Appendix A: Nomenclature

The list of notations used in this paper is shown in Table 3.

Table 3: 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.

$\tilde{h}^t_{rsn}$	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

$\tilde{d}^t_s$	The number of drop-off in region s after re-routing in the time interval t.
Δ^t_s	The change of the NoPUDO in region s before and after re-routing in the time interval t.
$\tilde{y}^t_s$	Updated traffic speed each re-routing.
m^t_{rs}	Travel time from region r to region s before re-routing in the time interval t.
$\tilde{m}^t_{rs}$	Travel time from region r to region s after re-routing in the time interval t.
$\tilde{c}^t_{rsn}$	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 B: Conditional Average Treatment Effects of DSML (CATE)

To capture and estimate the nonlinear congestion effect of NoPUDO under different traffic conditions, we design and develop the conditional average treatment effects estimation in the DSML method. We extend the DSML method by adding conditional variables to obtain the dynamic congestion effect estimation under different traffic conditions. The CATE affords us to examine and explore the nonlinear relationship between the NoPUDO and traffic speed given different traffic conditions over time. We propose the Assumption 4 for estimating the CATE of the DSML method.

ASSUMPTION 4 **(Non-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^t is defined in Equation* (24).

$$y_v^t|_{\texttt{do}(d_v^t=d_1)} - y_v^t|_{\texttt{do}(d_v^t=d_2)} = \theta_v^t\left(\mathbf{Y}_v^t, \mathbf{Y}_{\mathcal{N}(v)}^t, \mathbf{W}_v^t\right) \cdot (d_1 - d_2) \tag{24}$$

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. The $\theta_v^t\left(\mathbf{Y}_v^t, \mathbf{Y}_{\mathcal{N}(v)}^t, \mathbf{W}_v^t\right)$ indicates that the congestion effect of NoPUDO at time interval t and region v is affected by the traffic dynamics and other external attributes.*

Mathematically, the CATE of the DSML method also contains three submodels: Model `Y`, Model `D`, and Model `Z`. The first two sub-models, Model `Y` and Model `D` in the CATE are the same as that of ATE, as shown in Equation (8) and Equation (10). The Model `Z` of CATE in the DSML method can refer to Equation (25).

$$\hat{\epsilon}_v^t = \mathcal{X}^T \beta_v^t \hat{\xi}_v^t + \hat{e}_v^t \tag{25}$$

where $\hat{e}_v^t$ represents the random error of the linear regression model.

First, we obtain the residual $\hat{\epsilon}_v^t$ from Model `Y` as shown in Equation (9) and residual $\hat{\xi}_v^t$ from Model `D` as shown in Equation (11). As for the Model `Z` for the CATE, it is a linear regression between the $\hat{\epsilon}_v^t$ and $\hat{\xi}_v^t$

given the conditions $\mathcal{X}$. The $\mathcal{X}$ represents the time-varying traffic conditions. In Equation (25), the value of $\hat{\epsilon}_v^t$, $\hat{\xi}_v^t$, $\mathcal{X}$ is obtained from the Model `Y`, Model `D`, and $\left(\mathbf{Y}_v^t, \mathbf{Y}_{\mathcal{N}(v)}^t, \mathbf{W}_v^t\right)$ respectively. The Model `Z` shown in Equation (25) is to learn the coefficient β_v^t. After obtaining the value of the coefficient β_v^t, we plug it into Equation (26) to calculate the estimated $\hat{\theta}_v^t$:

$$\hat{\theta}_v^t = \mathcal{X}^T \beta_v^t. \tag{26}$$

We further extend the numerical experiments to estimate the CATE for $\hat{\theta}_v^t$, and the estimated congestion effect is shown in Figure 12. We show the results for region 100, 186, 236, and 263, the same region example as shown in Figure 11. The CATE value of $\hat{\theta}_v^t$ is plotted from 16:00 to 20:00 across the whole study period. The x-axis is the time point and the y-axis represents the average value of $\hat{\theta}_v^t$. The purple line represents the mean of the $\hat{\theta}_v^t$ at these time points with the deviation. The four subfigures indicate that the average value of $\hat{\theta}_v^t$ is negative, which meets our assumption and matches with the ATE results. The fluctuation also indicates the dynamic congestion effect of NoPUDO under different traffic conditions.

