*e.g.*, UberX).
  - $x = 1, \cdots, X$ represents an SAV with $x$ ride-sharing travelers (*e.g.*, UberPool), where $X$ is the maximum number of travelers in a SAV.

TNCs are alerted by the platform when travelers place orders, and they then provide mobility services by routing SAVs. The platform pays TNCs the total trip fees, which comprise the fares travelers pay and the subsidy offered by the government. Furthermore, the platform shares parking vacancy information with the TNCs for SAV parking selection, and TNCs remit parking fees to the platform.

## 3. Model

We use the Double Queue (DQ) model to present the dynamic network loading process. The dynamics of the DQ model are captured by the two dependent queues: the downstream queue and the upstream queue. The downstream and upstream queue of SAVs is presented in Equation 1 and 2 for $(i, j) \in E$, $x = -1, 0, 1, \cdots, X$ and $k = 1, \cdots, T$:

$$q_{i,j}^{\mathcal{D},x}(k) = \sum_{a=0}^{k-\tau_{i,j}^{f}} \Delta_t \, u_{i,j}^{x}(a) - \sum_{a=0}^{k} \Delta_t \, v_{i,j}^{x}(a), \tag{1}$$

$$q_{i,j}^{\mathcal{U},x}(k) = \sum_{a=0}^{k} \Delta_t \, u_{i,j}^{x}(a) - \sum_{a=0}^{k-\tau_{i,j}^{w}} \Delta_t \, v_{i,j}^{x}(a), \tag{2}$$

The POCP design problem could be formulated as a nonlinear bi-level programming as follows:

- Upper level. This level involves the government's operation of the POCP with the goal of minimizing social costs, which equals maximizing social welfare. This is achieved by influencing travelers' behaviors via spatial-temporal discounts that subsidize their trip fees. The variables in this level, represented by $\mathbf{d}$, pertain to demand, while the spatial-temporal subsidy variables are represented by $\boldsymbol{\alpha}$.
- Lower Level. In this level, the TNCs handle the routing choices of SAVs upon receiving travel demand orders via a centralized platform. The aim is to maximize profit through the provision of mobility services. The variables in this level, denoted by $\mathbf{y}$, encapsulate the flow dynamics variables.

The formulation of the POCP design is presented as follows:

$$
\begin{aligned}
&\min_{\mathbf{y},\mathbf{d},\alpha} \quad Social\ Cost \\
&s.t. \quad \underset{\mathbf{d},\alpha}{Demand\ Constraints} \\
&\qquad \mathbf{y} \in \arg\max_{\mathbf{y}} \quad TNC\ Profit \\
&\qquad\quad s.t.\ Double\ Queue\ Constraints \\
&\qquad\qquad Double\ Flow\ Matching\ Constraints \\
&\qquad\qquad Node\ Conservation\ Constraints \\
&\qquad\qquad Parking\ Constraints \\
&\qquad\qquad Waiting\ Constraints \\
&\qquad\qquad Other\ Constraints
\end{aligned}
\tag{3}
$$

## 4. Numerical Experiments

We evaluate the effectiveness of the POCP framework and provide a comparison of two scenarios in real Hong Kong network: the privately-owned platform with no subsidy, and POCP applied with the subsidy obtained by proposed Algorithm. Table 1 presents the effect of applying optimal spatial-temporal discount by the proposed Algorithm in the Hong Kong network. Compared with original traffic conditions that no subsidy would be applied, the proposed POCP with spatial-temporal subsidy saves the $1.108 \times 10^6$ social cost by $\mathbf{9.085} \times 10^5$ subsidy. Moreover, Equations indicate that the decrease of travelers cost is the sum of the decrease of TNC profit, the increase of government cost and the decrease of social cost. Therefore, the reduction of travelers cost in POCP compared with the privately-owned platform is approximately $2.80 \times 10^6$. This can be interpreted as the transfer of TNC profits amounting to $9.06 \times 10^5$, government expenditures of $7.82 \times 10^5$, and the decrease of social cost as $1.11 \times 10^6$, which represents the reduction of congestion due to subsidy in POCP.

Then we calculate the average value of link flow under SO-DTA for all links and time intervals to explore the network-wide effects of applying subsidy in POCP in large-scale networks, as referred to in Figure 1.

| Stakeholder | Cost | Private-owned platform | POCP |
|---|---|---|---|
| Traveler | Traveler Travel Time Cost ($10^6$) | 2.141 | 1.468 |
| | Traveler Trip Fee Cost ($10^6$) | 3.166 | 1.022 |
| | Traveler Waiting Cost ($10^5$) | 4.137 | 2.268 |
| | Traveler Late Cost ($10^5$) | 1.686 | 0.349 |
| | Traveler Late Penalty ($10^5$) | 4.309 | 0.891 |
| | Traveler Cost ($10^6$) | 5.459 | 2.663 |
| TNC | SAV Travel Cost ($10^5$) | 5.226 | 4.088 |
| | SAV Travel Revenue ($10^6$) | 3.166 | 1.931 |
| | SAV Parking Cost ($10^5$) | 0.444 | 1.708 |
| | TNC Profit ($10^6$) | 2.168 | 1.262 |
| Government | Subsidy Amount ($10^5$) | 0.000 | **9.085** |
| | Social Cost ($10^6$) | **3.246** | **2.138** |

Table 1: Results of POCP on Hong Kong network.

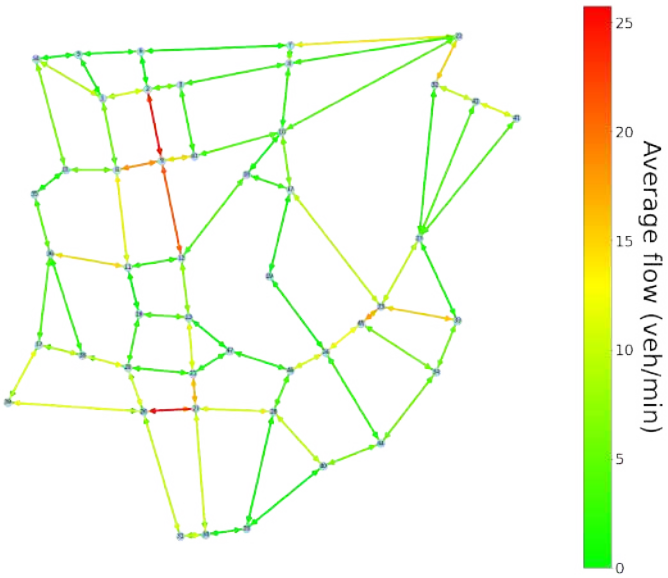



Figure 1: Overview of the link flow of POCP in Hong Kong network.

# Agent-Based Simulation Considering Parking Search and Choice Behavior Using Bayesian Inference

Prakash Ranjan[a], Fatemeh Fakhrmoosavi, Ph.D.[a*], Krishna M. Gurumurthy, Ph.D.[b], Abdelrahman Ismael, Ph.D.[b], Nazmul A. Khan, Ph.D.[b]

[a]*School of Civil and Environmental Engineering, University of Connecticut, *Corresponding Author, moosavi@uconn.edu*
[b]*Transportation and Power Systems Division, Argonne National Laboratory*

**MOTIVATION:** Finding a parking spot can be a challenging and time-consuming task, especially in dense urban areas where parking spaces are limited (*1*). Consequently, modeling travelers' parking search behavior is a critical element of urban transportation planning (*2*). In addition, cruising for parking can form a mobile queue of vehicles mixed with regular traffic, contributing to congestion and lowering the efficiency of the transportation system. Understanding how travelers search for parking at their destinations helps planners and policymakers identify factors influencing parking demand and informs the design of effective parking policies, such as dynamic pricing and demand-responsive parking management (*3*).

Several studies have explored parking search behavior using both mathematical and data-driven techniques (*4, 5*). These efforts often employ revealed or stated preference surveys to model parking type choice (e.g., on-street versus garage) and parking location choice. While surveys yield valuable information, integrating survey data with trip trajectories can offer additional insights into parking decisions and search patterns. Multiple studies have also simulated parking choices using traffic simulation tools, making various assumptions about parking search behavior for different vehicle types, including private human-driven vehicles and shared autonomous vehicles (SAVs). For example, Fakhrmoosavi et al. (*6*) examined two parking search scenarios for SAVs within a traffic simulation tool, assuming that private vehicles park at the nearest available location if they lack access to private parking. However, driver behavior often varies in terms of parking type and specific location choice, especially in highly populated urban areas. There remains a need to develop a model for parking type and specific parking location choice that can be implemented in agent-based simulators accounting for estimation uncertainties and driver heterogeneity.

To address these needs, we developed a parking choice model that estimates travelers' parking search behavior by integrating trip trajectories from GPS data with traveler surveys. A Bayesian nested logit model was employed to represent parking search behavior, including the decision between on-street and garage parking as well as the specific choice of a garage. The model accounts for factors such as parking fees, the distance between destinations and parking locations, driver and trip characteristics, and land use. The Bayesian framework provides a probabilistic basis that incorporates empirical information on real-world parking behavior and facilitates the quantification of uncertainties in model estimation. We then implemented the estimated model in a multiagent activity-based travel demand model with a dynamic traffic simulation tool, POLARIS, allowing for a comprehensive full-day simulation of vehicle cruising and parking activities for the entire region's population (*7*). Additionally, we developed an optimization framework for dynamic parking pricing and incorporated it into the same simulation tool; however, due to page limitations, details about the optimization framework and pricing results are not included here.

For modeling and simulation, this study utilizes several data sources, including GPS data collected by the Chicago Metropolitan Agency for Planning (CMAP). Land use information was drawn from the Environmental Protection Agency's Smart Location Database (SLD). Garage parking locations, along with associated fees and capacities, were matched to corresponding points derived from the GPS trajectory data to identify parking trips and measure cruising activity. This study advances previous research through three main contributions: 1) Combining trip trajectories from GPS data with a revealed preferences survey provides a more holistic view of drivers' decision-making processes and improves accuracy in predicting both parking type and specific parking choices. 2) Employing a data-driven Bayesian framework effectively captures the nested choice structure, accommodates uncertainties, and overcomes computational challenges, leading to more robust and reliable results. 3) By modeling parking choice and vehicle cruising behavior within an agent-based simulation tool for the entire Chicago region, the approach enables the examination of various parking policies and their impacts on regional transportation outcomes.

**DATA DESCRIPTION:** This study primarily utilized data collected by CMAP (*8*) as part of a comprehensive household travel survey. The survey aimed to analyze socio-demographic characteristics and travel choice behavior among individuals in the Chicago metropolitan area. A total of 30,683 individuals from 12,391 households participated, and expansion factors were applied to scale these responses to represent the region's 8.5 million residents in 2019. Data collection spanned from spring 2018 to spring 2019, capturing potential seasonal variations in travel and parking behavior. To obtain detailed and accurate travel records, the study employed smartphone-based GPS tracking that recorded trip trajectories and arrival/departure points for five weekdays. By combining survey responses with GPS tracking, the dataset provided information on parking decisions and behaviors, including parking type (on-site, off-site, on-street, garage) and instances where no parking was utilized. The survey also gathered information on parking fees, duration, and any subsidies associated with parking, importance of arriving on-time, and other factors.

**BAYESIAN STATISTICS NESTED LOGIT MODEL:** This study employs a nested multinomial logit model within a Bayesian framework to estimate two levels of travelers' parking choices: (1) the decision between on-street and garage parking, and (2) the selection of a specific garage. To construct the choice sets, all on-street parking locations are consolidated into a single alternative, with its fee and distance drawn randomly from appropriate distributions. Meanwhile, any garage within a five-kilometer radius of a traveler's destination is included in the garage choice set. The Bayesian estimation approach is well-suited to these hierarchically structured choices; it provides a posterior distribution for each coefficient, thereby explicitly accounting for estimation uncertainty. Additionally, when prior information is incorporated, Bayesian methods can converge more quickly than traditional maximum-likelihood estimation (*9*). Figure 1 illustrates the main steps in developing the parking choice model.

Convergence of the model parameters is assessed using the Metropolis–Hastings (MH) algorithm. For each proposed parameter vector ($\beta$, $\lambda$), the algorithm computes an acceptance probability, $\alpha$. If $\alpha$ exceeds a random number drawn from a uniform distribution (0–1), the proposed parameters replace the current ones, and the model outputs are updated. Otherwise, the proposal is rejected, and the algorithm retains the current parameter values for the next iteration. This iterative process continues until a large, predefined maximum number of iterations is reached. The MH algorithm is part of the broader class of Markov chain Monte Carlo (MCMC) methods. It is particularly well-suited for Bayesian model estimation due to its robustness in high-dimensional parameter spaces, relatively rapid convergence, and ease of implementation (*9*). In general, MCMC algorithms construct a Markov chain, $X_0$, $X_1$, ..., $X_n$, designed so that the chain's stationary distribution matches the target density $\pi$. In that case, given $X_n$ , a "proposed value", $Y_{n+1}$, is generated from a pre-specified density $q(X_n, y)$ and then approved with a probability of $\alpha(X_n, Y_{n+1})$ given by

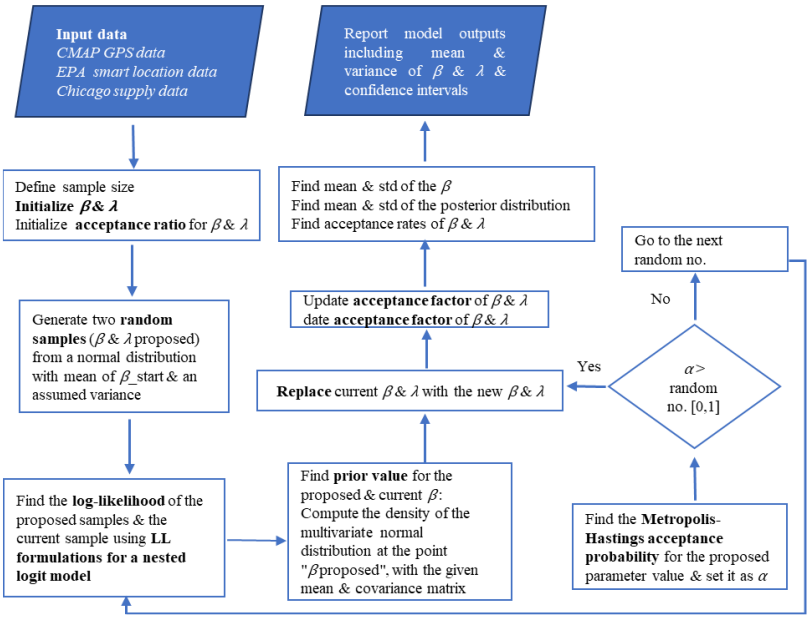



**FIGURE 1 Bayesian Statistic Nested Logit Model Algorithm**

$$\alpha(x,y) = \begin{cases} min\left\{\frac{\pi(y)}{\pi(x)}\frac{q(y,x)}{q(x,y)}, 1\right\}, & \pi(x)q(x,y) > 0 \\ 1, & \pi(x)q(x,y) = 0 \end{cases} \quad (1)$$

After proposing a value, $Y_{n+1}$, if it is accepted, the algorithm sets $X_{n+1} = Y_{n+1}$, and if it is rejected, it sets $X_{n+1} = X_n$ . The purpose of the function $\alpha$(x, y) is to ensure that the Markov chain is reversible concerning the target density $\pi$(y), so that the target density is stationary for the chain. In this study, a Gaussian distribution is used for hyperparameters, and Sims' priors are adopted for the inclusive value parameters ($\lambda$). The general form of Sims' priors is:

$$f(\lambda) = \begin{cases} 0, & if \ \lambda \le 0, \\ \alpha(s)\lambda^{s-a}\exp(-\lambda^s), & if \ \lambda > 0, \end{cases} \quad (2)$$

where $s$ and $a$ ($s \ge a$) are the hyperparameters, $\alpha(s)$ is the normalization constant. The prior mode of $\lambda$ can be expressed as:

$$\lambda = \left(\frac{s-a}{s}\right)^{\frac{1}{s}} \quad (3)$$

**SIMULATION MODELING FRAMEWORK:** The estimated parking choice model was integrated into POLARIS to simulate travelers' parking decisions. The decision process begins during trip planning: at this stage, the model determines whether a traveler will park on-street or in a garage and identifies the specific garage if that alternative is chosen. If a traveler decides on garage parking, they drive directly to the designated facility. If a traveler intends to use on-street parking, they begin traveling toward their destination and search for available on-street spaces along the

way, effectively cruising for parking. If they arrive at their destination and do not find any available on-street space, an R-tree algorithm is used to locate the nearest alternative parking option. With parking decisions derived from the implemented parking choice model, travelers query shortest paths to get from one location to another, and the multimodal router provides information on-time to walk from parking location to the activity and back. The schematic framework of the parking trip process inside the simulation tool is illustrated in Figure 2.

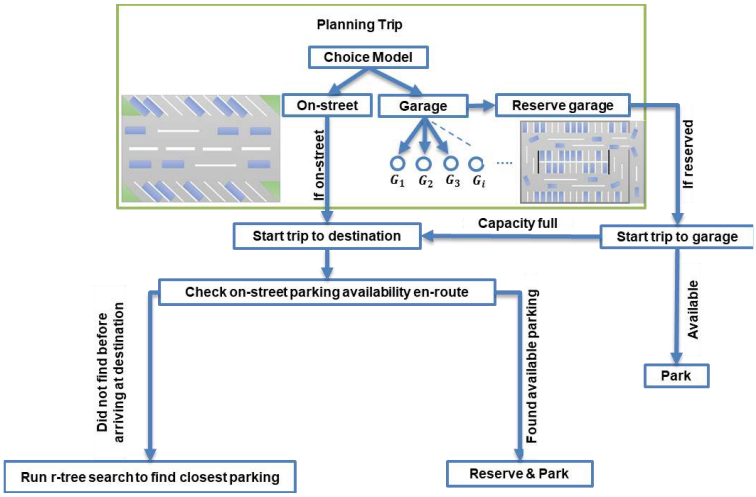



**FIGURE 2 Parking Framework Implemented inside POLARIS**

**RESULTS AND DISCUSSION:** To estimate the Bayesian model, we generated random on-street parking distances between 0 and 0.3 km (based on CMAP data) using a uniform distribution. On-street parking fees were drawn from a standard normal distribution, shifted and truncated to ensure positive values, with a mean of zero and a standard deviation of $3 per hour. The Metropolis–Hastings (MH) algorithm was then executed for 1,000 iterations (following the literature), and the resulting coefficients served as the initial values for 10,000 main iterations. The final model was applied to the Chicago network, which contains 48,379 links, 35,883 nodes, and 1,961 zones. To reduce computational time, 25% of the population was simulated, leading to approximately 2.6 million travelers making nearly 9 million one-way person-trips over a 24-hour period. Traffic scale factors were incorporated to align network performance with real-world conditions. Full details on the modeling, calibration, and results are provided in the complete paper.

**Parking Choice Model Estimation Results:** As expected, the likelihood of a traveler selecting a particular parking location decreases as the distance from the destination increases. The estimated standard deviations further indicate that distance has a stronger effect on travelers' garage choices than on-street choices. Similarly, higher parking fees reduce the probability of choosing that location, although fees had less influence on on-street parking decisions compared to garages. Time-of-day also affects parking preferences, with travelers more inclined to use garage parking during peak hours (7–10 AM and 5–8 PM). Additionally, travelers who consider arriving at the destination "very" or "extremely" important, according to the CMAP survey, tend to choose garage parking to save time. The model results show that travelers who are older than 60 years old prefer to park at on-street parking locations. Finally, in areas with a dense population and jobs, travelers prefer to park at garage locations due to a high demand for parking spaces. The credible intervals of the posterior distribution for the parking choice model attributes indicate that almost all attributes are statistically significant. The estimated parking choice model is implemented within POLARIS to explore the travelers' parking choice behavior under various policy-sensitive scenarios. Once implemented, the model was calibrated to ensure that the simulated parking trips resemble the observed trips in the CMAP data.

**Example of Sensitivity Analysis and Discussion:** Analysis of the simulated scenarios for the Chicago network (Figure 3) showed that, on average, only 50-70% of garage spaces were utilized during peak periods, suggesting potential for garage fee adjustments or capacity management to better accommodate specific vehicle types. When garage fees increased by 50%, the number of travelers who switched to on-street parking due to full garages dropped by 63%, while average garage occupancy during peak periods remained around 30% (Figure 4). Additionally, while adjusting garage fees citywide for the entire day significantly impacted the number of garage trips, fee

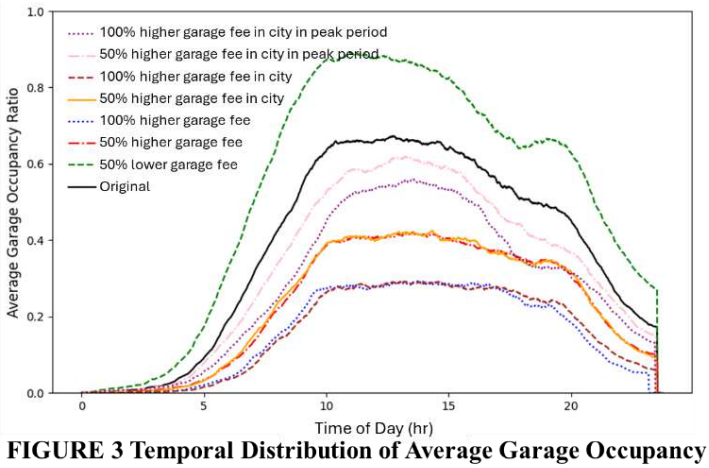



**FIGURE 3 Temporal Distribution of Average Garage Occupancy**

variations during peak periods had minimal effect due to the high importance of these trips and the limited availability of on-street parking in highly congested areas. The study's findings indicate that varying garage fees across different zones and time periods can effectively influence users' parking location choices. On-street parking and cruising for parking contribute to traffic congestion in large cities like Chicago. Therefore, dynamic garage pricing and capacity restrictions can serve as effective strategies for managing on-street parking demand. These insights can help develop a parking management strategy that maximizes the efficient use of available parking resources and optimizes curb space allocation. Additionally, the study's results can assist transport planners in designing a dynamic parking pricing model that effectively serves stakeholder interests.

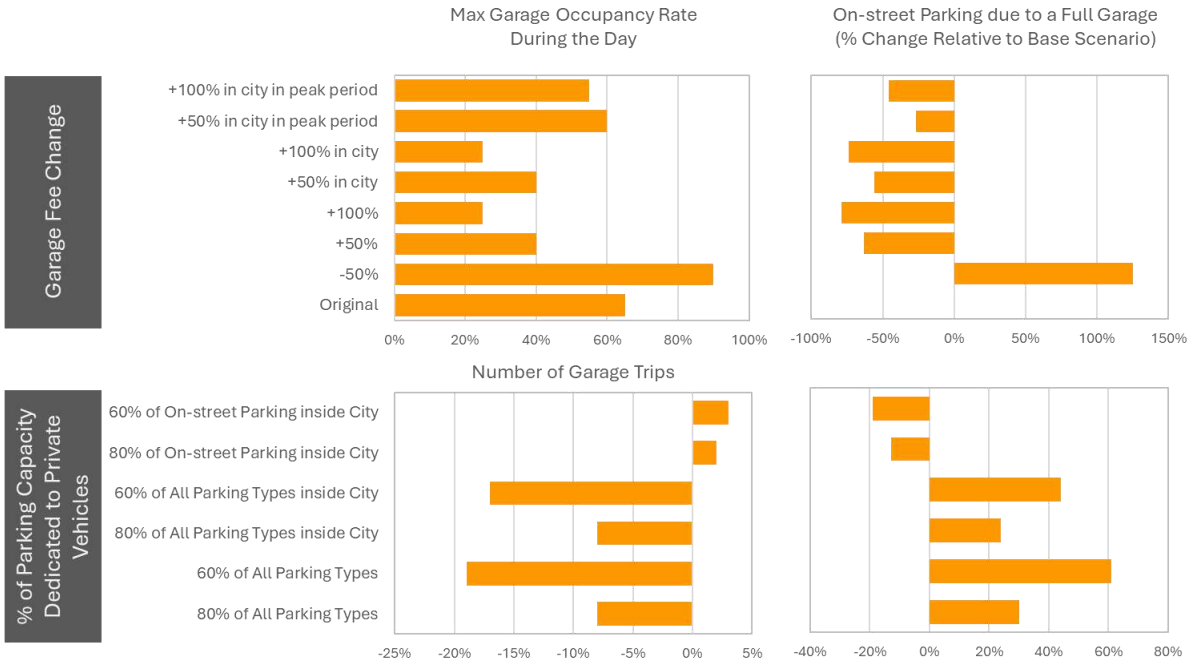



**FIGURE 4 Sensitivity of Parking Trips to Parking Fee and Capacity Constraints**

**ACKNOWLEDGMENT:** This study was sponsored by the U.S. Department of Energy (DOE) Vehicle Technologies Office (VTO) under the Pathways to Net-Zero Regional Mobility, an initiative of the Energy Efficient Mobility Systems (EEMS) Program.

## REFERENCES

1. Belloche, S. (2015). On-street Parking Search Time Modelling and Validation with Survey-based Data. *Transportation Research Procedia*, 6.
2. Ben Hassine, S., Mraihi, R., Lachiheb, A., & Kooli, E. (2022). Modelling parking type choice behavior. *International Journal of Transportation Science and Technology*, 11(3), 653–664.
3. Zhang, X., Liu, W., & Waller, S. T. (2019). A network traffic assignment model for autonomous vehicles with parking choices. *Computer-Aided Civil and Infrastructure Engineering*, *34*(12), 1100–1118.
4. Yang, S., & Qian, Z. (Sean). (2017). Turning meter transactions data into occupancy and payment behavioral information for on-street parking. *Transportation Research Part C: Emerging Technologies*, 78, 165–182.
5. Shoup, D. C. (2006). Cruising for parking. *Transport Policy*, 13(6), 479–486.
6. Fakhrmoosavi, F., Gurumurthy, K. M., Kockelman, K. M., Hunter, C. B., & Dean, M. D. (2024). Parking Strategies and Outcomes for Shared Autonomous Vehicle Fleet Operations. *Journal of Transportation Engineering, Part A: Systems*, *150*(4).
7. Auld, J., Hope, M., Ley, H., Sokolov, V., Xu, B., & Zhang, K., (2016). POLARIS: Agent-based modeling framework development and implementation for integrated travel demand and network and operations simulations. *Transportation Research Part C: Emerging Technologies*, 64, 101-116.
8. Chicago Metropolitan Agency for Planning (CMAP), (2019). My Daily Travel survey data: *https://www.cmap.illinois.gov/data/transportation/travel-survey*
9. Lahiri, K., & Gao, J. (2002). Bayesian analysis of nested logit model by Markov chain Monte Carlo. *Journal of Econometrics*, *111*(1), 103–133.
10. Roberts, G. O., & Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical science*, 16(4), 351-367.

# Optimizing the Placement of Dedicated Lanes for Autonomous Vehicles in Large-Scale Networks: A Simulation-based Framework

**Ehsan Kamjoo**
Doctoral Researcher, Michigan State University
428 S. Shaw Ln., East Lansing, MI, USA, 48824
Email: kamjooeh@msu.edu

**Alireza Rostami**
Doctoral Researcher, Michigan State University
428 S. Shaw Ln., East Lansing, MI, USA, 48824
Email: darzianr@msu.edu

**Fatemeh Fakhrmoosavi**
Assistant Professor, University of Connecticut
Francis L.Castleman Bldg, Storrs, CT, USA, 06269
Email: mousavi@uconn.edu

**Ali Zockaie** (Corresponding Author)
Associate Professor, Michigan State University
428 S. Shaw Ln., East Lansing, MI, USA, 48824
Phone: 517-355-8422; Fax: 517-432-1827; Email: zockaiea@egr.msu.edu

An Extended Abstract submitted to be considered for presentation at the 10th International Symposium on Dynamic Traffic Assignment (DTA) 2025.

## ABSTRACT

This study introduces a framework to maximize societal benefits associated with the AV dedicated lane implementation at large-scale transportation networks, considering the travel time savings and the required investments to prepare the infrastructure for their deployment. To this end, a bi-level optimization problem is formulated. The upper level determines the links for dedicated lane deployment, while at the lower level, a mesoscopic traffic simulation tool is employed to enable a realistic representation of these vehicles in a mixed traffic. The problem is solved using the Genetic Algorithm. To further reduce the computational burden, this study adopts a clustering method based on the snake algorithm to group the candidate links and reduce the size of the solution space. The proposed framework is successfully applied to the case study of Chicago downtown network, considering various demand levels, AV market penetration rates, and implementation approaches. The results highlight the need for optimizing the placement of AV dedicated lanes to ensure economically beneficial adoption of this strategy across different scenarios. This study provides transportation planners with key operational insights to facilitate the effective adoption of AV dedicated lanes during the transitional phase from HDVs to a fully AV environment.

**Keywords:** Autonomous Vehicles, Traffic Simulation, Dedicated Lanes, Market Penetration Rate, Bi-Level Optimization

## RESEARCH MOTIVATION

Autonomous Vehicle (AV) technology is increasingly being recognized for its potential to enhance mobility in transportation systems through Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communication. AVs operate with shorter headways, form platoons, and increase roadway capacity compared to Human-Driven Vehicles (HDVs) (1). However, during the transition period to a fully AV environment, the coexistence of AVs and HDVs presents challenges, such as increased driving complexity and hindered AV platoon formation.

To maximize AV benefits during this transition, deploying dedicated lanes for AVs (AVDLs) has been proposed to address some of the issues regarding the interactions between AVs and HDVs. Two approaches can be considered for their implementation. First, they can be added to the existing roadway infrastructure (ADL), and alternatively, they can be deployed by restricting HDVs from accessing one of the existing lanes and reserving that lane for AVs (EDL). While ADL requires substantial infrastructure investment, EDL is more cost-effective but may induce congestion and shockwaves at entry/exit points (2). This also brings up equity concerns regarding the experience of HDV users in the network (3), due to the potential longer travel times and reduced accessibility for the HDV users. Therefore, it is essential to investigate the effects of AVDL deployment in transportation networks.

The effectiveness of this strategy depends on several factors, including the AV Market Penetration Rate (MPR) and technological capabilities, network topology, and travel demand level (4). In this regard, extensive research has been conducted in the literature to study the impacts of AVDLs on improving the traffic performance at the corridor level (5). However, the impacts of these dedicated lanes can propagate over the network. The AV and HDV users' route choice can also change since more demand is attracted to the links with AVDLs and their adjacent links. Therefore, research is warranted to evaluate their impacts on traffic performance at the network-level. Besides, the AVDLs might not necessarily improve the average travel time of vehicles, as the HDVs would lose their access to a proportion of the network if the AVDL placement is not optimized (6). This calls for identifying optimal locations for their deployment.

To this end, this study aims to propose a framework for optimizing the locations for implementation of AVDLs at large-scale networks. As outlined in Table 1, while some studies examined the effects of AVDLs in small networks, this problem is not fully explored in the literature, especially in the context of large-scale networks, while incorporating the differences in microscopic behavior of AVs and HDVs, the effects of presence of AVs on traffic dynamics, and en-route behavior of the vehicles. In this study, the locations for AVDL deployment are optimized in the network to maximize societal benefits, considering travel time savings and infrastructure modification costs, through a bi-level optimization framework. In this regard, an updated version of DYNASMART-P, a mesoscopic traffic simulation tool (7), is used to assess the impacts of AVDLs at lower level. This simulation tool considers the interactions between different vehicle types and the impacts of AVs' presence on network performance. The unique features of this tool are presented later.

To mitigate the computational burden associated with traffic simulation and dynamic traffic assignment, this study adopts a clustering approach, originally introduced by Saeedmanesh and Geroliminis (8), to address the practical consideration of minimum length of consecutive links for AVDL deployment and effectively reduce the size of the solution space. By using this approach, the potential candidate links are grouped into clusters to constrain the size of the solution space and also the minimum length of AVDLs. The optimization problem is solved using the Genetic Algorithm (GA). The proposed framework is successfully applied to the Chicago downtown network. The main contributions of this study are as follows.

- Proposing a framework for optimizing the locations of implementing AVDLs at large-scale networks considering three specific aspects of this network design problem jointly: traffic simulation for mixed fleet of AVs and HDVs; identifying candidate locations based on practical constraints for AVDLs; and computational efficiency for large-scale applications
- Employing a clustering approach to group the candidate links for AVDL deployment
- Examining the effects of different travel demand levels, AV MPRs, and AVDL implementation approaches on the effectiveness of this strategy as well as optimal locations for their deployment
- Exploring the impacts of AVDL adoption on travel time for both HDVs and AVs

**TABLE 1 Summary of related works on optimizing the placement of AVDLs at the network-level**

| Authors | Strategy | Method (Lower-level) | Specifications | Scenarios |
|---|---|---|---|---|
| Chen et al. (13) | AVDL with endogenous demand | Analytical (Multiclass UE and Diffusion models) | • Per-lane capacity becomes tripled when converted to AVDL<br>• AVs have no impact on regular lanes | South Florida network with 2% initial MPR |
| Movaghar et al. (15) | AVDL | Analytical (Multiclass UE model) | • Link capacity is adjusted as a function of AV proportion on the link | Sioux Falls network with various MPRs |
| Madadi et al. (17) | AVDL and AV zone | Analytical (Multiuser class SUE model) | • Capacity of AVDL is assumed to be 1.5 times of regular lanes<br>• Use of AVDL is mandatory for AVs<br>• AVs have no impact on regular lanes | Amsterdam network with various MPRs |
| Seilabi et al. (16) | AVDL with demand uncertainty | Analytical (UE and Diffusion models) | • Per-lane capacity becomes tripled when converted to AVDL | Synthetic network with various potential MPRs |
| Pourgholamali et al. (6) | AVDL and pricing policy | Analytical (Multiclass UE and Diffusion models) | • Per-lane capacity is multiplied by a factor when converted to AVDL | Sioux Falls network |
| This study | AVDL | Simulation-based (Dynamic Traffic Assignment) | • Distinct micro-models for AVs & HDVs and en-route behavior<br>• Practical locations for AVDLs<br>• Intersection capacity variations AVs<br>• Incorporating mixed traffic dynamics | • Chicago network for different MPRs & demand levels<br>• EDL & ADL implementation |

## PROBLEM STATEMENT

This study considers a large-scale general network, $G(V, E)$, where $V$ denotes the set of nodes (i.e., interchanges or intersections) and $E$ represents the set of links (i.e., road segments). A subset of freeway links, $F \subseteq E$, is considered as potential candidate links for AVDL deployment. Two approaches ($a$) are considered for AVDLs implementation: 1) dedicating an existing lane on freeway links to AVs (EDL), and 2) adding a new lane to freeway links as dedicated lane for AVs (ADL). The costs related to constructing a new lane and the technological costs, involved in developing an AV lane are taken into account in the analyses. The objective function for determining the optimal placement for AVDLs, $\Psi$, considering the travel time benefits and infrastructure modification costs, is formulated as follows.

$$Max_{\mathbf{X}}\ \Psi = \sum_{\Delta t}[TT^0(\Delta t) - TT(\Delta t)]\beta - \sum_{\Delta t}\sum_{f \in F} x_f C_a(\Delta t) L_f \quad (1)$$

$$x_f \in \{0,1\}, \forall\ a \in \{EDL, ADL\}$$

The decision variable in the Equation (1) is X, which is a vector of binary variables $x_f$ for each candidate link $f \in F$, indicating whether that link is selected for AVDL deployment (0: not selected, and 1: selected). The first term represents the travel time savings due to the AVDL deployment, where $TT(\Delta t)$ and $TT^0(\Delta t)$ are the system travel time of all vehicles within time interval $\Delta t$ after and before the deployment of AVDLs, and $\beta$ is the average value of travel time for the users in the network. The second term represents the infrastructure modification costs required for deployment of these lanes. Here, $C_a(\Delta t)$ is the cost associated with the AVDL implementation, in unit of $ per mile per time period $\Delta t$, which varies based on the chosen implementation approach, $a$. In this regard, for both implementation approaches, cost associated with transforming a standard lane into an AV-enabled lane is considered. Besides, the construction cost as well as the operational and maintenance cost of the added lane are considered for the ADL approach. Finally, $L_f$ is the length of the candidate link $f$. Different demand levels, AV MPRs, and implementation approaches are considered for developing various scenarios to observe the effects of these factors on the optimal placement and the cost-effectiveness of AVDL strategy.

## METHODOLOGY

A bi-level optimization framework is defined to identify the optimal locations of AVDLs in the network to maximize the objective function (Equation (1)). The upper-level problem aims to determine the links, on

which AVDLs need to be deployed to maximize the benefits. For estimating the objective function at lower-level, an updated version of DYNASMART-P, a mesoscopic simulation tool, is used to obtain vehicle trajectories and estimate the total system travel time. In this regard, based on the selected links in the upper-level, the network structure is altered to accommodate AVDLs on selected links in traffic simulation.

Besides, to improve the computational efficiency, a state-of-the-art clustering approach is adopted to group the candidate links based on their expected impacts to prune the solution space and also constrain the minimum length of consecutive links for AVDLs deployment to avoid confusion for drivers by preventing frequent lane changes. This aligns the deployment strategy with practical considerations essential for the safe integration of AVDLs. The clusters of candidate links are then given as the decision variables to the optimization framework. The solution approach, traffic simulation, and the clustering method utilized in this study are described in subsequent subsections.

### Solution Approach

The optimization problem in this study is not solvable by exact methods due to the integration of a mesoscopic traffic simulation tool. Therefore, a Genetic Algorithm (GA) is employed to maximize the objective function in Equation (1). The GA's parallelization capability makes it well-suited for this problem (13). At each generation, the objective function value is evaluated at the lower level for each individual in the population. This involves modifying the network structure to include AVDLs at selected locations and running the traffic simulation.

### Traffic Simulation Tool

This study uses a recently updated version of DYNASMART-P to simulate the mixed traffic of AVs and HDVs in a large-scale network (7). The mesoscopic simulation tool integrates adaptive fundamental diagrams to account for the non-uniform distribution of AVs and HDVs. It incorporates distinct microscopic models: the stochastic acceleration model of Hamdar et al. (14) for HDVs and the model of Talebpour and Mahmassani (15) for AVs. The HDV model also considers driver heterogeneity during calibration. These microscopic models determine spacing-speed relationships, which are used to derive the congested section of the macroscopic fundamental diagram and dynamically update speed values across links and time intervals. Additionally, the tool refines traffic flow models for arterial links and intersection capacities based on vehicle type proportions passing each intersection.

The simulation tool allows real-time route switching (en-route behavior) for AVs and for a portion of HDVs. It is further customized to model AVDLs, which are implemented as parallel links restricting HDV access. The upper-level framework determines AVDL locations, which serve as inputs to the simulation tool to modify the network structure accordingly. For further details on the simulation tool, refer to (7,16). Figure 1 briefly outlines the framework used for simulating AVs and HDVs across a large-scale network.

### Clustering of Candidate Links for AVDL Deployment

As mentioned earlier, this study adopts the clustering method by Saeedmanesh and Geroliminis (8), to group candidate links into the clusters of links with homogeneous impacts after AVDL implementation. This technique, commonly used for network partitioning, accounts for the spatial correlation of congestion (17). This algorithm is chosen for its inherent capability to ensure connectivity within each cluster, a crucial requirement for practical AVDL deployment, while also satisfying the minimum cluster length constraint. Clustering also reduces the solution space, thereby leading to a notable reduction in computational complexity. For more information, please refer to the study by (8).

The measure for evaluating the homogeneity between the candidate links is set to be the percent change in the total throughput of a link over the simulation horizon, when no AVDL is deployed, and when the AVDLs are implemented on all the candidate links. This metric effectively captures key AVDL impacts, including traffic stabilization and vehicle platooning. Since the throughput difference varies with AV MPR, demand level, and AVDL implementation approach, freeway link clustering is performed separately for each scenario.

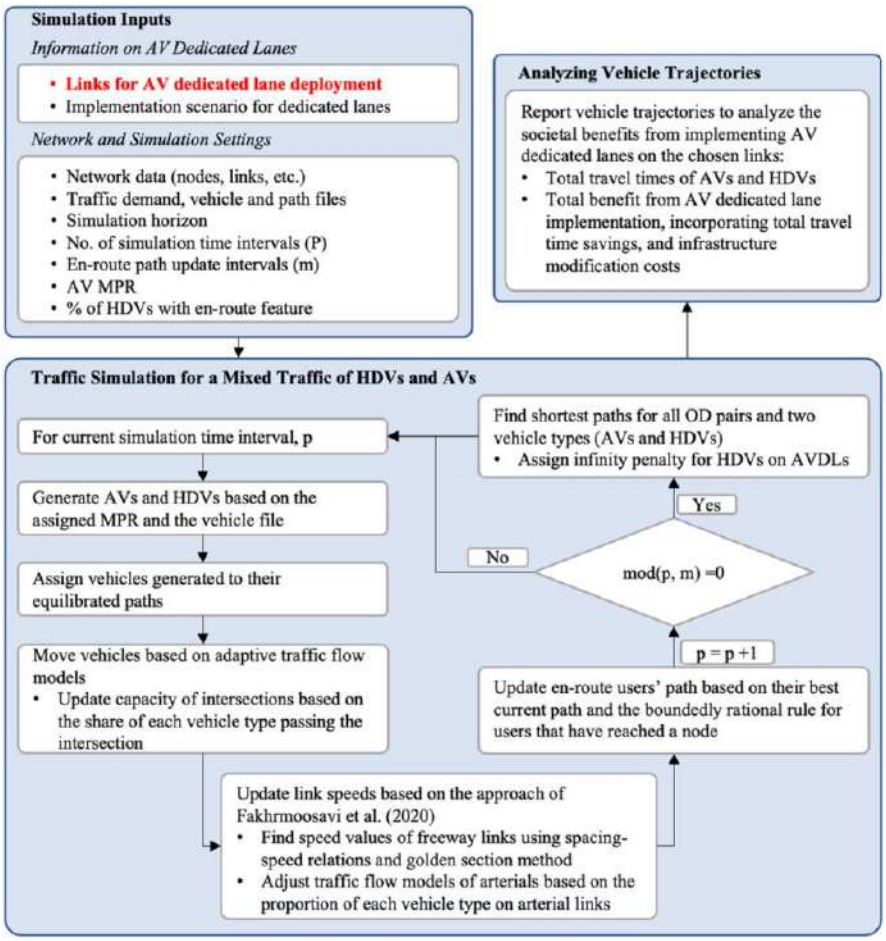



**FIGURE 1 Traffic simulation framework for mixed traffic of HDVs and AVs considering AVDLs at large-scale network**

Overall, in the proposed framework of this study, the clustering process is performed to group the candidate links in terms of the percent change in the total throughput of a link over the simulation horizon. Then, the formed clusters as well as the network data, demand level, AVDL implementation approach, infrastructure modification costs, and the GA hyperparameters are given to the GA. Within the GA, for each generated individual (potential solution), the objective function (Equation 1) is evaluated by obtaining the travel times of the vehicles using the traffic simulation tool and considering the infrastructure modification costs. The GA continues iterating until reaching the maximum generation count. Finally, the best solution found is reported, representing the optimized placement of AVDLs within the network.

## CASE STUDY

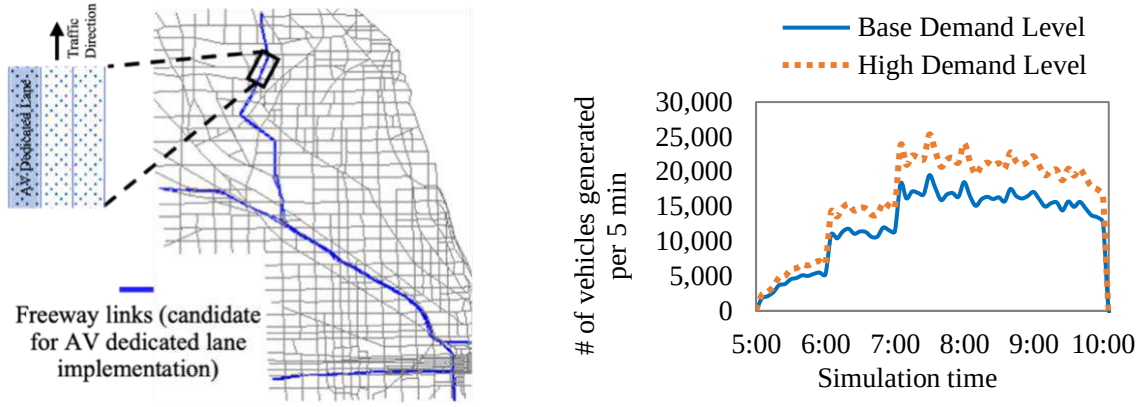
The proposed framework in this study is applied to the Chicago downtown network, which comprises 1578 nodes, 4805 links, and 218 traffic analysis zones. The network consists of a total of 150 freeway links (73 miles in total), which are considered as potential candidates for deploying AVDLs. Figure 2(a) demonstrates a schematic depiction of the network. As stated earlier, two implementation scenarios of EDL and ADL are explored AVDLs. These two implementation scenarios are compared to the base scenario (w/o AV dedicated lane), where no AVDL is deployed on the network.

The simulation is conducted for the AM peak period from 5:00 AM to 10:00 AM. An extra two hours of simulation is incorporated to unload the network. This study explores two demand levels: The "base demand level", which is calibrated based on historical real-world data (18), resulting in about 760,000 vehicles in the network. The second scenario is "high demand level", which is generated by increasing base demand with 30% additional demand. This scenario is included since the presence of AVs is anticipated to increase vehicle-miles traveled. Figure 2(b) shows the demand profiles associated with base and high

demand level scenarios. Various AV MPRs from 10% to 90% with increment of 10% are evaluated. The simulation is conducted considering a non-uniform distribution of AVs and HDVs, comprising 10 distinct driver classes, across the network.

The preferred headway of AVs is considered to be 1 second, when operating on the regular lanes. It was found that the preferred headway of AVs could be lower when operating within AV platoons, as opposed to when they are in a mixed traffic stream (6). In this regard, in this study, the preferred headway of AVs is considered to be 0.5 second on the dedicated lanes (6). The average computational time of performing the traffic simulation is about 25 minutes on an Intel(R) Xeon(R) Gold 6246R CPU @ 3.40 GHz (2 processors) with 256 GB RAM. Regarding the values of parameters related to the infrastructure modification costs associated with AVDLs, please refer to (6). The average value of travel time of $20 per hour for users is considered. The AM peak period travel time is multiplied by a factor to be converted to daily travel time. In addition, for clustering the candidate links, the number of clusters is considered to be 15 as the baseline, and the minimum length of clusters is assumed as 1 mile.

The net benefit, total length of selected links for AVDL deployment, and travel time improvements for AVs and HDVs are analyzed across different scenarios. Also, sensitivity analyses are conducted on infrastructure modification costs and number of clusters to evaluate the effects of these parameters on optimization results and optimal locations for AVDLs.




(a) (b)

**FIGURE 2 a) Schematic depiction of case study network, and b) Profiles of different demand scenarios for the AV peak period**

## KEY FINDINGS

The key findings of this study are as follows.

- The cost-effectiveness, optimal placement, and the benefits of AVDL strategy at the network level are sensitive to AV MPR, demand level, and the implementation approach for these lanes.
- Optimizing the placement of AVDLs in the network, considering the infrastructure modification costs, is essential to make the adoption of this strategy economically beneficial for different demand levels and AV MPRs.
- Incorporating the effects of AVs on the performance of the regular lanes is crucial when exploring the cost-effectiveness of AVDL strategy at the network level, particularly for high AV MPRs.
- The AVDL strategy is more effective for improving the mobility in the network when the AV MPR is relatively low (20% and 30%).
- Converting an existing lane to AVDL at optimal locations reduces the travel times of HDVs, which holds significant relevance when evaluating the equity and fairness of this strategy.
- The benefits of AVDLs deployment on optimal locations in the network is more evident for high demand levels, compared to when the congestion level is relatively low in the network.
- None of the explored implementation approaches is more cost-effective than the other one across various demand levels and AV MPRs, which calls for different analyses based on the predicted market share and travel demand in future.

## REFERENCES


1. Li Q, Li X. Trajectory planning for autonomous modular vehicle docking and autonomous vehicle platooning operations. Transp Res E Logist Transp Rev. 2022;166:102886.
2. He S, Ding F, Lu C, Qi Y. Impact of connected and autonomous vehicle dedicated lane on the freeway traffic efficiency. European Transport Research Review. 2022;14(1):1–14.
3. Pourgholamali M, Miralinaghi M, Ha PYJ, E Seilabi S, Labi S. Sustainable Deployment of Autonomous Vehicles Dedicated Lanes in Urban Traffic Networks. Sustain Cities Soc. 2023;99:104969.
4. Seilabi SE. Lane Management in the Era of Connected and Autonomous Vehicles Considering Sustainability [Internet]. 2022. Available from: https://www.proquest.com/docview/2838333451?pq-origsite=gscholar&fromopenview=true&sourcetype=Dissertations & Theses
5. Chen Y, Zhang H, Wang D, Wang J. Overall Influence of Dedicated Lanes for Connected and Autonomous Vehicles on Freeway Heterogeneous Traffic Flow. J Adv Transp. 2022;2022:1–15.
6. Fakhrmoosavi F, Kamjoo E, Zockaie A, Mittal A, Fishelson J. Assessing the Network-Wide Impacts of Dedicated Lanes for Connected Autonomous Vehicles. Transp Res Rec. 2023;2677(3):371–88.
7. Fakhrmoosavi F, Saedi R, Zockaie A, Talebpour A. Impacts of Connected and Autonomous Vehicles on Traffic Flow with Heterogeneous Drivers Spatially Distributed over Large-Scale Networks. Transp Res Rec. 2020;2674(10):817–30.
8. Saeedmanesh M, Geroliminis N. Dynamic clustering and propagation of congestion in heterogeneously congested urban traffic networks. Transportation research procedia. 2017;23:962–79.
9. Chen Z, He F, Zhang L, Yin Y. Optimal deployment of autonomous vehicle lanes with endogenous market penetration. Transp Res Part C Emerg Technol. 2016;72:143–56.
10. Movaghar S, Mesbah M, Habibian M. Optimum location of autonomous vehicle lanes: A model considering capacity variation. Math Probl Eng. 2020;2020:1–13.
11. Madadi B, Van Nes R, Snelder M, Van Arem B. Optimizing road networks for automated vehicles with dedicated links, dedicated lanes, and mixed-traffic subnetworks. J Adv Transp. 2021;2021:1–17.
12. Seilabi SE, Pourgholamali M, Homem de Almeida Correia G, Labi S. Robust design of CAV-Dedicated lanes considering CAV demand uncertainty and lane reallocation policy. Transp Res D Transp Environ [Internet]. 2023;121:103827. Available from: https://www.sciencedirect.com/science/article/pii/S1361920923002249
13. Chen S, Wang H, Meng Q. Designing autonomous vehicle incentive program with uncertain vehicle purchase price. Transp Res Part C Emerg Technol. 2019;103:226–45.
14. Hamdar SH, Treiber M, Mahmassani HS. Calibration of a Stochastic Car-Following Model Using Trajectory Data: Exploration and Model Properties. In: Transportation Research Board 88th Annual Meeting. 2009.
15. Talebpour A, Mahmassani HS. Influence of connected and autonomous vehicles on traffic flow stability and throughput. Transp Res Part C Emerg Technol. 2016;71:143–63.
16. Fakhrmoosavi F, Kamjoo E, Kavianipour M, Zockaie A, Talebpour A, Mittal A. A stochastic framework using Bayesian optimization algorithm to assess the network-level societal impacts of connected and autonomous vehicles. Transp Res Part C Emerg Technol. 2022;139:103663.
17. Saedi R, Saeedmanesh M, Zockaie A, Saberi M, Geroliminis N, Mahmassani HS. Estimating network travel time reliability with network partitioning. Transp Res Part C Emerg Technol. 2020;112:46–61.
18. Zockaie A, Chen Y, Mahmassani HS. Adaptive drivers and time-dependent origin-destination demand estimation: methodology and application to large-scale network. 2014.

# Experimental study on characteristics of autonomous vehicles under different loads

Shi-Teng Zheng[1], Jian Shang[1], Rui Jiang[1,*]

[1] *School of Systems Science, Beijing Jiaotong University, Beijing 100044, China*

Over the past decade, autonomous vehicle (AV) has undergone rigorous testing and evaluation (Feng et al., 2023; Feng et al., 2021; Yan et al., 2023), and has developed from concept to commercial reality. The AV technology is designed not only to alleviate driving tasks of human drivers, mitigate human errors (Braun and Randell, 2020), and enhance driving safety (Chirachavala and Yoo, 1994; Kikuchi et al., 2003), but also to enhance traffic efficiency and improve traffic capacity (Potluri, 2023). Due to their potential to revolutionize traffic safety and mobility, an increasing number of productive AVs are released on the road, paving the way for large-scale autonomous driving in future transportation.

Real-world experimental studies using commercial AVs, particularly those equipped with adaptive cruise control (ACC) systems, have been conducted to gain deeper insights into AV behavior and characteristics. For example, Milanés and Shladover (2014) conducted a four-vehicle platoon experiment with the leading vehicle executing the speed changes and pointed out that the multiple consecutive ACC vehicles are unstable. Knoop et al. (2019) carried out the experiment on the public freeway using seven vehicles equipped with SAE level-2 automation and also found that the ACC vehicles' control system is unstable. Makridis et al. (2020) studied a five-vehicle platoon and showed that the platoon is unstable in perturbation situations. Gunter et al. (2021) also tested the string instability of an eight-vehicle platoon with seven ACC vehicles following under a perturbation and observed the last vehicle disengaging from ACC due to the violent oscillations. Ciuffo et al. (2021) carried out an experimental study of ACC platoon under different perturbation situations and summarized that the ACC vehicle platoon under the minimum time gap setting exhibits more unstable behavior, while it is stable under the maximum time gap setting.

It is noteworthy that current experimental studies on AVs primarily focus on the AVs themselves, often neglecting the impact of external factors such as vehicle load conditions. For instance, a typical passenger car with a seating capacity of five weighs of approximately 1,400-2,000 kg, with a typical load of about 375-500 kg, roughly accounting for 1/4 of the total vehicle weight. This variation can significantly influence the power-to-weight ratio and, consequently, AV performance. Murthy and Masrur (2016) highlighted that different load conditions affect a deceleration capability, underscoring the need to reassess safe braking conditions within a platoon to establish appropriate following distances.

* Corresponding author: jiangrui@bjtu.edu.cn

Motivated by the fact, very recently, we performed an experimental study on the characteristics of AVs under varying load conditions on Nov. 7, 2024, on a about 1.5-kilometer straight track in the closed test field affiliated to Research Institute of Highway, Ministry of Transport, China. The snapshots of experiment are provided in Figure 1.

(a) (b)

Figure 1. Snapshot of the experiment. (a) The AV tracking the pre-specified speed. (b) The load applied using standard sandbags, each weighing 25 kg.

In the experiment, we employed a programmable AV, as opposed to a commercially available AV, allowing precise control over a pre-specified speed profile without the presence of leading vehicles. Our study focuses solely on longitudinal control, which is implemented in two levels: at the upper level, an acceleration command is generated based on the tracked speed profile, while at the lower level, the acceleration command is executed through the AV's motion control mechanisms.

To simulate real-world traffic oscillations, the AV was programmed to follow a carefully designed speed profile consisting of multiple target speeds of 0, 10, 20, 30, 40, and 50 km/h. The transitions between these stable speeds were executed with a uniform acceleration or deceleration of 1 m/s$^2$. The pre-specified speed profile is illustrated in Figure 2.

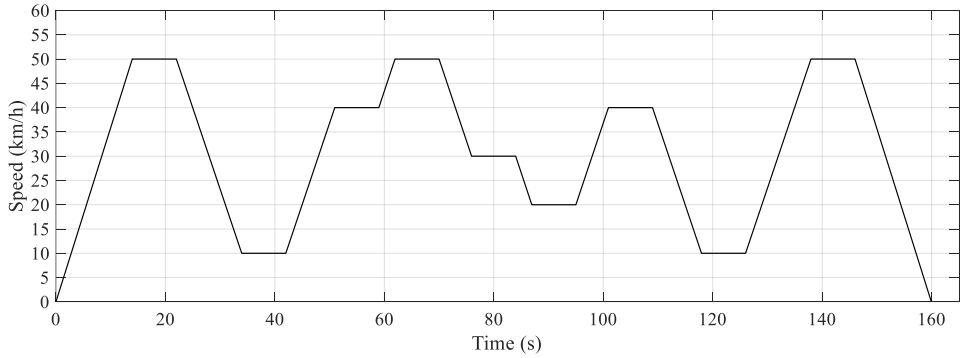



Figure 2. The pre-specified speed profile of the AV in the experiment.

The load is adjusted using standard 25 kg sandbags, which are added or removed at one end of the straight track. After completing a full speed profile—defined as one run—the AV performs a U-turn and prepares for the next run to return to the starting point, at which point the load is

modified. Consequently, each load condition is tested twice. The experimental run settings are detailed in Table 1. The AV is equipped with onboard sensors, including LiDAR and high-precision GPS, which facilitate its autonomous driving functions and enable continuous data collection throughout the experiments at a frequency of 20 Hz.

Table 1. Details of the experiment runs in load order.

| No. | Experiment time | Loads (kg)† |
|---|---|---|
| 1 | 10:58 - 11:03 | 0 |
| 2 | 11:05 -11:13 | 50 |
| 3 | 11:35 - 11:41 | 100 |
| 4 | 11:44 - 11:50 | 150 |
| 5 | 12:55 - 13:03 | 200 |
| 6 | 13:05 - 13:11 | 250 |
| 7 | 13:38 - 13:44 | 300 |
| 8 | 13:22 - 13:29 | 350 |

†Excluding the driver's weight, the driver is responsible for safety supervision and steering operations in the car, with a weight of 75 kg.

The experimental results for the command acceleration, actual acceleration, and actual speed of the AV are reported in Figure 3. It is evident that under varying load conditions, the AV is more significantly affected during the deceleration phase compared to the constant speed and acceleration phases. In this case, the vehicle adjusts its command acceleration to maintain the pre-specified speed, resulting in variations in actual acceleration and reduced speed control precise. Notably, as the load increases (represented by the color transition from cool to warm), the commanded acceleration decreases during deceleration, while the range of actual acceleration narrows — the peak acceleration is reduced, and the valley acceleration is elevated.

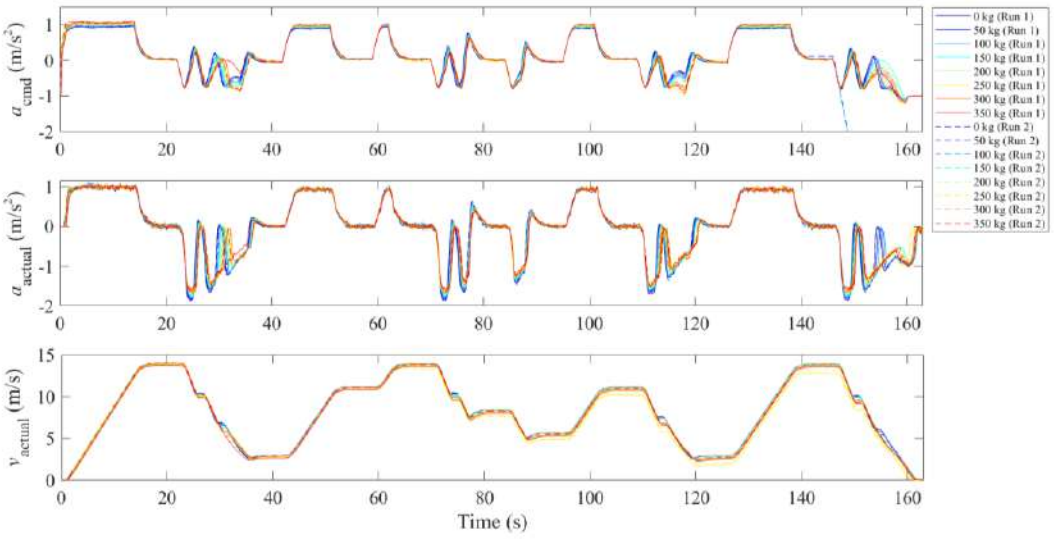



Figure 3. The experimental command acceleration, actual acceleration, and actual speed of the AV.

We would like to highlight that the contribution of this study is at least twofold. (i) It provides insights into the real-world performance of AVs under varying load conditions. (ii) The experimental results emphasize that vehicle load cannot be ignored, offering valuable information for the development of more precise AV control algorithms.

In future work, we will present further data analysis of the experiment, including a quantitative comparison of the differences across load conditions, to better understand the relationship between AV performance and vehicle load. Additionally, we will model the lower-level controller incorporating load factors and conduct a simulation study to analyze the impact of AV loads on traffic flow characteristics.

## References


Braun, R., Randell, R., 2020. Futuramas of the present: The "driver problem" in the autonomous vehicle sociotechnical imaginary. *Humanities & Social Sciences Communications* 7 (1), 163.

Chirachavala, T., Yoo, S. M., 1994. Potential safety benefits of intelligent cruise control-systems. *Accident Analysis and Prevention* 26 (2), 135-146.

Ciuffo, B., Mattas, K., Makridis, M., Albano, G., Anesiadou, A., He, Y. L., Josvai, S., Komnos, D., Pataki, M., Vass, S., Szalay, Z., 2021. Requiem on the positive effects of commercial adaptive cruise control on motorway traffic and recommendations for future automated driving systems. *Transportation Research Part C* 130, 103305.

Feng, S., Sun, H., Yan, X., Zhu, H., Zou, Z., Shen, S., Liu, H. X., 2023. Dense reinforcement learning for safety validation of autonomous vehicles. *Nature* 615 (7953), 620-627.

Feng, S., Yan, X., Sun, H., Feng, Y., Liu, H. X., 2021. Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. *Nature Communications* 12 (1), 748.

Gunter, G., Gloudemans, D., Stern, R. E., McQuade, S., Bhadani, R., Bunting, M., Delle Monache, M. L., Lysecky, R., Seibold, B., Sprinkle, J., Piccoli, B., Work, D. B., 2021. Are commercially implemented adaptive cruise control systems string stable? *IEEE Transactions on Intelligent Transportation Systems* 22 (11), 6992-7003.

Kikuchi, S., Uno, N., Tanaka, M., 2003. Impacts of shorter perception-reaction time of adapted cruise controlled vehicles on traffic flow and safety. *Journal of Transportation Engeneering* 129 (2), 146-154.

Knoop, V. L., Wang, M., Wilmink, I., Hoedemaeker, D. M., Maaskant, M., Van der Meer, E.J., 2019. Platoon of sae level-2 automated vehicles on public roads: Setup, traffic interactions, and stability. *Transportation Research Record* 2673 (9), 311-322.

Makridis, M., Mattas, K., Ciuffo, B., Re, F., Kriston, A., Minarini, F., Rognelund, G., 2020. Empirical study on the properties of adaptive cruise control systems and their impact on traffic flow and string stability. *Transportation Research Record* 2674 (4), 471-484.

Milanés, V., Shladover, S. E., 2014. Modeling cooperative and autonomous adaptive cruise control dynamic responses using experimental data. *Transportation Research Part C* 48, 285-300.

Murthy, D. K., Masrur, A., 2016. Braking in close following platoons: The law of the weakest, *19th Euromicro Conference on Digital System Design* (DSD), Limassol, Cyprus, pp. 613-620.

Potluri V. *Developing a cyber-physical system for real-time proactive traffic control and management of mixed fleet of vehicles with various levels of autonomy*. Dissertation: Arizona State University (2023).

Yan, X., Zou, Z., Feng, S., Zhu, H., Sun, H., Liu, H. X., 2023. Learning naturalistic driving environment with statistical realism. *Nature communications* 14 (1), 2037.

The Dynamic Spatial Price Equilibrium Problem

Terry L. Friesz

Cheng-Chang Lin

**Introduction.** The spatial price equilibrium problem explains how the demand for freight services is derived from spatially separated production and consumption activities. We are concerned herein with generalization of the notion of spatial price equilibrium from a static to a dynamic setting under highly specific circumstances to facilitate the practical modeling of truck borne freight flows. In that spatial price equilibrium is considered an expression of how the transport of commodities occurs in response to differences between origin and destination prices of transportable goods, it is an obvious paradigm for such modeling.

Spatial price equilibrium (SPE) is achieved when remote market price equals local market price plus the cost of transporting the goods of interest to the remote market. The essential elements of static SPE theory are generally thought to have first been articulated in the seminal work of Cournot (1927). Later Samuelson (1952), Beckmann et al. (1956), Smith (1963) and Takayama and Judge (1964) connected SPE theory to linear and nonlinear programming, thereby opening up the prospect of solving truly large SPE problems.

Further refinements of static SPE theory, especially the introduction of congestion, fully general supply and demand functions, and nonlinear complementarity and variational inequality formulations were made by Florian and Los (1982), Friesz et al. (1983), Friesz and Tobin (1983), Dafermos (1983), Dafermos and Nagurney (1984), Smith (1984), Chao and Friesz (1984), Smith and Friesz (1985), and several others over the last 70 years. Notably Friesz et al. (1983) and Dafermos and Nagurney (1984) present perhaps the earliest variational inequality formulations of static SPE; Friesz et al. (1983) also present the first nonlinear complementarity formulation.

It is Beckmann in Chapter 5 of Beckmann et al. (1956) who is the first to give the key insights for an extremal formulation of spatial price equilibrium; this occurs in the same book where Beckmann gives the much studied and widely employed mathematical programming formulation of static user equilibrium of passenger flows. Also found in Chapter 5 of Beckmann et al. (1956) are remarks about how dynamic equilibrium models might be constructed. Although Beckmann's suggestions regarding dynamic equilibrium have been followed with respect to dynamic user equilibrium (Wardrop's first principle), relatively little research has been done on extensions of spatial price equilibrium to a dynamic setting.

In this paper, we present a dynamic extension of spatial price equilibrium and, in the process, provide (*i*) a succinct definition of dynamic spatial price equilibrium (DSPE), (*ii*) an associated differential variational inequality (DVI), and (*iii*) an analysis of that DVI. These are provided in the hope of stimulating research on DSPE by other scholars, as well as applications by professionals involved in strategic freight planning. Our contribution is the first spatial price equilibrium formulation that explicitly treats lead times (time shifts) in a continuous-time context in the articulation of inventory dynamics, when such lead times are dictated by the transport of goods to distant markets. Intimately tied to our basic formulation is its reformulation as a differential variational inequality (DVI), which immediately provides necessary conditions that must be satisfied by equilibrium solutions.

In this paper we show that there is a differential variational inequality (DVI) that integrates the theory of spatial price equilibrium in a dynamic setting with the path delay operator notion used in the theory of dynamic user equilibrium as presented in Friesz and Han (2022). Our path delay operator is based on LWR theory and vetted in the published dynamic user equilibrium literature; see, in particular, the paper by Han et al. (2016). The aforementioned integration is original and allows truck borne freight to be modelled in way that recognizes for the interaction of freight and passenger vehicles by explicitly considering multiple classes of vehicles. As such, our model of DSPE to be presented at DTA2025 constitutes a significant addition to both the spatial price equilibrium and the freight network equilibrium modeling literatures.

**The Classical Spatial Price Equilibrium Problem**. We posit a freight network that transports several types of commodities between markets with positive excess supply and those with positive excess demand. The decision variables of the model correspond to the arabic letters $h$, $S$, $D$, and $I$. In particular, $h$ will refer to path flow, $S$ to rate of supply, $D$ to demand rate, and $I$ to inventory/backorder level. Furthermore, we will use $c$ to refer to transportation cost per unit of flow. Commodity prices will be $\pi$. The relevant subscripts/superscripts for variables, as is meaningful, are $p$ for a specific path, $k$ for a specific commodity, and $i$ or $j$ for a specific node. Other notation will be introduced as needed.

We will employ the following sets in discussing spatial price equilibrium and its extension from a static to a dynamic setting:

$\mathcal{N}$ = the set of nodes in the network of interest

$\mathcal{K}$ = the set of commodities

$\mathcal{W}$ = the set of origin-destination pairs

$\mathcal{P}$ = the set of all paths connecting the OD pairs

$\mathcal{P}_{ij}^{k}$ = the subset of paths connecting $(i,j) \in \mathcal{W}$ and suitable for commodity $k \in \mathcal{K}$

Let us consider a static spatial price equilibrium using the following vectors:

$$h = (h_p^k : p \in \mathcal{P}, k \in \mathcal{K})$$

$$S = (S_i^k : i \in \mathcal{N}, k \in \mathcal{K})$$

$$D = (D_i^k : i \in \mathcal{N}, k \in \mathcal{K})$$

$$\pi = (\pi_i^k : i \in \mathcal{N}, k \in \mathcal{K})$$

$$h(t) = (h_p^k : p \in \mathcal{P}, k \in \mathcal{K})$$

The essential characteristic of a spatial price equilibrium is that, if the shipping rate between a pair of supply and demand nodes is positive, the delivered price equals the local price. Moreover, if the delivered price exceeds local price, the shipping rate is zero. Static spatial price equilibrium may be stated in the following fashion:

$$h_p^k > 0, p \in \mathcal{P}_{ij}^k \Rightarrow \pi_i^k + c_p^k(h) = \pi_j^k \quad (1)$$

$$\pi_i^k + c_p^k(h) > \pi_j^k, p \in \mathcal{P}_{ij}^k \Rightarrow h_p^k = 0 \quad (2)$$

as first noted by Friesz et al. (1983).

It is well known and was first established by Friesz et al. (1983) that these conditions are satisfied by solutions of the following static, finite-dimensional variational inequality: find $(h^*, S^*, D^*) \in \Omega$ such that

$$c(t, h^*)^T(h - h^*) - \Theta(D^*)^T(D - D^*) + \Psi(S^*)^T(S - S^*) \geq 0 \quad \forall (h, S, D) \in \Omega \qquad (3)$$

where $\Theta$ and $\Psi$ are, respectively inverse commodity supply and demand vector functions. Of course, $\Omega$ is the set of feasible solutions as defined by market clearing (flow conservation) constraints and nonnegativity restrictions.

**The Dynamic Spatial Price Equilibrium Problem with Costless Inventorying**. Now we assume

$$h \in (L_+^2[t_0, t_1])^{|\mathcal{P}||\mathcal{K}|} \qquad \pi: [t_0, t_1] \in (L_+^2[t_0, t_1])^{|\mathcal{N}||\mathcal{K}|}$$

$$S \in (L_+^2[t_0, t_1])^{|\mathcal{N}||\mathcal{K}|} \qquad I: [t_0, t_1] \in (W^1[t_0, t_1])^{|\mathcal{N}||\mathcal{K}|}$$

$$D \in (L_+^2[t_0, t_1])^{|\mathcal{N}||\mathcal{K}|}$$

where $L_+^2[t_0, t_1]$ is the space of square integrable functions relative to the interval $[t_0, t_1]$ of the real line and $(L_+^2[t_0, t_1])^{|\mathcal{N}||\mathcal{K}|}$ is its $|\mathcal{N}||\mathcal{K}|$-fold product, while $W^1[t_0, t_1]$ is a Sobolev and space and $(W^1[t_0, t_1])^{|\mathcal{N}||\mathcal{K}|}$ is its $|\mathcal{N}||\mathcal{K}|$-fold product We take the variables $(h, S, D)$ to be controls; the inventory variables $I$ will be the state variables. We use $\tau_{ij}^k$ -to denote the travel time (delay) between origin-destination pair $(i, j) \in \mathcal{W}$ for commodity $k \in \mathcal{K}$.

The key DSPE conditions are stated in the following fashion:

$$h_p^k(t) > 0, p \in \mathcal{P}_{ij}^k \Rightarrow \pi_i^k(t) + c_p^k(t, h) = \pi_j^k\left(t + \tau_{ij}^k\right) \quad (4)$$

$$\pi_i^k(t) + c_p^k(t, h) > \pi_j^k\left(t + \tau_{ij}^k\right), p \in \mathcal{P}_{ij}^k \Rightarrow h_p^k(t) = 0 \quad (5)$$

Note also that

$$\pi_i^k(t) + c_p^k(t, h) \geq \pi_j^k\left(t + \tau_{ij}^k\right) \quad \forall (i, j) \in \mathcal{W}, k \in \mathcal{K}, p \in \mathcal{P}_{ij}^k \quad (6)$$

If (6) did not obtain, then, for some $(i, j) \in \mathcal{W}$, we would have

$$\pi_i^k(t) + c_p^k(t, h) < \pi_j^k\left(t + \tau_{ij}^k\right)$$

and need to consider two cases:

$(i)$ $h_p^k(t) > 0 \Rightarrow \pi_i^k(t) + c_p^k(t, h) = \pi_j^k\left(t + \tau_{ij}^k\right)$, which is a contradiction; or

$(ii)$ $h_p^k(t) = 0$, which is a failure to take advantage of an apparent spatial arbitrage (delivered price strictly less than local price). Since the presence of such arbitrage opportunities is inconsistent with equilibrium, we enforce (6). We also note that, in light of the nonnegativity of departure rates ($h \geq 0$), expression (5) is redundant since it is implied by (4).

For the special case of costless inventorying (holding and backordering), an intuitive extension of the VI for static SPE yields a mathematically rigorous DVI for DSPE. However, the variational analysis needed to establish the validity of the DVI is tedious. Nonetheless, it will be explained during our envisioned DTA2025 presentation using some simplifying "tricks" that make the proof accessible to those not familiar with differential game theory. The result (for costless inventory) is that the following differential variational inequality has solutions that are dynamic spatial price equilibria:

$$\int_{t_0}^{t_1} [c(t,h^*)^T (h-h^*) - \Theta(D^*)^T(D-D^*) + \Psi(S^*)^T(S-S^*)]dt \geq 0 \quad \forall (h,S,D) \in \Omega \quad (7)$$

where $\Omega$ is the set of feasible controls, defined as

$$\Omega = \{(h,S,D)\}: constraints\ (8)-(13)\ below\ hold\}$$

We will show during DTA2025 that the application of appropriate necessary conditions for differential variational inequality (7) yields the DSPE conditions, along with the other relevant considerations described above, and thereby establish that (7) is a correct DVI formulation.

The relevant constraints are given immediately below; in fact, their articulation relies on constants $A_i^k$ and $B_i^k$ the initial final and terminal inventory levels and the state variables $I_i^k(t)$

$$\frac{dI_i^k(t)}{dt} = S_i^k(t) + \sum_{j\in:(j,i)\in\mathcal{W}} \sum_{p\in\mathcal{P}_{ji}^k} h_p^k\left(t-\tau_{ji}^k\right) - \sum_{j:(i,j)\in\mathcal{W}} \sum_{p\in\mathcal{P}_{ij}^k} h_p^k(t) - D_i^k(t) \quad (8)$$

$$I_i^k(t_0) = A_i^k \quad (9)$$

$$I_i^k(t_1) = B_i^k \quad (10)$$

$$-h_p^k \leq 0 \quad (\mu_p^k) \quad \forall k\in\mathcal{K}, p\in\mathcal{P} \quad (11)$$

$$-S_i^k \leq 0 \quad (\alpha_i^k) \quad \forall k\in\mathcal{K}, i\in\mathcal{N} \quad (12)$$

$$-D_i^k \leq 0 \quad (\beta_i^k) \quad \forall k\in\mathcal{K}, i\in\mathcal{N}, \quad (13)$$

**Dynamic Spatial Price Equilibrium With Inventory Costs and State-Space Constraints**. When inventory costs are introduced the DVI possesses pure state-space constraints and its adjoint variables (dynamic dual variables) may possess jumps. Such problems have intrinsically difficult to analyze necessary conditions; indeed their necessary conditions may be intractable.

We will present our latest research on state-space constrained DPSE at DTA2025, including algorithms. Our research will continue up to the convening of the Symposium.

**References**


1. Beckmann, M., McGuire, C. B., & Winsten, C. B. (1956). *Studies in the Economics of Transportation*. Yale University Press, 232 pages.

2. Chao, G. S., & Friesz, T. L. (1984). Spatial price equilibrium sensitivity analysis. *Transportation research part B: methodological*, 18(6), 423-440.

3. Florian, M., & Los, M. (1982). A new look at static spatial price equilibrium models. *Regional science and urban economics*, 12(4), 579-597.

4. Friesz, T. L. (2010). *Dynamic optimization and differential games*. New York: Springer.

5. Friesz, T. L. (2024). Dynamic spatial price equilibrium, dynamic user equilibrium, and freight transportation in continuous time: A differential variational inequality perspective. *Transportation Research Part B: Methodological*, *190*, 103085.

6. Friesz, T. L., Tobin, R. L., Smith, T. E., & Harker, P. T. (1983). A nonlinear complementarity formulation and solution procedure for the general derived demand network equilibrium problem. *Journal of Regional Science*, 23(3), 337-359.

7. Friesz, T. L., & Han, K. (2022). *Dynamic network user equilibrium*, Springer, 363 pages (isbn-13: 978-3031255625).

8. Han, K., Piccoli, B., & Friesz, T. L. (2016). Continuity of the path delay operator for dynamic network loading with spillback. *Transportation Research Part B: Methodological*, *92*, 211-233.

9. Samuelson, P. A. (1952). Spatial price equilibrium and linear programming. *American economic review*, 42(3), 283-303.

10. Smith, T. E. (1984). A solution condition for complementarity problems: with an application to spatial price equilibrium. *Applied mathematics and computation*, 15(1), 61-69.

11. Smith, T. E., & Friesz, T. L. (1985). Spatial market equilibria with flow-dependent supply and demand: the single commodity case. *Regional science and urban economics*, 15(2), 181-218.

12. Takayama, T., & Judge, G. G. (1964). Equilibrium among spatially separated markets: a reformulation. *Econometrica: journal of the econometric society*, 510-524.

13. Tobin, R. L., & Friesz, T. L. (1983). Formulating and solving the spatial price equilibrium problem with transshipment in terms of arc variables. *Journal of regional science*, 23(2), 187-198.

# Incentive-Driven Utility-based Self-Rebalancing for Ride-Hailing Service

Jihu Kim[a], Jeongyun Kim[a], Hwasoo Yeo[a,b,]

[a]*Department of Civil and Environmental Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, South Korea*
[b]*Corresponding author: hwasoo@kaist.ac.kr*

## 1. Introduction

Mobility-on-Demand (MoD) systems, such as ride-hailing services, dynamically allocate vehicles in response to real-time user demand, enhancing user convenience and vehicle utilization. However, spatiotemporal demand-supply imbalances frequently occur, leading to vehicle shortages in high-demand areas and excess idle vehicles in low-demand regions. These inefficiencies increase service wait times, operational costs, and empty trips, emphasizing the need for effective vehicle rebalancing strategies([1][2][3]). While existing studies primarily focus on centralized control-based rebalancing, challenges such as demand uncertainty, cost efficiency, and real-time adaptability remain key concerns([4][5][6][7]).

This study considers a ride-hailing model where a central platform manages user requests and disseminates demand information. However, the platform does not directly control vehicles, as final movement decisions are made by drivers. In this structure, self-rebalancing—where drivers independently decide their repositioning—becomes a crucial factor in maintaining service efficiency([8]). Unlike theoretical centralized models, real-world ride-hailing platforms can only suggest repositioning strategies, making it essential to develop a realistic rebalancing framework that accounts for driver behavior.

To address this gap, this study explores self-rebalancing behavior using a utility-based model. Drivers make relocation decisions based on both platform-provided demand forecasts and personal experience, balancing factors such as expected service opportunities, relocation costs, and competition. Additionally, incentive mechanisms are introduced to influence driver choices, guiding repositioning movements toward system-wide efficiency. By integrating incentives into the self-rebalancing model, this study aims to enhance rebalancing efficiency while respecting driver autonomy.

## 2. Methodology

This study investigates self-rebalancing behavior in ride-hailing services, where drivers independently decide their repositioning based on perceived conditions. Unlike centrally

controlled rebalancing, self-rebalancing reflects decentralized decision-making influenced by multiple factors, including predicted demand, travel costs, competition, and incentives.

### *2.1. Utility-Based Driver Choice Model*

A utility function-based approach is employed to model self-rebalancing decisions. The function consists of four main components:

A utility function-based approach is employed to model self-rebalancing decisions. The function consists of four main components. The Attraction Term represents expected demand in each region, driving vehicles toward areas with potential service opportunities. The Cost Term reflects the travel cost from the origin to the destination, discouraging unnecessary long-distance relocations. The Competition Term accounts for vehicle concentration in a region and its surroundings, discouraging movements toward oversaturated areas. Lastly, the Incentive Term represents external incentives designed to encourage vehicle movement toward specific areas.

$$\begin{aligned} \mathrm{Attr}_{ij} &= \max\left(0, PredictedDemand_j - IdleVehicles_j\right) \\ \mathrm{Cost}_{ij} &= \mathrm{T}_{ij} \\ \mathrm{Comp}_{ij} &= IdleVehicles_j + \frac{ReachableVehicles_j}{n} \\ \mathrm{I}_{ij} &= f(Policy) \end{aligned}$$

where Policy refers to system-wide incentive strategies.

Each term is normalized to ensure balanced contributions to decision-making:

$$\begin{aligned} U_{ij} &= \alpha \cdot NormAttr_{ij} - \beta \cdot NormCost_{ij} - \gamma \cdot NormComp_{ij} + \delta \cdot NormI_{ij} \\ NormX_{ij} &= \frac{X_{ij} - X_{min}}{X_{max} - X_{min}} \end{aligned}$$

The probability of a driver relocating from region $i$ to region $j$ is determined using a logit model:

$$\mathrm{p}_{ij} = \frac{e^{U_{ij}}}{\sum e^{U_{ij}})}$$

### *2.2. Simulation Setup*

A simulation is conducted to analyze the impact of utility parameters on system performance under two demand scenarios:

- Scenario 1: High-demand setting with spatially imbalanced main flows (1,000 requests).
- Scenario 2: Low-demand setting with randomly distributed requests (250 requests).

A grid-based environment with 16 regions (2 km $\times$ 2 km) is used, where 200 vehicles operate. Travel times are predefined based on a constant speed assumption. Each simulation runs for two hours with a one-second timestep. This study consists of two experimental phases. In the first phase, driver-driven factors—attraction, cost, and competition terms—are analyzed by varying their parameter weights to examine their effects on service rate and idle time ratio. Based on the results, a representative parameter combination is selected. In the second phase, the selected parameter combination is used as a baseline for evaluating policy-driven factors, where the incentive term is introduced to examine its impact on vehicle repositioning. This two-phase experimental approach allows for a structured analysis, first examining how drivers naturally respond to different mobility conditions and then evaluating how external incentives can guide their repositioning decisions.

## 3. Driver-driven Factors Results

The simulation results show that service rate and idle time ratio are influenced by the weights of attraction, cost, and competition terms in the utility function, with varying impacts depending on demand conditions. This highlights the need for careful parameter adjustment to enhance system efficiency.

For service rate, the attraction term ($\alpha$) played a key role. In high-demand settings (Scenario 1), a lower $\alpha$ improved service rate by minimizing unnecessary repositioning, while in low-demand settings (Scenario 2), a higher $\alpha$ increased service rate by encouraging movement toward dispersed demand. The competition term ($\gamma$) also affected service rate, where moderate values ($\sim$0.5) improved efficiency, but excessive values discouraged movement to high-demand areas.

For idle time ratio, the cost term ($\beta$) significantly reduced idle time by discouraging long-distance repositioning. The $\alpha$ had contrasting effects: a lower $\alpha$ decreased idle time in Scenario 1 by keeping vehicles in demand-dense areas, while a higher $\alpha$ reduced idle time in Scenario 2 by facilitating proactive repositioning. The $\gamma$ helped distribute vehicles efficiently at moderate levels but increased idle time when overemphasized.

These findings suggest that adjusting the relative influence of attraction, cost, and competition based on demand patterns can enhance service performance while maintaining efficient vehicle distribution.

## 4. Framework for Policy-driven Factors Analysis

This phase presents the framework for analyzing policy-driven factors, which will be conducted in future experiments. Policy-driven factors refer to incentive terms that are determined based on the strategic objectives of the service system. Unlike driver-driven factors, which reflect individual decision-making based on perceived conditions, policy-driven factors introduce external interventions to influence vehicle movements and enhance overall service efficiency.

To conduct this experiment, the representative parameter combinations were selected based on the driver-driven factors analysis for each scenario. Scenario 1 involves large-scale demand with concentrated movement patterns. To achieve high service rates and reduced idle time, the selected parameter combination is $(\alpha, \beta, \gamma) = (0.25, 1.5, 0.5)$. Scenario 2 features smaller-scale demand with dispersed movement patterns. To encourage proactive movement toward demand locations while minimizing competition, the selected parameter combination is $(\alpha, \beta, \gamma) = (1.25, 1.0, 0.25)$. Using these fixed parameter settings, the incentive term $(\delta)$ will be incorporated into the utility function, and its effects will be examined under three conditions. The Baseline scenario excludes incentives, considering only attraction, cost, and competition terms. The Fixed Incentive scenario applies a constant incentive value to specific target regions. The Dynamic Incentive scenario adjusts incentives based on real-time demand-supply imbalances. These conditions will allow an assessment of how different incentive strategies influence vehicle repositioning and overall system performance.

## 5. Conclusion

This study examined the impact of self-rebalancing in ride-hailing services by analyzing driver decision-making through a utility-based approach. The simulation results demonstrated that the effects of attraction, cost, and competition terms on service performance vary depending on demand characteristics. In high-demand scenarios, limiting unnecessary repositioning improved service rates, while in low-demand environments, proactive movements helped balance supply. These findings highlight the importance of scenario-specific parameter adjustments for effective vehicle rebalancing.

Building on this framework, future work will extend the analysis to policy-driven factors by incorporating incentive mechanisms into the utility function. By evaluating fixed and dynamic incentive strategies, we aim to explore how external interventions can further improve service efficiency while maintaining driver autonomy. These insights will contribute to developing adaptive and data-driven rebalancing strategies that enhance overall mobility system performance.

## 6. Acknowledgements

This work was supported by the Institute of Information Communications Technology Planning Evaluation(IITP)-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government(MSIT)(IITP-2025-RS-2022-00156287)

## References

[1] M. Wittmann, L. Neuner, M. Lienkamp, A predictive fleet management strategy for on-demand mobility services: A case study in munich, Electronics 9 (2020) 1021.

[2] X. Guo, H. Xu, D. Zhuang, Y. Zheng, J. Zhao, Fairness-enhancing vehicle rebalancing in the ride-hailing system, arXiv preprint arXiv:2401.00093 (2023).

[3] C. Lu, M. Maciejewski, K. Nagel, Effective operation of demand-responsive transport (DRT): Implementation and evaluation of various rebalancing strategies, Technische Universität Berlin, 2023.

[4] S. Tavor, T. Raviv, Anticipatory rebalancing of robotaxi systems, Transportation Research Part C: Emerging Technologies 153 (2023) 104196.

[5] J. Yeo, S. Lee, K. Jang, J. Lee, Real-time operations of autonomous mobility-on-demand services with inter-and intra-zonal relocation, IEEE Transactions on Intelligent Vehicles (2023).

[6] R. Nair, E. Miller-Hooks, Fleet management for vehicle sharing operations, Transportation Science 45 (2011) 524–540.

[7] M. Repoux, M. Kaspi, B. Boyacı, N. Geroliminis, Dynamic prediction-based relocation policies in one-way station-based carsharing systems with complete journey reservations, Transportation Research Part B: Methodological 130 (2019) 82–104.

[8] T. Dong, Q. Luo, Z. Xu, Y. Yin, J. Wang, Strategic driver repositioning in ride-hailing networks with dual sourcing, Transportation Research Part C: Emerging Technologies 158 (2024) 104450.

# Efficient Gradient-Based Calibrations with A Differentiable Dynamic Traffic Assignment (DTA) Simulation Model by Iterative Backpropagation (IB) And Auto-Differentiation

Kejun Du[a], Qiru Ma[a], Enoch Lee[b] and Hong K. Lo[a]
[a] *Department of Civil and Environmental Engineering,*
*The Hong Kong University of Science and Technology, Hong Kong, China*
*Email: kduab@connect.ust.hk, qmaap@connect.ust.hk, cehklo@ust.hk*
[b] *Department of Logistics & Maritime Studies,*
*The Hong Kong Polytechnic University, Hong Kong, China*
*Email: enoch.lee@polyu.edu.hk*

## 1. INTRODUCTION

Dynamic Traffic Assignment (DTA) is a fundamental framework for modeling traffic flow dynamics and user equilibrium in transportation networks. Traditional DTA models often rely on heuristic perturbations or computationally expensive numerical approximations for optimization and calibration. This study introduces a differentiable DTA simulation model that enables gradient-based optimization by leveraging iterative backpropagation and auto-differentiation techniques. By formulating DTA simulations for user equilibrium (UE) conditions as a computational graph, we establish a robust framework for calibrating traffic models and optimizing route choice strategies efficiently. The UE conditions are approached with the method of successive averages (MSA) for passenger route flows (choices), and passengers' route choices are assumed to be based on approximated link travel times instead of recorded route travel times from previous iterations, to ensure the analytical differentiability of the simulated link flows throughout simulation iterations, whose partial derivatives with respect to arbitrary traffic parameters are analytically derived for illustration. With the iterative backpropagation (IB) algorithm and auto-differentiation, some preliminary experiments are conducted to validate the proposed framework of differentiable DTA simulation for UE in calibrating various traffic parameters. The proposed framework has high potential in simultaneous calibrations of high dimensional traffic parameters in large-scale transportation networks, as the analytical gradients enable gradient-based optimization algorithms to efficiently update parameters without redundant computations.

