Submitted to *Transportation Science*
manuscript (Please, provide the manuscript number!)

Estimating and Mitigating the Congestion Effect of Curbside Pick-ups and Drop-offs: A Causal Inference Approach

Xiaohui Liu
Department of Information Systems and Analytics, School of Computing, National University of Singapore, Singapore
xiaohuiliu@u.nus.edu

Sean Qian
Department of Civil and Environmental Engineering & H. John Heinz III Heinz College, Carnegie Mellon University, Pittsburgh, PA, USA, seanqian@cmu.edu

Wei Ma
Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong SAR
wei.w.ma@polyu.edu.hk

Curb space is one of the busiest areas in urban road networks. Especially in recent years, the rapid increase of ride-hailing trips and commercial deliveries has induced massive pick-ups/drop-offs (PUDOs), which occupy the limited curb space that was designed and built decades ago. These PUDOs could jam curb utilization and disturb the mainline traffic flow, evidently leading to significant societal externalities. However, there is a lack of an analytical framework that rigorously quantifies and mitigates the congestion effect of PUDOs in the system view, particularly with little data support and involvement of confounding effects. In view of this, this paper develops a rigorous causal inference approach to estimate the congestion effect of PUDOs on general networks. A causal graph is set to represent the spatio-temporal relationship between PUDOs and traffic speed, and a double and separated machine learning (DSML) method is proposed to quantify how PUDOs affect traffic congestion. Additionally, a re-routing formulation is developed and solved to encourage passenger walking and traffic flow re-routing to achieve system optimal. Numerical experiments are conducted using real-world data in the Manhattan area. On average, 100 additional units of PUDOs in a region could reduce the traffic speed by 3.70 and 4.54 mph on weekdays and weekends, respectively. Re-routing trips with PUDOs on curbs could respectively reduce the system-wide total travel time by 2.44% and 2.12% in Midtown and Central Park on weekdays. Sensitivity analysis is also conducted to demonstrate the effectiveness and robustness of the proposed framework.

Key words: Curbside management; Curbside pick-up and drop-off; Causal inference, Double and separated machine learning; Causal graph; Spatio-temporal data analytics; Machine learning

1. Introduction

1.1. Motivation

In addition to roads and intersections, curb space is becoming a new conflicting area where multiple traffic flow converges and interacts (Mitman et al. 2018). Curb space serves various traffic modes such as car parking, truck loading, scooters, and passenger pick-ups/drop-offs (Mitman et al. 2018, Jaller et al. 2021). In recent years, substantial concerns about the congestion effect caused by curbside passenger pick-ups/drop-offs (PUDOs) have arisen (Jaller et al. 2021, Erhardt et al. 2019, Golias and Karlaftis 2001), and this study focuses on mitigating such concerns. The PUDO refers to the behavior that passengers get on and off the vehicles on curb space. Although the action of the curbside PUDO only takes about one minute (Erhardt et al. 2019, Lu 2019, Jaller et al. 2021, Rahaim 2019), it could induce traffic congestion by disturbing traffic flow and occupying curb space, as shown in Figure 1. The reasons are two-fold: 1) PUDOs force vehicles to leave and rejoin the main traffic stream frequently, which disrupts vehicles on main roads (Goodchild et al. 2019, Golias and Karlaftis 2001, Erhardt et al. 2019, Chai and Rodier 2020); 2) PUDOs can be viewed as temporary parking on curb space (Schaller et al. 2011). If the curb space is extensively filled with PUDOs (Butrina et al. 2020), vehicles will spillover to main roads and induce extra delay.

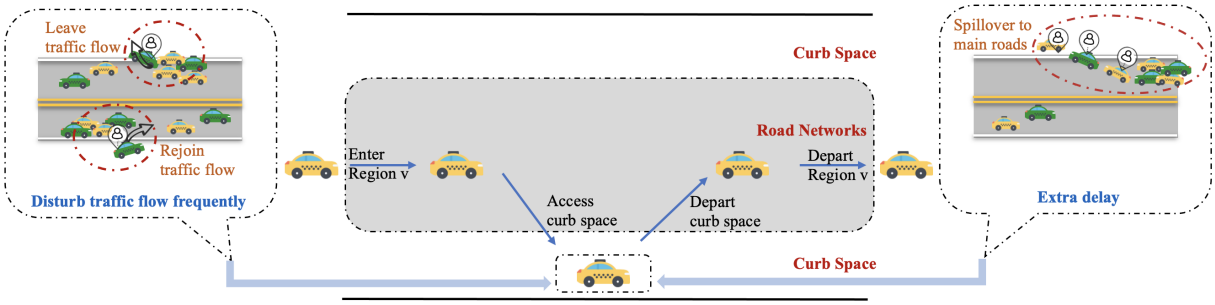


Figure 1: Illustration of congestion effect caused by PUDOs.

With the proliferation of ride-hailing services, massive orders generate excessive PUDOs on limited curb space, which further exacerbate the congestion effect caused by PUDOs. In 2019, Uber owns 111 million users and has completed 6.9 billion trip orders (Iqbal 2019). Each order always starts with a passenger's pick-up and ends with a drop-off. Some studies (Wijayaratna 2015, Erhardt et al. 2019) model the congestion effect of PUDOs as the average duration of each PUDO blocking or disturbing the traffic flow on curb space. In 2016, when the Transportation Network Companies (TNCs) started to provide services, the average duration of TNC PUDOs is 144.75s on major arterial and 79.49s on the minor arterial in San Francisco (Erhardt et al. 2019). However, the PUDO duration is around 60s when TNCs do not exist in 2010 (Erhardt et al. 2019, Lu 2019,

Jaller et al. 2021, Rahaim 2019). The time delay caused by PUDO activities has ranked the third among temporary loss of transportation capacity (TLC) events, right behind vehicular crashes and the presence of work zones (Han et al. 2005). With the further development of the Mobility as a Service (MaaS), it is foreseeable that the number of PUDOs will keep increasing in the near future (Smith et al. 2019). Therefore, it is challenging for public agencies to allocate the limited curb space to accommodate the massive PUDOs. In this paper, we focus on the PUDOs of ride-hailing vehicles as it accounts for the majority of all PUDOs in urban cities.

Although the omnipresent PUDOs play a significant role in traffic congestion (Goodchild et al. 2019), to the best of our knowledge, there are few studies aiming to understand the congestion effect of curbside PUDOs. From the perspective of the ride-hailing service, both PUDOs and vehicles cruising are its by-products. However, PUDOs have not received as much attention as the cruising (Xu, Yin, and Zha 2017, Xu, Chen, and Yin 2019, Zhang and Nie 2021). Many recent studies have pointed out that ride-hailing services are contributing to traffic congestion by exploiting more public resources than driving and public transits. The public resources are not just roads, but also curb space (Castiglione et al. 2016, 2018, Erhardt et al. 2019, Agarwal, Mani, and Telang 2019, Tirachini 2020, Beojone and Geroliminis 2021, Zhang et al. 2021). From the perspective of traffic operation and management, curbside parking has been extensively studied in recent years. For example, Arnott and Rowse (2013) consider parkers' heterogeneity in a high or low value of time and propose that adding curbside parking time limits can reduce the number of parkers with long-time parking and eliminate wasteful cruising for parking. Liu, Zhang, and Yang (2021) model the joint equilibrium of destination and parking choices given public curbside and private shared parking. In contrast, the PUDO, as a temporary form of parking, needs a better understanding and further investigation. The most recent study for evaluating the congestion effect of PUDOs is conducted by Goodchild et al. (2019), which shows that increasing one-unit PUDO could reduce the traffic speed by around 1%, and the estimation result is statistically significant. The proposed estimation method is for one single region, while approaches for estimating the network-wide congestion effects are still lacking.

The current practice to manage PUDOs relies on expert experience and heuristics. Ride-hailing PUDOs have not emerged as a major problem until 2012 (Zalewski, Buckley, and Weinberger 2012, Butrina et al. 2020), and currently, governments, TNCs, and researchers turn their attention to the management of curb space due to chaotic phenomenons caused by PUDOs (Smith et al. 2019, Zhang and Nie 2021, Castiglione et al. 2018, Goodchild et al. 2019, Schaller et al. 2011, Anurag et al. 2019, Lu 2019). For example, airports like JFK prepare a specific area for the PUDOs of ride-hailing vehicles (RVs). Some airports (*e.g.*, LAX) directly ban the curbside PUDOs by RVs. In general, various operation and management strategies can be adopted to mitigate the PUDOs'

effects, including traffic flow re-routing, pricing (Liu, Ma, and Qian 2022), curb space allocation (Goodchild et al. 2019), and curb space re-design (McCormack et al. 2019). Jaller et al. (2021) also propose to utilize curb space as a flex zone where multiple vehicles can occupy a different proportion of curb space at different time periods and locations. However, how to incorporate the precise estimation of the congestion effect of PUDOs into the management framework is worth further exploration. In this paper, we explore the possibility of using a traffic flow re-routing strategy to mitigate the overall congestion caused by PUDOs. The key idea is to shift the PUDOs from the areas with high congestion effects to the areas with low congestion effects so that the city-wide total travel time can be reduced.

In summary, this paper aims to estimate and reduce the congestion effect of PUDOs, and the following two research questions will be addressed:

- How to estimate the congestion effect caused by PUDOs from actual traffic data?
- How to manage PUDOs to minimize the city-wide total travel time based on the differences in congestion effects among regions?

1.2. Challenges and opportunities

This section explains the challenges and difficulties in estimating the congestion effect of PUDOs. With an accurate estimation of the congestion effect, the corresponding management strategies can be developed conveniently using network modeling approaches. First of all, we define the number of PUDOs (NoPUDO) as the total number of pick-ups and drop-offs in a region within a fixed time interval. Without loss of generalization, this paper focuses on the average congestion effect of PUDOs, while the proposed framework can be used for PU and DO separately. Secondly, we use average traffic speed in a region to represent its congestion levels. Specifically, lower traffic speed indicates a more serious congestion level. Therefore, the congestion effect of PUDOs can be quantitatively measured as the change of speed induced by the change of NoPUDO.

However, it is challenging to capture such congestion effects because both speed and NoPUDO are mutually affected by other latent factors, such as travel demands. An illustration of the relationship among travel demands, NoPUDO, and traffic speed is shown in Figure 2. In general, the PUDO has a negative effect on traffic speed, which is our estimation target. However, the growing travel demands can stimulate more ride-hailing requests, making PUDOs happen more frequently. Simultaneously, the increasing travel demands also slow down traffic speed because more vehicles are occupying roads (Yuan, Knoop, and Hoogendoorn 2015, Retallack and Ostendorf 2019).

If the latent effect of traffic demand is overlooked and we directly estimate the relationship between NoPUDO and traffic speed, the congestion effect can be overestimated. We use Example 1 to illustrate how the overestimation arises.

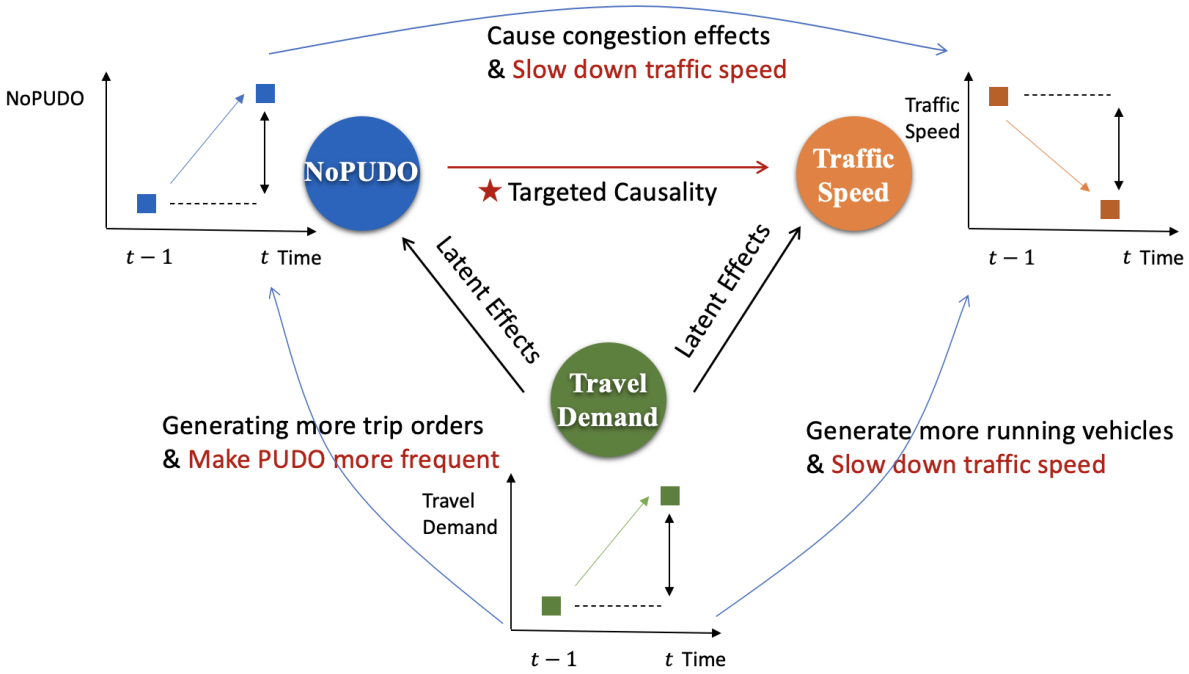


Figure 2: Relationship among travel demands, NoPUDO and traffic speed.

EXAMPLE 1. For a specific time interval t, suppose 100 additional travelers arrive in a specific region, 20 of them take RVs and the rest 80 travelers drive by themselves. The 20 travelers will get dropped off on curb space, while the 80 travelers directly park in the garage. Due to the increasing traffic demand, the average speed in the region reduces by 2 mph. The sources inducing speed reduction is actually two-fold: 1) the congestion induced by the 100 vehicles (both RVs and conventional cars) on the roads; 2) the congestion effect induced by the 20 PUDOs. For the speed reduction, we suppose that the former accounts for 1.5 mph, and the latter accounts for 0.5 mph, then the congestion effect of a PUDO can be calculated as $0.5/20 = 0.025$ mph/PUDO. However, if we directly evaluate the relationship between NoPUDO and traffic speed without identifying the causal relationship, then the congestion effect of a PUDO is *wrongly* calculated as $2/20 = 0.1$ mph/PUDO, and hence the congestion effect is over-estimated.

Essentially, what Example 1 demonstrates is the difference between correlation and causality between NoPUDO and traffic speed. The problem of estimating the congestion effect of PUDOs is indeed a causal effect estimation problem. Specifically, it can be formulated as the problem of quantifying how the change of NoPUDO will induce the changes in traffic speed after eliminating other latent factors (Greenwood, Wattal et al. 2017, Burtch, Carnahan, and Greenwood 2018, Babar and Burtch 2020). One intuitive solution to measure the congestion effect is to conduct field experiments in the real world, while it is practically demanding and costly. In recent years,

advanced machine learning (ML) models empower us to infer the causal effect from observational data without intrusive experiments (Pearl 2019).

Casual inference consists of two major tasks: 1) casual discovery; 2) casual estimation. For a comprehensive review of using ML models for causal inference, readers can refer to Yao et al. (2021). This paper focuses on estimating the causal effects, and we assume that the causal relationship has been identified. Some representative models for casuality estimation include inverse-propensity scoring (IPS) methods, meta-learners, deep representation-based methods, and double machine learning (DML). The IPS methods require estimating the probability of occurrence of each data point, which can be challenging in traffic applications. The meta-learner methods include T-learner, S-learner, X-learner, etc (Künzel et al. 2019), but these methods are more suitable for binary treatments. The deep representation-based methods lack theoretical guarantees (Yao et al. 2019), making them less reliable for engineering applications. The closest work to this paper is double machine learning (DML), which can estimate the casual effects by training two ML models (Wager and Athey 2018, Oprescu, Syrgkanis, and Wu 2019). This method has rigorous theoretical guarantees on the estimation quality (Chernozhukov et al. 2018), and hence it is suitable for engineering applications. However, the standard DML cannot model the interwoven causal relationship between NoPUDO and traffic speed, especially when such a relationship is convoluted with both time and space. A novel method needs to be developed to consider the spatio-temporal patterns of both NoPUDO and traffic speed when estimating the network-wide congestion effects of PUDOs.

1.3. Contributions

Overall, there lacks a quantitative method to estimate the congestion effect of PUDOs on the traffic speed using observational traffic data, and how the estimated congestion effect can be used for traffic management is worth investigating. To fill up the research gaps, this paper proposes a data-driven framework to evaluate and manage the spatio-temporal congestion effects of PUDOs using multi-source traffic data. This paper first rigorously analyzes the causal relationship between NoPUDO and traffic speed. Next, we develop the Double and Separated Machine Learning (DSML) method to estimate the congestion effect of PUDO. A re-routing strategy is further formulated and solved by re-distributing the PUDO from busy regions to less busy regions, thereby mitigating the overall traffic congestion. Lastly, the proposed framework is examined with real-world data in the large-scale network in the New York City to demonstrate its effectiveness.

The contributions of this paper can be summarized as follows:

- To the best of our knowledge, this is the first study to use the causal inference approach to estimate the congestion effect of PUDOs from a data-driven perspective.

- This study rigorously formulates a causal graph to articulate the spatio-temporal relationship between the NoPUDO and traffic speed. A novel double and separated machine learning (DSML) method is developed and theoretically analyzed for estimating the congestion effect of PUDOs based on the causal graph.
- We develop a re-routing formulation to re-distribute PUDOs to minimize the network-wide total travel time, and a customized solution algorithm is developed to effectively solve the formulation.
- The developed framework is validated with real-world data in a large-scale network in the Manhattan area. The estimation results obtained by the DSML method are consistent with actual traffic conditions, and re-routing trips with PUDOs can effectively reduce the network-wide total travel time.