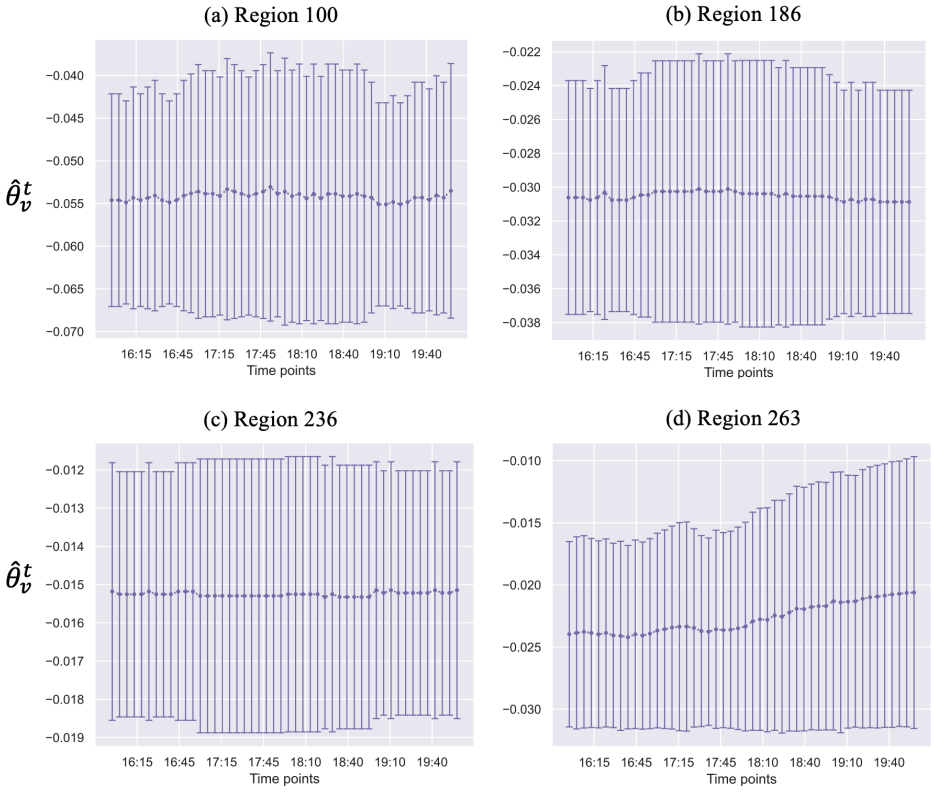


Figure 12: CATE results $\hat{\theta}_v^t$ on weekdays.

Appendix C: Robustness check of the contemporaneous time interval index setting

In the structural equation model shown in Equation (2), we use the same time interval index to capture the congestion effect between NoPUDO and traffic speed. To provide robust justifications for the use of the contemporaneous time interval index given the coarse time granularity of the current datasets, we quantify the time series correlation between traffic speed and NoPUDO over time by a distributed lag model, as shown in Equation (27) (Amemiya and Fuller 1967).

$$y_v^t = \iota_0 d_v^t + \iota_1 d_v^{t-1} + \cdots + \iota_I d_v^{t-I} + \kappa_v^t, \forall v, t, \tag{27}$$

where κ_v^t is the error term. The coefficients ι_0, ι_1 ... and ι_I quantify the correlation between traffic speed and NoPUDO with respect to different time intervals.

In the experimental setting, we set $I = 10$, and feed the whole dataset for different v, t into the above equation to capture the overall trend of the effect of NoPUDO on traffic speed over time. We separately estimate the coefficients based on datasets of weekdays and weekends. The estimation result is shown in the Table 4. First, the coefficients present a negative effect of NoPUDO on traffic speed over time. Second, the coefficient of d_v^t is statistically significant and the most negative compared to those at other time intervals, such as d_v^{t-1}, d_v^{t-2}... and d_v^{t-10}. Therefore, given the current data granularity based on 5 minutes, the empirical result validates that the congestion effect requires us to make an estimation based on traffic speed and NoPUDO with the same time interval in this study.

Table 4: Estimation result of time series correlation by the distributed lag model.