Simulation-based optimization (SBO) has been widely used in transportation studies to calibrate passenger route choice models and optimize network performance. Osorio & Bierlaire (2013) introduced a metamodel SBO approach that integrates traffic simulation information to enhance network efficiency and reliability, while Chong & Osorio (2018) extended this framework to dynamic problems with time-dependent decision variables. Meanwhile, computational graphs have also been adopted in transportation modeling, with Ma et al. (2020) formulating OD estimation problems as computational graphs, enabling scalable and efficient solutions through deep learning frameworks. Guarda & Qian (2024) further extended this approach to infer utility functions from traffic counts, speeds, and sociodemographic data, illustrating the versatility of computational graphs in transportation analytics. Additionally, some studies leveraged auto-differentiation and backpropagation techniques from machine learning for efficient optimizations. Du et al. (2025) introduced a fully differentiable end-to-end SBO framework with a differentiable metro system simulation model, using IB algorithm with auto-differentiation to efficiently calibrate passenger route choices in metro systems, showing capabilities of simultaneously optimizing for all passenger route choices between all OD pairs. This study is an extension of the fully differentiable SBO framework to DTA simulations for UE conditions on general transportation networks.

## 2. METHODOLOGY

### 2.1 Dynamic Traffic Assignment (DTA) Simulation for User Equilibrium (UE) Conditions as A Computational Graph

In this study, the DTA simulation model is designed for approaching user equilibrium (UE) conditions by leveraging the method of successive averages (MSA) for iteratively updating the passenger stochastic route choices (SRC) based on route travel times from the previous iteration of simulation. To formulate the iterative DTA simulation model as a computational graph, all calculations must be fully differentiable with respect to the concerned parameters such as passenger route choice model parameters. Therefore, to ensure the differentiability, instead of recorded route travel times from the previous iteration, the updated route choices are based on the approximated link travel times from the simulated link flows, as discussed further in the following sections.

The proposed iterative DTA simulation for UE is illustrated in Figure 1, designed in a computational graph structure. Firstly, from any randomly initialized feasible route choices (flows), the DTA simulation model is run to generate the time-dependent link flows and densities. Then, the link costs (travel times) are approximated with the link cost functions (such as BPR functions) based on time-dependent link flows or densities. Then, the auxiliary stochastic route choices of all passengers are generated as the Time Dependent Shortest Paths (TDSP) based on the approximated link costs. With the auxiliary route choices, the MSA interpolation is conducted to update the

passenger route choices towards the auxiliary route choices with a given step size $\lambda$. Finally, check convergence, if not, repeat the first step to run the DTA simulation with current route choices. Since every step of the iterative DTA simulation framework is differentiable, the partial derivatives of the converged link flows with respect to any analytically tractable traffic parameters can be derived by the Iterative Backpropagation (IB) algorithm with auto-differentiation.

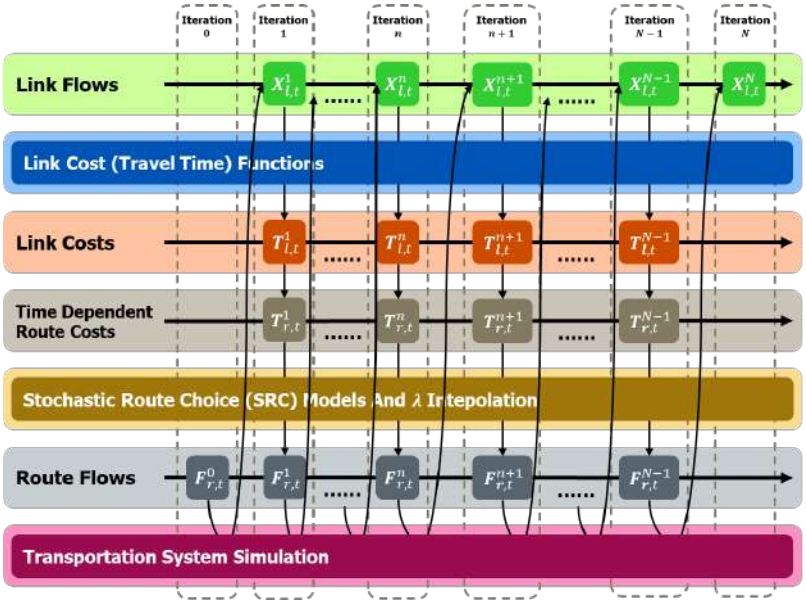



Figure 1. The Iterative DTA Simulation for UE Conditions as A Computational Graph.

## 2.2 Differentiable Flow-Based Mesoscopic DTA Simulation

To conduct the DTA simulation with given route choices and OD demands, this study proposes a differentiable flow-based mesoscopic DTA simulation model, where all processes are designed to be analytically tractable. In the simulation model, passengers or vehicles are simulated as continuous flow units instead of discreate agents. Each link is separated into two queues: the running queue and exiting queue, as shown in Figure 2. In the running queue, for each simulation time step, the flow units are moved forward by a distance calculated from the current link speed based on link density and free flow speed. Flow units which finish the link in the running queue would be sequentially moved to the exiting queue, waiting to be discharged when the next link has enough residual capacities. Signals are designed for each intersection, and various phase designs and timing plans are customizable. The sample pseudo code is provided in Algorithm 1.

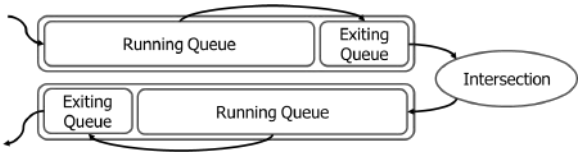



Figure 2. The Link-Intersection Structure in The Flow-Based Mesoscopic DTA Simulation Model.

**Algorithm 1. Differentiable Flow-Based Mesoscopic DTA Simulation**

$t = 0$
**FOR** each origin node $o$:
    Generate new flow units to each destination $d$
    Split the flow units to different routes with a route choice model
    Move flow units to origin queues for different outward links accordingly
    **IF** the residual capacity $c_l^t$ is still enough:
        Move each flow units to the running queue of the outward link $l$
**FOR** each link $l$:
    Calculate link speed $v_l^t$ by link density $\rho_l^t$: $v_l^t = v_l^{free} \cdot \left(1 - \frac{\rho_l^t}{\rho_l^{max}}\right)$
    Move all flow units in the running queue forward accordingly: $x_{n,l}^t = x_{n,l}^{t-1} + v_l^t \cdot \Delta t$
    **FOR** each flow unit in the running queue:
        **IF** it reaches the end of this link, move those do into the exiting queue
    **FOR** all flow units in the exiting queue:
        **IF** this flow unit arrives at the destination, remove it from the exiting queue

**FOR** each intersection $I$:
Get the current allowed movements between links
**FOR** each outward link $l$ of this intersection $I$:
Calculate the residual capacity $r_l^t$ of this out-link
**FOR** each inward link $k$ among the allowed movements to this outward link $l$:
Sum up flow unit volumes in the exiting queue **IF** next link as $l$
**IF** summed flow volumes are smaller than $r_l^t$, move these flow units from all inward links to the running queue of $l$
**ELSE** equally distribute the capacity to all inward links, and move the first flow units in each inward link up to the divided capacity to the running queue of $l$
$t = t + 1$, go back to the second step

### 2.3 Traffic Parameter Calibration with DTA Model by Iterative Backpropagation

This section discusses the scenario of calibrating traffic models or passenger route choice model parameters with the proposed differentiable DTA simulation for UE as a computational graph. Here the fundamental diagram relationship between speed and density is adopted for link travel time approximations as Equation 1.

$$v = v_f \cdot \left(1 - \frac{\rho}{\rho_{max}}\right), \ t = \frac{l}{v} = \frac{l \cdot \rho_{max}}{v_f \cdot (\rho_{max} - \rho)} \tag{1}$$

Then, to enable the gradient-based optimization algorithm, analytical partial derivatives of the converged link flows for $\boldsymbol{x_{l,t,N}}$ link $\boldsymbol{l}$ at time interval $\boldsymbol{t}$ after $\boldsymbol{N}$ iterations are derived with respect to arbitrary parameters $\boldsymbol{\theta}$, through the iterative simulations. Firstly, the MSA update of passenger route flows $\boldsymbol{f_{d,t,n}^{r}}$ of OD pair $\boldsymbol{d}$ starting from time interval $\boldsymbol{\tau}$ in iteration $\boldsymbol{n}$ is shown in Equation 2 with auxiliary route flows $\boldsymbol{f_{d,t,n}^{r,*}}$ and predetermined step size $\boldsymbol{\lambda_n}$. Therefore, the converged route choices can be decomposed as Equation 3.

$$f_{d,\tau,n}^{r} = f_{d,\tau,n-1}^{r} + \lambda_n \cdot \left(f_{d,\tau,n}^{r,*} - f_{d,\tau,n-1}^{r}\right) \tag{2}$$

$$f_{d,\tau,N}^{r} = f_{d,\tau,0}^{r} + \sum_{n=1}^{N} \Delta f_{d,\tau,n}^{r} = f_{d,\tau,0}^{r} + \sum_{n=1}^{N} \left(\lambda_n \cdot \left(f_{d,t,n}^{r,*} - f_{d,\tau,n-1}^{r}\right)\right) \tag{3}$$

Meanwhile, as the initial route flows can be chosen arbitrarily, which are not relevant to the any parameters for calibration, the partial derivatives of initial route flows with respect to the parameters are zero: $\frac{\partial f_{d,\tau,0}^{r}}{\partial \theta} = 0$. Then, the partial derivatives of the path flow with respect to the parameters can be recursively derived as Equation 4.

$$\begin{aligned} \frac{\partial f_{d,\tau,N}^{r}}{\partial \theta} &= 0 + \sum_{n=1}^{N} \frac{\partial \Delta f_{d,\tau,n}^{r}}{\partial \theta} = \sum_{n=1}^{N} \left(\lambda_n \cdot \left(\frac{\partial f_{d,\tau,n}^{r,*}}{\partial \theta} - \frac{\partial f_{d,\tau,n-1}^{r}}{\partial \theta}\right)\right) = \lambda_N \cdot \frac{\partial f_{d,\tau,N}^{r,*}}{\partial \theta} - \sum_{n=1}^{N-1} \left(\frac{\partial f_{d,\tau,n}^{r,*}}{\partial \theta} \cdot \prod_{m=n}^{N} \lambda_m\right) \\ &= 2 \cdot \lambda_N \cdot \frac{\partial f_{d,\tau,N}^{r,*}}{\partial \theta} - \sum_{n=1}^{N} \left(\frac{\partial f_{d,\tau,n}^{r,*}}{\partial \theta} \cdot \prod_{m=n}^{N} \lambda_m\right) \end{aligned} \tag{4}$$

Furthermore, after $N$ iterations, we denote $\delta_{d,\tau,l,t,N}^{r} \in \{0,1\}$ as the indicator of whether the route flow unit $f_{d,\tau,n}^{r}$ is located at link $l$ in time interval $t$. Therefore, the converged link flow $x_{l,t,N}$ can be expressed as Equation 5. The iterative backpropagation (IB) algorithm with auto-differentiation can deal with if-else conditions with binary indicators, as long as the forward process (the simulation) has been conducted and the indicators are already determined.

$$x_{l,t,N} = \sum_{\tau} \sum_{d} \sum_{r \in R_d} f_{d,\tau,N}^{r} \delta_{d,\tau,l,t,N}^{r} \tag{5}$$

Finally, the partial derivatives of the converged link flows are expressed as Equation 6. As the updated route choices are made based on approximated travel times from previous iterations, which are analytically tractable iteratively back to the link cost function parameters with link densities, $\frac{\partial f_{d,\tau,n}^{r,*}}{\partial \theta}$ will be obtained by auto-differentiation for parameters of both link cost functions and route choice models.

$$\begin{aligned} \frac{\partial x_{l,t,N}}{\partial \theta} &= \sum_{\tau} \sum_{d} \sum_{r \in R_d} \left(\delta_{d,\tau,l,t,N}^{r} \cdot \frac{\partial f_{d,\tau,N}^{r}}{\partial \theta}\right) \\ &= \sum_{\tau} \sum_{d} \sum_{r \in R_d} \left(\delta_{d,\tau,l,t,N}^{r} \cdot \left(2 \cdot \lambda_N \cdot \frac{\partial f_{d,\tau,N}^{r,*}}{\partial \theta} - \sum_{n=1}^{N} \left(\frac{\partial f_{d,\tau,n}^{r,*}}{\partial \theta} \cdot \prod_{m=n}^{N} \lambda_m\right)\right)\right) \end{aligned} \tag{6}$$

In practice, the parameters for calibration will be initialized randomly. With the given parameters, for each optimization iteration, the iterative DTA simulation for UE will be conducted until convergence, and the partial derivatives of the differences between simulated link flows and actual observations will be calculated by auto-differentiation, for the gradient-based optimization algorithm to update the parameters for the next iteration of

simulation, until the parameters converge.

## 3. PRELIMINARY EXPERIMENTS

To illustrate the capability of the proposed DTA simulation model in calibrating traffic parameters by minimizing the differences between simulation results and actual observations, preliminary experiments are conducted with the toy network in Figure 3. In the toy network, there are four nodes, five links and three OD pairs: AD, AC and BD. The simulation time is 1 hour, and link flows are recorded in time intervals of 10 minutes. For all links, the free flow speed $v_f$ is 15 $\left[\frac{m}{s}\right]$, the maximum link density $\rho_{max}$ is 1200 $\left[\frac{veh}{km}\right]$ and the passenger value of time $VOT$ is 1000 $\left[\frac{HKD}{hr}\right]$. Simulations are run with these parameters to generate synthetic flows for validation optimizations.

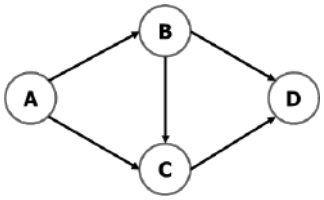



Figure 3. A Toy Network.

Then, with the synthetic link flows as actual observations, starting from randomly initialized traffic parameters, the proposed DTA simulation model is run iteratively with the Adam optimizer minimizing the mean squared errors (MSE) between the simulated link flows and actual observations based on the gradients from auto-differentiation. Three calibration scenarios are practiced for illustration: calibrating for $\rho_{max}$ only, $VOT$ only, $\rho_{max}$ and $VOT$ together. The optimization processes are shown in Figures 4(a), 4(b) and 4(c) respectively, where the MSEs are all stably decreasing, implying that the proposed framework with differentiable DTA simulation is potentially effective in calibrating traffic model parameters.

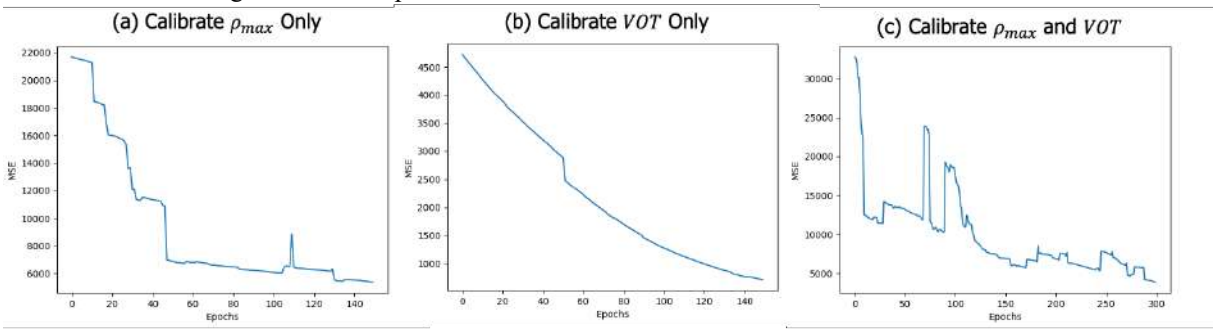



Figure 4. Optimization Processes of Calibrating Traffic Parameters with The Proposed DTA Simulation Model. (a) $\rho_{max}$ only. (b) $VOT$ only. (c) $\rho_{max}$ and $VOT$ together.

## 4. CONCLUSIONS

This study proposes a differentiable mesoscopic DTA simulation framework that enables efficient gradient-based optimization through iterative backpropagation. User equilibrium conditions are approached by iterative DTA simulations with MSA of stochastic route choices (SRC). By approximating the route travel times by link travel times derived from the fundamental diagrams, the proposed DTA simulation model is fully differentiable and formulated as a computational graph. The gradients of simulated link flows with respect to the predetermined traffic parameters are preserved throughout simulation iterations. Preliminary experiments validate the capability of the proposed framework in calibrating traffic parameters with auto-differentiation and gradient-based optimization algorithms. Future research will refine the DTA simulation model, introduce various link travel time approximations and extend the proposed framework to large-scale networks for practical purposes.

## 5. REFERENCES


Chong, L., & Osorio, C. (2018). A simulation-based optimization algorithm for dynamic large-scale urban transportation problems. *Transportation Science*, *52*(3), 637–656. https://doi.org/10.1287/trsc.2016.0717

Du, K., Lee, E., Ma, Q., Su, Z., Zhang, S., & Lo, H. K. (2025). Modeling Metro Passenger Routing Choices with a Fully Differentiable End-to-End Simulation-Based Optimization (SBO) Approach. *Transportation Science*. https://doi.org/10.1287/trsc.2024.0557

Guarda, P., & Qian, S. (2024). Statistical inference of travelers' route choice preferences with system-level data. *Transportation Research Part B: Methodological*, *179*. https://doi.org/10.1016/j.trb.2023.102853

Ma, W., Pi, X., & Qian, S. (2020). Estimating multi-class dynamic origin-destination demand through a forward-backward algorithm on computational graphs. *Transportation Research Part C: Emerging Technologies*, *119*. https://doi.org/10.1016/j.trc.2020.102747

Osorio, C., & Bierlaire, M. (2013). A simulation-based optimization framework for urban transportation problems. *Operations Research*, *61*(6), 1333–1345. https://doi.org/10.1287/opre.2013.1226

# Quantifying Traffic State Vulnerability at Highway Bottleneck Networks Using Lyapunov Exponents

Hye-young Tak[a] and Hwasoo Yeo[a,*]

[a] Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
hytak@kaist.ac.kr, hwasoo@kaist.ac.kr
* Corresponding author

*Extended abstract submitted for presentation at the 10th International Symposium on Dynamic Traffic Assignment (DTA 2025)*
*Sep 16-18, 2025, Salerno, Italy*

April 11, 2025

Keywords: Traffic State Vulnerability, Lyapunov Exponent, State-Space Reconstruction, Traffic Stability, Highway Bottleneck

## 1 INTRODUCTION

Highway bottlenecks exhibit complex and dynamic traffic conditions, where interactions between upstream and downstream flows often lead to significant disruptions and traffic congestion. These disruptions result from local perturbations, such as speed fluctuations and temporary slowdowns, which can develop into unstable traffic states and propagate through the network. Understanding traffic state vulnerability—the susceptibility of traffic flow to such perturbations—is critical for improving network resilience and mitigating congestion. Previous studies have explored traffic state transitions at bottlenecks, particularly the impact of on-ramp merging and lane-changing activity on flow stability (Yeo, 2008, Ahn *et al.*, 2010, Oh & Yeo, 2012). However, a robust method for quantifying traffic vulnerability based on system stability remains an open challenge.

Recent studies have approached traffic vulnerability assessment from different perspectives. Complex network theory and cloud models have been used to evaluate congestion-induced vulnerabilities, identifying critical road segments based on uncertainty in traffic states (Deng *et al.*, 2020). Additionally, vulnerabilities in connected transportation networks have been examined, particularly focusing on the cascading effects of cybersecurity threats and disruptions in intelligent transportation systems (ITS) (Hou *et al.*, 2019). Graph-theoretic approaches have also been applied to assess network capacity in emergency evacuation scenarios, identifying locations that restrict flow under extreme conditions (Ohi & Kim, 2021). While these studies provide valuable insights into network-level vulnerabilities, they primarily focus on static congestion metrics, network topology, and external risk factors, rather than capturing the dynamical evolution of traffic instabilities over time.

Traffic state vulnerability is inherently a dynamical problem, as speed fluctuations and density variations can propagate through the network, amplifying congestion beyond localized bottlenecks. Existing approaches often overlook the temporal evolution of traffic instability, limiting their ability to predict when and how disruptions escalate into system-wide congestion. To address this gap, this study employs the largest Lyapunov exponent (LLE) to quantify short-term and long-term vulnerability by assessing the sensitivity of traffic flow to small disturbances over time. Unlike conventional network-based assessments, this approach evaluates whether a traffic state stabilizes or remains unstable, providing deeper insights into traffic resilience at highway bottlenecks.

This study applies LLE-based analysis to a highway bottleneck segment, reconstructing the state space to analyze traffic flow stability under different perturbation conditions. The analysis offers a systematic method to quantify vulnerability, helping identify critical traffic conditions that are more prone to sustained instability. The paper is organized as follows: the methodology outlines state-space reconstruction and Lyapunov exponent estimation, the results provide a comparative statistical analysis of traffic vulnerability, and the conclusion discusses implications for network resilience and future research directions.

## 2 METHODOLOGY

The Lyapunov exponent quantifies the exponential divergence of initially close state-space trajectories, providing a measure of chaos within a system (Rosenstein *et al.*, 1993). The use of LLE to assess local dynamical stability has been applied to transportation systems (Lan *et al.*, 2008, Liu & Zhang, 2016).

Suppose traffic data is collected from loop detectors $N$ on the highway at timestamp $t$ as $\mathbf{x}_t \in \mathbb{R}^N$, with the complete dataset spanning $T$ timestamps represented as $\mathbf{X}_{1:T} = [\mathbf{x}_1, \cdots, \mathbf{x}_T] \in \mathbb{R}^{N \times T}$. In this study, we define the finite-time traffic state $\mathbf{x}_t$ as the lane-aggregated traffic speed, as it directly reflects the degree of congestion and disruptions. To reconstruct an appropriate state space from this single time series, we use the original data along with its time-delayed copies, as shown below (see Takens (2006) for more details):

$$Y(i) = [\mathbf{x}_t, \mathbf{x}_{t+\tau}, \ldots, \mathbf{x}_{t+(d_e-1)\tau}] \tag{1}$$

where $Y(i)$ is the a point in the reconstructed phase space from the delay time $\tau$, and the embedding dimension $d_e$. The number of points in $d_e$-dimensional subspace is $i = T - (d_e - 1)\tau$. To compute the Lyapunov exponent, the nearest neighbor points in the reconstructed phase space are identified for each trajectory at the initial condition:

$$\begin{aligned} j &= \arg\min_{j \neq i} \|Y(i) - Y(j)\|, \\ L_i &= \log\left(\frac{\|Y(i+n) - Y(j+n)\|}{\|Y(i) - Y(j)\|}\right) \end{aligned} \tag{2}$$

Here, $j$ represents the index of the nearest neighbor to $Y(i)$, minimizing the Euclidean distance between them while excluding $Y(i)$ itself. Each $L_i$ measures the logarithmic divergence between the trajectories beginning at $Y(i)$ and its nearest neighbor $Y(j)$ after $n$ time steps. The LLE can then be estimated by analyzing the average divergence rate of these closest neighbors as the slope of the divergence curve:

$$\lambda_i = \frac{1}{\tau}\left(\frac{1}{N}\sum_{i=1}^{N} L_i\right) \tag{3}$$

where $\lambda$ quantifies the average exponential separation rate between trajectories of nearest neighbors in phase space, capturing the system's sensitivity to initial conditions. If the $\lambda > 0$, the system is unstable and exhibits chaotic behavior, with its magnitude reflecting sensitivity to initial conditions. If the $\lambda < 0$, the system converges to a stable fixed point.

## 3 PRELIMINARY RESULTS

The proposed methodology is applied to a 2.2 km stretch of SR-91 in California, a freeway segment prone to congestion during morning rush hours. The dataset, collected from PeMS, includes speed and flow measurements at 5-minute intervals between 6:00 AM and 11:00 PM from

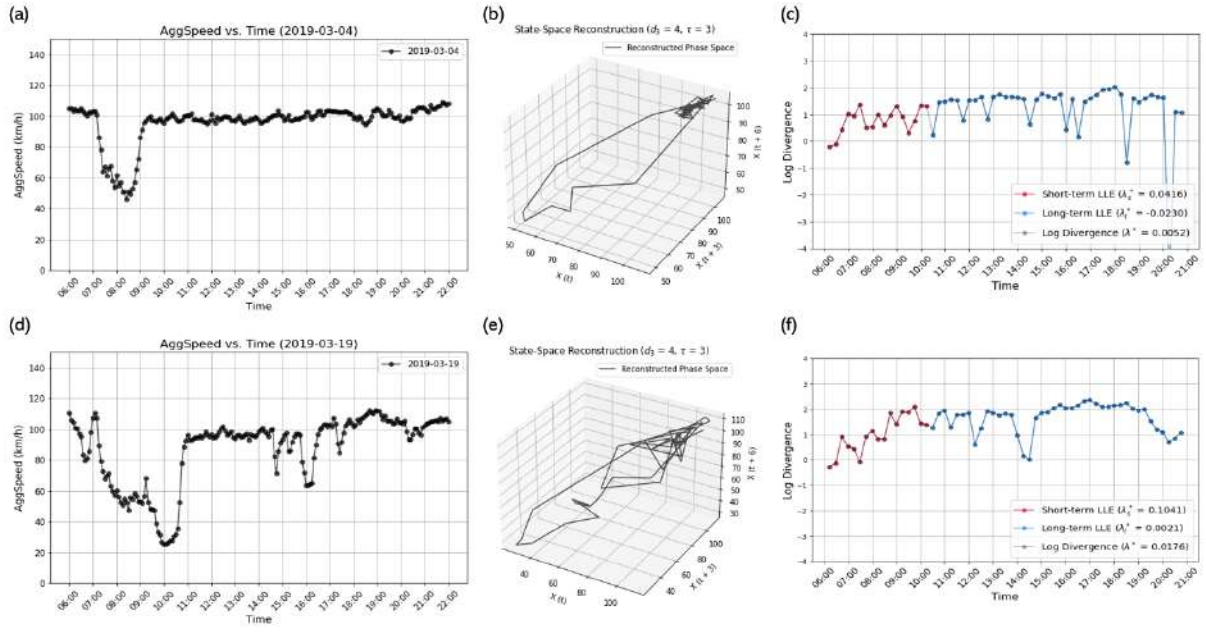



Figure 1 – *(a), (d) Aggregated speed over time on March 4, and March 19, 2019, respectively, illustrating temporal variations in traffic speed; (b),(e) State space reconstruction for different speed ($\tau$=3, $d_e$=4); and (c), (f) Short-term ($\lambda_s^*$), long-term ($\lambda_l^*$)), and overall LLEs over time, providing a quantitative assessment of local traffic stability.*

March 1 to September 30, 2019. The analysis focuses on identifying traffic state vulnerability using the Lyapunov exponent-based stability framework.

Figure 1 illustrates the speed fluctuations and the impact of initial perturbations (traffic breakdowns) on the upstream detector segment, analyzed using the proposed Lyapunov exponent-based stability framework. On March 4, 2019, significant speed fluctuations were observed between 6:00 AM and 9:00 AM, indicating instability. The state-space reconstruction (Figure 1(b)) reflects these perturbations, showing a dynamically evolving phase-space trajectory. The short-term LLE ($\lambda_s^*$) measures immediate sensitivity to disturbances, while the long-term LLE ($\lambda_l^*$) provides insight into whether the system stabilizes over time. On March 4, 2019, the short-term LLE was 0.0416, indicating initial chaotic behavior. However, as time progressed, the long-term LLE reached -0.0230, signifying that the system eventually regained stability.

On March 19, 2019, however, speed fluctuations persisted throughout the day, with pronounced afternoon perturbations (Figure 1d). The short-term LLE ($\lambda_s^*$) was 0.1041, indicating greater sensitivity to disturbances and the long-term LLE ($\lambda_l^*$) remained close to 0.0021, suggesting that the system did not fully stabilize. These findings highlight the variability in traffic state transitions, where some disturbances dissipate over time, while others sustain prolonged instability, impacting network performance.

# 4 DISCUSSION

This study evaluates traffic state vulnerability at a highway bottleneck using Lyapunov exponent-based stability analysis, focusing on short-term and long-term impacts of speed fluctuations. The findings suggest that traffic instability depends on the severity and persistence of perturbations, with higher short-term LLE values indicating greater immediate instability. While some disruptions naturally stabilize over time, others remain prolonged, emphasizing the need for a quantitative framework to assess vulnerability at a macroscopic level.

Future research could expand the Lyapunov exponent framework by expanding the analysis to network-wide traffic flow instability, capturing how local disruptions propagate across the system. Additionally, incorporating merging activity analysis and integrating microscopic perspectives will improve sensitivity to local perturbations and strengthen traffic vulnerability assessments. By refining these methodologies, this framework can contribute to more effective traffic management strategies, supporting intelligent transportation systems and improving network resilience against recurrent congestion.

# 5 ACKNOWLEDGEMENTS

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government(MSIT) (IITP-2025-RS-2022-00156287).

# References

Ahn, Soyoung, Laval, Jorge, & Cassidy, Michael J. 2010. Effects of merging and diverging on freeway traffic oscillations: theory and observation. *Transportation research record*, **2188**(1), 1–8.

Deng, Zhenping, Huang, Darong, Liu, Jinyu, Mi, Bo, & Liu, Yang. 2020. An assessment method for traffic state vulnerability based on a cloud model for urban road network traffic systems. *IEEE Transactions on Intelligent Transportation Systems*, **22**(11), 7155–7168.

Hou, Guangyang, Chen, Suren, & Han, Yan. 2019. Traffic performance assessment methodology of degraded roadway links following hazards. *Journal of Aerospace Engineering*, **32**(5), 04019055.

Lan, Lawrence W, Sheu, Jiuh-Biing, & Huang, Yi-San. 2008. Investigation of temporal freeway traffic patterns in reconstructed state spaces. *Transportation Research Part C: Emerging Technologies*, **16**(1), 116–136.

Liu, Yan, & Zhang, Jiazhong. 2016. Predicting traffic flow in local area networks by the largest Lyapunov exponent. *Entropy*, **18**(1), 32.

Oh, Simon, & Yeo, Hwasoo. 2012. Estimation of capacity drop in highway merging sections. *Transportation research record*, **2286**(1), 111–121.

Ohi, Sabrena Jahan, & Kim, Amy M. 2021. Identifying critical corridors during an area-wide disruption by evaluating network bottleneck capacity. *International Journal of Disaster Risk Reduction*, **64**, 102487.

Rosenstein, Michael T, Collins, James J, & De Luca, Carlo J. 1993. A practical method for calculating largest Lyapunov exponents from small data sets. *Physica D: Nonlinear Phenomena*, **65**(1-2), 117–134.

Takens, Floris. 2006. Detecting strange attractors in turbulence. *Pages 366–381 of: Dynamical Systems and Turbulence, Warwick 1980: proceedings of a symposium held at the University of Warwick 1979/80*. Springer.

Yeo, Hwasoo. 2008. *Asymmetric Microscopic Driving Behavior Theory*. Working Paper qt1tn1m968. University of California Transportation Center.

# A Link-Time-Based Method for Semi-Discretized Dynamic Traffic Assignment

Jiang Qian YING
Faculty of Regional Studies, Gifu University
1-1 Yanagido, Gifu 501-1193, Japan
Email: ying.jian.qian.i6@f.gifu-u.ac.jp

This research addresses the dynamic traffic assignment problem (DTA) within the context of user equilibrium. In most existing DTA models, time-varying flows - either path flows or link flows - are fundamental variables iteratively adjusted through a reassignment procedure in the solution algorithms. Link travel time costs, including link cruising time and queuing delays, are computed through a dynamic network loading (DNL) process based on these flows. Although link and path travel times are critical for adjusting flows, they are subordinate quantities determined from flow data by the physical mechanisms of the loading model. Using the current flow data and these subordinately determined travel times, flows are adjusted with appropriate reassignment algorithms. Exceptions to the above typical formulation can be found in the works of Ran and Boyce (1996) and Long et al. (2016). Ran and Boyce developed a link-time-based algorithm, where link times serve as fundamental variables iteratively adjusted. Long et al. use path travel times as the fundamental variables.

Another common feature often shared by DTA algorithms is the discretization of the time axis, which serves as the basis for traffic propagation, into small intervals. While some DTA models are inherently designed with discrete time intervals, others are initially formulated in a continuous-time framework. Although the continuous-time formulation provides a theoretically rigorous basis, time discretization becomes essential for numerical computation when applied to complex real-world networks. In this process, both traffic flows and travel times are defined over discrete time intervals. With cellular transmission model (CTM) being an exception, most dynamic flow models assume continuous travel times consistent with the underlying continuous traffic propagation physics. The dependency of continuous traffic propagation and travel times on discrete time intervals complicates the existing DAT algorithms in these models.

The aforementioned works by Ran and Boyce (1996) and Long et al. (2016) are also grounded in this framework of time-discretization. In such a framework, the link or path time-based formulation offers clarity in model formulation but seems to have limited algorithmic advantages compared to flow-based formulations.

Another noticeable exception can be found in Kuwahara and Akamatsu (1993) and Akamatsu (2001), where DTA for networks with special structures is addressed. For the case with a single origin, the traffic propagation spreading to the remaining nodes is depicted on the

departure time axis at the unique origin. Link flows and earliest node arrival times are sequentially calculated by incrementing the departure time steps from an initial state, without the need to discretize the travel time axis. Analogous formulation applies to a network with a single destination, but not easily to general networks.

In this paper, we propose a link-time-based method in a semi-discretized framework for dynamic traffic assignment. The method adopts link travel times as the fundamental variables to be directly adjusted, with flows treated as subordinate variables. This method does not need path enumeration. Moreover, the method is *semi-discretized* in the sense that only the arrival time horizon is discretized, while the time axis for traffic propagation remains continuous variables. These features together enhance greatly the ease of implementation of the proposed method.

Specifically, we address the problem of simultaneous route and arrival time choice (SRAT) during the morning peak traffic period. The time horizon for traffic arrivals is divided into discrete intervals, with one interval designated as the most desirable. Penalties for early and late arrivals are pre-specified for all other intervals. Travelers choose routes and arrival time intervals based on the sum of travel time costs and penalties, following a nested *logit type random choice model*. Departure time is then determined by the target arrival time and the average route travel time. As a result, our formulation is equivalent to a simultaneous route and departure time choice (SRDT) model.

**Arrival time choice**

Consider a network consisting of a set of links $\mathcal{A} = \{a, b, \dots\}$ and a set of nodes $\mathcal{N} = \{i, j, \dots\}$. Let $\{Q_{rs}\}$ denote the set of OD demands. (The ranges for indices, when easily identified from the context, will be omitted.) All travelers *arrive* at their destinations during a time horizon discretized into $M$ intervals: $[T_{\tau-1}, T_\tau]$, $1 \le \tau \le M$. In the following $\tau$ also denotes the interval $[T_{\tau-1}, T_\tau]$. The travelers desire to arrive at their destination at a common time interval $\tau_0$, incur a penalty for early or late arrival denoted by a function $EL(\tau)$, where $\tau$ is the actual arrival time interval.

Let $S_{rs}^\tau$ be the *expected* minimum travel time cost from $r$ arriving at $s$ at time $\tau$. The travelers choose their arrival time $\tau$ by the aggregated cost $S_{rs}^\tau + EL(\tau)$, obeying a *logit* model. Let $q_{rs}^\tau$ denote the travel demand from $r$ arriving at $s$ at time $\tau$. The vector $\boldsymbol{q} = \{q_{rs}^\tau\}$ is thus a function of $\boldsymbol{S} = \{S_{rs}^\tau\}$.

**Link-time-based model**

On a link $a$, travelers with a common destination $s$ and a common scheduled arrival time $\tau$, wherever they come from, are assumed to enter and leave the link at a same time and incur a

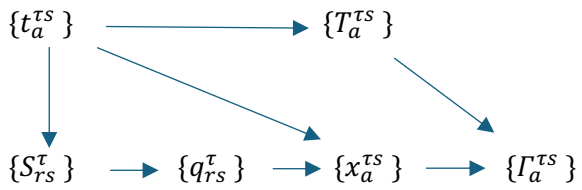
common link travel time cost $t_a^{\tau s}$. Let $x_a^{\tau s}$ denote the volume of these travelers, which will also referred to as a platoon in the following.

Now suppose that $\boldsymbol{t} = \{t_a^{\tau s}\}$ is given. Assume that travelers determine their routes according to a logit type random choice model. The vector of link flows $\boldsymbol{x} = \{x_a^{\tau s}\}$ can then be written as a function $\boldsymbol{x}\left(\boldsymbol{t}, \boldsymbol{q}\left(\boldsymbol{S}(\boldsymbol{t})\right)\right)$ of $\boldsymbol{t}$.

The time $T_a^{\tau s}$ for the platoon $x_a^{\tau s}$ to enter the link $a$ can be computed from $\boldsymbol{t}$. The vector $\{x_a^{\tau' s'}, T_a^{\tau' s'}\}_{\tau' s'}$, comprising information of the magnitudes and entry times of all the platoons that traverse a link $a$, characterizes the temporal profile of traffics on the link. We assume that this vector determines the traversal time of each platoon on the link, denoted by the function

$$\Gamma_a^{\tau s}\{x_a^{\tau' s'}, T_a^{\tau' s'}\}_{\tau' s'}$$

Combining the above equations, we have the following diagram depicting the relationship between link travel times and the other variables




Thus, the above diagram defines a map from the cube $[0, L]^{M\times|\mathcal{A}|\times|\mathcal{N}|}$ into itself, which can be written as

$$\boldsymbol{t} \to \boldsymbol{\Gamma}\left(\boldsymbol{x}\left(\boldsymbol{t}, \boldsymbol{q}\left(\boldsymbol{S}(\boldsymbol{t})\right)\right), \boldsymbol{T}(\boldsymbol{t})\right)$$

A fixed point of the map characterizes an equilibrium for the simultaneous arrival time and route choice dynamic traffic assignment problem. It is easy to verify that there exists an equilibrium if the link travel cost functions $\Gamma_a^{\tau s}\{x_a^{\tau' s'}, T_a^{\tau' s'}\}_{\tau' s'}$ are continuous in $x_a^{\tau' s'}$ and $T_a^{\tau' s'}$. A heuristic algorithm of the form, with step size $\alpha$ which may depend on the times of iterations,

$$t_a^{\tau s} \to (1-\alpha) t_a^{\tau s} + \alpha \Gamma_a^{\tau s}$$

can be adopted to compute the equilibrium.

**Link Travel Time Functions**

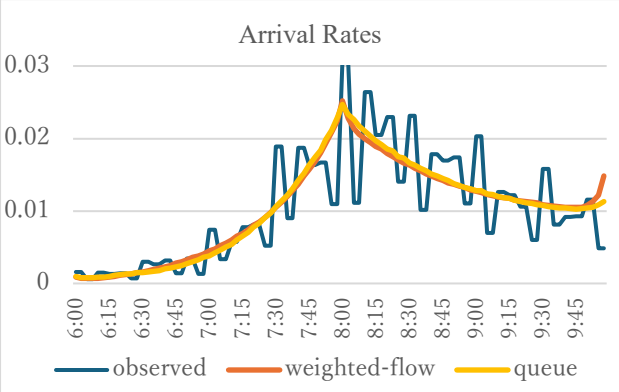
Several models for the function $\Gamma_a^{\tau s}\{x_a^{\tau' s'}, T_a^{\tau' s'}\}_{\tau' s'}$ can be easily incorporated into the proposed method. (i) the point queue model; (ii) the congestible capacitated BPR type link time function (see Bliemer et al. 2013 for a discussion of related models); (iii) weighted-flow model which modifies BPR functions by taking into account the strength of interference between traffics in accordance with their entry time differences:

$$\Gamma_a^{\tau s} = \gamma_a(X_a^{\tau s}) = \gamma_a\left(\sum_{\tau' s'} x_a^{\tau' s'}\, g_{\tau' s'}^{\tau s}\right)$$

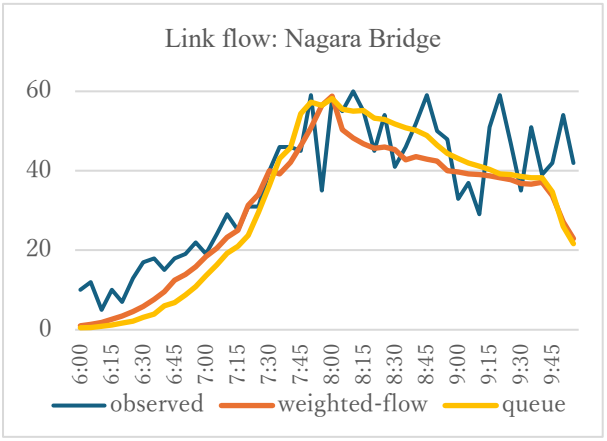
where $\gamma_a$ is a macroscopic function and $g_{\tau' s'}^{\tau s}$ are weights expressing the strength of influences of traffic destined for $s'$ to that destined for $s$, which depend on $\boldsymbol{T}(\boldsymbol{t})$. Models (ii) and (iii), referred to as *queue* and *weighted-flow* models, are adopted in the following case study.

**Case study**

The method was implemented for the road network of Gifu City, Japan, with about 394 thousand population. The study area is composed of 38 zones, covering Gifu city and surrounding smaller cities and towns, with a network of 762 links. The OD consists of commuting and free activity and business trips, with data available from the 2011 PT survey. We assume that during the morning 4 hours from 6:00 to 10:00, all the commuting trips and 40% of the free and business trips are finished. The arrival time horizon is discretized into $T = 96$ time intervals, each with a length of $2.5\ minute$. The most desirable arrival time interval is $[7{:}57{:}30, 8{:}00]$, for all the trips. The costs of early and late arrivals for the commuting trips are set at 0.18 and 0.12 times the travel time cost, respectively. For the free and business trips, they are set at 0.15 and 0.015 times the travel time cost, respectively. These parameters differ largely from those suggested for individual commuters (Small 1982), reflecting the variation of target arrival times. The *queue* and the *weighted-flow* link cost function models build on BPR functions calibrated for roads in Japan. The link capacities (car/minute) for the queuing model and the weighted flow macroscopic model are set at 1/630 and 1/63 of the daily capacity, respectively. In the left figure is shown the profiles of arrival rates during the study periods for the aggregated trips, calculated by the weighted flow model, the queuing model, along with the observed real data, respectively.




In the right figure, the computed link traffic profiles are compared with the real observed data, for a main bridge connecting the north and south parts of Gifu. The computed flows exiting the corresponding links are aggregated into 5-minute intervals, matching the time periods used for collecting real traffic data.



**Findings**

Our approach may be viewed as a mild modification of static models (e.g., Ying and Miyagi 2001) which inherits much of knowledge and experience accumulated in the practice of applying static models. For example, the calibration of the BPR functions has been intensively conducted to fit the cases in different countries and regions. The case study shows that the adoption of the BPR functions widely applied in Japan may yield reasonable results in the dynamic context. The computed arrival rates and traffic profiles in important links fit reasonably well with the real data. We think that the accuracy of the model can be further improved by a refinement of link performance functions considering the types of roads, distribution of traffic signals, and general network structure.

On the other hand, detailed traffic dynamics, including physical queues and spill backs, are not addressed in this paper. Extending our method to address these issues is a challenging yet interesting topic.

**References**

Akamatsu, T., 2001. An efficient algorithm for dynamic traffic equilibrium assignment with queues. Transportation Science 35 (4), 389 - 404.

Bliemer, M.C.J., Raadsen, M. P. H., Smits, E. S., Zhou, B., Bell, M. G. H., 2014, Quasi-dynamic traffic assignment with residual point queues incorporating a first order node model, Transportation Research B 68, 363–384.

Kuwahara, M., Akamatsu, T., 1993. Dynamic equilibrium assignment with queues for one-to-many OD pattern. In: Daganzo, C.F. (Ed.), Proceedings of the 12th International Symposium on Transportation and Traffic Theory. Elsevier, Berkeley, pp.185-204.

Long, J.C., Szeto, W.Y., 2019. Link-based system optimum dynamic traffic assignment problems in general networks. Operations Research 67, 167–182.

Ran, B. and D. Boyce. 1996. Modeling Dynamic Transportation Networks. Berlin: Springer-Verlag.

Ying, J. Q., Miyagi, T., 2001. Sensitivity analysis for stochastic user equilibrium network flows - A dual approach, Transportation Science 35, pp.124-133.

**Extended ABSTRACT for 10th International Symposium on Dynamic Traffic Assignment (DTA2025)**

# A Projection-Based Dynamic System Optimum Emulated using Macroscopic Fundamental Diagram dynamics on regional networks

Mostafa Ameli[a] , Sérgio F. A. Batista[b], J.-P. Lebacque[a] and M. Ménéndez[c]
[a] *University Gustave Eiffel, COSYS, GRETTIA, Paris, France*
[b] *Department of Civil and Environmental Engineering, Imperial College London, UK*
[c] *Division of Engineering, New York University Abu Dhabi, United Arab Emirates*

**Keywords:** System Optimum; Macroscopic Fundamental Diagram; Regional Networks; Projection Method; Dynamic Traffic Assignment.

## 1 INTRODUCTION

In the framework of Dynamic User Equilibrium (DUE), each traveler seeks to minimize individual travel time, whereas, under Dynamic System Optimal (DSO), the objective is to minimize the total travel time or overall system cost (Wardrop, 1952; Sheffi, 1985). Nonetheless, in extensive and congested networks, employing fully detailed Dynamic Traffic Assignment (DTA) models can become computationally prohibitive. Consequently, aggregated models, such as those derived from the Macroscopic Fundamental Diagram (MFD) (Daganzo, 2007), have been developed to encapsulate essential dynamics at the *regional* level. Within each region $i$ in a city or metropolitan area, dynamics are characterized by a singular accumulation $N_i$ and corresponding supply/demand functions $\Sigma_i(N_i)$, $\Delta_i(N_i)$.. In simulation-based DTA, a Dynamic System Optimal (DSO) state is achieved when the total travel time of the system cannot be further reduced by changing the routes selected by travellers. The emulation of dynamic network equilibria in large-scale networks presents numerous challenges, as previously discussed in existing literature (see e.g. Ameli *et al.*, 2020, 2022). Solving the DUE or DSO problems necessitates the determination of the global solution to a constrained convex optimization problem (Sheffi, 1985), with the DSO specifically requiring the computation of marginal travel times (see e.g. Ameli *et al.*, 2024). Although MFD-based DTA models can substantially decrease the state space, resolving the DSO remains complex. Many methodologies rely on the approximation of marginal costs or employ the classical method (see e.g. Peeta & Mahmassani, 1995; Yildirimoglu & Kahraman, 2018).

Although MFD-based DTA can significantly reduce the state space, solving the DSO remains non-trivial. Many approaches rely on approximating marginal costs or the classical method of successive averages (MSA) to find approximate solutions (Yildirimoglu *et al.*, 2015; Batista & Leclercq, 2019). Yildirimoglu *et al.* (2015) discussed a perimeter control framework with route guidance that targets the DSO conditions of the regional network. The authors determined the network equilibrium using the MSA and considered the calculation of the marginal travel times. Aghamohammadi & Laval (2020) provides a review of the application of dynamic traffic assignment models with MFD-based traffic models. The development of simulation-based dynamic traffic assignment models for regional networks is important not only for the design of traffic management strategies but also for setting environmental policies to name a few examples as evidenced in the review work done by (Johari *et al.*, 2021). In this regard, a suitable framework for emulating simulation-based DSO without the necessity to compute approximated path marginals remains absent in the literature. This paper aims to fill this gap. Therefore, in this work, we outline a **projection-based** solution algorithm for the DSO that *avoids* explicit marginal-time calculations. We aim to handle time-varying demand, desired arrival times, and region-to-region path flows within a dynamic simulation.

## 2 Regional Networks, MFD Dynamics, and DSO formulation

We represent a city as a regional network $\mathcal{G} = (R, U)$, where each node $r \in R$ represents a region, and an edge $(i, j) \in U$ indicates that travel is possible from region $i$ to region $j$. Any region can be labelled as origin ($O$), destination ($D$), or intermediate ($J$), depending on its role in the trips under study. When a traveler moves from $i$ to $j$, we write $j \in \Gamma^+(i)$ (so $i \in \Gamma^-(j)$ as well). Each ordered sequence of travelled regions from $O$ to $D$ constitutes a path in the regional network. At time $t \in [0, T]$, the demand of vehicles going from $O$ to $D$ with a desired arrival time $t_a$ is represented by $\Delta^{OD,t_a}(t)$. The total demand of that group is given and the departure times satisfy :

$$\Delta_{tot}^{OD,t_a} \;=\; \int_{\mathcal{T}_d} \Delta^{OD,t_a}(t)\,\mathrm{d}t, \quad \text{where } \Delta^{OD,t_a}(t) \;\geq 0 \text{ and } \mathcal{T}_d \subseteq [0, T]. \tag{1}$$

In this network, each region $i$ has an associated accumulation $N_i(t)$ (total number of vehicles in region $i$ at time $t$). By summing over all destinations $D$ and desired times $t_a$, one obtains

$$N_i(t) \;=\; \sum_{D\in\mathcal{D}} \sum_{t_a\in\mathcal{T}_a} N_i^{D,t_a}(t). \tag{2}$$

To capture congestion effects, each region $i$ is associated with a Macroscopic Fundamental Diagram (MFD), providing a pair of functions $\Delta_i(\cdot)$ (demand) and $\Sigma_i(\cdot)$ (supply), each depending on $N_i(t)$. In particular,

$$\delta_i(t) \;=\; \Delta_i(N_i(t)), \quad \sigma_i(t) \;=\; \Sigma_i(N_i(t)).$$

When $i$ is an origin region, it can be treated as a buffer zone: the buffer dynamics allow a certain inflow of new demand $\Delta^{OD,t_a}(t)$, while outflow is bounded by downstream supply. In the special case of a destination region $D$, the MFD requires only an exit supply constraint, often denoted $\sigma_D$. To update flows in each region, the methodology proceeds in three loops: *(i)* computing the supply and demand for every region based on its current accumulation; *(ii)* determining inflows ($q_{ij}(t)$) by matching demand from region $i$ to supply from region $j$; and *(iii)* applying conservation of vehicles to update the state $N_i(t)$. The flow into region $j$ from $i$ at time $t$ can be calculated by a simple model such as

$$q_{ij}(t) \;=\; \min\{\, \delta_{ij}(t),\; \sigma_{ij}(t)\},$$

where $\delta_{ij}$ is the partial demand (the share of $\delta_i$ wanting to go to $j$) and $\sigma_{ij}$ is the partial supply (the share of $\sigma_j$ allocated to traffic arriving from $i$). One can also employ a more general flow-optimization approach (using KKT conditions) if the invariance principle must be satisfied. Finally, region accumulations evolve through first-order differential equations:

$$\frac{\mathrm{d}N_i}{\mathrm{d}t}(t) \;=\; \sum_{\ell\in\Gamma^-(i)} q_{\ell i}(t) \;-\; \sum_{j\in\Gamma^+(i)} q_{ij}(t),$$

and similarly for each disaggregate class $N_i^{D,t_a}(t)$. By coupling these equations with the MFD relations and the definitions of outflows from origins (or inflows to destinations), the model captures the time evolution of traffic at the regional level in a consistent manner.

After defining how flows and accumulations evolve under MFD dynamics, we introduce a DSO formulation that seeks to minimize the total cost of travel while respecting regional supply/demand constraints. Concretely, we define the system state as all accumulations $\{N_i^{D,t_a}(t)\}$ grouped into a vector $N(t)$. The control (or decision) variables consist of the time-dependent inflows and path-split fractions, denoted by $\Delta_O^{D,t_a}(t)$ (the departure rate from origin $O$) and $\gamma_{ij}^{D,t_a}(t)$ (the fraction of traffic in region $i$ headed to region $j$, with destination $D$ and desired arrival $t_a$). The DSO is then posed as an optimal control problem of the form

$$\min \;\; \int_0^{t_f} \Big[\; c\,.N(t) \;+\; \sum_{t_a\in\mathcal{T}_a} Q^{t_a}(N(t),\gamma(t))\,\mathcal{L}(t_a,t)\Big]\mathrm{d}t \;+\; \tfrac{1}{2}\,\mu\,(N(t_f))^2, \tag{3}$$

where $c\,.N(t)$ represents the total time spent in the network, the sum term captures possible early/late arrival penalties resulting from the network outflows $Q^{t_a}(N(t),\gamma(t))$ and the unit penalty $\mathcal{L}(t_a,t)$, and a terminal term $\frac{1}{2}\,\mu\,(N(t_f))^2$ penalizes any remaining vehicles at final time $t_f$. Subject to network and MFD constraints.

## Gradient-Based DSO with Adjoint States and Projections

To solve the Dynamic System Optimum (DSO) problem, we employ a gradient (optimal control) methodology in discrete time. Let $k = 0, \ldots, K$ denote the time steps. Denote the *state* variables by $N_k$, which include both the total number of vehicles in each region $i$ and the more disaggregated numbers $N_i^{D,t_a}$. Denote the *control* (or decision) variables by $\beta_k \equiv (\gamma_k, \Delta_k)$, which gather the path-split fractions $\gamma_{ij}^{D,t_a}(k)$ and the origin departures $\Delta_O^{D,t_a}(k)$. Our objective function (the discretized version of the total cost) is

$$J(\{\beta_k\}) \;=\; \sum_{k=0}^{K-1} g_k(N_k,\beta_k) \;+\; g_K(N_K),$$

where each $g_k$ is the time-$k$ contribution to the cost (e.g., travel time plus schedule penalties), and $g_K$ is a terminal criterion.

**Calculation of the Partial Derivatives** This section discusses how we calculate the gradient for solving the previous optimal control formulation. Considering that the initial state is given, the criterion can be viewed as a function of the decision variables only. We consider the time steps $k = 0, \cdots, K$. Furthermore, we simplify some notation to derive the gradient of the criterion concerning the decision variables in a discrete-time setting.

Let us start by introducing some notation. We define: $N_k$ as the state in the time step $k$; $\beta_k$ as the command at time step $k$ ($\beta_k = (\gamma_k, \Delta_k)$); $g_k(N_k, \beta_k)$ as the contribution of the time step $k$ to the criterion; and $g_k(N_k)$ as the terminal criterion. Therefore:

$$g_k(N_k,\beta_k) = \left[c \cdot N_k + \sum_{t_a \in \mathcal{T}_a} s \cdot Q^{t_a}(N_k,\beta_k) \cdot \mathcal{L}(t_a, k \cdot \Delta t)\right] \times \Delta t \quad \forall k = 0..K-1 \tag{4}$$

$$g_K(N_K) = \frac{1}{2} \cdot N_K \cdot \mu \cdot N_K \tag{5}$$

$$N_{k+1} = E(N_k,\beta_k) \quad \forall k = 0..K-1 \tag{6}$$

$$N_{k+1} = N_k + \Delta t \cdot F(N_k,\beta_k) \tag{7}$$

i.e.:

$$E(N_k,\beta_k) = \Delta t \cdot F(N_k,\beta_k) + N_k \quad \forall k = 0..K-1 \tag{8}$$

Equations 6, 7, 8 constitute the discretized state equations of the system.

We define $J$ as the discretized version of the criterion:

$$J \stackrel{\text{def}}{=} \sum_{k=0}^{K-1} g_k(N_k,\beta_k) + g_K(N_K) \tag{9}$$

The generic calculation of the derivatives of $J$ concerning $\beta$, i.e. $\partial_{\beta_{K-1}} J$ and $\partial_{\beta_k} J$. First note that $N_h$ depends on $\beta_k$ if and only if $h > k$. More precisely the following apply

$$\begin{array}{rcl} \partial_{\beta_k} N_{k+1} & = & \partial_{\beta_k} E_k \\ \partial_{\beta_k} N_{h+1} & = & \partial_{N_h} E_h \cdot \partial_{\beta_k} N_h \quad \forall h > k \end{array} \tag{10}$$

Second, for all $k$:

$$\partial_{\beta_k} J = \partial_{\beta_k} g_k + \sum_{h=k+1}^{K-1} \partial_{N_h} g_h \cdot \partial_{\beta_k} N_h + \partial_{N_K} g_K \cdot \partial_{\beta_h} N_K \tag{11}$$

and the iterative calculation of the derivatives of $J$ concerning $\beta_k$ follows:

$$\begin{array}{rcl} \partial_{\beta_{K-1}} J & = & \partial_{\beta_{K-1}} g_{K-1} + \partial_{N_K} g_K \cdot \partial_{\beta_{K-1}} E_{K-1} \\ \partial_{\beta_k} J & = & \partial_{\beta_k} g_k + \sum_{h=k+1}^{K} \partial_{N_h} g_h \cdot \partial_{N_{h-1}} E_{h-1} \cdot \ldots \cdot \partial_{N_{k+1}} E_{k+1} \cdot \partial_{\beta_k} E_k \quad \forall k < K \end{array} \tag{12}$$

Thus let us introduce the adjoint state $\Upsilon$:

$$\begin{cases} \Upsilon_K = \partial_{N_K} g_K \\ \Upsilon_k = \partial_{N_k} g_k + \Upsilon_{k+1} \cdot \partial_{N_k} E_k, \quad k = K-1, \cdots, 0 \end{cases} \tag{13}$$

and the gradient of $J$ is expressed as:

$$\partial_{\beta_k} J = \partial_{\beta_k} g_k + \Upsilon_{k+1} \cdot \partial_{\beta_k} E_k, \quad k = 0, \ldots, K \tag{14}$$

Actually, given that the set of constraints applied to $\beta$ is convex, the discrete-time minimum principle applies. We will not apply this approach here.

**Projected Gradient Update.** Finally, we perform an update of the control variables in the negative gradient direction:

$$\beta_k^{(\tau+1)} = \mathcal{P}\Big[\beta_k^{(\tau)} - \alpha^{(\tau)} \partial_{\beta_k} J(\beta^{(\tau)})\Big],$$

where $\mathcal{P}$ is the projection operator that enforces nonnegativity, simplex constraints, and total demand constraints. In practice, these subproblems (simplex and demand projection) can be solved with rank-based cut values, ensuring that the computational cost remains modest for each iteration. In particular, each of these projection subproblems can be solved in $O(K \log K)$ time per region or OD class, using cut-value searching. Hence, despite the complexity of the overall DSO, the gradient steps remain computationally tractable when combined with the MFD-based region-state updates.

# 3 Numerical Proof of Concept

We implement the above framework on a synthetic four-region network. Each region has a basic triangular MFD with known critical density. We consider multiple destination classes and desired arrival times. The simulation runs from $t = 0$ to $t = T$, after which no new inflow arrives, but vehicles may still clear the network. Each iteration of the solution updates:

1. *Forward simulation*: compute demands/supplies and flows for each region via the MFD loops.
2. *Backward pass*: compute the adjoint $\Upsilon_k$ needed for $\partial_\beta J$.
3. *Projection step*: update $\gamma_{ij}^{D,t_a}$ and $\Delta_O^{D,t_a}$ under simplex constraints.

The objective function (total time spent plus schedule penalties) decreases steadily until numerical convergence. The results confirm that we reach the *global optimum*, matching a reference solution obtained via a small-scale enumerative check. CPU times remain reasonable, illustrating the feasibility of this approach for medium-sized networks.

# ACKNOWLEDGEMENTS

M. Ameli acknowledges support from the French ANR research project SMART-ROUTE (grant number ANR-24-CE22-7264). M. Menéndez acknowledges support from the NYUAD Center for Interacting Urban Networks (CITIES) funded by Tamkeen under the NYUAD Research Institute Award CG001.

# References

Aghamohammadi, Rafegh, & Laval, Jorge A. 2020. Dynamic traffic assignment using the macroscopic fundamental diagram: A Review of vehicular and pedestrian flow models. *Transportation Research Part B: Methodological*, **137**, 99–118. Advances in Network Macroscopic Fundamental Diagram (NMFD) Research.

Ameli, Mostafa, Lebacque, Jean-Patrick, & Leclercq, Ludovic. 2020. Simulation-based dynamic traffic assignment: Meta-heuristic solution methods with parallel computing. *Computer-Aided Civil and Infrastructure Engineering*, **35**(10), 1047–1062.

Ameli, Mostafa, Faradonbeh, Mohamad Sadegh Shirani, Lebacque, Jean-Patrick, Abouee-Mehrizi, Hossein, & Leclercq, Ludovic. 2022. Departure time choice models in urban transportation systems based on mean field games. *Transportation Science*, **56**(6), 1483–1504.

Ameli, Mostafa, Lebacque, Jean-Patrick, Alisoltani, Negin, & Leclercq, Ludovic. 2024. Collective departure time allocation in large-scale urban networks: A flexible modeling framework with trip length and desired arrival time distributions. *Transportation Research Part B: Methodological*, 102990.

Batista, S. F. A., & Leclercq, L. 2019. Regional dynamic traffic assignment framework for MFD multi-regions models. *Transportation Science*, **53**, 1563–1590.

Daganzo, C. 2007. Urban gridlock: Macroscopic modeling and mitigation approaches. *Transportation Research Part B: Methodological*, **41**, 49–62.

Johari, Mansour, Keyvan-Ekbatani, Mehdi, Leclercq, Ludovic, Ngoduy, Dong, & Mahmassani, Hani S. 2021. Macroscopic network-level traffic models: Bridging fifty years of development toward the next era. *Transportation Research Part C: Emerging Technologies*, **131**, 103334.

Peeta, Srinivas, & Mahmassani, Hani S. 1995. System optimal and user equilibrium time-dependent traffic assignment in congested networks. *Annals of Operations Research*, **60**(1), 81–113.

Sheffi, Y. 1985. *Urban Transportation networks: Equilibrium Analysis with Mathematical Programming Methods*. United States of America: Prentice Hall Inc. Chap. 10 and 11.

Wardrop, J. G. 1952. Some theoretical aspects of road traffic research. *Institution of Civil Engineering*, **1**, 325–362.

Yildirimoglu, Mehmet, & Kahraman, Osman. 2018. Searching for empirical evidence on traffic equilibrium. *PloS one*, **13**(5), e0196997.

Yildirimoglu, Mehmet, Ramezani, Mohsen, & Geroliminis, Nikolas. 2015. Equilibrium analysis and route guidance in large-scale networks with MFD dynamics. *Transportation Research Part C: Emerging Technologies*, **59**, 404 – 420. Special Issue on International Symposium on Transportation and Traffic Theory.

# Assessing the evolutionary process of self-financed shared mobility systems: A laboratory experimental approach

Haruki Takiguchi[1], Sowa Suzuki[1], Takamasa Iryo[2], Koki Satsukawa[3], and Haruko Nakao[4]

[1]Tohoku University, [2]University of Tokyo, [3]Kanazawa University, [4]University of Luxembourg

*Keywords: on-demand shared mobility, evolutionary process, community-owned transport*

## 1 Introduction

In rural areas and suburbs, the low density of transport demand often prevents the development of conventional public transport services or causes operators to withdraw. These areas have become public transport blank areas, leading to an increasing dependence on private cars. Elderly people, children, and mobility-impaired people who cannot use private cars are forced to suffer the inconvenience of restricted freedom of movement.

One way to ensure local transport in such circumstances is to share and make use of transport resources such as vehicles and drivers, which incur fixed costs, among local residents. Since these limited resources can be effectively utilised by multiple users by sharing them, it is expected that a public transport service can be provided to the community at a lower operating cost than conventional public transport. In this study, this type of public transport, in which resources are shared in the community, is named community-owned shared transport (CST).

One of the issues in operating CST is to establish a fair system for sharing the cost of securing transport resources. One such system is the average cost system, in which the residents who use the CST share the operating costs equally. As each individual can decide whether or not to use the CST independently, and as the operating costs of the CST are shared equally, there is no need for special consensus-building efforts. On the other hand, the most important issue in continuing to operate this type of CST is that if a sufficient number of subscribers cannot be obtained, each subscriber will be charged an excessive fee, and as a result, the number of CST users will decrease. This problem is particularly problematic in the early stages of introducing a CST with a small number of subscribers. In this case, the fixed costs of the CST will be covered by a small number of subscribers, and the high subscription fee will make the CST less attractive to potential subscribers.

In this study, we propose a measure to increase the attractiveness of CSTs in the early stages of introduction by providing an incentive to set up a meeting point (MP) in front of one's home by signing up early. In general, in shared mobility services such as CSTs, restricting the number of meeting points increases the efficiency of the service. By using such restrictions, we can expect not only an increase in efficiency but also an effect on promoting the spread of CSTs in the early stages of introduction as an incentive for early subscription. The effect of this incentive can be evaluated by constructing a day-to-day dynamic model that describes the diffusion of CST and assessing where the dynamics converge with and without the incentive. When the initial number of subscribers is set to zero, there are two possible outcomes of this dynamic: (1) the number of subscribers increases and the subscription fee per subscriber becomes lower, making the CST attractive to many people and stabilising at a high number of CST users. (2) The number of subscribers does not increase and the subscription fee per subscriber is too high, so no one uses CST. It is theoretically predicted that (2) will occur in the absence of incentives, and (1) will occur in the presence of incentives. However, this is only a theoretical prediction. Actual users may behave differently, and as a result, the outcome may differ from the above prediction.

To investigate the day-to-day diffusion dynamics with and without the incentive, we conduct a laboratory experiment simulating a CST. Laboratory experiments are a well-known technique

in experimental economics, but they are also a common tool in transport science (Dixit *et al.*, 2017). It has the advantage that it can be carried out at a much lower cost than experiments using real-world transport systems, and that the experimental conditions can be freely controlled. This paper reports on the day-to-day dynamics of the behaviour adjustment process observed in these indoor experiments, and analyses how the results match and differ from the theoretically predicted dynamics.

# 2 Methodology

## 2.1 Overview of the experiment

The experiment was conducted using a web-based system constructed using the online experimental platform oTree Chen *et al.* (2016). The experiments were conducted for 3 days on 19th, 20th and 21st December 2024 at the campus on Tohoku University. The participants were assembled in a single room and a desktop computer equipped in the room was assingned to each of them. The conversation among the participants was forbidden. On each day, 10 sessions of the experiment were conducted with different conditions. It was repeated for 3 days in total of 30 sessions. The participants were recruited among students of Tohoku University. The fixed reward was paid for each participant. In total, 49 people joined the experiments, 17 people on the first day, 16 people on the second day and 16 people on the third day. The experiments were conducted with the approval of the research ethics committee of the Graduate School of Information Science and Technology, Tohoku University.

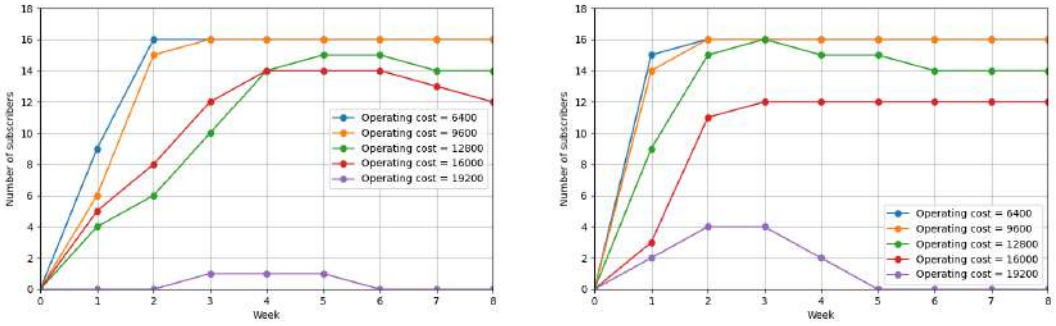
## 2.2 Experiment settings

For this experiment, we considered a residential area without access to public transport. The players were assumed to reside in this area without access to a private car. The scenario was set during the first eight weeks following the introduction of a community-owned shared transport (CST) service. The CST was designed to operate using designated meeting points (MPs), where users could be picked up and dropped off. In this experiment, the number of MPs was deliberately set to be smaller than the number of players, and the travel cost from a participant's home to an MP was assumed to depend on the distance between the two.

Among the 10 experimental sessions, five were conducted with an incentive scheme in which early subscribers had an MP located directly in front of their home. This allowed them to maintain zero travel costs to reach the MP over time. Another set of sessions was conducted without any incentives, where MP locations were predetermined. For each set of session with and withou incentive scheme, the operating cost of the CST varied across five different levels, ranging from 12,800 to 19,200 JPY, with each session featuring a different cost level. For sessions with 17 participants, the cost was adjusted by a factor of 17/16 to maintain consistency.

For each session, players were asked to decide whether to subscribe to the CST, based on the following rules:

- Players aimed to minimize their total transport costs over the eight-week period.
- They had the opportunity to decide whether to subscribe to the CST at the beginning of each week, meaning each player had up to eight decision points during the experiment.
- Once a player subscribed to the CST, they were required to remain in the service for a minimum contract period of three weeks, during which cancellation was not permitted.
- If a player declined CST subscription, they used an alternative transport mode.
- Locations of players' homes were randomly assigned, and their alternative transport costs were calculated based on these places. Their home locasionts remain consistent for the two sessions that have the identical CST operation costs—one with an incentive scheme and the other without.




(a) *Without incentive* (b) *With incentive*

Figure 1 – *Changes in the number of subscribers, 2nd day of experiment*

The total operating cost of the CST, including maintenance and purchasing costs, was equally shared among subscribers. Consequently, the weekly subscription fee was determined by dividing the total CST operating cost by the number of subscribers for that week. If the number of subscribers fell below a critical threshold, the CST service would cease operation for that week. In such cases, subscribers were required to pay the minimum maintenance cost while using alternative transport.

Players did not know the exact subscription fee in advance, as it fluctuated based on the number of subscribers each week. However, they were provided with the following information to aid their decision-making: 1) The total operational cost of the CST, 2) Subscription fees in the previous week, 3) Alternative transport costs, 4) Travel cost to the nearest MP, and 5) The minimum number of subscribers required for the CST to continue operating.

## 3 Experimental results

Figure 1 illustrates how the number of subscribers changes over eight weeks at five different levels of CST operational cost, without the incentive scheme (left) and with the incentive scheme (right) on the 2nd day of the experiment. Regardless of the incentive scheme, all participants eventually subscribed to CST when the operational cost was 6,400 JPY or 9,600 JPY. At an operational cost of 12,800 JPY and 16,000 JPY, 75% of participants subscribed to CST. In the case of the highest operational cost, there were no subscribers after eight weeks. Across all cases, some players joined the CST even when there were only a few subscribers and the subscription cost was high. This may be because they anticipated that the subscription cost would eventually decrease as the number of subscribers increased. Once the system reached a stable state, players' daily behavior generally aligned with the theoretical assumption of short-sighted decision-making. With the introduction of the incentive scheme, the number of subscribers reached equilibrium faster than in the case without incentives. However, it did not affect the total number of subscribers.The same trend was observed on other days as well

The experimental results also demonstrated the theoretically predicted dynamics of convergence to a closer stable equilibrium solution following an unstable equilibrium. Figure 2 shows the cost of using CST (y-axis) versus the number of subscribers (x-axis), represented by the blue line. It also depicts the user's willingness to pay, defined as the difference between the cost of alternative mode and the travel cost to to the MP, for each number of subscribers, represented by the yellow line. The left-hand intersection of the two lines represents the unstable equilibrium solution, while the right-hand intersection represents the stable equilibrium solution. The red dotted line indicates the number of subscribers each week, and the red circle represents the cost of using CST for that week, given the number of subscriber. The yellow and blue arrows at

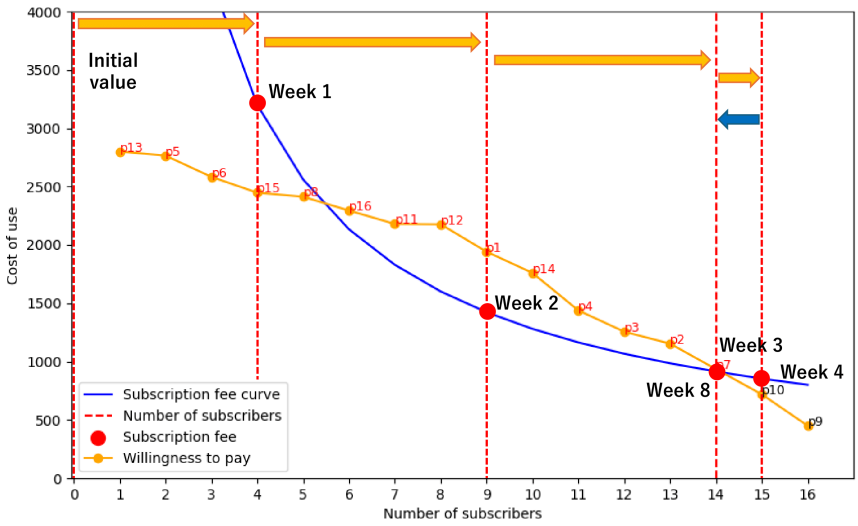



Figure 2 – *Process of convergence to stable equilibrium solution. 3rd day of experiment, no incentive, session with total cost of 12,800*

the top of the figure illustrate the process of convergence to the stable equilibrium solution over the eight-week period. As shown in the figure, the number of subscribers was initially below the unstable equilibrium solution in the first week but gradually increased, surpassing the stable equilibrium solution before eventually converging to it.

## 4 Conclusion

Under the conditions of this experiment, it is discovered that the results generally followed theoretical predictions. It was observed that CST was adopted even when the total operational cost was near the upper limit predicted by theory. Additionally, the incentive scheme was found to increase the number of subscribers in the early stages. We have also conducted a stated-preference survay and simulations based on this result, which will be explained at the presentation.

Future research directions include developing a theoretical framework to explain the non-myopic behavior observed in players during the initial weeks, constructing a user behavior model that better reproduces the convergence to the equilibrium solution based on myopic decision-making in later weeks, and investigating the impact of changes in conditions such as group size and minimum subscription period.

## Acknowledgement

This study was financially supported by the Kajima Foundation's Support Program for International Joint Research Activities.

## References

Chen, Daniel L., Schonger, Martin, & Wickens, Chris. 2016. oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, **9**, 88–97.

Dixit, Vinayak V., Ortmann, Andreas, Rutström, E. Elisabet, & Ukkusuri, Satish V. 2017. Experimental Economics and choice in transportation: Incentives and context. *Transportation Research Part C: Emerging Technologies*, **77**, 161–184.

# Markov-chain-based model for the evolutionary process of self-financed shared mobility systems: Theoretical assessments

Takamasa Iryo[1], David P. Watling[2], Koki Satsukawa[3], Richard D. Connors[4], Haruko Nakao[5], and Sowa Suzuki[6]

[1]University of Tokyo, [2]University of Leeds, [3]Kanazawa University, [4]Luxembourg Institute of Science and Technology, [5]University of Luxembourg, [6]Tohoku University

*Keywords: shared mobility, evolutionary process, community-owned transport, Markov chain*

## 1 Introduction

In rural areas, demand for transport is low and dispersed, making conventional public transport systems inefficient. In such regions, individuals who cannot use private vehicles, such as the elderly and children, tend to experience limited mobility. On-demand shared mobility services have the potential to serve as an effective alternative to traditional public transport. However, concerns regarding profitability may discourage private operators from investing in such services.

In this study, we consider a community-operated, self-funded on-demand shared mobility service (CST) as an alternative transport solution in such regions. We define CST as a means of transport established by sharing mobility resources within a community, where the cost of utilising these resources remains constant regardless of the number of users and does not vary with patronage. We could expect that an autonomous shared vehicle (SAV) is particularly well-suited for CST, as it eliminates driver wages, a major operational cost. Even if SAVs become widely available, their costs may still be unaffordable for individuals. However, sharing an SAV among community members could distribute these costs, making it a viable mobility solution.

One way to introduce CST in rural communities is through voluntary consensus-building among residents. However, this approach involves high communication costs and is not straightforward. To promote CST without relying on voluntary consensus-building within the community, it is necessary to devise an operational strategy that ensures CST adoption, even when individual residents make independent decisions about participation. One example is the approach proposed by Nakao *et al.* (2025), which involves initially allowing operational deficits by reducing fares and recovering the losses once CST adoption becomes widespread.

Nakao *et al.* (2025) formulated the CST adoption process using a Markov chain process model to represent day-to-day dynamics and demonstrated its effectiveness under certain conditions through numerical simulations. However, the implementation rules used in their simulations were constructed in an ad hoc manner and they are not necessarily optimal. Therefore, there is no guarantee of robustness in the solutions when the parameters change. Furthermore, it is known to be difficult to assess the long-run dynamics of whether adopted CST can be sustained over time through numerical experiments.

This study aims to comprehensively investigate the characteristics of a Markov-chain-based model for the evolutionary process of CST using a mathematical approach, supplemented by numerical methods where necessary. Specifically, the analysis focuses on the following aspects:

- Dynamics without policy interventions: Estimate transition probabilities between multiple equilibrium states and perform sensitivity analyses on key parameters.
- Dynamics under fare-reduction policies: Evaluating the probability of successful adoption of CST and assessing the overall financial balance of operational costs.
- Optimisation of fare reduction strategies: Formulating an optimisation problem to derive the optimal fare reduction policy and proposing a method to obtain its solution.

These insights are expected to enhance the practical effectiveness of the policy proposed by Nakao *et al.* (2025) providing a theoretical foundation for designing more reliable and effective strategies to promote CST adoption.

## 2 Methodology

In this study, we use a simplified model of the transport system proposed by Nakao *et al.* (2025), which includes CST. We consider a region consisting of $N$ residents indexed by integers $i$ from 1 to $N$. Each resident makes one trip per day using either CST or an alternative transport mode. If resident $i$ chooses CST on day $t$, we define $x_i(t) = 1$. If they use an alternative transport mode, $x_i(t) = 0$. The number of CST users on day $t$ is given by:

$$x(t) = \sum_{i=1}^{N} x_i(t).$$

The generalized transport cost of the alternative mode for the resident $i$ is indicated as $\pi_i$. The total operating cost of CST is given as a positive constant denoted by $C$. The CST fare on day $t$ is defined as $f(t)$, which is determined by a given fare policy. In the long run, the collected fares must support the operating cost, which means:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \{x(t)f(t) - C\} \geq 0,$$

which is preferably to be zero to maximise benefits of the residents. If daily income and cost must always be in agreement, then for all $t$, the fare must satisfy:

$$x(t)f(t) = C$$

Once residents choose between CST and an alternative transport mode, they retain their choice over days but occasionally update it. Let $r(t)$ be the probability that a resident updates their choice on day $t$. In the simplest case, $r(t)$ is assumed to be constant over time, denoted by $r$ in this case. When resident $i$ updates their choice on day $t$, the probability of choosing CST, $p_i$, follows the logit model:

$$p_i = \frac{\exp(\theta(f(t) + \rho_i))}{\exp(\theta(f(t) + \rho_i)) + \exp(\theta\pi_i)} = \frac{1}{1 + \exp(\theta(\pi_i - \rho_i - f(t)))}$$

where $\theta$ is a logit parameter and $\rho_i$ is an additive generalised travel cost of resident $i$, e.g. walking cost to a nearest meeting point of the CST. The formulated model can be interpreted as a Markov chain with $x_i(t)$ as the state variables.

To analyse the properties of the proposed model, we primarily use the following methods:

1. **Continuous Approximation:** Assuming $N$ is sufficiently large and $r(t)$ is small, we derive a continuous approximation similar to the mean dynamics in Hofbauer & Sandholm (2007). This leads to an ordinary differential equation formulation, making mathematical analysis more tractable than the original Markov chain.

2. **Simplified State Space:** When $p_i$ is constant across all residents, individual differences disappear, allowing the formulation of a Markov chain with $x(t)$ as the state variable. Alternatively, if $p_i$ takes only two distinct values, the model reduces to a Markov chain with two state variables, making analytical treatment more feasible.

3. **Stationary Distribution of the Markov Chain:** If $p_i$ is constant, it is expected that the stationary distribution can be derived analytically.

4. **Markov Chain Mixing Time (MCMT) Analysis:** MCMT evaluates how quickly the Markov chain converges to its stationary distribution. In this model, MCMT is expected to reflect the transition speed between CST adoption and non-adoption states. We adopt the methodologies proposed by Iryo *et al.* (2024) for assessing MCMT.

For the proposed model, MCMT represents the rate at which CST adoption shifts between widespread use and limited adoption. If MCMT is sufficiently large, once CST adoption becomes prevalent, it is expected to remain stable over time.

## 3 Current findings and future prospects for the presentation

We here present an example of results by numerical simulations. As an example of the fare-reduction policies, we set a the maximum fare as a given constant, denoted by $f_{\max}$, and an exponential-smoothing strategy that defers the change of the fare over days. It is defined as:

$$f(t) = \begin{cases} \min\left\{(1-\gamma)f(t-1) + \dfrac{\gamma C}{x(t)}, f_{\max}\right\} & \text{if } \ x(t) > 0 \\ f_{\max} & \text{if } \ x(t) = 0 \end{cases} \tag{1}$$

which gives a balanced fare if $x(t)$ is positive and stable over days and $C/x(t) \leq f_{\max}$.

Selected results are shown in Figure 1, where we let $N = 100$, $r(t) = r = 0.1$, $\pi_i - \rho_i$ be between 3.0 and 22.8. The daily income/loss is defined as $x(t)f(t) - C$ and the cumulative income/loss is defined as its summation since the first day. We can confirm that the debt cannot be repaid if $f_{\max}$ is too low and the CST cannot be popularised if $f_{\max}$ is too high, while a moderate $f_{\max}$ can lead the CST to a cost-neutral state in which the daily income/loss is zero with a sufficient number of users.

Based on the findings in the preliminary results above, we plan to demonstrate the following results in our presentation. We will use methods 1 and 2 for the first two items. We will use methods 3 and 4 for the last two items. We will also investigate the dependence on $r(t)$ particularly for these two items.

**A**. Method for determining the optimal $f(t)$: The $f(t)$ calculated by exponential smoothing has been successful in ensuring the use of CST while repaying the debt when $f_{\max} = 10$. However, looking at the results in Figure 1, cumulative income is largely positive. We propose a method for determining $f(t)$ that can make this zero. In addition, the maximum value of cumulative loss during popularisation should be as small as possible. We also propose a method for determining $f(t)$ that achieves this.

**B**. Method for determining the optimal $f_{\max}$ when $\pi_i$ is unknown: If $\pi_i$ of all users are known, the optimal $f_{\max}$ can be determined by a simulation. but such a situation is rare in the real world. We propose a method for automatically searching for an appropriate $f_{\max}$ when $\pi_i$ is unknown.

**C**. Calculating $x(t)$ in the stationary distribution: This is useful when evaluating the number of CST users based on their probability distribution. Moreover, if a stationary distribution is obtained with two major peaks in both the state where the CST is not popular and the state where it is popular, there is a possibility that even if the CST becomes popular, the state will not be maintained for a long time. Knowing the stationary distribution is also useful for detecting such situations.

**D**. Stability analysis using MCMT calculations: This is mainly carried out to evaluate the stability of the state in which CST is popularised.

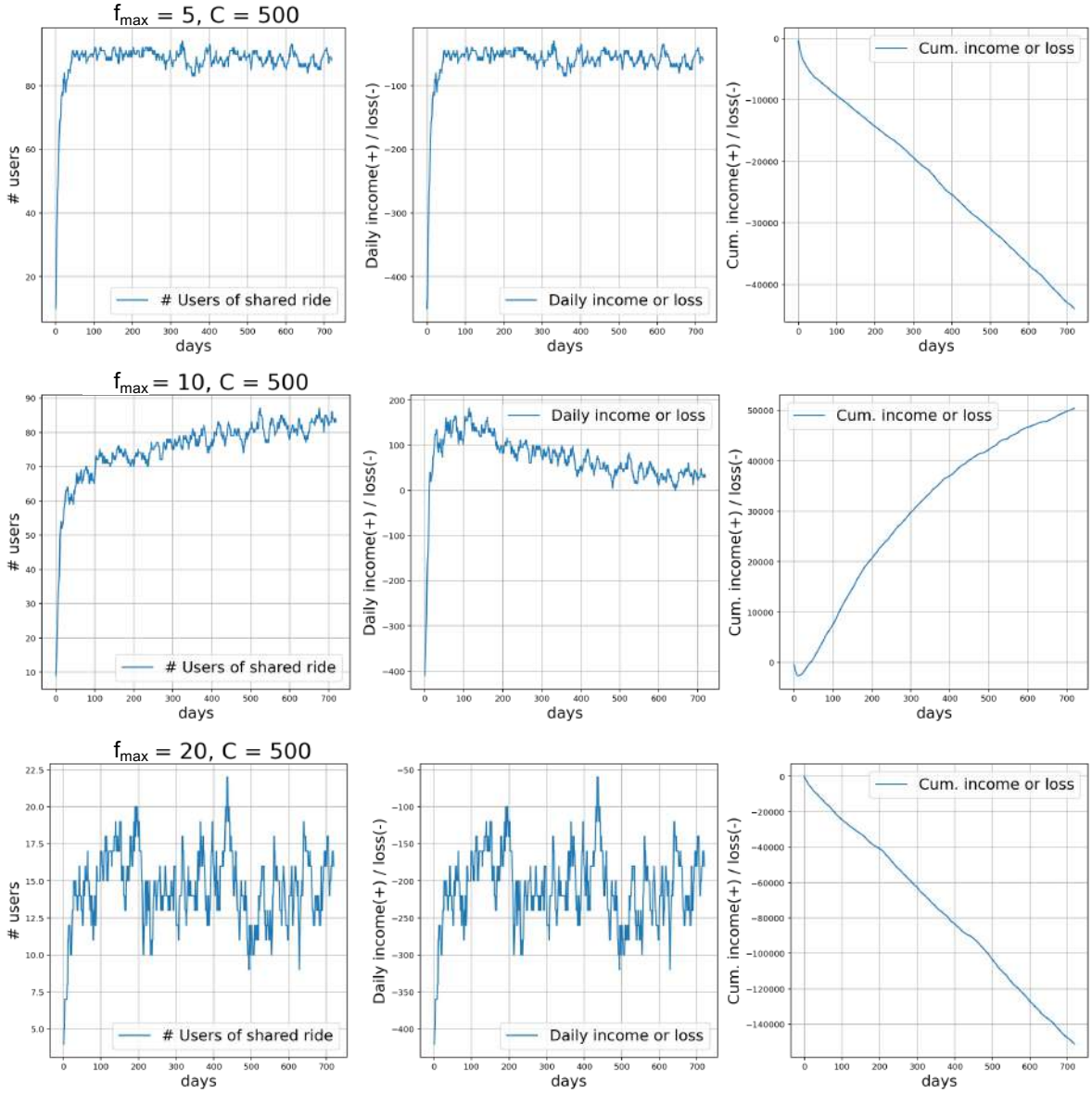



Figure 1 – Results of the evolutionary processes with different $f_{\max}$

## Acknowledgements

This study was financially supported by the Kajima Foundation's Support Program for International Joint Research Activities.

## References

Hofbauer, Josef, & Sandholm, William H. 2007. Evolution in games with randomly disturbed payoffs. *Journal of Economic Theory*, **132**(1), 47–69.

Iryo, Takamasa, Watling, David, & Hazelton, Martin. 2024. Estimating Markov Chain Mixing Times: Convergence Rate Towards Equilibrium of a Stochastic Process Traffic Assignment Model. *Transportation Science*, **58**(6), 1168–1192.

Nakao, H., Satsukawa, K., Iryo, T., Connors, R., & Suzuki, S. 2025. *Evolutionary process of self-financed shared mobility systems.* Abstract to be presented at the Triennial Symposium on Transportation Analysis (TRISTAN) XII.