The remainder of this paper is organized as follows. Section 2 discusses the causal estimation framework, which consists of the causal graph, DSML, and the re-routing formulation. Section 3 presents the solution algorithms, and section 4 exhibits the numerical experiments on the Manhattan area. Finally, conclusions are drawn in Section 5.

2. Model

In this section, we first develop a causal graph to model the spatio-temporal relationship between NoPUDO and traffic speed and mathematically formulate the structural equation models. Secondly, the Double and Separated Machine Learning (DSML) method is developed and analyzed for the causal graph. Thirdly, a system-optimal problem is formulated and solved to minimize the total travel time by re-routing PUDOs from the current region to neighboring regions. Notations used in this paper are summarized in Table 6, and each notation will also be introduced when it first appears.

2.1. Causal relationship between NoPUDO and Traffic Speed

In this section, we first analyze the causal graph of the NoPUDO and traffic speed. This causal relationship is then mathematically formulated.

2.1.1. Causal graph Traffic states of a city are modeled by a spatial random variable that evolves over time, $\{y_v^t \in \mathbb{R}^+, t \in \mathbb{T}\}$, where v is a region in $\mathcal{G}$, $v \in \mathcal{V}$, and $\mathcal{G}$ denotes the multi-community city consisting of the set of regions $\mathcal{V}$ (Liu, Zhang, and Yang 2021). $\mathbb{T}$ is the set of time intervals, and y_v^t is the quantitative measures of traffic states (*e.g.*, speed or flow) in the region v and the time interval t (He et al. 2016). Besides, we use d_v^t to denote the NoPUDO in the region v and time interval t. Without loss of generality, this paper models the average effect of the total NoPUDO on traffic speed in a given region v. We can further extend the proposed framework to consider other traffic states, such as flow, density, travel time, etc.

As discussed in the Section 1.2, the relationship between NoPUDO and traffic speed is convoluted with latent factors such as travel demands. In addition, the estimation of the congestion effect should also consider the temporal and spatial features of traffic scenes. Given a time interval t and a region v, we summarize the interwoven causal relationship between the NoPUDO and traffic states as follows:

- First, the traffic speed y_v^t is affected by its historical traffic speed records $\mathbf{Y}_v^{t-I:t-1}$. Because the traffic speed changes gradually throughout the day, the historical traffic speed $\mathbf{Y}_v^{t-I:t-1}$ can reflect the congestion levels, and passengers may refer to the past speed records to avoid picking up or dropping off in the congested regions. Hence $\mathbf{Y}_v^{t-I:t-1}$ is a critical factor for predicting the traffic speed.
- Second, the traffic speed y_v^t in the region v is also affected by the traffic speed of its surrounding regions during the past time window $\mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$, due to traffic flow exchanges. For example, if the neighboring regions $\mathcal{N}(v)$ of the region v is congested by traffic accidents, the accumulated vehicles will spillover to the region v. The consideration of the surrounding traffic state actually manifests the importance of spatial correlation in causality estimation.
- Third, the NoPUDO d_v^t is affected by its historical NoPUDO $\mathbf{D}_v^{t-I:t-1}$ in region v from the time interval $t-I$ to the time interval $t-1$. Similar to the traffic speed prediction, the historical NoPUDO $\mathbf{D}_v^{t-I:t-1}$ reflect the demand levels, and hence it is critical for predicting d_v^t.
- Fourth, external control variables $\mathbf{W}_v^t$, such as weather, holidays, peak hours, and so on, also affect the traffic speed and NoPUDO. For instance, rain and snow may limit drivers' sight, therefore making traffic speed slower and travel time longer (Ahmed and Ghasemzadeh 2018). Besides, holidays may stimulate more trip orders around places of interest than usual (Rong, Cheng, and Wang 2017), which accumulates more NoPUDO. Therefore, these external control variables should be considered to eliminate potential biases in causality estimation.

Additionally, we assume Assumption 1 holds as the congestion effect of PUDOs is immediate and the effect duration is short.

Assumption 1. *For region v in the network $\mathcal{G}$, the average traffic speed y_v^t in the time interval t is not causally affected by the historical records of the NoPUDO $\mathbf{D}_v^{t-I:t-1}$.*

In short, the continuity of time, interactivity in space, and extra influence caused by external variables make the causality estimation between NoPUDO and traffic speed more dynamic and intricate.

Combining the above discussion and Assumption 1, we develop the causal graph, as shown in Figure 3, to depict the causal relationship of PUDOs and traffic speed in both time and space dimensions. It is worth noting that, for region v, the NoPUDO in $\mathcal{N}(v)$ does not causally affect

$d_v^t, y_v^t, \forall t$, as travelers cannot go to two regions at the same time. We believe conditioning on $\mathbf{Y}_v^{t-I:t-1}$, $\mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$, $\mathbf{D}_v^{t-I:t-1}$, $\mathbf{W}_v^t$, $\mathbf{D}_{\mathcal{N}(v)}^{t-I:t-1}$ is independent of d_v^t and y_v^t.

The proposed causal graph contains two random variables y_v^t and d_v^t, as we omit $\mathbf{W}_v^t$ for simplicity. To show the causal relationship more clearly, in Figure 3, we expand to draw $\mathbf{Y}_v, \mathbf{Y}_{\mathcal{N}(v)}, \mathbf{D}_v$ and demonstrate how they affect y_v^t and d_v^t. We note that $\mathbf{Y}_v^{t-I:t-1}$, $\mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$, and $\mathbf{D}_v^{t-I:t-1}$ are actually combinations of $y_v^{t'}$ and $d_v^{t'}, \forall t' < t$.

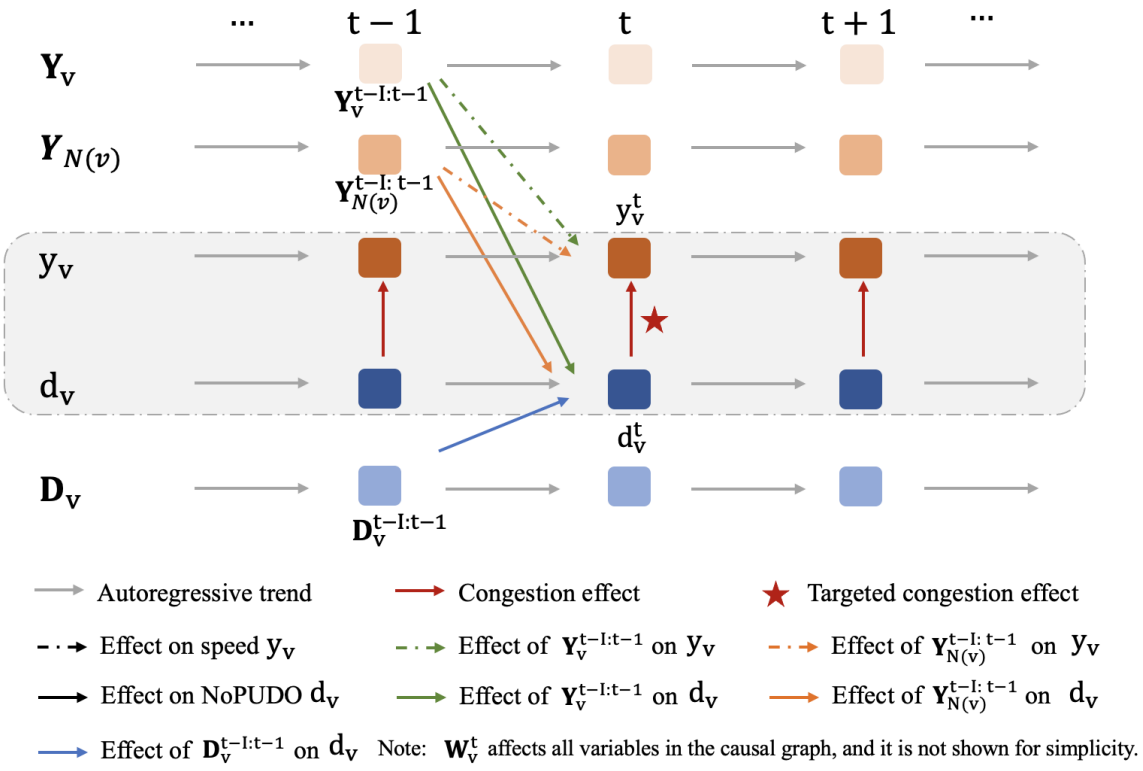


Figure 3: The causal graph of the NoPUDO and traffic speed.

The red solid line with a red star from d_v^t to y_v^t indicates the causal effect of PUDOs on traffic speed, which is the estimation target. Specifically, the effect θ_v is represented by the change of current speed y_v^t induced by increasing one additional unit of PUDO in the region v, given other variables unchanged. The green dotted line from $\mathbf{Y}_v^{t-I:t-1}$ to y_v^t denotes the effect of traffic speed during the past time windows $t-I:t-1$ in the time interval t, and the orange dotted line from $\mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$ to y_v^t represents the effect of the speed in surrounding regions $\mathcal{N}(v)$ on the current region v. The reason for both two dotted lines here is that traffic state y_v^t is affected by both historical traffic speed from the time interval $t-I$ to $t-1$ in the region v ($\mathbf{Y}_v^{t-I:t-1}$) and traffic speed from the time interval $t-I$ to $t-1$ in neighboring regions $\mathcal{N}(v)$ ($\mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$). The blue solid line represents the

effect of PUDOs in the past time windows on that in the current time interval, as the NoPUDO in the time interval t and region v (d_v^t) is influenced by its historical trends from the time interval $t-I$ to the time interval $t-1$ in the region v, denoted by $\mathbf{D}_v^{t-I:t-1}$. Additionally, the green and orange solid lines represent the effects of historical traffic speed ($\mathbf{Y}_v^{t-I:t-1}, \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$) on the NoPUDO.

2.1.2. Structural equation models In this section, we rigorously formulate the congestion effect caused by PUDOs. We define θ_v to be the changes in traffic speed y_v caused by a one-unit change of NoPUDO in the region v. In this paper, we use traffic speed to represent the traffic conditions, while we can also use other traffic-related variables, such as flow, density, and occupancy. In our case, lower speed indicates that traffic conditions tend to become congested. Mathematically, θ_v is defined based on Assumption 2.

Assumption 2 **(Linear effects)**. *For a specific region v, given fixed $\mathbf{Y}_v^{t-I:t-1}, \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}, \mathbf{W}_v^t$, the congestion effect θ_v is defined in Equation 1.*

$$y_v^t|_{\texttt{do}(d_v^t=d_1)} - y_v^t|_{\texttt{do}(d_v^t=d_2)} = \theta_v(d_1 - d_2) \tag{1}$$

where $\texttt{do}(\cdot)$ is the do-operation defined in Pearl (2009), and d_1 and d_2 are two arbitrarily positive integers representing the NoPUDO.

One can read from Equation 1 that the effect of PUDOs on traffic speed is linear in each region v. The linear relationship means that adding an additional unit of NoPUDO will make traffic speed increase by θ_v in the region v. Additionally, we expect that $\theta_v \leq 0$ because the increase of the NoPUDO could induce more congestion.

Generally, different regions in a city are equipped with different population densities, economic statuses, and traffic conditions. These factors will all contribute to the fluctuation of the estimated congestion effect caused by PUDOs in different regions. We assume the homogeneity within each region, and it means that the congestion effect caused by PUDOs (*i.e.*, θ_v) is constant within a region. Therefore, we conduct the causal analysis based on the regional level.

To better understand Equation 1, we note that the following two remarks hold for θ_v as a result of Assumption 2.

Remark 1 (Constant effects within a region). The congestion effect is constant within each region and across different time intervals. In other words, for each region v, θ_v does not depend on the time intervals in which the PUDO happens.

Remark 1 simplifies the problem of congestion effect estimation to a static problem, and the time variation is not considered. To estimate the time-varying congestion effect, we can run the proposed framework multiple times using the observed data in each time interval. In this paper, we estimate the θ_v for weekdays and weekends respectively.

Remark 2 (Independent effects across different regions). For each region v, θ_v is not affected by other regions, and θ_v is only related to the attributes and properties of region v. One can see that Remark 2 ensures that the estimation of θ_v can be conducted for each region v separately. If the remark is violated, it is also straightforward to extend the estimation framework presented in this paper to the conditional average treatment effect (CATE) (Abrevaya, Hsu, and Lieli 2015).

Given the causal graph in section 2.1.1 and Assumption 2, we are now ready to formulate the causal relationship between NoPUDO and traffic speed in Equation 2 and Equation 3.

$$y_v^t = \varphi_v(\mathbf{Y}_v^{t-I:t-1}; \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}; \mathbf{W}_v^t) + \theta_v \cdot d_v^t + e_v^t \tag{2}$$

$$d_v^t = \psi_v\left(\mathbf{D}_v^{t-I:t-1}, \mathbf{Y}_v^{t-I:t-1}, \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}; \mathbf{W}_v^t\right) + \xi_v^t \tag{3}$$

where φ_v predicts the traffic speed y_v^t using historical traffic speed records, and ψ_v predicts the NoPUDO d_v^t using historical traffic speed as well as the historical NoPUDO. Both e_v^t and ξ_v^t are zero-mean noise, which are defined in Equation 4 and 5.

Equation 2 and 3 can be viewed as a Structural Equation Model (SEM): traffic speed y_v^t is the outcome variable, the NoPUDO d_v^t is the treatment variable, and $\mathbf{D}_v^{t-I:t-1}, \mathbf{Y}_v^{t-I:t-1}, \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}, \mathbf{W}_v^t$ are control variables. Specifically, θ_v is the treatment effect that shows the effect of the NoPUDO d_v^t on traffic speed y_v^t. The inclusion of control variables can help to eliminate the biased influence of other factors on the estimation results.

One can see that Equation 2 and 3 characterize the causal relationship between NoPUDO and traffic speed in a spatio-temporal manner, and the above equations are consistent with the causal graph discussed in section 2.1.1. We further assume that the random errors e_v^t and ξ_v^t follow Assumption 3.

Assumption 3 **(Independent Noise)**. *For any time interval t and region v, we have the following equations hold.*

$$\mathbb{E}[e_v^t | \mathbf{Y}_v^{t-I:t-1}; \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}; d_v^t; \mathbf{W}_v^t] = 0 \tag{4}$$

$$\mathbb{E}[\xi_v^t | \mathbf{D}_v^{t-I:t-1}; \mathbf{Y}_v^{t-I:t-1}; \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}; \mathbf{W}_v^t)] = 0 \tag{5}$$

$$e_v^t \overset{iid}{\sim} U^e\left(\mathbf{Y}_v^{t-I:t-1}; \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}; d_v^t; \mathbf{W}_v^t\right) \tag{6}$$

$$\xi_v^t \overset{iid}{\sim} U^\xi\left(\mathbf{D}_v^{t-I:t-1}; \mathbf{Y}_v^{t-I:t-1}; \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}; \mathbf{W}_v^t\right) \tag{7}$$

where $\overset{iid}{\sim}$ means the noise is independent and identically distributed. U^e and U^ξ are unknown and parameterized zero-mean distributions.

Intuitively, Assumption 3 indicates that unknown random error in y_v^t and d_v^t are zero-mean and independent. Hence the two functions (φ_v, ψ_v) and congestion effect θ_v could capture the causal relationship between speed and NoPUDO.

Based on the above formulation, we prove that when the traffic speed y_v^t, NoPUDO d_v^t, and external control variables $\mathbf{W}_v^t$ are observable, it is theoretically sufficient to estimate θ_v, as presented in Proposition 1.

PROPOSITION 1 **(Identifiable)**. *Suppose that Equation 2, 3, 4, and 5 hold and y_v^t, d_v^t, and $\mathbf{W}_v^t$ are observable for all v, t, then θ_v is identifiable,* i.e., *θ_v can be uniquely estimated from $y_v^t, d_v^t, \mathbf{W}_v^t, \forall v, t$.*

Proof. First, given $y_v^t, d_v^t, \mathbf{W}_v^t, \forall v, t$ are observable, we have $\mathbf{Y}_v^{t-I:t-1}, \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}, \mathbf{D}_v^{t-I:t-1}$ are also observable. Second, in the time interval t and for any region v, we consider the ordered pair of variables (d_v^t, y_v^t), and we define $\mathcal{Z} = \{\mathbf{Y}_v^{t-I:t-1}, \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}, \mathbf{D}_v^{t-I:t-1}, \mathbf{W}_v^t\}$. We claim that $\mathcal{Z}$ satisfies the back-door criterion relative to (d_v^t, y_v^t). The reason is: in the causal graph presented in Figure 3:

- No node in $\mathcal{Z}$ is a descendant of d_v^t;
- $\mathcal{Z}$ blocks every path between d_v^t and y_v^t that contains an arrow into y_v^t.

Based on Theorem 3.3.2 in Pearl (2009), the congestion effect θ_v is identifiable, and hence θ_v can be uniquely estimated based on the Definition 3.2.3 in Pearl (2009). □

2.2. Double and Separated Machine Learning

In this section, we propose a novel method to estimate the congestion effect of PUDOs θ_v based on Equation 2 and 3. As we discussed in 2.1.1, the challenge in estimating θ_v lies in the complex spatio-temporal relationship between traffic speed and NoPUDO, as shown in Equation 2 and 3. To accurately model such a spatio-temporal relationship, both φ_v and ψ_v need to be generalized as non-linear functions that can model the arbitrary relationship between the traffic speed and NoPUDO. ML models shed light on modeling the non-linear relationship among variables with simple model specifications, and hence we propose to employ ML methods to learn both φ_v and ψ_v using massive data.