Variables	**Weekdays**	**Weekends**
NoPUDO at time interval t	-0.024***	-0.024***
	(0.001)	(0.001)
NoPUDO at time interval $t-1$	-0.012***	-0.012***
	(0.001)	(0.001)
NoPUDO at time interval $t-2$	-0.005***	-0.006***
	(0.001)	(0.001)
NoPUDO at time interval $t-3$	-0.001*	-0.004***
	(0.001)	(0.001)
NoPUDO at time interval $t-4$	0.001	-0.002**
	(0.001)	(0.001)
NoPUDO at time interval $t-5$	0.002***	-0.001
	(0.001)	(0.001)
NoPUDO at time interval $t-6$	0.002***	-0.001
	(0.001)	(0.001)
NoPUDO at time interval $t-7$	0.001	-0.002**
	(0.001)	(0.001)
NoPUDO at time interval $t-8$	-0.002***	-0.004***
	(0.001)	(0.001)
NoPUDO at time interval $t-9$	-0.007***	-0.007***
	(0.001)	(0.001)
NoPUDO at time interval $t-10$	-0.017***	-0.015***
	(0.001)	(0.001)
Intercept	18.622***	20.450***

Estimation result of time series correlation by the distributed lag model (continued)

Variables	Weekdays	Weekends
	(0.010)	(0.014)

Note. Standard errors are in parentheses.
*p < 0.1, **p < 0.05, ***p < 0.01.

Appendix D: 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.

D.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 (28).

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

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 (29) and (30).

$$y_v^t = \mathbf{A_y}^T \mathbf{Y}_v^{t-I:t-1} + \mathbf{B_y}^T \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1} + \hat{\varepsilon}_v^t \tag{29}$$

$$d_v^t = \mathbf{A_d}^T \mathbf{Y}_v^{t-I:t-1} + \mathbf{B_d}^T \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1} + \mathbf{C_d}^T \mathbf{D}_v^{t-I:t-1} + \hat{\xi}_v^t \tag{30}$$

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 $\hat{\psi}_v$.

We consider an alternative least-squares regression question:

$$\hat{\varepsilon}_v^t = \hat{\theta}_v \hat{\xi}_v^t + \hat{e}_v^t \tag{31}$$

To analyze the property of $\hat{\theta}_v$, we derive $\hat{\varepsilon}_v^t$ by substituting Equation (28) into Equation (29), as shown in Equation (32).

$$\hat{\varepsilon}_v^t = \theta_v d_v^t + (\mathbf{A} - \mathbf{A_y})^T \mathbf{Y}_v^{t-I:t-1} + (\mathbf{B} - \mathbf{B_y})^T \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1} + e_v^t \tag{32}$$

Then we plug the variable d_v^t in the Equation (30) into Equation (32). Eventually, we can formulate the $\hat{\varepsilon}_v^t$ in the Equation (33).

$$\hat{\varepsilon}_v^t = \theta_v \hat{\xi}_v^t + (\theta_v \mathbf{A_d} + \mathbf{A} - \mathbf{A_y})^T \mathbf{Y}_v^{t-I:t-1} + (\theta_v \mathbf{B_d} + \mathbf{B} - \mathbf{B_y})^T \mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1} + (\theta_v \mathbf{C_d})^T \mathbf{D}_v^{t-I:t-1} + e_v^t \tag{33}$$

As $\hat{\varepsilon}_v^t$ is the residual from the linear regression in Equation (29), 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, $\hat{\varepsilon}_v^t$ 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 (33). Consequently, we have Equation (34) holds.

$$\hat{\varepsilon}_v^t = \theta_v\hat{\xi}_v^t + e_v^t \tag{34}$$

By comparing Equation (31) and Equation (34), we have Equation (35) holds.

$$\begin{aligned}\hat{\theta}_v &= \theta_v\\ \hat{e}_v^t &= e_v^t\end{aligned} \tag{35}$$

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.

D.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 (36)

$$\omega(\theta_v,\eta_v)=(y_v^t-\theta_v d_v^t-\varphi_v)\,(d_v^t-\psi_v) \tag{36}$$

We note that $\omega(\theta_v,\eta)$ is insensitive to the small change of either φ_v or θ_v, as presented in Equation (37).

$$\partial_{\eta_v}\mathbb{E}\omega(\theta_v,\eta_v)[\eta_v-\eta_v^0]=0 \tag{37}$$

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 E: Causality analysis of DSML