# A stochastic transit assignment with adaptive route choice

Oskar A. L. Eikenbroek[a] , Eric C. van Berkum [a]

[a] *Transport Studies, University of Twente, Enschede, The Netherlands*

## Abstract

Transit assignment models estimate the distribution of passenger demand over a public transportation (PT) network. Schedule-based assignment models use a space-time graph to represent the timetable, yet oftentimes assume that the arrival and departure times of services are known with certainty, i.e., deterministic. In practice, travelers in PT are faced with uncertainty regarding their travel and use online information to adapt their decisions. This paper proposes a novel stochastic transit assignment model, where the arrival and departure times of PT vehicles are modeled as random. Riders react to and anticipate changing levels of information regarding arrivals and departures and crowding effects due to the adaptive choice behavior of others. While passengers are en route, the uncertainty is gradually realized and recourse decisions are made. As such, passenger route choice is modeled as a multi-stage stochastic program, leading to random passengers flows. We formulate user equilibrium conditions for networks with stochastic supply, and propose a stochastic variational inequality to efficiently find such equilibrium flows. Numerical experiments in realistic networks reveal the differences with a deterministic approach.

**Keywords:** Adaptive Route Choice, Transit assignment, Equilibrium, Public Transport

## 1 Introduction

Public transportation (PT) assignment models distribute passenger demand over a network, under assumptions regarding choice behavior. Such models are often a 'lower level' in hierarchical optimization, where a set of measures is evaluated while accounting for the behavioral responses of passengers to these interventions (Eikenbroek et al., 2018). In this paper, we focus on schedule-based assignment models using a timetable-based space-time graph. As noted by Kumar and Khani (2023), it is typically assumed in these models that PT vehicles arrive and depart on time. Consequently, the underlying space-time graph is deterministic, and passengers have thus full certainty regarding the feasibility of their travel. In practice, there is quite some uncertainty with respect to the realization of a timetable, and perturbations or events can lead to delayed arrivals and departures, which may eventually cause that passengers miss their transfer. Typically, there is an imbalance between information and decisions. Travelers have to make a share of their decisions before all information is available. During the travel, the uncertainty is dynamically revealed and passengers make recourse decisions in response to new information. This is either because they have to, e.g., if they have missed or are going to miss a transfer, or because they want to, if the intended path has become sub-optimal conditional on the information available. Hence, assignment models should consider both the a priori choices of passengers, but also their en-route adaptations (cf. Hall, 1986).

Various studies have aimed at capturing the adaptive choice behavior of passengers in assignment models. Recently, Kumar and Khani (2023) proposed a schedule-based uncapacitated and capacitated assignment where passengers are modeled as expected-cost minimizers, utilizing online bus arrival information to adapt their decisions. The steady-state flow distribution represents the average flow of passengers on each link in the network. Rambha et al. (2016) represent passenger decision making as a Markov decision process, where travelers are assumed to formulate a policy based on the least expected cost while guaranteeing that the destination is reached within a threshold. In both cases - in contrast to our study - an independence assumption is made regarding transit travel times. In addition, these studies ignore the potential influence of crowding or congestion on the perceived travel time, or generalized cost. Tirachini et al. (2016), among others, showed that passengers not only prefer uncrowded routes, but also dislike to stand. In fact, it can be hypothesized that crowding is becoming an increasingly important attribute when evaluating alternatives since passengers have nowadays access to real-time crowding information (Jenelius, 2020).

In this paper, we propose a novel schedule-based public transport assignment model explicitly accounting for emerging information, i.e., supply is inherently variable (e.g., delays and disruptions), and riders react to and anticipate changing levels of information regarding supply, but also about crowding and discomfort due to adaptive choices of others. Consequently, the assignment is modeled as random; passenger decisions are not fixed or modeled to be in steady-state but depend on the scenario.

## 2 Problem Formulation

We consider a stochastic schedule-based static transit assignment with fixed demand. In our stochastic setting, we consider probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with supply depending on the elements $\omega$ ('scenarios') of $\Omega$. Hereinafter, we assume $\Omega$ to be a finite set ($\mathcal{F} = 2^{\Omega}$) with mass $\mathbb{P} : \Omega \to [0, 1]$, so that $\sum_{\omega \in \Omega} \mathbb{P}(\omega) = 1$, and $\mathbb{P}(\omega) > 0$. We do not assume travel times to be independent: the only assumption we make is that the possible realizations of the timetable can be captured by a discrete probability measure with a finite number of scenarios, (numerical approaches in stochastic programming almost exclusively rely on discrete probability measures (Henrion and Römisch, 2022)).

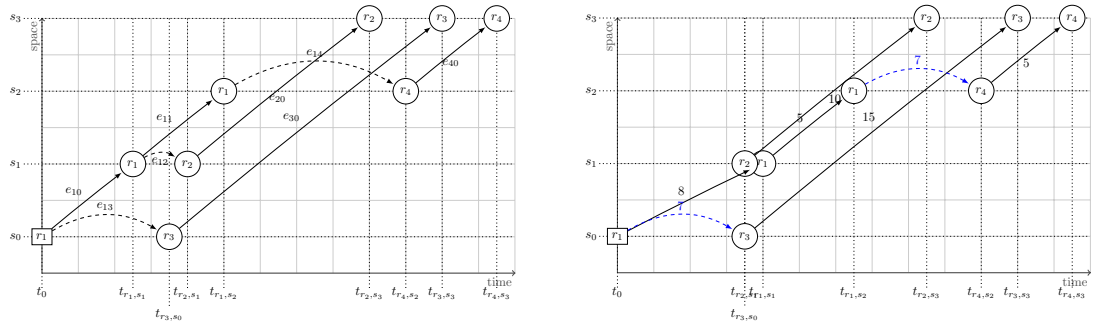
To model the PT network, we use a random time-space graph $G(\cdot) = (V(\cdot), E(\cdot)) : \omega \mapsto G(\omega) = (V(\omega), E(\omega))$, where - for each scenario $\omega$ - each node $v$ is positioned in time and space, that is $v = (t, s) \in [0, T] \times S$ with $[0, T]$ the operating time interval and $S$ the set of stations in the network.

Supply is modeled through a timetable with a set of services $\mathcal{R}$, and each service $r$ serves a set of $n_r$ stations $S_r = \{s_r^0, s_r^1, \ldots, s_r^{n_r}\} \subset S$. The timetable denotes the arrival and departure time at each station, but for simplicity we assume there are no dwell times. In our stochastic setting, arrival times at stations are uncertain and node positions in the time-space graph become $\omega$-dependent. Assuming none of the services are canceled and no stops are skipped, for a given scenario (realized timetable) $\omega$ and service $r$, arrival times are denoted by $t_r^i(\omega)$. As such, each service $r$ induces a random subgraph $G_r(\cdot) = (V_r(\cdot), E_r) : \omega \mapsto G_r(\omega) = (V_r(\omega), E_r)$, with node set $V_r(\omega) = \{(t_r^0(\omega), s_r^0), (t_r^1(\omega), s_r^1), \ldots, (t_r^{n_r}(\omega), s_r^{n_r})\}$, with $t_r^i(\omega) \leq t_r^{i+1}(\omega)$. The edges $e \in E_r$ connect 'consecutive' nodes $(t_r^i(\omega), s_r^i)$ and $(t_r^{i+1}(\omega), s_r^{i+1})$, representing PT vehicle movements between two stops with $\omega$-dependent travel time $t_e(\omega) = (t_r^{i+1}(\omega) - t_r^i(\omega)) \in \mathbb{R}_+$.

We have a set of multi-period origin destination pairs (OD pairs) $\mathcal{W}$ with $(t_w, s_w^o, s_w^d) \in [0, T] \times S \times S$ and demands $d_w > 0, w \in \mathcal{W}$. Each OD-pair is characterized by a departure time-station combination $(t_w, s_w^0) \in V$ and an egress station $s_w^d$. Apart from *travel arcs*, edge sets include *access*, *egress* and *transfer arcs* (Luan et al., 2024). Access arcs represent passengers entering the system from an origin, egress arcs model passengers arriving at their destination. Transfer arcs model passengers changing services, where we for now ignore the possible minimum transfer time at station. The scenario-based approach allows for modeling cases in which due to the uncertainty in arrival and departure times, some transfers and boardings become impossible. Access arcs $e = ((t_w, s_w^0), (t_r^i(\omega), s_r^i))$ are only included in $E(\omega)$ if $t_r^i(\omega) \geq s_w^0$. Similarly, transfer edge $e = ((t_r^i(\omega), s_r^i), (t_q^j(\omega), s_q^j))$ is only included if $t_r^i(\omega) \leq t_q^j(\omega)$.

### 2.1 Illustrative Example

We use an example to illustrate our approach to capture scenario-dependent timetables using a random space-time graph, and how passengers make adaptive routing decisions based on dynamic information. Figure 1(a) provides the time-space graph of a scheduled timetable. Travelers are assumed to arrive at time $t_0$ at station $s_0$, and all want to travel to station $s_3$. In order to do so, they can travel along path $p_1$, i.e., directly board service $r_1$, travel to station $s_1$, and transfer to service $r_2$. Alternatively, they can use service $r_1$ to station $s_2$ and transfer there to service $r_4$ (path $p_2$). The last possibility is to wait at station $r_1$ and use service $r_3$ directly to $s_3$ (path $p_3$).




(a) Scheduled timetable

(b) Realized timetable under scenario $\omega^4$

Figure 1: Scenario-dependent timetables

In our stochastic setting, arrival times are uncertain, and transfers may become impossible. In such a setting, passengers adopt a policy, that is an adjustable strategy based on revealed information, rather than a fixed path (Hall, 1986). Passengers do not know the realized scenario a priori but know the

| | $t_{r_1,s_1}$ | $t_{r_1,s_2}$ | $t_{r_2,s_1}$ | $t_{r_2,s_3}$ | $t_{r_3,s_0}$ | $t_{r_3,s_3}$ | $t_{r_4,s_2}$ | $t_{r_4,s_3}$ | $\mathbb{P}(\omega)$ |
|---|---|---|---|---|---|---|---|---|---|
| $\omega^1$ | 2 | 7 | 7 | 17 | 7 | 22 | 20 | 25 | $\frac{6}{20}$ |
| $\omega^2$ | 2 | 7 | 8 | 18 | 7 | 22 | 20 | 25 | $\frac{3}{20}$ |
| $\omega^3$ | 2 | 7 | 14 | 24 | 7 | 22 | 20 | 25 | $\frac{1}{20}$ |
| $\omega^4$ | 8 | 13 | 7 | 17 | 7 | 22 | 20 | 25 | $\frac{6}{20}$ |
| $\omega^5$ | 8 | 13 | 8 | 18 | 7 | 22 | 20 | 25 | $\frac{3}{20}$ |
| $\omega^6$ | 8 | 13 | 14 | 24 | 7 | 22 | 20 | 25 | $\frac{1}{20}$ |

Table 1: Scenario-dependent service arrival times at each station

probability mass function, and receive hints about the realized scenario over time. Although information is revealed on a continuous basis, travel decisions are only made upon arrival at a node.

Consider a setting with 6 scenarios regarding the arrival times, i.e., $\Omega = \{\omega^1, \omega^2, \ldots, \omega^6\}$. The probability mass function of the arrival times at each station is provided in the table above. As a result the 'expected' timetable corresponds to the one depicted in Figure 1(a), but the timetable as presented there is *never* realized. Particularly scenario $\omega^4$ is relevant since in this scenario the transfer from $r_1$ to $r_2$ at $s_1$ is not possible; service $r_1$ arrives after $r_2$ has departed (Figure 1(b)).

To formalize the scenario-dependent shortest path, we further introduce notations (Patriksson, 2015). A feasible *flow* for a given demand $d \in \mathbb{R}_+^{|\mathcal{W}|}$ and scenario $\omega$ is a pair of vectors $(x(\omega), v(\omega)) \in \mathbb{R}^{|E(\omega)|} \times \mathbb{R}^{|E(\omega)||\mathcal{W}|} = (x_e, e \in E(\omega); v_{e,w}, (e,w) \in E(\omega) \times \mathcal{W})$ so that

$$\Delta(\omega)v(\omega) = g,\ \Lambda(\omega)v(\omega) = x(\omega), \quad \text{and} \quad v(\omega) \geq 0.$$

We denote by $F(\omega)$ the set of all flows $(x(\omega), v(\omega))$ that satisfy above system. Here, $\Lambda(\omega)$ is the matrix so that $x_e(\omega) = \sum_{w \in \mathcal{W}} v_{e,w}(\omega)$. $\Delta(\omega) \in \{-1, 0, 1\}^{|V(\omega)| \times |E(\omega)|}$ is the node-link incidence matrix with $\Delta_{i,e}(\omega) = 1$ if node $i$ is the origin node of link $e$, $\Delta_{i,e}(\omega) = -1$ if node $i$ is the destination node of link $e$, and $\Delta_{i,e}(\omega) = 0$ otherwise. Further, $g_{i,w} = d_w$ if node $i$ is the origin of OD pair $w$, $g_{i,w} = -d_w$ if node $i$ is the destination of OD pair $w$, and $g_{i,w} = 0$ otherwise.

In the unrealistic case that the scenario $\omega$ is known beforehand, the optimization problem corresponding to finding the shortest path becomes $\omega$-dependent, i.e.,

$$LP(\omega): \quad \min_{(x(\omega),v(\omega))} z(\omega, x(\omega)) = \sum_{e \in E(\omega)} t_e(\omega)x_e(\omega) \quad \text{s.t.} \quad (x(\omega), v(\omega)) \in F(\omega).$$

In light of the scenarios, the shortest path becomes random, and the final link flows is a random pair of vectors $(x^*(\cdot), v^*(\cdot)) : \omega \mapsto (x^*(\omega), v^*(\omega))$, with $(x^*(\omega), v^*(\omega))$ a solution to $LP(\omega)$ for each $\omega \in \Omega$ [1]. A scenario-dependent shortest path cannot be chosen in practice, since knowledge about the realized scenario is not available when making boarding and routing decisions. The expected shortest path with recourse can be re-formulated as multi-stage stochastic program (Shapiro et al., 2021)

$$\min_{(x(\cdot),v(\cdot))} \mathbb{E}_\omega \left\{ \sum_{e \in E(\omega)} t_e(\omega)x_e(\omega) \right\} \quad \text{s.t.} \quad \begin{array}{l} (x(\cdot), v(\cdot)) \in F(\cdot) \\ v(\cdot) \text{ is "nonanticipative"} \end{array}$$

where $F(\cdot) = \{(x(\cdot), v(\cdot)) \mid (x(\omega), v(\omega)) \in F(\omega) \text{ for each } \omega \in \Omega\}$. We note that this is an optimization problem in a Hilbert space (i.e., a space of 'response' functions), and not in a conventional real space. The nonanticipativity constraint basically says that decisions cannot rely on information that has not yet been revealed. In our example, assuming one unit of flow is sent at $t_0$ from $s_0$ to $s_3$ this comes down to a constant random flow for links $e_{10}$ and $e_{13}$, i.e., $x_{e_{10}}(\omega) = 1$ or 0 for all $\omega$, and $x_{e_{13}}(\omega) = 1 - x_{e_{10}}(\omega)$. Note that $x^*_{e_{40}}(\omega^4) = 1$ and $x^*_{e_{40}}(\omega) = 0$ otherwise.

## 2.2 A stochastic variational inequality

We use the linear program of previous subsection to formulate a variational inequality condition that accounts for uncertainty in the supply, as well as for as increased discomfort as a function of passenger loads (e.g., in-vehicle crowding). Each arc $e$ has an accompanying flow-dependent generalized

[1] Note that we slightly abuse notation here, since the edge set is scenario dependent. We can circumvent this issue by expanding the space for each scenario, and subsequently add constraints to $F(\omega)$ that force some link flows to be 0.

cost $t_e(\cdot, x_e(\cdot)) : \omega \mapsto t_e(\omega, x_e(\omega))$ expressed as a (separable) strictly increasing continuous function of $x_e(\omega)$ modeling the increasing discomfort of travelers using arcs $e$. The cost of a path $p$ is the sum of travel costs of all edges in that path. Following our discussion, in equilibrium, passengers adopt a policy in which - upon arrival at a node - they potentially adapt their decision to proceed along the expected shortest path (in terms of generalized costs) conditional on the information available. Consequently, such an equilibrium is a solution of the stochastic variational inequality (Rockafellar and Wets, 2017).

$$\begin{bmatrix} t(\cdot, x(\cdot)) \\ 0 \end{bmatrix} \in N_{F(\cdot)\cap\mathcal{N}}(x(\cdot), v(\cdot)),$$

where $\mathcal{N}$ denotes the non-anticipativity space. $N$ is the normal cone, and since $F(\cdot)\cap\mathcal{N}$ is convex, the set of normals can be calculated through convex analysis, using the inner product between functions $x(\cdot)$ and $y(\cdot)$, understood as $\langle x(\cdot), y(\cdot)\rangle = \sum_{\omega\in\Omega} \mathbb{P}(\omega)\langle x(\omega), y(\omega)\rangle$. This information availability combined with the nonanticipativity constraint can equivalently be formulated using a filtration. As such, passengers receive hints about the scenario that is being realized over time, and can thereby adapt their decisions. In the example of Section 2.1, when passengers arrive at time $t = 2$ at node $s_1$, they know either scenario $\omega^1$, $\omega^2$ or $\omega^3$ will be realized in the end (with the accompanying conditional probabilities) .

## 3 Conclusion and outlook

This paper proposes a novel stochastic transit assignment model, assuming that supply is scenario dependent. Consequently, passengers use online information to adapt their decisions, modeled as a multi-stage stochastic program. In our approach, recourse decisions occur due to updated information regarding arrival times of PT vehicles, and the possibility of missed transfers, as well as due to probabilistic information about the adaptive choices of other passengers. The resulting flows are scenario-dependent, meaning that using the underlying probability mass function, the expected flow per link can be determined. The stochastic variational inequality can be used to numerically find equilibrium flows. We compare the resulting flows from our model with the results from other approaches.

**Acknowledgments:** This work was partially funded by the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101077049 (CONDUCTOR).

## References

Eikenbroek, O. A. L., Still, G. J., van Berkum, E. C., and Kern, W. (2018). The boundedly rational user equilibrium: a parametric analysis with application to the network design problem. *Transportation Research Part B: Methodological*, 107:1–17.

Hall, R. W. (1986). The fastest path through a network with random time-dependent travel times. *Transportation science*, 20(3):182–188.

Henrion, R. and Römisch, W. (2022). Problem-based optimal scenario generation and reduction in stochastic programming. *Mathematical Programming*, 191(1):183–205.

Jenelius, E. (2020). Personalized predictive public transport crowding information with automated data sources. *Transportation Research Part C: Emerging Technologies*, 117:102647.

Kumar, P. and Khani, A. (2023). Schedule-based transit assignment with online bus arrival information. *Transportation Research Part C: Emerging Technologies*, 155:104282.

Luan, X., Eikenbroek, O. A. L., Corman, F., and van Berkum, E. C. (2024). Passenger social rerouting strategies in capacitated public transport systems. *Transportation Research Part E: Logistics and Transportation Review*, 188:103598.

Patriksson, M. (2015). *The traffic assignment problem: models and methods*. Courier Dover Publications.

Rambha, T., Boyles, S. D., and Waller, S. T. (2016). Adaptive transit routing in stochastic time-dependent networks. *Transportation Science*, 50(3):1043–1059.

Rockafellar, R. T. and Wets, R. J. (2017). Stochastic variational inequalities: single-stage to multistage. *Mathematical Programming*, 165(1):331–360.

Shapiro, A., Dentcheva, D., and Ruszczynski, A. (2021). *Lectures on stochastic programming: modeling and theory*. SIAM.

Tirachini, A., Sun, L., Erath, A., and Chakirov, A. (2016). Valuation of sitting and standing in metro trains using revealed preferences. *Transport Policy*, 47:94–104.

# New stable responsive local gating strategies to control vehicle queues and flows in congested urban networks (Abstract for DTA 2025)

Michael J Smith[a], David Watling[b], Ronghui Liu[b], Koki Satsukawa[c], Takamasa Iryo[d], and Richard Mounce[e].
[a]Department of Mathematics, University of York, United Kingdom of Great Britain and Northern Ireland;
[b]Institute for Transport Studies, University of Leeds, 10, Woodhouse Lane, Leeds LS2 9JT, United Kingdom of Great Britain and Northern Ireland.
[c]Institute of Transdisciplinary Sciences for Innovation, Kanazawa University, Japan;
[d]Graduate School of Information Sciences, Tohoku University, Japan;
[e]RBM Traffic Solutions Ltd, 95 Princess Drive, York, YO26 5SX, UK.
Keywords: Responsive traffic signal control, Proportional Adjustment Process, Stability, Markov Chain Analysis.

## 1. Introduction

Allsop (1974) and Beckmann (1956) suggested that signal-settings should take account of drivers' route choices; so as to beneficially affect the distribution of traffic flows on an urban road network.

Smith and van Vuren (1993) compare IOA (iterative assignment and control) using $P_0$ control and the equisaturation policy. The joint dynamics of day-to-day route choice and adaptive signal control under real-time information has been considered by Hu and Mahmassani (1997). Cantarella (2010) presented a formal definition of the combined day-to-day signal control and traffic assignment problem based on a discrete time, deterministic process model. He proved fixed-point stability results. Cantarella et al. (2012) showed how equilibrium stability conditions can be embedded as a constraint in a day-to-day signal setting–route choice framework. Meneguzzer (1995, 1997, 2012) gives an interesting review and shows that the frequency of signal updating may significantly affect the duration of the day-to-day dynamic process needed to achieve a network flow–control equilibrium. Yang and Yagar (1997) study assignment and control on saturated networks. Bie and Lo (2010) study the stability and attraction domains of traffic equilibria in a day-to-day dynamical system. Xiao and Lo (2015) investigated the behavior of a joint route choice–signal control dynamical system.

More recently, He et al (2022) present a discrete day-to-day signal retiming problem for fine-tuning the green splits in a single-destination traffic network to mitigate the congestion induced by travelers' day-to-day adaptation to a new signal plan. Numerical examples demonstrate that the proposed signal retiming scheme can reduce the total system travel time over the traffic equilibration period.

Meneguzzer (2024) considers adaptive traffic signal control to promote the efficient use of road intersections, but comments that "the reaction of drivers to repeatedly updated signal settings and the ensuing route choice dynamics may trigger the emergence of various kinds of network instability". Meneguzzer suggests a Logit form signal control policy to protect the system from instabilities. This paper will prove some stability results concerning certain ($P_0$) responsive traffic signal control policies and certain dynamic day-to-day models of route choice; substantially developing some of the ideas already published in Liu et al. (2015). The responsive control policies developed in this paper are carefully designed to operate a responsive local gating strategy (holding some traffic back) aiming to reduce queues in selected locations and overall. The holding back responsive control policies considered are all, under certain conditions, still capacity-maximising. Smith (1979) shows that network capacity depends on the traffic control policy employed; and so this is important. *In this abstract we give just one initial stability result in outline only: the paper will contain several extensions. In the full paper delays become more part of "control" instead of being minimised.*

Varaiya (2013) presents a different view of responsive traffic control, assuming that route choices are fixed; Varaiya suggests a control called MaxPressure.

## 2. Notation.

We consider, in a day-to-day context, the dynamics of responsive signal control and routeing on a general network. For any signal stage *S* at a junction there corresponds an *anti-stage A* at the same junction. Anti-stage *A* is the set of links at the same junction shown red when stage *S* is shown green.

Suppose now that in a general network with $N$ routes, $M$ signal stages (and so $M$ antistages) and $n$ links:
$\mathbf{X} = (X_1, X_2, \ldots, X_N)$ is the $N$-vector of all route flows $X_1, X_2, \ldots, X_N$,
$s_i$ = saturation low at the link $i$ exit,
$x_i$ = the sum of the route flows which traverse link $i$ (for $i = 1, 2, 3, \ldots, n$),
$\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is the $n$-vector of link flows $x_1, x_2, \ldots, x_n$,
$\mathbf{R} = (R_1, R_2, \ldots, R_a, \ldots, R_N)$ is the $M$-vector of all anti-stage red-times $R_1, R_2, \ldots, R_N$,
$r_i$ = the sum of the anti-stage red-times which contain link $i$ (for for $i = 1, 2, 3, \ldots, n$),
$\mathbf{r} = (r_1, r_2, \ldots, r_n)$ is the $n$-vector of red-times $r_1, r_2, \ldots, r_n$; $r_i = 0$ if link $i$ is not signalised.
$c_i(\cdot)$ is a new non-decreasing cost function of only $x_i$ (for $i = 1, 2, 3, \ldots n$),
$b_i$ represents link $i$ delay,
$\mathbf{b} = (b_1, b_2, \ldots, b_n)$ is the $n$-vector of extra (beyond $c_i(\cdot)$) added bottleneck delays $b_i$ at the exit of link $i$,
link $i$ travel time = $c_i(x_i) + b_i$,
$f_i$ = a particular *arbitrary* non-decreasing non-negative smooth function defined on the interval $[0, s_i)$.
*MAJOR INITIAL ASSUMPTION: Given the arbitrary non-decreasing non-negative smooth functions $f_i$, for all i the link i extra bottleneck delay $b_i = f_i(x_i + s_i r_i)$.*

It then follows that for all links $i$ the link $i$ travel time = $c_i(x_i) + b_i = c_i(x_i) + f_i(x_i + s_i r_i) = c_i(x_i, r_i)$.

Now we define the $N$ route travel costs $C_u$ (which will be travel times here):

$$C_u(\mathbf{X}, \mathbf{r}) = \sum_{\text{link } i \text{ belongs to route } u} [c_i(x_i) + f_i(x_i + s_i r_i)].$$

Also we put $\mathbf{C(X, r)} = [C_1(\mathbf{X, r}), C_2(\mathbf{X, r}), \ldots, C_N(\mathbf{X, r})]$. These route costs determine future route flows which are obtained by swapping route-flows from more to less costly routes following *PAP*.

We also then define the $M$ anti-stage red-time-costs $AC_a$ by:

$$AC_a(\mathbf{X}, \mathbf{r}) = \sum_{\text{link } i \text{ belongs to anti-stage } a} s_i b_i = \sum_{\text{link } i \text{ belongs to anti--stage } a} s_i f_i(x_i + s_i r_i).$$

Also put $\mathbf{AC(X, r)} = [AC_1(\mathbf{X, r}), AC_2(\mathbf{X, r}), \ldots, AC_N(\mathbf{X, r})]$. These anti-stage costs determine future anti-stage red-times which are obtained by swapping red-time from more to less costly anti-stages following a method similar to *PAP*.

## 3. $P_0$ Responsive traffic control with the arbitrary functions $f_i$

Given the arbitrary function $f_i$, our major initial assumption is that $b_i = f_i(x_i + s_i r_i)$. In order to design a signal control policy we will now be motivated by our initial assumption that $b_i = f_i(x_i + s_i r_i)$. This is the added cost (= time spent) felt by vehicles exiting link $i$, and involves red-times. We also define, for each link $i$, the *red-time cost* to be $s_i b_i = s_i f_i(x_i + s_i r_i)$. This cost is to be felt by the signal controller.

In our general network: at each junction the policy $P_0$ swaps red-time from the more red-time costly to the less red-time costly antistages as follows.

*Definition 1. The signal timing proportional $P_0$ adjustment process is as follows.*
Given a small k > 0, for all ordered pairs of anti-stages, (anti-stage a, anti-stage e), at the same junction:
the proportional $P_0$ adjustment process *($PP_0AP$)*
swaps red time $k[R_a(AC_a - AC_e)]_+$ from antistage $A_a$ to antistage $A_e$. (1)

$P_0$ is built into the red-time adjustment process *($PP_0AP$)* or (1) using the $s_i b_i$ in the definition of red-time-costs $AC_a$ in section 2. If link $i$ has no signal then $r_i = 0$ and does not change. Swapping rule (1) is similar to the route flow adjustment *PAP* in Smith (1984) but uses red-times which react to red-cost differences instead of route-flows which react to route-cost differences.

*Definition 2. The PAP proportional routeing adjustment process is as follows.*
Given a small $k > 0$, for all ordered pairs of routes, (route $i$, route $j$), joining the same OD pair:
the proportional routeing adjustment process *(PAP)*
swaps route flow $kX_i[C_i(\mathbf{X}) - C_j(\mathbf{X})]_+$ from route $i$ to route $j$. (2)

## 4. For small k and arbitrary non-decreasing smooth $f_i$, *PAP with* $PP_0AP$ is stable

### 4.1. For arbitrary non-decreasing smooth $f_i$, [C(X, R), AC(X, R)] is a gradient.

Every link in our network has link flow $x_i$, red-time $r_i$ and a smooth non-decreasing $f_i(x_i + s_i r_i)$. Define

$$V_i(x_i, r_i) = \int_0^{x_i} c_i(u)du + \int_0^{x_i + s_i r_i} f_i(u)du.$$

Then

$$\partial V_i(x_i, r_i)/\partial x_i = c_i(x_i) + f_i(x_i + s_i r_i)$$
$$\partial V_i(x_i, r_i)/\partial r_i = s_i f_i(x_i + s_i r_i).$$

So

$$grad\, V_i(x_i, r_i) = \big(c_i(x_i) + f_i(x_i + s_i r_i), s_i f_i(x_i + s_i r_i)\big).$$

Thus

$$\big(c_i(x_i) + f_i(x_i + s_i r_i), s_i f_i(x_i + s_i r_i)\big) \text{ is the gradient of } V_i.$$

Adding over $i$, it now follows that

$$\big(\mathbf{c}(\mathbf{x}) + \mathbf{f}(\mathbf{x} + \mathbf{sr}), \mathbf{sf}(\mathbf{x} + \mathbf{sr})\big) \text{ is the gradient of } V(\mathbf{x}, \mathbf{r}) = \sum_{i=1}^{n} V_i(x_i, r_i).$$

Using standard arguments, it now follows that **[C(X, R), AC(X, R)]** is a gradient.

### 4.2. For arbitrary non-decreasing smooth $f_i$, (control and routeing) adjustment (1) with (2) is stable

In a general network; it is easy to check (using the results in 4.1 above) that: if we start at a feasible $(\mathbf{X}, \mathbf{R})$, then *PAP* route-flow **X** swapping (toward cheaper routes) (2) and $P_0$ red-time **R** swapping (toward cheaper antistages (1) both reduce *V* if *k* is small enough. Thus simultaneously following (1) and (2) reduces *V*.

The paper shows, in a general network:

**A:** that if **f** is bounded then under reasonable conditions these dynamics cause $(\mathbf{X}, \mathbf{R})$ to converge to the set of those $(\mathbf{X}, \mathbf{R})$ which minimize *V*, and

**B:** that if **f** is unbounded, representing capacity limits on links; then, under reasonable conditions, dynamics (2) and (1): (a) keep delays bounded and (b) cause $(\mathbf{X}, \mathbf{R})$ to converge to the set of those $(\mathbf{X}, \mathbf{R})$ which minimize *V*. *This does not happen with standard control policies; see Smith (1979) for example.*

### 4.3. Extension to a different control policy but still using the above functions $f_i$

The above stability results also hold if the link *i* red-time cost = $s_i f_i(x_i + s_i r_i) + h_i(r_i)$, where $h_i(\cdot)$ is a new non-decreasing bounded function of only $r_i$. In this case we have the $P_\mathbf{h}$ *adjustment process* $PP_\mathbf{h}AP$, instead of $PP_0AP$. Again here the $b_i$ are still given by the arbitrary continuous non-decreasing delay formulae $f_i(x_i + s_i r_i)$.

## 5. Extensions to allow for more ideal and more real link delay functions

*The link functions* $f_i(x_i + s_i r_i)$ *have been arbitrarily chosen.* (The only requirements are that they are continuous and non-decreasing.) It follows that:

(A) typically real delays $b_i$ will be very different to the above arbitrarily chosen $f_i(x_i + s_i r_i)$ and

(B) we have proved our stability results above for a *very wide variety* of arbitrary (smooth, non-decreasing) functions $f_i(x_i + s_i r_i)$. There are two directions which arise from A and B. These are:

1: to more carefully select the $f_i(x_i + s_i r_i)$ so as to more closely approximate real delay functions, and

2: to more carefully select the $f_i(x_i + s_i r_i)$ from the very wide range of functions available to help control traffic flows and queues most beneficially.

These two are conflicting directions; in the paper we show how to exploit *both*. We use *more real (MR) delay functions* and also *use also the more ideal (MI) (or more beneficial) delay functions.* So now:

A. We choose the functions $f^{MR}{}_i(x_i + s_i r_i)$ to best or better represent reality; these are now fixed;

B. We choose, by some unspecified method, more ideal function $f^{MI}_i\,(x_i + s_i r_i)$, *which would if imposed benefit the network most.*

**5.1. More real delay functions $f^{MR}_i\,(x_i + s_i r_i)$ and more ideal functions $f^{MI}_i\,(x_i + s_i r_i)$.**

For link *i* such possible different delay functions, solid curve (more ideal) and dotted curve (more real) are illustrated in figure 1. Now we seek to control the real network (with the dotted more real delay curve) using the solid more ideal curve as a "target" delay curve. Like this we aim not only retain the stability shown in section 4 but also aim to cause routeing / control adjustments which reduces queueing and other measures of network performance.

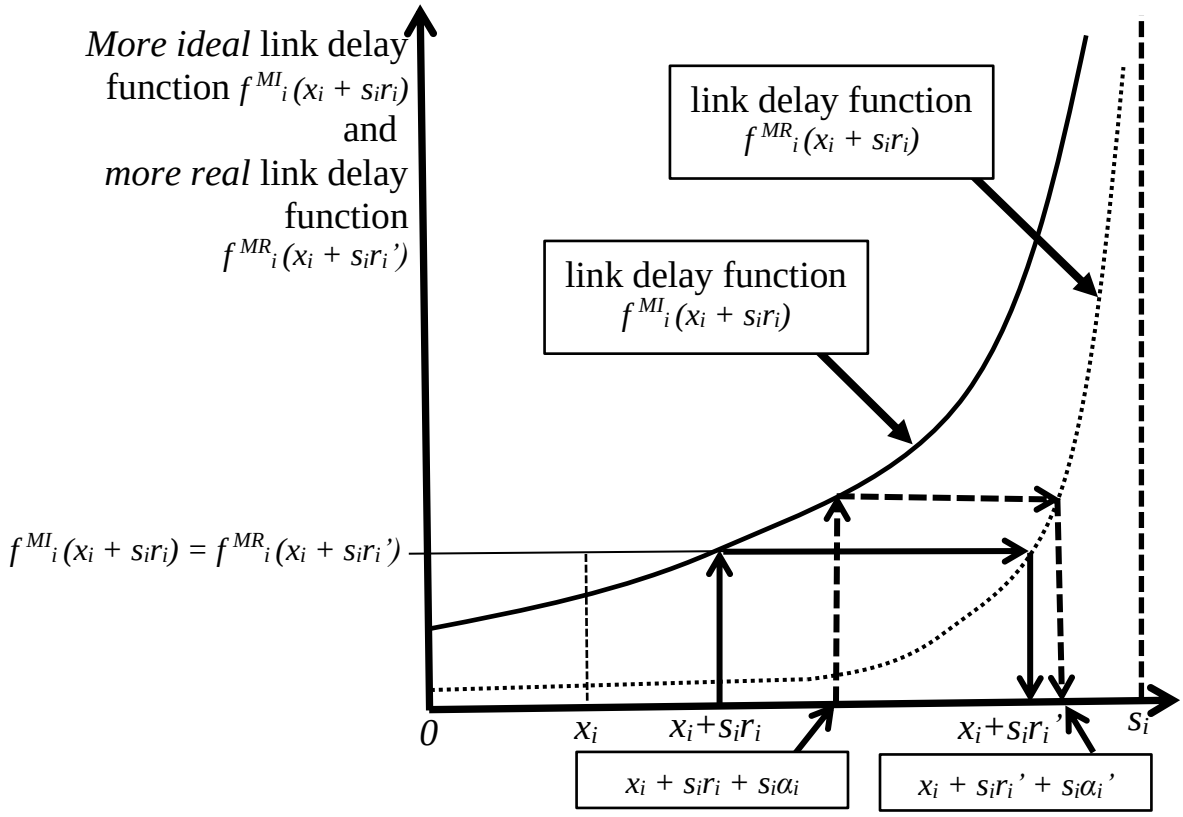



Figure 1. More ideal (solid) and more real (dotted) link delay functions. Connection between $r_i$ and $r_i'$.

Here our more real link *i* delay functions $f^{MR}_i$ still depend on $(x_i + s_i r_i)$. In this simple "different cost function shape only" case the choice of red-time re-allocation is to be a modified form of rule (1) in definition 1; the simple modification takes careful account of the more ideal functions $f^{MI}_i(x_i + s_i r_i)$ above, which we have been free to choose, *and* new different more "real" delay formulae $f^{MR}_i(x_i + s_i r_i)$.

*We here outline in this abstract the particular case where real delays are STILL given by formulae* $f^{MR}_i\,(x_i + s_i r_i)$ *depending only on* $(x_i + s_i r_i)$. See figure 1. It is natural to assume that each new more real delay function $f^{MR}_i\,(x_i + s_i r_i)$ is smaller than the more ideal delay $f^{MI}_i\,(x_i + s_i r_i)$ which we have already chosen. $f^{MR}_i\,(x_i + s_i r_i)$ may be *far* smaller than our previously chosen more ideal $f^{MI}_i\,(x_i + s_i r_i)$. *See figure 1.* Figure 1 shows both a reasonable $f^{MR}_i\,(x_i + s_i r_i)$ and a reasonable $f^{MI}_i\,(x_i + s_i r_i)$. Now we determine $r_i$' by following the three arrows (up, right, down) toward the bottom of figure 2. Generally, $r_i' > r_i$. Following the arrows in this way guarantees that $r_i'$ satisfies: $f^{MI}_i\,(x_i + s_i r_i) = f^{MR}_i\,(x_i + s_i r_i')$. Thus we may control the network with the more real cost functions $f^{MR}_i\,(x_i + s_i r_i')$, by choosing $r_i$' so that $r_i$ follows (1) above. Then provided $x_i$ follows (2) we retain stability shown in section 4.

In more detail, if we wish (in figure 1) to (say) increase $r_i$ to $r_i + \alpha_i$ then we would increase to $r_i' + \alpha_i'$ in the real network as shown in figure 1 where

*$(x_i + s_i r_i)$ determines $(x_i + s_i r_i')$ and $(x_i + s_i r_i + s_i \alpha_i)$ determines $(x_i + s_i r_i' + s_i \alpha_i')$.*
*So we may control the $r_i$ in accordance with (1) and still have stability even though the more real cost functions $f_i = f^{MR}_i$ (dotted) are not at all the same as the more ideal delay functions $f^{MI}_i$.* See figure 1.
*The full paper does this with more general delay formulations.*

## 6. A simple comparison

In figure 2, initially links 1, 2, 3 and 4 have the more real delay function $f^{MR}_i(x_i + s_i r_i)$. We consider on the figure 2 network the effects of swapping link 1 delay function from a reasonable $f^{MR}_i(x_i + s_i r_i)$ to a reasonable $f^{MI}_i(x_i + s_i r_i)$. It will be shown in the full paper, over a range of demands, that this switch substantially reduces vehicle flows and vehicle occupancies on link 1 and 4, increasing flows on links 2 and 3. This is achieved while

(A) still maximising network capacity and

(B) still retaining the stability shown in section 4.

The benefits estimated on links 1 and 4 arise because we use $f^{MI}_1(x_1 + s_1 r_1)$ instead of $f^{MR}_1(x_1 + s_1 r_1)$.
*Of course swapping the delay formula is achieved by following the process outlined in section 5.*

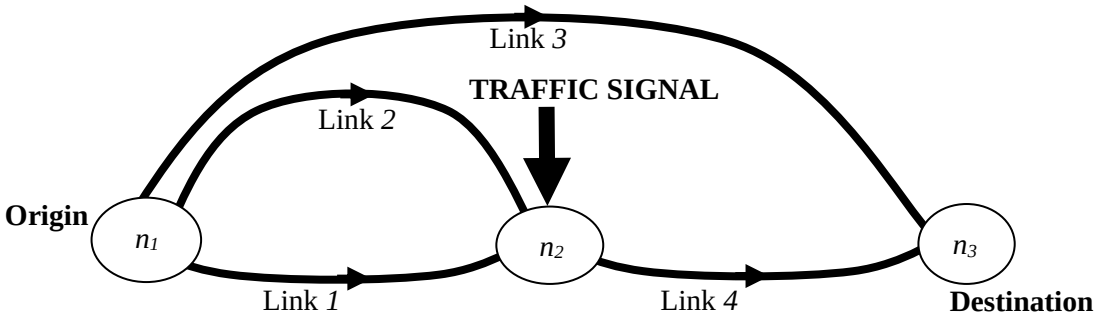



Figure 2. A signal-controlled network with an origin at $n_1$ and a destination at $n_3$.

**References.**


Allsop, R.E., 1974. Some possibilities of using traffic control to influence trip distribution and route choice. In: Proceedings of the 6th International Symposium on Transportation and Traffic Theory. pp. 345–374.

Beckmann, M., McGuire, C. B., and Winsten, C. B. (1956). *Studies in the Economics of Transportation*. Yale University Press.

Bie, J.; Lo, H.K., 2010. Stability and attraction domains of traffic equilibria in a day-to-day dynamical system formulation. Transp. Res. B. 44, 90–107.

Cantarella, G.E. 2010. Signal setting with dynamic process assignment. In New Developments in Transport Planning: Advances in Dynamic Traffic Assignment; Tampère, C.M.J., Viti, F., Immers, L.H., Eds.; Edward Elgar: Northampton, MA, USA, 2010; pp. 29–56.

Cantarella, G.E.; Velonà, P.; Vitetta, A., 2012. Signal setting with demand assignment: Global optimization with day-to-day dynamic stability constraints. J. Adv. Transp. 46, 254–268.

He, X., Wang, J., Peeta, S., Liu, H.X., 2022. Day-to-Day Signal Retiming Scheme for Single-Destination Traffic Networks Based on a Flow Splitting Approach. Networks and Spatial Economics 22, 855–882 https://doi.org/10.1007/s11067-022-09566-9

Hu, T.-Y., Mahmassani, H.S., 1997. Day-to-day evolution of network flows under real-time information and reactive signal control. Transp. Res. C. 5, 51–69.

Liu, R., Mounce, R., Smith, M.J., 2015. Route choice and signal control; a study of the stability and instability of traffic signal-controlled networks. Transp. Res. B 77, 123–145.

Meneguzzer, C. 1996. Computational experiments with a combined traffic assignment and control model with asymmetric cost functions. In Proceedings of the 4th International Conference on Applications of Advanced Technologies in Transportation Engineering, Capri, Italy, 27–30 June 1995. Published in the Proceedings: Editors Stephanedes, Y.J., Filippi, F., Eds.; ASCE: New York, NY, USA, 1996; pp. 609–614.

Meneguzzer, C. 1997. Review of models combining traffic assignment and signal control. J. Transp. Eng. 123, 148–155.

Meneguzzer, C. 2012. Dynamic process models of combined traffic assignment and control with different signal updating strategies. J. Adv. Transp., 46, 351–365.

Meneguzzer, C., 2024. Stability of Traffic Equilibria in a Day-to-Day Dynamic Model of Route Choice and Adaptive Signal Control. Appl. Sci. 2024, 14, 1891. https://doi.org/ 10.3390/app14051891

Smith, M.J., 1979. Traffic control and route-choice; a simple example. Transp. Res. 13B, 289–294.

Smith, M.J., 1984. The stability of a dynamic model of traffic assignment– an application of a method of Lyapunov. Transportation Science 18 (3), 245–252.

Smith, M.J.; van Vuren, T., 1993. Traffic equilibrium with responsive traffic control. Transp. Sci. 27, 118–132.

Varaiya, P., 2013. Max pressure control of a network of signalised intersections. Transp. Res. C 57, 177–195.

Xiao, L.; Lo, H.K., 2015. Combined route choice and adaptive traffic control in a day-to-day dynamical system. Netw. Spat. Econ. 15, 697–717.

Yang, H.; Yagar, S. 1995. Traffic assignment and signal control in saturated road networks. Transp. Res. A, 29, 125–139.

# Pricing is All You Need to Improve Traffic Routing

Yu Tang[1], Kaan Ozbay[1], and Li Jin[2]

[1]C2SMARTER Center, Department of Civil & Urban engineering, Tandon School of Engineering, New York University, 6 Metrotech Center, Brooklyn, New York, 11201, United States
[2]UM Joint Institute and Department of Automation, Shanghai Jiao Tong University, 800 Dongchuan Lu, Shanghai, 200240, China

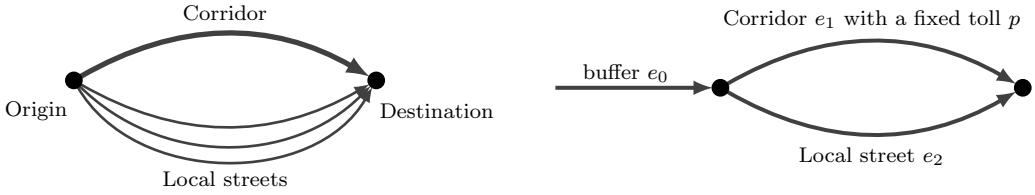
## 1 Introduction

### 1.1 Motivation

Traffic routing control aims at reducing congestion via providing drivers with route guidance. Nevertheless, it has been reported that driver non-compliance with routing instructions could undermine the performance of this management strategy [1], especially social routing advice that deliberately detours part of vehicles to achieve benefits in terms of road networks [2]. Besides, our previous paper showed in a theoretical manner that driver non-compliance could destabilize routed traffic systems [3].

Fortunately, it is promising to secure traffic routing control via pricing strategies. This is because monetary costs also play an important role in route choices [4]. Indeed, applying joint routing and pricing polices is not a new idea [5]. However, to the best of our knowledge, few studies have conducted an analytical analysis of these management approaches, particularly considering stochastic driver compliance influenced by congestion and tolls.

In this paper, we investigate the design of pricing policies that enhance driver adherence to route guidance, ensuring effective routing control. The major novelty lies in that we adopt a Markov chain to model drivers' compliance rates conditioned on both traffic states and tolls. By formulating the managed traffic network as a nonlinear stochastic dynamical system, we can quantify in a more realistic way the impacts of driver route choices and thus determine appropriate tolls. Specially, we focus on a network comprised of two parallel links; see Figure 1. Though simple, the two-parallel-link network serves as a typical scenario for studying routing control; it turns out to be an appropriate abstraction of multiple parallel links: one stands for an arterial and the other denotes a set of local streets [6]. We assume that a reasonable routing policy is specified in advance, which means both the corridor $e_1$ and the local street $e_2$ are fully utilized if drivers completely obey the routing control. However, drivers could be reluctant to be detoured to link $e_2$. Thus a fixed toll $p$ is set on the corridor $e_1$ to give drivers incentives to choose the local street.




(a) A network consisting of one corridor and multiple local streets.

(b) An abstract of networks consisting of parallel links.

Figure 1: Modeling parallel-link networks.

### 1.2 Our contributions

We try to address two main questions:

(i) How to determine whether the network can be stabilized by routing and pricing strategies, subject to driver non-compliance?

(ii) How to find the optimal pricing strategy that maximizes the network throughput, given a routing policy?

The first question aims to assess the effectiveness of the given routing and pricing policies, where instability signals inadequate traffic management. To address this, we derive a stability condition (Theorem 1) using the Foster-Lyapunov criterion [7] and an instability condition (Theorem 2) based on the transience of Markov chains [7].

It should be pointed out that the exact throughput cannot be determined, even for a simple two-parallel-link network, due to the randomness of driver non-compliance. To address the second question, we suggest using the stability and instability conditions to establish lower and upper bounds on throughput. This allows us to select suitable tolls that maximize these bounds.

## 2 Problem Statement

In this section, we first present our model formulation. Then we introduce the formal definitions of stability and throughput, which are closely related to our research problems.

### 2.1 Model formulation

The considered network comprises links $e_0$, $e_1$ and $e_2$, as shown in Figure 1(b). Each link $e \in \{e_0, e_1, e_2\}$ is characterized by a length $l_e$ and a state of traffic density $x_e(t)$ at time $t$. Particularly, link $e_0$ serves a buffer receiving the upstream demand $D(t) \in \mathcal{D}$. We assume that $D(t)$ is governed by a stationary stochastic process with $\mathbb{E}[D(t)] = \bar{D}$. Following the convention [8], the buffer is assumed to have infinite storage, indicating $x_{e_0}(t) \in [0, \infty)$. Link $e_0$ is also associated with a bounded and non-decreasing sending flow $f_{e_0} : [0, \infty) \to [0, Q_{e_0}]$, where $Q_{e_0}$ denotes the capacity. We define its *critical density* as

$$x_{e_0}^c := \inf\{x | f_{e_0}(x) = Q_{e_0}\}, \tag{1}$$

which represents the lowest traffic density at which the sending flow $f_{e_0}$ attains its capacity. As for link $e \in \{e_1, e_2\}$, it only has finite storage such that $x_e \in [0, x_e^{\max}]$ where $x_e^{\max}$ is interpreted as the *jam density*. In addition to the bounded and non-decreasing sending flow $f_e : [0, x_e^{\max}] \to [0, Q_e]$, link $e \in \{e_1, e_2\}$ also has a bounded and non-increasing receiving flow $r_e : [0, x_e^{\max}] \to [0, Q_e]$. For notational convenience, we let $\mathcal{X} := [0, \infty) \times [0, x_{e_1}^{\max}] \times [0, x_{e_2}^{\max}]$ and $x := [x_{e_0}, x_{e_1}, x_{e_2}]^{\mathrm{T}} \in \mathcal{X}$.

For traffic management strategies, we first consider a routing policy, denoted by $\alpha : \mathcal{X} \to [0, 1]$, specifying the proportion of vehicles assigned onto link $e_1$. The routing policy $\alpha(x)$ is assumed to be non-increasing with respect to $x_{e_1}$ and non-decreasing with respect to $x_{e_2}$. Besides, a fixed toll $p \geq 0$ is set on link $e_1$ for each passing vehicle. Then, we model drivers' response to the traffic management. Particularly, we denote by $C_e(x, p)$ the compliance rate regarding the routing instruction to link $e \in \{e_1, e_2\}$, which generally depends on the traffic state $x$ and the fee $p$. We consider that $C(x, p) := [C_{e_1}(x, p), C_{e_2}(x, p)]^{\mathrm{T}}$ is a random vector conditioned on $x$ and $p$, with a distribution $\Gamma_{x,p}$ supported on $\mathcal{C} \subseteq [0, 1]^2$. We also assume i) that $\mathbb{E}[C_{e_1}(x, p)]$ is non-increasing with respect to $x_{e_1}$ and $p$, and non-decreasing with respect to $x_{e_2}$, and ii) that $\mathbb{E}[C_{e_2}(x, p)]$ is non-decreasing with respect to $x_{e_1}$ and $p$, and non-increasing with respect to $x_{e_2}$.

The conservation law yields the following dynamics:

$$x_{e_0}(t+1) = x_{e_0}(t) + \frac{\delta_t}{l_{e_0}} \Big( D(t) - q_{e_1}(x, p, C) - q_{e_2}(x, p, C) \Big), \tag{2a}$$

$$x_e(t+1) = x_e(t) + \frac{\delta_t}{l_e} \Big( q_e(x, p, C) - f_e(x_e) \Big), \ e \in \{e_1, e_2\}, \tag{2b}$$

where $\delta_t$ is the time step size, and the flow from link $e_0$ to link $e \in \{e_1, e_2\}$, denoted by $q_e(x, p, C)$, is given below.

$$q_{e_1}(x,p,C) = \min\Big\{\Big(\alpha(x)C_{e_1}(x,p) + (1-\alpha(x))(1-C_{e_2}(x,p))\Big) f_{e_0}(x_{e_0}), r_{e_1}(x_{e_1})\Big\}, \tag{3a}$$

$$q_{e_2}(x,p,C) = \min\Big\{\Big(\alpha(x)(1-C_{e_1}(x,p)) + (1-\alpha(x))C_{e_2}(x,p)\Big) f_{e_0}(x_{e_0}), r_{e_2}(x_{e_2})\Big\}. \tag{3b}$$

Clearly, we obtain a nonlinear stochastic system (2a)-(2b) that is a Markov chain.

Now we briefly discuss how the toll $p$ influences the inter-link flows $q_{e_1}(x, p, C)$ and $q_{e_2}(x, p, C)$. As mentioned in previous section, this paper considers that drivers may resist being redirected to local streets. When the toll $p$ is low, the compliance rate $C_{e_1}(x, p)$ is high but $C_{e_2}(x, p)$ could be low, consequently compromising the routing policy $\alpha$. However, extremely high tolls may render low $C_{e_1}(x, p)$ and high $C_{e_2}(x, p)$, which also nullifies traffic routing.

### 2.2 Stability and throughput

The following gives the definition of stability considered in this paper.

**Definition 1** (Stability & Instability)**.** A stochastic process $\{Y(t) : t \geq 0\}$ with a state space $\mathcal{Y}$ is *stable* if there exists a scalar $Z < \infty$ such that for any initial condition $y \in \mathcal{Y}$

$$\limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t} \mathbb{E}[|Y(\tau)||Y(0) = y] \leq Z, \tag{4}$$

where $|Y(\tau)|$ denotes 1-norm of $Y(\tau)$. The network is *unstable* if there does not exist $Z < \infty$ such that (4) holds for any initial condition $y \in \mathcal{Y}$.

The stability above is widely used in studying traffic control [9]. It indicates that the time-average traffic density is bounded in the long term. Obviously, in practice one is more concerned about traffic performance within finite time (e.g. peak hours). Although Definition 1 simplifies the analysis of real-world traffic systems, our later numerical examples illustrate that methods based on this definition are sufficient to produce insightful results for evaluating and designing management strategies. Moreover, this establishes a foundation for future research on refined finite-time stability [10].

The *throughput* $\bar{D}^{\alpha,p}$ of the network, given the routing policy $\alpha$ and the toll $p$, is defined as the maximal expected demand that the network can accept while maintaining stability:

$$\bar{D}^{\alpha,p} := \sup \bar{D} \quad \text{subject to the system (2a)-(2b) is stable.}$$

Our research problems are i) how to verify whether the system (2a)-(2b) under the routing and pricing policies satisfies the condition (4), and ii) how to select appropriate $p$ to maximize the throughput.

## 3 Major Results

In this section, we present both theoretical and numerical results. We begin by introducing theorems that establish stability and instability conditions, followed by their application in stability verification. Next, we provide examples illustrating the alignment of our theorems with numerical simulations. We also discuss the insights gained from our proposed methods.

### 3.1 Stability & instability conditions

We present two theorems below. Theorem 1 states one stability condition, derived using the Foster-Lyapunov criterion [7], while Theorem 2 provides one instability condition, based on the transience property of Markov chains [7].

**Theorem 1.** *The system* (2a)-(2b) *is stable if there exists a vector* $\theta := [\theta_{e_1}, \theta_{e_2}]^{\mathrm{T}} \in [0,1]^2$ *and a negative scalar* $\gamma < 0$ *such that*

$$\bar{D} - \sum_{e\in\{e_1,e_2\}} (1-\theta_e)\mathbb{E}[q_e(x,p,C)] - \sum_{e\in\{e_1,e_2\}} \theta_e f_e(x) < \gamma,\ \forall x \in \{x \in \mathcal{X} | x_{e_0} = x^c_{e_0}\}, \tag{5}$$

*where* $x^c_{e_0}$ *is given by* (1) *and* $\mathbb{E}[q_e(x,p,C)] := \int_{\mathcal{C}} q_e(x,p,c))\Gamma_{x,p}(\mathrm{d}c)$.

**Theorem 2.** *The system* (2a)-(2b) *is unstable if there exists a vector* $\theta := [\theta_{e_1}, \theta_{e_2}]^{\mathrm{T}} \in [0,1]^2$ *and a non-negative scalar* $\gamma \geq 0$ *such that*

$$\bar{D} - \sum_{e\in\{e_1,e_2\}} (1-\theta_e)\mathbb{E}[q_e(x,p,C)] - \sum_{e\in\{e_1,e_2\}} \theta_e f_e(x) \geq \gamma,\ \forall x \in \{x \in \mathcal{X} | x_{e_0} = x^c_{e_0}\}. \tag{6}$$

Theorem 1 (resp. Theorem 2) essentially says that the network is stable (resp. unstable) if the weighted expected net flow is negative (resp. non-negative) over the traffic state space $\{x \in \mathcal{X} | x_{e_0} = x^c_{e_0}\}$. One can implement Theorem 1 by solving the following Semi-Infinite Programming (SIP [11]):

$$(P_1)\ \min_{\theta,\gamma}\ \gamma \text{ subject to (5)}.$$

If the optimal $\gamma^*$ is negative, the stability is concluded. Similarly, the instability verification requires solving the SIP:

$$(P_2)\ \max_{\theta,\gamma}\ \gamma \text{ subject to (6)}.$$

If the optimal $\gamma^*$ is non-negative, we say that the system (2a)-(2b) is unstable.

## 3.2 Numerical examples

The following presents settings in our numerical examples. First, the demand $D(t)$ is assumed to be uniformly distributed on $[D^{\min}, D^{\max}]$. Then, the sending and receiving flows are specified by $f_e(x) = \min\{v_e x_e, Q_e\}$ for $e \in \{e_0, e_1, e_2\}$ and $r_e(x) = \min\{R_e - w_e x_e, Q_e\}$ for $e \in \{e_1, e_2\}$, respectively. We consider a fixed routing ratio based on the link capacities $Q_{e_1}$ and $Q_{e_2}$, namely $\alpha := Q_{e_1}/(Q_{e_1} + Q_{e_2})$. The compliance rate $C_e(x,p)$ is uniformly distributed on $[\max\{\bar{C}_e(x,p) - \epsilon_e, 0\}, \min\{\bar{C}_e(x,p) + \epsilon_e, 1\}]$, where $\bar{C}_e(x,p)$ is given by

$$\bar{C}_e(x,p) := \frac{1}{1 + e^{\beta^0_e + \beta^1_e x_{e_1} + \beta^2_e x_{e_2} + \beta^3_e p}}.$$

The parameters are summarized in Table 1. Note that negative $\beta^0_{e_1}$ and positive $\beta^0_{e_2}$ indicate that drivers naturally prefer the corridor.

Table 1: Parameter settings.

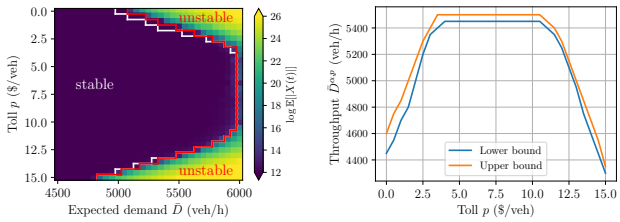
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $v_{e_0}$ | 80 (km/h) | $v_{e_1}$ | 100 (km/h) | $v_{e_2}$ | 50 (km/h) | $\beta^0_{e_1}$ | $-4$ | $\beta^0_{e_2}$ | 1 |
| $Q_{e_0}$ | 8000 (veh/h) | $Q_{e_1}$ | 4000 (veh/h) | $Q_{e_2}$ | 2000 (veh/h) | $\beta^1_{e_1}$ | 0.01 | $\beta^1_{e_2}$ | $-0.02$ |
| $D^{\min}$ | 4000 (veh/h) | $R_{e_1}$ | 4800 (veh/h) | $R_{e_2}$ | 2400 (veh/h) | $\beta^2_{e_1}$ | $-0.02$ | $\beta^2_{e_2}$ | 0.03 |
| $D^{\max}$ | [5000, 8000] (veh/h) | $w_{e_1}$ | 20 (km/h) | $w_{e_2}$ | 10 (km/h) | $\beta^3_{e_1}$ | 0.3 | $\beta^3_{e_2}$ | $-0.6$ |
| | | | | | | $\epsilon_{e_1}$ | 0.1 | $\epsilon_{e_2}$ | 0.1 |

### 3.2.1 Impacts of tolls and demands

Figure 2(a) shows the time-average traffic densities after $10^4$ steps and reveals the stability and instability regions. The white boundary is obtained from Theorem 1, while the red boundary is derived from Theorem 2. Therefore, we can conclude that the region to the left of the white boundary is stable, whereas the areas in the upper right and lower right corners are unstable. These findings are consistent with the numerical results. Note that there is a gap between the white and red boundaries, within which stability or instability cannot be determined. This is because we do not have sufficient and necessary stability conditions. In fact

the gap is not a concern, as it can be narrowed by using more advanced Lyapunov or test functions, though at the expense of increased computational cost.

The key findings from Figure 2 are summarized as follows. First, setting the toll $p$ either too low or too high can result in network instability. Second, in the case study, a toll of approximately 5 \$/veh is identified as optimal for maximizing the lower bound of throughput.



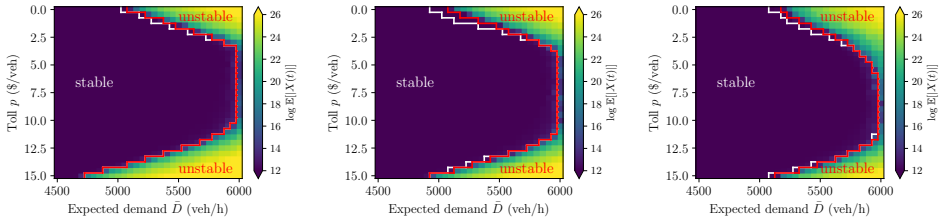

(a) Stability regions. (b) Throughput bounds.

Figure 2: Impact analysis of tolls and expected demands.

### 3.2.2 Impacts of variances of compliance rates

Figure 3 illustrates the impacts of compliance rate variances by keeping the same $\bar{C}_{e_2}(x, p)$ but selecting different $\epsilon_{e_2}$. From the upper right corners of Figures 3(a)-(c), we can see the instability regions enlarge as $\epsilon_{e_2}$ increases. This demonstrates that uncertainties in compliance rates may bring negative impacts on traffic management. From the lower right corners of Figures 3(a)-(c), it is interesting to observe that, for the same level of demand, increasing tolls can stabilize a previously unstable network as $\epsilon_{e_2}$ increases. This result is reasonable since more uncertainties indicate higher tolls to persuade drivers to choose link $e_2$.

More importantly, our white and red boundaries in Figures 3(a)-(c) capture those necessary changes. This demonstrates that our developed theorems offer practical yet powerful tools for evaluating traffic systems without the need for extensive simulations.




(a) $\epsilon_{e_2} = 0$. (b) $\epsilon_{e_2} = 0.2$. (c) $\epsilon_{e_2} = 0.4$.

Figure 3: Stability and instability regions under different $\epsilon_{e_2}$.

# 4 Future Work

This work offers several potential avenues for extension. First, more sophisticated pricing strategies, such as stepwise tolls, could be explored. Second, it would be valuable to investigate conditions under which the stability criterion is both necessary and sufficient. Third, the approach could be expanded to accommodate more complex network structures.

# References


[1] W. B. Powell, M. T. Towns, and A. Marar, "On the value of optimal myopic solutions for dynamic routing and scheduling problems in the presence of user noncompliance," *Transportation Science*, vol. 34, no. 1, pp. 67–85, 2000.

[2] M. Van Essen, O. Eikenbroek, T. Thomas, and E. Van Berkum, "Travelers' compliance with social routing advice: Impacts on road network performance and equity," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 1180–1190, 2019.

[3] Y. Tang, L. Jin, and K. Ozbay, "How does driver non-compliance destroy traffic routing control?" in *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 2023, pp. 1118–1123.

[4] K. Kerkman, T. Arentze, A. Borgers, and A. Kemperman, "Car drivers' compliance with route advice and willingness to choose socially desirable routes," *Transportation research record*, vol. 2322, no. 1, pp. 102–109, 2012.

[5] H. Yang, "Evaluating the benefits of a combined route guidance and road pricing system in a traffic network with recurrent congestion," *Transportation*, vol. 26, pp. 299–322, 1999.

[6] X. Pi and Z. S. Qian, "A stochastic optimal control approach for real-time traffic routing considering demand uncertainties and travelers' choice heterogeneity," *Transportation Research Part B: Methodological*, vol. 104, pp. 710–732, 2017.

[7] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[8] C. F. Daganzo, "The cell transmission model, part ii: network traffic," *Transportation Research Part B: Methodological*, vol. 29, no. 2, pp. 79–93, 1995.

[9] S. Barman and M. W. Levin, "Throughput properties and optimal locations for limited deployment of max-pressure controls," *Transportation Research Part C: Emerging Technologies*, vol. 150, p. 104105, 2023.

[10] Y. Hong, Z.-P. Jiang, and G. Feng, "Finite-time input-to-state stability and applications to finite-time control design," *SIAM Journal on Control and Optimization*, vol. 48, no. 7, pp. 4395–4418, 2010.

[11] O. Stein, "How to solve a semi-infinite optimization problem," *European Journal of Operational Research*, vol. 223, no. 2, pp. 312–320, 2012.

**Extended ABSTRACT for 9th International Symposium on Dynamic Traffic Assignment (DTA2023)**

# A Deep Learning Framework for Path-Level User Equilibrium Estimation in Traffic Networks

Van Anh Le[a, b], Mostafa Ameli[a, c, *] and Alexander Skabardonis [c]
[a] *University Gustave Eiffel, COSYS, GRETTIA, Paris, France*
[b] *Leonard de Vinci Engineering School, ESILV, Paris, France*
[c] *Institute of Transportation Studies, University of California, Berkeley, USA*
[*] *Corresponding author, Email: mostafa.ameli@univ-eiffel.fr*

**Keywords:** Traffic Assignment, Transformer, Network equilibrium, Traffic flow prediction, Path-based analysis

## 1 INTRODUCTION

The Traffic Assignment Problem (TAP) determines traffic flow across a network based on OD demand and capacity constraints (Yosef Sheffi. Prentice-Hall, 1986). It is traditionally solved using mathematical optimization under the User Equilibrium (UE) principle, assuming drivers have perfect information and act rationally (WARDROP, 1952). While this may not fully reflect reality, it provides reliable solutions (Jafari *et al.*, 2017). However, solving large networks is computationally expensive, as complexity increases non-linearly with OD pairs and feasible paths. Any change in demand or network structure requires re-solving, underscoring the need for an alternative approach to estimating new solutions efficiently.

UE is typically computed using optimization methods like the gradient method or fixed-point algorithms (Liu *et al.*, 2023; Ameli *et al.*, 2020). With advancements in machine learning, prediction offers a promising alternative to direct optimization. Supervised learning can be used to learn and predict optimal solutions for new scenarios. Recent studies have estimated UE link flows using machine learning (Liu & Meidani, 2024; Rahman & Hasan, 2023), but this study takes it further by directly predicting path flow from a user's perspective. Since extracting path flow from link flow is a complex equilibration problem with no unique solution, predicting equilibrium path flow directly provides better insights into demand propagation.

Deep neural networks (DNNs) model complex patterns across domains, including transportation (Wang *et al.*, 2022). Transformer, a state-of-the-art deep learning model, excels in sequence modeling tasks (Wen *et al.*, 2023). This study employs a Transformer-based architecture to estimate equilibrium path flow in multi-class traffic networks and explores its use in "what-if" analyses to assess network and demand changes.

The key research question is whether path flow distribution under UE can be accurately predicted at the network level. Since many methods fail to capture OD correlations, Transformer offers a promising alternative.

## 2 METHODOLOGY

Following the UE principle, the shortest path is the path that has the minimum cost. In the UE solution, all used paths must be at the minimum possible cost (or equal the shortest path). Mathematically, it is equivalent to the following conditions:

$$f_p^r(c_p^r - u_r) = 0 \quad \forall p \in \mathbf{P}_r, r \in R \tag{1}$$

$$c_p^r - u_r \geq 0 \quad \forall p \in \mathbf{P}_r, r \in R \tag{2}$$

$$\sum_{\forall p \in \mathbf{P}_r} f_p^r = x_r \quad \forall r \in R \tag{3}$$

$$f_p^r \geq 0, c_p^r \geq 0, u^r \geq 0 \quad \forall p \in \mathbf{P}_r, r \in R \tag{4}$$

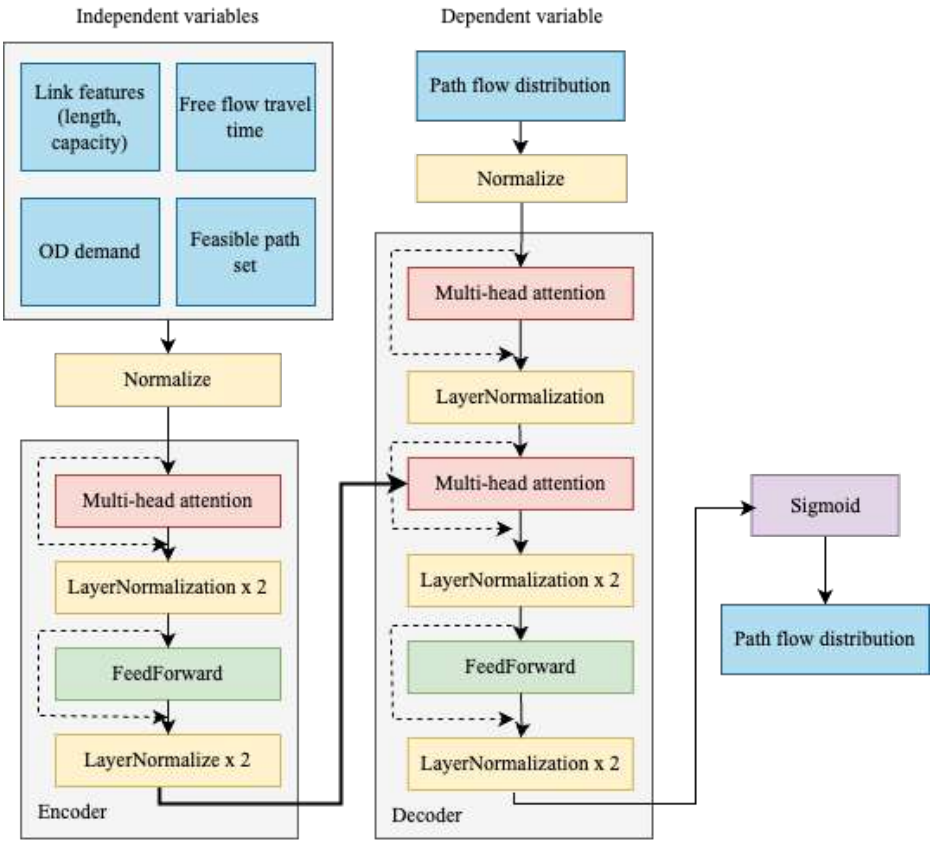



Figure 1: Model architecture

Table 1: Notation and definitions

| Notation | Description |
|---|---|
| $n$ | Number of classes in the network |
| $\mathbf{P}_r \in \mathbf{P}^k$ | Feasible path set containing $k$ feasible paths for each OD pair $r \in R$ |
| $\mathbf{X}$ | OD demand matrix |
| $x_r \in \mathbf{X}$ | Demand of OD pair $r$ |
| $f_p^r \in \mathbf{F}$ | Path flow of OD pair $r$ on path $p$ |
| $c_p^r$ | Cost of path $p$ for OD pair $r$ |
| $u^r$ | Minimum path travel cost of OD pair $r$ |
| $\mathbf{F}^*$ | Optimal path flow solution |
| $\widehat{\mathbf{F}}^*_{G,\mathbf{X}}$ | Predicted path flow distribution for graph $G$ with demand $\mathbf{X}$ |

The model proposed in this study aims to learn a function $\mathcal{F}(.)$ that maps OD demand matrix $\mathbf{X}$ and feasible path set $\mathbf{P}$ in graph $G$, to optimal path flow distribution $\mathbf{F}^*$, defined as:

$$\mathcal{F}(G, \mathbf{X}, \mathbf{P}) = \widehat{\mathbf{F}}^*_{G,\mathbf{X}} \tag{5}$$

To effectively learn the graph, all features must be on the same scale. Since the raw data is not normalized, we preprocess it by encoding raw features into a lower-dimensional representation. The input tensor is composed of three key tensors: (1) Graph information tensor – Stores link features, including length, capacity and free-flow travel time; (2) OD demand tensor – Represents the demand matrix for each class; (3) Feasible paths tensor – Encodes feasible paths for each OD pair. These tensors are stacked and normalized before being fed into the Encoder

block. The Encoder consists of two main components: a multi-head attention mechanism and a feed-forward neural network, both followed by residual connections and two layers of normalization. The Decoder generates the output path flow distribution by attending to both the previous Decoder layer's output and the Encoder's output. In Transformer architectures, the attention mechanism within both the Encoder and Decoder enables the model to selectively focus on relevant values while filtering out less meaningful ones (Vaswani *et al.* (2017)).

The model is trained using the Mean Squared Error (MSE) loss function and optimized with the Adam algorithm

To evaluate the performance of the model, we calculate the difference of average delay (AD) of the network between the predicted flow and the solution obtained from the optimizer as follows:

$$AD = \frac{\sum_{\forall r} \sum_{\forall p} \hat{f}_p^r (c_p^r - u^r)}{\sum_{\forall r} x_r} \tag{6}$$

# 3 NUMERICAL EXPERIMENT

We evaluate the proposed model's generalization ability on a Manhattan-like network by generating a 25-node, 80-link grid network. Link attributes, including capacity, length, and free-flow travel time, are randomly assigned based on a uniform distribution. The training process demonstrates smooth convergence after 80 epochs with minimal fluctuations. Training takes 72 minutes, whereas solving the UE problem using an optimizer requires 210 minutes. The model predicts path flows for a single OD demand matrix in just 0.001 seconds, making it approximately 5000 times faster than the Gurobi optimizer, which takes 5 to 7 seconds per OD matrix. Table 2 presents the model's prediction performance under varying OD demand missing ratios.

Table 2: Model performance under different OD demand missing ratios

| **Indicator** | **Missing ratio = 30%** | | **Missing ratio = 40%** | |
|---|---|---|---|---|
| | **Link flow** | **Path flow** | **Link flow** | **Path flow** |
| MAE | 236.61 | 10.95 | 307.01 | 18.54 |
| MAPE (%) | 2.56 | 4.50 | 4.35 | 5.57 |
| Average path cost (mins) | 216.62 | | 387.51 | |
| Predicted average delay (mins) | 2.64 | | 5.34 | |
| Delay percentage (%) | 1.22 | | 1.38 | |
| AD difference (mins) | 2.62 | | 4.61 | |

Building on the model's performance with the Manhattan-like network, we assess its effectiveness on an urban transportation network (Sioux Falls network). Table 3 presents the model's prediction performance at 30% of OD demand missing, across different scenarios, including no missing links, 5% missing links, and 10% missing links. The results indicate that while the model maintains the ability to predict path flow distribution even with missing links, its accuracy is highest when all links are available.

Table 3: Model performance with Sioux Falls network under different link missing ratios

| **Indicator** | **Missing ratio = 0%** | | **Missing ratio = 5%** | | **Missing ratio = 10%** | |
|---|---|---|---|---|---|---|
| | **Link flow** | **Path flow** | **Link flow** | **Path flow** | **Link flow** | **Path flow** |
| MAE | 256.27 | 5.02 | 425.32 | 6.93 | 522.35 | 7.37 |
| MAPE (%) | 1.61 | 2.25 | 5.54 | 2.98 | 6.87 | 3.07 |
| Predicted average delay (mins) | 0.71 | | 1.96 | | 2.15 | |
| Delay percentage (%) | 0.79 | | 1.52 | | 1.69 | |
| AD difference (mins) | 0.65 | | 1.9 | | 2.06 | |

# 4 CONCLUSION

This paper introduces a Transformer-based framework for predicting User Equilibrium (UE) solutions in the Traffic Assignment Problem (TAP). Unlike traditional optimization methods, which are computationally expensive, this approach uses deep learning for efficient and flexible predictions. While prior studies focus on link-level flow estimation, our model predicts path flows from a user perspective, offering a more detailed theoretical solution. Using an Encoder-Decoder architecture, it captures OD correlations to enhance traffic flow predictions (Figure 1). The key contribution is replacing traditional optimization with a predictive model, shifting the focus from solving UE to leveraging learned patterns for faster and more adaptable solutions. To validate the model, we test it on synthetic and real-world networks, including scenarios with incomplete OD demand and missing links. Unlike methods considering all possible equilibrium solutions, our approach prioritizes a single high-quality equilibrium based on selected criteria.

The proposed framework serves as a robust surrogate model for accelerating optimization in applications like resource allocation and infrastructure management. It reduces computational costs, adapts to network changes, and supports "what-if" analysis for transportation planning. The model generalizes across different test cases and reliably predicts path flows even with incomplete OD demand data.

# ACKNOWLEDGEMENTS

M. Ameli acknowledges support from the French ANR research project SMART-ROUTE (grant number ANR-24-CE22-7264).

# References

Ameli, Mostafa, Lebacque, Jean-Patrick, & Leclercq, Ludovic. 2020. Simulation-based dynamic traffic assignment: Meta-heuristic solution methods with parallel computing. *Computer-Aided Civil and Infrastructure Engineering*, **35**(10), 1047–1062.

Jafari, Ehsan, Pandey, Venktesh, & Boyles, Stephen D. 2017. A decomposition approach to the static traffic assignment problem. *Transportation Research Part B: Methodological*, **105**, 270–296.

Liu, Tong, & Meidani, Hadi. 2024. Heterogeneous Graph Neural Networks for End-to-End Traffic Assignment and Traffic Flow Learning. *Transportation Research Part C: Emerging Technologies*.

Liu, Zhichen, Yin, Yafeng, Bai, Fan, & Grimm, Donald K. 2023. End-to-end learning of user equilibrium with implicit neural networks. *Transportation Research Part C: Emerging Technologies*, **150**, 104085.

Rahman, Rezaur, & Hasan, Samiul. 2023. *Data-Driven Traffic Assignment: A Novel Approach for Learning Traffic Flow Patterns Using a Graph Convolutional Neural Network*.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Ł ukasz, & Polosukhin, Illia. 2017. Attention is All you Need. *In:* Guyon, I., Luxburg, U. Von, Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., & Garnett, R. (eds), *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc.

Wang, Senzhang, Cao, Jiannong, & Yu, Philip S. 2022. Deep Learning for Spatio-Temporal Data Mining: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, **34**(8), 3681–3700.

WARDROP, J G. 1952. ROAD PAPER. SOME THEORETICAL ASPECTS OF ROAD TRAFFIC RESEARCH. *Proceedings of the Institution of Civil Engineers*, **1**(3), 325–362.

Wen, Yanjie, Xu, Ping, Li, Zhihong, Xu, Wangtu, & Wang, Xiaoyu. 2023. RPConvformer: A novel Transformer-based deep neural networks for traffic flow prediction. *Expert Systems with Applications*, **218**, 119587.

Yosef Sheffi. Prentice-Hall, Englewood Cliffs. 1986. Urban transportation networks: Equilibrium analysis with mathematical programming methods. *Transportation Research Part A: General*, **20**(1), 76–77.

# Optimizing Volunteer Assignment and Routing to Service Centers in Non-Profit Food Banking Under Equity Constraints

Hiruni Niwunhella[a], Mehr Sadat Salami[a], Leila Hajibabai[a*] and Faisal Alkaabneh[b]

[a]North Carolina State University, Raleigh, North Carolina, United States
[b]American University of Sharjah, Sharjah, United Arab Emirates
* lhajiba@ncsu.edu

April 15, 2025

**Abstract**

Efficient volunteer assignment and routing are critical to the success of non-profit food banking systems, ensuring timely and equitable food access for food-insecure communities. This study introduces an optimization-based framework for volunteer assignment and routing to service centers in non-profit food banking under equity constraints. The goal is to minimize operational costs while ensuring fair workload distribution among volunteers, considering key logistical and capacity constraints. The proposed methodology formulates the problem as a Vehicle Routing Problem with Pickup and Delivery and Equity Constraints (VRPPD-e), integrating mixed-integer linear programming to optimize volunteer assignments and vehicle routing. The model accounts for volunteer availability, service center demand, vehicle capacity, and time-window constraints while incorporating fairness measures to balance workload distribution. To enhance computational efficiency, heuristic-based refinements and decomposition techniques are employed. Numerical experiments on a test network demonstrate the effectiveness of the model in balancing equity and efficiency. To further validate its real-world applicability, the framework is being expanded to a large-scale non-profit food banking system in Durham, NC, incorporating a greater number of volunteers, service centers, and operational constraints. The results will provide valuable insights for data-driven decision-making in non-profit logistics, improving food banking operations through optimized volunteer coordination.

Keywords: Volunteer Assignment, Vehicle Routing Problem with Pickup and Delivery, Equity, Column Generation, Labeling Algorithm, Lagrangian Relaxation

## 1 Introduction

With the world population growing, ensuring equitable access to food resources has become critically important since more than 820 million people suffer from hunger currently (Tkemaladze, 2025). Therefore, the food banks that are involved in organizing service locations with volunteers to serve food for food insecure people, must improve their decision making and streamline their operations. Past literature presents several studies on vehicle routing problem with pickup and delivery including time windows (VRPPD-TW). These include the study by Ropke & Pisinger (2006) where a meta-heuristic based on large neighborhood searching algorithm is used, and the exact solution technique presented by Ropke *et al.* (2007). Other efficient techniques are also used to solve VRPPD-TW such as augmented Lagrangian relaxation (Yang *et al.*, 2020) or tabu search (Qiu *et al.*, 2018). Furthermore, many studies focus on vehicle routing problem with pickup and delivery with equity (VRPPD-e) where fair resource allocation is done in the non-profit settings. These include the studies by Balcik *et al.* (2014), Rey *et al.* (2018), Nair *et al.* (2018), and Eisenhandler & Tzur (2019). Although the past literature efforts provide insights into various vehicle routing and resource allocation operations in non-profit settings,

there exists no work that jointly model fair volunteer assignment and routing considering equity, time windows and capacity constraints.

This study introduces a pioneering solution technique to VRPPD-e for non-profit food distribution. It concentrates on optimizing multiple volunteer assignments to service centers which deliver to food-insecure households via food banks, using cost-efficient routes for their pickups and drop-offs with equitable service distribution. This approach considers key constraints such as volunteers' available time windows and vehicle capacity to achieve a balanced and a fair solution, handling volunteer transportation to and from service locations efficiently.

## 2 Problem Statement

The objective of the proposed model is minimizing the cost of routing (handling efficiency) while maximizing the number of volunteers assigned across the food distribution network (dealing with effectiveness), and ensuring a fair distribution of volunteers across service centers depending on the demand (handling equity).

The three parties involved in the food distribution network are food bank, volunteers, and service locations. A service location is a program organized by the food bank through a partnering agency to distribute food to food-insecure people. Volunteers visit service locations to help food distribution and other tasks required. However, not all those volunteers would have means of traveling to and from the allocated service locations, and thus, the food bank provides transportation for those volunteers in need. The food bank uses its capacitated fleet of vehicles to transport volunteers from their homes to a service location, and back to their homes, depending on the demand at each service location ensuring equity. The food bank manages this network of volunteers and service locations by assigning volunteers, with efficient vehicle routes with predefined time windows and capacity constraints. A special requirement of this problem is the fact that the food bank has to arrange dropping off volunteers to their respective homes after completing their service, i.e., the pick-up and drop-off locations are the same (co-located), making the problem unique and more complex than the vehicle routing problem (VRP) models with classical flow conservation constraints. Figure 1 depicts an example of a feasible solution for the explained assignment-routing network optimization (for a small scale network with 1 food bank, 5 volunteers and 2 service locations). It depicts how the co-located pickup and delivery locations act in the model, where all service locations are served but not all volunteers are served.

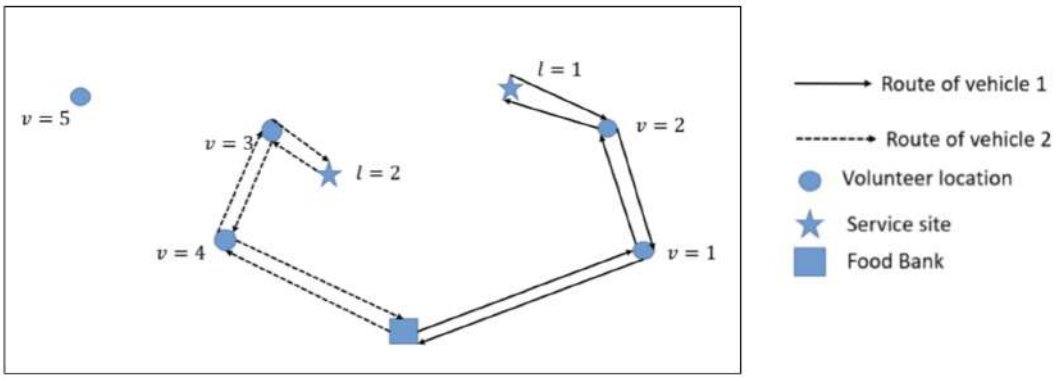



Figure 1 – *Example of assignment-routing network optimization*

## 3 Methodology

An initial approach to solve the aforementioned problem involves formulating a mixed integer linear programming (MILP) model. This basic model becomes computationally challenging when applied to large-scale instances. To address this scalability issue, the model is adapted

into a customized branch-and-price algorithm. This algorithm acts as the master problem and incorporates a Lagrangian relaxation technique to handle the equity constraints more efficiently. The main objective of the model remains threefold: minimizing the overall routing costs while simultaneously maximizing the number of volunteers and ensuring equity.

The branch-and-price algorithm refines the solution iteratively through a column generation procedure. At each iteration, a new feasible route is introduced into the master problem by identifying and adding routes with negative reduced costs. This process continues until no additional routes with negative reduced costs can be found. To accelerate the column generation, a bi-directional labeling algorithm is used for generating feasible routes, which effectively solves the shortest path pricing sub-problem. This step significantly enhances computational efficiency due to its ability to prune non-optimal paths early. It applies the strongest dominance rules in both the forward and backward directions, ensuring that only the most promising paths are remained. Once an optimal solution for the master problem is reached through the column generation process, the Lagrangian relaxation framework further refines the optimal solution. The sub-gradient method is employed to iteratively adjust the Lagrangian multipliers, progressively improving the bound on the objective function until convergence is achieved, which ultimately ensures the solution optimality.

## 4 Results

Numerical experiments on the proposed model are conducted for a small-scale network with one food bank, eight volunteers, and two service locations (a 10-node network) with five available vehicles. Preliminary results for initial model instances are presented in 1, where the average demand per service location is varied. The results illustrate how the model effectively balances equity in volunteer assignments while minimizing routing costs.

The model runtime remains significantly low due to the small problem size and the performance-boosting algorithms integrated into the solution approach. To thoroughly assess the scalability and effectiveness of the methodology, we are expanding the numerical experiments to a real-world large-scale food distribution network in Durham, NC. This extended analysis will incorporate a substantially larger number of volunteers, service locations, and operational constraints, providing a comprehensive evaluation of the model's computational efficiency and practical applicability in large-scale food rescue logistics.

Table 1 – *Preliminary results for initial instances of the model*

| **Avg demand** | **Runtime** | **Optimal cost** | **Optimal # routes** | **Avg optimal # volunteers/route** |
|---|---|---|---|---|
| 1 | 0.07 sec | $8.03 | 1 | 2 |
| 2 | 1.53 sec | $15.72 | 2 | 2 |
| 3 | 0.15 sec | $18.47 | 2 | 3 |
| 4 | 2.33 sec | $29.23 | 3 | 2.67 |
| 5 | 2.01 sec | $33.99 | 3 | 2.67 |

Figure 2 depicts the variation of optimal cost and average optimal no. of volunteers per route with the total number of nodes in the network considering the average demand per location as 3 and the available number of vehicles as 5.