When both φ_v and ψ_v are modeled as non-linear ML models, directly estimating θ_v becomes challenging. The main reason is that most ML models are biased due to model regularization (Hastie et al. 2009). With the biased estimation of φ_v and ψ_v, we need to estimate θ_v in an unbiased manner, and this presents challenges for the model formulation. To this end, we propose the Double and Separated Machine Learning (DSML) method with consideration of the potential biases in the ML models for φ_v and ψ_v. The proposed DSML method consists of three sub-models: 1) Model `Y` learns φ_v and predicts the traffic speed y_v^t; 2) Model `D` learns ψ_v and predicts the NoPUDO d_v^t; and 3) Model `Z` estimates the congestion effect of PUDOs on traffic speed.

The relationship among the three sub-models is presented in Figure 4. To be specific, we present each model as follows.

- Model `Y`, which is denoted as $\hat{\varphi}_v$, predicts speed y_v^t based on historical speed record $\mathbf{Y}_v^{t-I:t-1}$ in current region v, $\mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$ in the neighboring regions $\mathcal{N}(v)$, and external control variables $\mathbf{W}_v^t$, without considering the congestion effect of NoPUDO.
- Model `D`, which is denoted as $\hat{\psi}_v$, predicts the NoPUDO d_v^t based on historical record of NoPUDO $\mathbf{D}_v^{t-I:t-1}$, speed record $\mathbf{Y}_v^{t-I:t-1}$, $\mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$, and external control variables $\mathbf{W}_v^t$.
- Model `Z` fits a linear regression model from the residuals of Model `D` to the residuals of Model `Y`, and the slope is the estimation of θ_v. Proof and intuitive explanations will be provided in the following sections.

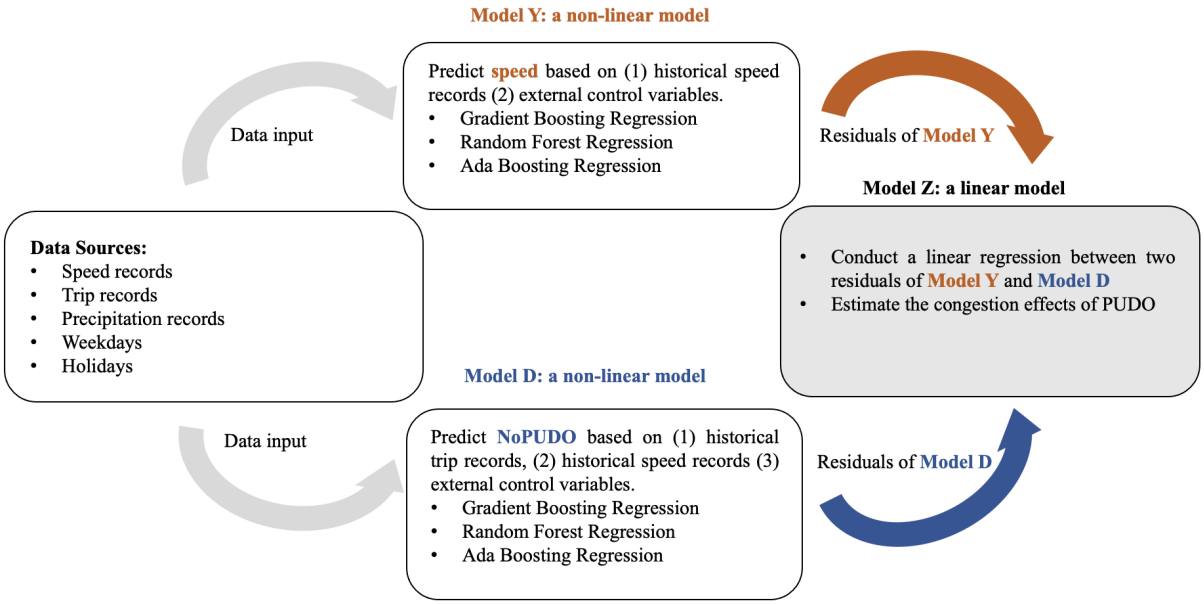


Figure 4: The framework of the DSML method.

In each sub-model, both spatial and temporal dependencies of variables are considered. One can see that in the DSML method, the task of estimating θ_v is decomposed into Model `Y`, `D`, and `Z` respectively. We note that the DSML method is an extension of the generalized Double Machine Learning (DML) method (Chernozhukov et al. 2018), and the DSML method is specifically designed for the congestion effect estimation using the causal graph in Figure 3. In the following sections, we present each sub-model in detail.

2.2.1. Model Y Model Y predicts the traffic speed using historical speed data without considering the congestion effect caused by PUDOs, as formulated in Equation 8.

$$\hat{y}_v^t = \hat{\varphi}_v(\mathbf{Y}_v^{t-I:t-1}; \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}; \mathbf{W}_v^t) \tag{8}$$

where, $\hat{y}_v^t$ is the predicted speed in the time interval t and region v. Three input variables include a vector of history speed record $\mathbf{Y}_v^{t-I:t-1}$ from the time interval $t-I$ to the time interval $t-1$ in the region v, historical average speed record $\mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$ from the time interval $t-I$ to the time interval $t-1$ at neighboring regions $\mathcal{N}(v)$, and the external control variables $\mathbf{W}_v^t$. $\hat{\varphi}_v$ is the function that maps these input variables to the speed y_v^t, which can be learned by ML models using massive observed data.

The residual of Model Y, $\hat{\epsilon}_v^t$ is the difference of the predicted value $\hat{y}_v^t$ and the true value y_v^t, as shown in Equation 9.

$$\hat{\epsilon}_v^t = y_v^t - \hat{y}_v^t \tag{9}$$

The residual $\hat{\epsilon}_v^t$ deserves more attention, as it is a random variable that consists of two sources of variation: 1) the changes of y_v^t due to the NoPUDO, and 2) the other random noise. Intuitively, $\hat{\epsilon}_v^t = \theta_v d_v^t + [\varphi_v(\cdots) - \hat{\varphi}_v(\cdots)] + e_v^t \approx \theta_v d_v^t + e_v^t$. To extract the θ_v from $\hat{\epsilon}_v^t$, we make use of the Model D to build the correlation between $\hat{\epsilon}_v^t$ and d_v^t.

2.2.2. Model D Model D aims to predict the NoPUDO using the historical traffic speed and NoPUDO, and the formulation is presented in Equation 10.

$$\hat{d}_v^t = \hat{\psi}_v \left(\mathbf{D}_v^{t-I:t-1}, \mathbf{Y}_v^{t-I:t-1}, \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}, \mathbf{W}_v^t \right) \tag{10}$$

where $\hat{d}_v^t$ is the predicted value of NoPUDO in the time interval t and region v.

Based on the causal graph in Figure 3, $\hat{d}_v^t$ not only includes the historical traffic speed ($\mathbf{Y}_v^{t-I:t-1}, \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$), but also considers the influence of the historical NoPUDO ($\mathbf{D}_v^{t-I:t-1}$).

Similarly, $\hat{\psi}_v$ is obtained by training with historical data. One important note is that the training data for Model D should be different from that is used for training Model Y, so that the learned $\hat{\varphi}_v$ and $\hat{\psi}_v$ are independent of each other (Chernozhukov et al. 2018). To this end, the data splitting technique is adopted, and details will be introduced in section 3.1.

The residual $\hat{\xi}_v^t$ of Model D can be computed as the subtraction of true value d_v^t and predicted value $\hat{d}_v^t$ of the NoPUDO, as shown in Equation 11.

$$\hat{\xi}_v^t = d_v^t - \hat{d}_v^t \tag{11}$$

The residual $\hat{\xi}_v^t$ is a random variable, and it contains the proportion of d_v^t that is not affected by the historical traffic speed. Intuitively, $\hat{\xi}_v^t$ and $\hat{\epsilon}_v^t$ are correlated because of the congestion effect of PUDOs, and we have Proposition 2 holds.

PROPOSITION 2. *Given a region v, suppose Equation 2, 3, and Assumption 3 hold, when $\mathbf{D}_v^{t-I:t-1}$, $\mathbf{Y}_v^{t-I:t-1}$, $\mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$, and $\mathbf{W}_v^t$ are observed for any t, we have*

$$\theta_v = 0 \iff \hat{\xi}_v^t \perp\!\!\!\perp \hat{\epsilon}_v^t, \tag{12}$$

where $\perp\!\!\!\perp$ means independence.

Proof. Based on Equation 2 and 11, we have $\hat{\epsilon}_v^t = \theta_v d_v^t + [\varphi_v(\cdots) - \hat{\varphi}_v(\cdots)] + e_v^t$ and $\hat{\xi}_v^t = \left[\psi_v(\cdots) - \hat{\psi}_v(\cdots)\right] + \xi_v^t$, where we use $\cdots$ to represent the input variables. We show the proposition from two directions:

- When $\theta_v = 0$, we have $\hat{\epsilon}_v^t = [\varphi_v(\cdots) - \hat{\varphi}_v(\cdots)] + e_v^t$. Additionally, $[\varphi_v(\cdots) - \hat{\varphi}_v(\cdots)] \perp\!\!\!\perp e_v^t$, which is because e_v^t is iid. Then we have $\hat{\xi}_v^t \perp\!\!\!\perp [\varphi_v(\cdots) - \hat{\varphi}_v(\cdots)]$ due to the data splitting technique, and $\hat{\xi}_v^t \perp\!\!\!\perp e_v^t$ due to Equation 6 and 7. Therefore, $\hat{\xi}_v^t \perp\!\!\!\perp \hat{\epsilon}_v^t$.
- When $\hat{\xi}_v^t \perp\!\!\!\perp \hat{\epsilon}_v^t$, we know $\hat{\xi}_v^t \perp\!\!\!\perp \theta_v d_v^t + [\varphi_v(\cdots) - \hat{\varphi}_v(\cdots)] + e_v^t$. Again $\hat{\xi}_v^t \perp\!\!\!\perp e_v^t$ and $\hat{\xi}_v^t \perp\!\!\!\perp [\varphi_v(\cdots) - \hat{\varphi}_v(\cdots)]$ hold, so $\hat{\xi}_v^t \perp\!\!\!\perp \theta_v d_v^t$. Because $\hat{\xi}_v^t$ and d_v^t are correlated due to Equation 3, then θ_v has to be zero.

Combining the above two directions, we have the proof completed. □

One can see from Proposition 2, the correlation of $\hat{\xi}_v^t$ and $\hat{\epsilon}_v^t$ is closely associated with the value of θ_v. Indeed, θ_v can be estimated from the two residuals, as presented in the next section.

2.2.3. Model `Z` Based on the discussions in the previous two sections, both Model `Y` and Model `D` depict the trends of traffic speed and NoPUDO using the spatio-temporal historical data, respectively. Importantly, all the edges were modeled in the causal graph in Figure 3, except for the congestion effect of PUDOs, which is marked with $\star$. To estimate the congestion effect θ_v, we develop Model `Z` that fits a linear regression model from the residual $\hat{\xi}_v^t$ of Model `D` to the residuals $\hat{\epsilon}_v^t$ of Model `Y`, as represented by Equation 13.

$$\hat{\epsilon}_v^t = \theta_v \hat{\xi}_v^t + \hat{e}_v^t \tag{13}$$

where $\hat{e}_v^t$ represents the random error of the linear regression model. We note that θ_v can be estimated using the Ordinary Least Square (OLS), as presented in Equation 14

$$\hat{\theta}_v = \arg\min_{\theta} \mathbb{E}\left[\sum_{t\in\mathbb{T}} \left(\hat{\epsilon}_v^t - \theta \hat{\xi}_v^t\right)^2\right] \tag{14}$$

We claim that $\hat{\theta}_v$ is an unbiased estimator of θ_v. Before the rigorous proof, we intuitively explain why this claim is true. To this end, the variable $\hat{e}_v^t$ can be derived in Equation 15.

$$\begin{aligned}\hat{e}_v^t = \hat{\epsilon}_v^t - \theta_v \hat{\xi}_v^t &= \theta_v d_v^t + [\varphi_v(\cdots) - \hat{\varphi}_v(\cdots)] + e_v^t - \theta_v \left(\left[\psi_v(\cdots) - \hat{\psi}_v(\cdots) \right] + \xi_v^t \right) \\ &= (\theta_v d_v^t - \theta_v (\psi_v(\cdots) + \xi_v^t)) + [\varphi_v(\cdots) - \hat{\varphi}_v(\cdots)] + \theta_v \hat{\psi}_v(\cdots) + e_v^t \\ &= [\varphi_v(\cdots) - \hat{\varphi}_v(\cdots)] + \left[\theta_v \hat{\psi}_v(\cdots) \right] + e_v^t \\ &\approx [-\theta_v d_v^t] + [\theta_v d_v^t] + e_v^t \\ &= e_v^t \end{aligned} \tag{15}$$

where $\varphi_v(\cdots) - \hat{\varphi}_v(\cdots) = -\theta_v d_v^t$ because $\hat{\varphi}_v(\cdots)$ is a ML model to predict y_v^t, and $\theta_v \hat{\psi}_v(\cdots) = \theta_v d_v^t$ because $\hat{\psi}_v$ is a ML model to predict d_v^t. Therefore, $\hat{e}_v^t$ is zero-mean, and hence θ_v can be estimated using linear regression from $\hat{\xi}_v^t$ to $\hat{\epsilon}_v^t$.

Now we are ready to present Proposition 3, which proves that $\hat{\theta}_v$ is an unbiased estimator of θ_v when φ_v and ψ_v are linear models.

PROPOSITION 3 **(FWL Theorem)**. *For any region v, we suppose Equation 2, 3, and Assumption 3 hold. When φ_v and ψ_v are linear models, $\hat{\theta}_v$ obtained from Equation 14 is an unbiased estimator of θ_v. Mathematically, we have $\hat{\theta}_v = \theta_v$.*

Proof. See Appendix A.1. □

We further extend to consider both φ_v and ψ_v are non-linear functions and can be learned by ML models, as presented in Proposition 4.

PROPOSITION 4. *For any region v, we suppose Equation 2, 3, and Assumption 3 hold. Given both φ_v and ψ_v are learnable by the ML models, we have Equation 16 holds.*

$$\begin{aligned} \tfrac{1}{|\mathbb{T}|} \textstyle\sum_{t \in \mathbb{T}} (\hat{\varphi}_v - \varphi_v)^2 &\xrightarrow{P} 0 \\ \tfrac{1}{|\mathbb{T}|} \textstyle\sum_{t \in \mathbb{T}} \left(\hat{\psi}_v - \psi_v \right)^2 &\xrightarrow{P} 0 \end{aligned} \tag{16}$$

where $\xrightarrow{P}$ represents the convergence in probability. If $\hat{\varphi}_v$ and $\hat{\psi}_v$ are learned with data splitting technique, then $\hat{\theta}_v$ obtained from Equation 14 follows Equation 17.

$$\hat{\theta}_v - \theta_v \sim \mathbf{N}\left(0, \frac{1}{|\mathbb{T}|}\right) \tag{17}$$

where $\mathbf{N}\left(0, \frac{1}{|\mathbb{T}|}\right)$ denotes the normal distribution with mean zero and variance $\frac{1}{|\mathbb{T}|}$.

Proof. See Appendix A.2. □

Both Proposition 3 and 4 support the claim that the DSML method can estimate θ_v in an unbiased manner. Proposition 3 is actually a special case of Proposition 4 with more intuitive explanations, which could help readers better understand the essential idea of the proposed DSML method.

2.3. Re-routing traffic flow with PUDOs to reduce total travel time

In this section, we present to re-route traffic flow with PUDOs to minimize the network-wide total travel time. Currently, PUDOs are mainly concentrated in busy regions such as office buildings, shopping malls, and residential areas. The uneven distribution of PUDOs concentrates congestion on several specific regions (Zhang et al. 2021, Dong et al. 2022). Consequently, one unit of the PUDO will generate a more significant congestion effect in those busy regions, which further exacerbates the congestion. Using the Manhattan area as an example, the $|\theta_v|$ in Midtown is typically higher than that in Upper West Side, and hence the congestion caused by PUDOs in Midtown is more severe.

To reduce total travel time on the entire network, this paper aims to re-route some of the traffic flow with PUDOs to the neighboring regions based on the differences of congestion effects in different regions. To be specific, we allow travelers to 1) walk from their origin regions to the nearby regions and get picked up, and/or 2) get dropped off in nearby regions, and then walk to their destination regions. The underlying idea behind the re-routing strategy is to re-distribute PUDOs from the busy regions to uncongested neighboring regions. Example 2 further illustrates how re-routing strategy reduces the total travel time.

Example 2. Consider a network with 6 regions, which are represented by 6 nodes in Figure 5. Values on each link represent the time cost to drive from the tail to the head of the link. Region 5 is busy, while Region 4 and 6 are less busy and they are neighboring regions of Region 5. Therefore we assume the absolute congestion effect of Region 5, $|\theta_5|$, is larger than that in Region 4 $|\theta_4|$ and Region 6 $|\theta_6|$. An additional passenger departs from Region 1 to Region 5: if the passenger arrives at Region 5 by taxi directly, the average speed in the Region 5 will decrease, and hence the travel time in this region will increase. Instead, if we let the passenger get dropped-off in Region 4 or 6 and walk to Region 5, traffic speed in Region 5 will increase. Although the traffic speed in Region 4 or 6 will be reduced, the caused congestion is less significant given that both regions are less busy.