The central idea of estimating causality is to construct a counterfactual world that matches the factual world (Lewis 1973, 1986, 2004, Durand and Vaara 2009), and the difference in the dependent variable between the counterfactual and factual world is the targeted causality. In general, there are two primary approaches to constructing the counterfactual world, including experimental and non-experimental approaches. RCT randomly separates participants into control and treatment groups and only conducts interventions in the treatment group. Regarding the treatment group, the status of the control group is considered as the counterfactual world. However, such controlled experiments are not always feasible for complex real-world transportation scenarios (Greene 2003, Angrist and Pischke 2009, Imbens and Rubin 2015, Gordon, Moakler, and Zettelmeyer 2023). The second approach is to leverage historical data to make a prediction of the dependent variable and construct the counterfactual world based on the predicted results. The DSML method, which falls under the second approach, formulates the counterfactual world from a data-driven standpoint. In our context, the predicted numerical value of NoPUDO and traffic speed can be regarded as the results from the counterfactual world. We unfold the congestion effect estimation using the second approach due to the unavailable experiment setting and infeasible treatment manipulation. Specifically, it is challenging to

conduct a field experiment by setting up the control and treatment groups due to the inevitably disruptive nature of the real traffic situation. Furthermore, the treatment in our setting is a continuous number of pick-ups and drop-offs. It is demanding to manipulate the treatment, NoPUDO, in the intricate and dynamic real-world road networks.

We consistently follow a well-established procedure of causality inference to conduct the congestion effect estimation in the study. Park, Tafti, and Shmueli (2023) propose that causal estimation can be derived from the following steps: (1) identifying related variables; (2) clarifying causal relationships; (3) constructing causal diagrams; (4) determining the possibility of transporting historical data from other existing causal scenarios; (5) formulating structural equation based on causal diagrams; and (6) making a final causality estimation. In the same vein, we first identify the research gap through an exhaustive literature review, focusing on investigating the congestion effect of NoPUDO on traffic speed. Second, we justify and conclude the causal relationship between NoPUDO and traffic speed based on empirical evidence, as shown in Figure 1. Third, we extrapolate and conceptualize the causal relationship between NoPUDO and traffic speed in the causal diagram as presented in Figure 3. This causal diagram integrates the transportation domain knowledge to demonstrate the intertwined temporal and spatial relationship among involved variables. Fourth, we skip the fourth step mentioned above because the congestion effect has not been extensively investigated in pioneering literature and there are few existing similar causal scenarios. Fifth, we mathematically formulate the causal graph and prove rigorously these assumptions and equations in Section 2.1.2. Last, we conduct the congestion estimation using the observational data of NoPUDO and traffic speed in Section 4.

The DSML method can guarantee that our estimates are of causality rather than correlation by dual channels: causal graph drawing and mathematical formulas validation. In this study, we first utilize the causal graph to delineate the causal relationship among the variables and apply the machine learning method to estimate the congestion effect based on the above causal graph. The causality analysis of the DSML method first needs to trace back to the well-built causal graph, as shown in Figure 3. An essential principle for ensuring that the estimation represents causality rather than correlation is that we should eliminate the additional influence of confounders on the estimation. The causal graph conveys the causal assumption and empowers us to identify and control confounding variables in the estimation (Manzoor et al. 2023).

Specifically, in Model `Y`, the confounders are the historical traffic speed record at the current region $\mathbf{Y}_v^{t-I:t-1}$, the historical traffic speed record at the neighboring regions $\mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$, and external control variables $\mathbf{W}_v^t$. In a similar vein, in the Model `D`, the confounders are the historical NoPUDO in the current region $\mathbf{D}_v^{t-I:t-1}$, the historical traffic speed record in the current region $\mathbf{Y}_v^{t-I:t-1}$, the historical traffic speed record in the neighboring region $\mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$, and external control variables $\mathbf{W}_v^t$. Essentially, the central idea is to capture and model the trend of the time series, and hence the historical records of speed and NoPUDO are the confounders. Furthermore, the congestion effect is estimated separately and independently for each region, which also has controlled implicit confounders induced by the region heterogeneity. Comprehensively controlling confounders in the DSML method can effectively ensure our estimates of causality.

Besides, from the perspective of structural causal modeling, we have rigorously proved that if the above Equations (2), (3), (4), (5), (6) and (7) hold, it can guarantee that we can accurately and unbiasedly

estimate the causal effects of NoPUDO on traffic speed. The machine learning-based DSML method is different from the traditional econometric methods, such as Difference-in-Difference (DID) which mainly emphasizes the difference in the value of the dependent variable before and after adding treatment. In the DSML method, we first formulate and capture the correlation between the traffic speed and NoPUDO in Equation (2). To distill the pure causality, we included and formulated the confounders in Model `Y` and Model `D`, as shown in Equation (8) and (10). The non-linear relationship between the confounding variables and traffic speed/NoPUDO is obtained by parameter learning based on machine learning methods. The residuals represent the remaining parts of traffic speed and NoPUDO after excluding the disturbance of confounders (Xu, Ghose, and Xiao 2023). In the Model `Z`, the regression between the above two residuals captures and quantifies the causality. The double-layer design of the DSML method can elegantly address the issue of confounding correlation between independent variables and covariates to obtain the causality, which remains challenging in one single regression as shown in Equation (2) (Dube et al. 2020).