## 5 Conclusion

It is vital for the food banks to adhere to performance improvement techniques for efficient and effective service distribution, with the rising levels of food insecurity across the world. This study provides a significant technique for the food distribution networks to optimally carry out

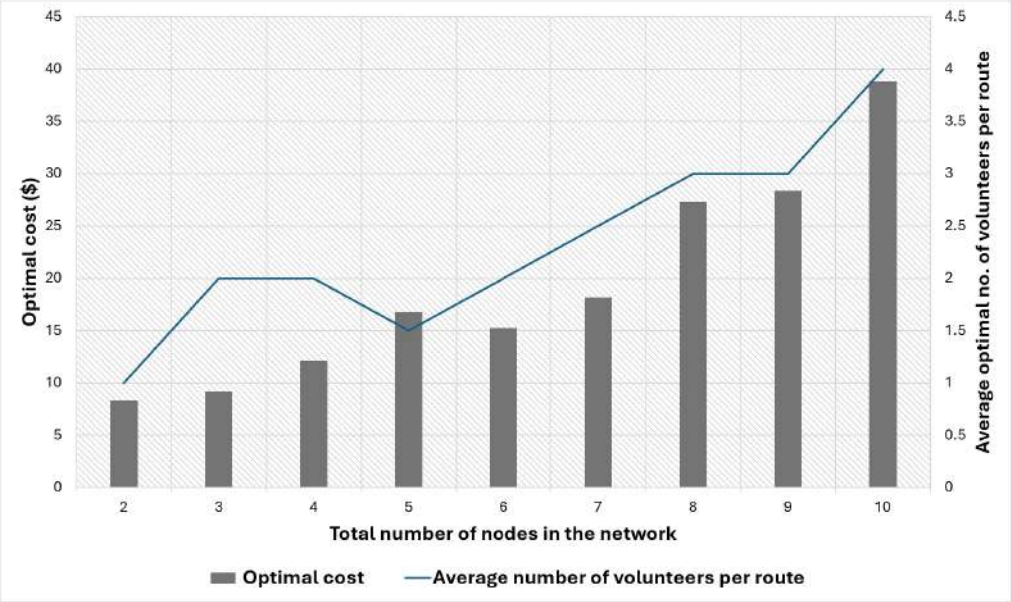



Figure 2 – *Variation of optimal cost and average optimal no. of volunteers per route with the total number of nodes*

the volunteer management and routing processes. The proposed model utilizes a novel approach to assign multiple volunteers under equity conditions, for a non-profit setting in the context of food distribution, while minimizing the routing costs. It follows the conversion of the mixed integer linear programming model to a Lagrangian relaxed model using branch-and-price algorithm. The performance is enhanced using column generation to generate the routes, where the shortest path pricing subproblem is solved using the labeling algorithm. This model depicts high computational efficiency with the aforementioned performance boosting techniques. By combining branch-and-price with column generation, the solution progressively improves linear programming Lagrangian relaxations. This integrated methodology not only redefines the VRPPD-e problems but also presents a new technique to manage workforce in non-profit settings. A promising future avenue for this study would be catering to stochastic availability of volunteers and dynamic routing in the non-profit food banking sector. Incorporating stochastic route choice models would enhance the realism of the proposed framework while introducing additional complexity, particularly in managing volunteers and their routing in real-time. Additionally, analyzing stochastic fundamental diagrams to account for the stochastic and dynamic nature of traffic – and its impact on route design – would be a valuable direction for future research.

## References

Balcik, Burcu, Iravani, Seyed, & Smilowitz, Karen. 2014. Multi-vehicle sequential resource allocation for a nonprofit distribution system. *IIE Transactions*, **46**(12), 1279–1297.

Eisenhandler, Ohad, & Tzur, Michal. 2019. A segment-based formulation and a matheuristic for the humanitarian pickup and distribution problem. *Transportation Science*, **53**(5), 1389–1408.

Nair, DJ, Grzybowska, H, Fu, Y, & Dixit, VV. 2018. Scheduling and routing models for food rescue and delivery operations. *Socio-Economic Planning Sciences*, **63**, 18–32.

Qiu, Meng, Fu, Zhuo, Eglese, Richard, & Tang, Qiong. 2018. A Tabu Search algorithm for the vehicle routing problem with discrete split deliveries and pickups. *Computers & Operations Research*, **100**, 102–116.

Rey, David, Almi'ani, Khaled, & Nair, Divya J. 2018. Exact and heuristic algorithms for finding envy-free allocations in food rescue pickup and delivery logistics. *Transportation Research Part E: Logistics and Transportation Review*, **112**, 19–46.

Ropke, Stefan, & Pisinger, David. 2006. An adaptive large neighborhood search heuristic for the pickup and delivery problem with time windows. *Transportation science*, **40**(4), 455–472.

Ropke, Stefan, Cordeau, Jean-François, & Laporte, Gilbert. 2007. Models and branch-and-cut algorithms for pickup and delivery problems with time windows. *Networks: An International Journal*, **49**(4), 258–272.

Tkemaladze, J. 2025. Concept to the Food Security. *Longevity Horizons*, **108**.

Yang, Senyan, Ning, Lianju, Shang, Pan, & Tong, Lu Carol. 2020. Augmented Lagrangian relaxation approach for logistics vehicle routing problem with mixed backhauls and time windows. *Transportation Research Part E: Logistics and Transportation Review*, **135**, 101891.

# Deep Reinforcement Learning and Active Traffic Flow Control Policies for Extending the Life of Multi-Asset Transportation Roadway Networks

Papakonstantinou, K., M. Saifullah, W. Zhou, S. Stoffels, E. Miller-Hooks (speaker)

10th International Symposium on Dynamic Traffic Assignment – Extended Abstract

February 14, 2025

## Abstract

In this paper, a state-of-the-art Deep Decentralized Multi-agent Actor-Critic reinforcement learning with Centralized Training and Decentralized Execution architecture is proposed for large-scale infrastructure maintenance and rehabilitation (M&R) activity optimization. Expanding on recent advances that capture deterioration uncertainty, a limited budget, the effects of M&R activity execution on traffic performance, multiple asset classes and other performance and risk constraints, the proposed method explicitly models traffic effects on the condition of the roadway network components and enables a proposed active traffic flow control strategy based on traffic assignment that can be taken to limit access of the vehicles to identified roadway segments. By diverting the traffic to other locations and thus modifying traffic patterns, the expected remaining life of these roadway assets can be extended. The potential of the proposed concept and solution methodology is applied on a hypothetical case study.

## 1. Motivation

Preserving structural integrity and ensuring longevity and reliability of our roadway surface and bridge networks through maintenance and rehabilitation (M&R) is a recurring and expensive process that involves complex temporal and spatial decisions with system-wide impacts. While prioritization decisions can be taken based on component-level health for each asset class (here roadway pavements and bridges), because individual contributions of each roadway component of any asset class contributes to traffic capacity only through its collective influence on system performance, M&R actions on individual assets have nonlinear system-level impact.

Assessing the potential impact of the various combinations of M&R actions is computationally burdensome, as this involves solving difficult combinatorial problems. Such problems require consideration of an exponentially growing number of action combinations and their impacts with increasing network size. Further complicating this action-plan scheduling problem is uncertainty in knowledge of current condition state and condition-state deterioration of system components into the future, as well as the need to evaluate traffic flow impacts.

Numerous works have proposed algorithms to support the scheduling of M&R actions (e.g., Bocchini & Frangopol, 2011; Chu & Chen, 2012; Saydam & Frangopol, 2014; Yang & Frangopol, 2019; Mendoza, et al., 2021; Zhou & Wang, 2012) using genetic algorithms, threshold-based approaches, risk-informed techniques, renewal processes, and decision-trees, among other methods. Despite their important merits, many of these approaches suffer from suboptimality and scalability issues, ignore key characteristics of deterioration dynamics, and assume certainty in knowledge of future infrastructure condition states.

In prior research, notable contributions by (Ouyang, 2007) and (Durango-Cohen & Sarutipand, 2009) incorporated traffic network effects by linking deterioration and maintenance actions with roadway conditions, such as pavement roughness. In the former, these traffic effects were integrated through a user equilibrium (UE) model embedded within their modeling framework, focusing on an optimal resurfacing plan for the highway network. However, the proposed policy iteration method was limited to small-scale problems. In the latter, demand was assessed as a function of condition and capacity loss was utilized to account for the impacts of deterioration and implemented improvements on roadway conditions. Both works assume the existence of a deterministic deterioration process.

In (Medury & Madanat, 2013), the downtime effects of pavement improvement actions during their implementation were considered by accounting for capacity differences due to various maintenance actions.

A TD-learning (SARSA) reinforcement learning solution was applied using a simplified traffic model. Ng et al. (2009) incorporated capacity loss during maintenance activities in a maintenance planning framework. They accounted for downtime associated with the activities over increments of years and proposed a genetic algorithm for problem solution. A bi-level optimization approach with UE at the lower-level capturing deterioration and downtime effects on traffic was proposed by Chu and Chen (2012), with a threshold-based maintenance plan used at the upper level. The model presumes a deterministic, continuous deterioration process and a tabu search algorithm is proposed to choose from various threshold options. More recently, pavement maintenance planning with integrated traffic related effects of the implementation of maintenance actions can also be seen in (Aad et al., 2022; Prajapati et al., 2024), with both works using deterministic models.

In recent work by Zhou et al. (2002), limitations of these approaches were addressed. Their work proposes a Deep Reinforcement Learning (DRL) methodology for optimizing transportation infrastructure maintenance planning while also considering traffic impacts, uncertainty in component deterioration mechanisms, multiple asset classes, and traffic capacity loss during improvement action execution. They conceptualized the maintenance planning problem through a bilevel formulation that incorporates traffic impacts of maintenance actions through a lower-level UE model of traffic conditions. They reformulated the problem through a Markov Decision Process (MDP) framing and proposed a powerful Deep Centralized Multi-agent Actor-Critic (DCMAC) DRL algorithm for its solution. By realistically capturing traffic, the impacts of maintenance activities are properly accounted for in action scheduling. Results of a case study demonstrate the effectiveness of their approach, showing significant reductions in both traffic congestion and total costs compared to other commonly used solution methods.

While these many advancements have been made toward scheduling methods for this complex multi-asset M&R application, the problem of continuous need for cost-prohibitive and labor- and material-intensive M&R actions toward extending the life of roadway networks of pavements and bridges remains. Improved scheduling methodologies that can account for real-world characteristics and system-state predictions can lead to more efficient use of limited M&R resources through careful action prioritization and even cross-asset coordination; however, more is needed to contend with the high costs and labor and material needs of maintaining our transportation infrastructure. In this extended abstract, this class of problems is revisited.

Herein, a state-of-the-art Deep Decentralized Multi-agent Actor-Critic (DDMAC) DRL (Androitis & Papakonstantinou, 2021; Morato et al., 2023) with Centralized Training and Decentralized Execution (CTDE) structure is introduced to address this scheduling problem, but with an added traffic flow control feature. The DDMAC-CTDE approach is a deep, off-policy actor-critic algorithm that utilizes experience replay. It was introduced in (Saifullah et al., 2022, 2024) and is a variation of DCMAC applied in (Zhou et al., 2022) and the Decentralized-Partially Observable Markov Decision Process (Dec-POMDP) concept described in (Oliehoek et al., 2008) to capture the reality of uncertainties in condition state that remain even after inspections.

In addition to the novelty of the DDMAC-CTDE solution framework, the scheduling problem is expanded in two key dimensions. It captures: (1) the impact of traffic on surface and bridge-deck deterioration mechanisms and (2) formulates the M&R problem as a joint M&R scheduling and active traffic flow control problem. Thus, the solution provides an M&R scheduling plan that accounts for deterioration uncertainty, a limited budget, the effects of M&R activity execution on traffic performance, multiple asset classes and other performance and risk constraints as in prior works, but adds two new dimensions: that traffic itself explicitly affects the condition of the roadway components and that decisions can be taken to limit access of the vehicles to identified roadway segments. By diverting the traffic to other locations and thus modifying traffic patterns, the expected remaining life of these roadway assets can be extended. The proposed tool is used to show the potential of controlling roadway capacity and, thus, traffic flows to better manage roadway system deterioration and repair needs.

## 2. Methodology

An integrated framework for large-scale infrastructure M&R activity optimization with active traffic flow

control is proposed. The devised approach combines POMDPs, UE modeling, and the DDMAC-CTDE to optimize maintenance decisions while implementing active traffic flow control to restrict traffic flows by selectively limiting link capacities. Figure 1 gives an overview of the general solution structure.

While a description of the proposed method is beyond the scope of this short abstract, several core elements are needed to address active traffic flow control, including specification of traffic model deterioration models for both pavements and bridge decks and cost estimation. Cost estimation is complicated here by a need to include traffic-related costs that are a function of traffic flow levels, which are both determined through solution of a UE problem herein and controlled through decisions to limit link capacities to preserve link health. Longevity is captured through a reliability-oriented approach, where transition from a state of non-failure to failure governs the operational lifespan of the system. The UE solution is obtained through solution of the traffic assignment problem under chosen capacity restrictions through variational inequality (VI) (Smith, 1979; Boyles et al., 2020) and solution by the Frank-Wolfe algorithm (Frank et al., 1956). The talk will describe the need for faster traffic assignment solution and the potential benefits of dynamic traffic assignment for this application.

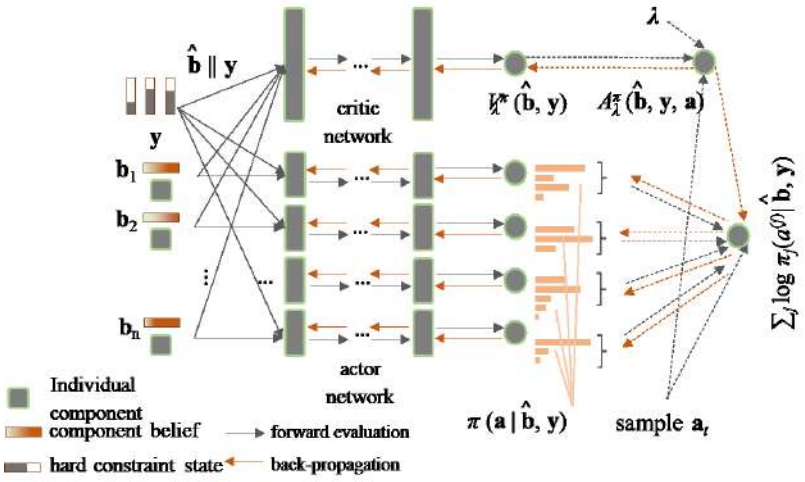



Figure 1. DDMAC-CTDE architecture for M&R scheduling with active traffic flow control via traffic assignment and capacity restrictions

## 3. Illustrative Case Study

The proposed concept and solution framework were applied on a 19-component roadway network (Figure 2). The problem is characterized by stochastic, non-stationary dynamics, and partially observable state conditions through uncertain outcomes of inspections. A decision period of 20 years with a discount rate of $\gamma = 0.970$ was used. The network includes 15 pavement sections with 8 designated as Type I, comparable to interstate highways with multiple lanes and high traffic volumes, and 7 as Type II, akin to primary highways. Four bridges are included: one Type I bridge situated on a Type I route and three Type II bridges on Type II routes. Four maintenance actions per component are considered: Do Nothing, Minor Repair, Major Repair, and Reconstruction actions. Active flow control actions through capacity changes are also incorporated into the action space, with 100 and 50% available capacities when the Do-Nothing action is selected, and 52.2% capacity reduction when maintenance actions are performed. The budget ranges from $0.6 to 1.1 billion based on a set of examined scenarios. Results of the application show the effectiveness of active traffic flow control at increasing longevity under a given budget and that there are chosen paths whose capacities are preserved at the cost of added M&R activity to maintain reliable traffic performance for system users.

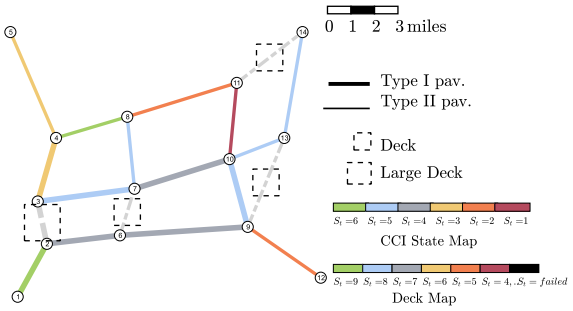



Figure 2: Network with 4 bridges and 15 pavement sections under given condition state

## 4. Conclusions

An active traffic flow control strategy combined with the M&R scheduling can significantly improve network-level roadway life-cycle performance, It allows for more efficient utilization of M&R monetary, labor and material resources by discouraging traffic patterns that include assets that are already deteriorated and postponing their repair or replacement for a prolonged period of time. This work contributes to larger national and international efforts to extend the life of the world's costly roadway infrastructure.

***Acknowledgment***. This works was sponsored by CIAMTIS under USDOT UTC Federal Grant 69A3551847103.

## 5. References

Aad, M., Abbas, M. & Wang, L., 2022. *A simulation-optimization framework for integrated infrastructure condition and traffic management strategies.* IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), 2869-2874.

Andriotis, C. P. & Papakonstantinou, K. G., 2021. Deep reinforcement learning driven inspection and maintenance planning under incomplete information and constraints. *Reliability Engineering & System Safety* 212, 107551.

Bocchini, P. & Frangopol, D. M., 2011. A probabilistic computational framework for bridge network optimal maintenance scheduling. *Reliability Engineering & System Safety* 96(2), 332-49.

Boyles, S. D., Lownes, N. E. & Unnikrishnan, A., 2020. *Transportation network analysis.* Austin, TX: Version 0.91.

Chu, J. & Chen, Y., 2012. Optimal threshold-based network-level transportation infrastructure life cycle management with heterogeneous maintenance actions. *Transportation Research Part B: Methodological* 46(9), 1123-1143.

Durango-Cohen, P. & Sarutipand, P., 2009. Maintenance optimization for transportation systems with demand responsiveness. *Transportation Research Part C: Emerging Technologies* 17(4), 337–348.

Frank, M., Wolfe, P. & others, 1956. An algorithm for quadratic programming. *Naval Research Logistics Quarterly,* 3(1-2), 95-110.

Medury, A. & Madanat, S., 2013. Incorporating network considerations into pavement management systems: A case for approximate dynamic programming. *Transportation Research Part C: Emerging Technologies 33,* 134-150.

Mendoza, J., Bismut, E., Straub, D. & Köhler, J., 2021. Risk-Based fatigue design considering inspections and maintenance. *Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering,* 7(1), 04020055.

Morato, P., Andriotis, C.P., Papakonstantinou, K.G. & Rigo, P., 2023. Inference and dynamic decision-making for deteriorating systems with probabilistic dependencies through Bayesian networks and deep reinforcement learning. *Reliability Engineering & System Safety* 235(0951-8320), 109144.

Ng, M., Lin, D. & Waller, S., 2009. Optimal long-term infrastructure maintenance planning accounting for traffic dynamics. *Journal of Computer-aided Civil and Infrastructure Engineering* 24(7), 459–469.

Oliehoek, F. A., Spaan, M. T. & Vlassis, N., 2008. Optimal and approximate Q-value functions for decentralized POMDPs. *Journal of Artificial Intelligence Research* 32, 289--353.

Ouyang, Y., 2007. Pavement resurfacing planning for highway networks: parametric policy iteration approach. *Journal of Infrastructure Systems* 13, 65-71.

Prajapati, A., Sengupta, A. & Guler, S. I., 2024. Accounting for traffic dynamics in pavement maintenance scheduling. *Transportation Research Record* 2678(7), 361-375.

Saifullah, M., Andriotis, C.P., Papakonstantinou, K.G. & Stoffels, S., 2022. *Deep reinforcement learning-based life-cycle management of deteriorating transportation systems.* IABMAS, 293-301.

Saifullah, M., Papakonstantinou, K. G., Andriotis, C. P. & Stoffels, S. M., 2024. Multi-agent deep reinforcement learning with centralized training and decentralized execution for transportation infrastructure management. *arXiv preprint arXiv:2401.12455*.

Saydam, D. & Frangopol, D., 2014. Risk-based maintenance optimization of deteriorating bridges. *Journal of Structural Engineering,* 141(4), 04014120.

Smith, M., 1979. The existence, uniqueness and stability of traffic equilibria. *Transportation Research 13B,* 295-304.

Yang, D. Y. & Frangopol, D. M., 2019. Life-cycle management of deteriorating civil infrastructure considering resilience to lifetime hazards: A general approach based on renewal-reward processes. *Reliability Engineering & System Safety,* Volume 183, 197-212.

Zhou, G. & Wang, L., 2012. Co-location decision tree for enhancing decision-making of pavement maintenance and rehabilitation. *Transportation Research Part C: Emerging Technologies,* 21(1), pp. 287--305.

Zhou, W., Miller-Hooks, E., Papakonstantinou, K.G., Stoffels, S., & McNeil, S., 2022. A reinforcement learning method for multiasset roadway improvement scheduling considering traffic impacts. *Journal of Infrastructure Systems* 28(4), 04022033.

**Congestion Management through Path Incentives and Tolls**

Ramin Niroumand[1], Leila Hajibabai[2], and Ali Hajbabaie*[3]

[1] Postdoctoral Researcher, Department of Civil, Construction, and Environmental Engineering, NC State University, USA

[2] Associate Professor, Department of Industrial and Systems Engineering, NC State University, USA

[3] Associate Professor, Department of Civil, Construction, and Environmental Engineering, NC State University, USA

## SHORT SUMMARY

This study introduces a framework for steering user equilibrium (UE) flow patterns toward a system optimal (SO) state by integrating road pricing with path-based incentives. The framework is formulated as an optimization model with complementarity constraints, incorporating equity considerations at both the network and origin-destination (OD) levels. A column generation (CG)-based solution approach iteratively explores previously unconsidered paths, enhancing solution quality and preventing convergence to local optima. The framework's effectiveness is demonstrated using the Nguyen-Dupuis network under three scenarios: Base, Network-Equity, and OD-Equity. Results indicate that while the Base scenario minimizes total travel time, it requires external funding and does not ensure equity. The Network-Equity scenario balances tolls and incentives at the network level with minimal increases in travel time but does not address OD-level fairness. In contrast, the OD-Equity scenario ensures equitable toll distribution within OD pairs at a negligible cost (<0.02% increase in travel time), underscoring the framework's potential for equitable and efficient traffic management.

**Keywords:** Incentive scheme, road pricing, system optimum, traffic assignment, traffic management, urban transportation network.

## 1. INTRODUCTION

Urbanization and the increasing number of private vehicles have significantly exacerbated traffic congestion and air pollution in recent years. While expanding transportation infrastructure could, in theory, alleviate congestion and improve air quality, it is often prohibitively expensive and impractical due to limited urban space. Additionally, infrastructure expansion tends to induce demand, which can ultimately worsen congestion and environmental issues over time. Consequently, researchers have shifted their focus toward optimizing existing infrastructure by redirecting traffic flow from the naturally occurring user equilibrium (UE) pattern to a system optimal (SO) flow pattern (Andersson et al., 2018; Cheng et al., 2020; Sunio and Schmöcker, 2017; Zangui et al., 2015).

In a UE state, all travelers between the same origin and destination experience equal travel costs, whereas an SO state minimizes total travel cost across the entire network, even if some drivers incur higher individual costs. Achieving SO, therefore, requires incentives to encourage drivers to take longer or less convenient routes for the greater benefit of the overall system. Traditionally, road pricing has been used to influence UE flow patterns toward SO (Bergendorff et al., 1997;

Ren et al., 2020; Yang and Huang, 2004; Zangui et al., 2015). However, road pricing often faces public resistance due to perceived unfairness and concerns about inequitable welfare distribution (May et al., 2010; Vosough et al., 2022).

Recently, incentive-based approaches have gained popularity as an alternative (Ettema et al., 2010; Leblanc and Walker, 2013; Sun et al., 2020). Incentives can be implemented using either link-based or path-based strategies (Luan et al., 2023), with research indicating that path-based incentives are more effective in achieving desired outcomes (Niroumand et al., 2024). A combined approach, integrating both monetary incentives and road pricing, can provide a balanced solution by leveraging the advantages of both methods. Furthermore, toll revenues collected through road pricing can be reinvested into the community as part of the incentivization scheme, fostering broader public acceptance and promoting equitable outcomes.

This study presents a framework for shifting UE flow patterns toward SO by employing a combination of road pricing and path-based incentives. The proposed approach is formulated as an optimization model with complementarity constraints, incorporating equity considerations at both the network and OD levels. A column generation (CG)-based solution method is used to ensure computational efficiency and scalability.

## 2. PROBLEM FORMULATION

This section presents an optimization model designed to determine the optimal path incentives and tolls required to shift the UE flow pattern toward the SO, while adhering to budget and maximum path toll constraints. The transportation network is modeled as a graph $G(V, A)$, where $V$ denotes the set of nodes, and $A \subset V \times V$ represents the set of links. The set of OD pairs is defined as $W \subset V \times V$, with a fixed travel demand $q_w$ for each OD pair $w \in W$. The set of all paths between a specific OD pair $w \in W$ is denoted by $P_w$. The decision variables in the optimization model are $f_w^p$, $y_w^p$, and $\hat{y}_w^p$, which correspond to the vehicle flow, toll, and incentive assigned to path $p \in P_w$ for OD pair $w \in W$, respectively. Additionally, the travel cost and flow on a link $a \in A$ are denoted as $t_a$ and $x_a$, respectively. The following formulation represents the proposed problem.

$$Z = \min_{\boldsymbol{f},\boldsymbol{y},\hat{\boldsymbol{y}}} \sum_{a \in A} (x_a t_a) \tag{1}$$

$$\sum_{p \in P_w} f_w^p = q_w \qquad \forall\, w \in W \tag{2}$$

$$\sum_{a \in A} (\delta_a^p t_a) + y_w^p - \hat{y}_w^p - u_w \geq 0 \qquad \forall\, p \in P_w,\ w \in W \tag{3}$$

$$\left(\sum_{a \in A} \left(\delta_a^p t_a\right) + y_w^p - \hat{y}_w^p - u_w\right) f_w^p = 0 \qquad \forall\, p \in P_w,\ w \in W \tag{4}$$

$$y_w^p \leq T \qquad \forall\, p \in P_w,\ w \in W \tag{5}$$

$$\sum_{w \in W} \sum_{p \in P_w} \hat{y}_w^p f_w^p \leq B \tag{6}$$

$$x_a = \sum_{w \in W} \sum_{p \in P_w} \delta_a^p f_w^p \qquad \forall\, a \in A \tag{7}$$

$$t_a = t_a(x_a) \qquad \forall\, a \in A \tag{8}$$

$$\sum_{w \in W} \sum_{p \in P_w} \hat{y}_w^p f_w^p = \sum_{w \in W} \sum_{p \in P_w} y_w^p f_w^p \tag{9}$$

$$\sum_{p\in P_w} \hat{y}_w^p f_w^p = \sum_{p\in P_w} y_w^p f_w^p \qquad \forall\, w \in W \qquad (10)$$

Objective function (1) minimizes the total travel time in the network with respect to path flows, incentives, and tolls. Constraints (2) guarantee flow feasibility by ensuring that all travelers get to their predefined destinations. Constraints (3) and (4) serve as complementarity constraints to enforce a UE flow pattern, ensuring that the generalized travel times are equal across all used paths. The generalized travel time of a path is calculated as the summation of the actual travel times of its constituting links ($\sum_{a\in A} \delta_a^p t_a$) plus the toll assigned to it ($y_w^p$) minus its assigned incentive ($\hat{y}_w^p$). Here $\delta$ represents the path-link incidence matrix, where $\delta_a^p$ equals one if link $a$ is on path $p$, and zero otherwise. Constraints (5) impose an upper limit on the toll assigned to each path to mitigate public dissatisfaction with excessive toll charges. Constraints (6) ensure that the total incentives provided do not exceed the available budget, maintaining financial feasibility. Constraints (7) define link flows as a function of path flows, while constraints (8) specify link travel times based on link flows.

Equity considerations are incorporated through constraints (9) and (10) at both the network and OD-pair levels, respectively. Constraints (9) guarantee that the total toll revenue collected is fully redistributed as incentives, upholding the principle of reinvesting funds collected from users back into the system to enhance overall user benefits. Constraints (10) further extend the equity by requiring that toll collection and incentive distribution are balanced for each OD pair. This provision guarantees that travelers have equitable access to both options: a shorter, tolled route or a longer route with monetary incentives, promoting fairness in route selection.

## 3. SOLUTION TECHNIQUE

A CG technique is introduced to solve the optimization model defined by equations (1) – (10), allowing for the dynamic creation of new paths as the algorithm advances. While CG can reach optimality by enumerating all paths (Desaulniers et al., 2006), it can be terminated when the improvements between iterations fall below a predefined threshold, effectively balancing computational efficiency and solution accuracy.

The solution framework begins with an empty set of working paths. At each iteration, the algorithm identifies and adds a new path for each OD pair as a column, continuing for a minimum of $N$ iterations. The process terminates when the relative difference in total travel time between two consecutive iterations falls below a predefined threshold $\epsilon$. The proposed CG-based approach ensures that the shortest path not yet included in the current working set is identified and added at each iteration. This guarantees that a new path is introduced to the solution space, even if it has a longer travel time, as such paths may benefit from incentives and contribute to a reduction in overall network travel time. Furthermore, the minimum iteration count $N$ is chosen to ensure that a sufficient number of paths are generated between each OD pair before the algorithm halts, enhancing solution robustness. The steps of the proposed algorithm are as follows.

1. **Initialization:**
   i. Define values for $N$ and $\epsilon$
   ii. Set $P_w = \emptyset$, $n = 1$, $y = 0$, $f = 0$, $x^0 = x(f)$, and $t^0 = t(x^0)$
2. **Finding Shortest paths:** for each OD pair $w \in W$,
   i. Find the shortest path $p$ such that $p \notin P_w$,
   ii. Set $P_w = P_w \cup p$

3. **Finding UE:** solve the optimization model and find path flows $\boldsymbol{f}$ along with their corresponding toll $\boldsymbol{y}$ and incentive $\widehat{\boldsymbol{y}}$.

4. **Update:** set $\boldsymbol{x^n} = \boldsymbol{x}(\boldsymbol{f})$ and $\boldsymbol{t^n} = \boldsymbol{t}(\boldsymbol{x^n})$.

5. **Stopping criteria**

   i. Calculate $\bar{\epsilon} = \frac{|\sum_{a\in A}(x_a^n t_a^n) - \sum_{a\in A}(x_a^{n-1} t_a^{n-1})}{\sum_{a\in A}(x_a^n t_a^n)}$

   ii. If $\bar{\epsilon} \leq \epsilon$ and $n \geq N$ stop, otherwise set $n = n + 1$ and go to Step 2.

In Step 2 of the proposed algorithm, a new path that is not part of the current working path set is identified to prevent convergence to a local optimum. Since this path cannot be found through conventional shortest path algorithms, we formulate and solve the integer optimization problem presented in equations (11) - (14). This approach ensures that the previously unexplored shortest path between each OD pair is systematically added to the working path set, enhancing the overall solution quality.

$$\tau_w = \min_{\boldsymbol{\beta}} \sum_{a\in A} (\beta_a t_a) \tag{11}$$

$$\sum_{a\in A: start(a)=v} \beta_a - \sum_{a'\in A: end(a')=v} \beta_{a'} = \begin{cases} +1 & if\ v\ is\ the\ origin\ node \\ -1 & if\ v\ is\ the\ destination\ node\ , \\ 0 & otherwise \end{cases} \quad \forall\, v \in V \tag{12}$$

$$\textstyle\sum_{a\in A}(\delta_a^p - \beta_a) \leq 1, \quad \forall\, p \in P_w \tag{13}$$

$$\beta_a \in \{0,1\}, \quad \forall\, a \in A \tag{14}$$

where, $\beta_a$ is a binary variable that equals one if link $a$ is part of the shortest path and zero otherwise. The objective function (11) aims to identify the shortest path by selecting a combination of links that minimizes travel time. Constraints (12) ensure that the resulting solution forms a continuous path rather than a disconnected set of links. In these constraints, start($a$) and end($a$) represent the starting and ending nodes of link $a \in A$, respectively. Constraints (13) prevent the selection of any path that already exists in the current working path set to ensure that a new, previously unexplored path is generated.

## 4. RESULTS AND DISCUSSION

The proposed methodology is implemented on the Nguyen-Dupuis network illustrated in Figure 1. This network comprises 13 nodes, 19 links, and 4 OD pairs. We are working on applying the approach to Sioux-Falls and Austin, TX network to be included in the full paper. The link travel time function is assumed to follow the BPR function (Manual, 2000), expressed as:

$$t_a(x_a) = t_a^0 \left(1 + 0.15 \left(\frac{x_a}{C_a}\right)^4\right), \quad \forall\, a \in A \tag{15}$$

where $t_a^0$ represents the free flow travel time and $C_a$ is the capacity of link $a \in A$. The link properties used in the BPR function and the demand between OD pairs are sourced from (Luan et al., 2023).

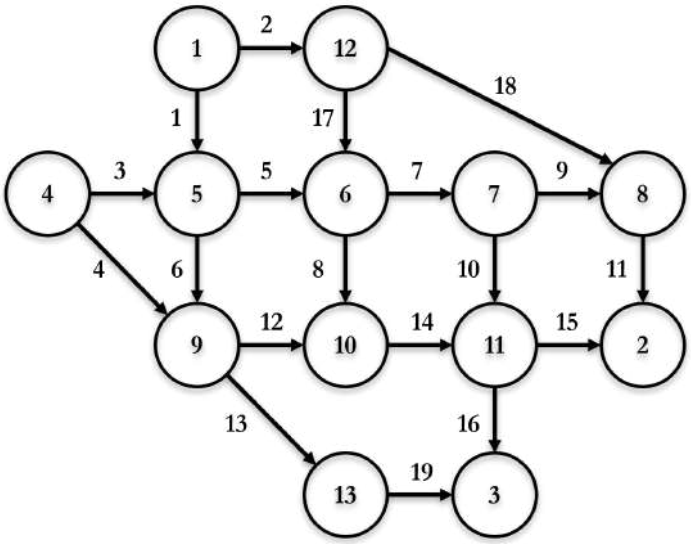



**Figure 1: Nguyen-Dupuis network**

We define three scenarios based on equity constraints: *Base*, *Network-Equity*, and *OD-Equity*. The *Base* scenario excludes equity considerations by solving the optimization model defined by equations (1) to (8), without incorporating the equity constraints specified in (9) and (10). The *Network-Equity* scenario promotes system-wide equity by redistributing toll revenues as incentives across the entire network, aligning with the optimization model outlines in equations (1) through (9). Finally, the *OD-Equity* scenario focuses on ensuring that toll revenues collected from users of a specific OD pair are redistributed among users of the same OD pair. This scenario is modeled using the optimization model defined by equations (1) to (8) and (10).

Table 1 summarizes the total travel time, total toll revenue, and total incentives expended under varying path toll caps and budget limits in the base scenario, where equity considerations are not incorporated. Notably, the scenario with zero budget and zero tolls corresponds to the naturally occurring UE flow pattern, resulting in a total travel time of 181,430.87 time units. In contrast, the SO flow pattern achieves a total travel time of 179,589.29 time units. The results demonstrate that increasing both the budget limit and the path toll cap leads to a reduction in total travel time. Furthermore, a combined approach utilizing both tolling and incentivization appears more effective in achieving optimal network performance. For example, achieving the minimum total travel time across the network requires 5,680.87 time units of monetary incentives. However, allowing a maximum toll of 2 time units per path significantly reduces the required budget for incentives to 2,978.21 time units.

We analyzed the impact of incorporating equity considerations using a case study with a budget limit of 2,000 time units and a path toll cap of 2 time units. The results, including total travel time, total toll revenue, and total distributed incentives, are presented in Table 2. As anticipated, the *Base* scenario yields the lowest total travel time across the entire network and between all OD pairs. However, under this scenario, the total toll revenue amounts to nearly half of the distributed incentives, necessitating external funding to sustain the incentivization scheme. Additionally, travelers between OD pairs 1-2 and 4-3 receive fewer incentives than the tolls they pay, highlighting an imbalance in the distribution of benefits. The *Network-Equity* scenario ensures that total toll revenue matches the total distributed incentives, achieved with a minimal increase in total travel time. Nevertheless, this scenario does not address the inequitable distribution of tolls and incentives at the OD pair level. In contrast, the *OD-Equity* scenario ensures that toll revenue equals distributed incentives for each OD pair, resulting in a fairer distribution across users. This equitable allocation offers travelers the choice between shorter routes with toll payments and longer routes with incentives. Notably, the equity-driven adjustments result

in a negligible increase in total travel time, accounting for less than 0.02% of the overall travel time.

**Table 1: Total travel time, total toll revenue, and total incentive with different toll cap and budget under the *base* scenario**

| | | | Budget (incentive limit) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | 0 | 1000 | 2000 | 3000 | 4000 | 5000 | 6000 |
| Maximum path toll | 0 | Total travel time | 181430.9 | 180107.1 | 179781.0 | 179666.9 | 179615.2 | 179593.5 | 179589.3 |
| | | Total toll | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | | Total incentive | 0.0 | 1000.0 | 2000.0 | 3000.0 | 4000.0 | 5000.0 | 5680.9 |
| | 1 | Total travel time | 180692.3 | 179791.8 | 179667.0 | 179606.4 | 179590.2 | 179589.3 | 179589.3 |
| | | Total toll | 1144.3 | 809.7 | 800.5 | 648.8 | 641.3 | 568.2 | 434.1 |
| | | Total incentive | 0.0 | 1000.0 | 2000.0 | 3000.0 | 4000.0 | 4249.1 | 4315.0 |
| | 2 | Total travel time | 180202.4 | 179653.8 | 179596.0 | 179589.3 | 179589.3 | 179589.3 | 179589.3 |
| | | Total toll | 2398.8 | 985.1 | 1004.0 | 897.3 | 264.3 | 868.2 | 264.3 |
| | | Total incentive | 0.0 | 1000.0 | 2000.0 | 2978.2 | 3545.3 | 2629.1 | 3545.2 |
| | 3 | Total travel time | 179892.1 | 179591.6 | 179589.3 | 179589.3 | 179589.3 | 179589.3 | 179589.3 |
| | | Total toll | 3327.4 | 1717.4 | 1689.9 | 608.0 | 1596.0 | 1404.1 | 608.0 |
| | | Total incentive | 0.0 | 1000.0 | 1610.8 | 2929.0 | 1796.9 | 1605.1 | 2928.9 |
| | 4 | Total travel time | 179713.8 | 179589.3 | 179589.3 | 179589.3 | 179589.3 | 179589.3 | 179589.3 |
| | | Total toll | 3641.7 | 1909.2 | 2289.2 | 2068.1 | 527.7 | 893.8 | 1828.9 |
| | | Total incentive | 0.0 | 1000.0 | 1250.3 | 1029.2 | 2048.7 | 2094.8 | 1270.0 |
| | 5 | Total travel time | 179722.9 | 179589.3 | 179589.3 | 179589.3 | 179589.3 | 179589.3 | 179589.3 |
| | | Total toll | 4397.8 | 1363.0 | 2591.7 | 2019.3 | 2121.0 | 2344.7 | 2344.8 |
| | | Total incentive | 0.0 | 964.0 | 872.7 | 980.4 | 1082.1 | 625.8 | 625.9 |
| | 6 | Total travel time | 179592.0 | 179589.3 | 179589.3 | 179589.3 | 179589.3 | 179589.3 | 179589.3 |
| | | Total toll | 4190.3 | 1895.6 | 1696.1 | 1228.9 | 1588.6 | 2631.6 | 1134.3 |
| | | Total incentive | 0.0 | 856.6 | 2000.0 | 2109.9 | 1189.7 | 712.6 | 2335.4 |

**Table 2: Total travel time, total toll revenue, and total incentive with a budget limit of 2000 and path toll cap of 2 time units under different equity consideration scenarios**

| Scenario | Metric | Network | Origin - Destination | | | |
|---|---|---|---|---|---|---|
| | | | 1-2 | 1-3 | 4-2 | 4-3 |
| *Base* | Total incentive | 2000.00 | 0.00 | 1280.32 | 616.52 | 103.16 |
| | Total toll | 1003.99 | 61.72 | 271.75 | 541.85 | 128.67 |
| | Total travel time | 179596.02 | 32630.73 | 86618.50 | 41935.62 | 18411.17 |
| *Network-equity* | Total incentive | 1924.06 | 40.70 | 381.91 | 491.61 | 1009.83 |
| | Total toll | 1924.06 | 632.24 | 745.79 | 546.03 | 0.00 |
| | Total travel time | 179615.21 | 32430.36 | 85791.58 | 41968.85 | 19424.42 |
| *OD-equity* | Total incentive | 1924.70 | 283.01 | 821.15 | 605.99 | 214.56 |
| | Total toll | 1924.70 | 283.01 | 821.15 | 605.99 | 214.56 |
| | Total travel time | 179635.95 | 32874.29 | 86456.90 | 41975.50 | 18329.26 |

## 5. CONCLUSIONS

This study presents a comprehensive framework for steering user equilibrium (UE) flow patterns toward a system optimal (SO) state by integrating road pricing and path-based incentives. The framework is formulated as an optimization model with complementarity constraints,

incorporating equity considerations at both the network and origin-destination (OD) levels. The proposed model is solved using a column generation (CG)-based solution technique, which iteratively generates previously unexplored paths to enhance solution quality and avoid local optima. The effectiveness of the proposed approach is demonstrated through its application to the Nguyen-Dupuis network, analyzed under three scenarios: Base, Network-Equity, and OD-Equity. Results indicate that while the Base scenario achieves the lowest total travel time, it requires external funding to sustain the incentive scheme and fails to ensure a fair toll-incentive distribution. The Network-Equity scenario balances total toll revenue and distributed incentives at the network level, with only a negligible increase in total travel time. However, this approach does not address fairness at the OD level. In contrast, the OD-Equity scenario ensures that toll revenues are redistributed equitably within each OD pair, allowing travelers to either choose shorter routes with tolls or opt for longer routes with incentives. Notably, the cost of achieving OD-level equity is minimal, resulting in less than a 0.02% increase in total travel time.

While the improvements observed in this study are promising, the difference between SO and UE solutions is relatively modest, at only 1%. To further evaluate the potential of this approach, we are currently applying the proposed method to the Sioux-Falls and Austin, TX networks, testing the formulation and solution technique under more complex and challenging conditions.

## REFERENCES


Andersson, A., Hiselius, L.W., Adell, E., 2018. Promoting sustainable travel behaviour through the use of smartphone applications: A review and development of a conceptual model. Travel Behav. Soc. 11, 52–61.

Bergendorff, P., Hearn, D.W., Ramana, M. V, 1997. Congestion toll pricing of traffic networks. Springer.

Cheng, Z., Pang, M.-S., Pavlou, P.A., 2020. Mitigating traffic congestion: The role of intelligent transportation systems. Inf. Syst. Res. 31, 653–674.

Desaulniers, G., Desrosiers, J., Solomon, M.M., 2006. Column generation. Springer Science \& Business Media.

Ettema, D., Knockaert, J., Verhoef, E., 2010. Using incentives as traffic management tool: empirical results of the" peak avoidance" experiment. Transp. Lett. 2, 39–51.

Leblanc, R., Walker, J.L., 2013. Which is the biggest carrot? comparing nontraditional incentives for demand management. In: Transportation Research Board 92nd Annual Meeting.

Luan, M., Waller, S.T., Rey, D., 2023. A non-additive path-based reward credit scheme for traffic congestion management. Transp. Res. part E Logist. Transp. Rev. 179, 103291.

Manual, H.C., 2000. Highway capacity manual. Washington, DC 2, 1.

May, A.D., Koh, A., Blackledge, D., Fioretto, M., 2010. Overcoming the barriers to implementing urban road user charging schemes. Eur. Transp. Res. Rev. 2, 53–68.

Niroumand, R., Vosough, S., Roncoli, C., Rinaldi, M., Connors, R., 2024. Beyond links: The power of path incentives in alleviating congestion and emissions in urban networks. In: Symposium of the European Association for Research in Transportation.

Ren, T., Huang, H.-J., Liu, T.-L., Nie, Y.M., 2020. Some analytical results on spatial price differentiation in first--best congestion pricing schemes. Transp. Res. Part C Emerg. Technol. 114, 425–445.

Sun, J., Wu, J., Xiao, F., Tian, Y., Xu, X., 2020. Managing bottleneck congestion with incentives. Transp. Res. part B Methodol. 134, 143–166.

Sunio, V., Schmöcker, J.-D., 2017. Can we promote sustainable travel behavior through mobile apps? Evaluation and review of evidence. Int. J. Sustain. Transp. 11, 553–566.

Vosough, S., de Palma, A., Lindsey, R., 2022. Pricing vehicle emissions and congestion externalities using a dynamic traffic network simulator. Transp. Res. Part A Policy Pract. 161, 1–24.
Yang, H., Huang, H.-J., 2004. The multi-class, multi-criteria traffic network equilibrium and systems optimum problem. Transp. Res. Part B Methodol. 38, 1–15.
Zangui, M., Aashtiani, H.Z., Lawphongpanich, S., Yin, Y., 2015. Path-differentiated pricing in congestion mitigation. Transp. Res. Part B Methodol. 80, 202–219.

**Extended ABSTRACT for the 10th International Symposium on Dynamic Traffic Assignment (DTA2025)**

# Dynamic traffic assignment in a trip-based multi-reservoir model

Jean-Patrick Lebacque[a] and Megan Khoshyaran[b]
[a] *University Gustave Eiffel, COSYS, GRETTIA, Paris, France*
[b] *Economics Traffic Clinic, ETC, Paris, France*

February 13, 2025

**Keywords:** System Optimum; User Optimum; Regional Network; GSOM model; Dynamical system; Macroscopic fundamental diagram; Advection equation; Generalized bathtub model

## 1 Introduction

As automation and communication progress, the management and planning of very large regional-sized transportation networks becomes an increasingly important issue. The ability to solve Dynamic Traffic Assignment (DTA) problems on such netwoks is crucially important for regional authorities. Solving DTA problems on very large networks turns out to be a very diffucult task, to which much research has been devoted, notably when using microscopic models. Refer for instance to Ameli et al. [2020]. Using simplified macroscopic models constitutes a promising approach Yildirimoglu et al. [2015], Ehteram et al. [2017], Balzer et al. [2023], Hassanin [2025]. Bidimensional models provide a first paradigm. This technique was initiated in [Taguchi and Iri, 1982] (operational research flow assignment problems), then adapted to static and dynamic mono-centric traffic flow problems Ho and Wong [2006], Jiang et al. [2011], Mollier et al. [2019]. Multi-directional flow modelling has been considered in Saumtally et al. [2013], Khoshyaran and Lebacque [2020, 2025]. The MFD (macroscopic fundamental diagram) approach provides a second, much studied, paradigm, Geroliminis and Daganzo [2008], Geroliminis et al. [2013], Aghamohammadi and Laval [2019], Huang et al. [2020]. MFD techniques have evolved into several directions, accumulation-based models, trip-based models and delay-based models, refer to Huang et al. [2024].
The model developped in this paper is connected to trip-based models Mariotte and Leclercq [2017], Mariotte et al. [2017] We will also use some tools from the related generalized bathtub model Jin [2020]. The considered network is described as a set of reservoirs, each of them endowed with a MFD, and traffic from one reservoir to an adjacent reservoir is described as propagating along a mean trip joining the reservoir centroids. This representation allows for FIFO dynamics and is thus well-adapted to DTA problems. Indeed in DTA problems it is necessary to keep track of the composition of traffic in terms of destination, assignment choices and possibly departure times of travellers. The DTA model proposed in the paper follows the approch developped in Khoshyaran and Lebacque [2023, 2025]: the route choice is based on instantaneous travel costs, whereas the departure time choice is based on predictive travel costs. It will be shown that these costs can readily be obtained within the framework of the model.

## 2 Outline of the model

Denote by $\mathcal{R}$ the set of reservoirs, including the reservoirs describing the considered network plus a special reservoir $\Omega$ representing all that is outside of the network. Given a reservoir $(i) \in \mathcal{R}$, we denote by $\Gamma_i^+ \subset \mathcal{R}$ the set of successor reservoirs of $(i)$, that is $(j) \in \Gamma_i^+$ if there is a non-empty interface $(i) - (j)$ through which vehicles can travel directly $(i) \to (j)$. By definition $(i)$ is then a predecessor of $(j)$, that is $(i) \in \Gamma_j^-$. The model assumes a specific trip length $\ell_i$ for travellers inside a reservoir $(i)$, bound for a reservoir $(j) \in \Gamma_i^+$. The number of these travellers is given by $M_{ij} + N_{ij}$ (functions of time), possibly disaggregated with respect to an attribute $e \in \mathcal{E}$:

$$M_{ij} = \sum_{e \in \mathcal{E}} M_{ij}^e \qquad N_{ij} = \sum_{e \in \mathcal{E}} N_{ij}^e$$

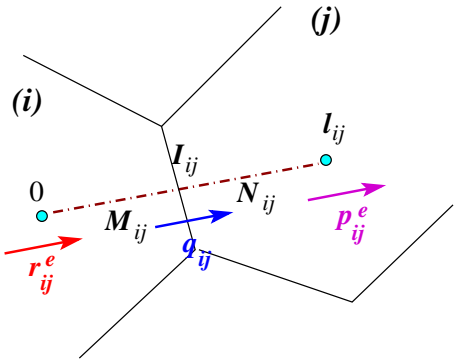



Figure 1: Reservoirs and associated quantities

Refer to Figure 1. Typically the traveller attribute $e$ is a vector, the components of which will be destination $d \in \mathcal{D}$ or route $p \in \mathcal{P}$ (depending on the assignment behaviour), and desired arrival time $t_a \in \mathcal{T}_a$. Thus $e = (d, t_a)$ or $e = (p, t_a)$. Let us define some further elements of description of the reservoirs.

- $N_i \overset{def}{=} \sum_{(j)\in\Gamma_i^+} M_{ij} + \sum_{(k)\in\Gamma_i^-} N_{ki}$ : the total accumulation in reservoir $(i)$;
- $\ell_{ij}$: the common mean trip length perceived by travellers travelling $(i) \to (j)$ . A refinement of the model could be to consider a distribution of trip lengths $(i) \to (j)$ as in the generalized bathtub model [Jin, 2020], Ameli et al. [2022].
- $\nu^e_{ij}(\ell)\, d\ell$ : the distribution of accumulation along the $(i) \to (j)$ trips pertaining to attribute $(e) \in \mathcal{E}$. We also denote $\nu_{ij}(\ell) \overset{def}{=} \sum_{e\in\mathcal{E}} \nu^e_{ij}(\ell)$. Thus we define

$$M^e_{ij} = \int_0^{\mathcal{I}_{ij}} \nu^e_{ij}(\ell)\, d\ell \qquad N^e_{ij} = \int_{\mathcal{I}_{ij}}^{\ell_{ij}} \nu^e_{ij}(\ell)\, d\ell\ ; \qquad M_{ij} = \int_0^{\mathcal{I}_{ij}} \nu_{ij}(\ell)\, d\ell \qquad N_{ij} = \int_{\mathcal{I}_{ij}}^{\ell_{ij}} \nu_{ij}(\ell)\, d\ell \tag{1}$$

  where $\mathcal{I}_{ij}$ denotes the mean trip length inside reservoir $(i)$ of travellers $(i) \to (j)$. This concept allows us to keep track of the progress of traffic per destination and to respect FIFO (first-in-first-out) discipline for traffic in reservoirs, which is necessary when solving DTA problems.
- The set $\mathcal{W}$ of ODs (origin-destination couples) is a subset of $\mathcal{R} \times \mathcal{R}$. If $w \in \mathcal{W}$ the set $\mathcal{P}_w$ denotes the set of paths joining the OD $w$. Thus a path is defined as a succession of reservoirs.

Let us now outline the traffic dynamics.

- First we need to define assignment coefficients $\gamma^e_{ij,k}$ (which are function of $\lambda \in [o, \ell_{ij}]$ and of time $t$): the fraction of traffic $(i) \to (j)$ chosing $k \in \Gamma_j^+$ given their attribute $e$.
- Let us define concepts of supply and demand for reservoir. One possibility would be to adopt the concepts of supply and demand introduced in [Khoshyaran and Lebacque, 2025], based on the capacity attributes of the interfaces between reservoirs. In this paper we propose a simpler approach. We call partial demand $\delta_{ij}$ the demand of traffic $M_{ij}$. We propose the following expression

$$\delta_{ij} \overset{def}{=} \frac{N_{ij}}{N_i}\, \Delta_i\,(N_i)\,. \left[ Mij + \sum_{k\in\Gamma_i^-, e\in\mathcal{E}} \gamma^e_{ki,j} N^e_{ki,j} \right] \tag{2}$$

  where $\Delta_i$ denotes the global demand of the reservoir $(i)$ as could be deduced from the MFD (macroscopic fundamental diagram) of the reservoir. (2) expresses a GSOM rule for partial flows (proportional to composition of traffic). Since we normalize with respect to $N_i$ we need to consider all travellers liable to travel $(i) \to (j)$.

- Similarly we define the donstream partial supply $\sigma_{ij} \overset{def}{=} \beta_{ij}\Sigma_j(N_j)$ as a fraction of the total supply of reservoir $(j)$ in accordance with the STRADA model Buisson et al. [1995]. We can deduce by the min principle the flow $q_{ij} : (i) \to (j)$ which applies naturally at location $\lambda = I_{ij}$ at the interface between $(i)$ and $(j)$, and from the flow $q_{ij}$ we deduce the speed $v_{ij}$ of traffic $(i) \to (j)$:

$$q_{ij} = \min\left[\delta_{ij}, \sigma_{ij}\right] \qquad v_{ij} = q_{ij}I_{ij}/M_{ij} \tag{3}$$

  In keeping with the idea of trip-based modelling this speed is assumed to be independent of $\lambda$. Given (2) it can be shown that the speed $v_{ij}$ is less than the maximum speed in reservoir $(i)$.

- The composition $\nu^e_{ij}$ is advected at speed $v_{ij}$, hence follows an advection equation

$$\partial_t \nu^e_{ij} + v_{ij}\partial_\lambda \nu^e_{ij} = 0 \tag{4}$$

  The outflows $(i) \to (j)$ result: $p^e_{ij} = v_{ij}\nu^e_{ij}(\ell_{ij})$. Let us denote by $\Pi^e_{ij}$ the inflow of travellers into the reservoir $(i)$ from outside the network (parkings etc) chosing $(j)$ as the next reservoir of their trip. Then the inflows $r^e_{ij}$ into $(i) \to (j)$ are given by

$$r^e_{ij} = \Pi^e_{ij} + \sum_{k\in\Gamma^-_i, e\in\mathcal{E}} \gamma^e_{ki,j} p^e_{ki,j} \tag{5}$$

  These inflows define the boundary conditions for (4): $\nu^e_{ij}(\lambda)_{|\,\lambda=0} = r^e_{ij}/v^e_{ij}$.

# 3 Dynamic traffic assignment.

The data of the problem is given by the total demand $D^{t_a}_w$, with $t_a \in \mathcal{T}_a$ the desired arrival times and $w \in \mathcal{W}$ the OD. This total demand must be disaggregated with respect to departure time $t \in \mathcal{T}_d$, the distribution of departure times is denoted $\varphi^{t_a}_w(t)dt$ (which is positive and of total mass equal to 1). This distribution will be the result of the departure time choice, based on predictive travel costs.
We also denote $\varpi^{t_a}_{wp}(t)$ the fraction of traffic with attributes $w \in \mathcal{W}$, desired arrival time $t_a \in \mathcal{T}_a$, chosing path $p \in \mathcal{P}_w$ at departure time $t$. The route choice will result from instantaneous travel costs. With these definitions

$$\Pi^e_{ij}(t) = \sum_{w/(i)=O(w)} \sum_{p\in\mathcal{P}_w,(ij)\in p} \varpi^{t_a}_{wp}(t)\varphi^{t_a}_w(t)D^{t_a}_w \tag{6}$$

($O(w)$ denotes the origin of $w$). Thus the main unknowns of the DTA problem are $\varphi, \varpi$.
Travel costs equal travel times plus early/late arrival time penalty. Thus it is necessary to estimate the travel costs. Consider a path $p = (a_0, a_1, \ldots, a_K)$ and denote $\ell_k \overset{def}{=} \ell_{a_{k-1},a_k}, v_k \overset{def}{=} v_{a_{k-1},a_k}$. The instantaneous travel time of $p$ is given by:

$$ITT_p(t) = \sum_{k=1}^{K} \ell_k/v_k \tag{7}$$

The predictive travel time $PTT$ is calculated a posteriori. One possible idea is to use the tools of Jin [2020] and Ameli et al. [2022]. First define

$$z_k(t) = \int_0^t v_k(s)ds \tag{8}$$

then the recursion

$$\left|\begin{array}{rcl} t_0 &=& t \\ &\vdots& \\ t_k &=& z_k^{-1}\left(z_{k-1}(t_{k-1}) + \ell_k\right) \\ &\vdots& \\ t_K &=& z_K^{-1}\left(z_{K-1}(t_{K-1}) + \ell_K\right) \end{array}\right. \tag{9}$$

yields the predictive travel time $PTT_p(t) = t_K$. The path predictive travel cost results by adding the late/early arrival time penalty, which will be a function of the difference $t_K - t_a$. The OD predictive travel cost will be obtained as the expectation of path predictive travel costs for all paths of the OD. The expectation is then calculated with respect to $\varpi$.
Finally the equilibrium is obtained following the method described in Khoshyaran and Lebacque [2023, 2025].

# 4 Numerical experiments, outlook

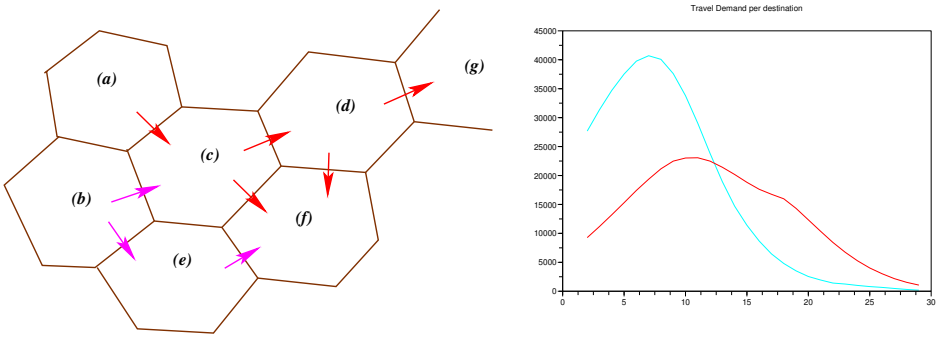



Figure 2: Left: network with 7 reservoirs, 2 destinations, 2 origins, 7 paths. Right: cumulated departure time distribution (for the two destinations) at equilibrium, with Logit assignment.
The model will be implemented and compared to the bidimensional model developed in Khoshyaran and Lebacque [2025], notably on the network depicted in Figure 2. The classical issues: existence of equilibria, discretization of the model, will also be studied. The right part of Figure 2 shows some results: stochastic user equilibrium was achieved after some 30 iterations and the resulting equilibrium departure time distributions (the $\varphi$ are depicted per destination and cumulated over the two origins).

# References


Rafegh Aghamohammadi and Jorge Laval. A continuum model for cities based on the macroscopic fundamental diagram: A semi-lagrangian solution method. *Transportation Research Procedia*, 38:380–400, 2019.

Mostafa Ameli, Jean-Patrick Lebacque, and Ludovic Leclercq. Cross-comparison of convergence algorithms to solve trip-based dynamic traffic assignment problems. *Computer-Aided Civil and Infrastructure Engineering*, 35(3):219–240, 2020.

Mostafa Ameli, Mohamad Sadegh Shirani Faradonbeh, Jean-Patrick Lebacque, Hossein Abouee-Mehrizi, and Ludovic Leclercq. Departure time choice models in urban transportation systems based on mean field games. *Transportation Science*, 56(6):1483–1504, 2022.

Louis Balzer, Mostafa Ameli, Ludovic Leclercq, and Jean-Patrick Lebacque. Dynamic tradable credit scheme for multimodal urban networks. *Transportation Research Part C: Emerging Technologies*, 149:104061, 2023.

C Buisson, JB Lesort, and JP Lebacque. Macroscopic modelling of traffic flow and assignment in mixed networks. In *Computing in Civil and Building Engineering, Proceedings of the sixth International Conference on Computing in Civil Engineering, Berlin, Germany 12-15 July 1995. Volume 2*, 1995.

Mohammad Ehteram, Hojat Karami, Sayed-Farhad Mousavi, Saeed Farzin, and Ozgur Kisi. Optimization of energy management and conversion in the multi-reservoir systems based on evolutionary algorithms. *Journal of cleaner production*, 168:1132–1142, 2017.

N. Geroliminis and C. Daganzo. Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transportation Research Part B: Methodological*, 42:759–770, 2008. URL https://dx.doi.org/10.1016/j.trb.2008.02.002.

N. Geroliminis, J. Haddad, and M. Ramezani. Optimal perimeter control for two urban regions with macroscopic fundamental diagrams: a model predictive approach. *IEEE Transactions on Intelligent Transportation Systems*, 14:348–359, 2013. URL https://dx.doi.org/10.1109/TITS.2012.2216877.

Omar Hassanin. Integrating queue dynamics into the trip-based macroscopic fundamental diagram. 2025.

HW Ho and SC Wong. Two-dimensional continuum modeling approach to transportation problems. *Journal of Transportation Systems Engineering and Information Technology*, 6(6):53–68, 2006.

YP Huang, JH Xiong, A Sumalee, N Zheng, WHK Lam, ZB He, and RX Zhong. A dynamic user equilibrium model for multi-region macroscopic fundamental diagram systems with time-varying delays. *Transportation Research Part B: Methodological*, 131:1–25, 2020.

Yunping Huang, Jianhui Xiong, Shu-Chien Hsu, Agachai Sumalee, William Lam, and Renxin Zhong. A comparison of the accumulation-based, trip-based and time delay macroscopic fundamental diagram models. *Transportmetrica A: Transport Science*, pages 1–37, 2024.

Yanqun Jiang, SC Wong, HW Ho, Peng Zhang, Ruxun Liu, and Agachai Sumalee. A dynamic traffic assignment model for a continuum transportation system. *Transportation Research Part B: Methodological*, 45(2):343–363, 2011.

Wen-Long Jin. Generalized bathtub model of network trip flows. *Transportation Research Part B: Methodological*, 136:138–157, 2020.

Megan M Khoshyaran and Jean-Patrick Lebacque. Continuum traffic flow modelling: network approximation, flow approximation. In *Traffic and Granular Flow 2019*, pages 505–513. Springer, 2020.

Megan M Khoshyaran and Jean-Patrick Lebacque. Complex dynamics generated by simultaneous route and departure time choice in transportation networks. In *Chaotic Modeling and Simulation International Conference*, pages 313–330. Springer, 2023.

Megan M Khoshyaran and Jean-Patrick Lebacque. Dynamic traffic assignment in a bi-dimensional model. *Transportmetrica B: Transport Dynamics*, 13(1):2450501, 2025.

G. Mariotte, L. Leclercq, and J. A. Laval. Macroscopic urban dynamics: Analytical and numerical comparisons of existing models. *Transportation Research Part B*, 101:245–267, 2017. URL https://dx.doi.org/10.1016/j.trb.2017.04.002.

Guilhem Mariotte and Ludovic Leclercq. The mfd trip-based approach applied to multi-reservoir systems. In *hEART2017-6th symposium of the European Association for Research in Transportation*, pages 3–p, 2017.

Stéphane Mollier, Maria Laura Delle Monache, and Carlos Canudas de Wit. A step towards a multidirectional 2d model for large scale traffic networks. In *98th TRB*, pages 1–7, 2019.

Tibye Saumtally, Jean-Patrick Lebacque, and Habib Haj-Salem. A dynamical two-dimensional traffic model in an anisotropic network. *Networks and Heterogeneous Media*, 8(3):663–684, 2013.

Azuma Taguchi and Masao Iri. Continuum approximation to dense networks and its application to the analysis of urban road networks. *Applications*, pages 178–217, 1982.

Mehmet Yildirimoglu, Mohsen Ramezani, and Nikolas Geroliminis. Equilibrium analysis and route guidance in large-scale networks with mfd dynamics. *Transportation Research Procedia*, 9:185–204, 2015.

# Integrating Dynamic Traffic Assignment in a Digital Twin to Support Intelligent Traffic Management Under Disruptions

Zerun Liu[a], Zilin Bian[a], Fan Zuo[a], Dachuan Zuo[a], Kaan Ozbay[a]
[a]Dept. of Civil and Urban Engineering, NYU

**Keywords:** Multi-scale DTA, DTALite, Real-time Data, AI models, Traffic Management

## 1. Introduction

Real-time urban traffic management has to operate in a complex, dynamic environment characterized by complex network conditions, various data types, multiple spatial scales, and diverse external disturbances such as accidents, incidents, and extreme events. In developing and deploying an intelligent traffic system to deal with that multifaceted environment, dynamic traffic assignment (DTA) plays an essential role. The main role of dynamic traffic assignment is to reproduce vehicles' movements according to certain behavioral rules given the prevailing traffic conditions [1]. By capturing time-varying route choices and network flows, DTA yields realistic traffic representations that underpin models for emissions, safety, and resilience, thereby supporting comprehensive policy analysis and decision-making.

However, there are two primary challenges when applying DTA in the context of intelligent traffic management systems. First, integrating multi-level DTA models in a digital twin remains difficult. Traffic modeling spans macroscopic models that capture aggregate flows, mesoscopic models that balance detail and efficiency, and microscopic models that simulate individual vehicle interactions. Seamlessly coupling these scales while producing results that are consistent with the real world is challenging due to differences in representation and computational demands. Second, real-time data integration can be problematic. Although traffic data is abundant, the inherent noise and heterogeneity complicate the continuous calibration, validation, and updating tasks of various components of the underlying components of DTA models.

To overcome these challenges and advance the practice of intelligent traffic management, we propose a hybrid DTA-based digital twin framework that integrates multi-scale DTA with AI-based real-time data processing and AI-driven decision-support functions for online deployment. The framework comprises four modules: AI-based data ingestion and fusion, multi-scale DTA, on and offline calibration of DTA, and AI-powered decision support. A key contribution is the integration of DTA Lite developed by Zhou et al. (2014) [2] into the DT-ATeam platform developed by Bian et al. (2024) [3, 4], transforming an offline planning tool into a dynamic, real-time system capable of generating and evaluating multiple scenarios, especially in the presence of major incidents and extreme events. By utilizing real-time sensing and AI predictions, this framework transforms diverse sensor data into structured DTA inputs. It connects various simulation scales and aligns outputs with field observations, providing actionable insights for traffic planning, control, and service. This approach significantly enhances real-time urban traffic management.

## 2. Related Work

Based on extensive studies [1, 5, 6, 7, 8], the network modeling models in DTA (traffic flow simulation software serves as the network loading component in DTA [9, 1]) can be classified into three levels: macroscopic (e.g., (M–N model) [10], link performance functions [11], METANET [12], the Lighthill–Whitham–Richards (LWR) model [13, 14], and simplified traffic flow representations as seen in MATSim [15]), mesoscopic (e.g., MATSim, DTALite[2], and certain modules of SUMO [16]), and microscopic (e.g., VISSIM [17], MITSIMLab[18], and SUMO). To better combine the advantages of multilevel DTA models, more researchers have advanced multiscale DTA frameworks [19, 20, 21, 22, 5, 8]. [23] developed a feedback loop that integrates strategic route planning with microsimulation for daily replanning under congestion. [8] proposed a co-simulation framework that connects traffic and network simulations via a run-time infrastructure and federation object model, enhancing interoperability and enabling analysis of wireless vehicle communication impacts. [5] extended [24]'s framework by incorporating vehicles and using the CRIO meta-model to automatically adjust simulation resolution

across microscopic, mesoscopic, and macroscopic scales through an influence-reaction mechanism. [9] introduced a simulator that directly integrates macroscopic and microscopic models into a differentiable framework by computing gradients for traffic states, thereby accelerating optimization and improving scalability for complex traffic control problems.

In summary, current research on hybrid DTA highlights several potential areas for improvement. First, despite advances in multi-level DTA, a comprehensive AI-driven framework that effectively bridges microscopic, mesoscopic, and macroscopic DTA results into a digital twin is still needed. Second, aligning DTA inputs and calibrations with real-time field data remains a challenge. Reliable methods are necessary to convert raw sensor data into actionable DTA inputs and to continuously adjust DTA models with real-time observations. Third, it is necessary to translate multiscale DTA outputs into actionable guidance for traffic real-time management, as systematic frameworks linking DTA results with ancillary models (e.g., for emissions, safety, or resilience) are absent. Addressing these areas for improvement is crucial for fully realizing the potential of hybrid DTA-based digital twin frameworks in both operational and strategic traffic.

## 3. Methodology

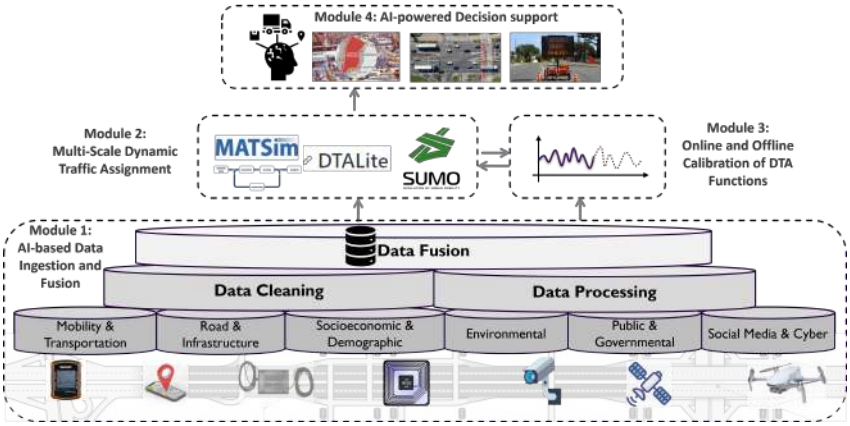



Figure 1: Framework

Based on DT-ATEAM developed by [3, 4], we propose a hybrid DTA-based digital twin framework for intelligent traffic management, as shown in Fig. 1. The framework is composed of four key modules: (1) AI-based data ingestion and fusion, (2) multi-scale DTA, (3) on- and off-line calibration of DTA functions, (4) AI-powered decision support.

### 3.1 Data Ingestion and Fusion

Our framework begins with a data ingestion task that continuously acquires real-time information from heterogeneous sensors, including mobile devices (e.g., smartphone GPS, fleet telematics), infrastructure sensors (e.g., inductive loop detectors, RFID, Bluetooth), and spatial sensors (e.g., cameras, UAVs, satellite imagery). The ingested data are then cleaned and standardized using anomaly detection, imputation, and normalization. Outliers are identified with clustering or autoencoder models. We use Kalman filtering and interpolation methods (e.g., linear, spline, and polynomial interpolation) to addresses the issues related to missing or erroneous values. Camera feeds are processed with CNN-based object detection to classify incidents, and text sources are parsed with natural language processing to extract key details. Finally, AI-driven fusion methods aggregate these data into unified traffic-state variables (e.g., link speed, density, flow) at specified temporal and spatial resolutions. Ensemble techniques, such as gradient boosting, random forests, and neural networks, synthesize redundant inputs to yield robust predictions, which serve as the primary input for our hybrid DTA environment and as a reference for calibrating DTA outputs.

### 3.2 Multi-Scale Dynamic Traffic Assignment with Bridging Mechanisms

To capture both regional congestion patterns and localized vehicle interactions, our framework integrates a macroscopic DTA model (e.g., MATSim) for network-wide demand and route assignment, a mesoscopic model (e.g., DTALite) for intermediate resolution, and a microscopic model (e.g., SUMO) for vehicle-level dynamics. Achieving consistency across these scales in near real time requires a dedicated *bridging mechanism* that periodically exchanges information between the finer (micro/meso) and coarser (macro) simulations.

Within each synchronization interval $[t, t + \Delta t]$, the microscopic or mesoscopic DTA produces high-resolution traffic states for every link $l \in \mathcal{L}$ in the network. Let $n_l(t)$ be the number of vehicles on link $l$ at time $t$, and let $v_i(t)$ be the instantaneous speed of vehicle $i$. The model also tracks $\Delta n_l(t)$, the count of vehicles that completely traverse link $l$ in $[t, t + \Delta t]$. These raw outputs are condensed into average speed, density, and flow:

$$\bar{v}_l(t) = \frac{1}{n_l(t)} \sum_{i=1}^{n_l(t)} v_i(t), \tag{1}$$

$$\bar{k}_l(t) = \frac{n_l(t)}{\ell_l}, \tag{2}$$

$$\bar{q}_l(t) = \frac{\Delta n_l(t)}{\Delta t}, \tag{3}$$

where $\ell_l$ is the length of link $l$. We denote these aggregated quantities collectively as $\mathbf{z}_l(t) = \big(\bar{v}_l(t), \bar{k}_l(t), \bar{q}_l(t)\big)$. A bridging function $\Phi_{\boldsymbol{\theta}}$ then maps $\mathbf{z}_l(t)$ to macroscopic parameters for each link:

$$\mathbf{x}_l^{\text{macro}}(t) = \Phi_{\boldsymbol{\theta}}\big(\mathbf{z}_l(t)\big), \tag{4}$$

where $\mathbf{x}_l^{\text{macro}}(t)$ could include fundamental diagram coefficients (e.g., free-flow speed, critical density, jam density) or link performance function parameters. By updating these parameters at each synchronization interval, the macro model remains informed of fine-scale congestion and incident impacts.

Conversely, the macroscopic or mesoscopic model outputs aggregated flows and route assignments for each link $l$. Let $F_l(t)$ (vehicles per hour) be the assigned flow and $\{\alpha_r(t)\}$ the route choice splits. We define a disaggregation function $\Psi_{\boldsymbol{\omega}}$ that interfaces with a well-calibrated SUMO network model [25] to generate individual vehicle injections. First, we convert the hourly flow into an instantaneous arrival rate

$$\lambda_l(t) \;=\; \frac{F_l(t)}{3600}, \tag{5}$$

which determines how often vehicles are spawned in SUMO. Specifically,

$$\mathbf{x}_l^{\text{micro}}(t + \Delta t) \;=\; \Psi_{\boldsymbol{\omega}}\big(F_l(t), \{\alpha_r(t)\}\big) \tag{6}$$

encodes the set of vehicle injection events, including spawn times (when the vehicle enters the system) based on calculated arrival time, alternative routes, and other behavioral attributes. Inside SUMO—which requires detailed lane configurations, calibrated car-following models, and signal timing plans—each newly created vehicle follows realistic microscopic rules, such as time-varying lane changes and intersection queueing. These behavioral models are calibrated to be consistent with local driving conditions and geometry to ensure realistic speed and headway distributions. If network-level flows or route splits $\{\alpha_r(t)\}$ change significantly, SUMO can also trigger route updates during dynamic traffic assignment, thus preserving consistency with macro-level re-routing decisions. This two-way link between macro and micro ensures that large-scale demand shifts and localized traffic dynamics remain synchronized over time.

### 3.3 Online and Offline Calibration

While domain knowledge can provide an initial approximation of $\Phi_{\boldsymbol{\theta}}$ and $\Psi_{\boldsymbol{\omega}}$, we employ a two-phase calibration strategy—off-line deep learning [25] followed by on-line stochastic approximation[26]—to improve accuracy in reflecting real-world conditions.

First, we compile a labeled dataset $\mathcal{D}$ that associates micro-level aggregates $\mathbf{z}_l = (\bar{v}_l, \bar{k}_l, \bar{q}_l)$ with the corresponding macro-level parameters $\mathbf{x}_l^{\text{macro}}$, and macro-level flows $\mathbf{x}_l^{\text{macro}}$ with micro-level injection patterns $\mathbf{x}_l^{\text{micro}}$. We then train two neural networks, $\Phi_{\boldsymbol{\theta}}$ and $\Psi_{\boldsymbol{\omega}}$, to minimize mean-squared error losses of the form

$$\mathcal{L}_{\Phi}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{z}_l, \mathbf{x}_l^{\text{macro}}) \in \mathcal{D}} \left\| \Phi_{\boldsymbol{\theta}}(\mathbf{z}_l) - \mathbf{x}_l^{\text{macro}} \right\|^2, \tag{7}$$

$$\mathcal{L}_{\Psi}(\boldsymbol{\omega}) = \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_l^{\text{macro}}, \mathbf{x}_l^{\text{micro}}) \in \mathcal{D}} \left\| \Psi_{\boldsymbol{\omega}}(\mathbf{x}_l^{\text{macro}}) - \mathbf{x}_l^{\text{micro}} \right\|^2. \tag{8}$$

These models are trained using gradient-based optimization (e.g., Adam), with domain-specific constraints (e.g., maximum flow capacities) embedded as penalty terms or hard limits. This offline initialization produces parameter sets $(\boldsymbol{\theta}_0, \boldsymbol{\omega}_0)$ that capture normal traffic conditions prior to live deployment.

Once deployed, the bridging functions are further refined to accommodate unanticipated events or shifts in traffic demand. We employ *Simultaneous Perturbation Stochastic Approximation* (SPSA) [25], which approximates the gradient of an online cost function $\mathcal{J}(\boldsymbol{\theta}, \boldsymbol{\omega})$ by perturbing parameters in random directions and observing changes in $\mathcal{J}$. The cost function measures discrepancies between observed field measurements and the hybrid DTA's predicted traffic states after applying $\Phi_{\boldsymbol{\theta}}$ and $\Psi_{\boldsymbol{\omega}}$. For each synchronization interval,

$$\mathcal{J}(\boldsymbol{\theta}, \boldsymbol{\omega}) = \sum_{t \in \mathcal{T}} \left\| \mathbf{z}_{\text{obs}}(t) - \hat{\mathbf{z}}(t \mid \Phi_{\boldsymbol{\theta}}, \Psi_{\boldsymbol{\omega}}) \right\|^2 + \lambda R(\boldsymbol{\theta}, \boldsymbol{\omega}), \tag{9}$$

where $\mathbf{z}_{\text{obs}}(t)$ are sensor-derived traffic states, $\hat{\mathbf{z}}(t)$ are simulated states, and $R(\cdot)$ is a regularization term that prevents physically infeasible outcomes. By iteratively adjusting $(\boldsymbol{\theta}, \boldsymbol{\omega})$ to minimize $\mathcal{J}$, the bridging functions remain aligned with real-time conditions, even under incident or off-peak scenarios.

## 4. Results

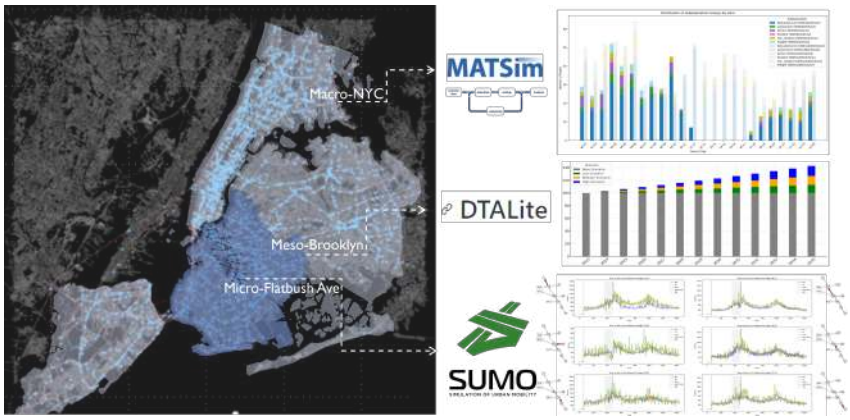



Figure 2: Hybrid DTA

The proposed hybrid DTA-based digital twin framework was applied to New York City (NYC). The data ingestion module captured and aggregated real-time traffic data from multiple sources across NYC. Traffic information on travel speeds and volumes was obtained from the NPMRDS dataset [27] at 5-, 15-, and 60-minute intervals, while infrastructure data from NPMRDS and the NYSDOT GIS provided roadway geometry and attributes such as lane count and width. Incident data were acquired from real-time feeds (e.g., 511NY [28]) and offline NYCDOT [29] files, supplying event type, location, and duration. Raw data were processed using AI-based anomaly detection, clustering, and classification; for instance, a deep-learning object detection model automatically recognized and quantified work zones

[30]. Finally, event-road and traffic-road matching methods fused these outputs into an integrated data format.

The hybrid DTA module models traffic dynamics across three scales. At the macroscopic level, MATSim simulates the entire NYC and provides activity patterns for different population groups, capturing large-scale travel demand and behavior. At the mesoscopic level, DTALite focuses on the Brooklyn borough, generating multiple scenario analyses to evaluate the impacts of various policies and situations. At the microscopic level, SUMO simulates detailed vehicle interactions along Flatbush Avenue, yielding high-fidelity vehicular information. The outputs from these scales are then integrated to ensure consistency and provide a comprehensive multilevel view of urban traffic dynamics, as shown in Fig. 2. The calibration module continuously adjusts the DTA in real-time by using CV-enabled detectors to update incidents, network sensors to synchronize the traffic assignment, and AI models to forecast traffic flows and guide parameters [31].

Finally, the decision support module turns calibrated DTA outputs into actionable insights. The activity comparisons, scenario analyses, and detailed vehicle movement data produced by the hybrid DTA model are integrated with additional models like major incidents and extreme events. An example of decision-support application is shown in Fig. 3. Based on aggregated multi-scale DTA results (e.g., travel time and density), users can apply a customized AI model (e.g., deep recurrent neural network) to predict traffic crashes. The predictions generated offer spatial and temporal estimates of crash occurrences, which can be visualized on the system dashboard. This information supports proactive policy interventions, such as targeted infrastructure improvements, revised speed limits, and focused enforcement strategies. In parallel, the digital twin reflects real-time traffic impacts of crashes (e.g., reduced speeds and increased densities) to enable immediate operational responses. AI methods (e.g., deep Q-learning, PPO) process this potential real-time feedback to continuously adjust adaptive signal control and dynamic rerouting strategies. This enables traffic managers to swiftly mitigate congestion and enhance safety during incident situations.

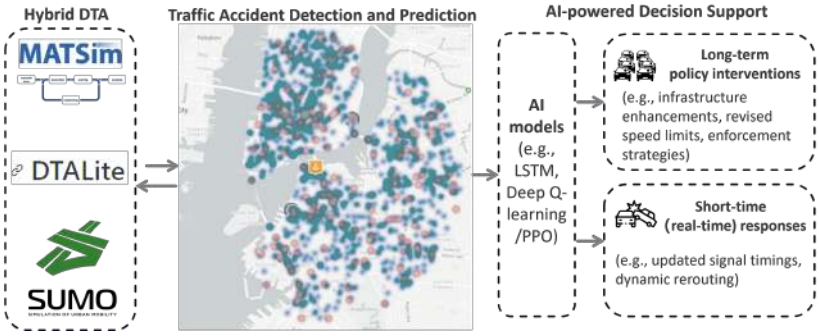



Figure 3: An example of AI-powered decision support in traffic accident management

## 5. Conclusion and Future Work

This study proposed a hybrid DTA-based digital twin framework for intelligent traffic management. The framework integrates multi-scale DTA models—employing MATSim at the macroscopic level, DTALite at the mesoscopic level, and SUMO at the microscopic level—with real-time sensor data and AI-enhanced calibration. By using AI-driven fusion techniques, our framework can convert heterogeneous sensor inputs into structured traffic-state variables. In addition, our framework ensures seamless aggregation and disaggregation between different DTA scales, while continuous calibration via online stochastic approximation aligns DTA outputs with field observations.

Future work will concentrate on improving the robustness of bridging mechanisms by utilizing advanced AI techniques. We will expand the framework to include additional traffic management scenarios and integrate more ancillary models, such as emissions, resilience, and safety. This will support comprehensive, real-time, multi-objective decision-making for urban transportation systems.

## References


[1] Yi Wang, Wai Y Szeto, Ke Han, and Terry L Friesz. Dynamic traffic assignment: A review of the methodological advances for environmentally sustainable road transportation applications. *Transportation Research Part B: Methodological*, 111:370–394, 2018.

[2] Xuesong Zhou and Jeffrey Taylor. Dtalite: A queue-based mesoscopic traffic simulator for fast model evaluation and calibration, 2014.

[3] Tao Li, Zilin Bian, Haozhe Lei, Fan Zuo, Ya-Ting Yang, Quanyan Zhu, Zhenning Li, Zhibin Chen, and Kaan Ozbay. Digital twin-based driver risk-aware intelligent mobility analytics for urban transportation management. *arXiv preprint arXiv:2407.15025*, 2024.

[4] Zilin Bian, Dachuan Zuo, Jingqin Gao, Kaan Ozbay, and Matthew D Maggio. A-team: Advanced traffic event analysis and management platform for transportation data-driven problem solving. *arXiv preprint arXiv:2406.15452*, 2024.

[5] Igor Tchappi Haman, Vivient Corneille Kamla, Stéphane Galland, and Jean Claude Kamgang. Towards an multilevel agent-based model for traffic simulation. *Procedia Computer Science*, 109:887–892, 2017.

[6] Hesham Rakha, Aly M Tawfik, and RA Meyers. Traffic networks: Dynamic traffic routing, assignment, and assessment., 2009.

[7] Serge P Hoogendoorn and Piet HL Bovy. State-of-the-art of vehicular traffic flow modelling. *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, 215(4):283–303, 2001.

[8] Moritz Gütlein, Reinhard German, and Anatoli Djanatliev. Towards a hybrid co-simulation framework: Hla-based coupling of matsim and sumo. In *2018 IEEE/ACM 22nd international symposium on distributed simulation and real time applications (ds-rt)*, pages 1–9. IEEE, 2018.

[9] Sanghyun Son, Yi-Ling Qiao, Jason Sewall, and Ming C Lin. Differentiable hybrid traffic simulation. *ACM Transactions on Graphics (TOG)*, 41(6):1–10, 2022.

[10] Deepak K Merchant and George L Nemhauser. A model and an algorithm for the dynamic traffic assignment problems. *Transportation science*, 12(3):183–199, 1978.

[11] Bruce N Janson. Dynamic traffic assignment for urban road networks. *Transportation Research Part B: Methodological*, 25(2-3):143–161, 1991.

[12] Albert Messmer and Markos Papageorgiou. Metanet: A macroscopic simulation program for motorway networks. *Traffic engineering & control*, 31(9), 1990.

[13] Michael James Lighthill and Gerald Beresford Whitham. On kinematic waves ii. a theory of traffic flow on long crowded roads. *Proceedings of the royal society of london. series a. mathematical and physical sciences*, 229(1178):317–345, 1955.

[14] Paul I Richards. Shock waves on the highway. *Operations research*, 4(1):42–51, 1956.

[15] Kay W Axhausen, Andreas Horni, and Kai Nagel. *The multi-agent transport simulation MATSim*. Ubiquity Press, 2016.

[16] Daniel Krajzewicz, Michael Bonert, and Peter Wagner. The open source traffic simulation package sumo. *RoboCup 2006*, 2006.

[17] Martin Fellendorf. Vissim for traffic signal optimisation. *Traffic Technology International'96. Annual Review Issue*, 1996.

[18] Qi Yang, Haris N Koutsopoulos, and Moshe E Ben-Akiva. Simulation laboratory for evaluating dynamic traffic management systems. *Transportation Research Record*, 1710(1):122–130, 2000.

[19] Pushpendra Kumar, Rochdi Merzouki, Blaise Conrard, Vincent Coelen, and Belkacem Ould Bouamama. Multilevel modeling of the traffic dynamic. *IEEE Transactions on Intelligent Transportation Systems*, 15(3):1066–1082, 2014.

[20] Daniel Zehe, David Grotzky, Heiko Aydt, Wentong Cai, and Alois Knoll. Traffic simulation performance optimization through multi-resolution modeling of road segments. In *Proceedings of the 3rd ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, pages 281–288, 2015.

[21] Luca Crociani, Gregor Lämmel, and Giuseppe Vizzari. Multi-scale simulation for crowd manage-

ment: a case study in an urban scenario. In *Autonomous Agents and Multiagent Systems: AAMAS 2016 Workshops, Best Papers, Singapore, Singapore, May 9-10, 2016, Revised Selected Papers*, pages 147–162. Springer, 2016.

[22] Pitu B Mirchandani, Pengfei Li, Xuesong Zhou, et al. Integrating meso-and micro-simulation models to evaluate traffic management strategies, year 2. 2017.

[23] Ana LC Bazzan, Kai Nagel, and Franziska Klügl. Integrating matsim and itsumo for daily replanning under congestion. In *Proceedings of the 35th Latin-American Informatics Conference, CLEI*, 2009.

[24] Nicolas Gaud, Stéphane Galland, Franck Gechter, Vincent Hilaire, and Abderrafiaa Koukam. Holonic multilevel simulation of complex systems: Application to real-time pedestrians simulation in virtual urban environment. *Simulation Modelling Practice and Theory*, 16(10):1659–1676, 2008.

[25] Di Sha, Kaan Ozbay, and Yue Ding. Applying bayesian optimization for calibration of transportation simulation models. *Transportation Research Record*, 2674(10):215–228, 2020.

[26] Yu Tang, Li Jin, and Kaan Ozbay. Physics-informed machine learning for calibrating macroscopic traffic flow models. *Transportation Science*, 58(6):1389–1402, 2024.

[27] University at Albany Visualization And Informatics Lab. National performance management research dataset (npmrds). https://npmrds.availabs.org/landing/, 2022. Accessed: July 1, 2022.

[28] New York Department of Transportation. New york state's 511 services. https://511ny.org/map/, 2022. Accessed: July 1, 2022.

[29] New York Department of Transportation. Nys gis clearinghouse. http://gis.ny.gov/, 2022. Accessed: July 1, 2022.

[30] Fan Zuo, Jingqin Gao, Kaan Ozbay, Zilin Bian, and Daniel Zhang. Urban work zone detection and sizing: A data-centric training and topology-based inference approach. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 3235–3240. IEEE, 2023.

[31] Tao Li, Zilin Bian, Haozhe Lei, Fan Zuo, Ya-Ting Yang, Quanyan Zhu, Zhenning Li, and Kaan Ozbay. Multi-level traffic-responsive tilt camera surveillance through predictive correlated online learning. *Transportation Research Part C: Emerging Technologies*, 167:104804, 2024.