One can see that this example utilizes the uneven geographical distribution of PUDOs, which is attributed to the common phenomenon of uneven travel demands (Zhang et al. 2021, Dong et al. 2022). The differences of congestion effects in less busy and busy regions can be exploited to re-distribute PUDOs, finally resulting in a decrease of the overall travel time. Specifically, a great number of passengers flock to the same Central Business District (CBD). Even subtle improvement in travel time for each passenger will bring obvious improvements to the entire network.

We consider travelers from region r to region s in the time interval t, and their quantity is denoted as q_{rs}^t. These travelers are divided into two groups according to whether they are re-routed or not. As shown in Figure 6, we assume $\tilde{h}_{rsn}^t$ indicates the number of travelers whose original path is from region r to region s, will be re-routed to drop off in region n, and these travelers

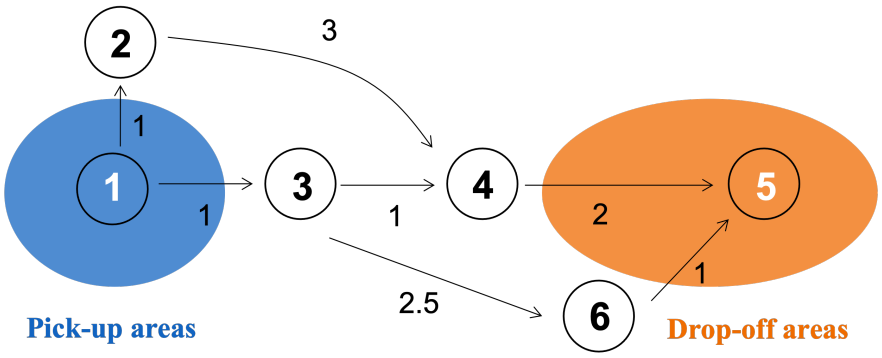


Figure 5: An example of the traffic flow re-routing with PUDOs.

need walk from region n to their final destination s. Other travelers, which is denoted as $\tilde{f}^t_{rs}$, will keep their original routes by vehicles directly. After re-routing, the NoPUDO in each region will be changed, and hence the travel time in each region will adjust according to the congestion effect θ_v. Ultimately, we expect the re-routing of the traffic flow will reduce the total travel time (TTT) on the network.

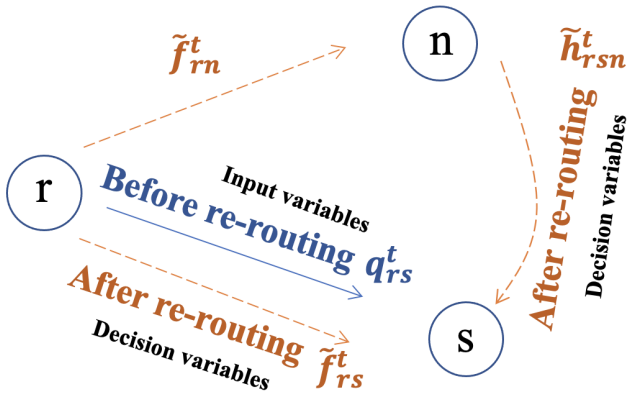


Figure 6: Illustration of variables related to the re-routing formulation.

To ensure the feasibility of the re-routing strategy, we limit the walking distance and assume that the drop-off region n belongs to destination region's neighboring regions, *i.e.*, $n \in \mathcal{N}(s)$, where $\mathcal{N}(s)$ represents the set of neighboring regions of region s. The mathematical formulation for re-routing the traffic flow with PUDOs in the time interval t is as presented in Formulation 18.

$$
\begin{aligned}
\min_{\{\tilde{f}^t_{rs}\}_{rst},\{\tilde{h}^t_{rsn}\}_{rsnt}} \quad & \sum_{r\in\mathcal{R}}\sum_{s\in\mathcal{R}} \tilde{f}^t_{rs}\tilde{m}^t_{rs} + \sum_{r\in\mathcal{R}}\sum_{s\in\mathcal{R}}\sum_{n\in\mathcal{N}(s)} \tilde{h}^t_{rsn}\tilde{c}^t_{rsn} \\
\text{s.t.} \quad & \tilde{f}^t_{rs} + \sum_{n\in\mathcal{N}(s)} \tilde{h}^t_{rsn} = q^t_{rs} && \forall r,s \\
& \tilde{f}^t_{rs} \geq 0 && \forall r,s \\
& \tilde{h}^t_{rsn} \geq 0 && \forall r,s,n \\
& \left(\{\tilde{m}^t_{rs}\}_{rst},\{\tilde{c}^t_{rsn}\}_{rsnt},\{\tilde{d}^t_{s}\}_{st}\right) = \Psi\left(\{\tilde{f}^t_{rs}\}_{rst},\{\tilde{h}^t_{rsn}\}_{rsnt}\right)
\end{aligned}
\tag{18}
$$

The objective function of the formulation is to minimize the total travel time (TTT) consisting of two branches of traffic flow $\tilde{f}^t_{rs}$ and $\tilde{h}^t_{rsn}$ in the time interval t, which are the decision variables. $\tilde{f}^t_{rs}$ represents the traffic flow that remains on the original routes and $\tilde{h}^t_{rsn}$ presents the traffic flow whose final destination is region s and the drop-off location is region n, $n \in \mathcal{N}(s)$. Ψ translates the two branches of traffic flow $(\tilde{f}^t_{rs},\tilde{h}^t_{rsn})$ into $(\{\tilde{m}^t_{rs}\}_{rst},\{\tilde{c}^t_{rsn}\}_{rsnt},\{\tilde{d}^t_{s}\}_{st})$, where $\tilde{m}^t_{rs}$ is the travel time of $\tilde{f}^t_{rs}$, $\tilde{c}^t_{rsn}$ is the travel time of $\tilde{h}^t_{rsn}$, and $\{\tilde{d}^t_{s}\}_{st}$ is the NoPUDO in region s and time interval t. To understand the objective function more accurately, we decompose it into three parts, as discussed in Proposition 5.

PROPOSITION 5 **(Total travel time decomposition)**. *The change of total travel time (TTT) after the re-routing using Formulation 18 can be decomposed into four parts, as presented in Equation 19.*

$$
\Delta TTT = \Delta_{Counterfactual} + \Delta_{PUDO,\ Remain} + \Delta_{PUDO,\ Detour} \tag{19}
$$

where ΔTTT denotes the change of TTT after the re-routing (after minus before), $\Delta_{Counterfactual}$ represents the change of the TTT after re-routing if the congestion effect of PUDOs is zero, Δ_{Remain} represents the change of the TTT after re-routing for the travelers staying on their original routes, and Δ_{Detour} represents the change of TTT after re-routing for the travelers taking the detours. To be specific, we have Equation 20 holds.

$$
\begin{aligned}
\Delta_{Counterfactual} &= \textstyle\sum_{r\in\mathcal{R}}\sum_{s\in\mathcal{R}}\left(\tilde{f}^t_{rs}m^t_{rs} + \sum_{n\in\mathcal{N}(s)}\tilde{h}^t_{rsn}\left(m^t_{rn}+m^t_{ns}\right) - q^t_{rs}m^t_{rs}\right) \\
\Delta_{PUDO,\ Remain} &= \textstyle\sum_{r\in\mathcal{R}}\sum_{s\in\mathcal{R}}\tilde{f}^t_{rs}\left(\tilde{m}^t_{rs}-m^t_{rs}\right) \\
\Delta_{PUDO,\ Detour} &= \textstyle\sum_{r\in\mathcal{R}}\sum_{s\in\mathcal{R}}\sum_{n\in\mathcal{N}(s)}\tilde{h}^t_{rsn}\left(\tilde{m}^t_{rn}-m^t_{rn}+w_{ns}-m^t_{ns}\right)
\end{aligned}
\tag{20}
$$

Proof. See Appendix B. □

We expect $\Delta TTT < 0$, which means that the TTT after the re-routing is smaller than the current situation without re-routing. In general, $\Delta_{\text{Counterfactual}} > 0$ because travelers prefer selecting the shortest paths. Additionally, $\Delta_{\text{PUDO, Detour}} > 0$ because the traffic flow increases on the detour routes, and walking usually takes longer time than driving. To make $\Delta TTT < 0$, we need to make $\Delta_{\text{PUDO, Remain}} < -|\Delta_{\text{Counterfactual}} + \Delta_{\text{PUDO, Detour}}| < 0$. That means, the re-routing should reduce the

travel time for travelers staying on their original routes. The reduced TTT for the travelers staying on the original routes should be larger than the increased TTT for the travelers taking the detours.

We further discuss Ψ, which can be formulated as a series of constraints, as shown in Equation 21.

$$d_s^t = \sum_{r\in\mathcal{R}} q_{rs}^t \tag{21a}$$

$$\tilde{d}_s^t = \sum_{r\in\mathcal{R}} \tilde{f}_{rs}^t + \sum_{r\in\mathcal{R}} \sum_{n\in\mathcal{N}(s)} \tilde{h}_{rns}^t \tag{21b}$$

$$\beta d_s^t \le \tilde{d}_s^t \le \gamma d_s^t \tag{21c}$$

$$\Delta_s^t = \tilde{d}_s^t - d_s^t \tag{21d}$$

$$\tilde{y}_s^t = y_s^t + \hat{\theta}_s \Delta_s^t \tag{21e}$$

$$\tilde{m}_{rs}^t = \sum_{v\in\mathcal{L}_{rs}} l_v / \tilde{y}_v^t \tag{21f}$$

$$u_{ns} = \sum_{v\in\mathcal{L}_{rs}} l_v / k \tag{21g}$$

$$\tilde{c}_{rsn}^t = \tilde{m}_{rn}^t + u_{ns} \tag{21h}$$

where $\mathcal{L}_{rs}$ is the set of regions listed in the shortest path from origin r to destination s, indexed by v. Ψ consists of two parts: 1) limiting the change of NoPUDO; 2) calculating the travel time after re-routing, as discussed below:

- Before re-routing, the NoPUDO d_s^t equals to the summation of all traffic flow whose destination is in region s, as shown in Equation 21a. After re-routing, the updated NoPUDO $\tilde{d}_s^t$ sums two branches of traffic flow, *i.e.*, $\tilde{f}_{rs}^t$ and $\tilde{h}_{rns}^t$, whose PUDOs location is region s, as shown in Equation 21b. The updated NoPUDO $\tilde{d}_s^t$ is within $[\beta d_s^t, \gamma d_s^t]$, as presented by Equation 21c, where $\beta \le 1, \gamma \ge 1$ are hyper-parameters to limit the change of NoPUDO.
- The change of NoPUDO is calculated as the difference of d_s^t and $\tilde{d}_s^t$, as shown in Equation 21d. Then the traffic speed in the region v after the re-routing can be updated using θ_v, as shown in Equation 21e. Based on the updated speed, the travel time from region r to region s after the re-routing can be calculated in Equation 21f. l_v is the average trip distance in the region v. For the re-routed flow, we first calculate the walking time u_{ns} from region n to region s in Equation 21g, where k is the average walking speed. Lastly, the travel time for the re-routed flow $\tilde{h}_{rsn}^t$ is calculated as the summation of travel time from r to n and from n to s, as shown in Equation 21h.

Overall, Formulation 18 belongs to non-linear programming as the objective function contains the product of $\tilde{f}_{rs}^t$ and $\tilde{m}_{rs}^t$, as well as the product of $\tilde{h}_{rsn}^t$ and $\tilde{c}_{rsa}^t$. The travel time $\tilde{m}_{rs}^t$ is also proportional to the reciprocal of $\tilde{y}_v^t$, as shown in Equation 21f. Given a large-scale network, the number of decision variables $\{\tilde{f}_{rs}^t\}_{rst}, \{\tilde{h}_{rsn}^t\}_{rst}$ can be large, making it difficult to solve by applying

standard non-linear programming solvers. In the following sections, we will present a customized solution algorithm to solve Formulation 18 effectively.

3. Solution algorithms

This section presents two solution algorithms. First, we design and implement the solution algorithm to the DSML method according to its theoretical structures. Then we develop a new algorithm to solve the re-routing formulation, which splits the solving process into two sub-processes and solves both sub-processes iteratively.

3.1. Solving the DSML method

To align with the proof of the DSML method, $\hat{\varphi}_v$ and $\hat{\psi}_v$ should be independently trained, which is similarly required in the standard DML (Chernozhukov et al. 2018). To this end, we always divide a dataset into two disjoint parts: one for training model `Y` and the other for training model `D`. At the same time, we make use of the b-fold cross-validation to select the optimal ML models and hyper-parameters in DSML. The detailed algorithm for DSML is presented in Algorithm 1.

Algorithm 1: Solution algorithm to the DSML method.

Input : $\{y_v^t\}_{vt}, \{d_v^t\}_{vt}, \{\mathbf{W}_v^t\}_{vt}$, candidates of ML models, ranges of hyper-parameters
Output: $\{\hat{\theta}_v\}_v$

for $v \in \mathcal{V}$ **do**
  Construct $\mathbf{Y}_v^{t-I:t-1}$ with y_v^{t-I}, y_v^{t-I+1}, $\cdots$, y_v^{t-1}, $\forall t$.
  Construct $\mathbf{D}_v^{t-I:t-1}$ with d_v^{t-I}, d_v^{t-I+1}, $\cdots$, d_v^{t-1}, $\forall t$.
  Construct $\mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$ by averaging the speed of the neighboring regions $\mathcal{N}(v)$, $\forall t$.
  Combine y_v^t, $\mathbf{Y}_v^{t-I:t-1}$, $\mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$, d_v^t, $\mathbf{D}_v^{t-I:t-1}$, $\mathbf{W}_v^t$ for all t to construct the entire dataset $\mathcal{D}$.
  Split the constructed dataset into b sub-datasets randomly. We denote $\mathcal{D}_i$ as the i-th sub-dataset, and $\mathcal{D}_{-i} = \mathcal{D}/\mathcal{D}_i$, where $\mathcal{D}$ is the entire dataset, $i = 0, \cdots, b-1$.
  for $i = 0; i < b; i++$ **do**
    Train Model `Y` by applying each of the candidate ML models with different hyper-parameter settings on the first half of $\mathcal{D}_{-i}$.
    Train Model `D` by applying each of the candidate ML models with different hyper-parameter settings on the second half of $\mathcal{D}_{-i}$.
    Select the optimal candidate ML model and hyper-parameter setting for Model `Y` and Model `D` respectively based on the performance on $\mathcal{D}_i$.
    Obtain the predicted values of $\hat{y}_v^t$ and $\hat{d}_v^t$ by running the trained Model `Y` and Model `D` on $\mathcal{D}_i$.
    Calculate the residual $\hat{\epsilon}_v^t$ of the Model `Y` on $\mathcal{D}_i$.
    Calculate the residual $\hat{\xi}_v^t$ of the Model `D` on $\mathcal{D}_i$.
  end
  Merge residuals $\hat{\epsilon}_v^t$ and $\hat{\xi}_v^t$ from each sub-dataset $\mathcal{D}_i, \forall i$.
  Estimate $\hat{\theta}_v$ by OLS between $\hat{\epsilon}_v^t$ and $\hat{\xi}_v^t$.
end
Return $\{\hat{\theta}_v\}_v$.

In this paper, the candidate ML models include Gradient Boosting Regression, Random Forest Regression, and Ada Boosting Regression. The ranges of hyper-parameters are set based on the recommendation of `scikit-learn`.

3.2. Solving the re-routing formulation

As discussed above, Formulation 18 is a non-linear program with high-dimensional decision variables on large-scale networks. To solve the formulation, we view $\tilde{m}^t_{rs}$ as an intermediate variable. With $\tilde{m}^t_{rs}$ known and fixed, Formulation 18 reduces to a linear program, which is easy to solve. Additionally, $\tilde{m}^t_{rs}$ can be updated using the decision variables $(\tilde{f}^t_{rs}, \tilde{h}^t_{rsn})$ with closed-form equations. Based on the above observations, we develop a solution algorithm to conduct the following two steps iteratively until convergence: 1) fix $\tilde{m}^t_{rs}$, solve the simplified Formulation 18 as a linear program to obtain $(\tilde{f}^t_{rs}, \tilde{h}^t_{rsn})$; 2) use the solved $(\tilde{f}^t_{rs}, \tilde{h}^t_{rsn})$ to update $\tilde{m}^t_{rs}$ based on Equation 21. Details of the algorithm are presented in Algorithm 2.

Algorithm 2: Solution algorithm to the re-routing formulation.

Input : $\{m^t_{rs}\}_{rst}$, $\{c^t_{rsn}\}_{rsnt}$, $\{q^t_{rs}\}_{rst}$, $\{d^t_s\}_{st}$, $\{y^t_s\}_{st}$, $\{\hat{\theta}_s\}_s$, $\{\mathcal{L}_{rs}\}_s$.
Output: $\{\tilde{f}^t_{rs}\}_{rst}$ and $\{\tilde{h}^t_{rsn}\}_{rsnt}$.

for $t \in \mathbb{T}$ **do**
  Initialize $\{\tilde{f}^t_{rs}\}_{rs}$ and $\{\tilde{h}^t_{rsn}\}_{rsn}$ such that the constraints of Formulation 18 are satisfied.
  while *changes of* $\tilde{f}^t_{rs}$ *and* $\tilde{h}^t_{rsn}$ *are within tolerances* **do**
    Update $\{\tilde{m}^t_{rs}\}_{rs}$ and $\{\tilde{c}^t_{rsn}\}_{rsn}$ based on Equation 21.
    Calculate and record the objective function $\sum_r \sum_s \tilde{f}^t_{rs} \tilde{m}^t_{rs} + \sum_r \sum_s \sum_n \tilde{h}^t_{rsn} \tilde{c}^t_{rsn}$.
    Solve Formulation 18 as a linear program problem by fixing $\{\tilde{m}^t_{rs}\}_{rs}$ and $\{\tilde{c}^t_{rsn}\}_{rsn}$.
    Obtain the solution results $\{\check{f}^t_{rs}\}_{rs}$ and $\{\check{h}^t_{rsn}\}_{rsn}$ after solving the above linear program.
    Update $\{\tilde{f}^t_{rs}\}_{rs}$ and $\{\tilde{h}^t_{rsn}\}_{rsn}$ by gradient descent with momentum, and the gradients are calculated as $\tilde{f}^t_{rs} - \check{f}^t_{rs}, \forall rs$ and $\tilde{h}^t_{rsn} - \check{h}^t_{rsn}, \forall rsn$, respectively.
  end
end
Return $\{\tilde{f}^t_{rs}\}_{rst}$ and $\{\tilde{h}^t_{rsn}\}_{rsnt}$.