The DSML method mainly differs from the traditional econometric models in partially linear assumption and means of parameters learning. First, traditional econometric models, such as ordinary least squares (OLS), assume everything linear, while the DSML method emphasizes the partially linear assumption, and models the non-linear relationship between covariates and the independent variable by machine learning models (Chernozhukov et al. 2018). The partially linear assumption is more adaptive and applicable to real-world scenarios and can cater to the emerging complex estimation demands. Second, the DSML method adopts machine learning methods to formulate the non-linear relationship in Model `Y` and `D`, which delve deep into data and enable greater flexibility in estimation. In the traditional econometric methods, there are pre-defined mathematical formulas defined before we feed the data set to estimate the targeted coefficient. In the DSML method, two functions, $\hat{\varphi}_v$ and $\hat{\psi}_v$, are fully trained based on the data, and the parameters are learned by the machine learning models. The DSML method guarantees an unbiased estimation by implementing Neyman-orthogonal scores and cross-validation to overcome the biases caused by using machine learning methods. We will further discuss the implementation details of the DSML method in Section 3.1.

To have an intuitive understanding of the causality in the DSML method, we use an example of a factual world and a counterfactual world to unfold and illustrate the underlying mechanism (Morgan and Winship 2015). Specifically, the treatment in this study is the amount of the PUDO. The factual world is that we increase or decrease the number of PUDO, and mainly then observe the change of the actual traffic speed. The counterfactual world is if the number of PUDO did not increase or decrease, we can observe the change in the actual traffic speed accordingly. An example of the factual and counterfactual world is shown in Table 5. Given a region, we identify the world at the time interval $t-1$ as the time period before applying the treatment. The world at the time interval t is regarded as the time period after receiving the treatment. As shown in Table 5, the NoPUDO increases by 5. The additional 5 units of PUDO is the treatment, and the actual treatment is continuous. As for the factual world in the time interval $t-1$ and t, we can observe the difference in the traffic speed in the current region. As for the counterfactual world, there is no difference in NoPUDO for both the before and after time periods. Similarly, we can observe the numerical value of the traffic speed. The difference in traffic speed in the time interval $t-1$ and t is due to the change in the

time series. In this case, the targeted causality is the difference between two differences: one is the disparity of traffic speed at the "Before" and "After" time periods in the factual world; another is the difference of traffic speed at the "Before" and "After" time periods in the counterfactual world.

Table 5: An example of factual and counterfactual world

World (Group)	Before: at time interval $t-1$		After: at time interval t	
	Traffic speed (mph)	NoPUDO	Traffic speed (mph)	NoPUDO
Factual world (Treated)	15	25	14	30
Counterfactual world (Controlled)	15	25	14.5	25

Appendix F: 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 (38).

$$
\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{38}
$$

The above decomposition completes the proof of Proposition 5.

Appendix G: Sensitivity analysis regarding the selection 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.

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 6. 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`.

Table 6: Sensitivity analysis of ML models used in the DSML method.

ML models	GB	RF	Ada
Correlation coefficient	0.99	0.94	0.83

Table 7: Correlation analysis for DSML vs. DML and DSML vs. LR.

Models	DML	LR
Correlation coefficient	-0.14	0.27

Appendix H: Sensitivity analysis regarding the region size

In this section, we conduct the sensitivity analysis to examine the effect of region size on the estimation result in the DSML method. The definition of the region in this paper follows the NYC Department of City Planning's Neighboring Tabulation Areas. The above estimation result is analyzed and obtained with a single region as a unit. How the region size affects the estimation result in the DSML method is not well studied and examined. Conducting the sensitivity analysis of region size lays a solid foundation for us to extend and promote the DSML method into other scenarios and enhance its generalization. In this sensitivity analysis, we first merge the neighboring regions into an aggregate region, which has a larger zone size than a single region. Subsequently, the DSML method can be re-applied to the aggregate region to estimate the congestion effect of NoPUDO for the entire aggregate region.

We take regions 48, 68, and 100 as a case study to examine how the region size affects the estimation result. Region 100, belonging to the Midtown, has been studied in the residual analysis in Figure 11. As shown in Figure 7, region 48 belongs to the Clinton area, region 68 is within the Hudson Yards area, and both regions are adjacent to region 100. We run the DSML method by viewing the two or three regions as a whole region with a large size. The estimation result of these aggregate regions can justify the effectiveness of the DSML method in different region sizes.