# Localized Queue Spillback Modeling for Dynamic Traffic Assignment with Uncertain Demand

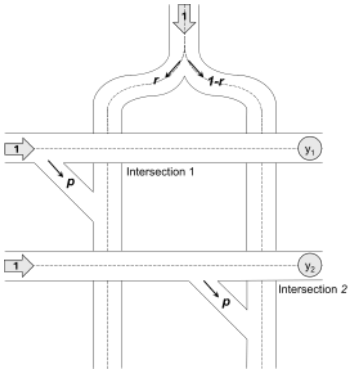
Jake Robbennolt, Debojjal Bagchi, and Stephen D. Boyles

## 1 Introduction

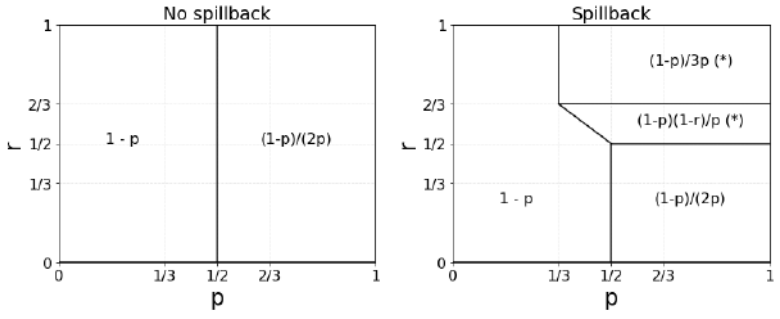
The use of dynamic traffic assignment (DTA) models has increased in recent years as they can capture temporal fluctuations in demand and passenger flows and accurately represent phenomena like shockwaves and queue spillback [3]. Given exogenous demand and turning proportions, the goal is to model the temporal evolution of traffic conditions, including congestion patterns, travel times, and queue formation, while adhering to traffic flow principles like flow conservation and capacity constraints [6]. In particular, queue spillback is a major concern with far-reaching impacts. When queues propagate upstream and block additional links, the consequences can cascade through entire transportation networks, potentially triggering widespread gridlock. These issues are also well documented on larger networks where queue spillback has been studied in relation to roadway geometry, ramp metering, signal timing, and work zones [2, 5, 7]. An extreme case is large-scale evacuations when extreme demand fluctuations can be much higher than the capacity of the roadway. The substantial reductions in efficiency due to queue spillback make it critical to incorporate into planning models.

However, methodological research suggests that this can cause convergence concerns and paradoxical behavior, even on small networks [1, 4]. Recent research suggests that more realistic models may be less robust to errors in input data [1]. In particular, Boyles and Ruiz Juri [1] found that models which include queue spillback can show worse performance when there is high uncertainty in input demand. This counterintuitive result suggests a fundamental vulnerability in current traffic modeling approaches, particularly those that attempt to make planning-level predictions far in the future. These findings indicate that while queue spillback is crucial to model in many situations, models of this phenomenon require a higher level of accuracy in input data. Unfortunately, there are many situations when it is impossible to precisely predict input data. In these scenarios, more care should be taken to evaluate where spillback should be incorporated to increase model robustness.

In this paper, we operationalize the key insight from Boyles and Ruiz Juri [1] to create a more nuanced approach to modeling spillback when errors in input data can be quantified. Our proposed approach seeks to strike a delicate balance: creating a model that can accommodate spillback at locations where it is highly likely to occur while maintaining robustness against errors in input data. Boyles and Ruiz Juri [1] present several experiments that suggest that when demand is sufficiently high relative to errors in input data, models with spillback are still preferred. To capitalize on this insight, we develop a probabilistic procedure that draws individual demand estimates from known distributions to systematically analyze spillback potential. We hypothesize that the actual distribution of errors has minimal impacts on our proposed approach as incorporating more information into the modeling process can improve estimations even if that information is incomplete, though this will be tested in the full paper. The key innovation lies in our ability to estimate the likelihood of spillback occurring on individual network links. By allowing spillback only at those locations we demonstrate that our models can be tailored to accurately predict traffic flows across a wide range of scenarios.




(a) Freeway interchanges: mainline capacities are 1, ramp capacities are 1/2, $r$ and $p$ are model parameters.




(b) Steady state $y_1$ values for spillback and no-spillback cases as a function of $r$ and $p$. This solution is the same for $y_2$ when $r$ is replaced by $1 - r$. Asterisks indicate queue spillback from the ramp.

Figure 1: (a) Network schematic and (b) steady state flow values

## 2 Motivating Example

Our proposed approach finds an intermediate solution between the model without queue spillback and the model with full queue spillback. We suggest that in some networks, even if there is uncertainty in the incoming demand, some queues are likely to spill back. At other locations, queue formation may be less consistent. By modeling spillback at only certain locations we can achieve a solution that is robust to errors in input data and is more realistic than the no-spillback alternative.

To demonstrate this idea, we develop a toy example that has two interchanges that are linked (See Figure 1a). Each intersection consists of a diverge and then a merge. Both ramps have the same splitting proportion from their Eastbound mainline $p$ which we will allow to vary. The ramps have a capacity of 1/2 and all mainlines have a capacity of 1. Each eastbound mainline has a fixed inflow of 1 unit. The southbound roads have a total inflow of 1 unit, which is split with $r$ headed towards intersection 1 and $1 - r$ headed towards intersection 2. We assume the diverge of the two southbound mainlines will never become congested (i.e. it is very long relative to the length of the ramps). This means that the only locations where spillback will cause predictions to differ are the ramps (at either intersection 1 or 2). Then, we will try to predict the outflows $y_1$ and $y_2$ as a function of $p$ and $r$ (the outflows on the southbound links will be the same regardless of queue spillback). Figure 1b shows the steady state outflow solution for intersection 1. Intersection 2 has the same solution except incoming flows are based on $1 - r$ from the north.

To determine the best model, we first take 20 evenly spaced values of $\hat{p}$ and $\hat{r}$ inside the unit square. For each of the resulting 400 scenarios, we generate $n = 2,500$ samples of $p$ and $r$ from independent normal distributions using $\hat{p}$ and $\hat{r}$ as the means and a provided standard deviation. We assume that the model with spillback predicts the vector of flows $m$ perfectly at the input values with no error. Then, for each sample, we can calculate the root mean squared error $\epsilon^N$ between the vector of actual flows and predicted flows, where $N$ represents the network state (spillback, no spillback, spillback at intersection 1, or spillback at intersection 2). Using the 2500 samples for each scenario, we calculate the expected value $\delta$ and standard deviation $s$ of the difference in errors $\epsilon^\alpha - \epsilon^\beta$ between each pair of models $\alpha$ and $\beta$. These values are used to compute the t-score $t = \frac{\delta}{s/\sqrt{n}}$ for determining statistical significance.

Using a significance level of 5%, Figure 2 shows all locations where each model is not statistically significantly worse than any other model and all locations where one model is significantly better than all other models. When error is low, many models perform similarly across most of the scenarios, with the spillback model generally being the most versatile. However, as errors increase, there are many cases when it is better to model spillback only at a single intersection or not at all. In fact, the root mean squared error in flows is a conservative metric because when demand is over-predicted, spillback may reduce the outflow. Other metrics such as total system travel time will increase both as demand is over-predicted and as more queues spill back, exacerbating the results shown in 2 and suggesting that the models which apply spillback selectively have even better performance. This effect is shown on the Austin, Texas network in Section 4.

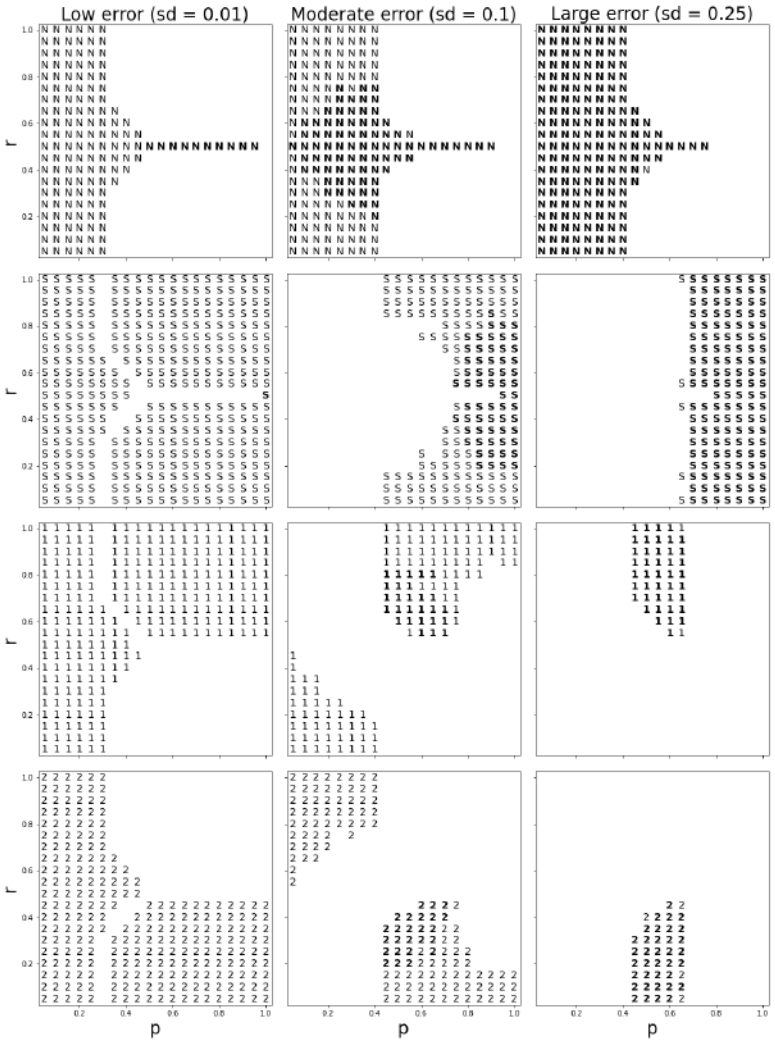



Figure 2: Scenarios where each model (S – spillback everywhere, N – no spillback, 1 – spillback on ramp 1, 2 – spillback on ramp 2) is not statistically significantly worse than any other model. Bold means statistically significantly better than *all* other models.

## 3 Localized Queue Spillback Algorithm

The example presented in the previous section suggests the need for a more flexible approach than allowing or disallowing spillback on the entire network. Queue spillback is a critical traffic phenomenon that must be modeled at locations where it is likely to occur. However, the discontinuities introduced by spillback can cause large prediction errors in certain locations under certain demand patterns. With these issues in mind, we present an algorithm to detect locations where spillback is likely. We assume that the modeler has some information about the distribution of possible demand matrices, rather than a single prediction. These could be generated through Monte Carlo simulation from other steps in the four-step planning model [8]. Using several demand samples, we can run dynamic network loading (DNL) and determine the likelihood of spillback occurring. Then, we can allow spillback only at those locations for the final DTA model implementation.

Algorithm 1 outlines the localized queue spillback (LQS) algorithm. In this algorithm, we start by running DTA at the

**Algorithm 1** Selective Queue Model Classification (LQS) for Dynamic Traffic Assignment

**Require:** Network $\mathcal{G}(\mathcal{N}, \mathcal{A})$, target demand $\hat{d}$, demand scenarios $\mathcal{D}$, number of levels $n$, spillback threshold $\delta$, convergence level $\epsilon$
**Ensure:** Sets of spillback links $S$ and probabilities $P$, path set $\Pi$, set of spillback counts $\mathcal{C}$

Initialize $S \leftarrow \emptyset$, $P[l] \leftarrow 0 \quad \forall l \in \mathcal{A}$
$\Pi \leftarrow$ RunDTA(network, $\hat{d}$, $S$)
**for** level$\leftarrow 1$ **to** $n$ **do**
    Initialize $\mathcal{C}[l] \leftarrow 0 \quad \forall l \in \mathcal{A}$
    **for all** $d \in \mathcal{D}$ **do**
        densities $\leftarrow$ RunDNL(network, $\Pi$, $d$, $S$)
        **for all** $l \in \mathcal{A}$ **do**
            **if** DetectSpillback($l$, densities) **then**
                $\mathcal{C}[l] \leftarrow \mathcal{C}[l] + 1$
    $\alpha = 1/(\text{level}+1)$
    **for all** $l \in \mathcal{A}$ **do**
        $P_{new}[l] = (1 - \alpha)P[l] + \alpha(\mathcal{C}[l]/n)$
    $S = \{l$ **for** $l \in \mathcal{A}$ **if** $P_{new}[l] \geq \delta\}$
    **if** $|P| - |P_{new}| < \epsilon$ **then**
        **break**
    $P \leftarrow P_{new}$
results $\leftarrow$ RunDTA(network, $\hat{d}$, $S$)
**return** results

target demand prediction under the assumption of no queue spillback. This will create a path set and turning proportions that will be used for all DNL runs, where demand is adjusted proportionally across all paths. Then, for each demand scenario, we run DNL and identify links where queue spillback should have occurred by examining the density of vehicles on that link. If this density exceeds the actual jam density, then queues would have spilled back under realistic jam density conditions. By aggregating results across all scenarios, we compute the probability of queues spilling back for each link. We then update the links to use the spillback model if the probability of spillback is greater than some threshold $\delta$. We repeat the entire process for several levels to refine the probability values. At each level, the same demand scenarios are rerun and probabilities are updated using the successive average of the probabilities from the previous iteration. The algorithm terminates either if the probabilities don't change by more than $\epsilon$ between levels or if a certain number of levels have been processed. Finally, using the final set of spillback links we rerun DTA at the target demand prediction to get the final solution.

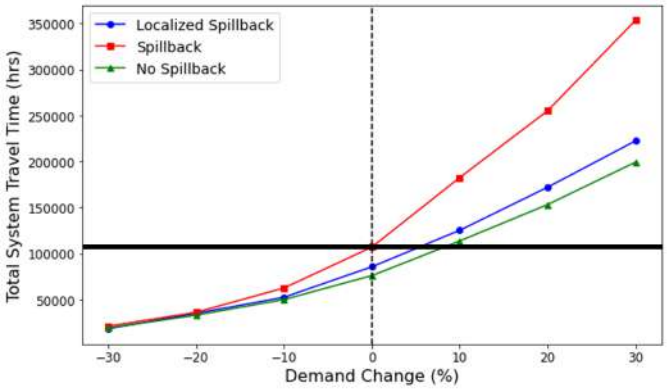
The algorithm allows many scenarios to be tested with only two full DTA runs, and additional DNL runs for other demand predictions. For many practical applications it is common to use the output of a single demand prediction, so this approach uses that same structure (ending with a model that is run for the target demand). However, by incorporating additional demand scenarios to modify where queue spillback is modeled, we make the results more robust to errors in input data. In addition, by using DNL rather than additional full DTA runs at each scenario, the algorithm can still be run quickly to evaluate many scenarios which may be computationally infeasible otherwise. Initial tests presented below replace the DNL with full DTA runs for initial testing, though this will be relaxed in the future.

## 4 Large-Scale Implementation

We now show the performance of the proposed model on the downtown Austin, TX network, a larger network with 1251 links, 546 nodes, and a total demand of 228,361 vehicles (depicted in Figure 3a). We use the VISTA software to solve the DTA model [9]. VISTA uses the cell transmission model for link models, modeling the available space at the end of the link based on the propagation of shockwaves. To model the no spillback scenario, we continue to use the cell transmission model, but set a very high jam density to allow arbitrarily long queues. To study the impact of queue spillback and test the LQS algorithm, we vary the incoming demand uniformly between -30% and +30%. We calculate the total system travel time (TSTT) using the model with spillback, without spillback, and using the localized spillback algorithm.

As a baseline, we solve the network to four iterations of the method of successive averages DTA. We use 7 demand samples (every 10% between -30% and 30%) and 6 levels in the LQS algorithm to select which links should have spillback. This requires a total of 42 DTA runs. Then, Figure 3b shows the resulting TSTTs for each model. **The model closest to the horizontal line (representing the TSTT when demand is predicted perfectly) is preferred.** As expected, the model with spillback performs poorly when demand is overestimated substantially. However, when demand is underestimated, the spillback model makes the best prediction. Of particular value is the slope of the localized spillback line as demand increases. While the slope of the spillback model line increases rapidly, the localized spillback line matches the no-spillback line when over-predictions are made. This is valuable in practice, particularly when demand is highly uncertain. In addition, when the demand is predicted perfectly, the localized spillback model outperforms the no-spillback model.

(a) Downtown Austin, TX network with spillback links in red and no-spillback links in black based on the localized queue spillback algorithm.




(b) TSTT for the downtown Austin, TX network as demand varies using the spillback, no spillback, and localized spillback models. Horizontal/vertical lines indicate the true TSTT value. **The model closest to the true value (horizontal line) is preferred.**

Figure 3: (a) Austin, TX Network schematic and (b) algorithm results.

## 5 Conclusion

Building on research demonstrating the challenges of modeling queue spillback in dynamic traffic assignment, this paper has presented a novel framework for selectively implementing spillback at locations where it is most likely to occur. Through both theoretical analysis and practical implementation, we show that, while spillback is important to model accurately, the phenomenon creates fundamental vulnerabilities in traffic modeling, particularly when there is uncertainty in demand predictions. The key innovation of our approach lies in recognizing that some locations are more prone to spillback than others, regardless of demand uncertainty. Our localized queue spillback algorithm capitalizes on this insight by systematically analyzing multiple demand scenarios to identify links where spillback is highly probable. The algorithm's effectiveness was demonstrated on the Austin, Texas network. When tested across a range of demand scenarios varying from -30% to +30%, the localized spillback model showed superior performance characteristics compared to both full spillback and no-spillback alternatives. Particularly noteworthy was the model's ability to match the spillback model's accuracy during demand under-prediction while maintaining the robustness of the no-spillback model during over-prediction scenarios. The input spillback probability threshold $\delta$ allows the model to be adapted to the modeler's risk tolerance.

The framework presented could be extended to incorporate additional stochastic elements such as stochastic fundamental diagrams to capture inherent traffic flow variability, probabilistic route choice models beyond Wardropian equilibrium, and time-varying reliability metrics. Such extensions would further enhance the model's ability to represent uncertainty in both supply and demand characteristics while maintaining computational tractability. In addition, future research should examine computational efficiency as the proposed approach requires many DNL runs. Other adaptations of link models or integrations of similar procedures directly into the DTA iterations may reduce the computational costs of this procedure.

## References


[1] Boyles, S. D. and Ruiz Juri, N. (2019). Queue Spillback and Demand Uncertainty in Dynamic Network Loading. *Transportation Research Record*, 2673(2):38–48.

[2] Chen, Y.-Y., Chen, Y.-H., and Chang, G.-L. (2021). Optimizing the integrated off-ramp signal control to prevent queue spillback to the freeway mainline. *Transportation Research Part C: Emerging Technologies*, 128:103220.

[3] Elimadi, M., Abbas-Turki, A., Koukam, A., Dridi, M., and Mualla, Y. (2024). Review of Traffic Assignment and Future Challenges. *Applied Sciences*, 14(2):683. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[4] Nie, Y. M. and Zhang, H. M. (2008). Oscillatory Traffic Flow Patterns Induced by Queue Spillback in a Simple Road Network. *Transportation Science*, 42(2):236–248.

[5] Noaeen, M., Mohajerpoor, R., H. Far, B., and Ramezani, M. (2021). Real-time decentralized traffic signal control for congested urban networks considering queue spillbacks. *Transportation Research Part C: Emerging Technologies*, 133:103407.

[6] Technical Activities Division, Transportation Research Board, and National Academies of Sciences, Engineering, and Medicine (2011). Dynamic Traffic Assignment: A Primer. ISBN: 9780309435291 Washington, D.C.

[7] Wong, C., Wong, S., and Lo, H. K. (2010). A spatial queuing approach to optimize coordinated signal settings to obviate gridlock in adjacent work zones. *Journal of Advanced Transportation*, 44(4):231–244.

[8] Zhao, Y. and Kockelman, K. M. (2002). The propagation of uncertainty through travel demand models: An exploratory analysis. *The Annals of Regional Science*, 36(1):145–163.

[9] Ziliaskopoulos, A. K. and Waller, S. T. (2000). An Internet-based geographic information system that integrates data, models and users for transportation applications. *Transportation Research Part C: Emerging Technologies*, 8(1):427–444.

DTA 2025 10TH INTERNATIONAL SYMPOSIUM ON DYNAMIC TRAFFIC ASSIGNMENT

# Optimizing Traffic Flow in Connected and Autonomous Vehicle Environments through Lane Drop Strategies

Mario Marinelli[a], Salvatore Viscio[a], Nadia Giuffrida[a], Mario Binetti[a], Roberta Di Pace[b], Franco Filippi[b], Stefano de Luca[b]

[a]*DICATECh, Polytechnic University of Bari - Italy*
[b]*Department of Civil Engineering, University of Salerno - Italy*

**Extended Abstract.**

The on-ramp area is a high-risk conflict zone where traffic accidents frequently occur. Connected and automated vehicles (CAVs) offer the potential to enhance merging safety through effective cooperative control strategies.
One possible solution for optimising the merging process involves managing the on-ramp merging lane by shielding it from the main traffic flow. To achieve this, in addition to conventional control strategies such as variable speed limits and ramp metering, geometric lane drops (e.g., work-zone or design-based) or virtual lane drops on the main road could be considered. These measures can help reduce conflicts in the on-ramp merging zone when a hard shoulder is unavailable and Hard Shoulder Running is not a viable option. However, implementing such solutions introduces a bottleneck on the main road, necessitating effective merging strategies to enhance safety upstream [1].
The literature proposes various lane merging strategies, including early/late merge techniques, lane-changing advisory systems, and optimization strategies for mixed and fully automated CAV environments [2-5].
Based on this premise, this study explores various network management approaches to identify key factors that influence efficient traffic control in a CAV environment. We conduct a real-scale case simulation using a microscopic traffic model to evaluate the impact of geometric and virtual lane drops upstream of an on-ramp zone. Additionally, both static and dynamic merging systems are analyzed, considering the presence of CAVs.
The case study focuses on the Tangenziale di Napoli (TaNa), a major arterial road in Naples, Italy. The methodology involves implementing a lane drop of the outermost lane on the main road (TaNa) to mitigate conflicts in the merging zone near the entrance ramp (Figure 1).
A What-If analysis is conducted using the PTV Vissim microscopic traffic simulation model, considering the following scenario parameters:

- fixed downstream closure distance ($l_v$);
- variable upstream closure distances ($l_m$);
- variable warning distances ($l_a$);
- variable lane drop durations ($t_c$).

Preliminary results indicate negligible variations in overall network performance, with the most noticeable effects on the ramp. At the same time, a greater impact is observed on safety indicators, particularly in reducing conflicts. Future research will focus on

formulating a multi-objective optimization problem to determine the optimal strategy, validating compliance with regulatory frameworks and real-world lane drop operations, and integrating control strategies such as variable speed limits while expanding the spatial scope of the analysis.


**Figure 1**. A scheme of the lane drop applied to the case study

**Keywords:** Lane Drop, CAV, Traffic Management.

## References

1. Lammers-Staats, E., Pigman, J.G., Howell, B.K., Kirk, A.J.: Applicability of Zipper Merge versus Early Merge in Kentucky Work Zones. Kentucky Transportation Center, Lexington (2018).
2. Federal Highway Administration (FHWA): Guidance for the Use of Dynamic Lane Merging Strategies. Washington, D.C.: U.S. Department of Transportation (2012).
3. Chy, A., Sisiopiku, V.: Quantifying Operational Impacts of Variations in Work Zone Setups, Traffic Demand, and Traffic Composition: A Case Study. Journal of Transportation Technologies, 13, 18-37 (2023).
4. Ramadan, O., Sisiopiku, V.: Bottleneck Merge Control Strategies for Work Zones: Available Options and Current Practices. Open Journal of Civil Engineering, 5, 428-436 (2015).
5. Ramadan, O.E. and Sisiopiku, V.P.: Modeling Highway Performance under Various Short-Term Work Zone Configurations. Journal of Transportation Engineering, Part A: Systems, 144 (2018).

For DTA 2025
Equilibrium and Day-to-Day assignment with Weibit route choice model
G.E. Cantarella, C. Fiori
DICIV, UNISA, IT

Route choice behaviour for each o-d pair $i$ can be described by applying any discrete choice modelling theory so that route choice proportions $\mathbf{p}_i$ depend on route systematic disutilities (costs) $\mathbf{w}$:

$$\mathbf{p}_i = \mathbf{p}_i(-\mathbf{w}_i; \boldsymbol{\theta}_i) \qquad \forall i \qquad (2.12)$$

where $\boldsymbol{\theta}_i$ is the choice function parameter vector, whose meaning depends on the choice model specification. If a utility scale parameter is present, it is considered included in the utility parameter $\rho_i$ (or vice versa).

DEFINITION 1. A choice function is defined *regular*, if:

- it is continuous and monotone increasing with respect to systematic utility,
- it is continuously differentiable with symmetric positive semi-definite (with respect to real vectors) Jacobian, formally $\nabla\mathbf{p}_i(\mathbf{v}_i)$ 0,
- resulting choice proportions depend on differences between systematic utility values only [HYP ®].

Most often Random Utility Theory (RUT) is applied, assuming that (i) each user, travelling between o-d pair $i$, associates to each available route a *perceived utility*, (ii) the perceived utility is modelled by a continuous random variable, with mean given by the *systematic utility*, due to several sources of uncertainty regarding the users or the modeller, and (iii) chooses the maximum perceived utility route; thus the choice probability of an alternative is given by the probability that its perceived utility is equal to maximum among all alternatives; hence the route choice proportions are assumed defined by the route choice probabilities.

When the perceived utility co-variance matrix is non singular, a *probabilistic route choice function* is obtained; it is also called *strictly positive* if each alternative gets a strictly positive probability, whichever are the systematic utility values; examples of strictly positive probabilistic route choice functions are the Logit, Weibit, Probit, Gammit choice functions, usually adopted for route choice modelling. [Strictly positive choice functions may sound somehow unrealistic, as any model they have to be considered suitable mathematical approximations.] If the parameters of the perceived utility pdf do not depend on systematic utility values, the resulting choice function is called *invariant*, if continuous and continuously differentiable, it is regular.

Anyhow equation (2.12) is generally enough to include choice models derived from other discrete choice theories, such as Fuzzy Utility Theory, Bounded Rationality, Prospect Theory, ... (some of them are described in Appendix 2). In any case, a choice function combined with an utility function gives a choice model.

---

An example of choice function (2.12) derived from RUT is the well known Logit choice function, often used as benchmark. For each o-d pair $i$, connected by routes in the route choice set $R_i$, the choice proportion / probability of using route $r$ is given by:

$$p_r = \exp(v_r / \theta_i) / \sum_{k \in R_i} \exp(v_k / \theta_i) \qquad \forall\ r \in R_i$$

where $\theta_i = (6^{0.5}/\pi)\ \sigma_i \cong 0.78\ \sigma_i > 0$ is a dispersion parameter proportional to the standard deviation $\sigma_i$ common to the perceive utilities of all routes connecting o-d pair $i$; the above Logit function is invariant if the route choice set $R_i$ and the dispersion parameter $\theta_i$ do

not depend on systematic utility values. Combing the above choice function with the utility function: $v_r = -\psi_i w_r$ , leads to:

$$p_r = \exp(-w_r / \theta_i) / \sum_{k \in Ri} \exp(-w_k / \theta_i) \quad \forall\ r \in R_i$$

The utility scale parameter $\psi_i$ is included in the dispersion parameter $\theta_i$.

# A dynamic model for route choices updating in urban areas with risk conditions

Francesco Russo, Marialuisa Moschella

DIIES, Università Mediterranea di Reggio Calabria – 89100 Reggio Calabria, Italy

Risk management from disastrous events, both natural and man-made, has become an increasingly central issue in recent years. Disasters such as Hurricane Katrina and the tsunami that hit Fukushima have highlighted the need for evacuation planning, which is closely linked to risk exposure management. While several actions can be taken to reduce vulnerability and the likelihood of damaging events, exposure can be reduced with an evacuation plan.

The increase in global conflicts and the growth in disastrous events caused by climate change over the last fifty years [1] make the need to develop risk reduction strategies even more urgent. For this reason, the Sendai Framework for Disaster Risk Reduction 2015-2030, together with Sustainable Development Goals 1, 11 and 13 of the 2030 Agenda, underlines the importance of developing evacuation plans with common standardised procedures in order to strengthen the resilience of communities around the world [2–5].

An evacuation plan contains well-defined strategies that aim to ensure the safety of people in emergency situations. Within the plan, it is determined how to move the population from a risk area, such as a city or a vulnerable region. To make this happen as quickly and efficiently as possible, transport systems are essential. They are in fact the means by which people can move from one place to another where there is the safety. Transport Risk Analysis (TRA) is an evolution of Quantitative Risk Assessment (QRA), originally applied to the chemical industry. The risk model is based on three main components [6]:

- Occurrence: Probability of an event happening.
- Vulnerability: level of damage that the event may cause.
- Exposure: Number of people or property that could be affected by the disaster.

The exposure component is reduced by the evacuation plan. Transport planning, therefore, becomes a key factor in reducing risk.

In order to reduce the risk of an emergency situation, models must be developed to simulate the evacuation procedures contained in the plan. These models can be developed in the Transport System Models (TSMs) field [7], and include the study of travel demand, which is the choices made by users according to the network and the services available, of transport supply, which analyses the costs associated with using the infrastructure, and of the assignment given by the interaction between supply and demand, with which it is possible to estimate the traffic flows resulting from user choices and network performance. Travel demand is modelled through behavioural or descriptive approaches. Behavioural models explicitly consider users' decisions on the basis of utility criteria, considering variables such as travel time, travel cost and individual preferences, Random Utility Model (RUM), Quantum Utility Model (QUM) and Fuzzy Utility Model (FUM) belong to this category. Descriptive models are based on aggregate relationships between demand and network characteristics, without making explicit the decision-making process of individual users [8–12]. Transport supply is represented through topological models, which describe the infrastructure network through nodes and links, associating cost functions to each arc that depend on traffic flows. The interaction between supply and demand is modelled through the traffic assignment process, which determines the distribution of travel flows on the network based on user choices. Assignment models can be classified into static and dynamic.

The development of TSMs requires a methodological approach that integrates survey tools to analyse user behaviour in an emergency context. This is important because it is not possible to study user behaviour during

a real emergency event, but it is necessary to simulate a hypothetical and at the same time realistic emergency scenario in order to make ex-ante assessments. There are two main types of surveys used for analysing travel decisions:

- Revealed Preference (RP), collects data based on users' actual choices under normal conditions, analysing their daily mobility behaviour;
- Stated Preference (SP), uses hypothetical scenarios allowing the estimation of user behaviour in hypothetical situations that cannot be directly observed.

The integration of RP and SP data permits the development of more accurate predictive models that can simulate the dynamic behaviour of the population during an evacuation.

The information gathered from the survey is the basis on which the TSMs, and in particular the path choice models that allow the risk exposure component to be studied, are built. Traditionally, static models are used, however, these models are forced for representing scenarios in which the risk event and the availability of transport infrastructure change rapidly over time. On the other hand, dynamic models allow the evolution of the system over time and the updating of user choice behaviour to be taken into account [13–16].

Under emergency conditions, the model must consider temporal variations endogenous (traffic) or exogenous (event occurred). Under these conditions, the models that predict the propagation of flows on the network are within-day models, which analyse traffic evolution over a day, and day-to-day models, which consider traffic evolution over several days [17, 18].

The risk condition is different from the usual one. There are multiple risk conditions in which the time between occurrence of the event and effect on users is sufficiently large, and such as to allow evacuation. In this condition the user may have information that allows him to update his choice in successive time periods. It is then possible to divide the day into time periods and then use a dynamic process structure of the type day-to-day to develop a period-to-period analysis.

In the dynamic approach, a key aspect to be considered is the learning process of users who may modify their path choices based on past experiences or real-time information received on the evolution of the event and the network. In this case, the approach used in the literature is sequential analysis that uses users' past decisions to determine future ones.

Referring to an evacuation plan, in order to analyse the behaviour of transport users under risk conditions, can be developed dynamic models that can estimate the updating of choices with a sequential approaches [19].

This work presents an application example of the development of transition matrices that consider the evolution of users' route choices under emergency conditions, relative to a generic transport network.

In order to understand user behaviour, a survey was conducted on a prototype sample collecting information on both Revealed Preferences (RP) and Stated Preferences (SP). The questionnaire was divided into two parts. The first part (RP) collects socio-economic information of the users and analyses their travel habits within the urban area. The second part (SP) introduces the hypothetical but realistic emergency scenario, which was a flood with the risk of two torrents overflowing, and describes the vulnerability conditions of the transport system. Each participant is provided with three key information: the location of the critical event, the position where s/he would be at the time of the emergency and the safe zone to be reached.

This study showed a practical application based on a transport network representative of many urban areas, analysing a hypothetical risk scenario. The results obtained demonstrate the potential of the approach adopted, highlighting how, in the development, a behavioural model can be applied through calibration and validation processes, exploiting the data of stated preferences of users. The proposed approach is particularly interesting for policy makers and experts engaged in risk reduction planning. Finally, the proposed approach can be developed for ordinary conditions and could provide useful guidance for the development of models relating to urban areas with ICT presence.

## References


1. United Nations Office for Disaster Risk, U.N.O. for D.R.: Global Assessment Report on Disaster Risk Reduction 2022: Our World at Risk: Transforming Governance for a Resilient Future. United Nations Publications, Bloomfield (2022).
2. United Nations. UN: The 2030 Agenda for Sustainable Development, https://sdgs.un.org/publications/transforming-our-world-2030-agenda-sustainable-development-17981, (2015).
3. Sendai Framework for Disaster Risk Reduction 2015-2030 | UNDRR, https://www.undrr.org/publication/sendai-framework-disaster-risk-reduction-2015-2030, last accessed 2025/02/06.
4. United Nations. UN: The Sustainable Development Goals Report 2018. (2018).
5. United Nations. UN: United Nation. UN: Sustainable Development Goals (SDG), https://unstats.un.org/sdgs/indicators/Global-Indicator-Framework-after-2024-refinement-English.pdf, (2024).
6. Russo, F., Vitetta, A.: Safety of users in road evacuation: general methodology and main results. In: Urban Transport XIII: Urban Transport and the Environment in the 21st Century. pp. 763–772. WIT Press, Coimbra, Portugal (2007). https://doi.org/10.2495/UT070721.
7. Cascetta, E.: Transportation systems engineering: theory and methods. Springer Science & Business Media (2013).
8. McFadden, D.: The measurement of urban travel demand. Journal of Public Economics. 3, 303–328 (1974). https://doi.org/10.1016/0047-2727(74)90003-6.
9. Ben-Akiva, M.E., Lerman, S.R.: Discrete choice analysis: theory and application to travel demand. MIT Press, Cambridge, Mass (1985).
10. Tversky, A., Shafir, E.: The Disjunction Effect in Choice under Uncertainty. Psychological Science. 3, 305–309 (1992).
11. Train, K.E.: Discrete choice methods with simulation. Cambridge university press, Cambridge (2009).
12. Vitetta, A.: A quantum utility model for route choice in transport systems. Travel Behaviour and Society. 3, 29–37 (2016). https://doi.org/10.1016/j.tbs.2015.07.003.
13. Newell, G.F.: Traffic flow on transportation networks. MIT Pr, Cambridge, Mass (1980).
14. Russo, F., Chilà, G.: Safety of users in road evacuation: Modelling and DSS for demand. Presented at the 4th International Conference on Sustainable Development and Planning (2009). https://doi.org/10.2495/SDP090431.
15. Marcianò, F.A., Musolino, G., Vitetta, A.: Within-day traffic assignment and signal setting in road evacuation: a procedure with explicit path enumeration. Presented at the SAFE 2011 , Antwerp, Belgium July 4 (2011). https://doi.org/10.2495/SAFE110351.
16. Vitetta, A.: Network Design Problem for Risk Reduction in Transport System: A Models Specification. Int. J. TDI. 6, 283–297 (2022). https://doi.org/10.2495/TDI-V6-N3-283-297.
17. Cantarella, G.E., Watling, D.P., De Luca, S., Di Pace, R.: Dynamics and stochasticity in transportation systems: tools for transportation network modeling. Elsevier, Amsterdam Oxford Cambridge, MA (2020).
18. Cantarella, G.E., Fiori, C., Mussone, L.: Dynamics and stochasticity in transportation systems: equations and examples (Part II). ELSEVIER, S.l. (2024).
19. Russo, F., Comi, A., Chilà, G.: Dynamic Approach to Update Utility and Choice by Emerging Technologies to Reduce Risk in Urban Road Transportation Systems. Future Transportation. 4, 1078–1099 (2024). https://doi.org/10.3390/futuretransp4030052.

**Title**: "*A performance-impact based strategy for multi-level speed harmonization in mixed traffic conditions*"

**Authors**: Maria Nadia Postorino[1], Luca Mantecchini[1], Caterina Malandri[1], Roberta Di Pace[2], Stefano De Luca[2]

[1]Department of Civil, Chemical, Environmental and Materials Engineering (DICAM), University of Bologna, Viale del Risorgimento 2, Bologna, Italy

[2]Laboratory of Transportation Systems Engineering and Sustainable Mobility (ISTMoS), Department of Civil Engineering, University of Salerno, via Giovanni Paolo II, Fisciano, SA, Italy

# Impact of Electric Vehicle Subsidies on a Mixed Urban Delivery Fleet

Yongling Gao[a], Yuan Qiao[a,b], Meng Xu[c]*

[a] Business School, Central University of Finance and Economics, Beijing, China

[b] HSBC Business School, Peking University, Shenzhen, China

[c] School of Systems Science, Beijing Jiaotong University, Beijing, China

*Extended abstract submitted for presentation at the 10th International Symposium on Dynamic Traffic Assignment (DTA2025)*
*September 16-18, 2025, Salerno, Italy*

Keywords: City logistics; Mixed fleet; Electric vehicles; Vehicle routing; Mileage subsidy

# 1 INTRODUCTION

Urban freight traffic makes up around 10-15% of the total distance traveled and is responsible for about 6% of greenhouse gas emissions related to transportation (Civitas, 2020). Currently, using internal combustion engine vehicles (ICEVs) for distribution contradicts the carbon emission reduction goal. In comparison, electric vehicles (EVs) can lower motor noise and emissions, and making them fundamental to the green development of urban logistics (Pelletier et al., 2016).

There are different types of financial support for promoting the adoption and operations of EVs, such as EV purchase subsidies, fleet electrification subsidy (FS), mileage subsidy (MS), tax rebates, charging facility subsidies, and charging discounts, as well as non-financial incentives, including traffic incentives and convenient parking measures (Morganti and Browne, 2018). **The EV purchase subsidy (*PS*)** is a widely utilized policy tool to make EVs more affordable compared to ICEVs and promote EV adoption. **The fleet electrification subsidy (*FS*)** aligns subsidy eligibility with the fleet electrification ratio and determines the subsidy amount based on the number of EVs to support LSPs that are making significant strides towards sustainable transportation practices. To incentivize the actual use of EVs, **the mileage subsidy (*MS*)** policy has been applied to each EV that meets a threshold for single-vehicle mileage, ensuring that the financial support goes toward vehicles that actively reduce emissions.

While the *PS* policy is widely studied for its role in adoption, the impact of the *FS* and *MS* policies on the EV utilization, total cost of the fleet, carbon emissions, and social welfare are less understood. Unlike the *PS* policy, the *FS* and *MS* policies impose additional requirements for fleet electrification ratios and EV operating mileage, which could significantly influence fleet routing decisions. As suggested by Yang and Hyland (2024), potential regulations concerning the proportion of EVs in delivery fleets and the miles driven by EVs are necessarily explored. To address this gap, it is expected to develop a fleet composition and vehicle routing model to investigate the impact of the *PS*, *FS*, and *MS* policies.

From the government's perspective, EV subsidies should help LSPs reduce costs and carbon emissions. From a business perspective, reducing costs and carbon emissions is attractive to the LSPs' customers and investors. van Gogh et al. (2021) find that a rapidly increasing number of companies are taking action to reduce supply chain emissions, prioritizing low-carbon supply chains, and are willing to pay a 5-10% premium for sustainable logistics services. Considering that both governments and firms aim to reduce costs and carbon emissions, this paper presents a bi-objective optimization framework to minimize costs and emissions in a mixed fleet of EVs and ICEVs, focusing on their composition and routing decisions. It also explores the impact of *PS*, *FS*, and *MS* policies on the fleet, aiming to answer the following three research questions:

1. How should an LSP reduce the total cost and carbon emissions of the fleet by optimizing EV and ICEV deployment and route selection under the *FS* and *MS*?

2. How to select the *PS*, *FS*, and *MS* policies to maximize social welfare?

3. What are the different impacts of the *FS* and *MS* policies, compared to the *PS* policy, on EV usage, total cost, and carbon emissions for an LSP?

* Corresponding author.
*E-mail addresses:* mengxu@bjtu.edu.cn

By studying these issues, this paper aims to understand how government agencies rank EV subsidy measures in terms of fleet electrification, EV mileage, carbon emissions, and social welfare. It also offers LSPs references for optimizing fleet composition and vehicle routing, and provides guidance on how urban delivery fleets can develop greenly in the context of carbon peak and carbon neutrality.

# 2 METHOLOGY

## 2.1 Description

The government aims to evaluate the impact of EV subsidies (PS, FS, and MS) on a mixed urban fleet of the LSP. Under each subsidy policy, the LSP makes fleet composition and route decisions with two objectives: reducing costs and lowering carbon emissions. Carbon emission reduction is regarded as a primary optimization goal, as many leading logistics companies focus on environmental, social, and governance initiatives, incorporating carbon reduction into their operations to attract customers and investors. Additionally, it is imperative for the government to evaluate the outcomes of subsidizing EVs while mandating LSPs to prioritize carbon emission reduction in their operations.

The LSP dispatches ICEVs and EVs from the distribution center to deliver to multiple superstores and return every day. Each customer's time window is known. The use of ICEVs is constrained by vehicle access restrictions, and the use of EVs requires consideration of the impact of their driving range and charging strategy. The EVs depart from the distribution center on a full charge, with the option to partially charge them en route. These assumptions are based on the case of Sinotrans, a leading LSP in China, responsible for daily deliveries to a number of supermarkets in Beijing. It is assumed that ICEVs and EVs can serve multiple customers, with each customer point being served only once. Furthermore, the maximum carrying capacity of vehicles with the same type of energy is assumed to be the same. We use the fleet composition and vehicle routing model results to compare the impacts of PS, FS, and MS, respectively. Figure 1 presents an overview of the optimization modeling.

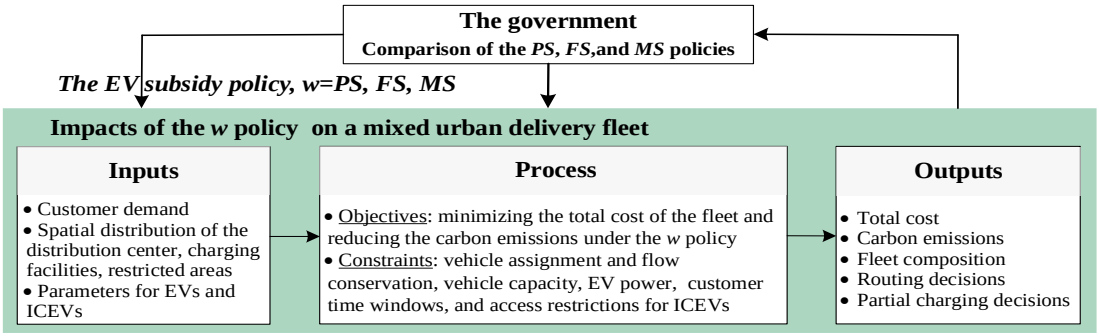



**Fig. 1** Overview of optimization modeling

## 2.2 A bi-objective MFVRP model under the *PS* policy

We now define notations for the bi-objective MFVRP. This problem is defined on a complete directed graph $G=(\Lambda, A)$, where $\Lambda = I \cup R \cup \{0\}$ is the set of nodes, $I$ is the set of customers, $R$ is the set of charging facilities, and $\{0\}$ denotes the distribution center. The set of arcs is characterized by $A=\{(i,\ j), i,\ j \in N \text{ and } i \neq j\}$. Let $K$ denote the set of vehicles and the subscript $k$ denote the vehicle number. The symbol $I^*$ represents the set of nodes with access restrictions for ICEVs. The fixed cost and variable cost of EV (ICEV) are denoted by $f^{EV}$ ($f^{ICEV}$) and $c^{EV}$ ($c^{ICEV}$), respectively. The EV's and ICEV's maximum carrying capacities are $Q^{EV}$ and $Q^{ICEV}$, respectively. Given the widespread adoption of the *PS* policy in logistics operations. Note that $y_k^{PS}=0$ ($y_k^{PS}=1$) represents the vehicle $k$ is an EV (ICEV). The total number of EVs and ICEVs under the *PS* can be determined by $N_{EV}^{PS}=\sum_{k\in K}(1-y_k^{PS})$ and $N_{ICEV}^{PS}=\sum_{k\in K}y_k^{PS}$, respectively. The total fixed cost of vehicles under the *PS* policy $C_F(y_k^{PS})$ is

$$C_F(y_k^{PS}) = \sum_{k\in K}[\, f_{EV}(1-y_k^{PS}) + f_{ICEV} y_k^{PS}\,] \quad (1)$$

The variable cost of vehicles with the *PS* $C_V(x_{ijk}^{PS}, y_k^{PS})$ is a nonlinear function, depending on the vehicle type decision $y_k^{PS}$ and the routing decision $x_{ijk}^{PS}$.

$$C_V(x_{ijk}^{PS}, y_k^{PS}) = \sum_{k\in K}\sum_{i\in\Lambda}\sum_{j\in\Lambda}\Big[\, c_{EV}(1-y_k^{PS}) + c_{ICEV} y_k^{PS}\,\Big] d_{ij} x_{ijk}^{PS} \quad (2)$$

Note $[e_i, l_i]$ denote the time window for customer $i$ and $t_{ik}^{PS}$ represent the vehicle $k$ 's arriving time at node $i$ under the *PS* policy. We assume that the LSP is required to pay compensation $p_E(e_i - t_{ik}^{PS})$ if vehicle $k$ arrives early and compensation $p_L(t_{ik}^{PS} - l_i)$ if vehicle $k$ arrives late, which represents a soft time window scenario. The penalty cost for violating the customer's time window under the *PS* policy is

$$\begin{aligned} C_P(z_{rk}^{PS}) &= \sum_{k\in K}\sum_{i\in I}\Big[\max\left\{p_E(e_i - t_{ik}^{PS}), 0\right\} + \max\left\{p_L(t_{ik}^{PS} - l_i), 0\right\}\Big] \\ &= \sum_{k\in K}\sum_{i\in I}\max\left\{p_E(e_i - t_{ik}^{PS}), 0, p_L(t_{ik}^{PS} - l_i)\right\} \end{aligned} \quad (3)$$

As a per-vehicle subsidy $s^{PS}$ is granted to the LSP for EV purchases, the LSP's total cost of the mixed fleet under the *PS* policy becomes $C(x_{ijk}^{PS}, y_k^{PS}, z_{rk}^{PS})$:

$$C(x_{ijk}^{PS}, y_k^{PS}, z_{rk}^{PS}) = C_F(y_k^{PS}) + C_V(x_{ijk}^{PS}, y_k^{PS}) + C_P(z_{rk}^{PS}) - s^{PS} N_{EV}^{PS} \quad (4)$$

This paper considers two objectives for the LSP: minimizing the total cost of the fleet and reducing the carbon emissions produced by ICEVs. Since EVs have zero tailpipe emissions, this paper mainly focuses on the ICEVs' carbon emissions during their operations. According to road transportation practices and existing studies such as Toro et al. (2017), carbon emissions of an ICEV can be estimated based on fuel consumption and carbon emissions of the fuel used. Then the carbon emissions from ICEV $k$ on arc $(i, j)$ (i.e., $E(x_{ijk}^{PS}, y_k^{PS})$) can be estimated using its fuel consumption $FC(x_{ijk}^{PS}, y_k^{PS})$ and the fuel emission factor $FE$ as follows:

$$E(x_{ijk}^{PS}, y_k^{PS}) = FC(x_{ijk}^{PS}, y_k^{PS})\, FE \quad (5)$$

where $FE$ represents the efficiency of converting fuel into carbon emissions and is defined as the amount of $CO_2$ emitted per liter of fuel. This efficiency is influenced by the vehicle type, the fuel type, and other related variables.

Further, we will present the bi-objective MFVRP model under the *FS* and *MS* policies.

## 2.3 Social welfare

A key issue for the government is understanding how subsidy policies affect social welfare. The social welfare under the $w$ policy ($w = PS, FS, MS$), denoted by $SW^w$, includes the following terms: (1) the LSP's profit $\pi_1(x_{ijk}^w, y_k^w, z_{rk}^w)$; (2) the customer's profit $\pi_2^w = (V_c - V_{lsp})\sum_{j\in I}\hat{q}_j$; (3) the profits of the EV and ICEV suppliers; (4) the profits of the electricity and fuel suppliers; (6) the total subsidies $S^w$. The total subsidies in social welfare can be eliminated as they are transferred from the government to the LSP. Then the social welfare under the $w$ policy can be defined by

$$SW(x_{ijk}^w, y_k^w, z_{rk}^w) = \pi_1(x_{ijk}^w, y_k^w, z_{rk}^w) + \pi_2^w + \sum_{g=3}^{4} \pi_g(y_k^w) + \sum_{g=5}^{6} \pi_g(x_{ijk}^w, y_k^w) - \Upsilon(x_{ijk}^w, y_k^w) - S^w$$

$$= V_{lsp} \sum_{j \in I} \hat{q}_j - C_F(y_k^w) - \hat{C}_V(x_{ijk}^w, y_k^w) - \hat{C}_P(z_{rk}^w) + \pi_2^w + \sum_{g=3}^{4} \pi_g(y_k^w) + \sum_{g=5}^{6} \pi_g(x_{ijk}^w, y_k^w) - \Upsilon(x_{ijk}^w, y_k^w)$$

# 3 NUMERICAL RESULTS

This paper considers that an LSP employs EVs and ICEVs to deliver goods from a distribution center to 30 customers daily, with a maximum demand of no more than 0.5 t per customer. Utilizing the improved NSGA-II, we observed an average time of 71 seconds for each run. After executing the model at least five times, we obtain the following average results: 72 Pareto frontier solutions, an average of 3.91 EVs and 2.26 ICEVs, an average total cost of 3612.33 yuan per day, an average carbon emission of 9631.41 g per day, and an average EV charging power of 204.96 kWh per day. The vehicle routing optimization results are randomly selected in the Pareto frontier solution as shown in Fig. 2.

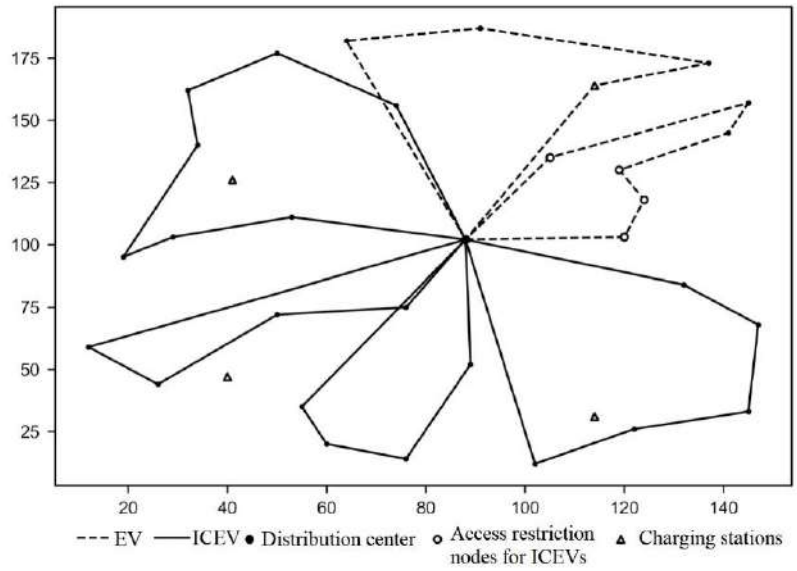



**Fig. 2.** Simulation results.

Further investigations on the sensitivity analysis of EV subsidies, and the comparison of EV subsidy policies will be implemented.

# 4 DISCUSSION

To promote EV adoption in city logistics, the *FS* and *MS* policies have implemented by several Chinese cities, providing eligible EVs with financial support. This study is expected to propose a bi-objective optimization model for MFVRP under three subsidy policies: *PS*, *FS*, and *MS*. The objectives are to minimize the total cost and carbon emissions of the LSP's fleet, while considering the impacts of the EV on-the-go charging decision and the ICEV's carbon emissions. As an early attempt, this paper introduces the *FS* and *MS* policies into the bi-objective MFVRP. Effects of *PS*, *FS*, and *MS* will be compared regarding the total cost of the fleet, the proportion of EVs, fleet carbon emissions, and social welfare.

The following results in this study are expected:

- The impact of the three policies, i.e., *PS*, *FS*, or *MS*, will be investigated upon the potential to maximize social welfare, depending on the number of EVs in the fleet and the environmental impact of EV production. When the environmental impact of producing an EV is sufficiently small (sufficiently large), the policy (*PS*, *FS*, or *MS*) that results in the maximum (minimum) number of EVs in the fleet is expected to achieve the greatest social welfare.
- The increasing of the *PS*, *FS*, and *MS* can reduce the total cost of the fleet but has varying effects on fleet carbon emissions, the proportion of EVs, and the average mileage of EVs. Specifically,

increasing both the *PS* and *FS* can reduce fleet carbon emissions and increase the proportion of EVs, but it does not necessarily lead to an increase in the average EV mileage.

- When the total fleet subsidy changes, the total cost of the fleet may be lowest under the *PS*, *FS*, or *MS* policy. Both the *PS* and *FS* policy can result in a higher proportion of EVs compared to the *MS*, leading to potentially lower fleet carbon emissions.

# References

CIVITAS, 2020. Smart choices for cities: making urban freight logistics more sustainable. URL: https://civitas.eu/sites/default/files/civ_pol-an5_urban_web.pdf.

Morganti, E., Browne, M., 2018. Technical and operational obstacles to the adoption of electric vans in France and the UK: An operator perspective, Transport Policy 63, 90-97.

Pelletier, S., Jabali, O., Laporte, G., 2016. 50th anniversary invited article—goods distribution with electric vehicles: review and research perspectives. Transportation Science 50(1), 3-22.

Toro, E. M., Franco, J. F., Echeverri, M. G., Guimarães, F. G., 2017. A multi-objective model for the green capacitated location-routing problem considering environmental impact. Computers & Industrial Engineering 110, 114-125.

van Gogh M., Farrag-Thibault A., Hausmann L., 2022. Driving decarbonization: Accelerating zero-emission freight transport. URL: https://www.mckinsey.com/industries/travel-logistics-and-infrastructure/our-insights/driving-decarbonization-accelerating-zero-emission-freight-transport.

Yang, D., Hyland, M. F., 2024. Electric vehicles in urban delivery fleets: How far can they go?. Transportation Research Part D: Transport and Environment, 129, 104127.

# Assessing the resilience of a road transport network: a dynamic approach

Massimo Di Gangi, Orlando Marco Belcore, Antonio Polimeni
Università degli Studi di Messina, Dipartimento di Ingegneria, Contr. di Dio – Villaggio S. Agata 98166 Messina, Italy

Evaluation of the resilience of a transport network is the basis for understanding the functionality of the network in the presence of disturbance. The objective of this work is to provide a measure of the resilience of a road network considering the case where, due to a disturbance situation, the network capabilities are reduced. The evaluation of the resilience of a transport network has been covered by many authors from different points of view, ranging from methods to quantify the resilience ([1], [2]) to resilience assessment ([3], [4]) to the network optimization ([5], [6]). The contribution of this work to the literature is in the dynamic approach developed to measure the resilience of a road network considering a definition of resilience based on mechanics (most of the works in the literature are based on a concept of resilience derived from ecological systems).

The concept of resilience is defined variously in the literature, in this work the definition is based on the concept of mechanical resilience, understood as the ability of a material to absorb elastic energy when subjected to stress and to release it when such stresses cease. The concept of resilience is measured both in terms of immediate resistance capacity (robustness) and in the speed of recovery (recovery time), just as in resilient materials it is evaluated as in mechanics, a resilient transport network is able to absorb a shock (disturbing event) and quickly restore its operation, minimizing efficiency losses. From the previous premises, in this work, the resilience of a transport network is defined as the capacity of the network to absorb disturbances (adverse events such as accidents, failures, extreme weather events, cyberattacks) while maintaining or quickly restoring its level of service. It is measured by indicators that evaluate the efficiency of the system before, during and after the disturbance event. To better clarify this definition, Table 1 reports a parallel between mechanical resilience and the resilience of a transport network.

*Table 1 – Comparison between the concept of resilience from mechanic systems to road transport network*

| Mechanics | Transport Networks | Meaning |
|---|---|---|
| Energy absorbed | Loss of network capacity | Impact of the disruptive event (e.g. reduction in traffic flow, congestion) |
| Applied stress | Critical event on the network | Accident, roadworks, landslides, floods, snowfalls, cyber-attacks on traffic management systems |
| Elastic capacity of the material | Traffic adaptation capacity | Ability of the network to redistribute vehicles on alternative routes (system elasticity) |
| Energy released | Traffic recovery | The speed with which traffic returns to pre-disturbance levels |
| Resilience test (Charpy/Izod) | Measurement of performance indicators | Analysis of the network's ability to absorb shocks and recover |

A resilient road network is able to handle disruptions, reducing their impact on traffic, and quickly restoring normal operation. Just as a resilient material absorbs energy without breaking, a resilient road network absorbs critical events, minimizing congestion and recovery times.
The indicators used to measure resilience in the transport network are as follows:

- Recovery time ($T_r$): it is the time required for traffic parameters (average speed, density, flow) to return to normal values after a disturbance;
- Capacity reduction index (Cr): represents the reduction (in percentage) of network capacity due to disruption;
- Robustness (Rb): measures the network's ability to maintain connections even under stress;
- Vulnerability index ($V$): it indicates how many interruptions at critical points can affect the entire network;
- Road network flexibility ($F$): it is the ability to distribute traffic over alternative routes without causing system collapse.

In the context of a road network, residual resistance represents the capacity of the network to continue to function, albeit with reduced performance, after a disturbance (accidents, road closures, natural events). Starting from the previous indicators, some resistance indicators can be derived:

- Residual transport capacity ($C_{res}$): it represents how much traffic capacity remains available after a disruption;
- Residual efficiency ($E_{res}$): is the percentage of network performance compared to the normal condition;
- Residual connectivity index ($I_{con}$): it indicates how many alternative routes remain available after a disruption.

Figure 1 reports schematically the adopted procedure. Starting from *demand* and *supply*, the system is simulated by means of a dynamic traffic assignment (DTA), and the previous *resilience indicators* are calculated for each time interval. The *test* is performed on the network load conditions: if the network is empty, the procedure ends, otherwise it goes to the next time interval. More in detail, the DTA model considered in this work is mesoscopic ([7], [8]), the users (vehicles) are assembled in packets moving on the road network, which means that the demand is discretized for each origin-destination pair. A graph $G(N,A)$ represents the transport network, $N$ is the set of nodes, and $A$ is the set of arcs. Each arc $a$ belonging to $A$ is composed of two parts: a running segment and a queuing segment, the length of such segments depending on the flow conditions (these values are evaluated at the beginning of each time interval). Furthermore, each arc is associated with a width ($W_a$) and a capacity ($C_a$). A packet $p\{u, od, \tau\}$ is defined as an element that contains vehicles of the same class $u$, moving on the same $od$ pair and starting from the origin at the same instant $\tau$. The movement of a packet is simulated considering a directed acyclic graph $\Gamma_{od,\tau}(M, E)$ ($M \subset N$, $E \subset A$), his arcs belonging to the feasible routes from $o$ to $d$ at instant $\tau$.

An application will be provided to test the proposed procedure. The test area chosen is the Sicily Island (South Italy). The area covers an area of about 25000 square kilometers and has about 5 million inhabitants. To make the test, an emergency scenario is generated by assuming a disruptive event that affects the road network with related reduction in performance.

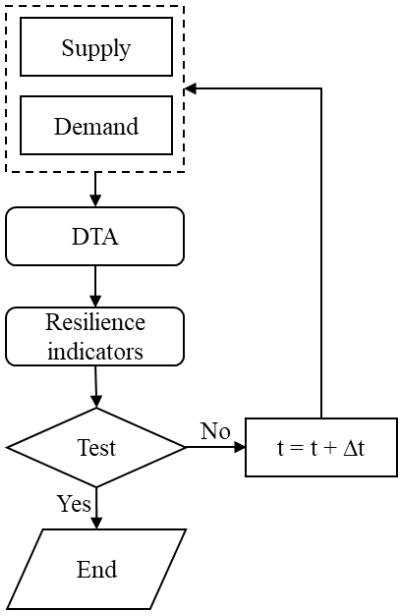



*Figure 1 – The proposed procedure*

## References

[1] P. Gauthier, A. Furno, and N.-E. El Faouzi, "Road Network Resilience: How to Identify Critical Links Subject to Day-to-Day Disruptions," *Transportation Research Record*, vol. 2672, no. 1, pp. 54–65, Dec. 2018, doi: 10.1177/0361198118792115.

[2] N. Y. Aydin, H. S. Duzgun, H. R. Heinimann, F. Wenzel, and K. R. Gnyawali, "Framework for improving the resilience and recovery of transportation networks under geohazard risks," *International Journal of Disaster Risk Reduction*, vol. 31, pp. 832–843, Oct. 2018, doi: 10.1016/j.ijdrr.2018.07.022.

[3] H. Zhang, G. Song, F. He, Z. Tan, and J. Huang, "Dynamic Evaluation Method for Urban Road Network Resilience Based on Congestion Propagation and Dissipation," *Transportation Research Record*, p. 03611981231183714, Jul. 2023, doi: 10.1177/03611981231183714.

[4] M. Ilbeigi, "Statistical process control for analyzing resilience of transportation networks," *International Journal of Disaster Risk Reduction*, vol. 33, pp. 155–161, Feb. 2019, doi: 10.1016/j.ijdrr.2018.10.002.

[5] A. Kaviani, R. G. Thompson, and A. Rajabifard, "Improving regional road network resilience by optimised traffic guidance," *Transportmetrica A: Transport Science*, vol. 13, no. 9, pp. 794–828, Oct. 2017, doi: 10.1080/23249935.2017.1335807.

[6] R. Faturechi and E. Miller-Hooks, "Travel time resilience of roadway networks under disaster," *Transportation Research Part B: Methodological*, vol. 70, pp. 47–64, Dec. 2014, doi: 10.1016/j.trb.2014.08.007.

[7] M. Di Gangi, G. E. Cantarella, R. Di Pace, and S. Memoli, "Network traffic control based on a mesoscopic dynamic flow model," *Transportation Research Part C: Emerging Technologies*, vol. 66, pp. 3–26, May 2016, doi: 10.1016/j.trc.2015.10.002.

[8] M. Di Gangi, D. Watling, and R. D. Salvo, "Modeling Evacuation Risk Using a Stochastic Process Formulation of Mesoscopic Dynamic Network Loading," *IEEE Trans. Intell. Transport. Syst.*, vol. 23, no. 4, pp. 3613–3625, Apr. 2022, doi: 10.1109/TITS.2020.3038478.

# Tradable Credit Scheme for Hybrid Traffic of Electric and Gasoline Vehicles under Elastic OD Demand

Guangmin Wang[a], Shiyi Liu[a], Meng Xu[b]

a) School of Economics and Management, China University of Geosciences, Wuhan, 430000, PR China

b) School of Systems Science, Beijing Jiaotong University, Beijing, 100044, PR China

**Extended Abstract:**

**Background:**

With the rapid development of electric vehicles (EVs), gasoline vehicles (GVs) and EVs will exist together for a long time. The unique charging behavior characteristics of EVs, the widespread range anxiety effect among drivers, and the significant differences in energy consumption between the EVs and GVs have a profound impact on the traffic network. Moreover, traffic congestion has always been a core research problem in transportation field. The widespread adoption of EVs further complicates the issue, requiring innovative strategies to ensure the efficient and sustainable transportation system. Compared with existing traffic management measures, the tradable credit scheme (TCS) has significant advantages in terms of fairness and efficiency and has the potential to solve the congestion problem in hybrid traffic of EVs and GVs. TCS internalizes the negative externalities that travelers impose on the transportation system, thereby guiding travelers to proactively adjust their travel behavior and effectively alleviate traffic congestion. Considering that the origin-destination (OD) demand between OD pairs in real traffic scenarios is not constant but fluctuates with the operation status of the traffic network, showing dynamic elastic characteristics, this paper considers the elastic OD demand situation to make the research more in line with the complex dynamic characteristics of actual traffic. In the actual operation of the traffic system, there is an aggregation phenomenon in the spatial distribution of EV charging demands, which easily leads to nonlinear congestion effects at charging stations. Therefore, this paper incorporates this complex factor into the research scope. The M/G/k queuing theory is used to model the queuing time at charging stations as a function of charging flow and demand when considering the limited capacity of charging stations and the different charging demands of EVs. The feedback effects on the dynamic changes in OD demand and the overall operational efficiency of the traffic network are analyzed and quantified when network congestion and charging station congestion are superimposed.

**Model and algorithm:**

This paper studies TCS for hybrid traffic of EVs and GVs under elastic OD demand, and constructs a mathematical programming with equilibrium constraints (MPEC). In the proposed model, the traffic management authority determines the amount of credit on congested links to maximize social welfare; the variational inequality is used to describe route and charging choices of network travelers to minimize their generalized travel costs under the given tradable credit scheme. The generalized travel costs include travel time, energy expenses (electricity for EVs, gasoline for GVs), credit costs, and charging time and queuing delays at charging stations. The model is as follows:

$$\max_{\boldsymbol{\kappa},\mathbf{f},\mathbf{d}} TSB(\kappa, f, d) = TUB - TC$$

$$s.t. \quad \underline{\kappa}_a^m \le \kappa_a^m \le \overline{\kappa}_a^m, \forall a \in A_m, m \in M$$

$$\sum_{m\in M}\sum_{w\in W}\sum_{r\in R_w^m} c_{r,w}^{m\,*}\left(f_{r,w}^m - f_{r,w}^{m\,*}\right) + \left(\sum_{m\in M}\sum_{a\in A_m}\kappa_a^m v_a^{m*} - K\right)(p_c - p_c^*) - \sum_{m\in M}\sum_{w\in W} D_w^{m-1}(d_w^{m*})(d_w^m - d_w^{m*}) \ge 0,$$

$$\forall (\mathrm{v}, \mathrm{d}) \in \widetilde{\Omega}(K, \kappa), \forall p_c \ge 0$$

where, $TC = (TCE + TCG)$

$TCE = \sum_{w\in W}\sum_{r\in R_w^e}\left[f_{r,w}^e * \left(\sum_{a\in A}(\beta_e t_a(v_a) + \varpi_e p_e l_a)\delta_{a,r,w} + \beta_e E_{i,r,w}\xi_i + \sum_{i\in\widehat{N}}\beta_e t_i \delta_{i,r,w}\right)\right]$ is the sum of the total cost of all EVs, $TCG = \sum_{w\in W}\sum_{r\in R_w^g}\left[f_{r,w}^g * \sum_{a\in A}\left(\beta_g t_a(v_a) + \varpi_g p_g l_a\right)\delta_{a,r,w}\right]$ is the sum of the total cost of all GVs, $TUB = \sum_{m\in M}\sum_{w\in W}\frac{1}{\beta_m}\int_0^{d_w^m} D_w^{m-1}(\omega)d\omega$ is the benefit of all travelers for travel, $\widetilde{\Omega}(K,\kappa) = \left\{(\mathrm{v},\mathrm{d}) \middle| \sum_{m\in M}\sum_{a\in A_m}\kappa_a^m v_a^m \le K, (\mathrm{v},\mathrm{d}) \in \Omega\right\}$ is the feasible tradable credit scheme, $t_i(x_i, E_i|\xi_i, k_i) = \begin{cases} \frac{\xi_i x_i \cdot diag(E_i E_i^T)}{2x_i \cdot E_i} \cdot \sum_{n=0}^{6}\alpha_{n,k_i}\left(\frac{\xi_i x_i \cdot E_i}{k_i}\right)^n, x_i \cdot E_i > 0 \\ 0, x_i \cdot E_i = 0 \end{cases}, \forall i \in \widehat{N}$ is the queue time of EVs at charging station $i \in \widehat{N}$

The genetic algorithm is designed to solve the optimization problem, and the gradient projection method and the improved continuous average method are used to solve the hybrid equilibrium problem. Particularly, considering the behavior characteristics of EV drivers who only charge the minimum amount of electricity to reach their destinations as soon as possible, a shortest path generation model with a range anxiety threshold is designed to simultaneously optimize route selection and charging plans, and it is embedded in the iterative process of the continuous average method.

**Results and contributions:**

This paper conducts numerical experiments on the Nguyen-Dupuis network. The results under user equilibrium (UE) and TCS are listed in Table 1:

**Table 1: The results under user equilibrium and TCS**

| Links | Results under UE | | | | Results under TCS | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $v_a^e$ | $v_a^g$ | $v_a$ | $v_a/g_a$ | $\kappa_a^e$ | $\kappa_a^g$ | $v_a^e$ | $v_a^g$ | $v_a$ | $v_a/g_a$ |
| 1-5 | 64.92 | 46.70 | 111.62 | 1.24 | 8 | 16 | 55.31 | 35.18 | 90.49 | 1.01 |
| 1-12 | 25.05 | 34.87 | 59.92 | 0.86 | - | - | 16.28 | 28.32 | 44.60 | 0.64 |
| 4-5 | 34.50 | 11.63 | 46.13 | 0.66 | - | - | 33.31 | 0.05 | 33.36 | 0.48 |
| 4-9 | 31.90 | 52.03 | 83.93 | 0.93 | - | - | 29.10 | 57.43 | 86.53 | 0.96 |
| 5-6 | 99.42 | 11.94 | 111.36 | 1.39 | 0 | 55 | 88.62 | 0.01 | 88.63 | 1.11 |
| 5-9 | 0.00 | 46.39 | 46.39 | 0.77 | - | - | 0.00 | 35.22 | 35.22 | 0.59 |
| 6-7 | 124.40 | 11.94 | 136.34 | 1.51 | 19 | 68 | 104.90 | 0.01 | 104.91 | 1.17 |
| 6-10 | 0.10 | 0.08 | 0.18 | 0.00 | - | - | 0.03 | 0.02 | 0.05 | 0.00 |
| 7-8 | 52.58 | 11.75 | 64.33 | 0.86 | - | - | 55.18 | 0.01 | 55.19 | 0.74 |
| 7-11 | 71.79 | 0.19 | 71.98 | 1.03 | 11 | 74 | 49.69 | 0.00 | 49.69 | 0.71 |
| 8-2 | 52.58 | 46.53 | 99.11 | 1.42 | 17 | 34 | 55.18 | 28.31 | 83.49 | 1.19 |
| 9-10 | 31.90 | 22.77 | 54.67 | 0.78 | - | - | 29.10 | 33.31 | 62.41 | 0.89 |
| 9-13 | 0.00 | 75.65 | 75.65 | 1.26 | 60 | 31 | 0.00 | 59.34 | 59.34 | 0.99 |
| 10-11 | 32.01 | 22.85 | 54.86 | 0.78 | - | - | 29.13 | 33.33 | 62.46 | 0.89 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 11-2 | 24.76 | 22.85 | 47.61 | 0.68 | - | - | 18.99 | 33.32 | 52.31 | 0.75 |
| 11-3 | 79.03 | 0.19 | 79.22 | 1.13 | 38 | 35 | 59.84 | 0.02 | 59.86 | 0.86 |
| 12-6 | 25.05 | 0.08 | 25.13 | 0.50 | - | - | 16.28 | 0.02 | 16.30 | 0.33 |
| 12-8 | 0.00 | 34.78 | 34.78 | 0.50 | - | - | 0.00 | 28.30 | 28.30 | 0.40 |
| 13-3 | 0.00 | 75.65 | 75.65 | 1.08 | 56 | 0 | 0.00 | 59.34 | 59.34 | 0.85 |
| $TSB$ | 43042.31 | | | | 49048.11 | | | | | |
| $TUB$ | 122176.13 | | | | 109185.19 | | | | | |
| $TC$ | 79133.82 | | | | 60137.07 | | | | | |
| $p_c$ | - | | | | 1.33 | | | | | |

To illustrate the sensitivity analysis, we obtain the results under different electricity and gasoline prices by increasing and decreasing 10% in Table 2:

**Table 2: The results under different electricity and gasoline prices**

| | Change in electricity prices | | | | | | |
|---|---|---|---|---|---|---|---|
| | -30% | -20% | -10% | 0 | +10% | +20% | +30% |
| $TSB$ | 50100.35 | 49724.46 | 49348.37 | 49048.11 | 48605.27 | 48187.28 | 47873.78 |
| $TUB$ | 109391.62 | 109962.25 | 109268.19 | 109185.19 | 108829.69 | 108089.25 | 107533.81 |
| $TC$ | 59291.27 | 60237.79 | 59919.82 | 60137.07 | 60224.43 | 59901.98 | 59660.03 |
| $p_c$ | 1.436 | 1.373 | 1.376 | 1.330 | 1.325 | 1.327 | 1.326 |
| | Change in gasoline prices | | | | | | |
| | -30% | -20% | -10% | 0 | +10% | +20% | +30% |
| $TSB$ | 52189.11 | 51060.93 | 50004.05 | 49048.11 | 47959.67 | 46934.00 | 46102.49 |
| $TUB$ | 112002.79 | 110545.88 | 110570.24 | 109185.19 | 107881.95 | 105805.57 | 105242.74 |
| $TC$ | 59813.68 | 59484.94 | 60566.18 | 60137.07 | 59922.28 | 58871.56 | 59140.25 |
| $p_c$ | 1.534 | 1.486 | 1.378 | 1.330 | 1.308 | 1.302 | 1.243 |

The TC and TUB are obtained by choosing different average time value of EV user, demonstrated in the following Figure 1:

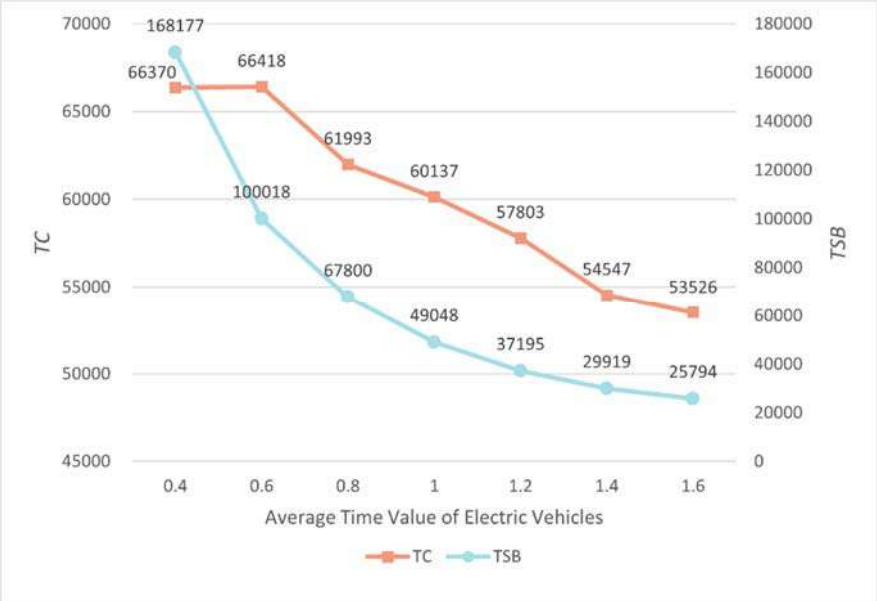



**Figure 1: The TC and TUB by choosing different average time value of EV user**

From the results, the following conclusions can be drawn:

(1) After implementing TCS, the total social welfare of the transportation system is

significantly improved, while the total travel cost of the system is greatly reduced, and the traffic conditions on congested links are significantly improved. However, the total demand of the traffic network has decreased, indicating that the credit charging inevitably have a certain inhibitory effect on users' travel. This scheme not only alleviates traffic congestion by guiding travelers to change their route choices but also further reduces traffic congestion by lowering OD demand. Notably, TCS also effectively increases the proportion of EVs in the equilibrium travel demand, encouraging some travelers to abandon GVs and switch to more environmentally friendly EVs.

(2) Both the decrease in electricity prices and gasoline prices increase social welfare, and the fluctuation of gasoline prices has a more significant impact on social welfare. Both raising the gasoline price and lowering the electricity price can increase the penetration rate of EVs in the equilibrium OD demand. Overall, raising the gasoline price has a more significant effect. This provides a reference for the traffic management authorities when formulating policies on how to balance the adjustment of gasoline prices and electricity prices.

(3) The increase in the average time value of EV users leads to an exponential decline in the total social welfare and significantly suppresses the overall OD demand. The proportion of EVs in the total travel also drops sharply. This is mainly because the value of travel time, charging time, and queuing time becomes more expensive, for EV users with a high time value concept. Although EVs have an advantage in energy consumption, their longer charging and queuing times may make the overall travel cost higher than those of GVs. Therefore, most travelers will abandon EVs and choose GVs for travel. The total OD demand is also suppressed due to this change in travel mode and the increase in travel costs. In addition, the queuing time at charging stations is also significantly affected. By analyzing the changes in charging flow, the utilization rate of charging stations is evaluated, which helps charging station operators and traffic management authorities to plan and manage charging station resources more scientifically, thereby better meeting the changing needs of EV users.

The innovation of this paper is as follows: Firstly, this paper researches the mixed equilibrium problems of hybrid traffic of EVs and GVs under the elastic OD demand to reflect the dynamic characteristics of OD demands in reality. Secondly, this paper adopts an approximate queuing time function related to charging demand and charging flow to more accurately reflect the complex scenario of the interaction between queuing at charging stations and traffic flow.

Therefore, the contribution of this paper lies in two sides: (1) Determining the TCS for hybrid traffic of EVs and GVs under elastic OD demand integrating the approximate average waiting time of the M/G/k system. (2) Exploring the impact of electricity prices, gasoline prices, and travel time value on OD demand and overall operational efficiency in the transportation network. The results can provide references for traffic management authorities to regulate traffic and optimize the allocation of charging station resources for hybrid EVs and GVs.

The deterministic user equilibrium is used in this paper to describe the behavior of the travelers. A more realistic and general situation is that OD demand and travel times are random variables or travel times are perceived by travelers in imperfect, stochastic manner. Thus, the stochastic user equilibrium (SUE) principle can be considered in the future to research the uncertain cases.

**Keywords:** Tradable Credit Scheme (TCS); Hybrid Traffic of Electric and Gasoline Vehicles; Elastic OD Demand; Mathematical Programming with Equilibrium Constraints (MPEC)

DTA 2025 10TH INTERNATIONAL SYMPOSIUM ON DYNAMIC TRAFFIC ASSIGNMENT

# Gating Traffic Control for Wide Area Applications

Franco Filippi[a], Facundo Storani[a], Roberta Di Pace[a], , Stefano de Luca[a]
[a]*Department of Civil Engineering, University of Salerno - Italy*

**Extended Abstract**

Effective traffic management in wide area networks, considering both urban and extra-urban, remains a critical challenge, especially in the presence of dynamic and unpredictable congestion patterns. This study proposes a gating traffic control strategy that integrates Day-to-Day (DTD) and Within-Day control mechanisms to optimize traffic flow in large-scale applications, such as highways connecting multiple urban areas. The DTD control regulates traffic based on historical and expected static flow patterns, ensuring balanced long-term demand distribution. In contrast, the Within-Day control dynamically adjusts gating measures in response to real-time disruptions, such as traffic incidents, lane closures, or sudden congestion spikes. The intended model functions as a Traffic Management Centre (TMC), continuously monitoring and adjusting traffic flow based on evolving conditions. It is expected that the proposed approach will enhance traffic efficiency by mitigating congestion, improving travel time reliability, and providing a more adaptive response to traffic fluctuations. Future research will focus on validating these hypotheses through simulations and real-world case studies.

**Key words:** Gating Control, TMC, Traffic Management

# Exploring multiple traffic assignment solutions resulting from economies of scale

Gunnar Flötteröd and David Watling

February 2025

## 1 Motivation

Economies of scale arise when the utility of a good increases with its usage. We give two examples in the transportation context. (i) In freight assignment, increasing the demand for one shipment alternative allows to deploy larger vehicles and terminals, which reduces the unit shipment cost and may further increase the demand for this alternative. (ii) Including shared mobility services with adaptive pricing schemes in public transport assignment means that the high usage of a shared service allows the provider to reduce its price and further increases its attractiveness.

Including such scale effects in traffic assignment easily leads to a multiplicity of solution points (equilibria). The objective of this ongoing work is to devise techniques that support the credible deployment of simulation-based traffic assignment in settings where a multiplicity of solutions can be expected. See Schmöcker et al. (2014) and Bar-Yosef et al. (2013) for further illustration.

Technically, we consider a discrete-time stochastic assignment process with a stationary distribution of several islands of (weakly connected) probability mass. The process spends, on average, enough iterations in one such region to settle in an apparently stationary local distribution before eventually (stochastically) leaving that region and settling somewhere else. This means that observing the process over a limited number of iterations may give a starting point dependent impression of its stationary distribution. This effect has already been described by Watling (1996); our work is more geared towards the study of scale effects.

## 2 Basic setup

Consider a population of decision-making agents indexed by $n = 1, \ldots, N$. Denote the choice set of agent $n$ by $C_n$. Let $C_{1:N} = C_1 \times C_2 \times \ldots \times C_N$ be the population's joint choice set, and let $i_{1:N} \in C_{1:N}$ represent the choices of all agents.

We are interested in studying models for policy analysis and introduce a policy parameter $\lambda \in [0, 1]$ where zero means "base case", one means "policy

case", and values in-between represent a partical policy implementation. A policy may affect agent choices as well as the network loading, as described immediately below.

Let $\mathcal{X}$ be the set of physical system states, and let $g : (C_{1:N}, [0,1]) \to \mathcal{X}$ be the policy-sensitive network loading where $g(i_{1:N}, \lambda)$ computes the network conditions resulting from the choices $i_{1:N}$ given the policy $\lambda$. Each agent is equipped with a policy-sensitive choice model $P_n : (C_n, \mathcal{X}, [0,1]) \to [0,1]$ where $P_n(i \mid x, \lambda)$ is the probability that agent $n$ chooses alternative $i$ given the network conditions $x$ and policy $\lambda$. Let $P : (C_{1:N}, \mathcal{X}, [0,1]) \to [0,1]$ represent the population choices in that $P(i_{1:N} \mid x, \lambda) = \prod_{n=1}^{N} P_n(i_n \mid x, \lambda)$.

A simple simulation version of the considered process model is given in Algorithm 1. For simplicity, agents only remember their most recent network condition experience.

**Algorithm 1** Stochastic process assignment model

Set policy parameter $\lambda$.
Set initialize initial network conditions $x^{(0)}$.
**for** $k = 1, 2, \ldots$ **do**
  Simulate choices $i_{1:N}^{(k)} \sim P(\cdot \mid x^{(k-1)}, \lambda)$.
  Simulate network conditions $x^{(k)} = g(i_{1:N}, \lambda)$.
**end for**

# 3 Small example

We consider a two-alternative scenario with $N = 1000$ homogeneous agents who all face identical choice sets $C_1 = \ldots = C_N = \{1, 2\}$. The network loading is policy-insensitive and merely counts the number of agents using alternative one:

$$g(i_{1:N}, \lambda) = \sum_{n=1}^{N} \mathbf{1}(i_n = 1). \tag{1}$$

A binomial logit choice model, common to all agents, reflects utilities of scale:

$$P_n(1 \mid x, \lambda) = \frac{e^{\mu \cdot x/N}}{e^{\mu \cdot x/N} + e^{\mu \cdot (1 - x/N)}} \tag{2}$$

with $\mu \geq 0$. Policy-sensitivity will be established further below by making $\mu$ a function of $\lambda$.

The positive $\mu$ parameter in combination with utility increasing with the use of an alternative suggests that this scenario has a tendency to attain stationarity either at $x$ values relatively near zero or relatively near $N$, with possibly very rare switches in-between. Figure 2 illustrates the number of alternative-1 users over 10'000 assignment iterations for $\mu = 2.08$. One observes that the system may stay for thousand of iterations in a state where most decision makers use one particular alternative before flipping to the opposite situation of most decision makers using the other alternative.

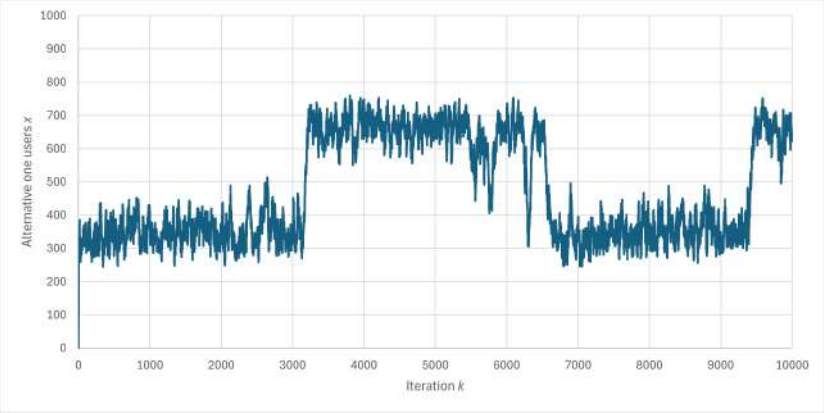



Figure 1: Small example. One process realization.

# 4 Metropolis Hastings based model analysis

To explore solution multiplicity, we deploy the Metropolis-Hastings (MH) algorithm (Hastings, 1970). We define the state space explored by the MH process as $\mathcal{X} \times C_{1:N}, \times [0, 1]$, where one element $(x, i_{1:N}, \lambda)$ of this state space consists of a realization of the network conditions, choices for all agents in the population, and a setting of the policy parameter.

Letting $\phi : \mathcal{X} \to \mathbb{R}_+$ denote a zero-centered probability density function over the network conditions, we construct the proposal distribution

$$q\left((x, i_{1:N}, \lambda) \to (x', i'_{1:N}, \lambda')\right) = \phi_{\text{proposal}}(x - x') \cdot P(i'_{1:N} \mid x', \lambda'), \tag{3}$$

meaning that simulating a proposal amounts to (i) uniformly drawing $\lambda$ between zero and one, (ii) drawing $x'$ from a distribution centered at $x$, and drawing $i'_{1:N}$ by evaluating the choice model given $x'$ and $\lambda'$.

As the target weights, we use

$$w(x, i_{1:N}, \lambda) = P(i_{1:N} \mid x, \lambda) \cdot \phi_{\text{target}}(x - g(i_{1:N}, \lambda)), \tag{4}$$

where the first factor evaluates the likelihood of obtaining the choices $i_{1:N}$ given the network conditions $x$ and the second factor measures the deviation between the network conditions $x$ and the result from a network loading of the choices $i_{1:N}$. In combination, these terms measure the mutual consistency of travel choices and network conditions; this aims to approximate the stationary conditions attained by Algorithm 1.

Given these choices of proposal distribution and target weights, we obtain the simple acceptance probability

$$\alpha\left((x, i_{1:N}, \lambda) \to (x', i'_{1:N}, \lambda')\right) = \min\left\{1, \frac{\phi_{\text{target}}(x' - g(i'_{1:N}, \lambda'))}{\phi_{\text{target}}(x - g(i_{1:N}, \lambda))}\right\}. \tag{5}$$

Important for practical use with a black-box simulator, this setting requires only drawing choices and loading the network; no knowledge of the underlying choice distribution or concrete properties of the network loading is needed.

# 5 Small example, continued

We create a policy-sensitive version of the small example by letting

$$\mu = \mu(\lambda) = (6\lambda - 3)^2 + b \tag{6}$$

with $b$ a structural model parameter; we explore below cases with $b \in [1, 4]$. For any value of $b$ within that range, $\mu$ is strictly positive; it is largest for $\lambda = 0$ and $\lambda = 1$ and smallest for $\lambda = 0.5$. By construction, $\mu(0) = \mu(1)$ so that there is no difference between the base case and the policy case.

We assume that data collected from reality indicates that the base case solution is the one with relatively few users on alternative one, i.e. $x$ near zero for $\lambda = 0$. Given this starting point, we would like to anticipate the effect of gradually introducing the policy measure, i.e. moving $\lambda$ from zero to one. For this, we deploy the MH algorithm where we instantiate both $\phi_{\text{proposal}}$ and $\phi_{\text{target}}$ as univariate Gaussian distributions with zero mean and a standard deviation of five.