We set the parameter for momentum to be 0.8, and the tolerance is set to be $1e-3$ in terms of ℓ_2-norm. To improve the running efficiency of Algorithm 2, all the variables involved will be vectorized. For example, to efficiently evaluate Equation 21, we use an incidence matrix to replace the summations in Equation 21b, 21f, and 21g. Matrix multiplications are much faster than loop-based summations on multi-core CPUs and GPUs, and hence the solution efficiency can be enhanced. For details of the vectorization procedures, readers are referred to Ma and Qian (2018), Ma, Pi, and Qian (2020).

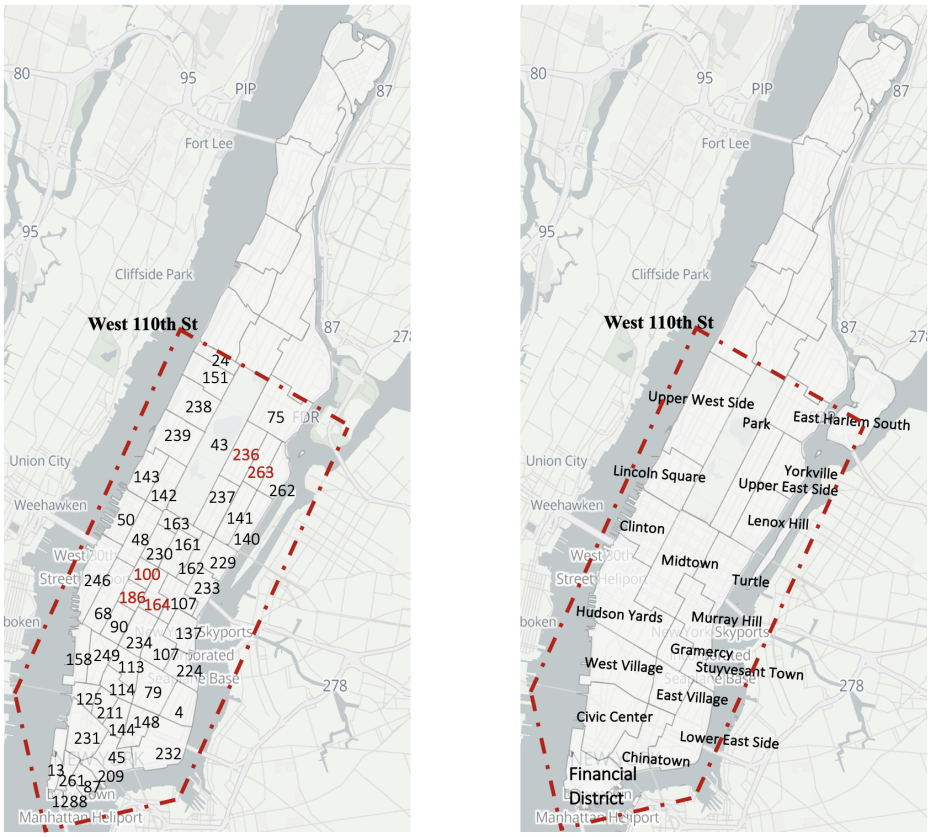
4. Numerical Experiments

In this section, we examine the effectiveness of the DSML method and re-routing formulation in the Manhattan area. We will first present the estimation results obtained by the DSML method, followed by the optimization results in the re-routing formulation.

4.1. Estimating the congestion effect of PUDOs

In this section, numerical experiments regarding the DSML method are presented. We first describe the datasets used in the study, which contain the NoPUDO, traffic speed, and precipitation. Then the estimation results in the Manhattan area are presented and discussed. Additionally, the effectiveness of the DSML method is compared with traditional methods, such as DML and Linear Regression (LR).

4.1.1. Data description We fence 52 regions below West 110th Street in the Manhattan area to be our study area, as shown in Figure 7. Because travel demands are mainly concentrated in these fenced regions, estimating the congestion effect of PUDOs in these regions is more meaningful.



(a) Region ID in the Manhattan area (b) Neighborhood name in the Manhattan area

Figure 7: Map of the study area.

We focus on the congested traffic conditions during afternoon peak hours from 16:00 to 20:00. Datasets used in this study include traffic speed, trip records, and precipitation from Feb 2019 to Jun 2020, as shown in Table 1. In this study, we only consider the PUDOs generated by passengers of taxi cabs and ride-hailing vehicles, due to the data limitation. The developed framework can be extended trivially to incorporate the PUDOs from public transit and private cars if data allows.

Table 1: Data Description

Datasets	Time Range	Resolution	Quantity	Descriptions
NYC traffic speed	Feb 2019 - Jun 2020	every 5 mins	404,351,029	Road segment, traffic speed, reference speed, time stamp
NYC trip records	Feb 2019 - Jun 2020	every 5 mins	18,157,071	Pick-up region ID, drop-off region ID, time stamp
NYC precipitation	Feb 2019 - Jun 2020	every 1 hour	11,987	precipitation, time stamp

The detailed descriptions and data processing procedures for each dataset are as follows:

• NYC speed data: The speed data contains several key fields including road segment ID, traffic speed, reference speed, and timestamp. The road speed is obtained based on probe vehicles, and the reference speed is the maximum speed limit on the road segment. To normalize the data for the DSML method, we calculate the relative speed, as shown in Equation 22.

$$y_v^t = \frac{\text{road speed} \times \text{traffic flow}}{\text{reference speed}} \tag{22}$$

Note that the relative speed is only used for training the DSML method. When calculating the TTT, we will transform the relative speed back to the actual traffic speed.

• NYC trip records: Trip order information from New York City Taxi & Limousine Commission (NYC-TLC) covers timestamps, vehicle types, pick-up locations, and drop-off locations. These orders come from four types of vehicles: yellow taxis, green taxis, For-Hire Vehicles (FHV), and High Volume For-Hire Vehicles (HVFHV). The NoPUDO in the region v every 5 minutes can be extracted to construct the value of d_v^t.

• NYC precipitation: Iowa Environmental Mesonet monitors the precipitation information in the Manhattan area every hour, and we use the volume of rainfall as the indicator of weather, denoted as $\mathbf{W}_v^t$.

4.1.2. Estimation results by the DSML method We apply the DSML method in the Manhattan area. An illustration of the variable relation in the DSML method is shown in Figure 8. The upper table records the observed traffic data, and the lower table denotes the predicted values and residuals obtained from Model `Y` and Model `D`. In our experiments, we set $I = 10$, *i.e.*, the historical data ranges from time $t-11$ to $t-1$. The residuals ϵ_v^t, ξ_v^t obtained from the differences of prediction and true values will be viewed as the dependent and independent variables into a linear regression in Model `Z`, and the congestion effect $\hat{\theta}_v$ can be estimated based on Algorithm 1.

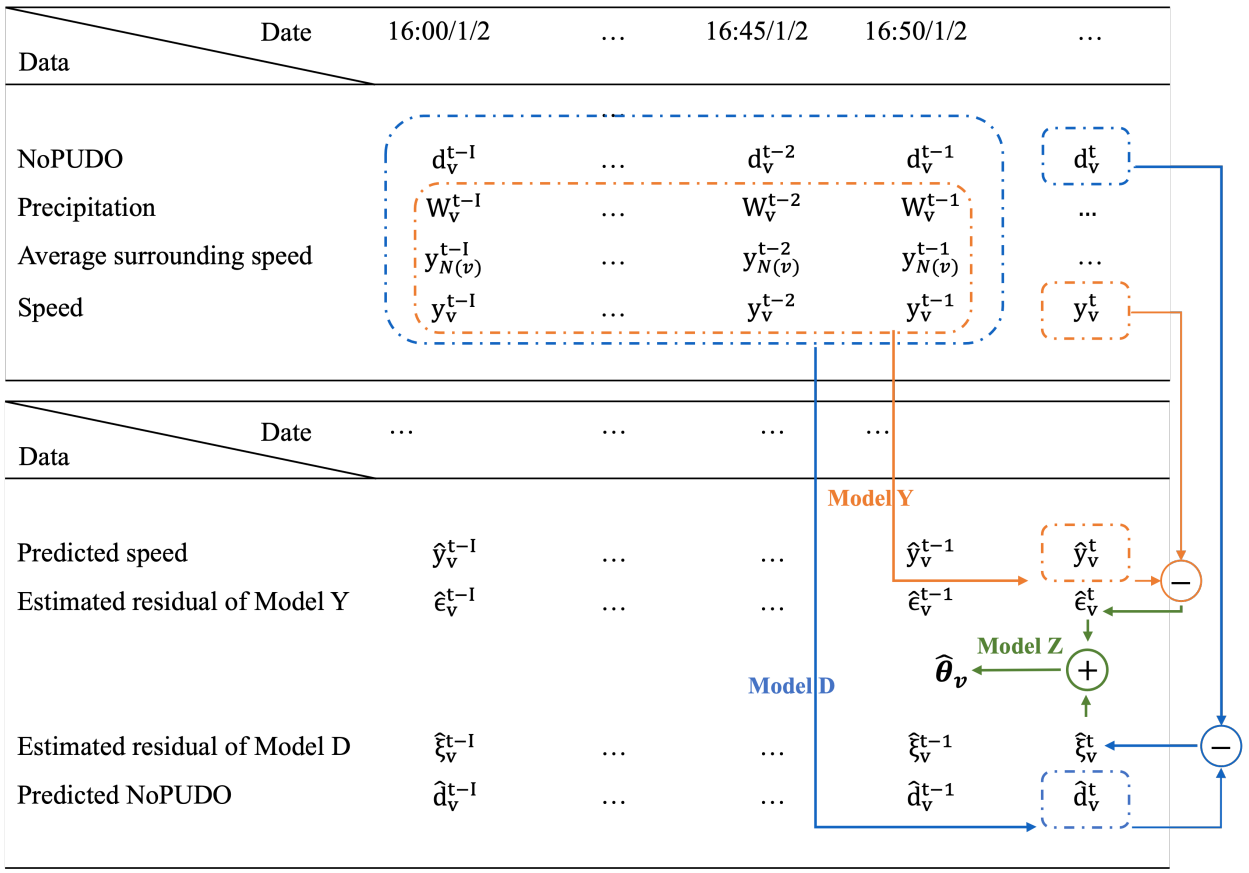


Figure 8: Variable relation of DSML for NYC data.

After running Algorithm 1, we found that the average θ_v is -0.0370 on weekdays and -0.0454 on weekends for all v in the Manhattan area. It means that when there are additional 100 unit PUDOs happening in a single region, the average traffic speed in that region will decrease by 3.70 miles/hour (mph) on weekdays and 4.54 mph on weekends.

We visualize the spatial distribution of the estimated $\hat{\theta}_v$ on weekdays and weekends in Figure 9, respectively. In Figure 9 (a) and (b), deeper color indicates higher values of $|\theta_v|$ and more severe congestion effects of PUDOs. The overall distribution of $\hat{\theta}_v$ is consistent with our common sense for the Manhattan area, as deeper color generally concentrates on busy regions in the Downtown and Midtown areas. In Figure 9 (c), we use purple points to mark the locations of some important points of interest (POIs), including the Empire State Building, Brooklyn Bridge, Time Square, Central Park, and so on. One can see that the distribution of $\hat{\theta}_v$ aligns well with those POIs, as shopping malls and office buildings usually generate more trips.

The distributions of $\hat{\theta}_v$ on weekdays and weekends also vary significantly, as shown in Figure 9(a) and (b). The congestion effect of PUDOs is more severe around POIs (*e.g.*, Times Square, Chinatown, and Brooklyn Bridge) on weekends than on weekdays, which is probably attributed to

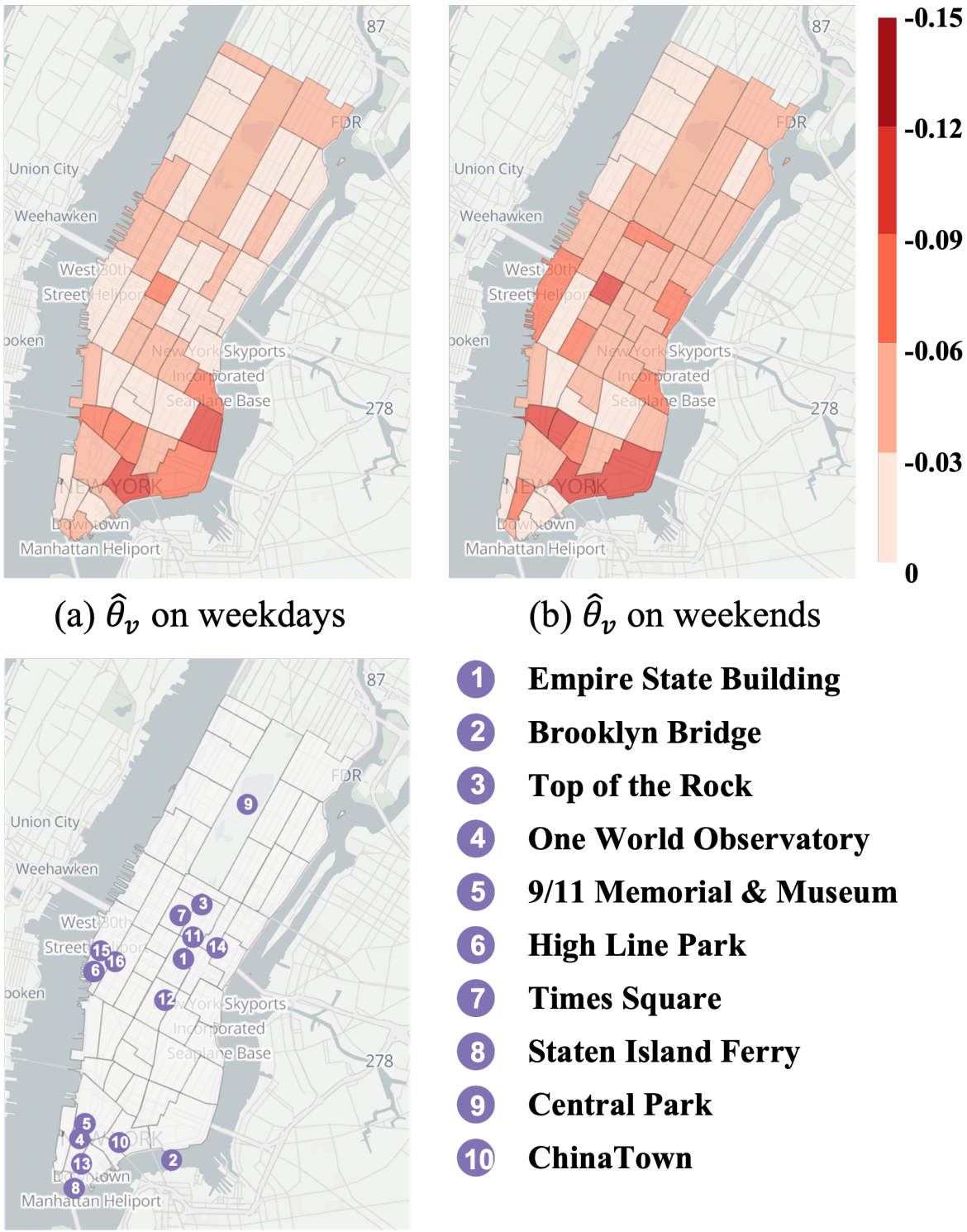


Figure 9: Overview of $\hat{\theta}$ learned from the DSML method and attractions in the Manhattan area.

the frequent activities around sightseeing attractions during weekends. We further present the histogram of the estimated $\hat{\theta}_v$ for weekdays and weekends in Figure 10. One can see that $\hat{\theta}_v$ on weekends is more probable to be below -0.10, and the mode of $\hat{\theta}_v$ on weekends is smaller than that on weekdays.

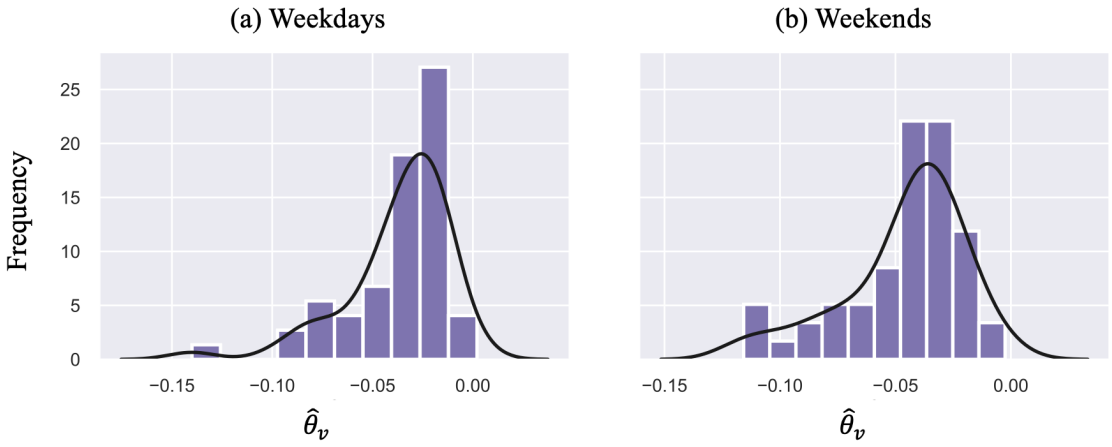


Figure 10: Histogram of $\hat{\theta}_v$ on weekdays and weekends.