As for the regions 48, 68, and 100, there are four types of combinations, denoted as case 1 (aggregating regions 48 and 68), case 2 (aggregating regions 48 and 100), case 3 (aggregating regions 68 and 100), and case 4 (aggregating regions 48, 68, and 100). The estimation result of the above four cases on weekdays and weekends is shown in Table 8. As for the four cases, the numerical value of the estimated congestion effect on the aggregated regions is within the range of that of its composing regions. The estimated effects of the aggregate region are generally the average of the effects of those disaggregated regions. The variation in the estimation result of the above cases indicates that the region size is a critical factor affecting the estimation result.

The effect of the region size on the congestion effect estimation can be disentangled from the perspective of the underlying mathematical mechanism and calculation principle. In this paper, we adopt the NYC Taxi zone system to divide regions in Manhattan and conduct the congestion effect estimation based on the region level. Specifically, when we preprocess data to obtain the average traffic speed, y_v^t and NoPUDO d_v^t, we firstly fence each region following the NYC Taxi zone system and calculate the average traffic speed and count the total number of PUDO in the current region. The numerical value of these variables will be used to estimate the congestion effect at the region level in the DSML method. If different analysis units are used, such as the street or city level, the data generation for the above variables would also be different. The estimation result on the smaller analysis unit would be more precise due to finer-grained data collection and effectively

controlled variables regarding each smaller area, but the fine-grained estimation might suffer the issue of limited trip record data. An intuitive example is that trip records are not evenly distributed in the streets. There are few trip records at the street level during non-peak hours.

Generally, the sensitivity analysis regarding the region size demonstrates that the zone scale affects the congestion effect estimation. Theoretically, the DSML method can be adaptive to estimating the congestion effect for different analysis units, including the street level, the region level, the city level, and so on. Compared to the coarse-grained level, the fine-grained unit requires running the DSML method based on the data at a more disaggregate level. The estimation based on the fine-grained and small area implicitly controls other covariates, which guarantees a more precise estimation result. However the limited data raises the issue of data lacking for small areas. Trading off between analysis granularity and data size limitation should depend on specific research questions to be solved. Specifically, in our study, we aim at designing and taking the region-based re-routing strategy to mitigate the congestion effect, and hence the estimation of the congestion effect of NoPUDO should also be conducted based on the region level.

Table 8: Estimation result of aggregate regions by DSML

Regions ID	Weekdays θ	Weekends θ
48	-0.040***	-0.036***
	(0.001)	(0.001)
68	-0.025***	-0.021***
	(0.001)	(0.001)
100	-0.075***	-0.090***
	(0.001)	(0.001)
Case 1 (48 & 68)	-0.035***	-0.032***
	(0.001)	(0.001)
Case 2 (48 & 100)	-0.052***	-0.057***
	(0.001)	(0.001)
Case 3 (68 & 100)	-0.048***	-0.050***
	(0.001)	(0.001)
Case 4 (48 & 68 & 100)	-0.043***	-0.051***
	(0.001)	(0.001)

Note. Standard errors are in parentheses.
*p $<$ 0.1, **p $<$ 0.05, ***p $<$ 0.01.

Appendix I: Comparison among DSML, DML, and LR

We compare the developed DSML method with the standard DML and classical LR methods in terms of involved features, outcome variables, and methods, as shown in Table 9. Both DSML and DML methods consist of three sub-models: the first and second sub-models use machine learning models to predict y_v^t and d_v^t separately and the third sub-model runs a linear regression between the above residuals from the first two sub-models to estimate the congestion effect. One of the biggest differences between DSML and DML methods lies in the involved features. Specifically, Model `Y` does not include $\mathbf{D}_v^{t-I:t-1}$ to predict traffic speed y_v^t.

Table 9: 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

The estimation result of $\hat{\theta}_v$ using the DML method on weekdays is shown in Figure 13 (a). The estimation results on weekends are similar and hence omitted. On average, $\hat{\theta}_v$ is -0.008 estimated by DML. The absolute numerical value of the estimated $\hat{\theta}_v$ by DML is generally smaller than that by DSML, and $\hat{\theta}_v$ for all regions is almost identically small. From the spatial distribution of the estimated results in Figure 13 (a), it is obvious that the DML method cannot accurately capture the heterogeneity of the congestion effects among different regions. The failure in this estimation of the DML method is attributed to that it 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.