We run four experiments. All experiments initialize the MH process with $x = 0$ as an approximation of the observable base case. Extracting one sample every 10'000 iterations and collecting in total 1000 samples per experiment (we did not attempt to fine-tune the algorithm) yields the results shown in Figure 2.

Each single point $(\lambda, x)$ represents a policy parameter $\lambda$ and a realization of the approximate corresponding stationary number of alternative-1 users $x$. For $b = 3$ and $b = 2.5$, a unique path to the policy case $\lambda = 1$ is identified. At $b = 2.0$, a bifurcation arises, and the MH algorithm succeeds to explore all branches of this bifurcation. At $b = 1.0$, two consecutive bifurcations can be observed. The solutions around $x = 0$ resp. $x = 1000$ are locally stable (may persist for many iterations), and those around $x = 500$ are unstable.

Even though available real data may allow to select the right solution in the base case (i.e. to push the simulation model by calibration towards a state that is compatible with the data), this is not possible in the policy case. Which solution a simulation model will predict in the policy case is, without further analysis, a matter of chance (in that it depends on how the simulation model is internally initialized).

The MH analysis thus yields (also practically) valuable insight. For $b = 3, 2.5$, it supports the prognosis that the policy case will not differ from the base case. For $b = 1, 2$, it indicates that the result of implementing the policy is diametrically ambiguous, indicating a need for policy refinement or at least a very careful approach to assessing its implications.

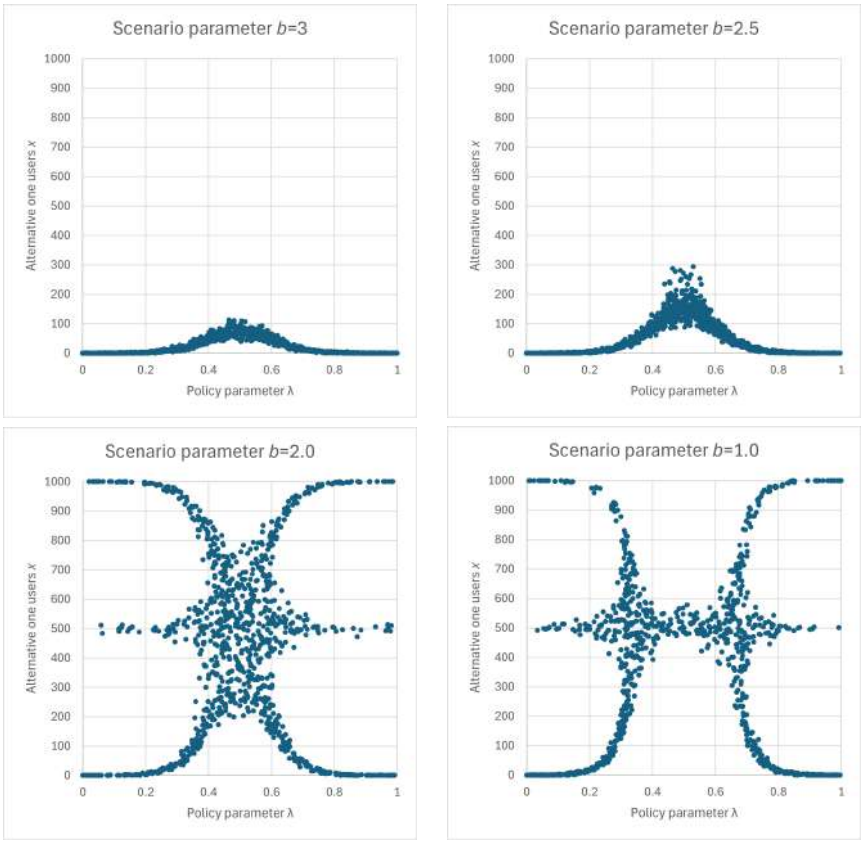



Figure 2: Small example. Alternative-1 users over policy parameter.

# 6 Outlook

We aim to operationalize the analysis presented here for large-scale simulation models in complex scenarios where solution multiplicity can no longer be analyzed by inspecting model structure (as in Iryo and Watling, 2019) or with anecdotal simulation evaluations. The MH machinery is versatile but computationally demanding; this needs to be addressed. We work towards the Swedish national freight model Samgods (Westin et al., 2016) as a real-world test case.

# References


A. Bar-Yosef, K. Martens, and I. Benenson. A model of the vicious cycle of a bus line. *Transportation Research Part B: Methodological*, 54:37–50, 2013.

W. K. Hastings. Monte Carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

T. Iryo and D. Watling. Properties of equilibria in transport problems with complex interactions between users. *Transportation Research Part B*, 126: 87–114, 2019.

J.-D. Schmöcker, T. Hatori, and D. Watling. Dynamic process model of mass effects on travel demand. *Transportation*, 41:279–304, 2014.

D. Watling. Asymmetric problems and stochastic process models of traffic assignment. *Transportation Research Part B*, 30(5):339–357, 1996.

J. Westin, I. Vierth, G. de Jong, R. Karlsson, N. Krüger, and M. Johansson. Analyzing model uncertainty and economies of scale of the swedish national freight model to changes in transport demand. *European Journal of Transport and Infrastructure Research*, 16(4), 2016.

**An operational alternative to origin/destination matrices**

Gunnar Flötteröd[1]

[1]Department of Communication and Transport Systems Linköping University, Sweden
& The Swedish National Road and Transport Research Institute

2025-01-19

## SHORT SUMMARY

A round-trip based alternative to origin/destination matrices is presented. The formal approach treats the round-trips of a population as a distributed quantity and adopts statistical techniques to the evaluation of this distribution. An operational Bayesian calibration approach is presented as a round-trip-based counterpart to origin/destination matrix estimation. Several concrete applications of the framework are presented. We believe that the vastly increased representative power of a round-trip-based model when compared to origin/destination matrices, in combination with its compatibility with modern agent-based simulation packages, outweighs its somewhat more involved technical development.

## 1 INTRODUCTION

Origin/destination (OD) matrices are ubiquitous in both person and freight network assignment modeling. The rows of a static OD matrix represent origins, its columns represent destinations, and its entries represent (freight or person) transport between origins and destinations. Dynamic OD matrices add a (usually discrete) time index, meaning that they map the triplet (origin, destination, time index) onto a transport demand. OD matrices are covered in any transport modeling textbook (e.g., Ortuzar and Willumsen, 2004; Cascetta, 2001).

The widespread use of OD matrices may be traced back to three ingredients: their simple and intuitive structure, their compatibility with mainstream network assignment models, and the availability of a comprehensive mathematical machinery for estimating OD matrices. Still, their appealing simplicity comes with limitations: An OD matrix does not encode any relationship between the individual movements it represents, neither does a dynamic OD matrix encode any temporal relationships. The resulting independence assumption across movements may violate mass conservation (e.g., more travelers leave a shopping mall than enter it) and causality (e.g., even if the time-sum of entries and exits to a shopping mall is zero, the entries may occur after the exits). In addition, the cell entries of a single OD matrix represent a homogeneous number of anonymous movements, and capturing heterogeneity by introducing group-specific OD matrices quickly reaches the computational limits of representing a correspondingly large number of OD matrices with correspondingly small entries. For similar reasons, the number of time steps resp. origins and destinations (usually traffic analysis zones) is limited, introducing a temporal resp. spatial aggregation bias.

These observations are not new and were one of the drivers for developing agent-based models (Horni et al., 2016; Nagel and Flötteröd, 2012). An agent-based transport demand is represented based on one trip-list per agent, with each trip being (minimally) annotated by origin, destination, and departure time. This allows, at the agent level, to ensure both spatial consistency (origin of one trip must be destination of the previous trip) and temporal consistency (departure

time of one trip must not be earlier than arrival time of the previous trip). Agent-based models for strategic planning (where a long-term, stationary state of the transport system of interest is considered) may consider circular trip lists: Person agents return to their homes on a daily basis, freight vehicle agents return to their depots at possibly larger time intervals (consider, for example, weekly train schedules).

This document describes an operational simulation/estimation framework for trip-list-based transport representations. It makes concrete and develops further the ideas sketched in Flötteröd (2024), going beyond that reference by (i) specifying in detail a round-trip sampling approach, (ii) considering not just one but arbitrarily many round-trips simultaneously, and (iii) presenting a Bayesian calibration framework. We believe that the vastly increased representative power of this approach when compared to OD matrices, in combination with its compatibility with modern agent-based simulation packages, outweighs the somewhat more involved technical developments.

## 2 METHOD

### *Round-trip-lists*

We discretize time into intervals (time bins) $k \in \mathcal{K} = \{1, \dots, K\}$ and consider a population of agents $n = 1, \dots, N$. A location set $\mathcal{L}_n$ of size $L_n$ is available to agent $n$. We define agent $n$'s location list $l_n = [l_n^{(1)}, \dots, l_n^{(J_n)}]$ and departure time bin list $d_n = [d_n^{(1)}, \dots, d_n^{(J_n)}]$ where $J_n \in \{1, \dots, \min\{J^{\max}, K\}\}$ is the finite length of both lists, $l_n^{(j)}$ is the origin location of the $j$th trip, and $d_n^{(j)}$ is its departure time bin. The location and departure time list of agent $n$ constitute its trip list $x_n = (l_n, d_n)$. To treat trip lists independently of an agent, we introduce the function $J(\cdot)$ that maps a trip list onto its length; obviously, $J(x_n) = J_n$. Causality is established by requiring

$$d_n^{(j)} < d_n^{(j+1)} \quad \text{for all } j \in \{1, \dots, J_n - 1\}. \tag{1}$$

(Note that this implies the above requirement $J_n \leq K$.) Repeated trips with the same departure location are allowed for to enable the representation of intra-zonal travel when locations are traffic analysis zones.

We subsequently consider *round-trips* where the destination location of the last trip is the departure location of the first trip.[1] For instance, given an hourly time discretization, the round-trip

$$([\text{home, office, shopping mall}], [6,\ 16,\ 17])$$

means that the considered agent leaves home at 6 am for office work, departs at 4 pm from the office, makes a short stop at the shopping mall that is planned to end at 5 pm, and then returns back home. The departure time bins represent a *desired* time structure that may or may not be compatible with a given physical reality of finite travel speeds; more on this further below.

We consider *distributed* population round-trips $X = [X_1, \dots, X_N]$ where $\Pr(X = x) = \Pr(X_1 = x_1, \dots, X_N = x_N)$ is a discrete probability with (for finite population and maximum trip-list length) finite support. To evaluate this distribution, a method to sample round-trips from a given target distribution $p(x)$ is required. For generality, we do not yet make assumptions about where this target distribution comes from (ample examples further below) but merely assume it to be given. We rely here on the Metropolis-Hastings (MH) algorithm (Hastings, 1970). This algorithm has well-known advantages (generality) and disadvantages (possibly long run-times), the latter having been addressed in a substantial body of literature. Ross (2012) offers a detailed introduction. We focus subsequently on a basic round-trip-specific instance of the MH algorithm,

[1] To capture trip-lists that are not round-trips within the same framework, one may add a $(J_n+1)$th trip where $l_n^{(J_n+1)}$ defines the destination of the $J_n$th trip and $d_n^{(J_n+1)}$ is arbitrary.

omitting general-purpose discussions around its practicalities (convergence tests, extraction of statistics, etc.).

The state space of the considered MH algorithm is composed of population round-trips $x = [x_1, \ldots, x_N]$. The algorithm requires an irreducible proposal distribution $p(x, y)$ from any state $x$ to any state $y$, which is subsequently developed.

***MH proposal distribution for a single round-trip ($N = 1$)***

Since a single round-trip is considered, the agent index $n$ is suppressed in this section. Four operations on a single round-trip are subsequently defined; their combination will yield the desired proposal distribution.

**INS(ERT)** Given a round-trip $x$ of length $J < J^{\max}$, an insertion index $i$ is uniformly drawn from $\{1, \ldots, J+1\}$. A new location is drawn uniformly from the location set $\mathcal{L}$, and a new departure time bin is drawn uniformly from the set of not yet used departure time bins $\mathcal{K} \setminus \cup_{j=1}^{J} d^{(j)}$. For $i \leq J$, these values are inserted into the location and departure time list at index $i$. For $i = J+1$, they are appended to the end of the lists. To comply with (1), the departure time list is subsequently ordered by increasing magnitude *without changing the ordering of the location list*. Letting the round-trip $y$ be the result of an insert operation at index $i$ in round-trip $x$, the probability of obtaining $y$ from $x$ is

$$q_{\text{INS}}(x, y) \quad = \quad \frac{n_{\text{LOC-INS}}(x, y)}{J+1} \cdot \frac{1}{L} \cdot \frac{1}{K-J} \tag{2}$$

where $n_{\text{LOC-INS}}(x, y)$ is the number of location indices in round trip $x$ where inserting a location allows to recover the location sequence of round trip $y$. Since the length of the round-trip increases by one, the new round-trip differs from the old one.

**REM(OVE)** Given a round-trip $x$ of length $J > 0$, two indices $i$ and $j$ are independently and uniformly drawn from $\{1, \ldots, J\}$, the $i$th element is removed from the location list, and the $j$th element is removed from the departure time list. Letting $y$ be the result of a remove operation from round-trip $x$, the probability of obtaining $y$ from $x$ is

$$q_{\text{REM}}(x, y) \quad = \quad \frac{n_{\text{LOC-REM}}(x, y)}{J^2} \tag{3}$$

where $n_{\text{LOC-REM}}(x, y)$ is the number of location indices in round trip $x$ where removing the location recovers the location sequence of round trip $y$. Since the length of the round-trip decreases by one, the new round-trip differs from the old one.

**FLIP_LOC(ATION)** Given a round-trip $x$ of length $J$, a location flip index $i$ is uniformly drawn from $\{1, \ldots, J\}$. A new location is uniformly drawn from $\mathcal{L} \setminus l^{(i)}$ to replace $l^{(i)}$. Letting $y$ be the result of flipping a location in round-trip $x$, the probability of obtaining $y$ from $x$ is

$$q_{\text{FLIP\_LOC}}(x, y) \quad = \quad \frac{1}{J} \cdot \frac{1}{L-1}. \tag{4}$$

Since the newly drawn location must differ from the original value, the new round-trip differs from the old one.

**FLIP_DEP(ARTURE)** Given a round-trip $x$ of length $J$, a departure time flip index $i$ is uniformly drawn from $\{1, \ldots, J\}$. A random departure time is drawn from $\mathcal{K} \setminus \cup_{j=1}^{J} d^{(j)}$ to replace $d^{(i)}$. The resulting departure time list is sorted by increasing magnitude without

changing the ordering of the location list. Letting $y$ be the result of flipping a departure time in round-trip $x$, the probability of obtaining $y$ from $x$ is

$$q_{\text{FLIP\_DEP}}(x, y) \quad = \quad \frac{1}{J} \cdot \frac{1}{K - J}. \tag{5}$$

Since the newly drawn time must differ from the original value, the new round-trip differs from the old one.

Given a round-trip $y$ that results from applying any of the above operations to a given round-trip $x$, the applied operation is uniquely given: $J(y) > J(x)$ can only result from INS; $J(y) < J(x)$ can only result from REM; $J(x) = J(y)$ and $l_x \neq l_y$ can only result from FLIP_LOC; $J(x) = J(y)$ and $d_x \neq d_y$ can only result from FLIP_DEP; $J(x) = J(y)$ and $l_x = l_y$ and $d_x = d_y$ is impossible. Let $\phi_{\text{INS}}, \phi_{\text{REM}}, \phi_{\text{FLIP\_LOC}}, \phi_{\text{FLIP\_DEP}}$ be the selection probabilities of the respective operations; they are strictly positive and sum up to one.[2] Drawing an operation, applying it to round-trip $x$ and receiving round-trip $y$ hence occurs with the following probability:

$$q_{[1]}(x, y) \quad = \quad \begin{cases} \phi_{\text{INS}} \cdot q_{\text{INS}}(x, y) & \text{if } J(y) > J(x) \\ \phi_{\text{REM}} \cdot q_{\text{REM}}(x, y) & \text{if } J(y) < J(x) \\ \phi_{\text{FLIP\_LOC}} \cdot q_{\text{FLIP\_LOC}}(x, y) & \text{if } J(y) = J(x) \wedge l_x \neq l_y \\ \phi_{\text{FLIP\_DEP}} \cdot q_{\text{FLIP\_DEP}}(x, y) & \text{if } J(y) = J(x) \wedge d_x \neq d_y \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

This distribution is irreducible because any state (round-trip) $b$ can be reached from any other state $a$ with positive probability: (i) If $J(a) = J(b)$, a sequence of FLIP_LOC and FLIP_DEP operations that turn $a$ into $b$ is possible. (ii) If $J(a) < J(b)$ resp. $J(a) > J(b)$, a sequence of INS resp. REM operations is possible that yields an intermediate state $a'$ with $J(a') = J(b)$, which then can be turned into $b$ according to (i).

***MH proposal distribution for multiple round-trips*** ***($N > 1$)***

We return to indexing the different round-trips in $x$ by $n = 1, \ldots, N$ and compose the proposal distribution for population round-trips from that for single round-trips. We sweep once over all round-trips $n = 1, \ldots, N$ and modify round-trip $x_n$ with probability $\phi_{\text{MOD}} > 0$ into $y_n \neq x_n$ according to $q(x_n, y_n)$ in (6), otherwise we let $y_n = x_n$.[3] If this results in $x = y$ (meaning that no modification has taken place), we repeat the process until $x \neq y$. This implements the proposal distribution

$$q_{[1:N]}(x, y) \quad = \quad \begin{cases} \frac{1}{1-\phi_{\text{MOD}}^N} \prod_{n=1}^{N} \begin{cases} \phi_{\text{MOD}} \cdot q_{[1]}(x_n, y_n) & \text{if } x_n \neq y_n \\ 1 & \text{otherwise} \end{cases} & \text{if } x \neq y \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

We have already established that any single round-trip can be turned into any other single round-trip by a suitable operation sequence. Consider now the problem of turning any population round-trip $a = (a_1, \ldots, a_N)$ into any other population round-trip $b = (b_1, \ldots, b_N)$. This can be achieved by only selecting $n = 1$ for modification until $a_1$ has been turned into $b_1$, then doing the same for $n = 2$, etc. This sequence of selecting round-trips for modification arises with positive probability, and so does every single modification from $a_n$ into $b_n$. This establishes the irreducibility of the population proposal distribution $q_{[1:N]}$.

[2] Our experimentation so far suggests that a uniform selection distribution performs well.

[3] Letting $\phi_{\text{MOD}} = 1/N$ has performed well in our experimentation so far.

### *MH target weights*

A major advantage of the MH algorithm in the given context is that it only requires an un-normalized version $t(x)$ of a given target distribution $p(x)$, meaning that any $t(x)$ with $p(x) = t(x)/\sum_{x'} t(x')$ for all $x$ is sufficient to use the algorithm to draw from $p(x)$. This avoids a normalization of the target distribution over the possibly gigantic state space.

The time structure of a round-trip represents an ambition; its physically feasibility depends on the movement duration between its locations. When there is a need to evaluate the physical realization of a round-trip, a deterministic mapping $g$ from the population round-trips $x$ onto some data structure representing the realized movement experiences (a movement simulation) can be used. The corresponding target distribution would still only depend on $x$ but internally also evaluate the movement simulation $g$. We subsequently suppress the possible use of $g$ when writing out a target distribution.

The concrete form of the target distribution is application-specific. We discuss below the problem of calibrating a population round-trip distribution against some given data-set. This constitutes the round-trip-based counterpart of OD matrix estimation.

### *Round-trip calibration and uninformed prior distribution*

We adopt a Bayesian approach. Given the data $y$, a likelihood function $p(y \mid x)$ is formulated that expresses the probability of observing the data $y$ given the population round-trips $x$. A Bayesian approach to conditioning the round trip distribution onto the data amounts to sampling from the (un-normalized) target distribution

$$t(x \mid y) \quad \sim \quad p(y \mid x) p^{\text{prior}}(x) \tag{8}$$

where the prior distribution $p^{\text{prior}}(x)$ represents available knowledge about the population round-trips before having seen the data $y$; if available, this may comprise travel behavioral assumptions (Flötteröd et al., 2011).

The construction of an uniformed prior distribution requires some care. Considering any agent $n$, the number of round-trip configurations of length $J$ available to that agent is

$$\#_n(J) \quad = \quad L_n^J \binom{K}{J} \tag{9}$$

where the first factor accounts for all possible location configurations and the second factor represents the number of possible departure time configurations, given their ordering by increasing magnitude. This means that the number of available round-trip configurations of a given length grows combinatorically with that length. A naive uniform prior distribution over all possible round-trips of an agent would hence be biased towards longer round-trips.

A maximum entropy (ME) approach is hence adopted to represent a maximally uninformed round-trip distribution. The prior distribution $p_n^{\text{ME}}(x_n)$ over all possible round-trips of agent $n$ is chosen to maximize entropy subject to the constraint that the expected round-trip length equals some exogenously given parameter $\overline{J}_n$.[4] The corresponding Lagrangian reads

$$L(p_n^{\text{ME}}, \lambda, \gamma) \quad = \quad -\sum_{x_n} p_n^{\text{ME}}(x_n) \ln p_n^{\text{ME}}(x_n) \ldots \tag{10}$$

$$+\lambda \cdot \left( \sum_{x_n} p_n^{\text{ME}}(x_n) - 1 \right) + \gamma \cdot \left( \sum_{x_n} J(x_n) p_n^{\text{ME}}(x_n) - \overline{J}_n \right) \tag{11}$$

[4] The arguably simplest setting of this parameter is $J_n = J$ for all $n$, with $J$ the estimated total number of trips made in a study region divided by its population.

where the first constraint with multiplier $\lambda$ requires probabilities to sum to one and the second constraint with multiplier $\gamma$ enforces the expected length. Evaluating first-order conditions on $p_n^{\mathrm{ME}}$:

$$\frac{dL}{dp_n^{\mathrm{ME}}(x_n)} = 0 \quad \Rightarrow \quad e^{\lambda-1}e^{\gamma J(x_n)}. \tag{12}$$

Inserting this into the probability sum constraint yields

$$p_n^{\mathrm{ME}}(x_n) \quad = \quad \frac{e^{\gamma J(x_n)}}{\sum_{J=0}^{\min\{K,J^{\max}\}} \#_n(J)e^{\gamma J}}, \tag{13}$$

which has, unsurprisingly (Anas, 1983), the form of a multinomial logit round-trip choice model with scale parameter $\gamma$. This parameter is obtained by inserting (13) into the mean length constraint and numerically solving the one-dimensional problem

$$\frac{\sum_{J=0}^{\min\{K,J^{\max}\}} J \cdot \#(J)e^{\gamma J}}{\sum_{J=0}^{\min\{K,J^{\max}\}} \#(J)e^{\gamma J}} \quad = \quad \overline{J}_n \tag{14}$$

for $\gamma$.

Overall, the ME population round-trip distribution becomes

$$p^{\mathrm{ME}}(x) \quad = \quad \prod_{n=1}^{N} p_n^{\mathrm{ME}}(x_n) \tag{15}$$

in combination with (13) and (14). The symmetry of this development to the long-standing ME-based estimation of OD matrices may be noteworthy (Van Zuylen and Willumsen, 1980).

## 3 RESULTS AND DISCUSSION

A primary result of this work is the above described modeling/estimation framework, which is accompanied by a freely available software implementation at `https://github.com/vtisweden/matsim-projects/tree/master/roundtrips`. The remainder of this section summarizes existing and ongoing applications of the framework, illustrating its versatility and applicability.

An early version of the framework was applied by Flötteröd (2024) to study daily electrical vehicle charging patterns in the Swedish municipality Skellefteå. To accommodate location-specific charging decisions, the physical network was extended into one where each location was duplicated into one (location, charging) and one (location, no-charging) node; sampling paths on the extended network hence also induced temporal charging patterns. The target distribution aimed to represent basic time- and land use assumptions. Figure 1(left) shows part of the study region, and Figure 1(right) exemplifies the estimated home- and time-of-day-dependent en-route charging patterns.

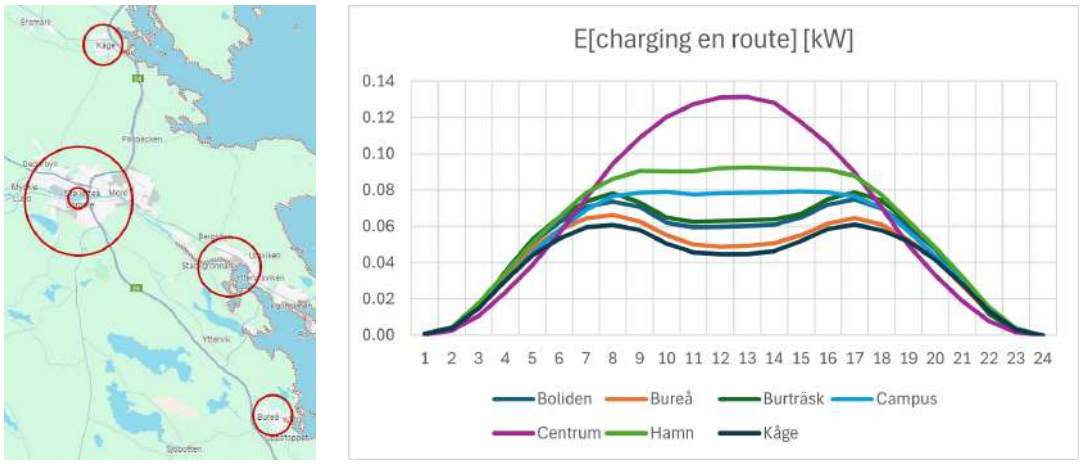



Figure 1: Electric vehicle charging analysis

A more recent application is that of Charalampidou et al. (2025) where implications of shopping location placement and opening times in a fifteen-minute-city setting are investigated. Here, the round trip encodes weekly activity/travel patterns. To encode activity participation, the physical network is extended into one where each location is duplicated into several (location, activity) nodes, representing participation in a particular activity at that location. The target distribution is set such that round trips respecting weekly activity-specific time use assumptions receive high probabilities. Figure 2(left) shows the study region (the Liesing district in the city of Vienna, Austria) and Figure 2(right) exemplifies the resulting shopping activity participation time structure.

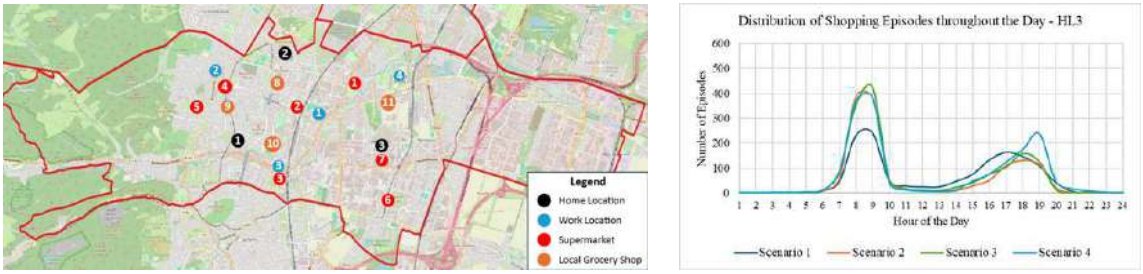



Figure 2: Time use and activity participation in a 15-minute-city

An instance of a multi-round trip calibration for the entire city of Vienna is presented by Rupprecht et al. (2025). The approach starts out from the ME prior (15) and combines it with two likelihood terms, one representing the all-day reproduction of an available static target OD matrix, and the other one representing spatio-temporal travel and activity participation summary statistics. The scatter-plot in Figure 3(left) displays the almost perfect reproduction of the static target OD matrix from a representative population of 50'000 round-tips, while Figure 3(right) illustrates the within-day time structure of the same 50'000 round-trips. The resulting population round-trips serve as initial travel plans file for an agent-based route/mode/departure time network assignment model.

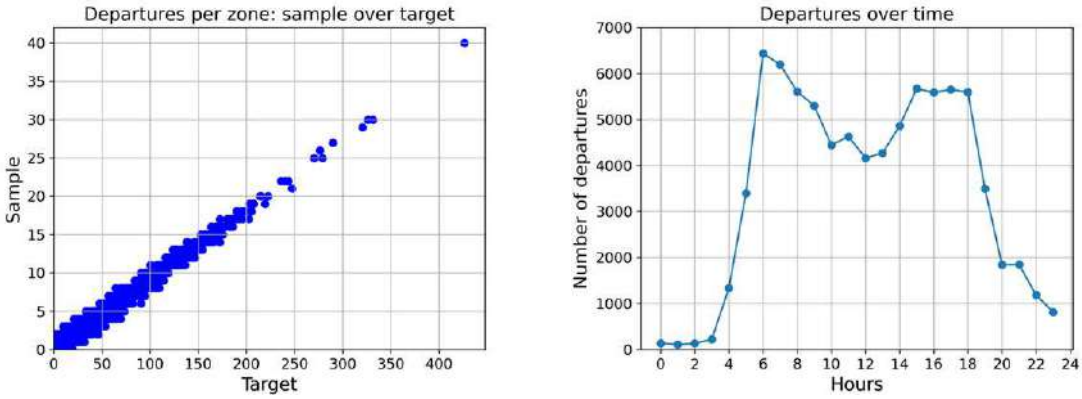



Figure 3: All-day travel pattern synthesis for agent-based traffic assignment

One ongoing but not yet documented application is the inclusion of travel survey data in the population round-trip estimation problem, aiming to complement the Swedish national person transport model with time-of-day dynamics. Another application is the generation of freight vehicle round-trips for Sweden and neighboring countries, aiming at an improved representation of freight consolidation in the Swedish national freight model.

## 4 SUMMARY

An operational round-trip based alternative to OD matrices has been presented, comprising a formal framework, an estimation method, and several application examples. The idea of replacing OD matrices by (round-)trip-lists has been around for a long time, and any agent-based transport simulation packages inevitably relies on some kind of traveler- or vehicle (round-)trip-list discretization of a possibly given OD matrix. This document contributes to this a formal and operational machinery to sampling population round trips from general target distributions. A concrete Bayesian calibration approach resting on this machinery is developed, offering a round-trip based alternative to OD matrix estimation.

Ample further developments of the presented method are possible; these comprise the specification of other than uniform MH proposal distributions, refinements of the basic MH approach (e.g. into Gibbs sampling where the round trip distribution of any single agent is conditioned on the round trips of all other agents), and the not obvious exploitation of parallel computing facilities to speeding up the inherently sequential MH process. A comparison to generative AI approaches such as that of Shone and Hillel (2024) may be attempted, even though much of the appeal of the presented method is its derivation from first principles, while generative AI appears to be more of a black-box approach.

## ACKNOWLEDGMENTS

This work was supported by the Swedish Transport Administration, grants TRV 2023/33572 and TRV 2023/109327.

## BIBLIOGRAPHY

Alex Anas. Discrete choice theory, information theory and the multinomial logit and gravity models. *Transportation Research Part B: Methodological*, 17(1):13–23, 1983. ISSN 0191-2615.

doi: https://doi.org/10.1016/0191-2615(83)90023-1. 2

E. Cascetta. *Transportation Systems Engineering: Theory and Methods*. Kluwer Academic Publishers, 2001. 1

G. Charalampidou, G. Flötteröd, and Y. Susilo. Sampling weekly activity schedules to analyze shopping behavior under the 15-minute city concept. In *(submitted) 13th Symposium of the European Association for Research in Transportation*, Munich, Germany, 2025. 3

G. Flötteröd. Charging demand analysis in the face of overwhelming uncertainty. In *12th Symposium of the European Association for Research in Transportation*, Aalto, Finland, 2024. 1, 3

G. Flötteröd, M. Bierlaire, and K. Nagel. Bayesian demand calibration for dynamic traffic simulations. *Transportation Science*, 45(4):541–561, 2011. 2

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 04 1970. ISSN 0006-3444. doi: 10.1093/biomet/57.1.97. 2

Andreas Horni, Kai Nagel, and Kay Axhausen, editors. *Multi-Agent Transport Simulation MATSim*. Ubiquity Press, London, Aug 2016. ISBN 978-1-909188-75-4, 978-1-909188-76-1, 978-1-909188-77-8, 978-1-909188-78-5. doi: 10.5334/baw. 1

K. Nagel and G. Flötteröd. Agent-based traffic assignment: going from trips to behavioral travelers. In R.M. Pendyala and C. Bhat, editors, *Travel Behaviour Research in an Evolving World*, chapter 12, pages 261–293. Emerald Group Publishing, Bingley, United Kingdom, 2012. 1

J. Ortuzar and L.G. Willumsen. *Modelling Transport*. Wiley, 2004. 1

S.M. Ross. *Simulation*. Knovel Library. Elsevier Science, 2012. ISBN 9780124158252. 2

F.-X. Rupprecht, G. Flötteröd, and Y. Susilo. Metropolis-hasting based synthesis of all-day round-trips from od-matrices and limited population segmentation. In *(submitted) 13th Symposium of the European Association for Research in Transportation*, Munich, Germany, 2025. 3

F. Shone and T. Hillel. Activity sequence modelling with deep generative models. In *12th Symposium of the European Association for Research in Transportation*, Aalto, Finland, 2024. 4

Henk J. Van Zuylen and Luis G. Willumsen. The most likely trip matrix estimated from traffic counts. *Transportation Research Part B: Methodological*, 14(3):281–293, 1980. ISSN 0191-2615. doi: https://doi.org/10.1016/0191-2615(80)90008-9. 2

# A Safety-Based Traffic Network Equilibrium Model Considering Vehicle Coordination in Mixed Traffic Flow

Huimin Tang*, Chunyang Han‡, Yiping Liu†, Zhanbo Sun*, and Tiantian Chen†
* The School of Transportation and Logistics, Southwest Jiaotong University, Chengdu 611756, China.
† The Cho Chun Shik Graduate School of Mobility, Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea.
‡ The Faculty of Transportation Engineering, Kunming University of Science and Technology, Kunming 650500, China.
*Corresponding author: Tiantian Chen (email: nicole.chen@kaist.ac.kr)*

***Abstract*—This paper develops a network-level equilibrium framework for analyzing mixed traffic involving human-driven vehicles (HVs) and autonomous vehicles (AVs), emphasizing both efficiency and safety. A Crash Risk Cost (CRC) model, incorporating probabilistic safety risks at road links and intersections, is integrated into a generalized travel cost function. We formulate a variational inequality (VI) model to capture mixed traffic equilibrium under various AV penetration rates and behavioral strategies: selfish user equilibrium (UE), private monopolist (fleet-based), and social planner (network-wide optimization). Numerical experiments on the Nguyen-Dupuis network indicate that moderate AV penetration (10%) initially increases system costs due to mixed-traffic inefficiencies, while higher adoption rates lead to notable cost reductions. The social planner scenario demonstrates the greatest improvement, underscoring the importance of coordinated AV deployment strategies to achieve system-wide safety and efficiency benefits.**

## I. Introduction

Autonomous vehicles (AVs) are seen as a promising solution to improving urban traffic efficiency and reducing accidents [1]. AVs help manage traffic by adjusting speed and following distances, reducing congestion waves. On safety, studies report that over 90% of road crashes result from human error [2]. By shifting driving tasks to Automated Driving Systems (ADS), AVs can significantly reduce these errors, leading to fewer crashes and fatalities [3]. However, these benefits rely on high AV adoption rates, which will take time [4]. Until then, AVs will coexist with human-driven vehicles (HVs) in mixed traffic flow, where AV-HV interactions may lead to conflicts, especially at intersections [5]. These challenges highlight the need for models assessing AV-HV interactions and their network-wide impacts.

HVs typically make route choices based on Stochastic User Equilibrium (SUE), minimizing personal costs with limited traffic information [6]. AVs, in contrast, can access real-time traffic data through V2V and V2I communication, enabling more informed decisions [7]. Unlike HVs, AVs can be programmed to optimize system-wide performance, but their objectives vary. Private AVs minimize personal costs, fleet operators like Tesla or Baidu optimize for their vehicle fleets, and governments may use AVs to improve overall traffic efficiency . These different objectives shape AV route choices, influencing traffic patterns and risk distribution.

With AV adoption increasing in cities like Shanghai and Los Angeles, research on mixed traffic has expanded [8]. Past studies on human-driven vehicles (HVs) explored factors like travel time [9], fuel consumption [10], route reliability [11], and road safety [12]. However, studies on AV route choice remain limited. Most research focuses on travel time [13], with some considering fuel costs [14] or AV ownership costs [15]. Yet, research suggests that in mixed traffic, safety is a primary concern.

To balance safety and efficiency, previous studies explored strategies like dedicated AV lanes and adaptive control systems [16]. Other studies analyzed AVs' impact on highway safety [17], particularly regarding rear-end collision risks. However, these studies mainly focus on road links, neglecting intersections, where AV-HV conflicts are more likely [5]. Some studies optimize single intersections with multi-lane control [18], while others analyze multiple intersections considering upstream-downstream interactions. However, focusing only on intersections or road links may miss the broader impact of mixed traffic on the entire network. Since transportation planners prioritize system-wide performance, understanding how AV-HV travel behavior influences overall traffic patterns is critical.

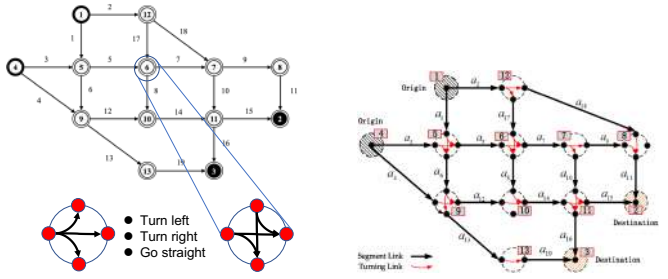
This paper addresses this gap by developing a network-level equilibrium model that captures AV-HV interactions, considering efficiency and safety at both road links and intersections. Using Crash Risk Cost (CRC) models, we analyze mixed traffic equilibrium under different AV adoption levels. Our previous research on network equilibrium has combined safety and efficiency by modeling risk costs through probability distributions [12]. However, two key challenges remain: (i) Integrating mixed traffic into equilibrium models – Most studies focus separately on AVs or HVs, lacking a unified model for their shared network interactions. (ii) Modeling adaptive AV behavior – AVs adjust driving strategies dynamically, requiring a more sophisticated approach to reflect their effects on network-wide equilibrium.

## II. Methodology

This section develops a mixed traffic equilibrium framework that integrates both efficiency and safety considerations. We first introduce the CRC model to quantify safety risks on road links and intersections, capturing both reliability and uncertainty. These safety measures are incorporated into a generalized travel cost function, which combines travel time and crash risk costs. Based on this, we formulate a variational inequality (VI) model to analyze route choice behavior under different AV adoption scenarios, where HVs follow a stochastic user equilibrium (SUE) and AVs exhibit varying travel strategies. This framework provides a comprehensive approach to evaluating mixed traffic flow dynamics and network-wide impacts.

### A. Extend network structure

This study utilizes the commonly used Nguyen and Dupuis network as shown in Fig. 1a. Typically, this network consists of links and nodes. However, in this study, we focus on the impact of intersection (i.e., node) turning movements on safety performance. Therefore, we expand the nodes into left-turn, right-turn, and through movements. For details, see Fig. 1b.




(a) Original Network (b) Extended Network

Fig. 1. Comparison of the original and extended Nguyen and Dupuis network.

### B. Safety models

The uncertainty of road crash risk subconsciously influences travelers' behaviors. Travelers evaluate road safety not only based on an exact value but also on their knowledge of how uncertain the safety of the road is. A single deterministic value is inadequate to reflect this safety perception. Thus, road crash risk $r$ is assumed to be a random variable that follows a distribution with a mean $\mathbb{E}(r_p)$ and variance $\mathrm{Var}(r_p)$.

*1) Crash Risk Cost:* Crash Risk Cost (CRC) describes the potential cost (e.g., economic or time loss) caused by traffic accidents on a specific route $p$. It is assumed that the CRC $r_p$ follows a log-normal distribution. For a path $p$, the mean and variance of CRC are determined by the CRC of its road links and intersection turns (as shown in Figure 1):

$$r_p \sim \mathcal{N}(\mathbb{E}(r_p), \mathrm{Var}(r_p)) \tag{1}$$

$$\mathbb{E}(r_p) = \sum_{a\in A}(\mathbb{E}(r_a)\cdot\delta_a^p) + \sum_{a\to b\in A'}(\mathbb{E}(r_{a\to b})\cdot\delta_{a\to b}^p), \tag{2}$$

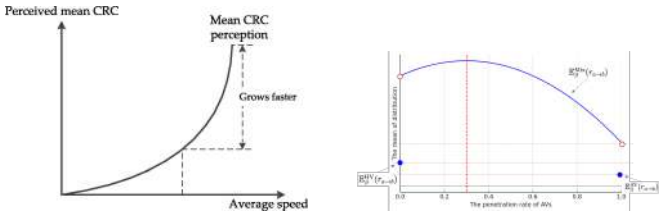
$$\mathrm{Var}(r_p) = \sum_{a\in A}(\mathrm{Var}(r_a)\cdot\delta_a^p) + \sum_{a\to b\in A'}(\mathrm{Var}(r_{a\to b})\cdot\delta_{a\to b}^p), \tag{3}$$

where $\mathbb{E}(r_p)$ is the expected value of CRC, representing the long-term average safety characteristic, and $\mathrm{Var}(r_p)$ is the variance of CRC, capturing the uncertainty in safety.

**For links**, it is assumed that AVs and HVs have the same average speed on a given link in this study. This is because vehicles typically travel in a car-following manner, and additional overtaking behavior is not considered in this paper. Therefore, the CRC distribution for a link is expressed as Equation 4, Equation 5, and Fig.2a:

$$\mathbb{E}(r_a) = t_a\gamma(v_a)^\eta \tag{4}$$

$$\mathrm{Var}(r_a) = t_a^2\bar{\gamma}(v_a)^{\bar{\eta}} \tag{5}$$




(a) Relationship between mean CRC and average speed for a link.

(b) Relationship between mean CRC and AV adoption for an intersection.

Fig. 2. Mean CRC for a link and an intersection.

where $t_a$ is the travel time on link $a$, $v_a$ is the average speed, $\gamma$ and $\bar{\gamma}$ are adjustment coefficients, and $\eta$ and $\bar{\eta}$ are exponential coefficients greater than 1.

**For intersections**, assuming that turning movements and mixed traffic flows are independent, the mean of the CRC distribution under mixed traffic conditions is given by the product of the means of two distributions: the safety cost associated with mixed traffic flow, $\mathcal{N}(\mathbb{E}_\alpha(r_{a\to b}), \mathrm{Var}_\alpha(r_{a\to b}))$, and the safety cost related to AV adoption, $\mathcal{N}(\mathbb{E}_\beta(r_{a\to b}), \mathrm{Var}_\beta(r_{a\to b}))$.

The safety cost associated with mixed traffic flow are shown in Equation (6), Equation (7).

$$\mathbb{E}_\alpha(r_{a\to b}) = \tau(x_{AV}+x_{HV})^\psi \tag{6}$$

$$\mathrm{Var}_\alpha(r_{a\to b}) = \tau'(x_{AV}+x_{HV})^{\psi'} \tag{7}$$

where $\tau$ and $\bar{\tau}$ are scaling coefficients, and $\varphi_n$ and $\bar{\varphi}_n$ are exponential coefficients for different turning types.

The data shows that traffic accidents are more likely to occur at road intersections, with differing levels of risk for left turns, right turns, and straight movements [12, 19].

$$\mathbb{E}_\beta^n(r_{a\to b}) = \begin{cases} \mathbb{E}_\beta^l(r_{a\to b}), & \text{if turning left, } n=1,\\ \mathbb{E}_\beta^r(r_{a\to b}), & \text{if turning right, } n=2,\\ \mathbb{E}_\beta^s(r_{a\to b}), & \text{if going straight, } n=3. \end{cases} \tag{8}$$

$$\mathrm{Var}_\beta^n(r_{a\to b}) = \begin{cases} \mathrm{Var}_\beta^l(r_{a\to b}), & \text{if turning left, } n=1,\\ \mathrm{Var}_\beta^r(r_{a\to b}), & \text{if turning right, } n=2,\\ \mathrm{Var}_\beta^s(r_{a\to b}), & \text{if going straight, } n=3. \end{cases} \tag{9}$$

For the safety cost related to AV adoption, the performance of AVs and HVs at intersections differs significantly, and the safety risk changes with the penetration rate of AVs. Additionally, it has been demonstrated that the impact of AV adoption rates on enhancing system efficiency and reducing risk costs is not always linear or straightforward; instead, it presents relatively complex trends and unclear. therefore, we assume the relationship between the mean of the safety cost related to AV adoption $\mathbb{E}_\beta(r_{a\to b})$ is shown in Equation. 10, 11, and Fig.2b.

$$\mathbb{E}_\beta(r_{a\to b}) = a\left(\frac{x_{AV}}{x_{AV}+x_{HV}} - b\right)^2 + c \tag{10}$$

$$\mathrm{Var}_\beta(r_{a\to b}) = a'\left(\frac{x_{AV}}{x_{AV}+x_{HV}} - b'\right)^2 + c' \tag{11}$$

Furthermore, it satisfies:

$$\mathbb{E}_\beta^{\mathrm{Mix}}(r_{a\to b}) > \mathbb{E}_\beta^{\mathrm{HV}}(r_{a\to b}) > \mathbb{E}_\beta^{\mathrm{AV}}(r_{a\to b}) \tag{12}$$

The variance is determined using second-order moments. Based on this, the CRC at intersections is assumed to follow a specific distribution.

$$\mathbb{E}(r_{a\to b}) = \mathbb{E}_\alpha(r_{a\to b})\cdot\mathbb{E}_\beta(r_{a\to b}), \tag{13}$$

$$\mathrm{Var}(r_{a\to b}) = \mathbb{E}_\alpha(r_{a\to b}^2)\cdot\mathbb{E}_\beta(r_{a\to b}^2) - \left(\mathbb{E}_\alpha(r_{a\to b})\cdot\mathbb{E}_\beta(r_{a\to b})\right)^2 \tag{14}$$

The calculation of second-order moments is :

$$\mathbb{E}_\alpha(r_{a\to b}^2) = \mathrm{Var}_\alpha(r_{a\to b}) + \left(\mathbb{E}_\alpha(r_{a\to b})\right)^2, \tag{15}$$

$$\mathbb{E}_\beta(r_{a\to b}^2) = \mathrm{Var}_\beta(r_{a\to b}) + \left(\mathbb{E}_\beta(r_{a\to b})\right)^2. \tag{16}$$

*2) Effective Crash Risk Cost:* The cost threshold at a given confidence level $\rho$, ensuring a safe budget for travel, provides a measure of route safety reliability. It helps travelers choose routes that meet their safety requirements.

$$\begin{aligned} R_p(\rho) &= \min\{R \mid \Pr(r_p \le R) \ge \rho\} \\ &= \mathbb{E}(r_p) + \lambda_p(\rho) \end{aligned} \tag{17}$$

where $R_p(\rho)$ is the required Effective Crash Risk Cost (ECRC) of route $p$ at confidence level $\rho$, and $\lambda_p(\rho)$ is the added CRC margin of route $p$ at confidence level $\rho$, which can be seen as the safety premium paid to ensure a safe trip.

*3) Mean-excess Crash Risk Cost:* The conditional expected value of CRC exceeding the ECRC comprehensively considers both safety reliability and unreliability. It reflects the uncertainty in route safety more comprehensively, especially in extreme high-risk scenarios, and avoids underestimating risk by relying solely on the ECRC.

$$\begin{aligned} \overline{R_p(\rho)} &= \mathbb{E}(r_p \mid r_p \ge R_p(\rho)) \\ &= R_p(\rho) + \mathbb{E}(r_p - R_p(\rho) \mid r_p > R_p(\rho)) \\ &= \mathbb{E}(r_p) + z_q \cdot \sqrt{\mathrm{Var}(r_p)} + \sqrt{\mathrm{Var}(r_p)} \cdot \frac{\phi(z_q)}{1-\Phi(z_q)} \end{aligned} \tag{18}$$

where $\overline{R_p(\rho)}$ is the Mean-excess Crash Risk Cost (MCRC) of route $p$ at confidence level $\rho$, and $\mathbb{E}(r_p - R_p(\rho) \mid r_p > R_p(\rho))$ is the conditional expectation of the excess part.

### C. Traffic assignment models

*1) Generalized Travel Cost:* Generalized Travel Cost includes travel time cost, calculated using the BPR function, and safety cost:

$$c_p^d = t_p^d + \theta^d \cdot \overline{R_p(\rho)}, \quad \forall a \in A,\ \forall d \in D \tag{19}$$

$$t_p^d = \sum_{a \in A} t_a \delta_a^p \tag{20}$$

$$t_a = t_0 \left[ 1 + \alpha \left( \frac{x_{AV} + x_{HV}}{c_a} \right)^{\beta} \right] \tag{21}$$

where $t_a^d$ is the travel time cost, $\overline{R_p}$ is the MCRC for path $p$, and $\theta^d$ is the conversion factor between route travel safety cost and time cost for class $d$ travelers.

*2) Travel Behavior Scenario:* Traditional traffic assignment assumes homogeneity among users in the network. Hence, classical traffic assignment models typically consider only Wardrop User Equilibrium (UE) and System Optimality (SO) scenarios. However, in real-world networks, travel behavior varies among different user groups. For autonomous vehicles (AVs), their travel behavior can be modeled either as individual vehicles or as a coordinated fleet. In mixed traffic flows consisting of human-driven vehicles (HVs) and AVs, treating AVs as a fleet introduces distinct mixed travel behavior patterns.

In this study, HVs are assumed to follow the logit-based Stochastic User Equilibrium (SUE) principle. For AVs, three travel behavior scenarios are considered: noitemsep, topsep=0pt

- **Scenario 1 ($h = 1$):** Selfish user behavior, where AVs follow the User Equilibrium (UE) principle, minimizing the travel time for individual vehicles.
- **Scenario 2 ($h = 2$):** Private monopolist behavior, where AVs minimize the total travel time of the autonomous vehicle fleet.
- **Scenario 3 ($h = 3$):** Social planner behavior, where AVs minimize the total travel time of both human-driven and autonomous vehicle fleets.

Based on the above formulations, the variational inequality (VI) for the mixed traffic equilibrium model under the three scenarios ($h = 1$, $h = 2$, $h = 3$) is expressed as:

$$\begin{aligned} &\sum_{w \in \mathcal{W}} \sum_{p \in P^w} \tilde{C}^w_{HV,p}(\mathbf{f}^*_{HV}, \mathbf{f}^*_{AV}) \cdot (f^w_{HV,p} - f^{w*}_{HV,p}) \\ &+ \sum_{w \in \mathcal{W}} \sum_{p \in P^w} \tilde{C}^{h,w}_{AV,p}(\mathbf{f}^*_{HV}, \mathbf{f}^*_{AV}) \cdot (f^w_{AV,p} - f^{w*}_{AV,p}) \ge 0 \end{aligned} \tag{22}$$

subject to:

$$\text{s.t.} \quad \sum_{p \in P^w} f^w_{HV,p} = q^w_{HV},\ \forall w \in \mathcal{W}, \tag{23}$$

$$\sum_{p \in P^w} f^w_{AV,p} = q^w_{AV},\ \forall w \in \mathcal{W}, \tag{24}$$

$$f^w_{HV,p} \ge 0,\ \forall p \in P^w,\ w \in \mathcal{W}, \tag{25}$$

$$f^w_{AV,p} \ge 0,\ \forall p \in P^w,\ w \in \mathcal{W}. \tag{26}$$

According to the different travel objectives, the route costs for HVs and AVs under various scenarios are defined as follows:

$$\begin{aligned} &\tilde{C}^w_{HV,p}(\mathbf{f}^*_{HV}, \mathbf{f}^*_{AV}) \\ &= \sum_{a \in A} c_a \delta^w_{a,p} + \sum_{a \to b \in A^-} c_{a \to b} \delta^w_{a \to b,p} + \frac{1}{\theta} \ln \frac{f^w_{HV,p}}{q^w_{HV}} \end{aligned} \tag{27}$$

$$\begin{aligned} &\tilde{C}^w_{AV,p}(\mathbf{f}^*_{HV}, \mathbf{f}^*_{AV}) \\ &= \begin{cases} \sum_{a \in A} c_a \delta^w_{a,p} + \sum_{a \to b \in A^-} c_{a \to b} \delta^w_{a \to b,p}, & h = 1, \\ \sum_{a \in A} c_a \delta^w_{a,p} + \sum_{a \to b \in A^-} c_{a \to b} \delta^w_{a \to b,p} \\ \quad + \frac{\partial c_a}{\partial x_{AV,a}} \cdot x^*_{AV,a} \delta^w_{a,p} + \frac{\partial c_{a \to b}}{\partial x_{AV,a \to b}} \cdot x^*_{AV,a \to b} \delta^w_{a \to b,p}, & h = 2, \\ \sum_{a \in A} c_a \delta^w_{a,p} + \sum_{a \to b \in A^-} c_{a \to b} \delta^w_{a \to b,p} \\ \quad + \frac{\partial c_a}{\partial x_{AV,a}} \cdot (x^*_{AV,a} + x^*_{HV,a}) \delta^w_{a,p} \\ \quad + \frac{\partial c_{a \to b}}{\partial x_{AV,a \to b}} \cdot (x^*_{AV,a \to b} + x^*_{HV,a \to b}) \delta^w_{a \to b,p}, & h = 3. \end{cases} \end{aligned} \tag{28}$$

where $\delta^w_{a,p}$ and $\delta^w_{a \to b,p}$ are indicator variables that denote whether link $a$ or turning link $a \to b$ is part of route $p$ for OD pair $w$.

## III. Contributions

The key contributions of this paper are:

(i) Developing a comprehensive mixed traffic equilibrium model that explicitly incorporates the behavior and interactions between HVs and AVs, considering both efficiency and safety at road links and intersections.

(ii) Proposing a dynamic AV route adaptation mechanism that accounts for varying levels of AV market penetration and their influence on traffic patterns.

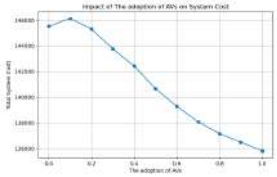
(iii) Validating the proposed models through numerical experiments on realistic network scenarios, including the Nguyen-Dupuis network network, to assess the impact of AV adoption on system-wide performance.

## IV. Computational Results

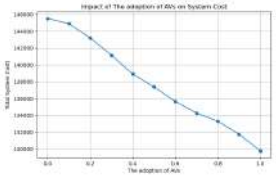
In Fig. 3a, which depicts Selfish User Equilibrium (UE) behavior, the total system cost initially hovers around 146,000 at zero AV penetration, exhibits a slight uptick at low levels of AV adoption, and then steadily declines to approximately 135,000 as AV penetration reaches 100%. This transient increase at low AV penetration rates may reflect temporary inefficiencies associated with mixed fleets under purely self-interested routing decisions.

By contrast, Fig. 3b, representing Private Monopolist behavior (fleet-based SO), shows a more consistent downward trend in system cost as the AV share increases, though the total cost remains higher than in the fully optimized scenario. Here, a single private operator coordinates fleet-based decisions, reducing some inefficiencies compared to purely selfish routing, yet still prioritizing the operator's objectives over global cost minimization.

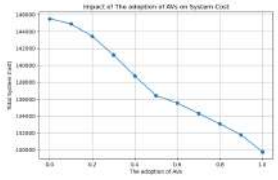
Finally, Fig. 3c, illustrating Social Planner (SO) behavior, demonstrates the most pronounced cost reduction: system cost falls from around 146,000 at zero AV adoption to approximately 130,000 at full penetration. This improvement reflects the planner's ability to assign routes to minimize overall network cost, rather than focusing on individual or monopolistic objectives.



(a) Selfish user behavior



(b) Private monopolist behavio



(c) Social planner behavior

Fig. 3. Comparison of three traffic scenarios: $h = 1$, $h = 2$, and $h = 3$.

## References


[1] S. Bahrami and M. J. Roorda, "Optimal traffic management policies for mixed human and automated traffic flows," *Transportation Research Part A: Policy and Practice*, vol. 135, pp. 130–143, 2020.

[2] NHTSA, "National motor vehicle crash causation survey: Report to congress," National Highway Traffic Safety Administration (NHTSA), Technical Report, 2008.

[3] Z. Wadud, "Fully automated vehicles: A cost of ownership analysis to inform early adoption," *Transportation Research Part A: Policy and Practice*, vol. 101, pp. 163–176, 2017.

[4] Y. Li, Z. Chen, Y. Yin, and S. Peeta, "Deployment of roadside units to overcome connectivity gap in transportation networks with mixed traffic," *Transportation Research Part C: Emerging Technologies*, vol. 111, pp. 496–512, 2020.

[5] R. Arvin, A. J. Khattak, M. Kamrani, and J. Rio-Torres, "Safety evaluation of connected and automated vehicles in mixed traffic with conventional vehicles at intersections," *Journal of Intelligent Transportation Systems*, vol. 25, no. 2, pp. 170–187, 2020.

[6] J. N. Prashker and S. Bekhor, "Route choice models used in the stochastic user equilibrium problem: a review," *Transport reviews*, vol. 24, no. 4, pp. 437–463, 2004.

[7] S. Gong and L. Du, "Cooperative platoon control for a mixed traffic flow including human drive vehicles and connected and autonomous vehicles," *Transportation research part B: methodological*, vol. 116, pp. 25–61, 2018.

[8] H. Yang and H.-J. Huang, "The multi-class, multi-criteria traffic network equilibrium and systems optimum problem," *Transportation Research Part B: Methodological*, vol. 38, no. 1, pp. 1–15, 2004.

[9] X. Xu, A. Chen, L. Cheng, and H. K. Lo, "Modeling distribution tail in network performance assessment: A mean-excess total travel time risk measure and analytical estimation method," *Transportation Research Part B: Methodological*, vol. 66, pp. 32–49, 2014.

[10] Y. M. Nie and Q. Li, "An eco-routing model considering microscopic vehicle operating conditions," *Transportation Research Part B: Methodological*, vol. 55, pp. 154–170, 2013.

[11] H. A. Aziz and S. V. Ukkusuri, "Integration of environmental objectives in a system optimal dynamic traffic assignment model," *Computer-Aided Civil and Infrastructure Engineering*, vol. 27, no. 7, pp. 494–511, 2012.

[12] C. Han, G. Xu, A. Pervez, F. Gao, H. Huang, X. Pei, and Y. Zhang, "Modeling traveler's speed-route joint choice behavior with heterogeneous safety concern," *Analytic methods in accident research*, vol. 37, p. 100253, 2023.

[13] Q. Guo, X. J. Ban, and H. A. Aziz, "Mixed traffic flow of human driven vehicles and automated vehicles on dynamic transportation networks," *Transportation research part C: emerging technologies*, vol. 128, p. 103159, 2021.

[14] J. Wang, S. Peeta, and X. He, "Multiclass traffic assignment model for mixed traffic flow of human-driven vehicles and connected and autonomous vehicles," *Transportation Research Part B: Methodological*, vol. 126, pp. 139–168, 2019.

[15] T. Xie and Y. Liu, "Impact of connected and autonomous vehicle technology on market penetration and route choices," *Transportation Research Part C: Emerging Technologies*, vol. 139, p. 103646, 2022.

[16] C. Hu, C. Han, A. Pervez, J. Hao, G. Xu, J. Tang, and H. Huang, "Optimal deployment of connected and autonomous vehicle dedicated lanes: A trade-off between safety and efficiency," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[17] A. Papadoulis, M. Quddus, and M. Imprialou, "Evaluating the safety impact of connected and autonomous vehicles on motorways," *Accident Analysis & Prevention*, vol. 124, pp. 12–22, 2019.

[18] S. Aoki and R. Rajkumar, "Safe intersection management with cooperative perception for mixed traffic of human-driven and autonomous vehicles," *IEEE Open Journal of Vehicular Technology*, vol. 3, pp. 251–265, 2022.

[19] H. Huang, C. Han, G. Xu, M. Jiang, S. Wong, and M. M. Haque, "Incorporating safety reliability into route choice model: Heterogeneous crash risk aversions," *Analytic methods in accident research*, vol. 25, p. 100112, 2020.

# Coexistence of human drivers and CAVs - behavioral models and long-term properties

Grzegorz Jamróz*[1], Gunnar Flötteröd[2], Rafał Kucharski[1], and David Watling[3]

[1] *Faculty of Mathematics and Computer Science, Jagiellonian University, Kraków, Poland*
[2] *Communications and Transport Systems, Linköping University, Sweden*
[3] *Institute for Transport Studies, University of Leeds, United Kingdom*
* *Corresponding author, e-mail: grzegorz.jamroz@uj.edu.pl*

February 27, 2025

## 1 Introduction

Interaction between autonomous vehicles and human drivers is a prolific research area that has many interesting aspects such as microscopic driving behaviour [6, 7]. The interaction between human drivers and coordinated fleets of autonomous vehicles in future transportation systems, however, seems to be understudied. The research related to adoption of AVs, namely, has so far focused mostly on individual factors determining CAV uptake and discussing static levels of adoption in relation to automation level (see e.g. [2, 8]) or different routing outcomes [14, 10]. The dynamic market evolution in systems with mature CAV as a service offer where drivers are free to switch between driving/routing independently and subscribing to a coordinated fleet of autonomously driving and routing vehicles remain unexplored.

Importantly, the dynamics of human-only (no CAVs) systems can already exhibit complex long run phenomena such as unstable or multiple attractors with research in this direction ranging from consideration of simple differential dynamical systems to full-scale agent-based or aggregated stochastic processes, e.g. [11, 9, 13, 5, 12]. Adding autonomous vehicles into such a system adds a further layer of complexity: the autonomous agents may behave selfishly, cooperatively or semi-cooperatively; they may interact differently with human vehicles than they do with other autonomous vehicles; and they may be 'tuned' to influence the system evolution in different ways, whether for the benefit or not of the different groups of agents.

Studying complex interactions such as HDV-CAV dynamics, one usually accounts for certain aspects while aggregating the others. In [4], e.g., the authors investigated microscopic driving conflict interactions and made long-term predictions based on the evolution of the system using (evolutionary) game theory. Nevertheless, a comprehensive theory of adoption of CAVs allowing for forecasting long term properties in mature markets with CAVs treated as equal road-users has been lacking. In this contribution we fill this gap, proposing dynamic behavioral models of CAV uptake derived from factors such as perceived efficiency of routing and studying their properties such as equilibria/stable states, long-term convergence etc. We use both theoretical methods based on dynamical systems, game theory, statistics etc. and experimental methods using agent-based simulations as well as fleet strategy execution based on optimization or reinforcement learning. We propose notions of equilibria in this bi-level (route choice and fleet-HDV switching) setting and discuss their existence, uniqueness and stability in aggregated as well as disaggregatged individual-based settings. An exemplary conclusion is that anti-social routing strategies adopted by fleet operators may turn out to be the most evolutionary successful, leading to maximization of long-term market share.

## 2 Setup

We consider traffic networks (ranging from a single OD pair to complex topologies) which are initially populated by human drivers only. Considering day-to-day route choice dynamics, we identify stable states

(theoretically or by means of agent-based simulations) and then let an initial share of CAVs replace human drivers. From this moment on, each driver, whether in a coordinated fleet (CAV) or routing independently (HDV), has the freedom to switch from HDV to CAV or revert to being an independent driver. Simultaneously, HDVs choose routes every day, while CAVs follow the routes assigned by the fleet operator. We let the system evolve (e.g. for 200 days in simulations), studying it convergence to stable states.

Heterogenous human drivers choose routes according to standard behavioural models based on expected utility (perceived travel time). CAV fleet operators adopt certain behavioural strategies such as (but not limited to) selfish - optimizing collective fleet travel time, social - optimizing the total travel time in the system, malicious - aiming to *maximize* the travel time of human drivers, disruptive - aiming to maximize human drivers' travel time at bounded own cost and altruistic - aiming to minimize human drivers' travel time (compare [3] for adversarial network response scenario).

In this dynamic setting drivers not only learn and make everyday route choices (compare [10, 1] for previous research) but also consider switching between HDV and CAV, the latter occuring potentially on a different time scale and/or incurring some cost or being harder to execute, depending on the scenario (see Results). We propose models of this switching behaviour and study the equilibrium and long-term properties of this dynamic system.

## 3 Results

**Sample switching model and simulation results**

Here we present the simulation results of a sample model in a simplified static setting, see [10], leaving the full theoretical and simulation discussion of more detailed models to the full paper. In the considered model there is one OD pair with two available routes A and B and travel times expressed by BPR functions $t_A = 5min * \left(1 + (\frac{q_A}{500})^2\right)$ and $t_B = 15min * \left(1 + (\frac{q_B}{800})^2\right)$, where $q_A, q_B$ are the number of vehicles choosing $A$ and $B$ on a given day and the total demand equals 1000. Heterogenous HDVs make decisions based on learning from experienced travel times and maximization of utility. In contrast, being part of CAV fleet, the vehicles choose the route prescribed by the fleet operator which minimizes a collective objective function expressed as a combination of collective fleet travel times and HDV travel times:

$$\Phi = \lambda^{CAV} t^{CAV} + \lambda^{HDV} t^{HDV},$$

where $\lambda^{CAV}, \lambda^{HDV}$ prescribe the collective goal. E.g. collective selfish routing is obtained for $\lambda^{CAV} = 1$ and $\lambda^{HDV} = 0$ and malicious collective CAV routing occurs for $\lambda^{CAV} = 0$ and $\lambda^{HDV} = -1$.

The sample model of HDV↔CAV switching assumes that the expected disutilities (for every agent $i$) of using an HDV or CAV are expressed by formulas:

$$\begin{aligned} D^{HDV} &= \min_{r\in\{A,B\}} -U_r(i) = \min_{r\in\{A,B\}} (T_r(i) + \epsilon_r(i)) \\ D^{CAV} &= \kappa \frac{q_A^{CAV} * (-U_A(i)) + q_B^{CAV} * (-U_B(i))}{q_A^{CAV} + q_B^{CAV}} \end{aligned}$$

where $U_r(i) = -(T_r(i) + \epsilon_r(i))$ is the utility of route $r$ to driver $i$, where $T_r(i)$ is the expected (obtained by learning from experience) travel time on route $r$ and $\epsilon_r(i)$ accounts for taste preferences between the routes based on other factors (different for different drivers but fixed in time). $D^{CAV}$ is the expected disutility (i.e. expected perceived travel time) of agent $i$ belonging to the fleet, scaled by a discount factor $\kappa$. Typically, $\kappa \leq 1$, which expresses the gain in perceived travel time due to factors such as no need to drive etc., however we consider also $\kappa > 1$, which could be relevant for agents preferring driving and routing independently. $q_A^{CAV}, q_B^{CAV}$ are the number of agents routed via $A$ and $B$, respectively on a given day. We assume that switching $HDV \leftrightarrow CAV$ incurs some cost and so an HDV (CAV) considers switching on a given day only if $D^{HDV} - D^{CAV} > \theta$ $(< -\theta)$, for some threshold $\theta$, which we assume to be equal $1\,min$. The probability of switching is given by

$$p^{switch} = \frac{2}{\pi} \arctan(|D^{HDV} - D^{CAV}|/(10 * \theta)).$$

If, for instance, $D^{HDV} = D^{CAV} + 10\ min$ an HDV's probability of switching would be $p^{switch} = 0.5$ and a driver using a CAV would not consider switching at all.

This simplified model has non-trivial consequences. Fig. 1 shows that the antisocial malicious strategy of fleet operators and its soft version 'Disruptive' seem to be the most successful evolutionarily which is an important hint for policy-makers.

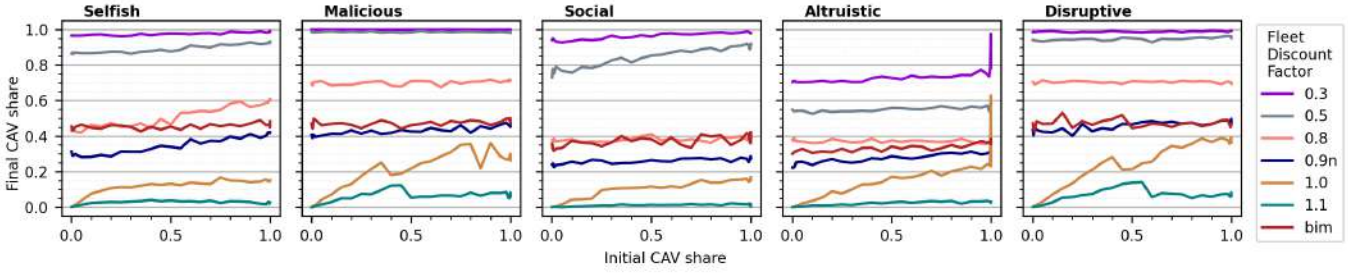



Figure 1: Average final CAV share as a function of initial CAV share (on the day of fleet introduction) for various fleet discount factors and fleet behaviours. Low discount factors imply, not surprisingly, high eventual CAV uptake. Discount factor of 1.0 (no discount) results in limited eventual uptake of $10 - 20\%$ for most initial CAV shares. The dependence on Initial CAV share is rather weak. $0.9n$ stands for discount factors sampled independently for each agent from the normal distribution $N(0.9, (0.2)^2)$ while for 'bim' from the bimodal distribution $0.5N(0.5, (0.1)^2) + 0.5N(1.5, (0.1)^2)$, with two peaks at 0.5 and 1.5. Malicious and disruptive strategies are most successful across a range of parameters yielding highest final CAV shares.

**Fleet-human dynamic equilibrium and strong equilibrium**

At the theoretical level we propose several notions of equilibria such as dynamic fleet-human equilibrium and strong fleet-human equilibirum. As the dynamics occur at two interdependent levels (route choice and switching HDV $\leftrightarrow$ CAV), with different equilibration (if any) rates, we can consider various equilibria. We introduce two of them below. The first one encompasses only equilibration of switching dynamics with route-choice dynamics possibly still taking place and can be conceived as a sort of self-confirming equilibrium.

**Definition 1.** *A mixed traffic system is in* **dynamic fleet-human equilibrium** *if no driver has an incentive to switch from HDV to CAV or vice versa.*

Note that this definition does not imply that every driver's route choice pattern is constant (i.e. selecting the same route every day or having constant probabilities of choosing different alternatives). In fact, the route choice probabilities of every individual agent can vary wildly as long as they are guaranteed to never make this agent switch to CAV if it is HDV or vice versa. For instance, a system composed initially (after introduction of CAVs) of 20% CAV who, no matter what, never want to revert to HDV and 80% HDVs who are completely opposed to switching to CAV is in dynamic fleet-human equilibrium even though the dynamics at the level of route choice may be ongoing and be perhaps very complex resulting in e.g. oscillations or final equilibration. A stronger notion is the two-tiered nested equilibrium introduced below.

**Definition 2.** *A mixed traffic system is in* **strong fleet-human equilibrium** *if no driver has an incentive to switch from HDV to CAV or vice versa and the system composed of atomic individual (or infinitesimal) human driver players and a group fleet player is in Nash (Wardrop) or similar ($\epsilon$-Nash etc.) equilibrium at the level of route choice.*

The equilibration at the route choice level only with some agents still considering switching (and perhaps taking many days before they switch) precludes a strong fleet-human equilibrium. On the other hand, e.g. a system in (stochastic) user equilibrium with human drivers only and no option to switch to CAV is in strong equilibrium. Finally, the system considered above in dynamic fleet-human equilibrium may not be in strong fleet-human equilibrium. However, it may eventually converge to a strong fleet-human equilibrium once the route choices equilibrate.

Let us stress that the precise definitions of the fleet-human equilibria depend on the level of aggregation of drivers we are considering. The detailed properties of these and related equilibria are discussed in detail in the full version of the paper. Here we only note that the properties of the equilibria introduced above heavily depend on human learning and route-choice model and fleet strategy.

## 4 Discussion

Future composition of urban traffic systems and interactions in mixed autonomous vehicle - human driver context remain largely unexplored. This paper contributes to shedding some light on possible evolution of dynamic CAV uptake based on studying certain factors such as travel time-based utilities and aggregating other unaccounted for factors. Even though further research seems necessary to make the predictions even more reliable, this paper contributes to understanding the collective routing phenomena and urges caution in deploying CAV fleets which might e.g. obtain unfair market advantage by applying antisocial routing policies.

## 5 Acknowledgement

This research was supported by the ERC Starting Grant number 101075838: COeXISTENCE.

## References

[1] Akman, A. O., Psarou, A., Varga, Z. G., Jamróz, G., and Kucharski, R. (2024). Impact of Collective Behaviors of Autonomous Vehicles on Urban Traffic Dynamics: A Multi-Agent Reinforcement Learning Approach. EWRL 17.

[2] Ardeshiri, A., & Vij, A. (2022). Consumer preferences for different CAV technologies and service models. Available at SSRN 4162473.

[3] Bell, M. G. (2000). A game theory approach to measuring the performance reliability of transport networks. Transportation Research Part B: Methodological, 34(6), 533-545.

[4] Bitar, I., Watling, D., and Romano, R. (2023). Sensitivity analysis of the spatial parameters in modelling the evolutionary interaction between autonomous vehicles and other road users. SN Computer Science, 4(4), 336.

[5] Cantarella, G. E., and Cascetta, E. (1995). Dynamic processes and equilibrium in transportation networks: towards a unifying theory. Transportation Science, 29(4), 305-329.

[6] Farah, H., Postigo, I., Reddy, N., Dong, Y., Rydergren, C., Raju, N., and Olstam, J. (2022). Modeling automated driving in microscopic traffic simulations for traffic performance evaluations: aspects to consider and state of the practice. IEEE Transactions on Intelligent Transportation Systems, 24(6), 6558-6574.

[7] Gora, P., Katrakazas, C., Drabicki, A., Islam, F., and Ostaszewski, P. (2020). Microscopic traffic simulation models for connected and automated vehicles (CAVs)–state-of-the-art. Procedia Computer Science, 170, 474-481.

[8] Harrison, G., Shepherd, S. P., and Chen, H. (2021). Modelling uptake sensitivities of connected and automated vehicle technologies. International Journal of System Dynamics Applications (IJSDA), 10(2), 88-106.

[9] Horowitz, J. L. (1984). The stability of stochastic equilibrium in a two-link transportation network. Transportation Research Part B: Methodological, 18(1), 13-28.

[10] Jamróz, G., Akman, A. O., Psarou, A., Varga, Z. G., and Kucharski, R. (2025). Social implications of coexistence of CAVs and human drivers in the context of route choice. Sci Rep 15, 6768. https://doi.org/10.1038/s41598-025-90783-w

[11] Smith, M. J. (1984). The stability of a dynamic model of traffic assignment—an application of a method of Lyapunov. Transportation science, 18(3), 245-252.

[12] Smith, M., Hazelton, M. L., Lo, H. K., Cantarella, G. E., and Watling, D. P. (2014). The long term behaviour of day-to-day traffic assignment models. Transportmetrica A: Transport Science, 10(7), 647-660.

[13] Smith, M. J., and Watling, D. P. (2016). A route-swapping dynamical system and Lyapunov function for stochastic user equilibrium. Transportation Research Part B: Methodological, 85, 132-141.

[14] Wang, J., Peeta, S., and He, X. (2019). Multiclass traffic assignment model for mixed traffic flow of human-driven vehicles and connected and autonomous vehicles. Transportation Research Part B: Methodological, 126, 139-168.

# DTA 2025 Abstract.
# New stable responsive local gating strategies to control vehicle queues and flows in congested urban networks.

Michael J Smith[a], David Watling[b], Ronghui Liu[b], Koki Satsukawa[c], Takamasa Iryo[d], and Richard Mounce[e].
[a]Department of Mathematics, University of York, United Kingdom of Great Britain and Northern Ireland;
[b]Institute for Transport Studies, University of Leeds, 10, Woodhouse Lane, Leeds LS2 9JT, United Kingdom of Great Britain and Northern Ireland.
[c]Institute of Transdisciplinary Sciences for Innovation, Kanazawa University, Japan;
[d]Graduate School of Information Sciences, Tohoku University, Japan;
[e]RBM Traffic Solutions Ltd, 95 Princess Drive, York, YO26 5SX, UK.
Keywords: Signal Control Gating, Responsive traffic signal control, Proportional Adjustment Process, Stability.

## 1. Introduction

### 1.1. The impacts of motor traffic in towns

The distribution of motor vehicle traffic flows and queues both have a very important impact on the effective use of urban road-space. For example:

(i) active travel is much more likely to thrive where motor vehicle flows and queues are small,
(ii) public transport is likely to run more efficiently where motor vehicle flows and queues are small,
(iii) enjoying the view of inspiring architecture is often much enhanced by reductions in the visual intrusion of motor vehicles, and
(iv) the use of streets for play and other social interactions becomes possible when traffic flows and queues are small or very small.

Traffic signals may be regarded as taps controlling both flows and queues. It follows that the interaction between traffic signal timings and motor traffic flows and queues on a road network is key to many important aspects of urban traffic. For example, if specific targets for flows and queues are agreed then signal timings may be used to help achieve these specified flow and queue targets.

Since traffic signals may be regarded as taps, there is the option of reducing flow out of a link by reducing the green-time at the link exit. This will usually reduce flows and queues downstream and this is the standard way of looking at gating: as an upstream control which affects downstream conditions. There are *very many* papers on gating (see Quinn (1992) for a substantial informative review). Most focus on upstream gating.

Our gating proposal here looks instead at the upstream effects. These upstream effects arise because reduced green-times at the link *i* exit increase delays and divert traffic flows away from link *i* and, usually, at least some upstream links. This gating may be thought of as "downstream gating", affecting upstream queues and flows.

Of course any gating action will in practice probably have both upstream and downstream effects.

The interaction between *adaptive* or *responsive* traffic signal control systems and the distribution of flows and queues on a network is very important; yet this interaction has received little substantial study. This paper considers elements of this interaction. The paper proposes a new sensitive distributed gating strategy for carefully and efficiently reducing flows and queues at specific locations, both upstream and downstream. There is substantial opportunity for development of the basic gating strategy introduced in this paper.

### 1.2. The interaction between traffic signal control and flows and queues of motor traffic

Allsop (1974) and Beckmann (1956) suggested that traffic signal-settings should take account of drivers' route choices; so as to beneficially affect the distribution of motor vehicle traffic flows on an urban road network. This paper will be concerned with responsive or adaptive signal controls where the green-times are responsive to traffic flows and queues. The paper addresses the question:

How should green-times respond to flows and queues?

The paper makes several suggestions in response to this question.

The responsive control policies developed in this paper are carefully designed to operate *a sensitive and responsive local gating strategy* (holding some traffic back) aiming to reduce queues in selected locations and overall. The holding back responsive control policies considered are all, under certain conditions, still capacity-maximising; because they build directly on the $P_0$ control policy which is capacity-maximising.

#### 1.2.1. Studies of the interaction between traffic signal control and flows and queues

Smith (1979a, b, c)

(i) proposed the local $P_0$ traffic signal control policy and

(ii) proved that, under certain conditions, this policy maximises overall network travel capacity.

The new control strategies are variations on this basic $P_0$ control policy, building on control policies $P_h$ introduced in Smith et al (2019a). These variations of $P_0$ are designed to allow more inputs from policy-makers, who may wish to reduce flows and queues in certain areas or on certain important road links.

Bentley and Lambe (1980) created a model embracing traffic assignment (and so vehicle routeing and vehicle flows) and traffic control. Smith and van Vuren (1993) compare IOA (iterative optimisation assignment) using the $P_0$ control policy and the equisaturation policy.

The joint dynamics of day-to-day route choice and adaptive signal control under real-time information has been considered by Hu and Mahmassani (1997). Mounce (2009) proves the existence of equilibrium in a continuous dynamic queueing model with responsive signal control. Cantarella (2010) presented a formal definition of the combined day-to-day signal control and traffic assignment problem based on a discrete time, deterministic process model. He proved fixed-point stability results. Cantarella et al. (2012) showed how equilibrium stability conditions can be embedded as a constraint in a day-to-day signal setting–route choice framework.

Varaiya (2013) presents a different view of responsive traffic control, assuming that route choices are fixed; Varaiya suggests a control called MaxPressure.

Meneguzzer (1996, 1997, 2012) gives an interesting review and shows that the frequency of signal updating may significantly affect the duration of the day-to-day dynamic process needed to achieve a network flow–control equilibrium. Yang and Yagar (1995) study assignment and control on saturated networks. Bie and Lo (2010) study the stability and attraction domains of traffic equilibria in a day-to day dynamical system. Xiao and Lo (2015) investigated the behavior of a joint route choice–signal control dynamical system.

More recently, He et al (2022) present a discrete day-to-day signal retiming problem for fine-tuning the green splits in a single-destination traffic network to mitigate the congestion induced by travelers' day-to-day adaptation to a new signal plan. Numerical examples demonstrate that the proposed signal retiming scheme can reduce the total system travel time over the traffic equilibration period.

Meneguzzer (2024) considers adaptive traffic signal control to promote the efficient use of road intersections, but comments that "the reaction of drivers to repeatedly updated signal settings and the

ensuing route choice dynamics may trigger the emergence of various kinds of network instability". Meneguzzer suggests a Logit form signal control policy (somewhat similar to a policy suggested by Le et al (2015)) to better protect the system from instabilities which may arise with other responsive traffic signal control strategies.

Satsukawa et al (submitted) consider a static traffic assignment problem with the $P_0$ control policy in a simple signal-controlled network. The equilibrium states consistent with $P_0$ are calculated for changing demand levels, from unsaturated to near saturated; to investigate whether uniqueness holds for the different traffic demand levels. Stability of the equilibria are studied using a graphical approach. The paper demonstrates (i) the occurrence of multiple equilibria with different total travel times at certain moderate traffic demand levels and (ii) the existence of a hysteresis loop, where different stable equilibria with different total travel times emerge when traffic demand increases and then decreases.

### 1.3. New sensitive distributed gating strategies to reduce flows and queues in sensitive locations

This paper states new sensitive distributed [green-split plus gating] strategies which aim to reduce flows and queues in urban road networks, especially in sensitive locations.

Gating strategies are usually designed to reduce queues and flows downstream; the gating is upstream of the beneficial effects. This paper shows how gating strategies may also be used to reduce queues and flows upstream.

Often traffic control systems are judged by measures such as "total vehicle travel time" or "total vehicle delay time" or "total number of vehicle stops" or "total person travel time" or "total person delay time" or "total network vehicle travel capacity" or "total network person travel capacity". Then these measures may be used to find control systems which reduce or even minimise "total vehicle travel time" etc.

*These standard measures do not consider especially bad impacts in especially important places.* So it is natural to consider other measures such as "vehicle queues on link *i*" or "the number of vehicles on link *i*" or "vehicle flows on link *i*" or "vehicle stops on link *i*". These may be considered for all links *i* separately or added together over all links *i* in (say) the city centre or in a residential district. These link *i* measures may also be multiplied by (say) estimates of the number of pedestrians on the link. These measures may be used to find control systems which reduce pollution or accident-risk to pedestrians; often by encourage motor traffic to switch to longer but less pedestrian-populated routes.

This paper proposes new responsive control strategies which maximise network capacity *and also* reduce "vehicle queues on link *i*" or "the number of vehicles on link *i*" or "vehicle flows on link *i*" or "vehicle stops on link *i*"; or sums of such measures. The proposed new control strategies are all developments of $P_0$; these new developments add careful sensitive gating strategies to extensions of $P_0$, and the paper demonstrates that the new strategies do still maximise network capacity and reduce some of the measures listed above.

The paper also proves some stability results concerning some of these new ($P_0$-augmented) responsive traffic signal control policies and certain dynamic day-to-day models of route choice; substantially developing some of the stability ideas already published in Liu et al. (2015).

### 1.4. A simple illustrative example; showing what we mean by "gating" and how traffic queues on a network may be reduced by using gating

Figure 1 illustrates a simple network with one traffic signal at a bottleneck at node $n_2$. This may represent a village with a high pedestrian flow on link 1 and perhaps on link 4. Let
$T$ = steady flow from the origin to the destination = flow on link 1 plus flow on link 3,
$g_1$ = link *1* green-time, $0 \leq g_1 \leq 1$.
For $i = 1, 3, 4$,
$s_i$ = saturation flow of link $i$,

$c_i$ = link *i* free-flow travel time,
$b_1$ = bottleneck delay at the exit of link *1* = travel time along link *1* – link *1* free-flow travel time.
Suppose that

$$s_1 < T < s_3 \text{ and } c_1 + c_4 < c_3.$$

The equilibrium queue on link 1 causes a queueing time which at equilibrium equalises the two route travel times $c_1 + b_1 + c_4$ seconds and $c_3$ seconds.

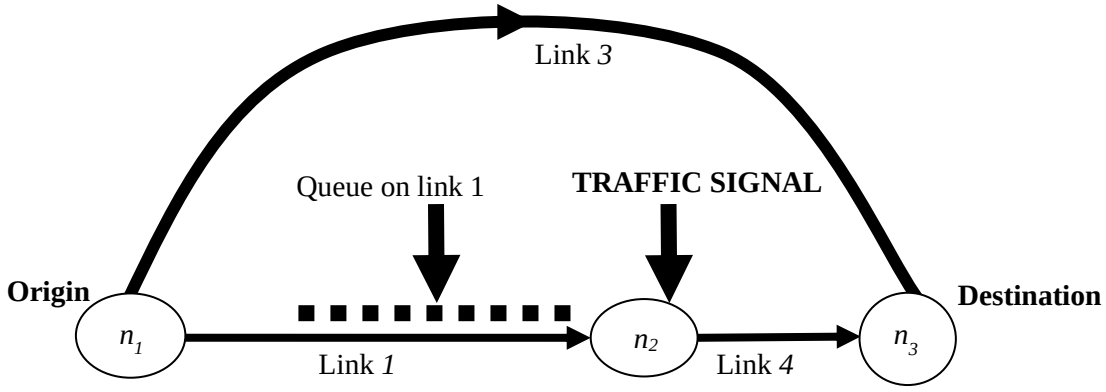



Figure 1. No gating. Link 1 green-time proportion $g_1 = 1$.

Figure 2 illustrates the same network with a link 1 green-time proportion *reduced from 1 to ½*. This is the simplest example of gating where the sum of green-times < 1 rather than = 1.

The queue on link 1 again causes a queueing time which at equilibrium equalises the travel times over the two routes. But by virtue of the gating; reducing the green-time $g_1$ from 1 to ½; this queue is now *more than halved*. See appendix A in the full paper for a detailed justification of this statement. It is anyway clear that reducing the link 1 green-time from 1 to ½ means that the link 1 queue volume needed to balance the network is significantly reduced.

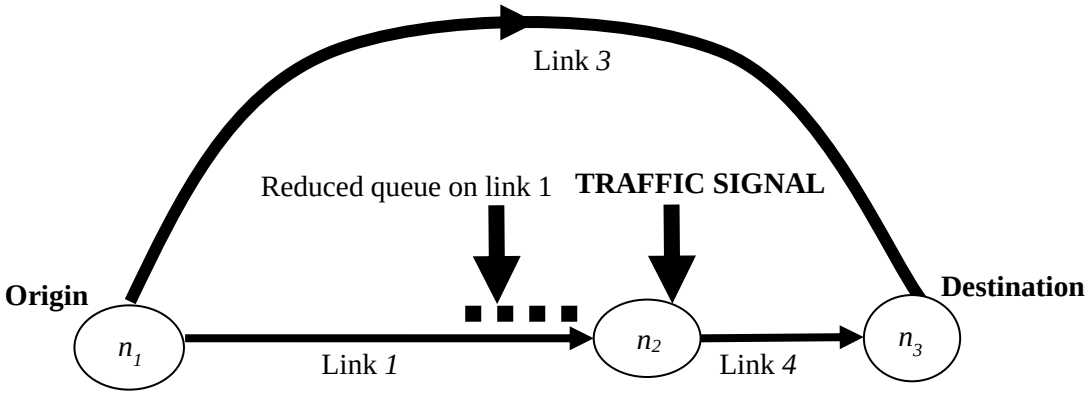



Figure 2. With gating: Link 1 green-time proportion $g_1 = ½ < 1$. Gating ensures that the queue is *more than halved*. See appendix A in the full paper for a detailed justification of this statement.

Here gating substantially reduces link 1 equilibrium queues. The gating strategy here also substantially reduces traffic flows along links 1 and 4 and substantially increases flow along link 3.

## 2. Stability of day-to-day signal-setting / traffic routeing using signal stages, anti-stages and red-times.

### 2.1. Signal stages and anti-stages.

We consider, in a day-to-day context, the dynamics of responsive signal control and routeing on a general network. For any signal stage *S* at a junction there corresponds an *anti-stage AS* at the same junction. Anti-stage *AS* is the set of links at the same junction shown red when stage *S* is shown green.