4.1.3. Analyzing residuals of the DSML method We select two areas to validate the estimation results from the DSML method in detail. The first area is located in Midtown, which consists of the Region 100 and 186; the second area is around Central Park, which consists of the Region 236 and 263. The four regions will also be studied in the re-routing formulation.

The fitted linear lines in Model `Z` for the four regions are separately shown in Figure 11.

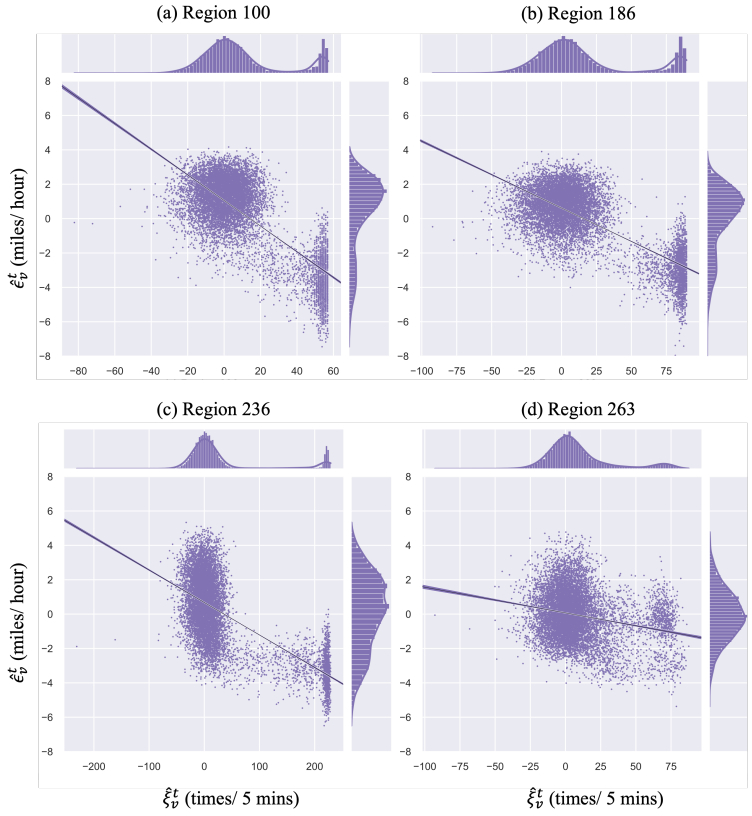


Figure 11: The fitted linear line in Model `Z`.

The residuals of both Model `Y` and Model `D` are centered at the origin, which indicates that both models achieve a good fitting. More importantly, the two residuals $\hat{\epsilon}_v^t$ and $\hat{\xi}_v^t$ are negatively correlated, and this suggests a negative value of θ_v. Indeed, the slope of the fitted line is $\hat{\theta}_v$, and the t-test can be conducted to evaluate the significance of the estimated $\hat{\theta}_v$.

We list the estimated $\hat{\theta}_v$ and the corresponding p-value for each region in Table 7. One can see that all the p-values of the DSML method are below 0.001, which indicates the estimated $\hat{\theta}_v$ is highly significant. Besides, the value of $\hat{\theta}_v$ is negative, which shows the NoPUDO has a negative effect on the traffic speed. Furthermore, the $\hat{\theta}_v$ is varied with different regions depending on unique attributes and properties in each region.

4.1.4. Sensitivity analysis regarding the choice of ML models We examine the robustness of different ML models used in Model `Y` and Model `D`. In Algorithm 1, the optimal ML model is selected from Gradient Boosting, Random Forest, and Ada Boosting Regression using cross-validation. In this section, we specify the ML model used in Model `Y` and Model `D` and evaluate how the estimation results are different from the original ones. In general, we believe a smaller difference indicates a more robust DSML method in terms of the choice of ML models.

Table 2: Sensitivity analysis of ML models used in the DSML method.

ML models	GB	RF	Ada
Correlation coefficient	0.99	0.94	0.83

Table 3: Correlation analysis for DSML vs. DML and DSML vs. LR.

Models	DML	LR
Correlation coefficient	-0.14	0.27

Table 4: Different features and outcome variables of DSML, DML, and LR.

Models	Features	Outcome Variable	Methods
DSML	$\mathbf{Y}_v^{t-I:t-1}$, $\mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$, $\mathbf{W}_v^t$	y_v^t	ML models
	$\mathbf{D}_v^{t-I:t-1}$, $\mathbf{Y}_v^{t-I:t-1}$, $\mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$, $\mathbf{W}_v^t$	d_v^t	ML models
	$\hat{\xi}_v^t$	$\hat{\epsilon}_v^t$	linear regression
DML	$\mathbf{D}_v^{t-I:t-1}$, $\mathbf{Y}_v^{t-I:t-1}$, $\mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$, $\mathbf{W}_v^t$	y_v^t	ML models
	$\mathbf{D}_v^{t-I:t-1}$, $\mathbf{Y}_v^{t-I:t-1}$, $\mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$, $\mathbf{W}_v^t$	d_v^t	ML models
	ξ_v^t	ϵ_v^t	linear regression
LR	d_v^t	y_v^t	linear regression

To this end, we run the DSML method by fixing Model `Y` and Model `D` to be either Gradient Boosting Regression, or Random Forest Regression, or Ada Boosting Regression. Then we compare the difference between the newly estimated and the original $\hat{\theta}_v$ through Pearson correlation coefficients, and the results are presented in Table 2. One can see that the correlation coefficients for Gradient Boosting, Random Forest, and Ada Boosting Regression are 0.99, 0.94, and 0.83, respectively. All the correlation coefficients are high, indicating that the proposed DSML method is robust to the choice of the ML models for Model `Y` and Model `D`.

4.1.5. Comparison among DSML, DML, and LR We compare the developed DSML method with the standard DML and LR methods. The comparison among the features and outcome variables of DSML, DML, and LR is shown in Table 4.

The estimated $\hat{\theta}_v$ by DML and LR on weekdays are shown in Figure 12. Results on weekends are similar and hence omitted. On average, $\hat{\theta}_v$ is -0.008 by DML and -0.055 by LR. The estimated $\hat{\theta}_v$ by DML is generally smaller than that by DSML, and $\hat{\theta}_v$ for all regions are almost identically small. One can see that DML cannot capture the congestion effect accurately, which is mainly because DML additionally considers the non-existing relationship from $\mathbf{D}_v^{t-I:t-1}$ to y_v^t based on the causal graph in Figure 3. In contrast, LR overlooks the complex spatio-temporal relationship between y_v^t and d_v^t, and the estimated $\hat{\theta}_v$ is smaller (the absolute value is larger) than that estimated from DSML, which is consistent with Example 1. Importantly, the estimated $\hat{\theta}_v$ is inconsistent with those commonly known busy regions. For example, on the upper west side, there are several regions in deep red near West 110th Street in Figure 12(b), while these regions usually generate a few travel demands and are not congested. We further compare the estimated $\hat{\theta}_v$ by DML and LR with

that estimated by DSML using the correlation coefficient, and the results are shown in Table 4. The low correlation between DML/LR and DSML indicates that the estimated $\hat{\theta}_v$ by DSML is completely different from that estimated by DML or LR. Since the distribution of $\hat{\theta}_v$ by DSML is more reasonable, we have a stronger belief that DSML can estimate the true congestion effects by PUDOs.

Additionally, we conduct the t-test for the estimated $\hat{\theta}_v$ from DML and LR as well, and the results are shown in Table 7. One can see that some estimated $\hat{\theta}_v$ are not significant, which might be due to the influence of the confounding factors $\mathbf{D}_v^{t-I:t-1}$ in DML. Though the significance levels for LR are high, the estimated $\hat{\theta}_v$ reflects not only causality but also correlation, based on our discussions in Example 1.

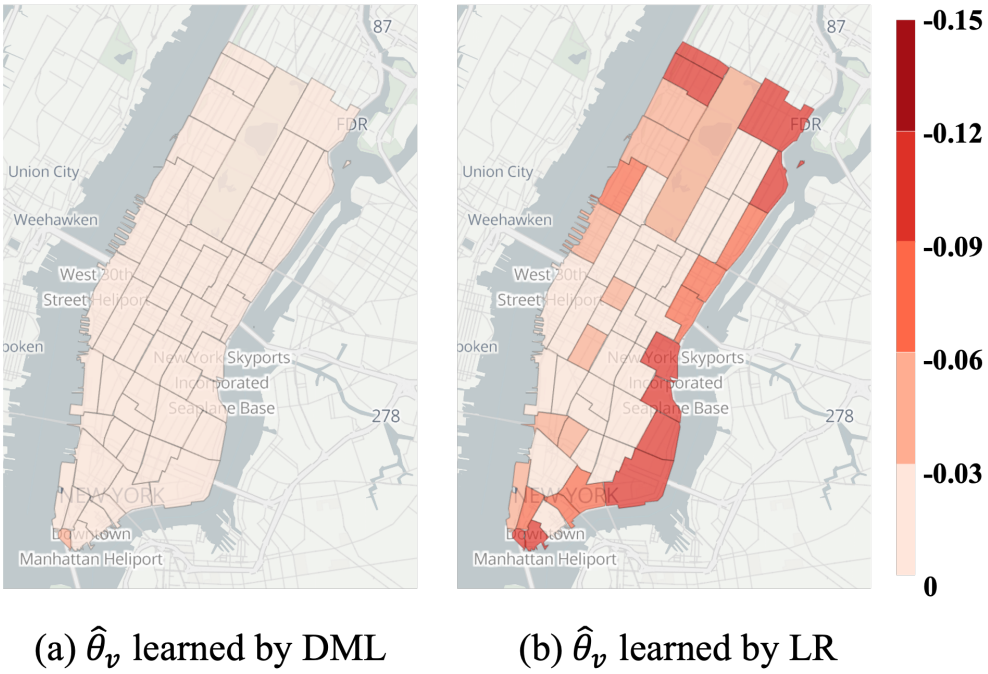


Figure 12: Comparison of estimated $\hat{\theta}_v$ by DML and LR on weekdays.

4.2. Solving the re-routing formulation

In this section, we solve the re-routing formulation for some busy regions in the Manhattan area.

4.2.1. Settings Four regions near Midtown and Central Park are selected as study areas, which are also used in section 4.1.3. We consider all the trips to these four regions. We consider the study period from 1st Jul 2019 to 30th Sep 2019, and time intervals 16:00–17:00, 17:00–18:00, 18:00–19:00, and 19:00–20:00 during the afternoon peaks are considered separately. The total number of vehicles on the roads is set to be λ times of the trip orders in the NYC datasets. We set $\beta = 0$, l_v

is calculated as the average travel distance with each region, k is set as 3.5 miles/hour, and $\hat{\theta}_v$ are estimated by DSML in the previous section.

We examine the improvement rate before and after re-routing based on Equation 23.

$$\text{improvement rate} = \frac{\text{TTT before re-routing} - \text{TTT after re-routing}}{\text{TTT before re-routing}} \times 100\% \tag{23}$$

To evaluate the TTT after re-routing, we follow the steps in Ma and Qian (2020). We assume the hypothetical traffic conditions (in terms of travel time) after re-routing are calculated based on the changes of NoPUDO in each region, as presented in Equation 21f. Only weekdays are considered as the results on weekends are similar.

4.2.2. TTT after re-routing We run Algorithm 2 with $\lambda = 15$, and statistics for TTT are shown in Table 5.

Table 5: TTT and improvement rates after re-routing on weekdays. (Mean±Std, $\lambda = 15$)

	Before re-routing ($\times 10^3$ hours)	After re-routing ($\times 10^3$ hours)	Improvement rate (%)
Midtown ($\gamma = 2.3$)			
Average	4.41 ± 0.71	4.30 ± 0.65	2.44 ± 1.55
16:00–17:00	4.60 ± 0.63	4.45 ± 0.58	3.01 ± 2.06
17:00–18:00	4.74 ± 0.75	4.60 ± 0.69	2.86 ± 2.42
18:00–19:00	4.50 ± 0.80	4.38 ± 0.75	2.47 ± 2.22
19:00–20:00	3.81 ± 0.84	3.76 ± 0.81	1.20 ± 1.15
Central Park ($\gamma = 1.6$)			
Average	3.63 ± 0.75	3.54 ± 0.70	2.12 ± 1.61
16:00–17:00	2.94 ± 0.56	2.84 ± 0.51	2.98 ± 1.85
17:00–18:00	3.75 ± 0.78	3.64 ± 0.72	2.58 ± 1.75
18:00–19:00	4.23 ± 0.90	4.13 ± 0.84	2.25 ± 2.01
19:00–20:00	3.57 ± 0.82	3.54 ± 0.80	0.76 ± 1.09

We note that the mean and standard deviation in Table 5 are calculated based on the TTT of each day. One can see that the average improvement rate is 2.44% in Midtown and 2.12% in Central Park on weekdays. The improvements are more significant during 16:00–19:00 for both areas. The standard deviation is roughly half the mean, indicating the high randomness of network conditions. Overall, re-routing traffic flow with PUDO has great potential in reducing the total travel time for both areas and across all time periods.

4.2.3. Sensitivity analysis To evaluate the sensitivity with respect to demands level, we first perturb λ to be 5, 10, 20, 25, 30 and evaluate the improvement rate. Note that λ indicates the level

of total traffic demands, and higher λ represents more traffic demand. The mean and standard deviation of the improvement rates on different λ for the Midtown and Central Park are shown in Figure 13 and Figure 14, respectively.

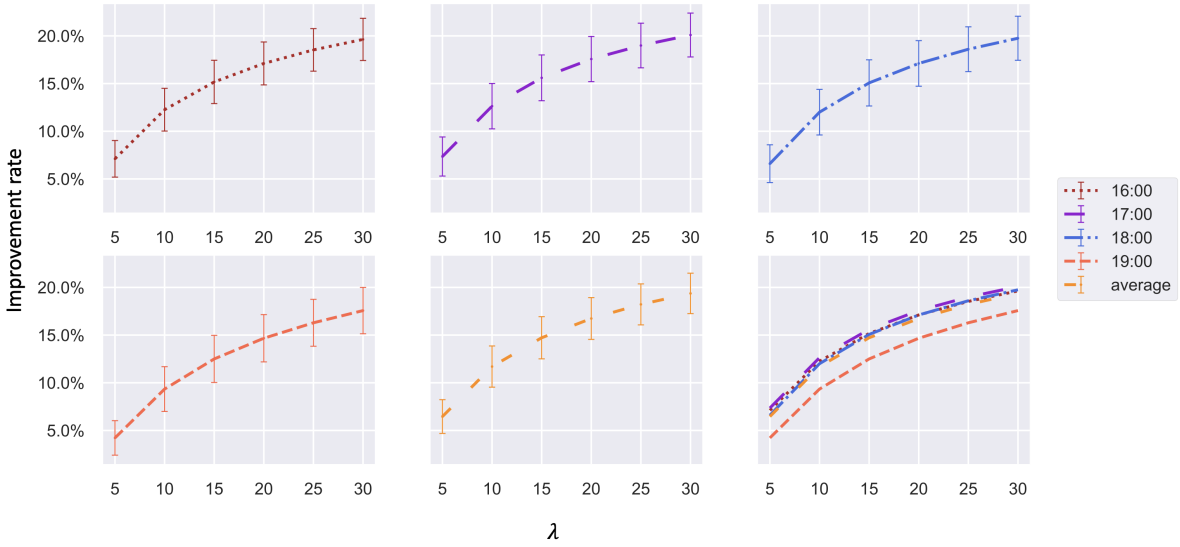


Figure 13: Improvement rates on different λ in Midtown (error bar represents the standard deviation).

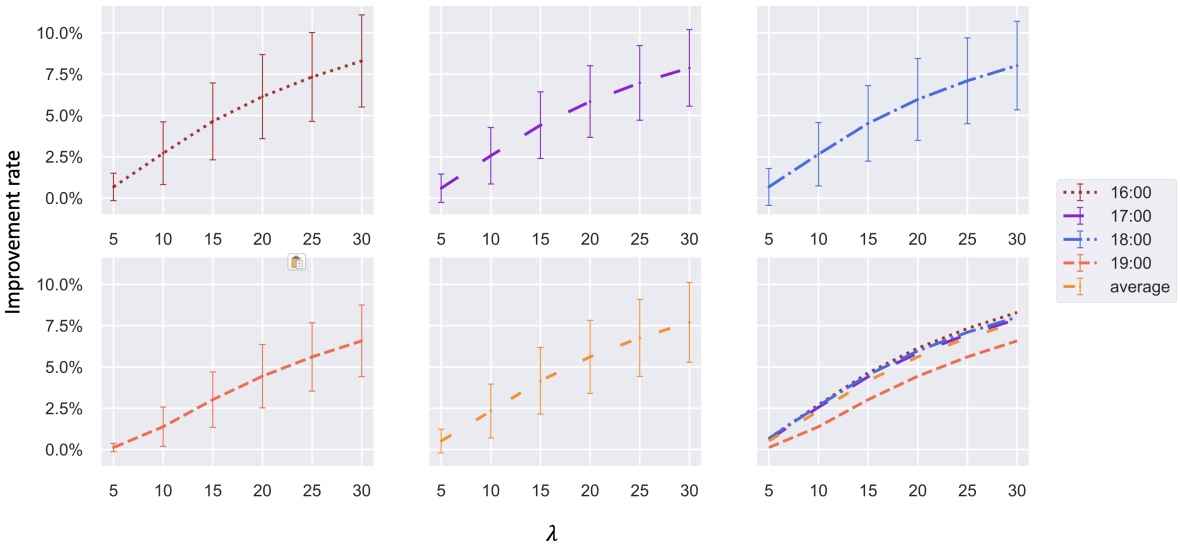


Figure 14: Improvement rates on different λ in Central Park (error bar represents the standard deviation).