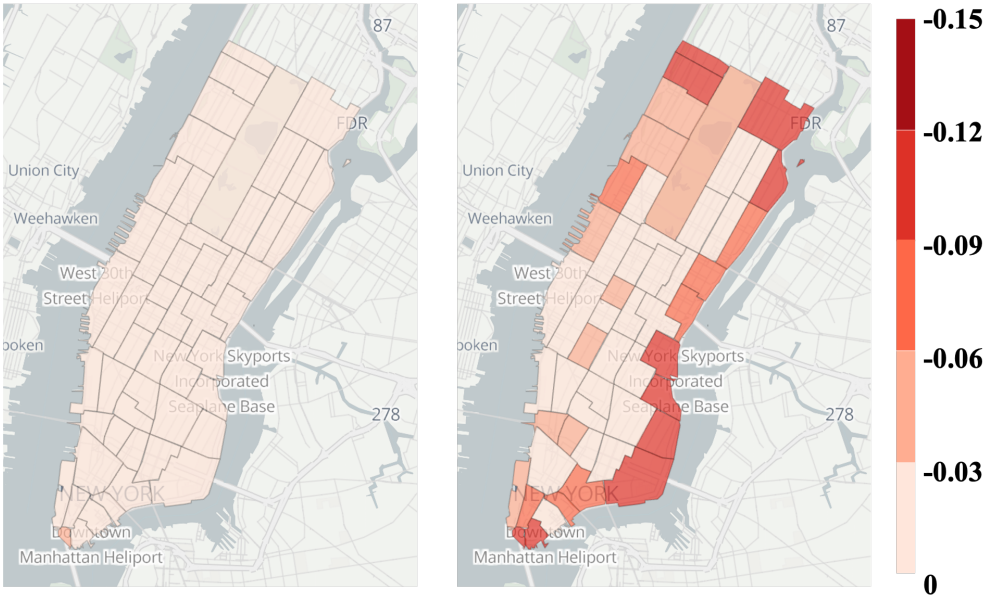
We also compare the LR and DSML methods from multiple aspects of the theoretical underpinnings and estimation results. LR emphasizes modeling the linear relationship between the dependent variable y_v^t and independent variable d_v^t directly in the linear regression, and the DSML method first focuses on using advanced machine learning methods to make predictions separately and obtain residuals from the first two sub-models before the final linear regression. The adoption of machine learning methods in the DSML method significantly enhances flexibility and improves the effectiveness of the estimation, but it might also induce biased estimation because of regularization and overfitting (Chernozhukov et al. 2018, Gordon, Moakler, and Zettelmeyer 2023). As we present in Section 3.1 regarding the underlying solution algorithm, the DSML method overcomes the bias by using Neyman-orthogonal moments/scores in the regularization and eliminating the bias from overfitting through implementing the cross-validation.

Traditional LR, while appealing for its formulation understandability and coefficient interpretability, will face challenges when addressing the high-dimensional variables and separating the correlation between the covariates and independent variables. First, intricate transportation scenarios require us to consider and include many control variables representing the intertwined spatial-temporal relationship in the congestion effect estimation. It triggers highly complex computations to estimate the coefficient in the presence of high-dimensional nuisance parameters. The actual computation violates the traditional assumption (e.g. Donsker properties) that limits the complexity of the parameter space in the classical semi-parametric setting (Chernozhukov et al. 2018). Secondly, the intuition of using two separate structural equations in the DSML method is to address the endogeneity issues raised by the correlation between the covariates $\mathbf{Y}_v^{t-I:t-1}$, $\mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$ and the independent variable d_v^t (Xu, Ghose, and Xiao 2023). One contained in $\mathbf{Y}_v^{t-I:t-1}$ or $\mathbf{Y}_{\mathcal{N}(v)}^{t-I:t-1}$ could be correlated with the independent variable d_v^t in Equation (2). The residuals of Model `Y` and Model `D` are the remaining parts that have not been explained by the covariables. Leveraging the residuals to run the

final regression can mitigate the disturbance from covariates, resulting in extracting and distilling the pure causal effect.

The estimation result of $\hat{\theta}_v$ using the LR method on weekdays is shown in Figure 13 (b). The average numerical value of $\hat{\theta}_v$ estimated by LR is -0.055. Based on the spatial distribution of the estimated congestion effect, LR overlooks the intertwined 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. In conclusion, compared to the DSML method, LR is appealing for its explainability due to its understandable operation mechanism. However, it cannot be applied in the high-dimensional parameters due to the existing inherent assumption. A conflict arises between the practical requirements to address practical issues and the constraints of the traditional LR method.