**2.2. Central variables.**

Suppose now that in a general network with $N$ routes, $M$ signal stages (and so $M$ antistages) and $n$ links:
$s_i$ = saturation flow at the link $i$ exit,
$\mathbf{X} = (X_1, X_2, \ldots, X_a, \ldots, X_N)$ is the $N$-vector of all route flows $X_1, X_2, \ldots, X_N$ ,
$\mathbf{R} = (R_1, R_2, \ldots, R_a, \ldots, R_M)$ is the $M$-vector of all anti-stage red-times $R_1, R_2, \ldots, R_M$ ,
$x_i$ = the sum of the route flows over all routes which contain link $i$ (for $i = 1, 2, 3, \ldots, n$),
$r_i$ = the sum of the anti-stage red-times over all anti-stages which contain link $i$ (for $i = 1, 2, 3, \ldots, n$), thus:

$$x_i = \sum_{\{u;\ \text{route } u \text{ contains link } i\}} X_u \text{ and } r_i = \sum_{\{u:\ \text{anti-stage } u \text{ contains link } i\}} R_u$$

(and therefore $r_i = 0$ if link $i$ is not signalised),
$\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is the $n$-vector of link flows $x_1, x_2, \ldots, x_n$ ,
$\mathbf{r} = (r_1, r_2, \ldots, r_n)$ is the $n$-vector of red-times $r_1, r_2, \ldots, r_n$ (where $r_i = 0$ if link $i$ is not signalised).
$c_i(\cdot)$ is a new non-decreasing cost function of only $x_i$ (for $i = 1, 2, 3, \ldots n$),
$b_i$ represents the vertical link $i$ delay due to exiting the link (this is the whole link travel time minus $c_i(\cdot)$),
$\mathbf{b} = (b_1, b_2, \ldots, b_n)$ is the $n$-vector of extra (beyond $c_i(\cdot)$) added bottleneck delays $b_i$ at the exit of link $i$, and
the whole link $i$ travel time = $c_i(x_i) + b_i$.

**2.3. Delay functions involving both flow and red-time; route-costs and anti-stage costs.**

In this paper we utilise link delay functions which involve link red-times as well as link flows. Given link $i$, we now let:

$f_i$ = a particular fixed <u>*but arbitrary*</u> non-decreasing non-negative function defined on at least the interval $[0, s_i) = \{x_i;\ 0 \le x_i < s_i\}$.

The functions $f_i$ will eventually be very carefully chosen by the signal controller but for now

*these $f_i$ are fixed but arbitrary non-decreasing non-negative functions*

defined on at least the set of feasible link flows = $[0, s_i) = \{x_i;\ 0 \le x_i < s_i\}$. We now make a major assumption.

***MAJOR ASSUMPTION:***
*Given the fixed non-decreasing non-negative functions $f_i$, we assume that for each link i:*

*the link i (vertical) bottleneck delay $b_i = f_i(x_i + s_i r_i)$.*

We then, for each link $i$, define

the link $i$ travel cost = the link travel time = $c_i(x_i) + b_i = c_i(x_i) + f_i(x_i + s_i r_i)$.

In order to design a suitable control policy we define the link *i red-time cost* as follows:

*the link i red-time cost = $s_i b_i = s_i f_i(x_i + s_i r_i)$.*

Bearing in mind that

$$x_i = \sum_{\{u;\ \text{link } i \text{ belongs to route } u\}} X_u \text{ and } r_i = \sum_{\{u;\ \text{link } i \text{ belongs to anti-stage } u\}} R_u,$$

we now define the $N$ route travel costs $C_u$; *which will be travel times here*; and the $M$ anti-stage red-time-costs $A_u$ as follows:

$$C_u(\mathbf{X}, \mathbf{R}) = \sum_{\{i;\ \text{link } i \text{ belongs to route } u\}} [c_i(x_i) + f_i(x_i + s_i r_i)],$$

$$A_u(\mathbf{X}, \mathbf{R}) = \sum_{\{i;\ \text{link } i \text{ belongs to anti-stage } u\}} s_i b_i = \sum_{\{i;\ \text{link } i \text{ belongs to anti--stage } u\}} s_i f_i(x_i + s_i r_i).$$

Also we put

$$\mathbf{C}(\mathbf{X}, \mathbf{R}) = [C_1(\mathbf{X}, \mathbf{R}), C_2(\mathbf{X}, \mathbf{R}), \ldots, C_N(\mathbf{X}, \mathbf{R})] \text{ and}$$
$$\mathbf{A}(\mathbf{X}, \mathbf{R}) = [A_1(\mathbf{X}, \mathbf{R}), A_2(\mathbf{X}, \mathbf{R}), \ldots, A_M(\mathbf{X}, \mathbf{R})].$$

We assume here that the $C_u(\mathbf{X}, \mathbf{R})$ are all determined by the route flow vector and the anti-stage red-time vector and so are not directly under our control; although we control $\mathbf{R}$ we do not directly control

**X**. On the other hand we will consider some simple variations on the definitions of the $A_u$**(X, R)**; this will allow a variety of new control policies to arise.

Throughout this paper the current or *today's* route flow vector **X** and cost vector **C(X, R)** determine future or *tomorrow's* route flows; these are obtained from today's route-flow vector **X** by swapping route-flow from more to less costly routes following the proportional adjustment process, or *PAP*, see Smith (1984).

Also the current or *today's* anti-stage red-time vector **R** and anti-stage cost vector **A(X, R)** determine future or *tomorrow's* anti-stage red-times which are obtained from today's red-time vector **R** by swapping red-time from more to less costly anti-stages following a method similar to *PAP*.

In this paper all route-flow swaps and anti-stage red-time swaps will take place from one day to the next, or from today to tomorrow.

### 2.4. Start points and ultimate objectives.

*Start point.* There may also be a specified start point $(\mathbf{X}_0, \mathbf{R}_0)$. This point must in this paper be feasible:

$\mathbf{X}_0$ must satisfy a given demand,
$\mathbf{R}_0$ must also satisfy given constraints and also
$(\mathbf{X}_0, \mathbf{R}_0)$ must be such that $x_{0i} + s_i r_{0i} \leq s_i$ for all $i$.

In this paper we think that today's (route-flow, red-time) pair is feasible, and this pair might also be thought of as a start point.

*Ultimate objectives.* The ultimate objectives of the control policies here are

(i) to ensure that the dynamical adjustments remain feasible, hence our stability results and
(ii) to reduce queueing as illustrated in section 1.4 and section 6.

## 3. $P_0$ Responsive traffic control and dynamic route choice with arbitrary $f_i$

### 3.1. $P_0$ Responsive traffic control with the functions $f_i$

Given the fixed but arbitrary functions $f_i$, our major assumption (stated above) is that

the real link $i$ bottleneck cost = link $i$ bottleneck delay time = $b_i = f_i(x_i + s_i r_i)$ seconds.

This delay time or delay cost (in seconds) is felt by the car driver.

We also define, for each link $i$,

the link $i$ red-time cost = $s_i b_i = s_i f_i(x_i + s_i r_i)$.

This cost is to be felt or measured by the signal controller.

To design the $P_0$ signal control dynamic and the day-to-day re-routeing dynamic we utilise

(a) the major initial assumption,
(b) the route cost vector $\mathbf{C(X, R)} = [C_1(\mathbf{X, R}), C_2(\mathbf{X, R}), \ldots, C_N(\mathbf{X, R})]$ and
(c) the anti-stage red-time cost vector $\mathbf{A(X, R)} = [A_1(\mathbf{X, R}), A_2(\mathbf{X, R}), \ldots, A_M(\mathbf{X, R})]$.

These are all defined in section 2.3 above.

At each junction:

the policy $P_0$ swaps red-time from more red-time costly to less red-time costly anti-stages.

A more precise, proportional, version of this rather general dynamical control policy is as follows.

*Definition 1. The simplest dynamic signal timing proportional $P_0$ adjustment process, $PP_0AP$.*
Given a small $k > 0$, for all ordered pairs of anti-stages, (anti-stage *a*, anti-stage *e*), at the same junction:
the proportional $P_0$ adjustment process ($PP_0AP$):
swaps red time $k[R_a(A_a - A_e)]_+$ from anti-stage *a* to anti-stage *e*. (1)

$P_0$ is built into the red-time adjustment process ($PP_0AP$) or (1) using the $s_ib_i$ in the definition of anti-stage red-time-costs $A_a$ in section 2. Swapping rule (1) is similar to the route flow adjustment *PAP* in Smith (1984) but uses red-times reacting to red-cost differences instead of route-flows reacting to route-cost differences.

See section 2.4 for two ultimate objectives.

### 3.2. Dynamic route choice with the fixed but arbitrary functions $f_i$

*Definition 2. Probably the simplest dynamic proportional routeing adjustment process, PAP, is:*
Given a small $k > 0$, for all ordered pairs of routes, (route *i*, route *j*), joining the same OD pair:
the proportional routeing adjustment process *(PAP)*
swaps route flow $kX_i[C_i - C_j]_+$ from route $i$ to route *j*. (2)
See Smith (1984).

## 4. For fixed demand, small $k$ and arbitrary non-decreasing $f_i$, *PAP* with $PP_0AP$ is stable

### 4.1. For arbitrary non-decreasing $f_i$, [C(X, R), A(X, R)] is a gradient.

Every link in our network has saturation flow $s_i$, link flow $x_i$, red-time $r_i$ and an arbitrary non-decreasing $f_i(x_i + s_ir_i)$. For each link define the Lyapunov function $V$ as follows. First, for each link $i$

$$V_i(x_i, r_i) = \int_0^{x_i} c_i(u)du + \int_0^{x_i + s_i r_i} f_i(u)du$$

and then we put

$$V(\mathbf{x}, \mathbf{r}) = \sum_{i=1}^{n} V_i(x_i, r_i). \quad (3)$$

Then

$$\partial V(\mathbf{x}, \mathbf{r})/\partial x_i \; = \partial \, V_i(x_i, r_i)/\partial x_i = c_i(x_i) + \; f_i(x_i + s_i r_i)$$

and

$$\partial V(\mathbf{x}, \mathbf{r})/\partial r_i = \partial V_i(x_i, r_i)/\partial r_i = s_i f_i(x_i + s_i r_i).$$

So

$$\big(c_i(x_i) + \; f_i(x_i + s_i r_i), s_i f_i(x_i + s_i r_i)\big) \text{ is the gradient of } V_i$$

and

$$\big(\mathbf{c}(\mathbf{x}) + \mathbf{f}(\mathbf{x} + \mathbf{sr}), \mathbf{sf}(\mathbf{x} + \mathbf{sr})\big) \text{ is the gradient of } V(\mathbf{x}, \mathbf{r}) = \sum_{i=1}^{n} V_i(x_i, r_i). \quad (4)$$

Using standard arguments, it now follows that **[C(X, R), A(X, R)]** is a gradient.

### 4.2. For arbitrary non-decreasing $f_i$, (control, routeing) adjustment (1) with (2) is stable; one natural Lyapunov function is given by (3).

In a general network, by virtue of the results in 4.1 above: if we start at a feasible $(\mathbf{X}, \mathbf{R})$, then *PAP* route-flow **X** swapping (toward cheaper routes) (2) and $P_0$ red-time **R** swapping (toward cheaper anti-stages) (1) both reduce $V$ if $k$ is small enough. Thus simultaneously following (1) and (2) reduces $V$.

The paper shows, in a general network:
**A:** that if **f** is bounded then under reasonable conditions these dynamics cause $(\mathbf{X}, \mathbf{R})$ to converge to the set of those $(\mathbf{X}, \mathbf{R})$ which minimize *V*, given by (3), and
**B:** that if **f** is unbounded, representing capacity limits on links; then, under reasonable conditions, dynamics (2) and (1):

(a) keep delays bounded and

(b) cause $(\mathbf{X}, \mathbf{R})$ to converge to the set of those $(\mathbf{X}, \mathbf{R})$ which minimize *V*.

This does not always happen with standard control policies; see Smith (1979c) for example.

### 4.3. Extension to a different control policy but still using the above functions $f_i$

Given the currently arbitrary functions $f_i$, our major initial assumption (stated in section 2.3 above) is still to be that

the link *i* bottleneck delay time or bottleneck delay cost = $b_i = f_i(x_i + s_i r_i)$ seconds.

This delay time or delay cost (in seconds) is felt by the car driver.

Now we design a variation of the $P_0$ signal control policy. To do this we suppose given *n* non-decreasing functions $h_i = h_i(r_i)$. This variation of policy $P_0$, which we call $P_\mathbf{h}$, is defined as follows.

First we define, for each link *i*,

the new link *i* red-time cost = $s_i b_i + h_i = s_i f_i(x_i + s_i r_i) + h_i(r_i)$;

this cost is again to be felt or measured by the signal controller, and the new anti-stage cost $A_{\mathbf{h}u}$ is then defined for all $(\mathbf{X}, \mathbf{R})$ to be:

$$A_{\mathbf{h}u}(\mathbf{X}, \mathbf{R}) = \sum_{\text{link } i \text{ belongs to anti-stage } u} [s_i b_i + h_i] = \sum_{\text{link } i \text{ belongs to anti-stage } u} [s_i f_i(x_i + s_i r_i) + h_i(r_i)].$$

Using this new definition of anti-stage red-time cost, control policy $P_\mathbf{h}$ swaps red-time from more costly anti-stages to less costly anti-stages. This is given in more detail in the following definition (4) which is a slightly changed version of definition 1.

*Definition 3. The signal timing proportional $P_\mathbf{h}$ adjustment process $PP_\mathbf{h}AP$.*
Given a small $k > 0$, for all ordered pairs of anti-stages, (anti-stage *a*, anti-stage *e*), at the same junction, the proportional $P_\mathbf{h}$ adjustment process ($PP_\mathbf{h}AP$):

swaps red time $k[R_a(A_{\mathbf{h}a} - A_{\mathbf{h}e})]_+$ from anti-stage *a* to anti-stage *e*. (5)

$P_\mathbf{h}$ is built into the red-time adjustment process ($PP_\mathbf{h}AP$) or (5) using the $s_i b_i$ and $h_i$ in the definition of red-time-costs in this section 4.

As with swapping rule (1), swapping rule (5) is similar to the route flow adjustment *PAP* in Smith (1984) but uses red-times which react to red-cost differences instead of route-flows which react to route-cost differences.

*Remark.* In this case the new control policy depends on *n* fixed arbitrary non-decreasing functions $h_i$.

Every link *i* in our network has saturation flow $s_i$, link flow $x_i$, red-time $r_i$ (which may be zero for many links), an arbitrary non-decreasing $f_i(x_i + s_i r_i)$, and an arbitrary non-decreasing function $h_i(r_i)$. For each link *i* define:

$$V_{\mathbf{h}i}(x_i, r_i) = \int_0^{x_i} c_i(u)du + \int_0^{x_i + s_i r_i} f_i(u)du + \int_0^{r_i} h_i(u)du \tag{6}$$

and then put

$$V_\mathbf{h}(\mathbf{x}, \mathbf{r}) = \sum_{i=1}^{n} V_{\mathbf{h}i}(x_i, r_i). \tag{7}$$

It follows that

$$\partial V_{\mathbf{h}}(\mathbf{x},\mathbf{r})/\,\partial x_i \;= \partial\, V_{\mathbf{h}i}(x_i,r_i)/\partial x_i = c_i(x_i) + \; f_i(x_i + s_i r_i)$$

and

$$\partial V_{\mathbf{h}}(\mathbf{x},\mathbf{r})/\,\partial r_i = \partial V_{\mathbf{h}i}(x_i,r_i)/\partial r_i = s_i f_i(x_i + s_i r_i) + h_i(r_i).$$

So that

$$\big(c_i(x_i) + \; f_i(x_i + s_i r_i), s_i f_i(x_i + s_i r_i) + h_i(r_i)\big) \text{ is the gradient of } V_{\mathbf{h}i}$$

and

$$\big(\mathbf{c}(\mathbf{x}) + \mathbf{f}(\mathbf{x} + \mathbf{sr}), \mathbf{sf}(\mathbf{x} + \mathbf{sr}) + \mathbf{h}(\mathbf{r})\big) \text{ is the gradient of } V_{\mathbf{h}}(\mathbf{x},\mathbf{r}) = \sum_{i=1}^{n} V_{\mathbf{h}i}(x_i,r_i). \qquad (8)$$

Using standard arguments, it now follows that **[C(X, R), $\mathbf{A_h}$(X, R)]** is a gradient.

#### 4.3.1. For arbitrary non-decreasing $f_i$ and $h_i$ (control, routeing) adjustment (4) with (2) is stable.

In a general network, by virtue of the results in 4.3 above: if we start at a feasible $(\mathbf{X},\mathbf{R})$, then *PAP* route-flow **X** swapping (toward cheaper routes) (2) and $P_{\mathbf{h}}$ red-time **R** swapping (toward cheaper anti-stages) (5) both reduce $V_{\mathbf{h}}$ if $k$ is small enough. Thus, under natural conditions, simultaneously following (5) and (2) reduces $V_{\mathbf{h}}$.

In appendix B in the full paper it is shown that, in a general network:
(i): if **f** is bounded then under reasonable conditions these dynamics cause $(\mathbf{X},\mathbf{R})$ to converge to the set of those $(\mathbf{X},\mathbf{R})$ which minimize $V_{\mathbf{h}}$, given by (7), and
(ii): if **f** is unbounded, representing capacity limits on links; then, under reasonable conditions, dynamics (2) and (5): (a) keep delays bounded and (b) cause $(\mathbf{X},\mathbf{R})$ to converge to the set of those $(\mathbf{X},\mathbf{R})$ which minimize $V_{\mathbf{h}}$, given by (7). This does not always happen with standard control policies; see Smith (1979) for example.

## 5. Extensions to allow increased link delay functions as well as the original real link delay functions

The results obtained in the previous sections apply for *any* fixed collection of non-decreasing non-negative link functions $f_i(x_i + s_i r_i)$.

Given an arbitrarily chosen collection of $n$ non-decreasing non-negative functions $f_i(x_i + s_i r_i)$, it follows that:
**(A)** typically real bottleneck delays $b_i$ actually experienced will be very different to the above arbitrarily chosen $f_i(x_i + s_i r_i)$ and, somewhat on the other hand,
**(B)** we have proved stability results (see section 4 above) for a *very wide variety* of arbitrary (non-decreasing) functions $f_i(x_i + s_i r_i)$ and $h_i(r_i)$.

There are two directions which arise when we consider the two statements **A** and **B** above. These are:
**Direction A**: Select the $f_i(x_i + s_i r_i)$ so as *to most closely approximate real delay functions*, and
**Direction B**: Select the $f_i(x_i + s_i r_i)$ and the $h_i(r_i)$ so as *to most help control traffic flows and queues.*

Directions **A** and **B** are very likely to be conflicting directions; it is very unlikely that the real delay functions $f_i(x_i + s_i r_i)$ will ALSO be the delay functions which control the network best.

*Now assume that real delay functions have been determined; we denote these by $f_i(x_i + s_i r_i)$.*

Given these real delay functions, which give delays which are actually felt by drivers, we show here

how to exploit *both* directions A and B even when these directions are conflicting (which is the usual case).

Of course we use the real delay functions $f_i(x_i + s_i r_i)$ (following direction **A**), but also we follow direction **B** here by using *increased* delay functions $f^I_i(x_i + s_i r_i)$. These are defined as follows:

$f^I_i(x_i + s_i r_i) = max\{m_i(x_i + s_i r_i), f_i(x_i + s_i r_i)\} \geq f_i(x_i + s_i r_i)$ for all $(x_i + s_i r_i) \geq 0$.

*In this paper we utilise only these very simple linear "potentially better" increased delay functions.* Here the $m_i$ are fixed and $\geq 0$. We do not discuss in this paper how the $m_i$ or the $h_i$ the are selected.

We emphasise that the functions $f_i(x_i + s_i r_i)$ are now supposed here to be the real delay functions, actually felt by drivers. The paper does not suggest how these might be chosen. But it would be reasonable to begin by looking at Webster's delay formula. The $m_i$ and so the increased delay functions

$$f^I_i(x_i + s_i r_i) = max\{m_i(x_i + s_i r_i), f_i(x_i + s_i r_i)\}$$

are now also fixed in the rest of the document.

We have fixed both

the real delay functions $f_i(x_i + s_i r_i)$ and the increased delay functions $f^I_i(x_i + s_i r_i)$.

*We do not say here how the $f_i(x_i + s_i r_i)$ or the $m_i$ are chosen.*

### 5.1. Real delay functions $f_i(x_i + s_i r_i)$ and increased delay functions $f^I_i(x_i + s_i r_i)$

For link *i* possible different delay functions, solid curve (increased) and dotted curve (real) are illustrated in figure 3.

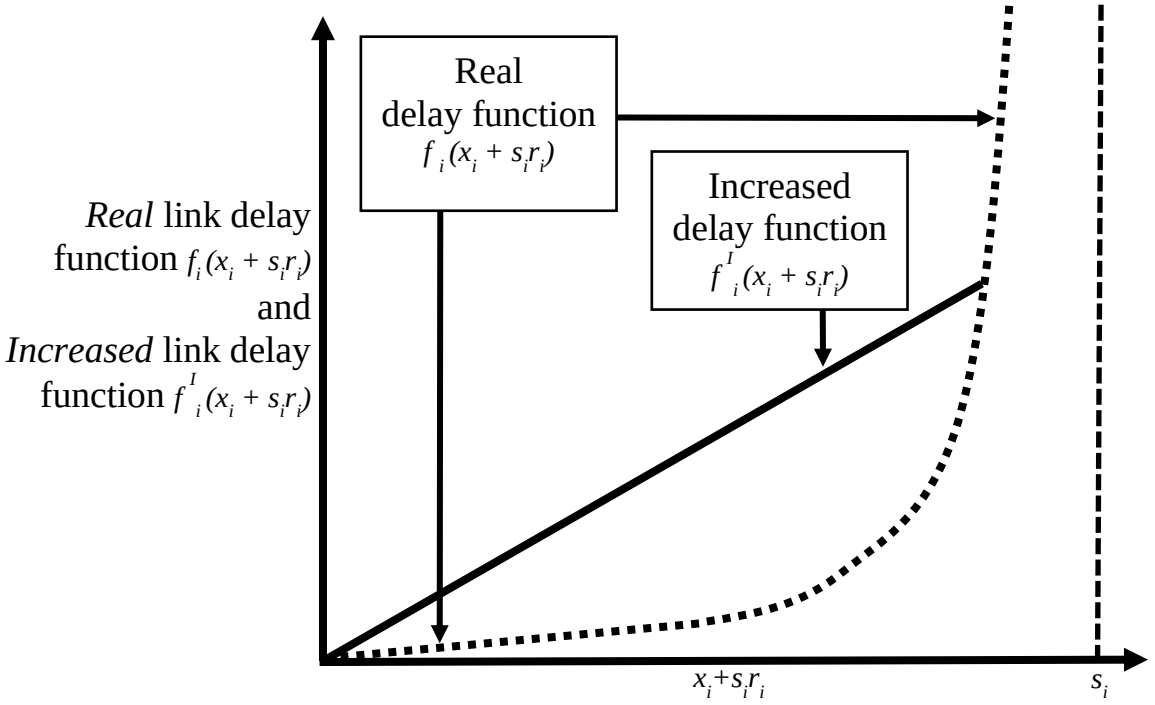



Figure 3. This shows both a reasonable real delay function $f_i(x_i + s_i r_i)$ (dotted) and a reasonable increased delay function $f^I_i(x_i + s_i r_i)$ (solid).

Now we seek to control the real network (using the dotted real delay curve) employing the solid increased delay curve as a "target" delay curve.

Like this we

(a) retain the stability shown in section 4 using the increased delay function $f^I_i(x_i + s_i r_i)$ and

(b) MAY cause routeing / control adjustments which reduce agreed measures of network performance (such as queue volumes and flows); using $f^{I}_{i}(x_i + s_i r_i)$ and $f_i(x_i + s_i r_i)$ together.

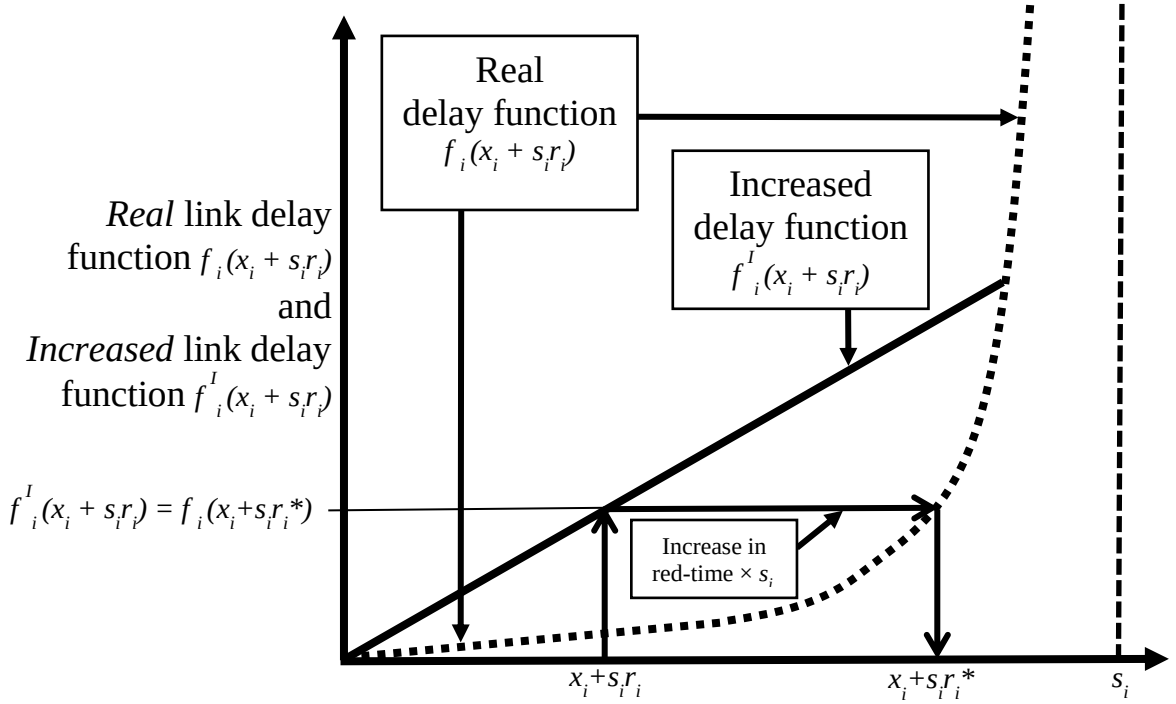



Figure 4. Real (dotted) and Increased (solid) link delay functions. Here $r_i^*$ or $x_i + s_i r_i^*$ is determined by increasing the red time $r_i$ by $r_i^* - r_i$; this red-time increase $r_i^* - r_i$ is designed to ensure that the real delay $f_i(x_i + s_i r_i^*)$ actually felt by drivers does equal $f^{I}_{i}(x_i + s_i r_i)$.

With these two functions $f^{I}_{i}$ and $f_i$, the red-time re-allocation is to be a simply modified form of rule (1) in definition 1; the simple modification takes careful account of both the real delay formulae $f_i(x_i + s_i r_i)$ and the increased delay functions $f^{I}_{i}(x_i + s_i r_i)$ above.

### 5.2. Ensuring that delays required by the increased delay functions $f^{I}_{i}(x_i + s_i r_i)$ are felt, using the real delay functions $f_i(x_i + s_i r_i)$

As illustrated in figure 4, given the real delay formula $f_i(x_i + s_i r_i)$ and given the increased delay function $f^{I}_{i}(x_i + s_i r_i)$ above, we determine $x_i + s_i r_i^*$ and so $r_i^*$, by following the three arrows (up, right, down) toward the bottom of figure 4. It is always the case that $r_i^* \geq r_i$. Following the three arrows in this way guarantees that $r_i^*$ satisfies:

$$f^{I}_{i}(x_i + s_i r_i) = f_i(x_i + s_i r_i^*);$$

*ensuring that the delays* $f^{I}_{i}(x_i + s_i r_i)$ *are actually felt as* $f_i(x_i + s_i r_i^*)$*; provided we choose the real red-time to be* $r_i^*$

Thus we may control the network using the real cost functions $f_i(x_i + s_i r_i^*)$, by choosing $r_i^*$ so that $r_i$ follows (1) above with the increased delay function. Then provided **x** follows (2) we retain the stability shown in section 4, with the increased delay function $f^{I}_{i}(x_i + s_i r_i)$.

In more detail, if we wish $r_i$ to follow (1), and hence (say) increase $r_i$ to $r_i + \alpha_i$, with the increased delay function then we would increase $r_i^*$ to $(r_i + \alpha_i)^*$ with the real delay function as shown in figure 5 where $(x_i + s_i r_i)$ determines $(x_i + s_i r_i^*)$ and $(x_i + s_i(r_i + \alpha_i))$ determines $(x_i + s_i(r_i + \alpha_i)^*)$.

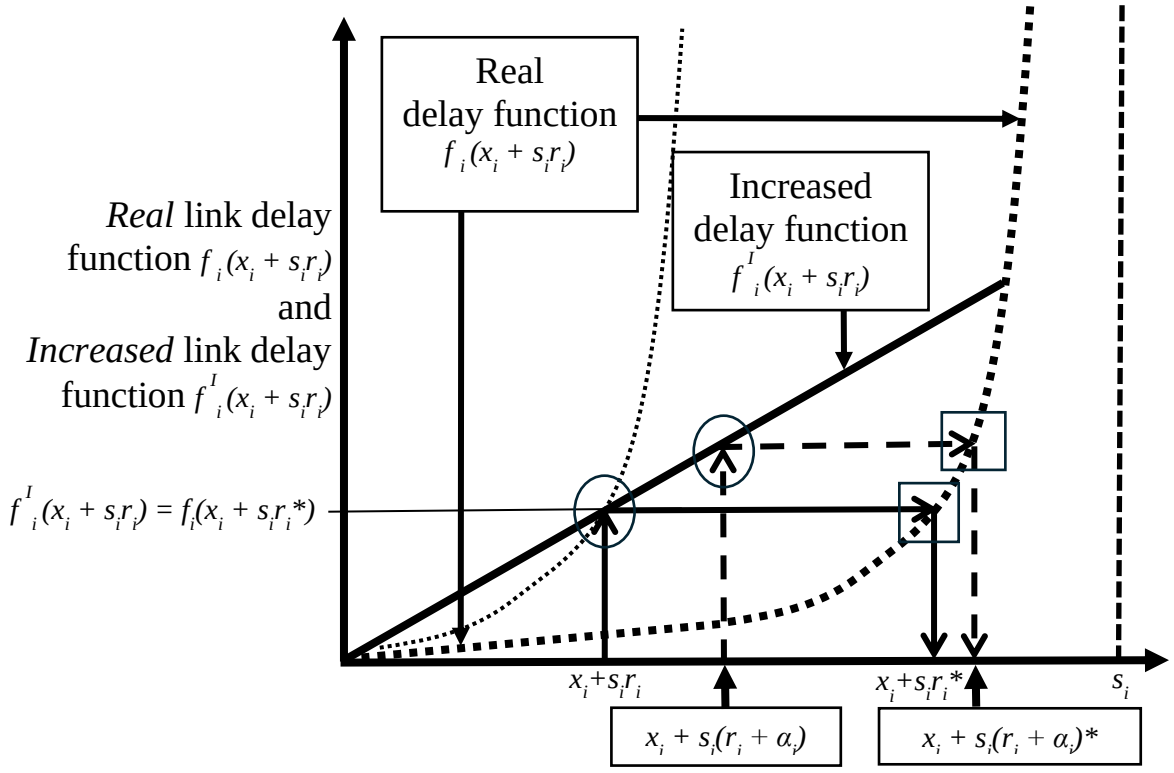



Figure 5. Increased (solid) and real (dotted) link delay functions. Showing also the connections between $x_i + s_i r_i$ and $x_i + s_i r_i^*$, and between $x_i + s_i(r_i + \alpha_i)$ and $x_i + s_i(r_i + \alpha_i)^*$.

In figure 5 $r_i^*$ is determined by increasing the red time $r_i$ by $r_i^* - r_i$ to ensure that

$$f^I_i(x_i + s_i r_i) = f_i(x_i + s_i r_i^*).$$

Also $(r_i + \alpha_i)^*$ is determined by increasing the red time $(r_i + \alpha_i)$ by $(r_i + \alpha_i)^* - (r_i + \alpha_i)$ to ensure that

$$f^I_i(x_i + s_i(r_i + \alpha_i)) = f_i(x_i + s_i(r_i + \alpha_i)^*).$$

So we may use the $r_i^*$ and the real delay curve to control the $r_i$ in accordance with (1) with the increased delay function; and still have stability even though the real cost functions $f_i$ (dotted) are not at all the same as the increased delay functions $f^I_i$.

### 5.1.1 Summary.

Suppose we are using the increased delay function at $x_i + s_i r_i$. The increased delay function then gives a delay of $f^I_i(x_i + s_i r_i)$; this is *much larger* than the real delay $f_i(x_i + s_i r_i)$ also at $x_i + s_i r_i$, as shown by the real dotted curve. Somehow we must arrange that the real delays felt are actually $f^I_i(x_i + s_i r_i)$. To achieve this we utilise the dotted real delay function with an increased red time as shown above.

### 5.1.2 An alternative view.

An alternative way of thinking about the above addition of red time is as follows. The above addition of red-time may be thought of as shifting the dotted real curve to the left; this is illustrated in figure 5 by the light dotted curve in figure 5. To shift the dotted curve to the left as shown we must increase the red-time from $r_i$ to $r_i^*$ shown in figure 5. The light left-shifted dotted curve shown in figure 5 is just the bold dotted curve moved to the left by $s_i(r_i^* - r_i)$. The shift we need, at $(x_i + s_i r_i)$, in the dotted real curve is illustrated in figure 5: the shift is the length of the arrow joining the lower circled point to the lower point surrounded by a square. Thus the added red-time must be $(r_i^* - r_i)$. Provided this is added to the red-time $r_i$ the dotted real delay curve, showing the delay actually felt, will indeed be shifted to the left by $s_i(r_i^* - r_i)$ and

$$f_i(x_i + s_i[r_i + (r_i^* - r_i)]) = f_i(x_i + s_i r_i^*) = f^I{}_i(x_i + s_i r_i).$$

Thus we have achieved our aim to cause traffic to feel a *real* delay $f^I{}_i(x_i + s_i r_i)$; at $x_i + s_i r_i^*$. Further the shift we need at $x_i + s_i(r_i + \alpha_i)$, in the dotted real curve, is illustrated in figure 5: the shift is the length of the arrow joining the upper circled point to the upper point surrounded by a square.

## 6. Illustrative example; gating is required to reduce equilibrium queues.

Figure 1 illustrates a simple network with one traffic signal at a bottleneck at node $n_2$. As is shown in section 1.4 gating at the traffic signal reduces equilibrium queues. We reconsider this network here.

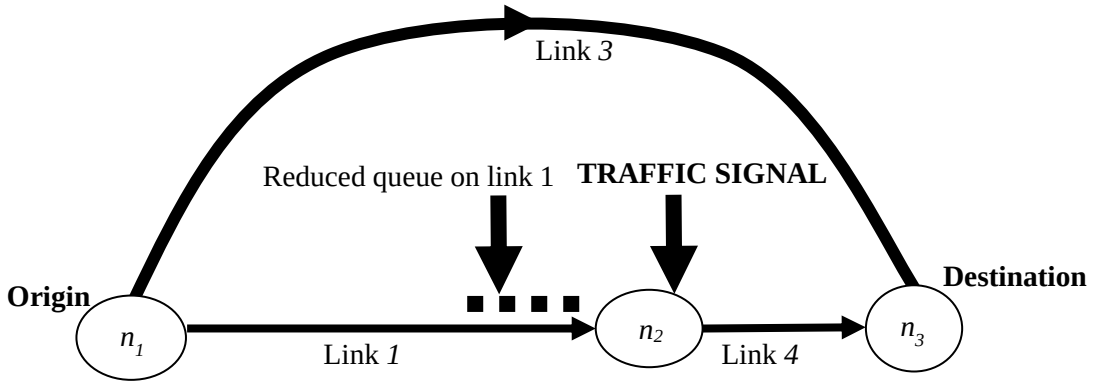



Figure 6. With gating: Link 1 green-time proportion $g_1 = ½ < 1$. Provided the bypass link 3 has a large capacity, gating ensures that the link 1 queue is *more than halved*. See appendix A in the full paper for a detailed justification of this statement. Much of the traffic originally using link 1 is diverted to the bypass link 3, away from links 1 and 4.

Now we examine this network without and with gating in more detail, as $T$ increases, using the real and increased delay-flow curves illustrated in figures 3, 4, and 5. $T$ increasing may be thought of as an increasing part of a peak period.

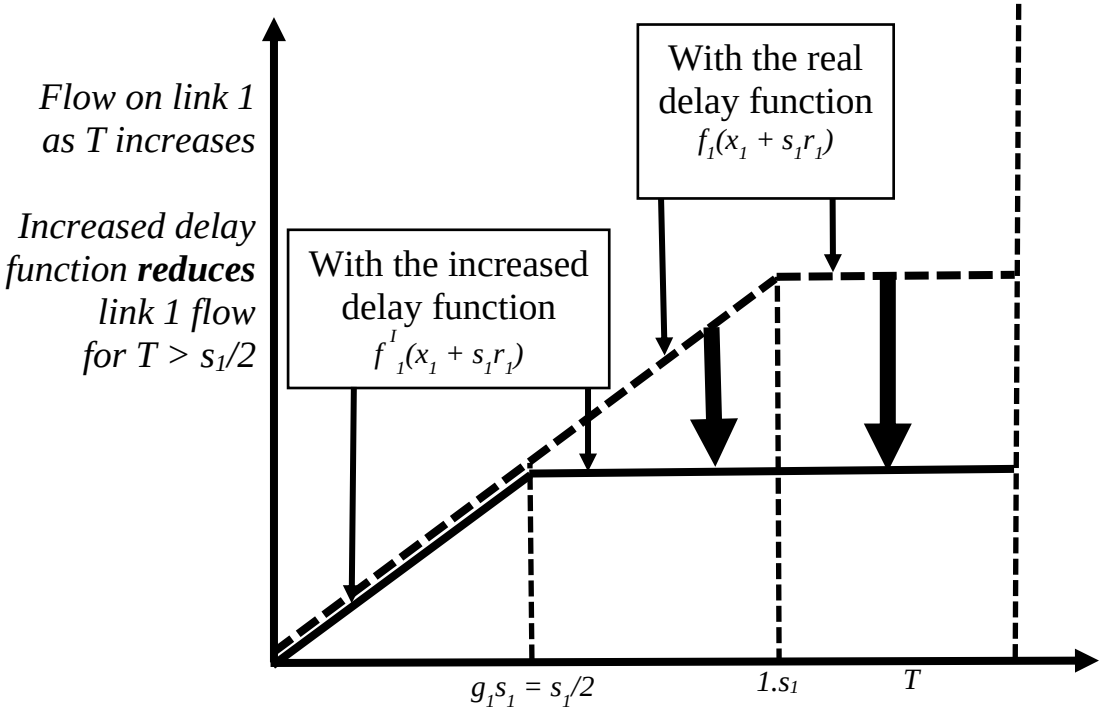



Figure 7. This shows how flow on link 1 varies with the OD flow $T$; with both a reasonable real delay function $f_i(x_i + s_i r_i)$ (dotted) and a reasonable increased delay function $f^I{}_i(x_i + s_i r_i)$ (solid). If the real delay function is switched to the increase delay function by using the gating mechanism in this paper,

flows on link 1 are reduced and the reduction is greater closer to the peak of the peak.

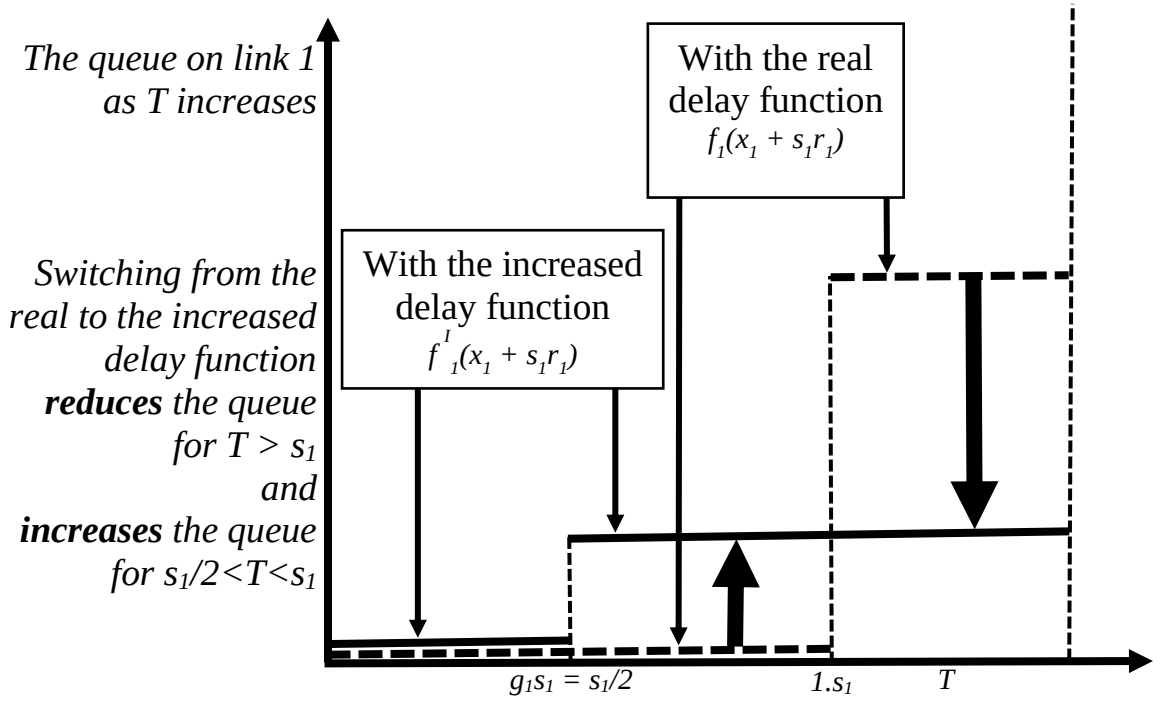



Figure 8. This figure shows how the queue on link 1 varies as the OD flow $T$ increases, with both a reasonable real delay function $f_i(x_i + s_ir_i)$ (dotted) and a reasonable increased delay function $f^I_i(x_i + s_ir_i)$ (solid). If the real delay function is switched to the increased delay function by using the gating mechanism in this paper, queues on link 1 are reduced near the peak of the peak, but increased earlier.

## 7. Conclusion.

This paper states a new sensitive gating method, and shows how this method may reduce queues and flows on sensitive links. The method fits into the control policy $P_{\mathbf{h}}$ which is capacity maximising, and the [$P_{\mathbf{h}}$ + gating] combination described here is still, under certain conditions, capacity maximising. There are many opportunities for developing the method described here.

### 7.1. The major assumption.

The main motivation behind the MAJOR ASSUMPTION, that link $i$ delays have the form of an increasing function of $(x_i + s_ir_i)$, is that with this major assumption we can obtain the stability results in section 4. The main "next step" must be to seek relaxations on this major assumption.

### 7.2. The mainly linear increased delay function.

We have chosen the increased delay function

$$f^I_i(x_i + s_ir_i) = max\{m_i(x_i + s_ir_i), f_i(x_i + s_ir_i)\}.$$

The linear function $m_i(x_i + s_ir_i)$ has been chosen for a variety of reasons:

(i) our linear function $m_i(x_i + s_ir_i)$ is simple;

(ii) the need to divert traffic from sensitive links increases as flows and red times increase so $m_i$ should be > 0 (and the added delays should not be constant);

(iii) our linear function $m_i(x_i + s_ir_i)$ is small if $x_i + s_ir_i$ is small, and so has little effect when flows and red-times are small, which is desirable; and we do not need to increase real-life link delays substantially when flow and red-times are small, which could be both difficult technically and also controversial in practice.

The increased function $f^I_i(x_i + s_ir_i) = max\{m_i(x_i + s_ir_i), f_i(x_i+s_ir_i)\} = f_i(x_i+s_ir_i)$ beyond a certain $x_i+s_ir_i$.

Perhaps one aim would be to choose the $m_i$ so that this is a reasonably rare occurrence.

### 7.3. Future developments.

This paper has introduced a new sensitive gating strategy. The strategy has a vector of parameters **m**. The possibility of adjusting **m** from day to day merits exploration; perhaps using AI or ordinary intelligence. The greatest, and most immediate, need is for *a relaxation of the major assumption.*

### References.


Allsop, R.E. 1974. Some possibilities of using traffic control to influence trip distribution and route choice. In: Proceedings of the 6th International Symposium on Transportation and Traffic Theory. pp. 345–374.

Bentley, R. W. and Lambe, T. A. 1980. Assignment of traffic to a network of signalized city streets. Transportation Research Part A, 14(1), 57-65.

Beckmann, M., McGuire, C. B., and Winsten, C. B. (1956). *Studies in the Economics of Transportation*. Yale University Press.

Bie, J., Lo, H.K. 2010. Stability and attraction domains of traffic equilibria in a day-to-day dynamical system formulation. Transp. Res. B. 44, 90–107.

Cantarella, G.E. 2010. Signal setting with dynamic process assignment. In New Developments in Transport Planning: Advances in Dynamic Traffic Assignment; Tampère, C.M.J., Viti, F., Immers, L.H., Eds.; Edward Elgar: Northampton, MA, USA, 2010; pp. 29–56.

Cantarella, G.E.; Velonà, P.; Vitetta, A. 2012. Signal setting with demand assignment: Global optimization with day-to-day dynamic stability constraints. J. Adv. Transp. 46, 254–268.

He, X., Wang, J., Peeta, S., Liu, H.X. 2022. Day-to-Day Signal Retiming Scheme for Single-Destination Traffic Networks Based on a Flow Splitting Approach. Networks and Spatial Economics 22, 855–882 https://doi.org/10.1007/s11067-022-09566-9

Hu, T.-Y., Mahmassani, H.S. 1997. Day-to-day evolution of network flows under real-time information and reactive signal control. Transp. Res. C. 5, 51–69.

Le,T., Kovacs, P., Walton, N., Vu, H.L., Andrew, L.L.H., Hoogendoorn, S.S.P. 2015. Decentralized signal control for urban road networks. Transp. Res. C, 58, 431–450.

Liu, R., Mounce, R., Smith, M.J. 2015. Route choice and signal control; a study of the stability and instability of traffic signal-controlled networks. Transp. Res. B 77, 123–145.

Meneguzzer, C. 1996. Computational experiments with a combined traffic assignment and control model with asymmetric cost functions. In Proceedings of the 4th International Conference on Applications of Advanced Technologies in Transportation Engineering, Capri, Italy, 27–30 June 1995. Published in the Proceedings: Editors Stephanedes, Y.J., Filippi, F., Eds.; ASCE: New York, NY, USA, 1996; pp. 609–614.

Meneguzzer, C. 1997. Review of models combining traffic assignment and signal control. J. Transp. Eng. 123, 148–155.

Meneguzzer, C. 2012. Dynamic process models of combined traffic assignment and control with different signal updating strategies. J. Adv. Transp., 46, 351–365.

Meneguzzer, C. 2024. Stability of Traffic Equilibria in a Day-to-Day Dynamic Model of Route Choice and Adaptive Signal Control. Appl. Sci. 2024, 14, 1891. https://doi.org/ 10.3390/app14051891

Quinn, D. 1992. A Review of Queue Management Strategies. Deliverable number 1 in the DRIVE II PROJECT V2016: PRIMAVERA Priority Management for Vehicle Efficiency, Environment and Road Safety on Arterials. ITS University of Leeds, Mizar Automazione Spa Città di Torino, HETS, Peek Traffic Ltd.

Mounce, R. 2009. Existence of Equilibrium in a Continuous Dynamic Queueing Model for Traffic Networks with Responsive Signal Control. In: Lam, W., Wong, S., Lo, H. (eds) Transportation and Traffic Theory 2009: Golden Jubilee. Springer, Boston, MA. https://doi.org/10.1007/978-1-4419-0820-9_16

Satsukawa, K., Iryo, T., Yoshizawa, N., Smith, M. J., and Watling, D. Adjustment process of decentralised signal control policies with route choices: a case study with the $P_0$ policy. Submitted.

Smith, M. J., 1979a. A local traffic control policy which automatically maximises the overall travel capacity of an urban road network. Proceedings of the International Symposium on Traffic Control Systems, University of California, Berkeley, Volume 2A, 11-32; and in Traffic Engineering and Control, 21, 1980, 298-302.

Smith, M., J. 1979b. The existence, uniqueness and stability of traffic equilibria. Transportation Research Part B, 13(4), 295-304.

Smith, M.J. 1979c. Traffic control and route-choice; a simple example. Transp. Res. 13B, 289–294.

Smith, M.J. 1984. The stability of a dynamic model of traffic assignment– an application of a method of Lyapunov. Transportation Science 18 (3), 245–252.

Smith, M.J., van Vuren, T. 1993. Traffic equilibrium with responsive traffic control. Transp. Sci. 27, 118–132.

Smith, M.J., Huang, W., Viti, F., Tampere, C.T.M., Lo, H.K. 2019a. Quasi-dynamic traffic assignment with spatial queueing, control and blocking back. Transportation Research Part B, 122, 140–166.

Smith, M., J., Viti, F., Huang, H., Mounce, R. 2023. Upstream-gating merge-control for maximising network capacity: with an application to urban traffic management. Transportation Research Part C 155, 104287.

Varaiya, P. 2013. Max pressure control of a network of signalised intersections. Transp. Res. C 57, 177–195.

Xiao, L., Lo, H.K. 2015. Combined route choice and adaptive traffic control in a day-to-day dynamical system. Netw. Spat. Econ. 15, 697–717.

Yang, H., Yagar, S. 1995. Traffic assignment and signal control in saturated road networks. Transp. Res. A, 29, 125–139.

# Sensitivity Analysis Through A Multi-objective Dynamic User Equilibrium Assignment Formulation

Sadullah Goncu[a,c], Mehmet Ali Silgu[d,e], Hilmi Berk Celikoglu[b,c]

[a]*Fatih Sultan Mehmet University, Department of Civil Engineering, Istanbul, 34445, Turkey*
[b]*Technical University of Istanbul, Department of Civil Engineering, Istanbul 34469, Turkey*
[c]*Technical University of Istanbul (ITU), ITU ITS Research Lab, Istanbul 34469, Turkey*
[d]*Bartın University, Department of Civil Engineering, Bartın, 74100, Turkey*
[e]*Koc University, Department of Industrial Engineering, Istanbul, 34450, Turkey*

**Abstract**: **This study investigates the impact of varying emission weight coefficients in the objective function of a Dynamic User Equilibrium (DUE) model. The primary aim is to analyze the sensitivity of route choice, travel costs, and emission reductions to changes in the emission operator's weight. An analytical framework is developed to explore equilibrium properties under different weighting scenarios. The findings provide insights into the trade-offs between travel time and environmental impact, offering theoretical foundations for emission-based traffic control policies.**

**Keywords**: Dynamic Traffic Assignment, Emissions, Green Campus

## 1. Introduction

Due to environmental degradation, transportation planning has increasingly incorporated an environmental perspective to mitigate the negative impacts of vehicular traffic on road networks. Road transportation has been identified as one of the largest sources of carbon emissions, air pollution, and fuel consumption [1]. To address these concerns, researchers have focused on integrating emission considerations into traffic assignment models, leading to developing environmentally friendly traffic assignment frameworks.

Traffic assignment models are essential for analyzing and optimizing vehicular flow across road networks while considering various constraints, including environmental impacts. These models aim to replicate real-world traffic behavior under simplified assumptions, enabling transportation planners to assess congestion patterns, estimate emissions, and develop effective mitigation strategies. Based on their treatment of temporal variations in traffic flow, traffic assignment methods can be classified into three categories: static traffic assignment (STA), semi-dynamic traffic assignment (semi-STA), and dynamic traffic assignment (DTA).

STA methods assume that traffic conditions remain constant over a given period, disregarding variations in congestion levels and travel demand fluctuations. This approach is computationally efficient and widely used for long-term planning but lacks the ability to model time-dependent congestion and queuing effects. DTA, on the other hand, models traffic flow as an evolving system where vehicles respond dynamically to congestion, route availability, and travel time variations. DTA integrates time-varying origin-destination (O-D) matrices, allowing it to capture key traffic phenomena such as queue spillback, shockwaves, and speed fluctuations [2]. As a result, DTA provides a more realistic representation of network conditions and is particularly valuable for studying real-time traffic management strategies and emission-reduction policies

[3]. Given its ability to reflect detailed traffic dynamics, integrating environmental objectives into DTA models has become a crucial research area [4]. By incorporating emission models into dynamic cost functions, researchers can evaluate how traffic control strategies impact both congestion and air quality, ultimately informing the development of sustainable urban mobility solutions. Dynamic User Equilibrium (DUE) models traditionally prioritize travel time minimization, but recent advances incorporate environmental objectives to achieve a balance between efficiency and sustainability [4]. The aim is to evaluate how varying emission weights influence route choices and system-wide emissions, providing insights into the trade-offs between travel time and environmental impact.

In our previous study [5], we presented a methodology to explicitly consider emissions in the DUE objective function alongside delays. This study extends our previous work on emission-integrated DUE models by conducting a sensitivity analysis on the weight of the emission operator in the objective function. The contributions of this study are as follows: i) Development of an analytical framework for evaluating the sensitivity of emission weights in DUE models ; ii) Examination of equilibrium properties under different weighting scenarios, and ; iii) Insights into the trade-offs between travel time and emissions for policy implications.

## 2. Methodology

We consider the following dynamic objective function [5]–[7].

$$\begin{aligned} &\min_{h^* \in \Lambda} \quad \alpha \sum_{p \in \mathcal{P}} \Delta t \sum_{k=1}^{n} T_p^k(h^*) h_p^k + \beta \sum_{p \in \mathcal{P}} \Delta t \sum_{k=1}^{n} E_p^{m,k}(h^*) h_p^k \\ &s.t. \quad h \in \Lambda_d = \left\{ h \geq 0 : \sum_{p \in \mathcal{P}_{ij}} \Delta t \sum_{k=1}^{n} h_p^k = Q_{ij}, \forall (i,j) \in \mathcal{W} \right\} \end{aligned} \tag{1}$$

where, $\mathcal{P}$ is the set of paths in the network, $\mathcal{W}$ is the set of O-D pairs in the network, $Q_{ij}$ is the O-D demand between $(i, j) \in \mathcal{W}$, $\mathcal{P}_{ij}$ is the subset of paths that connect O-D pair $(i, j)$, $t$ is the time parameter in a time horizon $[t_0, t_f]$, $h(t)$ is the complete vector of departure rates $h(t) = (h_p(t) : p \in \mathcal{P})$, $h_p(t)$ is the departure rate along path $p$ at time $t$, $\psi_p(t,h)$ is the travel cost along path $p$ with the departure time $t$ under the departure profile $h$, $v_{ij}(h)$ is the minimum travel cost between O-D pair $(i, j)$ for all paths and departure profiles, $h_p$ is the flow on path $p$, $T_p(h_p)$ represent the travel time operator based on the flow on path $p$ ( $T_p(\cdot) : \Re_+ \mapsto \Re_+, \forall p \in \mathcal{P}$), $M$ is the set of all pollutant types, $E_{m,p}(h_p)$ denotes the emissions of pollutant $m$ on path $p$, $\alpha$ and $\beta$ is the weight coefficients. Here, $\alpha \geq 0, \beta \geq 0$ and $\alpha + \beta = 1$.

The sensitivity analysis of the presented objective function is conducted on Istanbul Technical University (ITU) campus network [8]. ITU network consists of 78 nodes, 178 links, 344 O-D pairs, and a total of 6039 paths. The demand data is collected through real-world measurements (i.e., observed speeds on links,

turning ratios at intersections), which are used to obtain demand profiles $Q_{ij}$ (for morning peak hours). We analyze the sensitivity of DUE solutions through the parameters $\alpha$ , and $\beta$ . We change the values of $\alpha$ and $\beta$ from 0 to 1 with 0.05 increments. The sensitivity analysis is conducted on an analytical model in MATLAB. Each solution is for 18000 seconds with 1 second time step. The emission results are obtained through the integrated emission model of Barth and Boriboonsomsin [9].

## 3. Results and Conclusion

The sensitivity analysis conducted on the ITU campus network examines the impact of varying emission weight coefficients on total emissions (grams) and total delay (hours). The results demonstrate a clear trade-off between travel efficiency and environmental impact, as illustrated in Figure 1. The results indicate that increasing the emission weight ( $\beta$ ) in the objective function shifts route choices towards more environmentally friendly paths, reducing emissions but at the cost of increased total delay. Conversely, prioritizing delay minimization ( $\alpha$ ) results in higher emissions due to traffic concentration on faster but more congested routes. The pure delay-based DUE solution produces the lowest total delay ($6.09 \times 10^6$ hours) but results in the highest total emissions ($5.03 \times 10^9$ grams). Introducing a small emission weight already results in a measurable reduction in emissions, demonstrating the effectiveness of eco-routing with minimal impact on delay. At $\alpha$=0.90, $\beta$=0.1, emissions reach their lowest recorded level in the sensitivity analysis ($4.995 \times 10^9$ grams), while the total delay remains within a reasonable range ($6.11 \times 10^6$ hours).

The relationship between emissions and delay exhibits nonlinear characteristics. Moderate emission prioritization ($\beta$ between 0.05 - 0.1) significantly reduces emissions without drastically increasing delay. Beyond a certain threshold ( $0.2 \geq \beta \geq 0.1$), emission reductions plateau, while the total delay continues to rise, making further emission prioritization inefficient from a traffic performance perspective. From a policy standpoint, the results suggest an optimal range of values (0.05. - 0.2) where sustainable routing can be effectively integrated into traffic assignment models without excessively impacting travel time. Prioritizing emissions entirely ($\alpha$=0, $\beta$=1) leads to the highest total delay ($6.19 \times 10^6$ hours) while achieving marginal emission reduction gains beyond the optimal trade-off point.

Building on the analytical framework established in this study, future extensions will focus on incorporating stochastic elements to better reflect the inherent variability in real-world traffic systems. This includes the integration of stochastic fundamental diagrams to capture uncertainties in flow-density relationships and stochastic route choice models to represent variability in driver behavior beyond the deterministic Wardrop equilibrium assumptions. Additionally, the framework could be expanded to explore multi-class user dynamics, real-time adaptive control strategies, and uncertainty propagation in emissions modeling.

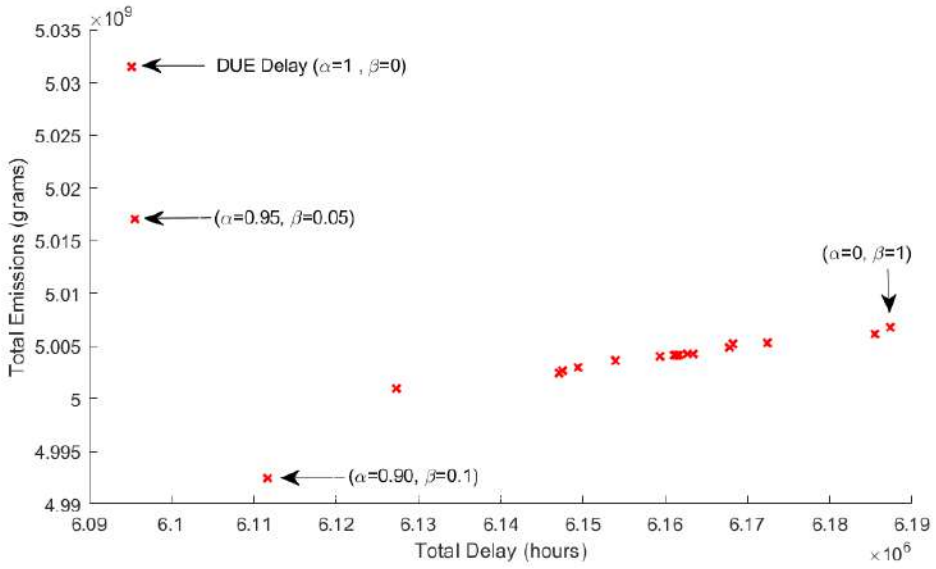



**Figure 1.** Obtained results of the sensitivity analysis

## References


[1] Y. Wang, W. Y. Szeto, K. Han, and T. L. Friesz, "Dynamic traffic assignment: A review of the methodological advances for environmentally sustainable road transportation applications," *Transp. Res. Part B Methodol.*, vol. 111, pp. 370–394, May 2018, doi: 10.1016/j.trb.2018.03.011.

[2] Y.-C. et al. Chiu, *Dynamic Traffic Assignment: A Primer*. Washington, D.C.: Transportation Research Board, 2011.

[3] T. Fontes, S. R. Pereira, P. Fernandes, J. M. Bandeira, and M. C. Coelho, "How to combine different microsimulation tools to assess the environmental impacts of road traffic? Lessons and directions," *Transp. Res. Part D Transp. Environ.*, vol. 34, pp. 293–306, Jan. 2015, doi: 10.1016/j.trd.2014.11.012.

[4] K. Han, G. Eve, and T. L. Friesz, "Computing Dynamic User Equilibria on Large-Scale Networks with Software Implementation," *Networks Spat. Econ.*, vol. 19, no. 3, pp. 869–902, Sep. 2019, doi: 10.1007/s11067-018-9433-y.

[5] M. A. Silgu *et al.*, "A Methodology to Consider Explicitly Emissions in Dynamic User Equilibrium Assignment," 2022, pp. 362–369.

[6] T. L. Friesz, D. Bernstein, T. E. Smith, R. L. Tobin, and B. W. Wie, "A Variational Inequality Formulation of the Dynamic Network User Equilibrium Problem," *Oper. Res.*, vol. 41, no. 1, pp. 179–191, Feb. 1993, doi: 10.1287/opre.41.1.179.

[7] T. L. Friesz and K. Han, "The mathematical foundations of dynamic user equilibrium," *Transp. Res. Part B Methodol.*, vol. 126, pp. 309–328, Aug. 2019, doi: 10.1016/j.trb.2018.08.015.

[8] "Istanbul Technical University 2021 - 2048 Climate Action Plan (CAP)," Istanbul, 2021. [Online]. Available: https://sustainability.itu.edu.tr/sustainability-office/climate-action-plan.

[9] M. Barth and K. Boriboonsomsin, "Real-World Carbon Dioxide Impacts of Traffic Congestion," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2058, no. 1, pp. 163–171, Jan. 2008, doi: 10.3141/2058-20.

# Population Synthesis using Deep Generative Models for Agent-based Simulations

Sangkuk Kim[1], Simon Oh[1*]

[1]*Department of Mobility Science and Engineering, Korea University*
*2511 Sejong-ro, Sejong 30019, Republic of Korea*
[*]*Corresponding author: simonoh@korea.ac.kr*

## 1. Introduction

Agent-based modeling (ABM) has been an important tool for both researchers and practitioners in various fields, including transportation planning. It captures the interactions between individuals and provides a simulation environment to test transportation and mobility-related policies based on the decisions of millions of agents within a transportation system (Bowman and Ben-Akiva (2001)). It is crucial to provide a realistic and representative set of individuals or households (HH) as a key input for ABM. This set of agents is called a synthetic population, which is often formulated as a fitting problem over different combinations of socio-demographic attributes based on survey data (Muller and Ivt (2011); Harland et al. (2012)).

Early studies on population synthesis focused on iterative proportional fitting (IPF) (Beckman et al. (1996); Pritchard and Miller (2012); Zhu and Ferreira (2014)), simulation-based approaches (Farooq et al. (2013)), Bayesian networks (Sun et al. (2018)), Hidden Markov Models (HMM) (Saadi et al. (2016)). And these methods have been pointed out with issues on scalability, algorithmic complexity, dimensionality, sampling and structural zeros (Borysov et al. (2019); Garrido et al. (2020); Kim and Bansal (2023)). To tackle the issues, researchers adopted deep generative models (DGMs) such as Variational Autoencoders (VAEs) (Kingma and Welling (2013)), Generative Adversarial Networks (GANs) (Goodfellow et al. (2014)) in generating the synthetic population. This research applies DGMs (VAE, GAN, Wasserstein GAN (WGAN), and Conditional Tabular GAN (CTGAN)) to the travel survey data collected from the metropolitan region of two adjacent cities, Sejong and Daejeon in Korea, and verifies the generated result based on the population distribution and commuting patterns.

## 2. Method

### 2.1 Data

The household travel survey (HTS) data was collected in the Daejeon and Sejong region, which has a population of around 2 million and covers an area of 1,000 square kilometers in 2016 (publicly available on KTDB (Korea Transport Database)). We have sorted out key attributes (including addresses of activity locations defined at an administrative zonal level) that are relevant to the travel behavior modeling as in Table 1.

Table 1. List of socio-demographic attributes

| # | Attributes | Type | Details |
|---|---|---|---|
| 1 | Region | Categorical | Region of respondents |
| 2 | HH Size (HHS) | Int | Size of a household |
| 3 | Kids in household (HHK) | Int | Number of Child under 5 in a HH |
| 4 | House type (HHT) | Categorical | 1: Apartment, 2: Multiplex, 3: Multi-family, 4: Single-family, 5: Condo, 6: Other |
| 5 | Income of HH (INC) | Categorical | 1: <1M, 2: 1-2M, 3: 2-3M, 4: 3-5M, 5: 3-10M, 6: >10M (unit: KRW) |
| 6 | Car Ownership (OWN) | Binary | 0: with, 1: without |
| 7 | Age | Categorical | [0, 5), [5, 10), [10, 15), ⋯ , [45,50), ⋯, [90,95), [95,100), [100 |
| 8 | Gender | Binary | 0: Male, 1: Female |

| | | | |
|---|---|---|---|
| 9 | Car license (LIC) | Binary | 0: with, 1: without |
| 10 | Students (STU) | Categorical | 1: Preschool, 2: Elementary, 3: Middle/High, 4: Undergraduate/Graduate, 5: Etc. |
| 11 | Job category (WTY) | Categorical | 1: Professional, 2: Service, 3: Retails, 4: Managers and Administration, 5: Agriculture, 6: Laborer, 7: housewife, 8: Unemployed (Student), 9: Other |
| 12 | Work from home (WFH) | Binary | 0: yes, 1: no |
| 13 | Num. of working days (WDY) | Categorical | 1: > 6 days, 2: 5days, 3: 3-4 days, 4: 1-2 days, 5: 0 |
| 14 | Address of HH | Categorical | Defined at an administrative zonal level |
| 15 | Address of Work | Categorical | Defined at an administrative zonal level |
| 16 | Address of Edu | Categorical | Defined at an administrative zonal level |

## 2.2 DGMs: VAE, GAN, WGAN, CTGAN

The VAE synthesizes population data by first encoding real individual-level records into a continuous latent space. During training, the model learns to approximate the data distribution by minimizing the reconstruction error and the Kullback-Leibler Divergence between the encoded latent distribution and a prior distribution. After training, synthetic individuals are generated by sampling from the latent space and decoding them back into the original data space. GANs generate synthetic population data through a game-theoretic framework involving a generator and a discriminator. The generator attempts to produce realistic individual records, while the discriminator learns to distinguish between real and synthetic records. Through adversarial training, the generator learns to approximate the true data distribution. WGANs improve upon the standard GAN by using Wasserstein distance as a more stable and meaningful measure of distributional difference. In population synthesis, WGANs allow for more robust training and better convergence by replacing the discriminator with a critic that scores the "realness score" of samples. This approach is particularly effective for high-dimensional population data with rare attribute combinations, as it reduces mode collapse and ensures diversity among the generated individuals. CTGAN is specifically designed for generating tabular data with mixed data types, including highly imbalanced categorical variables. CTGAN uses a conditional vector to guide the generation process, allowing the model to produce synthetic individuals with specific demographic characteristics. It applies mode-specific normalization for continuous variables and Gumbel-Softmax sampling for categorical variables, enabling accurate replication of both common and rare combinations of population attributes. The architectures of each network are summarized in Table 2.

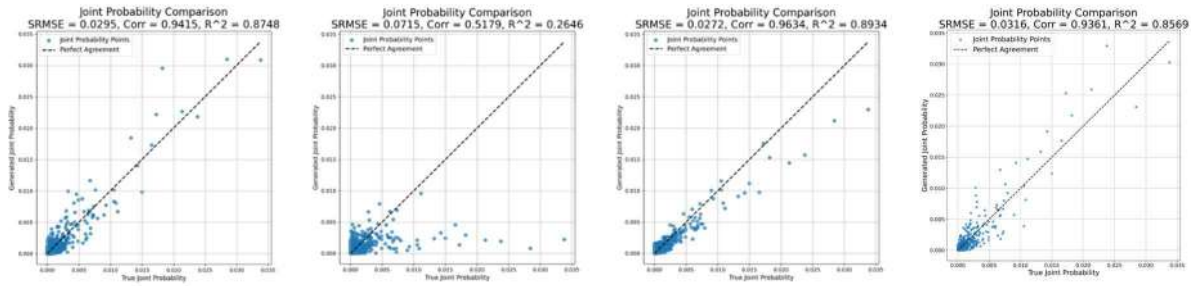
Table 2. Hyperparameters settings

| | VAE | GAN | WGAN | CTGAN |
|---|---|---|---|---|
| **Latent dim** | 128 | 128 | 128 | 128 |
| **Hidden dim** | 1024 | 1024 | 1024 | 256 |
| **Activation** | ReLU | ReLU | ReLU | LeakyReLU |
| **Learning rate** | 5e-5 | Generator: 1e-4<br>Discriminator: 1e-4 | Generator: 1e-4<br>Discriminator: 1e-4 | Generator: 2e-4<br>Discriminator: 2e-4 |
| **Optimizer** | Adam | Adam | RMSprop | Adam |

# 3. Results

## 3.1 Model Performances

Figure 1 shows the generated results for joint distribution for key attributes (Income, Job category, and Locations of household/work). One can notice VAE, WGAN, and CTGAN show a good agreement with the estimation from real data (0.0272-0.0316 of SRMSE), while the standard GAN results in higher error (standardized root mean square error, SRMSE) and less correlation measure. This result is also consistent with earlier works (Garrido et al. (2020)). Also note that details of marginal distributions are presented in Figure 2.




(a) VAE (b) GAN (c)WGAN (d) CTGAN

Figure 1. Comparison of multidimensional joint distributions


Figure 2. Marginal distribution of each attribute (Example: WGAN)

## 3.2 Verification on Travel Patterns

For the verification purpose, we compare the travel patterns for different types of trips. The trips in the study area can be categorized into internal trips (within a city), inter-city trips (between two cities), and

incoming/outgoing trips from/to other regions. Comparing the generated and observed trips for each OD pair (Table 3), overall travel patterns from two data are consistent. Accordingly, large portions of trips are made within Daejeon (~80%) and Sejong city (~10%) while inter-city, incoming/outgoing trips are captured in small.

Table 3. Generated and observed OD table (Example: WGAN)

| (unit: %) | Daejeon | | Sejong | | E1 | | E2 | | E3 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | WGAN | HTS | WGAN | HTS | WGAN | HTS | WGAN | HTS | WGAN | HTS |
| **Daejeon** | 77.93 | 84.18 | 4.70 | 0.66 | 0.23 | 0.07 | 0.88 | 2.13 | 2.66 | 0.09 |
| **Sejong** | 0.97 | 0.35 | 9.74 | 10.06 | 0.02 | 0 | 0.02 | 0.11 | 0.03 | 0 |
| **E1** | 0.28 | 0.39 | 0.44 | 0.43 | - | - | - | - | - | - |
| **E2** | 0.06 | 0.90 | 0.44 | 0.55 | - | - | - | - | - | - |
| **E3** | 0.72 | 0.09 | 0.33 | 0 | - | - | - | - | - | - |

Note: E1 (Seoul- Gyeonggi region), E2 (Chungcheong region), E3 (Other region)

## 4. Summary

This paper presented how DGMs can handle high-dimensional attributes, including socio-demographic and geographic information on individual activities, within a large-scale metropolitan area. The current research stream focuses on: (i) the application of recent AI models—such as large language models—to better represent socio-demographic descriptions in local or transportation-specific contexts, and (ii) the evaluation of activity patterns using agent-based modeling simulations for a given synthetic population. These enhancements aim to improve the credibility of agent-based simulations in replicating real-world traffic network dynamics.

## References


1. Bowman, J. L., & Ben-Akiva, M. E. (2001). Activity-based disaggregate travel demand model system with activity schedules. Transportation research part a: policy and practice, 35(1), 1-28.
2. Müller, K., & Axhausen, K. W. (2010). Population synthesis for microsimulation: State of the art. Arbeitsberichte Verkehrs-und Raumplanung, 638.
3. Harland, K., Heppenstall, A., Smith, D., & Birkin, M. H. (2012). Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques. Journal of Artificial Societies and Social Simulation, 15(1).
4. Beckman, R. J., Baggerly, K. A., & McKay, M. D. (1996). Creating synthetic baseline populations. Transportation Research Part A: Policy and Practice, 30(6), 415-429.
5. Pritchard, DR., & Miller, EJ. (2012). Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. Transportation, 39, 685-704.
6. Zhu, Y., & Ferreira, J. (2014). Synthetic population generation at disaggregated spatial scales for land use and transportation microsimulation. Transportation Research Record, 2429(1), 168-177.
7. Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013). Simulation based population synthesis. Transportation Research Part B, 58, 243-263.
8. Sun, L., Erath, A., & Cai, M. (2018). A hierarchical mixture modeling framework for population synthesis. Transportation Research Part B, 114, 199-212.
9. Saadi, I., Mustafa, A., Teller, J., Farooq, B., & Cools, M. (2016). Hidden Markov Model-based population synthesis. Transportation Research Part B: Methodological, 90, 1-21.
10. Borysov, S. S., Rich, J., & Pereira, F. C. (2019). How to generate micro-agents? A deep generative modeling approach to population synthesis. Transportation Research Part C, 106, 73-97.
11. Garrido, S., Borysov, S. S., Pereira, F. C., & Rich, J. (2020). Prediction of rare feature combinations in population synthesis: Application of deep generative modelling. Transportation Research Part C, 120, 102787.
12. Kim, E. J., & Bansal, P. (2023). A deep generative model for feasible and diverse population synthesis. Transportation Research Part C, 148, 104053.

# Run choice modelling in High Speed Rail (HSR) services from day-to-day ticket evolution

Giuseppe Musolino[1] and Domenico Sgro[1]

[1] DIIES, Università Mediterranea of Reggio Calabria – 89100 Reggio Calabria, Italy.
domenico.sgro@unirc.it

## 1. Introduction

There are more than 120,000 kilometres of planned, operative and under construction High-Speed Rail (HSR) lines worldwide, and in particular up to 60,000 kilometres of lines in operation (UIC, 2023). Passengers travelling with the HSR services in Europe have grown from the 15 billion passenger-kilometres in 1990 to 124 billion in 2016 (EU, 2024).

High Speed Rail (HSR) lines generate effects on passengers' travel demand which are generally segmented into the three components (Ben Akiva at al., 2010; Givoni and Dobruszkes, 2013; Cascetta and Coppola, 2014; Russo et al., 2023): diverted, from other modes or other rail services; induced, in terms of trip frequency and destination generated by high level of service (e.g. reduction of travel time) due to HSR; economy-based, in terms of trip frequency and destination generated by the economy in the cities, or areas, served by HSR. Several studies were conducted on models for the estimation of diverted demand, due to the attempt of capturing the demand diversion from the air mode and conventional rail services, as the entering of HSR caused competition in the inter-modal (e.g., mode choice) and in intra-modal levels (e.g., service, company and run choice) (Givoni and Dobruszkes, 2013; Cascetta and Coppola, 2012).

As far as concerns diverted demand, fare structures of HSR services play an important role in the intra-modal and inter-modal competition of the intercity passenger mobility. In broad terms, it is possible to classify the fare structures in static and dynamic, depending on the variation (or not) over time. On one hand, static fares are mainly referred to the conventional rail and bus services. On the other hand, dynamic fares historically characterized the air services, and, in the last decades, this typology of structure has been using in the High Speed Rail (HSR) services (Russo et al, 2024).

Two important themes are present in the intra-modal competition:

- the run choice model;
- the evolution of fares in the days before the journey.

The two themes are studied separately in the literature. The possibility of using innovative methods for collecting data relating to the tickets evolution allows the two themes to be combined, providing an important contribution for the development of sustainable transport policies in line with the goals and targets of Agenda 2030.

The present work aims to put together the two above research lines that have been historically developed independently:

- the analysis and estimation of dynamic fares of HSR services;
- the development of run choice models to estimate users' behaviour in the choice of run in the schedule- based intercity transport services.

The research contribution concerns the definition of a framework for the identification of users' choice of HSR run, as well as the development of choice model in the above-mentioned dimension.

The following sections are organized as follows. Section 2 briefly sketches the characteristics of the proposed framework. Section 3 regards the results and the discussion about the experimental testing of the proposed framework on a case study. Section 4 contains some remarks and research perspectives.

This work is aimed to support transport planners and decision-makers to inform sustainable transport policies in the evaluation of investment in HSR lines and services, by means of methodological and modelling tools to assess actual and potential HSR travel demand.

## 2. Proposed framework

The proposed framework is composed by two main phases. The former phase deals with the identification of user's choices by analysing the patterns of dynamic fares of HSR services. The latter phase deals with the specification, calibration and validation of a run choice model.

### *2.1. Identification of user's choice*

The hypothesis considered is that the users' choice of run can be identified in relation to the day-to-day ticket evolutions observed between two consecutive days, considering when the user buys the ticket in relation to the day of trip.

The monitoring of ticket evolution is an innovative data collection because it has a huge quantity of information that needs specific collection, but gives the possibility to update the model in a very cheap, repeatable and easy way.

The phase is sub-divided into three steps.

Step 1. Ticket coding, where every fare structure of every company is described by means of a vector;

Step 2. Day-to-day ticket costs' updating, where the day-to-day fare variations are compared;

Step 3. Identification of number of ticket mutations for each run in each day and association to user's choice of a run.

### *2.2. Run choice model*

The development of the run choice model implied two elements: the choice set and the run choice model building.

The choice set building at path choice dimension has been widely studied for the literature (Ben-Akiva et al.; 1984; Russo and Vitetta, 1995; Morikawa, 1996; Ben-Akiva et al., 2002; Cascetta et al., 2002). The case considered in this paper concerns, without affecting generality, the choice set of runs is generated taking into account the Desired Arrival Time (DAT) of a user at destination.

The choice models are based on the random utility (RU) (Ben-Akiva and Lerman, 1985; Cascetta, 2013; Train 2009; McFadden, 2001; Ben-Akiva, McFadden, Train, 2019). One of the main assumptions concerns that a user chooses a run, among the set of perceived alternatives, which maximize his/her perceived utility associated to each run**.** The calibration phase allows to obtain the estimates vector of models' parameters, according to users' choices. It is possible to estimate the vector of parameters with different method, for instance the maximum likelihood, which returns the values of the unknown parameters, the latter maximize the probability of observing the users' choices.

## 3. Results

The proposed procedure has been tested on the HSR relationship Rome-Milan, which is the main HSR line in the national rail network. The line belongs to traditional basic network for a short infrastructural stretch and belongs to the High Speed/High Capacity basic network for the Florence-Milan infrastructural stretch equipped with ERMTS system. The line is completely electrified with double track (www.rfi.it). The frequency of HSR services is currently 45 runs/day for Trenitalia and 33 runs/day for Italo Nuovo Trasporto Viaggiatori (NTV).

A survey has been conducted aimed at investigating the evolution over time of ticket evolution, by considering a set of days of trip and days of ticket purchasing. The survey period was the month of August of two consecutive years: 2022 and 2023.

The application of the proposed framework allowed to identify a sample of users' choices of runs and to calibrate of disaggregate run choice model attributes: a fare, or ticket cost associated to a run; penalty, or time interval between desired arrival time and scheduled arrival time; on-board travel time at the railway stations; number of intermediate stops between the departure and arrival railway stations. The sign of the calibrated parameter are in line with the nature of the attributes.

## 4. Discussion and conclusion

The research proposes an original framework for the identification of users' run service choice from the collection of data related to the day-to-day evolution of ticket costs a sample of HSR services.

The proposed procedure has been validated for along the relationship Rome-Milan (Italy), by the specification-calibration-validation of a run choice model belonging to the class of RU models. The aim of the validation has been to verify the possibility to ground the proposed procedure into a specified and calibrated choice model.

The procedure presents the following advantages and limitations today. On the side of the advantages, the procedure ensures the identification of users' choices without the execution of surveys on users that are quite expensive in terms of time and cost. On the side of the limitations, the procedure does not allow the emerging of the number of users that determined the change of the ticket cost from one day to another.

The conclusion is that the procedure has limitations that need to be overcame in the future but, nevertheless, the first results reported in this study seem encouraging. Future developments regard the extension of the data-base related to number of users and the development of stochastic route choice models instead of the wardropian ones.

## References

High Speed Lines In The World (Summary), Updated 1st October 2023. https://uic.org/passenger/highspeed/article/high-speed-data-and-atlas. Accessed 26 Feb 2025.

Europeran Union, 2024. A European high-speed rail network: not a reality but an ineffective patchwork. Special Reporto n.19 (2024).

Ben-Akiva M., Cascetta E., Coppola P., Papola A., & Velardi V. (2010). "High Speed Rail demand forecasting in a competitive market: the Italian case study" Proceedings of the World Conference of Transportation Research (WCTR), Lisbon, Portugal.

Givoni, M., Dobruszkes, F. (2013) A Review of Ex-Post Evidence for Mode Substitution and Induced Demand Following the Introduction of High-Speed Rail. Transport Re-views. 33, 720–742. https://doi.org/10.1080/01441647.2013.853707.

Cascetta E., Coppola P. (2014). High speed rail (HSR) induced demand models. Procedia-Social and Behavioural Science, 111, 147-156.

Russo, F., Sgro, D., Musolino, G.(2023). Sustainable Development of Railway Corridors: Methods and Models for High Speed Rail (HSR) Demand Analysis. In: Gervasi, O., et al. Computational Science and Its Applications–ICCSA 2023 Workshops. Lecture Notes in Computer Science, vol 14110. Springer, Cham. https://doi.org/10.1007/978-3-031-37123-3_36.

Cascetta E., Coppola P. (2012). An elastic demand schedule-based multimodal assignment model for the simulation of high-speed rai(HSR) systems, The Association of European Operational Research Societies2012,1,3-27.

Russo, F., Sgro, D., Musolino, G. (2024). Dynamic Structure of Fares for High Speed Rail Services. In: Gervasi, O., Murgante, B., Garau, C., Taniar, D., C. Rocha, A.M.A., Faginas Lago, M.N. (eds) Computational Science and Its Applications – ICCSA 2024 Workshops. Lecture Notes in Computer Science, vol 14822. Springer, Cham. https://doi.org/10.1007/978-3-031-65318-6_13.

Ben-Akiva M. and Lerman S. R. (1985). Discrete choice analysis: theory and application to travel demand. Cambridge, MA: MIT Press.

Cascetta E. (2013) Transportation systems engineering: theory and methods, Springer Science & Business Media.

Train, K.E., 2009. Discrete choice methods with simulation, 2nd ed. ed. Cambridge university press, Cambridge.

McFadden, D., 2001. Disaggregate Behavioral Travel Demand's RUM Side - A 30-Year Retrospective.

Ben-Akiva, M., McFadden, D., Train, K., 2019. Foundations of Stated Preference Elicitation: Consumer Behavior and Choice-based Conjoint Analysis. FNT in Econometrics 10, 1–144. https://doi.org/10.1561/0800000036.

# Accelerating convergence in Capacitated Schedule-Based Dynamic Transit Assignments

Lory Michelle Bresciani Miristice[1]*, Guido Gentile[1], Daniele Tiddi[2], Lorenzo Meschini[2]

[1]Department of Civil, Building and Environmental Engineering,
University of Rome La Sapienza, Rome, Italy

[2]PTV Group Sistema, Rome, Italy

lorymichelle.brescianimiristice@uniroma1.it, guido.gentile@uniroma1.it,
daniele.tiddi@ptvgroup.com, lorenzo.meschini@ptvgroup.com
* Corresponding author

*10th International Symposium on Dynamic Traffic Assignment (DTA 2025)*

June 17, 2025

Keywords: dynamic user equilibrium, congested public transport, passenger overcrowding and queuing, hyper-paths, fixed-point algorithms.

## Abstract

Public transportation is essential for sustainable urban mobility, yet overcrowding and queuing often degrade service quality. To improve short-term passenger flow forecasts, this paper introduces a schedule-based Dynamic Transit Assignment (DTA) framework that integrates fail-to-board probabilities and hyperpath-based strategic route choices. Unlike conventional models that define deterministic equilibrium using one-to-many network loading maps, we reformulate it as a one-to-one fixed-point problem, leveraging a Gradient Projection (GP) approach to enhance numerical stability and convergence speed.

To further address non-linearities from capacity constraints and fail-to-board events, we incorporate an Adaptive Trust Contraction (TC-A) algorithm, which dynamically adjusts step sizes for greater computational efficiency. Numerical experiments on small- and medium-scale networks show that the GP plus TC-A framework achieves linear convergence and significantly outperforms the Method of Successive Averages (MSA) in both precision and speed. These results underscore its potential for real-time transit optimization under high-demand conditions.

## 1 Introduction

Accurate simulation of passenger flows in dynamically congested transit systems is essential for optimizing performance and enhancing the passenger experience. By forecasting these flows promptly, transit agencies can proactively manage disruptions and reduce congestion, thereby alleviating overcrowding and shortening wait times.

DTA models are instrumental in replicating how passengers choose routes and interact within public transit networks, capturing the evolving patterns of travel demand and congestion throughout the day. In macroscopic models, individual travellers are aggregated into flows, offering a comprehensive view of system dynamics while improving computational efficiency for large-scale networks (Bellei *et al.*, 2005).

*Schedule-based DTA models* refine the macroscopic approach by incorporating the temporal dimension directly into the network topology via diachronic graphs (Nuzzolo & Russo, 1998).

This technique assumes that passengers possess full knowledge of run timetables, a key difference from hyperpath-based optimal strategies (Nguyen & Pallottino, 1988, Spiess & Florian, 1989) that rely solely on the distribution of line frequencies for route choice. Moreover, to accurately reflect the complexities of passenger decision-making under uncertainty, these models must incorporate fail-to-board probabilities (Hamdouch & Lawphongpanich, 2008). This addition captures scenarios in which capacity constraints prevent boarding, thereby introducing an alternative formulation of hyperpaths as detailed in Gentile & Noekel (2016).

A robust comparison of design scenarios necessitates high algorithmic precision, which can be attained by accelerating convergence to an equilibrium between demand and supply. However, the extensive size of diachronic graphs and the slow convergence of conventional fixed-point algorithms, such as the MSA, pose significant computational challenges, thereby hindering real-time short-term forecasts of passenger volumes on vehicles and at stops.

## 2 Methodology

To capture passenger flows under congested conditions, we reformulate the DTA as a fixed-point problem. The iterate, denoted by $\mathbf{p}$, represents the vector of arc conditional probabilities (or destination-specific splitting rates). The mapping function, $\mathbf{p} = [\mathbf{r}, \mathbf{x}]$ encapsulates the processes of demand loading, network congestion, and route choice. Specifically, $\mathbf{r}$ corresponds to the route choices made by passengers at standard nodes, while $\mathbf{x}$ represents the outcomes at stops, i.e., the fail-to-board probabilities arising from capacity limitations.

Following the implicit path enumeration approach proposed in Gentile (2016), we define the arc flows $mathbfq$ as the result of loading the origin-destination demand $\mathbf{d}$ onto the network based on the given arc conditional probabilities $\mathbf{p}$:

$$\mathbf{q} = q(\mathbf{p}; \mathbf{d}). \tag{1}$$

Arc flows $\mathbf{q}$ are used to compute the arc costs $\mathbf{c}$ and the fail-to-board probabilities $\mathbf{x}$. These quantities account for passenger congestion (i.e, overcrowding and queuing) based on the supply characteristics $\mathbf{s}$ (primarily the capacity constraints, as line speeds are inherently captured within the diachronic graph structure):

$$\mathbf{c} = c(\mathbf{q}; \mathbf{s}), \tag{2}$$

$$\mathbf{x} = x(\mathbf{q}; \mathbf{s}). \tag{3}$$

Local (deterministic) route choices, $\mathbf{r}$, are determined via shortest hyper-tree computations for each destination. These computations yield the arc weights $\mathbf{w}$, which represent the cost to reach the destination using a given arc, conditional on being at its tail:

$$\mathbf{w} = w(\mathbf{c}, \mathbf{x}), \tag{4}$$

$$\mathbf{r} \in r(\mathbf{w}). \tag{5}$$

When two or more local alternatives exhibit the same minimum cost, there exist infinitely many user distributions that satisfy Wardrop's principles (Wardrop, 1952). Consequently, the deterministic mapping $r(\mathbf{w})$ yields a set of points. Finally, the arc conditional probabilities are updated using a convergence method, and the process is repeated iteratively until the convergence criteria are met.