In general, higher traffic demands encourage a larger improvement rate for both areas. Rerouting traffic flow with PUDOs turns out to be a promising and robust tool for system optimal

under different demands levels. Additionally, an interesting finding is that the standard deviation of the improvement rate is also increasing. This suggests that when the demand increases, network conditions become more random, and the TTT improvement becomes more stochastic.

Secondly, we vary γ from 2.1 to 2.5 for Midtown, and from 1.4 to 1.8 for Central Park, to examine the sensitivity regarding the limitation of NoPUDO changes. The resulted improvement rate curves are shown in Figure 15 and Figure 16.

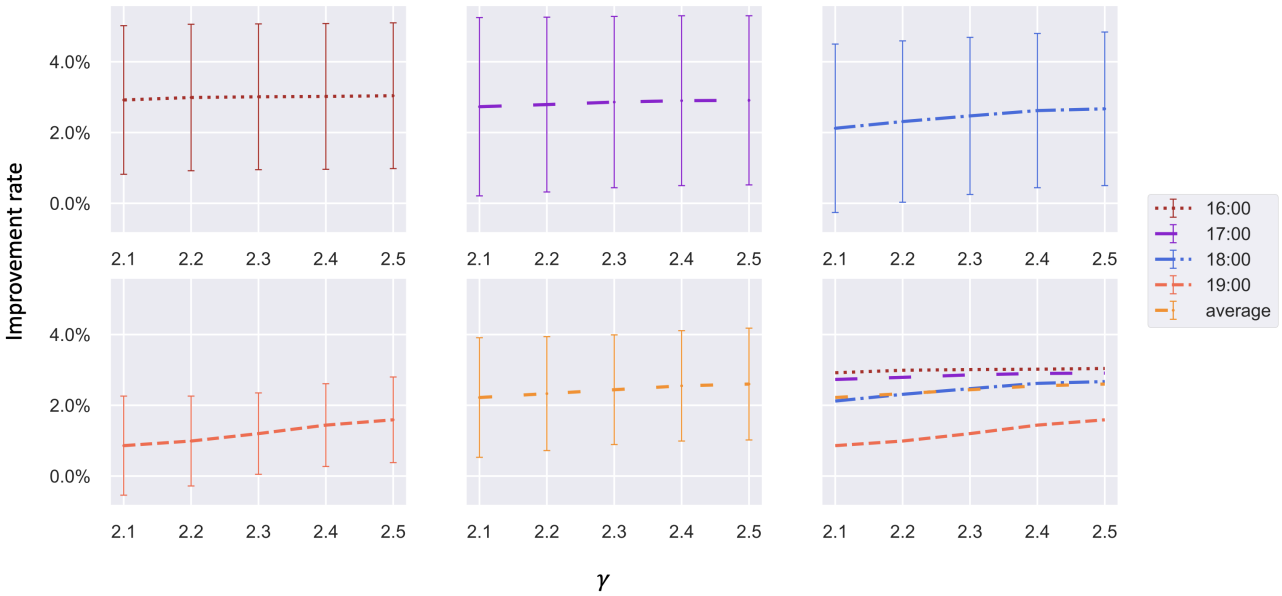


Figure 15: Improvement rates on different γ in Midtown (error bar represents the standard deviation).

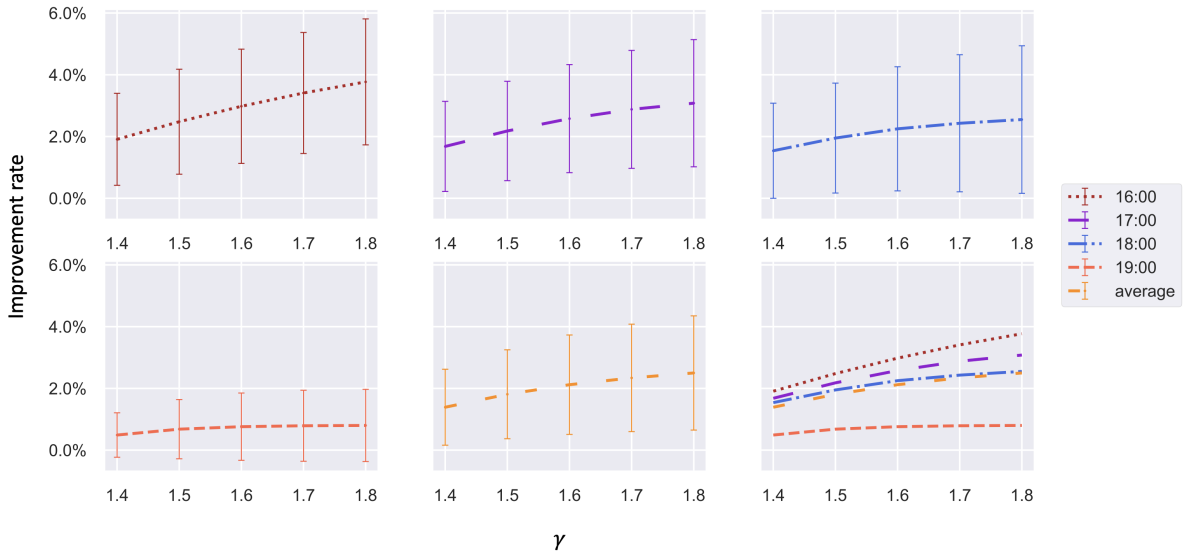


Figure 16: Improvement rates on different γ in Central Park (error bar represents the standard deviation).

The improvement rate increases when γ increases, and the reason is straightforward: increasing γ will relax the limitation on the changes of NoPUDO in each region, and hence the search space for the re-routing formulation becomes larger. Another noteworthy point is that the standard deviation of the improvement rates remains the same when γ changes in Midtown, while the standard deviation increases with respect to γ in Central Park. This might be because of the unique characteristics and demand levels in each region.

5. Conclusion

This paper first time makes use of the causal inference to estimate the congestion effect of PUDOs with observational traffic data, and the estimated congestion effect can be further used to mitigate the congestion induced by PUDOs. To this end, the causal relationship between NoPUDO and traffic speed is identified through a causal graph, and the novel DSML method is developed to estimate the congestion effect of PUDOs based on the causal graph. Theoretical guarantees regarding the estimation results of DSML are also provided. To reduce the network-wide travel time, a re-routing formulation is developed and the corresponding solution algorithm is proposed.

Experiments with real-world data in the Manhattan area demonstrate the effectiveness of the developed DSML method, and the estimation results align well with the actual traffic situations. On average, 100 additional units of the PUDO will decrease traffic speed by 3.70 mph on weekdays and 4.54 mph on weekends. The re-routing formulation also demonstrates great potential in reducing the total travel time. The improvement rate regarding the total travel time can reach 2.44% in Midtown and 2.12% in Central Park during weekdays.

As for the future research directions, it is worth considering different road attributes and properties when estimating the congestion effect. For example, PUDOs can cause more congestion on a one-way and one-lane road with narrow curb space, while the congestion effect on large curb space might be negligible. This paper estimates the congestion effect of PUDOs on regional levels. If there are more detailed data of PUDOs on the road levels, we can explore the congestion effects of PUDOs on each road segment separately, and the road-level congestion effects can be used for curb pricing (Liu, Ma, and Qian 2022) and the design of curb space. In addition, it would be interesting to identify the congestion effects of PUDOs from heterogeneous vehicle types, and the re-routing formulation can also be customized for different vehicle types. For example, PUDOs from ride-sharing vehicles may generate more congestion as the PUDO usually lasts a longer time, compared to a single-rider vehicle. Based on this principle, developing re-routing strategies for different types of vehicles could further reduce the total travel time. Additionally, as drop-offs usually take less time than pick-ups, we may consider modeling PU and DO separately when estimating the congestion effect and developing the re-routing formulations.

Supplementary Materials

The DSML method is implemented and the re-routing problem is solved in Python and open-sourced on GitHub (https://github.com/LexieLiu01/DSML).

Acknowledgments

The work described in this paper was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU/25209221) and a grant funded by the Hong Kong Polytechnic University (Project No. P0033933). The second author was supported by a National Science Foundation grant CMMI-1931827. The contents of this paper reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein.

Nomenclature

The list of notations used in this paper is shown in Table 6.

Table 6: List of notations.

Notations	Description
	Regions Related Variables
$\mathcal{V}$	Set of regions.
v, n	An index of a region in $\mathcal{V}$.
$\mathcal{R}$	Set of origin regions.
r	An index of an origin region in $\mathcal{R}$.
$\mathcal{S}$	Set of destination regions.
s	An index of a destination region in $\mathcal{S}$.
$\mathcal{N}(s)$	Set of neighboring regions for region s.

Observed Variables	
y_v^t	Traffic speed in the region v in the time interval t.
$\mathbf{Y}_v^{t-I:t-1}$	Vector of speed during the time intervals $t-I,\cdots,t-1$ in the region v. I is a constant that determines the length of historical data.
$\mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$	Vector of average speed of all regions $n \in \mathcal{N}(v)$ during the time intervals $t-I,\cdots,t-1$.
d_v^t	NoPUDO in the region v in the time interval t.
$\mathbf{D}_v^{t-I:t-1}$	Vector of the NoPUDO in region v during the historical time intervals $t-I...t-1$.
$\mathbf{W}_v^t$	External control variables in region v in the time interval t.
θ_v	Congestion effect of PUDOs in region v. One additional PUDO will make speed y_v^t increase by θ_v in region v.
Functions and Residuals of DSML	
φ_v	A function used to predict y_v^t without consideration of the congestion effect of PUDOs.
e_v^t	The residual of φ_v and $\theta_v d_v^t$ when predicting y_v^t.
ψ_v	A function used to predict d_v^t.
ξ_v^t	The residual of ψ_v when predicting d_v^t.
Estimated Variables	
$\hat{\varphi}_v$	Model `Y`.
$\hat{\psi}_v$	Model `D`.
$\hat{y}_v^t$	Prediction of the speed y_v^t in region v and time interval t.
$\hat{d}_v^t$	Prediction of the NoPUDO d_v^t in region v and time interval t.
$\hat{e}_v^t$	Estimation of the residual of the linear regression in Equation 13.
$\hat{\epsilon}_v^t$	Estimation of the residual of Model `Y`, which is obtained by subtracting the prediction $\hat{y_v^t}$ and the true value of y_v^t.
$\hat{\xi}_v^t$	Estimation of the residual of Model `D`, which is obtained by the subtracting the prediction $\hat{d}_v^t$ and the true value of d_v^t.
$\hat{\theta}_v$	Estimation of θ_v.
Network Flow Related Variables	
q_{rs}^t	Total traffic flow from region r to region s before re-routing in the time interval t.
$\tilde{f}_{rs}^t$	Traffic flow that stays on the original routes from origin region r to destination region s in the time interval t.
$\tilde{h}_{rsn}^t$	Traffic flow that departs from region r to one temporary destination n, $n \in \mathcal{N}(s)$ by vehicles, and from n to the final destination s by walking.
Iterated Variables in Re-routing	
$\tilde{d}_s^t$	The number of drop-off in region s after re-routing in the time interval t.
Δ_s^t	The change of the NoPUDO in region s before and after re-routing in the time interval t.

$\tilde{y}_s^t$	Updated traffic speed each re-routing.
m_{rs}^t	Travel time from region r to region s before re-routing in the time interval t.
$\tilde{m}_{rs}^t$	Travel time from region r to region s after re-routing in the time interval t.
$\tilde{c}_{rsn}^t$	Travel time for traffic flow depart from region r to region n by vehicles, and from region n to region s by walking after re-routing in the time interval t.

Constant Variables

k	Average walking speed.
u_{ns}	Walking time cost from region n to region s.
$\mathcal{L}_{rs}$	Set of regions in the shortest path from origin r to destination s, indexed by v.
l_v	The average travel distance in region v.

Appendix A: Property of $\hat{\theta}_v$

In this section, we first prove Proposition 3 for the case of linear models, then Proposition 4 is proved for the generalized cases.

A.1. Proof of Proposition 3

Based on the settings presented in Proposition 3, we prove $\hat{\theta}_v$ is an unbiased estimator of θ_v. To demonstrate the essential idea, we first use linear models for φ_v, as shown in Equation 24.

$$y_v^t = \theta_v d_v^t + \mathbf{A}^T \mathbf{Y}_v^{t-I:t-1} + \mathbf{B}^T \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1} + e_v^t \tag{24}$$

where we assume $\mathbf{A}, \mathbf{B}, \mathbf{Y}_v^{t-I:t-1}, \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$ are flattened vectors, and both $\mathbf{A}$ and $\mathbf{B}$ are parameters of φ_v.

Following the steps in DSML, we build additional regression models for y_v^t and d_v^t, as presented in Equation 25 and 26.

$$y_v^t = \mathbf{A_y}^T \mathbf{Y}_v^{t-I:t-1} + \mathbf{B_y}^T \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1} + \hat{\varepsilon}_v^t \tag{25}$$

$$d_v^t = \mathbf{A_d}^T \mathbf{Y}_v^{t-I:t-1} + \mathbf{B_d}^T \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1} + \mathbf{C_d}^T \mathbf{D}_v^{t-I:t-1} + \hat{\xi}_v^t \tag{26}$$

where $\mathbf{A_y}$, $\mathbf{B_y}$, $\mathbf{A_d}$, $\mathbf{B_d}$ and $\mathbf{C_d}$ are vectors of coefficients. $(\mathbf{A_y}, \mathbf{B_y})$ are the parameters for $\hat{\varphi}_v$, and $(\mathbf{A_d}, \mathbf{B_d}, \mathbf{C_d})$ are the parameters for $\hat{\phi}_v$.

We consider an alternative least-squares regression question:

$$\hat{\varepsilon}_v^t = \hat{\theta}_v \hat{\xi}_v^t + \hat{e}_v^t \tag{27}$$

To analyze the property of $\hat{\theta}_v$, we derive $\hat{e}_v^{ty}$ by substituting Equation 24 into Equation 25, as shown in Equation 28.

$$\hat{\varepsilon}_v^t = \theta_v d_v^t + (\mathbf{A} - \mathbf{A_y})^T \mathbf{Y}_v^{t-I:t-1} + (\mathbf{B} - \mathbf{B_y})^T \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1} + e_v^t \tag{28}$$

Then we plug the variable d_v^t in the Equation 26 into Equation 28. Eventually, we can formulate the $\hat{e}_v^t$ in the Equation 29.

$$\hat{\varepsilon}_v^t = \theta_v \hat{\xi}_v^t + (\theta_v \mathbf{A_d} + \mathbf{A} - \mathbf{A_y})^T \mathbf{Y}_v^{t-I:t-1} + (\theta_v \mathbf{B_d} + \mathbf{B} - \mathbf{B_y})^T \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1} + (\theta_v \mathbf{C_d})^T \mathbf{D}_v^{t-I:t-1} + e_v^t \tag{29}$$

As $\hat{e}_v^t$ is the residual from the linear regression in Equation 25, it is not correlated with $\mathbf{Y}_v^{t-I:t-1}$ or $\mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$ given both variables are the attributes of the linear regression. Additionally, $\hat{e}_v^t$ is not correlated with $\mathbf{D}_v^{t-I:t-1}$ due to the causal graph in Figure 3. Therefore, we have the coefficients $\theta_v \mathbf{A_d} + \mathbf{A} - \mathbf{A_y}$, $\theta_v \mathbf{B_d} + \mathbf{B} - \mathbf{B_y}$, and $\theta_v \mathbf{C_d}$ equal to zero in Equation 29. Consequently, we have Equation 30 holds.

$$\hat{\varepsilon}_v^t = \theta_v \hat{\xi}_v^t + e_v^t \tag{30}$$

By comparing Equation 27 and Equation 30, we have Equation 31 holds.

$$\begin{aligned} \hat{\theta}_v &= \theta_v \\ \hat{e}_v^t &= e_v^t \end{aligned} \tag{31}$$

The above proof is extended from the Frisch-Waugh-Lovell (FWL) theorem (Fiebig and Bartels 1996, Lovell 2008), and we show $\theta_v \mathbf{C_d} = 0$ based on the specific problem setting for the causal graph in this study.

A.2. Proof of Proposition 4

To prove Proposition 4, we rely on Theorem 3.1 in Chernozhukov et al. (2018). To this end, we verify that both Assumption 3.1 and 3.2 in Chernozhukov et al. (2018) hold. For region v, we set $\eta_v = (\varphi_v, \psi_v)$, and the inputs for both functions are omitted. Then the Neyman score function can be defined in Equation 32

$$\omega(\theta_v, \eta_v) = (y_v^t - \theta_v d_v^t - \varphi_v)(d_v^t - \psi_v) \tag{32}$$

We note that $\omega(\theta_v, \eta)$ is insensitive to the small change of either φ_v or θ_v, as presented in Equation 33.

$$\partial_{\eta_v} \mathbb{E}\omega(\theta_v, \eta_v)[\eta_v - \eta_v^0] = 0 \tag{33}$$

Then $\omega(\theta_v, \eta)$ is Neyman orthogonal, which satisfies Assumption 3.1. Additionally, Assumption 3.2 is satisfied because Equation 17 holds. Given that the data splitting technique presented in section 3.1 is adopted to train φ_v and ψ_v separately, then based on Theorem 3.1 in Chernozhukov et al. (2018), Proposition 4 is proved.