Traditionally, the problem is formulated by considering a fixed-point mapping defined by the *one-to-many* operator provided by the deterministic network loading map described above. This operator is given by the composition of equations 1, 2, 3, and 4:

$$\mathbf{p} \in [r(\hat{w}(\mathbf{p})), x(q(\mathbf{p}))], \tag{6}$$

where

$$\hat{w}(\mathbf{p}) = w(c(q(\mathbf{p})), x(q(\mathbf{p}))). \tag{7}$$

The issue is that the typical all-or-nothing assignment to the shortest hyper-paths, denoted $r^*(\mathbf{w})$, which is adopted to implement the deterministic network loading map, often leads to solutions that deviate from equilibrium. The Method of Successive Averages (MSA) addresses these fixed-point problems by averaging the solutions obtained from applying the map. However, it suffers from slow convergence due to decreasing step sizes:

$$\mathbf{p}^{(i+1)} = \mathbf{p}^{(i)} + \frac{1}{1+\gamma \cdot i} \cdot \left( f^*(\mathbf{p}^{(i)}) - \mathbf{p}^{(i)} \right), \tag{8}$$

$$f^*(\mathbf{p}) = [r^*(\hat{w}(\mathbf{p})), x(q(\mathbf{p}))] \tag{9}$$

where $\mathbf{p}^{(i)}$ denotes the solution at iteration $i$ and $\gamma < 1$ is a reduction factor used to decrease the weight of older iterates and speed up convergence.

Nonetheless, the introduction of the fail-to-board mechanism adds further non-linearities, increasing the model's complexity and causing oscillations in cost updates, which in turn further slow convergence.

To overcome these limitations, we adopt the GP method as fixed-point mapping. Equilibrium is reached when the difference between the route choices $\mathbf{r}$ and the arc weights $\mathbf{w}$ (scaled by a factor $\sigma$), when projected onto the feasible region defined by flow conservation, yields the same point. Formally, this condition is expressed as:

$$\mathbf{p} = f(\mathbf{p}) \equiv [\mathrm{Proj}_R \left( r(\hat{w}(\mathbf{p})) - \sigma \cdot \hat{w}(\mathbf{p}) \right), x(q(\mathbf{p}))], \tag{10}$$

where the arc weights $\hat{w}(\mathbf{p})$ are interpreted as the gradient of the Beckmann integral corresponding to our local deterministic model. In other words, they represents the objective function's gradient in an optimization program whose solution is an equilibrium. Due to non-separable congestion effects in transit networks, the traditional Beckmann formulation is not directly applicable, and the equilibrium conditions can instead be expressed as a Variational Inequality problem. In this paper, however, we prefer to introduce directly the above equivalent fixed-point formulation.

Pure GP algorithms may still face challenges with convergence, which is why the step size is further scaled by an MSA-like factor, proportional to the number of iterations (Gentile, 2016).

We enhance the GP approach with the TC-A algorithm (Gentile *et al.*, 2024). TC-A applies the feasible direction method (i.e., iterating a step along a descent direction) for nonlinear optimization on convex sets, where the objective function is the sum of squared residuals. It operates under the assumption that the residual $f(\mathbf{p}) - \mathbf{p}$ indicates a descent direction with respect to the sum of squared residuals.

TC-A determines the step size $\alpha$ from $\mathbf{p}$ to $f(\mathbf{p})$ by increasing it by a factor of $1+\gamma_1 > 1$ if the current iteration improves the sum of squared residuals compared to the previous one; conversely, if no improvement is observed, the step size is decreased by a factor of $1 + \gamma_2 > 1$. Both the current iterate and the search direction are continuously updated, under the assumption that the method will converge for some fixed, though initially unknown, step size. The adjustment rule for $\alpha^{(i)}$ is given by:

$$\alpha^{(i)} = \begin{cases} \min\left(1, \alpha^{(i-1)} \cdot (1+\gamma_1)\right), & \text{if } y^{(i)} < y^{(i-1)} \\ \dfrac{\alpha^{(i-1)}}{1+\gamma_2}, & \text{if } y^{(i)} \geq y^{(i-1)} \end{cases} \tag{11}$$

where $y^{(i)}$ denotes the sum of squared residuals at iteration $i$. Suggested values for the adjustment factors are $\gamma_1 = 0.1$ and $\gamma_2 = 0.5$.

## 3 Results and Discussion

We evaluated the GP method with TC-A against MSA using a relative gap convergence criterion. Initial tests were conducted on a simple network with 60 passengers arriving at 1 per minute over one hour. The network comprised a single transit line connecting two stops, with four departures every 20 minutes starting at minute 11. In an unconstrained scenario (20-passenger capacity per run), passengers distributed as expected: 11 boarded the first run, 20 each on the second and third, and 9 on the last. Without boarding delays, both MSA and TC-A converged in one iteration. To simulate congestion, we imposed capacity limits of 18, 15, and 10 passengers per run, representing slight, moderate, and severe congestion. Figure 1 illustrates the convergence trends, highlighting TC-A's superior adaptability. Its adaptive step sizing effectively mitigates non-linearities from fail-to-board probabilities, reducing oscillations and computation time.

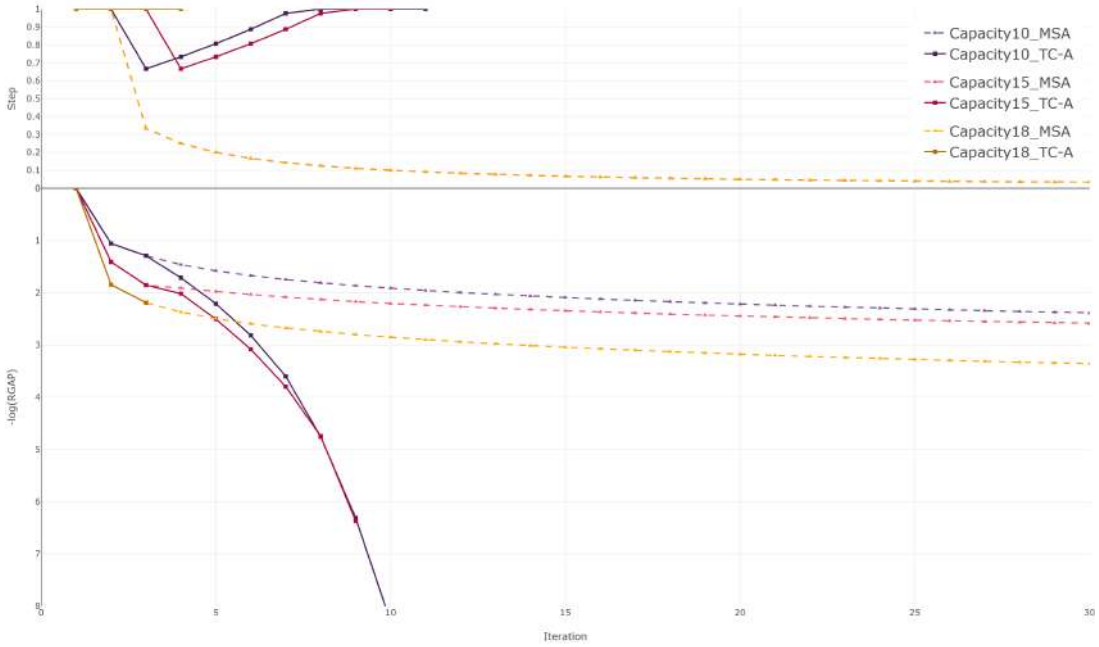



Figure 1 – *Convergence and step size trends of MSA and TC-A under different congestion levels. Note that the step size trends for congested MSA (capacities 18, 15, and 10) overlap.*

Under an 18-passenger capacity, fail-to-board events emerged as passengers on the second and third runs queued for later departures. TC-A converged in 3 iterations, while MSA required 15 iterations to reach a relative gap of $10^{-3}$. With a 15-passenger capacity, queuing intensified across the second, third, and fourth runs, causing 4 passengers to be unserved. TC-A converged in 9 iterations, whereas MSA required 80 iterations. Under severe congestion (10-passenger capacity), 20 passengers remained unserved. TC-A reached convergence in 10 iterations, while MSA exceeded 100 iterations to reach a relative gap of $10^{-3}$.

We further assessed TC-A on a large-scale network simulating Rome's morning peak (6–10 A.M.), comprising 60,000 passengers, 50 zones, 6,748 transit stops, and 1,417 transit routes. The assignment graph featured 2.33 million arcs and 800 capacitated vehicle runs. Both methods reached a relative gap of $10^{-2}$ in 14 seconds, but TC-A required only 6 iterations compared to 11 for MSA. Over 100 iterations, TC-A achieved a gap of 0.001 in 241 seconds, whereas MSA reached 0.003 in 122 seconds. TC-A also showed a more linear and stable convergence trend, consistent with its gradient-based formulation. These results highlight TC-A's potential, particularly under capacity constraints, though further algorithm refinement and testing on operational networks are needed to validate its practical applicability.

# References


Bellei, Giuseppe, Gentile, Guido, & Papola, Natale. 2005. A within-day dynamic traffic assignment model for urban road networks. *Transportation Research Part B: Methodological*, **39**(1), 1–29.

Gentile, Guido. 2016. Solving a Dynamic User Equilibrium model based on splitting rates with Gradient Projection algorithms. *Transportation Research Part B: Methodological*, **92**(10), 120–147.

Gentile, Guido, & Noekel, Klaus. 2016. Modelling Public Transport Passenger Flows in the Era of Intelligent Transport Systems.

Gentile, Guido, Eldafrawi, Mohamed, & Bresciani Miristice, Lory Michelle. 2024. *Applying the Trust Contraction Algorithm for Solution of Stochastic User Equilibrium.* Presented at SIDT XXXVI Scientific Seminar, Cagliari, Italy, June 12–14, 2024.

Hamdouch, Younes, & Lawphongpanich, Siriphong. 2008. Schedule-based transit assignment model with travel strategies and capacity constraints. *Transportation Research Part B: Methodological*, **42**(7), 663–684.

Nguyen, S., & Pallottino, S. 1988. Equilibrium traffic assignment for large scale transit networks. *European Journal of Operational Research*, **37**, 176–186.

Nuzzolo, A., & Russo, F. 1998. A dynamic network loading model for transit services. *Proceedings of Tristan III.*

Spiess, Heinz, & Florian, Michael. 1989. Optimal strategies: A new assignment model for transit networks. *Transportation Research Part B: Methodological*, **23**(2), 83–102.

Wardrop, John Glen. 1952. Road Paper. Some theoretical aspects of road traffic research. *Proceedings of the Institution of Civil Engineers*, **1**(3), 325–362.

# On the Control of Connected and Automated Vehicles as Moving Bottlenecks in Freeway Traffic

Nikolas Sacchi and Antonella Ferrara

***Abstract*— This work explores the potential of controlling at microscopic scale Connected and Automated Vehicles (CAVs) forming moving bottlenecks (MBs) to positively impact the macroscopic freeway traffic dynamics. To achieve this, we propose a neural network (NN)-based approach to design the individual CAV control system, which effectively addresses the traction control problem for the vehicle even when its model is not perfectly known. Specifically, two NNs are used to approximate the dynamics of each CAV, and an integral sliding mode (ISM) control strategy is employed to generate the necessary longitudinal force for the vehicles within the formation. This ensures anti-skid braking and anti-spin acceleration across all road conditions. Then, a direct benefit of controlling CAVs at the microscopic level is enhanced safety, even in adverse weather. Additionally, the formations of controlled CAVs function as artificial MBs within the traffic flow, helping to alleviate congestion and improve the overall traffic efficiency.**

***Index Terms*— Traffic Dynamics, Neural Networks, Traction Control**

## I. Introduction

Starting from the final decades of the previous century, a wide range of strategies have been developed to control freeway traffic, with the goal of minimizing congestion and improving the overall throughput of road traffic systems [1]–[4]. More recently, particular attention has been paid to developing control strategies that also aim to increase sustainability, hence minimizing factors such as emissions and fuel consumption [5], [6]. On the other hand, technology advancements in the automotive field have enabled the definition of the concept of connected and automated vehicles (CAVs), that is, vehicles characterized by a high level of automation and communication capability [7], [8]. The presence of CAVs can have a significant impact on freeway traffic dynamics, as discussed in [9], [10]. For this reason, in recent years, particular attention has been given to the development of strategies that rely on CAVs as active control elements within traffic systems [11]–[16].

Traction control (TC) is undoubtedly one of the most widely used driving assistance systems. Its primary objective is to improve vehicle performance during braking and acceleration, thereby enhancing stability and safety, even in challenging external conditions [17]. In general, the traction force produced at the vehicle wheels is directly related to the so-called wheel slip ratio, which, in turn, depends on the relative difference between the wheel angular speed and the vehicle longitudinal speed. Hence, the traction control problem can be treated as a wheel slip ratio control problem and is usually solved using a model-based approach. However, in the majority of cases, the vehicle model is affected by parameter uncertainties and disturbances, making it not sufficiently accurate for the traction controller design.

On dealing with perturbed and uncertain systems, an effective approach is sliding mode control (SMC), as it ensures the robustness of the controlled system when the system state evolves on the so-called sliding manifold [18], [19]. Specifically, SMC relies on a discontinuous control law, the amplitude of which is directly related to the magnitude of the uncertainty affecting the model. SMC has already been adopted to design traction control strategies [20]–[22]. Yet, this has been done under the assumption of perfectly known model of the considered vehicle.

To better account for the heterogeneity typically present in real-world road traffic, this work considers a more realistic scenario in which CAVs are integrated into the freeway traffic flow. In this scenario, each individual CAV may represent a different type of vehicle, and the CAV models could exhibit significant uncertainty.

To tackle this issue, inspired by [23], [24], we propose to design the CAVs traction controllers relying on a DNN-based Integral Sliding Mode (ISM) control approach. In particular, two neural networks (NNs) are utilized to approximate the dynamics of each CAV, while an integral sliding mode (ISM) control strategy is applied to generate the required longitudinal force for the vehicles creating the formation. This approach ensures effective anti-skid braking and anti-spin acceleration of the considered CAVs under all road conditions. The formation of CAVs moving in the traffic act as artificial moving bottlenecks (MBs). They can be controlled with the aim of regularizing the traffic flow, thus producing a beneficial effect on the macroscopic traffic. The proposed control architecture, schematized in Fig. 1, is assessed in simulation using the METANET model to represent the macroscopic dynamic of the traffic system.

## II. Multi-scale modeling

The aim of this section is to introduce the macroscopic model of the traffic, formulated with the METANET model, and the microscopic model of the individual CAVs.

Nikolas Sacchi and Antonella Ferrara are with the Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy (e-mail: *nikolas.sacchi01@universitadipavia.it*, *antonella.ferrara@unipv.it*). This work was supported by PRIN 2022 2022BB9JC9 (CUP: F53D23000540006) and PRIN PNRR P2022XJ9SX (CUP: F53D23010110001), M4C2, Investimento 1.1., European Union – Next Generation EU.

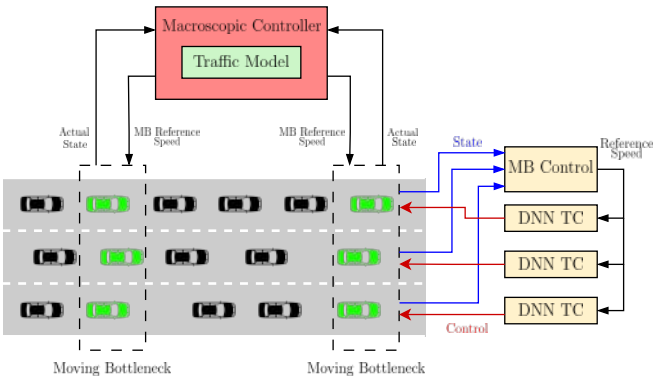



Fig. 1: The proposed control architecture.

### A. The traffic model with CAVs

Consider a freeway stretch characterized by $n_{\text{lanes}} \in \mathbb{N}_{\geq 1}$ lanes and on which both regular vehicles and $n_{\text{cav}} \geq n_{\text{lanes}}$ CAVs travel. Each CAV $c \in \{1, 2, \dots, n_{\text{cav}}\}$, the dynamics of which will be described in Section II-B, is characterized by speed $v^c$ and position $p^c = p^c(t_0) + \int_{t_0}^{t} v^c(\tau) d\tau$. Given a set of CAVs $\mathcal{C} \subseteq \{1, 2, \dots, n_{\text{cav}}\}$, it constitutes a MB if: 1) $|\mathcal{C}| = n_{\text{lanes}}$; 2) $|p^a - p^b| < \epsilon_{\text{mb}}$ for all $a, b \in \mathcal{C}$, with $a \neq b$, and where $\epsilon_{\text{mb}} \in \mathbb{R}_{>0}$ is a small design threshold; 3) the CAVs in $\mathcal{C}$ travel in a formation which occupies $n_{\text{lanes}}$ lanes, as depicted in Fig. 1.

Coherently with the METANET model [25], the freeway stretch can be subdivided into $N \in \mathbb{N}_{>0}$ sections, each identified by an index $i \in \mathcal{S}$, with $\mathcal{S} = \{1, 2, \dots, N\}$ and characterized by a length $L_i \in \mathbb{R}_{>0}$ [km]. Then, the traffic density in section $i$, without considering the CAVs, is denoted as $\rho_i \in \mathbb{R}_{\geq 0}$ [veh/km] and characterized by continuous time dynamics

$$\dot{\rho}_i(t) = \frac{q_{i-1}(t) - q_i(t) + r_i(t) - s_i(t)}{L_i} \tag{1}$$

where $q_i \in \mathbb{R}_{\geq 0}$, $r_i \in \mathbb{R}_{\geq 0}$, and $s_i \in \mathbb{R}_{\geq 0}$ represent the traffic flows leaving section $i$ towards section $i+1$, entering section $i$ from the on-ramp, and leaving section $i$ through the off-ramp, respectively. The presence of CAVs in the freeway stretch can be taken into account defining the overall section density

$$\bar{\rho}_i(t) = \rho_i(t) + \sum_{c=1}^{n_{\text{cav}}} \delta_i^c(t) o^c, \tag{2}$$

where $\delta_i^c \in \{0, 1\}$ indicates the presence of CAV $c$ in section $i$, while $o^c \in \mathbb{R}_{>0}$ is the occupancy [veh/km] of $c$ [26].

As for the mean speed $v_i \in \mathbb{R}_{\geq 0}$ [km/h] of section $i$, it is characterized by dynamics

$$\begin{aligned} \dot{v}_i(t) = \frac{V(\bar{\rho}_i(t)) - v_i(t)}{\alpha_1} + \frac{v_i(t)[v_{i-1}(t) - v_i(t)]}{L_i} + \\ - \frac{\alpha_2 \big(\rho_{i+1}(t) - \rho_i(t)\big)}{\alpha_1 L_i \big(\rho_i(t) + \alpha_3\big)} - \frac{v_i(t) r_i(t)}{L_i \big(\rho_i(t) + \alpha_3\big)}, \end{aligned} \tag{3}$$

where $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}_{>0}$ are model parameters, while $V : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is the so-called steady-state speed-density relation, computed as

$$V(\bar{\rho}_i(t)) = \bar{v}_i(t) \exp\left\{ -\frac{1}{\alpha_4} \left( \frac{\bar{\rho}_i(t)}{\rho_i^{\text{cr}}} \right) \right\}, \tag{4}$$

with $\rho_i^{\text{cr}} \in \mathbb{R}_{>0}$ being the critical density and $\alpha_4 \in \mathbb{R} > 0$ a model parameter. The time-varying term $\bar{v}_i \in \mathbb{R}_{>0}$ in (4) is the maximum speed that the vehicles can reach in section $i$ and it is defined as

$$\bar{v}_i(t) = \delta_i^{\mathcal{C}}(t) v_{\mathcal{C}} + \big(1 - \delta_i^{\mathcal{C}}(t)\big) v^{\text{ff}}, \tag{5}$$

where $\delta_i^{\mathcal{C}} \in \{0, 1\}$ denotes the presence of MB $\mathcal{C}$ in section $i$, $v_{\mathcal{C}} \in \mathbb{R}_{>0}$ is the speed of the MB, while $v^{\text{ff}} \in \mathbb{R}_{>0}$ is the free-flow speed.

### B. Vehicle modelling

Each CAV $c \in \mathcal{C}$ is characterized by measurable state $x^c = \begin{bmatrix} x_1^c & x_2^c & x_3^c \end{bmatrix}^\top = \begin{bmatrix} v^c & \omega_f^c & \omega_r^c \end{bmatrix}$, where $v^c \in \mathbb{R}$ is the longitudinal speed [m/s], while $\omega_f^c, \omega_r^c \in \mathbb{R}$ represent the angular speed [rad/s] of the front and rear wheels, respectively. Omitting the superscript $c$ for the sake of readability, the dynamics of the vehicle are defined as

$$\dot{x} = \begin{bmatrix} m^{-1} \Big( F_{f_x}(\lambda_f, F_{f_z}, \mu) + F_{r_x}(\lambda_r, F_{r_z}, \mu) - F_{\text{loss}}(x_1) \Big) \\ -J_f^{-1} R_f F_{f_x}(\lambda_f, F_{f_z}, \mu) + J_f^{-1} T_f \\ -J_r^{-1} R_r F_{r_x}(\lambda_r, F_{r_z}, \mu) + J_r^{-1} T_r \end{bmatrix} \tag{6}$$

where $m$ is the mass [kg], $T_f$ and $T_r$ are the input torques [Nm] of the front and rear wheels, $J_f$, $R_f$, $\lambda_f$ are the moment of inertia, the radius, and the slip ratio of the front wheel, respectively, while $J_r$, $R_r$, $\lambda_r$ represent the same quantities for the rear wheel. Then, $F_{f_z}$ and $F_{r_z}$ are the normal force on the front and rear wheels, computed by ignoring the influence of suspensions as $F_{f_z} = \frac{l_r mg - l_h m \dot{v}_x}{l_f + l_r}$ and $F_{r_z} = \frac{l_f mg + l_h m \dot{v}_x}{l_f + l_r}$, where $l_f$ is the distance from the front axle to the center of gravity, $l_f$ is the distance from the rear axle to the center of gravity, and $l_h$ is the height of the center of gravity, while $g$ is the gravity acceleration. Moreover, $F_{f_x}$ and $F_{r_x}$ are the front and rear longitudinal tire–road contact forces, and they are strictly dependent on the tire-road friction coefficient $\mu$ [27]. Finally, the slip ratios $\lambda_f$ and $\lambda_r$ are defined considering the relative difference between the wheel angular speed and the vehicle absolute speed, distinguishing between acceleration phase and braking phase. In particular, for $p \in \{f, r\}$, one has $\lambda_p = \frac{\omega_p R_p - v}{\omega_p R_p}$ if $\omega_p R_p > v$ and $\lambda_p = \frac{\omega_p R_p - v}{v}$ if $\omega_p R_p < v$.

## III. The Proposed Control Architecture

The aim of this section is to describe the proposed control architecture, depicted in Fig. 2. In particular, the architecture has a multi-scale structure. At a higher level, a model-based centralized controller generates the reference speed profiles for the MBs in the freeway stretch, while at a low level, references for the individual CAVs in each MB are computed and followed employing DNN-based traction control controllers.

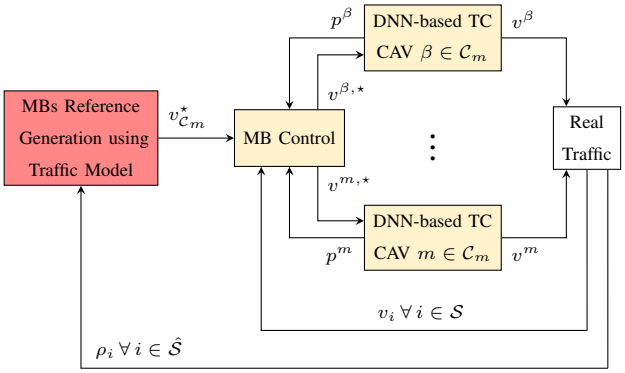



Fig. 2: Block Diagram of proposed control architecture for the MB $\mathcal{C}_m$. The red block represents the macroscopic scale control, the yellow ones the microscopic scale one, while the yellow represent the real traffic system.

### A. MB Reference Generation

Consider a stretch modeled as in Section II-A, in which $M \in \mathbb{N}_{>0}$ MBs are traveling. Then, the objective of the macroscopic scale controller, which works considering the macroscopic traffic model, is to generate a reference speed profile for each MB $\mathcal{C}_m$, with $m \in \{1, 2, \dots, M\}$, so that a beneficial effect on the traffic flow dynamics is obtained. To this end, an optimal control problem needs to be formulated. In this work, the objective is to minimize the total travel time (TTT) associated with the macroscopic traffic system. Hence, the optimal control problem can be formulated as follows:

$$\begin{aligned} \bar{v}^{\star} = \arg\min_{\bar{v}} & \int_{t_k}^{t_k+T} \sum_{i=1}^{N} L_i \rho_i(t) \\ \text{s.t.} \quad & \text{Traffic Dynamics (1)-(3)} \\ & 0 \leq \bar{v}_{\mathcal{C}_m} \leq v_i^m(t), \quad \forall m \in \{1, 2, \dots, M\} \end{aligned}$$

where $\bar{v} = \begin{bmatrix} v_{\mathcal{C}_1} & v_{\mathcal{C}_1} & \dots & v_{\mathcal{C}_M} \end{bmatrix}^{\top} \in \mathbb{R}^M$ is the vector containing the desired longitudinal speeds for the MBs, $v_i^m$ is the traffic mean speed of the section $i$ in which the MB $m$ is located, while $\bar{v}_{\mathcal{C}_m}$ denotes the desired longitudinal speed of the vehicles constituting the MB $\mathcal{C}_m$.

### B. Microscopic Scale MB Control

Once the reference speed for the MB is defined by the macroscopic level controller, as described in the previous section, the individual references for the CAVs composing the MB must be defined so that they keep the formation and then they all travel at speed $v^{\star}_{\mathcal{C}_m}$.

Note that, as in Fig. 1, the CAVs in a MB may be not perfectly aligned. Then, let $\bar{c}, \underline{c}$ be indexes of the first and last CAVs in the MB $\mathcal{C}_m$, respectively, defined as

$$\bar{c} = \arg\max_{c \in \mathcal{C}_m} p^c, \quad \underline{c} = \arg\min_{c \in \mathcal{C}_m} p^c,$$

Then, the reference speed for $\bar{c}$ is given by

$$v^{\bar{c},\star}(t) = \min\{v^{\star}_{\mathcal{C}_m}(t) + k_3\big(p^{\underline{c}}(t) - p^{\bar{c}}(t)\big), v_i^{\bar{c}}(t)\}, \quad (7)$$

where $k_3 \in \mathbb{R}_{>0}$ is a design parameter. As for the CAVs $c \in \mathcal{C}_m \setminus \bar{c}$, the reference speeds are defined as

$$v^{c,\star}(t) = \min\{v^{\star}_{\mathcal{C}_m}(t) + k_4\big(p^{\bar{c}}(t) - p^{c}(t)\big), v_i^{c}(t)\}, \quad (8)$$

with $k_4 \in \mathbb{R}_{>0}$ being a design constant.

Then, each CAV $c \in \mathcal{C}_m$ is controlled by means of a traction controller, as depicted in Fig. 3, so that $v^c \to v^{c,\star}$. In particular, following a reasoning similar to the one in [20], it is possible to map a desired value for the slip ratios, denoted as $\lambda_f^{\star}$ and $\lambda_r^{\star}$, starting from the desired longitudinal speed $v^{c,\star}$. The torque control signal for the wheel motors, denoted as $u = \begin{bmatrix} T_f & T_r \end{bmatrix}^{\top}$ is determined via the DNN-based slip controller in Fig. 3, such that the slip ratios reach their desired value in finite time.

*Remark 1:* It is important to notice that the choice of the desired slip ratios is strictly related to the value of the tire-road friction coefficient $\mu$, which in general is not perfectly known. A possible solution, which is the one adopted in this work, is to provide an estimate of $\mu$ via a SMC observer, as suggested in [20].

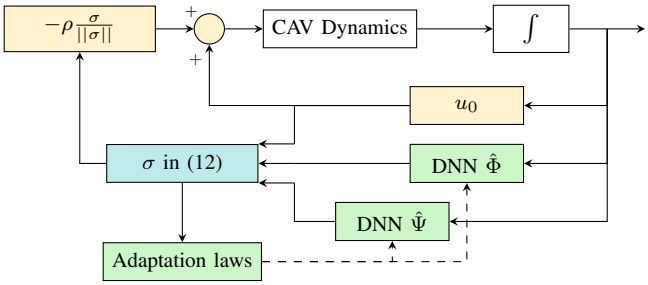



Fig. 3: Graphical representation of the DNN-ISM based traction controller.

Since the traction control problem can be reformulated as a slip ratio control problem, it is worth defining an auxiliary system characterized by states $\zeta = \begin{bmatrix} \lambda_f & \lambda_r \end{bmatrix}^{\top}$, with dynamics

$$\dot{\zeta} = \begin{bmatrix} h_f \\ h_r \end{bmatrix} + \begin{bmatrix} b_f & 0 \\ 0 & b_r \end{bmatrix} u_{\lambda} = h + B\, u_{\lambda}, \quad (9)$$

where $h_f, h_r, b_f$, and $b_r$ are defined differently in the acceleration and in the braking phases. In particular, for $p \in \{f, r\}$ it holds that

$$h_p = \begin{cases} -\frac{\dot{v}}{R_p \omega_p} - \frac{v F_{p_x}}{J_p \omega_p^2} & \text{if } \omega_p R_p > v, \\ -\frac{R_p \omega_p \dot{v}}{v^2} - \frac{R_p^2 F_{p_x}}{J_p v} & \text{if } \omega_p R_p < v, \end{cases}$$

$$b_p = \begin{cases} \frac{v}{J_p R_p \omega_p^2} & \text{if } \omega_p R_p > v, \\ -\frac{R_p}{J_p v} & \text{if } \omega_p R_p < v. \end{cases}$$

Then, it is possible to design an ISM control for the slip ratio as

$$u = u_0 - \varrho \frac{\sigma}{\|\sigma\|}, \quad (10)$$

where the second component aims at rejecting possible model mismatches and disturbances, while $u_0$ is a continuous

control law that makes the state $\zeta$ track a desired value $\zeta^\star$, and $\sigma$ is the so-called integral sliding variable, defined as

$$\sigma(t) = (\zeta(t) - \zeta^\star) - \zeta(t_0) - \int_{t_0}^{t} h + Bu_0(\tau)d\tau. \quad (11)$$

As it is clear from (11), the full knowledge of the dynamics (9) is required to design the ISM control law. Since such knowledge is assumed unavailable, an approximated version of (9) is employed. Specifically, two DNNs $\hat{\Phi} : \mathbb{R}^3 \to \mathbb{R}^2$ and $\hat{\Psi} : \mathbb{R}^3 \to \mathbb{R}^{2\times 2}$ are employed to provide an estimate of $h$ and $B$ in (9), so that $\sigma$ can be rewritten as

$$\sigma(t) = \zeta(t) - \zeta(t_0) - \int_{t_0}^{t} \hat{\Phi} + \hat{\Psi}u_0(\tau)d\tau. \quad (12)$$

Accordingly to the theory of DNN based ISM control developed in [23] and [24], the weights of the two DNNs are adjusted online, according to adaptation laws that are derived directly from Lyapunov analysis. Hence, properly selecting the weight dynamics and the gain $\varrho$ in (10), according to [23, Theorem 1], allows one to achieve in finite time the evolution of the state of the controlled system in a boundary-layer of $\|\sigma\| = 0$. This, along with a suitable choice of $u_0$, makes $\zeta$ sufficiently close to the desired value $\zeta^\star$.

## IV. Conclusions and Future Works

This work discusses the potential of controlling CAVs, forming MBs, at microscopic scale to positively impact macroscopic freeway traffic dynamics. A DNN-based approach is introduced for designing the individual CAV control systems, addressing traction control even with uncertain vehicle models. The approach ensures anti-skid braking and anti-spin acceleration of the CAVs, while leveraging CAVs formations as artificial moving bottlenecks to mitigate congestion and optimize the overall travel time in the traffic system. Future research directions may include the use to macroscopic models which employ a stochastic fundamental diagram (see, for instance, [28]), the extension to large scale traffic networks, and the explicit inclusion of route choice models.

## References


[1] S. Siri, C. Pasquale, S. Sacone, and A. Ferrara, "Freeway traffic control: A survey," *Automatica*, vol. 130, p. 109655, 2021.

[2] A. Ferrara, S. Sacone, and S. Siri, "An overview of traffic control schemes for freeway systems," in *Freeway traffic modelling and control*. Cham, Switzerland: Springer, 2018, ch. 8, pp. 193–234.

[3] M. Papageorgiou and A. Kotsialos, "Freeway ramp metering: An overview," *IEEE transactions on intelligent transportation systems*, vol. 3, no. 4, pp. 271–281, 2002.

[4] A. Hegyi, B. De Schutter, and H. Hellendoorn, "Model predictive control for optimal coordination of ramp metering and variable speed limits," *Transportation Research Part C: Emerging Technologies*, vol. 13, no. 3, pp. 185–209, 2005.

[5] C. Pasquale, S. Sacone, S. Siri, and B. De Schutter, "A multi-class model-based control scheme for reducing congestion and emissions in freeway networks by combining ramp metering and route guidance," *Transportation Research Part C: Emerging Technologies*, vol. 80, pp. 384–408, 2017.

[6] A. Jamshidnejad, I. Papamichail, M. Papageorgiou, and B. De Schutter, "A mesoscopic integrated urban traffic flow-emission model," *Transportation Research Part C: Emerging Technologies*, vol. 75, pp. 45–83, 2017.

[7] C. Diakaki, M. Papageorgiou, I. Papamichail, and I. Nikolos, "Overview and analysis of vehicle automation and communication systems from a motorway traffic management perspective," *Transportation Research Part A: Policy and Practice*, vol. 75, pp. 147–165, 2015.

[8] G. Abdelkader, K. Elgazzar, and A. Khamis, "Connected vehicles: Technology review, state of the art, challenges and opportunities," *Sensors*, vol. 21, no. 22, p. 7712, 2021.

[9] H. S. Mahmassani, "50th anniversary invited article—autonomous vehicles and connected vehicle systems: Flow and operations considerations," *Transportation Science*, vol. 50, no. 4, pp. 1140–1162, 2016.

[10] H. U. Ahmed, Y. Huang, P. Lu, and R. Bridgelall, "Technology developments and impacts of connected and autonomous vehicles: An overview," *Smart Cities*, vol. 5, no. 1, pp. 382–404, 2022.

[11] G. Piacentini, P. Goatin, and A. Ferrara, "Traffic control via platoons of intelligent vehicles for saving fuel consumption in freeway systems," *IEEE Control Systems Letters*, vol. 5, no. 2, pp. 593–598, 2021.

[12] ——, "Traffic control via moving bottleneck of coordinated vehicles," *IFAC-PapersOnLine*, vol. 51, no. 9, pp. 13–18, 2018.

[13] G. Piacentini, A. Ferrara, I. Papamichail, and M. Papageorgiou, "Highway traffic control with moving bottlenecks of connected and automated vehicles for travel time reduction," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 3140–3145.

[14] J. Li, C. Yu, Z. Shen, Z. Su, and W. Ma, "A survey on urban traffic control under mixed traffic environment with connected automated vehicles," *Transportation research part C: emerging technologies*, vol. 154, p. 104258, 2023.

[15] C. Zhao, H. Yu, and T. G. Molnar, "Safety-critical traffic control by connected automated vehicles," *Transportation research part C: emerging technologies*, vol. 154, p. 104230, 2023.

[16] Y. J. Zhang, A. A. Malikopoulos, and C. G. Cassandras, "Optimal control and coordination of connected and automated vehicles at urban traffic intersections," in *2016 American Control Conference (ACC)*. IEEE, 2016, pp. 6227–6232.

[17] P. Kachroo and M. Tomizuka, "Vehicle traction control and its applications," 1994.

[18] A. Ferrara, G. P. Incremona, and M. Cucuzzella, *Advanced and optimization based sliding mode control: Theory and applications*. SIAM, 2019.

[19] V. I. Utkin, *Sliding Modes in Control and Optimization*. Springer Berlin, Heidelberg, 1992.

[20] M. Amodeo, A. Ferrara, R. Terzaghi, and C. Vecchio, "Wheel slip control via second-order sliding-mode generation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 1, pp. 122–131, 2009.

[21] H. Lee and M. Tomizuka, "Adaptive vehicle traction force control for intelligent vehicle highway systems (ivhss)," *IEEE Transactions on Industrial Electronics*, vol. 50, no. 1, pp. 37–47, 2003.

[22] S. Drakunov, U. Ozguner, P. Dix, and B. Ashrafi, "Abs control using optimum search via sliding modes," *IEEE Transactions on control systems technology*, vol. 3, no. 1, pp. 79–85, 1995.

[23] N. Sacchi, E. Vacchini, G. P. Incremona, and A. Ferrara, "On neural networks application in integral sliding mode control," *Journal of the Franklin Institute*, vol. 361, no. 13, p. 106989, 2024.

[24] E. Vacchini, N. Sacchi, G. P. Incremona, and A. Ferrara, "Design of a deep neural network-based integral sliding mode control for nonlinear systems under fully unknown dynamics," *IEEE Control Systems Letters*, vol. 7, pp. 1789–1794, 2023.

[25] A. Ferrara, S. Sacone, and S. Siri, *Freeway traffic modelling and control*. Springer, 2018, vol. 585.

[26] C. Pasquale, S. Sacone, S. Siri, and A. Ferrara, "A new micro-macro metanet model for platoon control in freeway traffic networks," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 1481–1486.

[27] G. Genta, *Motor vehicle dynamics: modeling and simulation*. World scientific, 1997, vol. 43.

[28] X. Zhang, J. Sun, and J. Sun, "On the stochastic fundamental diagram: A general micro-macroscopic traffic flow modeling framework," *Communications in Transportation Research*, vol. 5, p. 100163, 2025.

# Online Application of Short-Term Traffic Prediction Integrating Simulation-Based Dynamic Traffic Assignment and Bayesian Filtering

Ernesto Cipriani[1], Chiara Colombaroni[2], Gaetano Fusco[2], Andrea Gemma[1], Natalia Isaenko[2]

[1] Department of Civil, Computer Science and Aeronautical Technologies Engineering, Roma Tre University, Via Vito Volterra 62, 00142 Rome, Italy

[2] Department of Civil, Constructional and Environmental Engineering, Sapienza University of Rome, Via Eudossiana 18, 00184 Rome, Italy

## 1 INTRODUCTION

Short-term traffic forecasting is crucial for intelligent transportation systems, enabling real-time decision-making to optimize traffic flow and reduce congestion (Lin, 2023). Traditional methods based on historical data and statistical models struggle with sudden disruptions. To overcome these limitations, advanced techniques integrating dynamic traffic assignment (DTA) and real-time data assimilation have emerged as effective solutions. This study presents an integrated framework for online traffic forecasting that combines simulation-based dynamic traffic assignment (DTA) with Bayesian filtering techniques updating demand and flows leverging on real-time sta. The main contributions of this work include:

- Development of an online simulation-based forecasting system.
- Integration of Bayesian filtering to enhance accuracy and manage uncertainties.
- Real-world application on a large highway network, demonstrating effectiveness

The paper is structured as follows. Section 2 details the methodology, Section 3 presents a case study, and Section 4 concludes with future research directions.

## 2 METHODOLOGY

The proposed methodology integrates an online traffic simulation model with dynamic route assignment and Bayesian filtering. The system continuously acquires real-time traffic data, updates origin-destination (O-D) matrices, and executes short-term simulations to predict network conditions. The approach follows a rolling horizon strategy, where periodic updates refine the predictions. The simulation model is designed to operate in two primary modes: offline and online. In offline mode, the model processes historical counts data to estimate a baseline dynamic O-D demand matrix, the equilibrium paths and flows, which are used as an initial condition in the online mode. The online mode continuously updates traffic states based on real-time observations, updating the current dynamic matrix, respecting historical one, and predicting the dynamic OD matrix for next interval. These new matrices are used to perform a new dynamic assignment. The estimated flow, related to the last interval are used to evaluate the accuracy of the model, comparing them with real-time data and predicted flows are used to predict the short-term fluctuation of traffic for a dynamic response to emerging traffic conditions. A diagram of this approach is shown in Figure 1. In the proposed methodology, the online assignment procedure is performed using a rolling horizon approach. In this framework, the system receives real-time data updates at regular intervals $\Delta t$ (e.g., 15 minutes) and utilizes this information to update the dynamic demand matrix over a time window extending from a preceding to a subsequent instant relative to the current time (e.g., 7:00–9:00 when the current time is 8:00). The updated demand

enables a dynamic assignment for the same period, allowing the prediction of traffic conditions. Generally, the simulation is completed within a time interval shorter than Δt. As a result, traffic estimates generated before the start of a new simulation improve the results of the previous one.

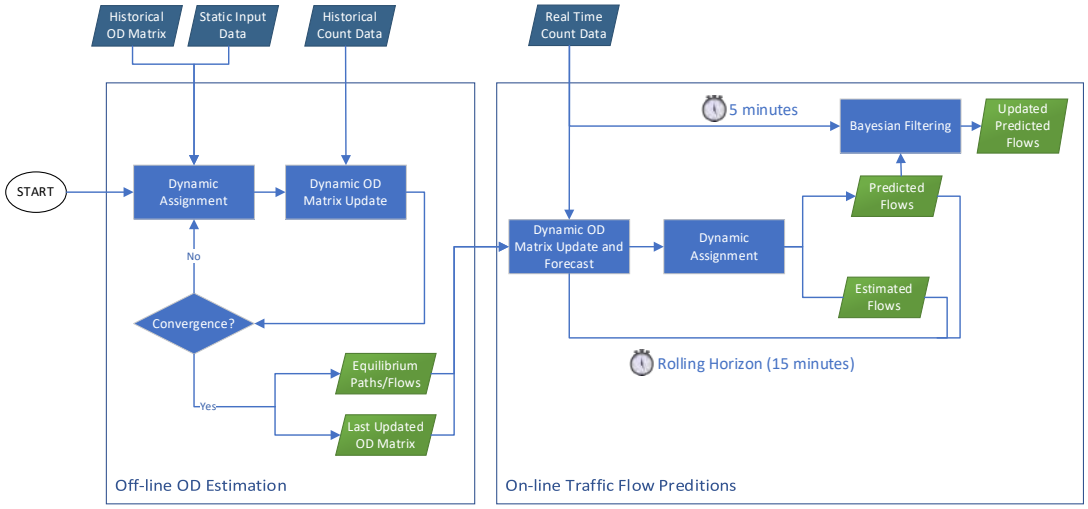



*Figure 1. Offline OD Estimation and On-Line Predictions.*
*Dark blue is used to highlight input data, green to highlight output data, and light blue for procedures.*

To compensate for this gap, during which traffic estimates are not updated, a hybrid approach incorporating a Bayesian filter has been introduced. This filter integrates the model-based traffic forecast as a prior estimate and refines it based on real-time traffic data collected at a higher frequency (e.g., every 5 minutes). This process applies a gradually adjusted correction over time, which varies according to the deviation from the standard conditions of a typical day.

## 2.1 OD Estimation

The offline OD estimation follows a simultaneous approach, correcting all time intervals of the dynamic OD matrix at once (Cascetta et al., 1993). This minimizes deviations from the initial seed demand while ensuring consistency with observed traffic counts. The process is formulated as a Generalized Least Squares (GLS) optimization problem, balancing differences between estimated and seed OD matrices and discrepancies in assigned vs. observed flows. Conversely, the online OD estimation adopts a sequential approach, where the OD matrix for the current interval is updated using real-time traffic counts before forecasting future matrices. This method enhances computational efficiency and responsiveness, making it suitable for real-time applications. Future OD estimates are generated through an autoregressive model, leveraging historical trends to predict demand variations ensuring real-time forecasting capability.

## 2.2 Dynamic Traffic Assignment

The Dynamic Traffic Assignment (DTA) model implemented in the online simulation framework integrates a parallel time-dependent path computation strategy, a second-order macroscopic traffic flow model, and an iterative demand correction mechanism. The combination of these components enables real-time traffic forecasting and decision support for traffic management systems. (Ben-Akiva et al., 2012, Florian et al, 2005)

### 2.2.1 Path Computation

The Shortest Path Problem (SPP) computation uses a parallelized time-dependent search algorithm (Sung et al., 2000), adapted from Dijkstra's method to handle time-dependent costs and prohibited maneuvers (Gutiérrez & Medaglia, 2007) via a link-labeled approach. In the offline phase, it computes optimal paths for each OD pair and departure interval, iterating until a minimum number (k) of paths is

reached, after which equilibrium paths are stored. In the online phase, precomputed paths serve as a baseline, but the system dynamically adapts to real-time congestion and incidents, ensuring adaptive route selection and preventing outdated path choices from distorting the simulation.

#### 2.2.2 Traffic Flow Model

The traffic flow component of the model is based on a second-order macroscopic approach, which provides a more accurate representation of congestion dynamics, especially in high-density conditions. The model performs a preliminary loading of path flows, which are redistributed according to the equilibrium proportions. Using equilibrium travel times, the flow propagation is computed, determining arc flows and splitting rates. Finally, these outputs, along with the path assignments, are used as inputs for the second-order highway flow model, which simulates the updated demand on the network and enables the estimation of new travel times for each arc.

#### 2.2.3 Online application of the simulation-dynamic assignment model

In the online implementation, the general model continuously updates demand and traffic conditions to align with real-time data, constantly adapting to the evolving network state. The dynamic traffic assignment (DTA) framework consists of a structured sequence of operations. First, the system loads the precomputed equilibrium paths from the offline phase, which serve as a reference for route selection. The updated demand is then dynamically distributed across these paths based on the real-time OD correction. Subsequently, the second-order macroscopic traffic flow model simulates the evolution of traffic conditions over the forecasting horizon, capturing congestion dynamics and flow propagation. As new simulated data become available, travel times are recalculated, and the route assignment is dynamically adjusted to reflect current conditions. In the presence of traffic anomalies, the system introduces alternative paths, enabling a rapid response to significant deviations from expected conditions.

### 2.3 Bayesian Filtering Logic

The Bayesian filtering approach is integrated into the simulation model to enhance the accuracy of real-time traffic forecasts by combining model-based estimates with observed traffic data. The methodology relies on an autoregressive model, which utilizes past observations to predict future traffic variables (flow, speed, and density). The forecast is updated using Bayes' theorem, where the *prior* estimate is derived from the simulation model, and the *posterior* estimate is obtained by combining this prediction with a statistical extrapolation of the most recent observations. The filter dynamically adjusts the weight assigned to observations relative to model predictions, adapting to traffic conditions. Under normal conditions, greater confidence is given to the model's forecast, ensuring stability in the prediction process. However, in the presence of anomalies such as incidents or unexpected congestion, the weight of real-time observations increases, allowing for a rapid correction of the forecast and improving the system's responsiveness to sudden changes in traffic dynamics. Additionally, the methodology incorporates a time-smoothed correction coefficient, which modulates the forecast update based on the prediction horizon. In the short term, real-time data dominate the correction process, while in the medium to long term, the forecast gradually aligns with the model output. (Castillo at al. 2008, Li & Xie, 2024)

## 3 Real Case Application

### 3.1 Case Study Description

The proposed traffic simulation and forecasting framework was embedded within the DSS of the control room of "Concessioni Autostradali Venete" (CAV), a highway network that includes the A4 "Passante di Mestre", the A57 "Tangenziale di Mestre", and the "Raccordo Autostradale" to "Marco Polo Airport". This infrastructure experiences significant traffic volumes, with frequent congestion, especially near

urban areas and major interchanges. The framework support the DSS in traffic management strategies such as variable speed limits (VSL) and dynamic lane control, including the opening of emergency lanes under specific traffic conditions. The simulation-forecasting model provides real-time estimations of traffic flow, speed, and congestion levels, allowing operators to proactively implement regulation measures to mitigate delays and improve network performance.

### 3.2 PREDICTION RESULTS

Results from real-world deployment demonstrated the system's high accuracy in traffic state prediction, with a deviation of less than 10% in flow estimates compared to sensor data. The integration of the Bayesian filtering technique enhanced anomaly detection and traffic state estimation, ensuring adaptive corrections to unexpected congestion patterns. The dynamic assignment model effectively redistributed demand under varying conditions, preventing the propagation of bottlenecks and improving overall network efficiency. The integration of this simulation-forecasting framework into the CAV DSS has proven instrumental in optimizing control measures, enhancing safety, and reducing travel times, making it a valuable tool for real-time highway traffic management.

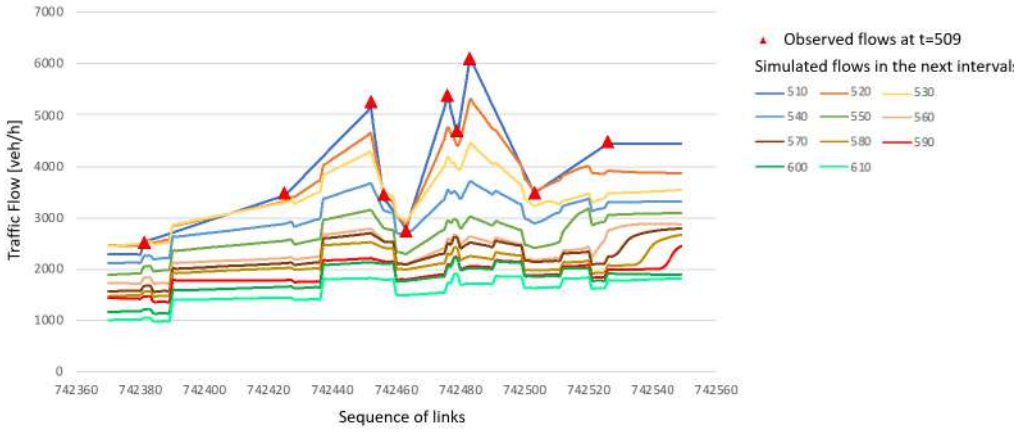



*Figure 2. Flows observed in a road section at time t=509 and flows simulated at subsequent intervals*

Figure 2 shows the flow forecasts on a road segment resulting from the application of the entire assignment fragment and the Bayesian filter for different time intervals with respect to space. It can be observed that for the interval following the application of the filter (interval t = 510) the forecast obtained is very similar to the observed values and as for long-term forecasts (100 minutes) the values are towards the observed historical values.

## 4 CONCLUSIONS

The integration of the traffic simulation and forecasting framework within the CAV Decision Support System (DSS) has demonstrated its effectiveness in real-time traffic management, enabling dynamic control strategies such as variable speed limits and emergency lane usage. The system's ability to continuously update OD demand, dynamically reassign traffic flows, and incorporate Bayesian filtering has significantly improved traffic state estimation and anomaly detection.

# 5 REFERENCES


Ben-Akiva, M. E., Gao, S., Wei, Z., & Wen, Y. (2012). A dynamic traffic assignment model for highly congested urban networks. *Transportation research part C: emerging technologies*, *24*, 62-82.

Cascetta, E., Inaudi, D., Marquis, G. (1993). Dynamic estimators of origin-destination matrices using traffic counts. Transportation Science 27(4): 363-373.

Castillo, E., Menéndez, J. M., & Sánchez-Cambronero, S. (2008). Predicting traffic flow using Bayesian networks. *Transportation Research Part B: Methodological*, *42*(5), 482-509.

Florian, M., Mahut, M., & Tremblay, N. (2005). Application of a simulation-based dynamic traffic assignment model. Simulation Approaches in Transportation Analysis: Recent Advances and Challenges, 1-22.

Gutiérrez, E., & Medaglia, A.L. (2007). Labeling algorithm for the shortest path problem with turn prohibitions with application to large-scale road networks. *Annals of Operations Research, 157*, 169-182.

Lin, Y. (2023). Models, Algorithms and Applications of DynasTIM Real-Time Traffic Simulation System. *Sustainability*, *15*(2), 1707. https://doi.org/10.3390/su15021707

Sung, K., Bell, M.G., Seong, M., & Park, S. (1998). Shortest paths in a network with time-dependent flow speeds. *Eur. J. Oper. Res., 121*, 32-39.

Li, J., & Xie, H. Y. (2024). Short term traffic flow prediction method for urban roads: improved Bayesian network. *Advances in Transportation Studies*.

# Vulnerability assessment of High-speed Rail (HSR) networks: a schedule-based approach with a case study

**Coppola P. Guglielmi F., Mariano P. (*)**

**DMEC – Politecnico di Milano**

(*) corresponding author

**Keywords:** *High-Speed Rail (HSR), Vulnerability, Schedule-based*

Transport networks are inherently complex systems that are highly susceptible to events of disruption and degradation (Sun et al., 2020). They are not only interconnected and interdependent but also intricately linked with broader socioeconomic and environmental systems. Consequently, disruptions in transport can trigger significant societal impacts. This is the case, for instance of railways systems which are relevant component of regional economies, underscoring the importance of studying their vulnerabilities in order to improve their resilience to disruptions in space and time.

There are diverse views on how to quantify vulnerability with respect to different hazards. In the context of this work, vulnerability is intended as the propension of the hazard to generate direct and indirect damages to things and people (Berdica, 2002). The quantification of such damage, in turn, represents the exposure to the hazard. Henceforth, the concept of vulnerability is closely tied to the transportation supply, whereas the exposure is related to travel demand. For example, a network with alternative and redundant path alternatives might be less vulnerable than a network with a limited number of paths which in case of disruption could fail to guarantee connections on some origin-destination (OD) pairs. The estimated number of travellers affected by the connection loss represents the exposure to the disruption.

Estimating vulnerability and exposure to hazards is particularly challenging for scheduled service systems (e.g., railways network), since the diversion to alternative services in case of disruption is constrained and limited by the scheduled nature of the service in space, described by the availability only at terminals, and also in time, described by the availability according to timetable only in certain time instants. In the literature, vulnerability assessment is approached in two predominant ways (Mattsson and Jenelius, 2015): topological and system-based. Topological approaches, rooted in graph theory, assess the vulnerability of a transportation system based solely on the physical configuration of the network. On the other hand, system-based approaches integrate supply-demand interactions to account for travellers' responses to disturbances and disruptions factors. The latter approach proves itself to be more effective in assessing the real-world impacts of disruptions, yet it requires also data for the travel demand estimation. Topological approaches are simpler in terms of methodological and data requirements, yet they prove inadequate for analysing scheduled service systems, such as rail system, as they fail to account for the time dimension of the supply. For instance, the same type of disruption (e.g. an interruption of a section of the network or in a station) may have different impacts according to time of the day, depending on the number of services operating along the network at that time.

This paper aims to assess, using a schedule-based modelling approach (Nuzzolo et al., 2002), railway networks vulnerability including disruptions not only in terms of structural damage to the physical infrastructure but also in relation to the temporal availability of services. Unlike traditional methods, this approach considers both where and when a disruption occurs, allowing for a more comprehensive analysis of its impact. Two types of indicators are proposed: topological indicators, which measure the number of affected services, and system-based indicators, which also account for the number of impacted travellers, estimated through a schedule-based assignment.

The proposed approach is applied to the Italian HSR network to compute and compare these indicators. Topological indicators rely on classical graph theory metrics—such as strength centrality, betweenness centrality, and clustering coefficient (Zhou et al., 2019)—calculated on a diachronic graph of HSR services. System-based indicators, on the other hand, incorporate passenger flows derived from a schedule-based HSR assignment model (Silvestri et al., 2024), capturing the interaction between supply and demand. By comparing these results, we highlight significant differences between the two perspectives, providing deeper insights into the key factors influencing disruption vulnerability and network exposure.

This research contributes to the broader discourse on transportation system vulnerability, providing a new methodological approach that integrates both topological and systemic analyses to diachronic networks. Findings suggest that the proposed approach offers valuable insights for assessing the vulnerability of the national HSR system, supporting the development of actions to make the system more resilient, i.e. capable of withstanding disruptive events.

## References


Berdica, K., 2002. An introduction to road vulnerability: what has been done, is done and should be done. Transp. Policy 9, 117–127. https://doi.org/10.1016/S0967-070X(02)00011-2

Mattsson, L.-G., Jenelius, E., 2015. Vulnerability and resilience of transport systems – A discussion of recent research. Transp. Res. Part Policy Pract. 81, 16–34. https://doi.org/10.1016/j.tra.2015.06.002

Nuzzolo, A., Lam, W.H.K., Bell, M.G.H., 2002. Transit Path Choice and Assignment Model Approaches(°), in: Advanced Modeling for Transit Operations and Service Planning. Emerald Group Publishing Limited, pp. 93–124.

Silvestri, F., Mariano, P., Montino, T., 2024. Estimating Schedule-Based Assignment Models for High-Speed Rail (HSR) Services Using Multiple Data Sources | springerprofessional.de [WWW Document]. URL https://link.springer.com/chapter/10.1007/978-3-031-53684-7_7 (accessed 2.11.25).

Sun, W., Bocchini, P., Davison, B.D., 2020. Resilience metrics and measurement methods for transportation infrastructure: the state of the art. Sustain. Resilient Infrastruct. 5, 168–199. https://doi.org/10.1080/23789689.2018.1448663

Zhou, Y., Wang, J., Yang, H., 2019. Resilience of Transportation Systems: Concepts and Comprehensive Review. IEEE Trans. Intell. Transp. Syst. 20, 4262–4276. https://doi.org/10.1109/TITS.2018.2883766

# Enhancing Operations through Integrated Heterogeneous Multi-depot Vehicle Scheduling and Transit Assignment

Prateek Agarwal[1], Kokkonda Prashanth[2], and Tarun Rambha[3]

[1]Civil Engineering, Indian Institute of Science (prateeka@iisc.ac.in)
[2]Civil Engineering, Indian Institute of Science (kprashanth@alum.iisc.ac.in)
[3]Center for infrastructure, Sustainable Transportation and Urban Planning, Indian Institute of Science (tarunarmbha@iisc.ac.in)

**Abstract**

A significant challenge in the public transit is integrating passenger route choice and vehicle scheduling. Passengers choose routes based on transit supply, while transit operators design timetables based on ridership. Existing studies often address these problems separately. Our research jointly solves them using a novel local search framework for the Heterogeneous Multi-Depot Vehicle Scheduling Problem (HMDVSP). User sensitivity to supply perturbations is assessed using an agent- and schedule-based Public Transit Assignment (PTA) framework that incorporates explicit capacity constraints. The proposed framework is parallelized and can simulate real-world transit networks with heterogeneous fleets and millions of door-to-door queries within a few hours. Using this model, we explore practical applications, such as (1) assigning trips to a heterogeneous fleet of buses (2) comparing passenger ridership and revenue for homogeneous and heterogeneous fleets, and (3) fleet planning to meet future demand.

***Keywords***: transit assignment, vehicle scheduling, local search, route choice

## 1 Introduction

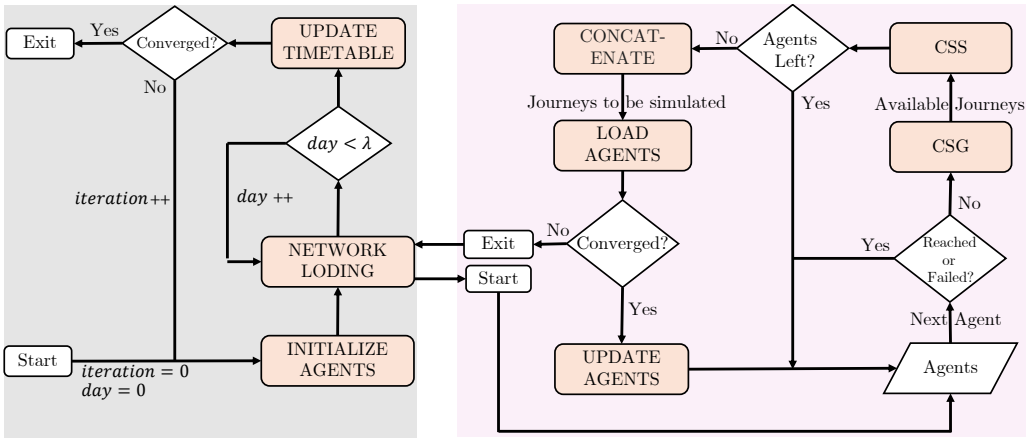
The planning process for public transportation systems typically involves four key stages: network design, timetabling, Vehicle Scheduling Problem (VSP), and crew scheduling. For a given timetable, Public Transit Assignment (PTA) can be used to predict travelers' route choices for a given supply configuration. Transit operators continually enhance service quality to make public transit an attractive travel and commuting option. However, the effectiveness of supply-side improvements is dependent on the complex nature of user behavior. Unlike most personal vehicle users, transit riders typically consider multiple criteria when selecting their journeys (or routes), such as in-vehicle travel time, waiting time, walking distance, cost, number of transfers, and crowding levels. Supply-side changes can impact many of these features and thus ridership. Ridership can also be used to tweak supply to improve transit operator efficiency. Studies that capture this dependence are relatively sparse owing to increased computational complexity. Though some works combine passenger behavior into vanilla VSPs (Liu and Ceder, 2017), its effect on Heterogeneous Multi-Depot Vehicle Scheduling Problem (HMDVSP) remains largely unexplored. To address this gap, we introduce an iterative HMDVSP and PTA framework that captures the user sensitivity to supply perturbations while maximizing the net profit/loss, defined as earnings from ticket sales minus the crew salaries and fuel expenses. We assume that the demand is captive and do not model elasticity. Although profit/loss should not be the primary goal of a public transit operator, our objective helps cut down costs by carefully allocating buses of different capacities to timetabled trips depending on the vehicle occupancies.

## 2 Methodology

**Preliminaries:** A bus *rotation* denotes a list of trips that can be performed in sequence, including the two deadhead trips (if any), i.e., from the starting depot to the initial stop of the first trip and from the final stop of the last trip to the ending depot. Buses start and end their rotation at the same depot. Each depot has a maximum capacity for housing buses. A list of bus rotations forms a valid *schedule* if (1) depot capacity limits are not exceeded and (2) each trip is assigned to a single bus. A pair of trips are time compatible if the bus has enough time to travel from the end point of the first trip to the start of the next trip. The goal of our work is to find a schedule that optimizes the *net profit/loss*, which deducts the operational costs from the revenue from ticket sales. The revenue is calculated as the product of total passenger km and fare per km. The operational cost include crew and fuel expenses. We set crew costs by multiplying a fixed crew pay rate and total working hours. The fleet comprises buses of different capacities,

with a fixed number of buses available in each category. We differentiate fuel costs based on capacity and calculate it by multiplying bus mileages with the total distance traveled. Bus acquisition costs are not considered but they can be trivially incorporated in our framework.

To evaluate the impact of changes in transit supply on ridership, we need a tractable PTA tool. To address this challenge, we developed a simulation-, agent-, and schedule-based PTA framework that incorporates explicit capacity constraints. Figure 1 presents two main modules of the PTA framework: *Equilibration* which allows time-table updates and tracks flow-based convergence and *Network Loading* for route choice and flow propagation with capacity restrictions. The simulation starts with the INITIALIZE AGENTS function in the equilibration module, which sets agent attributes. If an agent $a$ has not reached their destination, CSG identifies available journeys using the Trip-Based Public Transit Routing (TBTR) algorithm (Witt, 2015; Agarwal and Rambha, 2024). Next, CSS selects a preferred journey the available journey set using a mixed multinomial logit model combined with path size to account for overlapping journeys. Once each agent has a preferred journey, the CONCATENATE and LOAD AGENTS functions identify agents who are unable to complete their desired journey as they were denied boarding at some intermediate point due to capacity constraints. For such agents, the UPDATE AGENTS function resets their attributes. Due to stochasticity in the choice model, we repeat the NETWORK LOADING $\lambda$ times, to capture variability. Each network loading can be viewed as a single realization or a day of transit operations and the repetitions give a more stable estimate of passenger flow and load patterns across routes. After $\lambda$ cycles, we call UPDATE TIMETABLE in the equilibration module to adjust timetables based on transit trip dwell times. If convergence is not reached, the iteration counter increments, agents reinitialize, and the process repeats. This mimics a scenario where operators adjust schedules periodically to minimize deviations between planned and actual operations.




(a) Equilibration

(b) Network Loading

Figure 1: Proposed public transit assignment framework

**Integrated HMDVSP Framework**: Figure 2 illustrates the proposed local search framework. The framework consists of two phases: an *Initialization* phase, which generates an initial feasible schedule, and a *Main Process*, which iteratively perturbs and refines the solution.

**Initialization:** We start with a greedy algorithm called CONCURRENT SCHEDULER to assign trips to buses. The process begins by sorting all trips in ascending order based on their departure times from their first stop. The first trip is assigned to the first available bus. For each subsequent trip $j$, the algorithm evaluates its temporal compatibility with the most recently assigned trip $i$ in every existing bus rotation. If trip $j$ is compatible with trip $i$, it is appended to the first corresponding bus rotation that meets the compatibility criteria. If no compatible bus rotation is found, a new bus is assigned to handle trip $j$. Once the trips are assigned, bus capacities are determined in two steps. First, the PTA is solved under the assumption that each bus has unlimited capacity. This yields an uncapacitated estimate of the ridership for each trip. Second, bus capacities are assigned based on the ridership estimates using an integer program. Let $V$ denote the list of bus IDs currently in use. For a bus $u \in V$, say $R_u$ represents its *rotation*. Let $o(j, m)$ denote the ridership of the $m^{th}$ link (a link is a stop-to-stop movement without any intermediate halt) in trip $j$, and let $x_{u,c}$ be a binary decision variable that equals 1 if bus $u$ is assigned capacity $c \in C$, where $C$ is the list of unique bus capacities. For a trip $j$, $m \in (0, l_j - 1)$ , where $l_j$ denotes the number of stops in trip $j$. The objective function in (1) minimizes the number of passengers that exceed the assigned capacity of the bus. Constraint (2) ensures that each bus is assigned exactly one capacity type, while Constraint (3) limits the total number of buses assigned to a specific capacity type to the maximum number of buses available in that category.

$$\min \sum_{c \in C} \sum_{u \in V} x_{u,c} \underbrace{\sum_{j \in R_u} \sum_{m=1}^{l_j - 1} \max(0, o(j,m) - c)}_{\text{Demand Denied}} \tag{1}$$

$$\text{s.t.} \quad \sum_{c \in C} x_{u,c} = 1 \qquad \forall\, u \in V \tag{2}$$

$$\sum_{u \in V} x_{u,c} \leq \beta_c \qquad \forall\, c \in C \tag{3}$$

$$x_{u,c} \in \{0,1\} \qquad \forall\, u \in V, c \in C \tag{4}$$

**Main Process:** The valid schedule generated during the *Initialization* phase is iteratively refined using a local search framework until maximum runtime criterion is met. In each iteration, nine operators are applied sequentially on a list of the top-$\nu$ best schedules. Each operator selects a schedule from this list and generates a new candidate schedule. If the new schedule outperforms the worst schedule in the list, it replaces the latter, similar to genetic algorithms. This population approach helps explore the search space better and can prevent local optima. The operators can be categorized into three groups: Homogeneous, Heterogeneous, and Capacity-based.

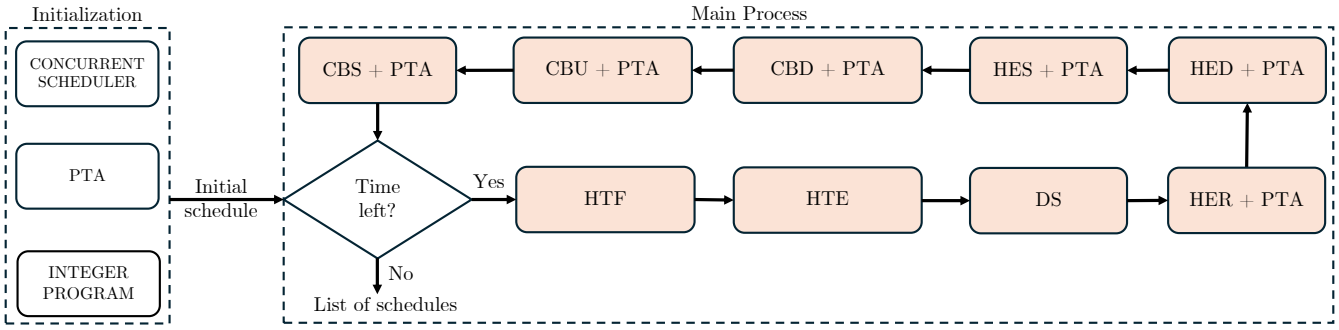



Figure 2: Local Search Architecture

**Homogeneous Operators**: These operators involve exchanging or shifting trips between buses of the same capacity. Since the trip capacities do not change, the ridership remains the same. Consequently, these are computationally efficient and can do an exhaustive search among all possible exchanges or shifts. They include three operators: Homogeneous Trip Shift (HTS), Homogeneous Trip Exchange (HTE), and Depot Swap (DS). HTS and HTE aims to minimizes the fuel cost by deadhead distance. For instance, in HTS buses are sorted in descending order based on the number of assigned trips. For each bus, the algorithm iterates through its trips and identifies the best valid shift to another bus of the same capacity. If no valid shift is found, a new bus of the same capacity is introduced (if available). DS minimizes the fuel cost by reducing the deadhead distance (to and from the depot). To do so, we iterate over all the rotations and assign them to the depot closest to the starting stop of the first stop (subject to depot capacity constraints).

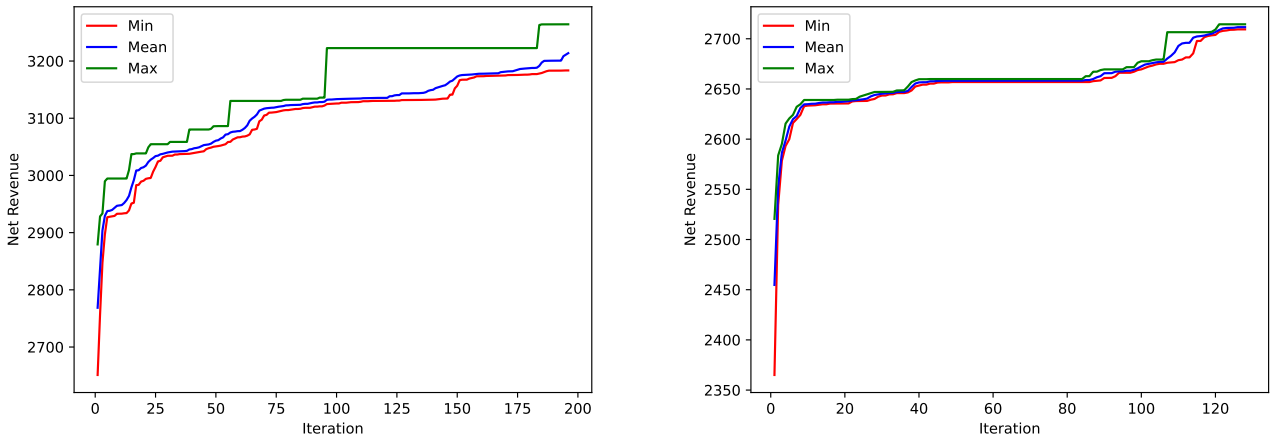
**Heterogeneous Operators**: This category comprises three operators: Heterogeneous Trip Exchange Revenue (HER), Heterogeneous Trip Exchange Deadhead (HED), and Heterogeneous Trip Shift (HES). The core idea behind these operators is to rank buses based on specific criteria and then explore potential shifts or exchanges. Heterogeneous operators are computationally expensive due to their impact on trip capacity, which requires invoking PTA. To manage this issue, we draw inspiration from Tabu Search by maintaining a list of tabu moves to avoid revisiting ineffective solutions. For example, HER focuses on maximizing revenue. It begins by copying the trips into two lists, say $T$ and $T'$. The trips in list $T$ are sorted by the total number of unoccupied seats across all links, while $T'$ is sorted by the number of fully occupied links (both in descending order). For each pair of trips $\{(t,t') : t \in T,\, t' \in T'\}$, the algorithm checks if $(t,t')$ is not in HER's tabu list. If not, PTA is invoked to evaluate the swap. This process continues for a fixed time budget. HED and HES optimize the deadhead distance by exchanging and shifting trips, respectively.

**Capacity-based Operators**: As the name suggests, these operators swap/change the capacity of the buses. Unlike the above two categories that exchange/swap trips between two busses, operators in this category perform larger perturbations as the capacities of all the trips being executed by the bus are changed at once. Specifically, we propose three operators: Capacity-Down (CBD), Capacity-Up (CBU), and Capacity-Swap (CBS). CBD reduces fuel costs by

lowering the capacity of a selected bus to the next lower level $c \in C$, provided buses of capacity $c$ are available. Buses with lower capacity typically have higher mileage. However, this may lead to revenue loss due to reduced seating. To mitigate this, buses are ranked by the number of unoccupied seats, prioritizing those with minimal potential revenue loss. CBU increases the capacity of a bus to the next upper level (if available), prioritizing buses where the additional revenue from ticket sales outweighs the increased fuel consumption. Buses are ranked based on the number of fully occupied links across all trips. Instead of modifying a single bus, we draw inspiration from variable neighborhood search and employ an adaptive block-size strategy to control the number of buses perturbed in each iteration.

# 3 Experiments

All operators described earlier were implemented in Python 3, while the PTA framework was developed in C++. Experiments were conducted on two U.S. transit networks: Santa Barbara, CA and Columbus, OH. Passenger route choice was modeled based on in-vehicle travel time, number of transfers, walk time, and crowding. Sensitivity parameters for the route choice model, along with the origin-destination matrix, fleet configuration, depot locations, and bus mileage, were synthetically generated using online sources. Each network was allocated a maximum runtime of 8 hours. For the PTA framework, the analysis was limited to a single network loading and one equilibration iteration. Figure 3 illustrates the variation in daily net revenue (in $) across iterations. The green, blue, and red lines denote the maximum, mean, and minimum net revenue, respectively.




(a) Santa Barbara

(b) Columbus

Figure 3: Daily net revenue in dollars for Santa Barbara (left) and Columns (right). The green, blue, and red lines denote the maximum, mean, and minimum net revenue, respectively

**Full Paper Outlook:** Full paper includes details on the proposed operators. An extensive set of experiments will be conducted to check the efficacy of the proposed framework. Further, we also aim to study policy implications, such as what happens if the fleet transitions from homogeneous to heterogeneous. By embedding the proposed framework inside a Bayesian Optimization framework, we also evaluate future scenarios such as given budget constraints, what buses should be acquired to meet future demand.

# References


Agarwal, P. and T. Rambha (2024). Scalable Algorithms for Bicriterion Trip-based Transit Routing. *IEEE Transactions on Intelligent Transportation Systems*.

Liu, T. and A. A. Ceder (2017). Integrated Public Transport Timetable Synchronization and Vehicle Scheduling with Demand Assignment: A Bi-objective Bi-level Model using Deficit Function Approach. *Transportation Research Procedia 23*, 341–361.

Witt, S. (2015). Trip-Based Public Transit Routing. In *Algorithms-ESA 2015*, pp. 1025–1036. Springer.

# Exploring multiple traffic assignment solutions resulting from economies of scale

Gunnar Flötteröd and David Watling

February 2025

## 1 Motivation

Economies of scale arise when the utility of a good increases with its usage. We give two examples in the transportation context. (i) In freight assignment, increasing the demand for one shipment alternative allows to deploy larger vehicles and terminals, which reduces the unit shipment cost and may further increase the demand for this alternative. (ii) Including shared mobility services with adaptive pricing schemes in public transport assignment means that the high usage of a shared service allows the provider to reduce its price and further increases its attractiveness.

Including such scale effects in traffic assignment easily leads to a multiplicity of solution points (equilibria). The objective of this ongoing work is to devise techniques that support the credible deployment of simulation-based traffic assignment in settings where a multiplicity of solutions can be expected. See Schmöcker et al. (2014) and Bar-Yosef et al. (2013) for further illustration.

Technically, we consider a discrete-time stochastic assignment process with a stationary distribution of several islands of (weakly connected) probability mass. The process spends, on average, enough iterations in one such region to settle in an apparently stationary local distribution before eventually (stochastically) leaving that region and settling somewhere else. This means that observing the process over a limited number of iterations may give a starting point dependent impression of its stationary distribution. This effect has already been described by Watling (1996); our work is more geared towards the study of scale effects.

## 2 Basic setup

Consider a population of decision-making agents indexed by $n = 1, \ldots, N$. Denote the choice set of agent $n$ by $C_n$. Let $C_{1:N} = C_1 \times C_2 \times \ldots \times C_N$ be the population's joint choice set, and let $i_{1:N} \in C_{1:N}$ represent the choices of all agents.

We are interested in studying models for policy analysis and introduce a policy parameter $\lambda \in [0, 1]$ where zero means "base case", one means "policy

case", and values in-between represent a partical policy implementation. A policy may affect agent choices as well as the network loading, as described immediately below.

Let $\mathcal{X}$ be the set of physical system states, and let $g : (C_{1:N}, [0,1]) \to \mathcal{X}$ be the policy-sensitive network loading where $g(i_{1:N}, \lambda)$ computes the network conditions resulting from the choices $i_{1:N}$ given the policy $\lambda$. Each agent is equipped with a policy-sensitive choice model $P_n : (C_n, \mathcal{X}, [0,1]) \to [0,1]$ where $P_n(i \mid x, \lambda)$ is the probability that agent $n$ chooses alternative $i$ given the network conditions $x$ and policy $\lambda$. Let $P : (C_{1:N}, \mathcal{X}, [0,1]) \to [0,1]$ represent the population choices in that $P(i_{1:N} \mid x, \lambda) = \prod_{n=1}^{N} P_n(i_n \mid x, \lambda)$.

A simple simulation version of the considered process model is given in Algorithm 1. For simplicity, agents only remember their most recent network condition experience.

**Algorithm 1** Stochastic process assignment model

```
Set policy parameter λ.
Set initialize initial network conditions x^(0).
for k = 1, 2, ... do
    Simulate choices i_{1:N}^(k) ~ P(· | x^(k−1), λ).
    Simulate network conditions x^(k) = g(i_{1:N}, λ).
end for
```

# 3 Small example

We consider a two-alternative scenario with $N = 1000$ homogeneous agents who all face identical choice sets $C_1 = \ldots = C_N = \{1, 2\}$. The network loading is policy-insensitive and merely counts the number of agents using alternative one:

$$g(i_{1:N}, \lambda) = \sum_{n=1}^{N} \mathbf{1}(i_n = 1). \tag{1}$$

A binomial logit choice model, common to all agents, reflects utilities of scale:

$$P_n(1 \mid x, \lambda) = \frac{e^{\mu \cdot x/N}}{e^{\mu \cdot x/N} + e^{\mu \cdot (1 - x/N)}} \tag{2}$$

with $\mu \geq 0$. Policy-sensitivity will be established further below by making $\mu$ a function of $\lambda$.

The positive $\mu$ parameter in combination with utility increasing with the use of an alternative suggests that this scenario has a tendency to attain stationarity either at $x$ values relatively near zero or relatively near $N$, with possibly very rare switches in-between. Figure 2 illustrates the number of alternative-1 users over 10'000 assignment iterations for $\mu = 2.08$. One observes that the system may stay for thousand of iterations in a state where most decision makers use one particular alternative before flipping to the opposite situation of most decision makers using the other alternative.

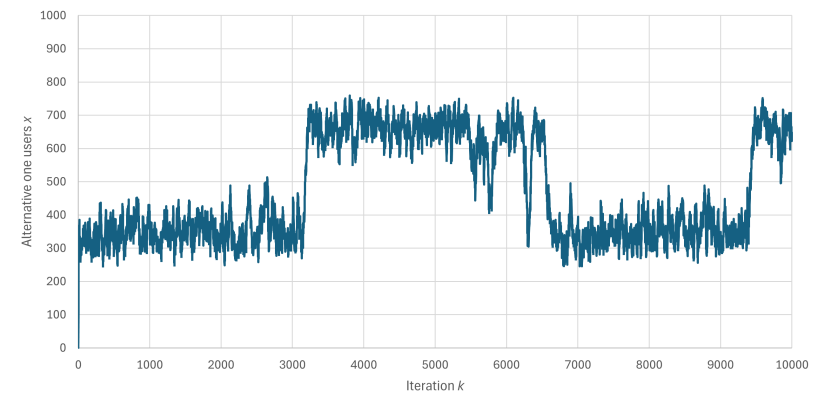



Figure 1: Small example. One process realization.

# 4 Metropolis Hastings based model analysis

To explore solution multiplicity, we deploy the Metropolis-Hastings (MH) algorithm (Hastings, 1970). We define the state space explored by the MH process as $\mathcal{X} \times C_{1:N}, \times [0,1]$, where one element $(x, i_{1:N}, \lambda)$ of this state space consists of a realization of the network conditions, choices for all agents in the population, and a setting of the policy parameter.

Letting $\phi : \mathcal{X} \to \mathbb{R}_+$ denote a zero-centered probability density function over the network conditions, we construct the proposal distribution

$$q\left((x, i_{1:N}, \lambda) \to (x', i'_{1:N}, \lambda')\right) = \phi_{\text{proposal}}(x - x') \cdot P(i'_{1:N} \mid x', \lambda'), \tag{3}$$

meaning that simulating a proposal amounts to (i) uniformly drawing $\lambda$ between zero and one, (ii) drawing $x'$ from a distribution centered at $x$, and drawing $i'_{1:N}$ by evaluating the choice model given $x'$ and $\lambda'$.

As the target weights, we use

$$w(x, i_{1:N}, \lambda) = P(i_{1:N} \mid x, \lambda) \cdot \phi_{\text{target}}(x - g(i_{1:N}, \lambda)), \tag{4}$$

where the first factor evaluates the likelihood of obtaining the choices $i_{1:N}$ given the network conditions $x$ and the second factor measures the deviation between the network conditions $x$ and the result from a network loading of the choices $i_{1:N}$. In combination, these terms measure the mutual consistency of travel choices and network conditions; this aims to approximate the stationary conditions attained by Algorithm 1.

Given these choices of proposal distribution and target weights, we obtain the simple acceptance probability

$$\alpha\left((x, i_{1:N}, \lambda) \to (x', i'_{1:N}, \lambda')\right) = \min\left\{1, \frac{\phi_{\text{target}}(x' - g(i'_{1:N}, \lambda'))}{\phi_{\text{target}}(x - g(i_{1:N}, \lambda))}\right\}. \tag{5}$$

Important for practical use with a black-box simulator, this setting requires only drawing choices and loading the network; no knowledge of the underlying choice distribution or concrete properties of the network loading is needed.

# 5 Small example, continued

We create a policy-sensitive version of the small example by letting

$$\mu = \mu(\lambda) = (6\lambda - 3)^2 + b \tag{6}$$

with $b$ a structural model parameter; we explore below cases with $b \in [1, 4]$. For any value of $b$ within that range, $\mu$ is strictly positive; it is largest for $\lambda = 0$ and $\lambda = 1$ and smallest for $\lambda = 0.5$. By construction, $\mu(0) = \mu(1)$ so that there is no difference between the base case and the policy case.

We assume that data collected from reality indicates that the base case solution is the one with relatively few users on alternative one, i.e. $x$ near zero for $\lambda = 0$. Given this starting point, we would like to anticipate the effect of gradually introducing the policy measure, i.e. moving $\lambda$ from zero to one. For this, we deploy the MH algorithm where we instantiate both $\phi_{\text{proposal}}$ and $\phi_{\text{target}}$ as univariate Gaussian distributions with zero mean and a standard deviation of five.

We run four experiments. All experiments initialize the MH process with $x = 0$ as an approximation of the observable base case. Extracting one sample every 10'000 iterations and collecting in total 1000 samples per experiment (we did not attempt to fine-tune the algorithm) yields the results shown in Figure 2.

Each single point $(\lambda, x)$ represents a policy parameter $\lambda$ and a realization of the approximate corresponding stationary number of alternative-1 users $x$. For $b = 3$ and $b = 2.5$, a unique path to the policy case $\lambda = 1$ is identified. At $b = 2.0$, a bifurcation arises, and the MH algorithm succeeds to explore all branches of this bifurcation. At $b = 1.0$, two consecutive bifurcations can be observed. The solutions around $x = 0$ resp. $x = 1000$ are locally stable (may persist for many iterations), and those around $x = 500$ are unstable.

Even though available real data may allow to select the right solution in the base case (i.e. to push the simulation model by calibration towards a state that is compatible with the data), this is not possible in the policy case. Which solution a simulation model will predict in the policy case is, without further analysis, a matter of chance (in that it depends on how the simulation model is internally initialized).

The MH analysis thus yields (also practically) valuable insight. For $b = 3, 2.5$, it supports the prognosis that the policy case will not differ from the base case. For $b = 1, 2$, it indicates that the result of implementing the policy is diametrically ambiguous, indicating a need for policy refinement or at least a very careful approach to assessing its implications.

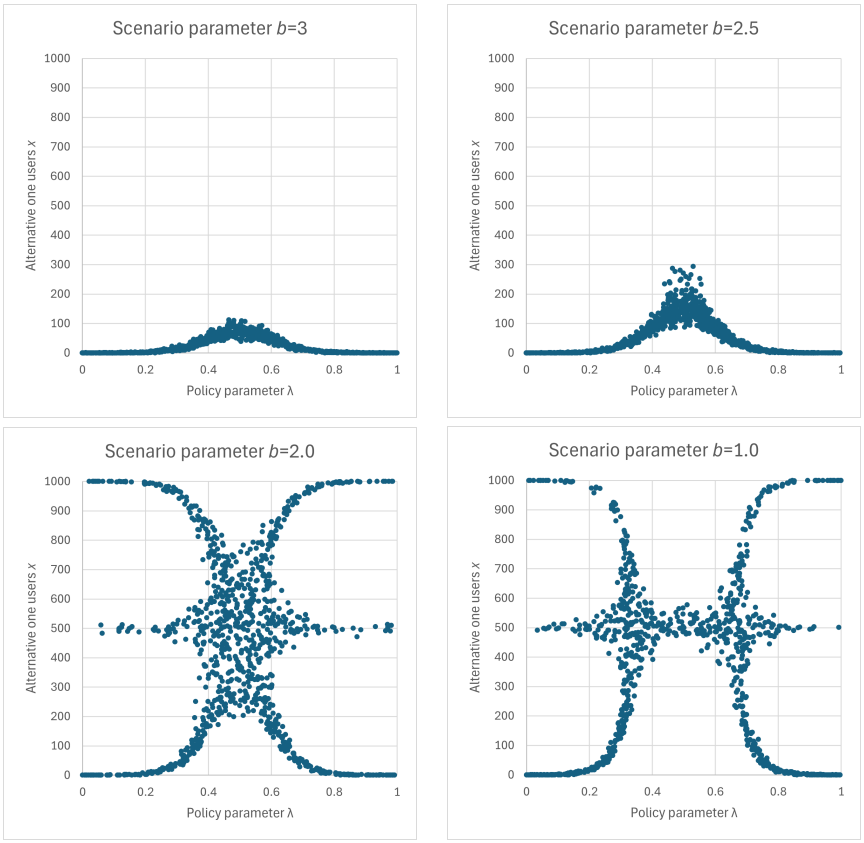



Figure 2: Small example. Alternative-1 users over policy parameter.

# 6 Outlook

We aim to operationalize the analysis presented here for large-scale simulation models in complex scenarios where solution multiplicity can no longer be analyzed by inspecting model structure (as in Iryo and Watling, 2019) or with anecdotal simulation evaluations. The MH machinery is versatile but computationally demanding; this needs to be addressed. We work towards the Swedish national freight model Samgods (Westin et al., 2016) as a real-world test case.

# References


A. Bar-Yosef, K. Martens, and I. Benenson. A model of the vicious cycle of a bus line. *Transportation Research Part B: Methodological*, 54:37–50, 2013.

W. K. Hastings. Monte Carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

T. Iryo and D. Watling. Properties of equilibria in transport problems with complex interactions between users. *Transportation Research Part B*, 126: 87–114, 2019.

J.-D. Schmöcker, T. Hatori, and D. Watling. Dynamic process model of mass effects on travel demand. *Transportation*, 41:279–304, 2014.

D. Watling. Asymmetric problems and stochastic process models of traffic assignment. *Transportation Research Part B*, 30(5):339–357, 1996.

J. Westin, I. Vierth, G. de Jong, R. Karlsson, N. Krüger, and M. Johansson. Analyzing model uncertainty and economies of scale of the swedish national freight model to changes in transport demand. *European Journal of Transport and Infrastructure Research*, 16(4), 2016.