Appendix B: Proof of Proposition 5

The total travel time (TTT) before the re-routing can be calculated as $\sum_{r,s\in\mathcal{R}} q_{rs}^t m_{rs}^t$, and the TTT after re-routing is represented as the objective function in Formulation 18. Therefore, the change of TTT (ΔTTT) can be written in Equation 34.

$$
\begin{aligned}
\Delta TTT &= \left(\sum_{r\in\mathcal{R}}\sum_{s\in\mathcal{R}} \tilde{f}_{rs}^t \tilde{m}_{rs}^t + \sum_{r,s\in\mathcal{R}}\sum_{n\in\mathcal{N}(s)} \tilde{h}_{rsn}^t \tilde{c}_{rsn}^t\right) - \sum_{r,s\in\mathcal{R}} q_{rs}^t m_{rs}^t \\
&= \sum_{r,s\in\mathcal{R}} \left(\tilde{f}_{rs}^t \tilde{m}_{rs}^t + \sum_{n\in\mathcal{N}(s)} \tilde{h}_{rsn}^t \tilde{c}_{rsn}^t - q_{rs}^t m_{rs}^t\right) \\
&= \sum_{r,s\in\mathcal{R}} \left(\tilde{f}_{rs}^t m_{rs}^t - \tilde{f}_{rs}^t m_{rs}^t - \tilde{f}_{rs}^t \tilde{m}_{rs}^t + \sum_{n\in\mathcal{N}(s)} \tilde{h}_{rsn}^t \tilde{c}_{rsn}^t - q_{rs}^t m_{rs}^t\right) \\
&= \sum_{r,s\in\mathcal{R}} \left(\tilde{f}_{rs}^t m_{rs}^t - \tilde{f}_{rs}^t m_{rs}^t + \tilde{f}_{rs}^t \tilde{m}_{rs}^t + \sum_{n\in\mathcal{N}(s)} \tilde{h}_{rsn}^t \left(\tilde{m}_{rn}^t + w_{ns}\right) - q_{rs}^t m_{rs}^t\right) \\
&= \sum_{r,s\in\mathcal{R}} \left(\tilde{f}_{rs}^t m_{rs}^t - \tilde{f}_{rs}^t m_{rs}^t + \tilde{f}_{rs}^t \tilde{m}_{rs}^t - q_{rs}^t m_{rs}^t\right) \\
&\quad + \sum_{r,s\in\mathcal{R}}\sum_{n\in\mathcal{N}(s)} \tilde{h}_{rsn}^t \left(\tilde{m}_{rn}^t - m_{rn}^t + m_{rn}^t + w_{ns} - m_{ns}^t + m_{ns}^t\right) \\
&= \sum_{r,s\in\mathcal{R}} \left(\tilde{f}_{rs}^t m_{rs}^t + \sum_{n\in\mathcal{N}(s)} \tilde{h}_{rsn}^t \left(m_{rn}^t + m_{ns}^t\right) - q_{rs}^t m_{rs}^t\right) + \sum_{r,s\in\mathcal{R}} \tilde{f}_{rs}^t \left(\tilde{m}_{rs}^t - m_{rs}^t\right) \\
&\quad + \sum_{r,s\in\mathcal{R}}\sum_{n\in\mathcal{N}(s)} \tilde{h}_{rsn}^t \left(\tilde{m}_{rn}^t - m_{rn}^t + w_{ns} - m_{ns}^t\right) \\
&= \sum_{r,s\in\mathcal{R}} \left(\tilde{f}_{rs}^t m_{rs}^t + \sum_{n\in\mathcal{N}(s)} \tilde{h}_{rsn}^t \left(m_{rn}^t + m_{ns}^t\right) - q_{rs}^t m_{rs}^t\right) + \sum_{r,s\in\mathcal{R}} \tilde{f}_{rs}^t \left(\tilde{m}_{rs}^t - m_{rs}^t\right) \\
&\quad + \sum_{r,s\in\mathcal{R}}\sum_{n\in\mathcal{N}(s)} \tilde{h}_{rsn}^t \left(\tilde{m}_{rn}^t - m_{rn}^t + w_{ns} - m_{ns}^t\right) \\
&= \Delta_{\text{Counterfactual}} + \Delta_{\text{PUDO, Remain}} + \Delta_{\text{PUDO, Detour}}
\end{aligned}
\tag{34}
$$

The above decomposition completes the proof of Proposition 5.

Appendix C: Estimation result of $\hat{\theta}_v$ by DSML, DML, and LR

We present the estimation results obtained by DSML, DML, and LR in Table 7.

Table 7: Estimation result by DSML, DML and LR

Regions ID	DSML		DML		LR	
	θ	p-value	θ	p-value	θ	p-value
4	-0.092	0.000***	-0.009	0.021*	-0.162	0.000***
12	-0.036	0.039*	-0.050	0.019*	-0.163	0.000***
13	-0.020	0.000***	-0.012	0.000***	-0.052	0.000***
24	-0.038	0.000***	-0.023	0.000***	-0.169	0.000***
43	-0.050	0.000***	-0.021	0.000***	-0.055	0.000***
45	-0.136	0.000***	0.012	0.008**	-0.080	0.000***
48	-0.040	0.000***	-0.008	0.000***	-0.014	0.000***
50	-0.035	0.000***	-0.011	0.000***	-0.040	0.000***
68	-0.025	0.000***	-0.011	0.000***	-0.013	0.000***
75	-0.043	0.000***	-0.025	0.000***	-0.107	0.000***
79	-0.021	0.000***	-0.002	0.014*	-0.009	0.000***
87	0.002	0.006**	0.001	0.689	-0.015	0.000***
88	-0.031	0.000***	-0.012	0.014*	-0.167	0.000***
90	-0.052	0.000***	-0.007	0.000***	-0.041	0.000***
100	-0.075	0.000***	-0.008	0.000***	-0.041	0.000***
107	-0.039	0.000***	-0.002	0.036*	-0.014	0.000***
113	-0.015	0.000***	-0.001	0.048*	-0.013	0.000***
114	-0.018	0.000***	-0.001	0.339	-0.010	0.000***
125	-0.083	0.000***	-0.015	0.000***	-0.040	0.000***
137	-0.019	0.000***	-0.003	0.293	-0.092	0.000***
140	-0.013	0.000***	-0.010	0.000***	-0.075	0.000***
141	-0.050	0.000***	-0.013	0.000***	-0.023	0.000***
142	-0.029	0.000***	-0.004	0.003**	-0.020	0.000***

Estimation result by DSML, DML and LR (continued)

Region ID	DSML		DML		LR	
	θ	p-value	θ	p-value	θ	p-value
143	-0.031	0.000***	-0.012	0.000***	-0.070	0.000***
144	-0.075	0.000***	-0.003	0.028*	-0.027	0.000***
148	-0.057	0.000***	-0.004	0.042*	-0.016	0.000***
151	-0.012	0.000***	-0.015	0.000***	-0.132	0.000***
158	-0.039	0.000***	-0.007	0.001***	-0.021	0.000***
161	-0.022	0.000***	-0.005	0.000***	-0.015	0.000***
162	-0.035	0.000***	-0.003	0.000***	-0.017	0.000***
163	-0.059	0.000***	-0.007	0.000***	-0.028	0.000***
164	-0.025	0.000***	-0.003	0.032*	-0.018	0.000***
170	-0.029	0.000***	-0.006	0.000***	-0.016	0.000***
186	-0.039	0.000***	-0.005	0.000***	-0.023	0.000***
209	-0.037	0.000***	-0.007	0.123	-0.075	0.000***
211	-0.061	0.000***	-0.007	0.003**	-0.030	0.000***
224	-0.067	0.000***	-0.003	0.570	-0.248	0.000***
229	-0.018	0.000***	-0.012	0.000***	-0.067	0.000***
230	-0.026	0.000***	-0.005	0.000***	-0.017	0.000***
231	-0.040	0.000***	-0.003	0.002**	-0.012	0.000***
232	-0.090	0.000***	-0.004	0.290	-0.150	0.000***
233	-0.019	0.000***	-0.010	0.000***	-0.078	0.000***
234	-0.035	0.000***	-0.000	0.953	-0.012	0.000***
236	-0.019	0.000***	-0.012	0.000***	-0.017	0.000***
237	-0.018	0.000***	-0.009	0.000***	-0.017	0.000***
238	-0.010	0.000***	-0.008	0.000***	-0.058	0.000***
239	-0.023	0.000***	-0.006	0.001***	-0.048	0.000***
246	-0.022	0.000***	-0.014	0.000***	-0.020	0.000***
249	-0.023	0.000***	0.001	0.504	-0.010	0.000***
261	-0.027	0.000***	-0.023	0.000***	-0.090	0.000***
262	-0.025	0.000***	-0.013	0.000***	-0.099	0.000***
263	-0.015	0.000***	-0.001	0.533	-0.023	0.000***

[a] ***$p \leq 0.001$, highly significant
[b] **$p \leq 0.001$, very significant
[c] *$p \leq 0.005$, significant
[d] $p > 0.05$, not significant

References

Abrevaya J, Hsu YC, Lieli RP, 2015 *Estimating conditional average treatment effects. Journal of Business & Economic Statistics* 33(4):485–505.

Agarwal S, Mani D, Telang R, 2019 *The impact of ride-hailing services on congestion: Evidence from indian cities. Available at SSRN 3410623* .

Ahmed MM, Ghasemzadeh A, 2018 *The impacts of heavy rain on speed and headway behaviors: an investigation using the shrp2 naturalistic driving study data. Transportation research part C: emerging technologies* 91:371–384.

Anurag K, Kalin P, Mohammadreza K, Pragun V, Arun K, 2019 *Mind the curb: Findings from commercial vehicle curb usage in california. Transportation Research Board (TRB) 98th Annual Meeting* .

Arnott R, Rowse J, 2013 *Curbside parking time limits. Transportation Research Part A: Policy and Practice* 55:89–110.

Babar Y, Burtch G, 2020 *Examining the heterogeneous impact of ride-hailing services on public transit use. Information Systems Research* 31(3):820–834.

Beojone CV, Geroliminis N, 2021 *On the inefficiency of ride-sourcing services towards urban congestion. Transportation research part C: emerging technologies* 124:102890.

Burtch G, Carnahan S, Greenwood BN, 2018 *Can you gig it? an empirical examination of the gig economy and entrepreneurial activity. Management Science* 64(12):5497–5520.

Butrina P, Le Vine S, Henao A, Sperling J, Young SE, 2020 *Municipal adaptation to changing curbside demands: Exploratory findings from semi-structured interviews with ten us cities. Transport Policy* 92:1–7.

Castiglione J, Chang T, Cooper D, Hobson J, Logan W, Young E, Charlton B, Wilson C, Mislove A, Chen L, et al., 2016 *Tncs today: a profile of san francisco transportation network company activity. San Francisco County Transportation Authority (June 2016)* .

Castiglione J, Cooper D, Sana B, Tischler D, Chang T, Erhardt GD, Roy S, Chen M, Mucci A, 2018 *Tncs & congestion* .

Chai H, Rodier C, 2020 *Automated vehicles and central business district parking: The effects of drop-off-travel on traffic flow and vehicle emissions [supporting dataset]* .

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J, 2018 *Double/debiased machine learning for treatment and structural parameters.*

Dong T, Luo Q, Xu Z, Yin Y, Wang J, 2022 *Strategic driver repositioning in ride-hailing networks with dual sourcing. Available at SSRN* .

Erhardt GD, Roy S, Cooper D, Sana B, Chen M, Castiglione J, 2019 *Do transportation network companies decrease or increase congestion? Science advances* 5(5):eaau2670.

Fiebig DG, Bartels R, 1996 *The frisch-waugh theorem and generalized least squares. Econometric Reviews* 15(4):431–443.

Golias I, Karlaftis M, 2001 *The taxi market in athens, greece, and its impacts on urban traffic. Transportation Quarterly* 55(1).

Goodchild A, MacKenzie D, Ranjbari A, Machado J, Chiara GD, 2019 *Curb allocation change project.*

Greenwood BN, Wattal S, et al., 2017 *Show me the way to go home: An empirical investigation of ride-sharing and alcohol related motor vehicle fatalities. MIS Q.* 41(1):163–187.

Han LD, Chin SM, Franzese O, Hwang H, 2005 *Estimating the impact of pickup-and delivery-related illegal parking activities on traffic. Transportation Research Record* 1906(1):49–55.

Hastie T, Tibshirani R, Friedman JH, Friedman JH, 2009 *The elements of statistical learning: data mining, inference, and prediction*, volume 2 (Springer).

He F, Yan X, Liu Y, Ma L, 2016 *A traffic congestion assessment method for urban road networks based on speed performance index. Procedia engineering* 137:425–433.

Iqbal M, 2019 *Uber revenue and usage statistics.* URL https://www.businessofapps.com/data/uber-statistics/.

Jaller M, Rodier C, Zhang M, Lin H, Lewis K, 2021 *Fighting for curb space: Parking, ride-hailing, urban freight deliveries, and other users* .

Künzel SR, Sekhon JS, Bickel PJ, Yu B, 2019 *Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the national academy of sciences* 116(10):4156–4165.

Liu J, Ma W, Qian S, 2022 *Optimal curbside pricing for managing ride-hailing pick-ups and drop-offs. Available at SSRN 4068718* .

Liu W, Zhang F, Yang H, 2021 *Modeling and managing the joint equilibrium of destination and parking choices under hybrid supply of curbside and shared parking. Transportation Research Part C: Emerging Technologies* 130:103301.

Lovell MC, 2008 *A simple proof of the fwl theorem. The Journal of Economic Education* 39(1):88–91.

Lu R, 2019 *Pushed from the curb: Optimizing curb space for use for ride-sourcing vehicles. Transportation Research Board (TRB) 98th Annual Meeting* .

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* 119:102747.

Ma W, Qian S, 2020 *Measuring and reducing the disequilibrium levels of dynamic networks with ride-sourcing vehicle data. Transportation Research Part C: Emerging Technologies* 110:222–246.

Ma W, Qian ZS, 2018 *Estimating multi-year 24/7 origin-destination demand using high-granular multi-source traffic data. Transportation Research Part C: Emerging Technologies* 96:96–121.

McCormack E, Goodchild A, Sheth M, Hurwitz D, Jashami H, Cobb DP, 2019 *Developing design guidelines for commercial vehicle envelopes on urban streets* .

Mitman MF, Davis S, Armet IB, Knopf E, 2018 *Curbside management practitioners guide.* Technical report.

Oprescu M, Syrgkanis V, Wu ZS, 2019 *Orthogonal random forest for causal inference. International Conference on Machine Learning*, 4932–4941 (PMLR).

Pearl J, 2009 *Causality* (Cambridge university press).

Pearl J, 2019 *The seven tools of causal inference, with reflections on machine learning. Communications of the ACM* 62(3):54–60.

Rahaim J, 2019 *Transportation impact analysis guidelines.* Technical report, San Francisco Planning Department.

Retallack AE, Ostendorf B, 2019 *Current understanding of the effects of congestion on traffic accidents. International journal of environmental research and public health* 16(18):3400.

Rong L, Cheng H, Wang J, 2017 *Taxi call prediction for online taxicab platforms. Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data*, 214–224 (Springer).

Schaller B, Maguire T, Stein D, Ng W, Blakeley M, 2011 *Parking pricing and curbside management in new york city.* Technical report.

Smith A, Wylie A, Salzberg A, Womeldorff E, Rubendall G, Ballus-Armet I, 2019 *A data driven approach to understanding and planning for curb space utility.* Technical report.

Tirachini A, 2020 *Ride-hailing, travel behaviour and sustainable mobility: an international review. Transportation* 47(4):2011–2047.

Wager S, Athey S, 2018 *Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association* 113(523):1228–1242.

Wijayaratna S, 2015 *Impacts of on-street parking on road capacity. Australasian Transport Research Forum*, 1–15.

Xu Z, Chen Z, Yin Y, 2019 *Equilibrium analysis of urban traffic networks with ride-sourcing services. Available at SSRN 3422294* .

Xu Z, Yin Y, Zha L, 2017 *Optimal parking provision for ride-sourcing services. Transportation Research Part B: Methodological* 105:559–578.

Yao L, Chu Z, Li S, Li Y, Gao J, Zhang A, 2021 *A survey on causal inference. ACM Transactions on Knowledge Discovery from Data (TKDD)* 15(5):1–46.

Yao L, Li S, Li Y, Xue H, Gao J, Zhang A, 2019 *On the estimation of treatment effect with text covariates. International Joint Conference on Artificial Intelligence.*

Yuan K, Knoop VL, Hoogendoorn SP, 2015 *Capacity drop: Relationship between speed in congestion and the queue discharge rate. Transportation Research Record* 2491(1):72–80.

Zalewski AJ, Buckley SM, Weinberger RR, 2012 *Regulating curb space: developing a framework to understand and improve curbside management.* Technical report.

Zhang K, Mittal A, Djavadian S, Twumasi-boakye R, Nie M, 2021 *RIde-hail VEhilce Routing (RIVER) as a congestion game. Available at SSRN 3974957* .

Zhang K, Nie M, 2021 *Mitigating traffic congestion induced by transportation network companies: a policy analysis. Available at SSRN 3882173* .