To further distinguish the difference in estimation results obtained from DSML, DML, and LR, we conduct the correlation analysis and t-test to delve deeper into our comparison and analysis. First, the correlation coefficients between DML and DSML, LR and DSML are shown in Table 7. The correlation coefficients can offer an overview regarding whether the spatial distribution of the estimated congestion effect is correlated. The small numerical value of the coefficients in Table 7 indicates that the estimated $\hat{\theta}_v$ by DSML is completely different from that estimated by DML or LR. Additionally, we also conduct the t-test and compare the significance level of the estimated coefficients, and the results show that some $\hat{\theta}_v$ estimated by the DML method are not statistically significant, which might be due to the superfluous consideration of the confounding factors $\mathbf{D}_v^{t-I:t-1}$ in Model Y. Although the significance levels of the estimated coefficient from LR are almost statistically significant, the estimated $\hat{\theta}_v$ not only contains the causality but also includes a correlation between NoPUDO and traffic speed, based on our discussion in Example 1.



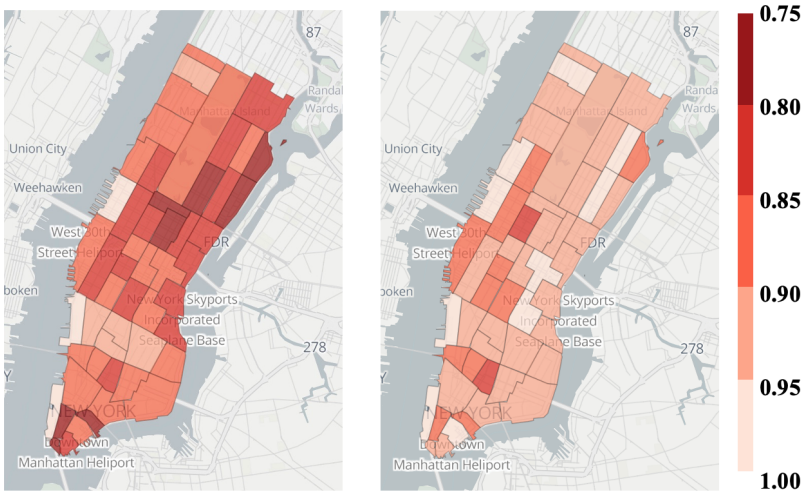
(a) $\hat{\theta}_v$ learned by DML (b) $\hat{\theta}_v$ learned by LR

Figure 13: Comparison of estimated $\hat{\theta}_v$ by DML and LR on weekdays.

Appendix J: Congestion map and correlation analysis

To reveal the congestion level of each region, we propose a traffic indicator, the speed index that calculates the ratio of the traffic speed and the free flow speed. The necessity for such a traffic indicator is because the same value of speed in different regions represents different congestion levels. The speed index can then reflect the relative congestion levels among different regions. The formula of the speed index is shown in Equation (39). The spatial distribution of the speed index of each region on weekdays and weekends is shown in Figure 14.

$$\text{speed index of a region} = \frac{\text{relative speed in the region}}{\text{free flow speed in the region}} \tag{39}$$



(a) Speed index on weekdays (b) Speed index on weekends

Figure 14: Congestion map in the Manhattan area.

As shown in Figure 14, the region marked with a redder color has a smaller value of the speed index, and it represents that the region is more congested. The congestion effect quantifies the negative effect of NoPUDO on traffic speed, and the speed index reflects the traffic situation of the transportation system. The correlation analysis between the speed index and the estimated congestion effect is shown in Table 10. The Pearson correlation coefficient between congestion effect and speed index on weekdays is 0.067, and that value on weekends is 0.403. This indicator measures the overall spatial correlation between the congestion effect and the speed index. The Pearson correlation coefficient shows that the speed index has a higher correlation with the estimated congestion effect on weekends rather than on weekdays.

Table 10: Correlation analysis between speed index and estimated congestion effect.

Speed index	Speed index on weekdays	Speed index on weekends
Correlation coefficient	0.067	0.403

Appendix K: Sensitivity analysis of re-routing formulation

To evaluate the sensitivity with respect to demand level, we first perturb λ to be 5, 10, 20, 25, and 30 and evaluate the according 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 15 and Figure 16, respectively.

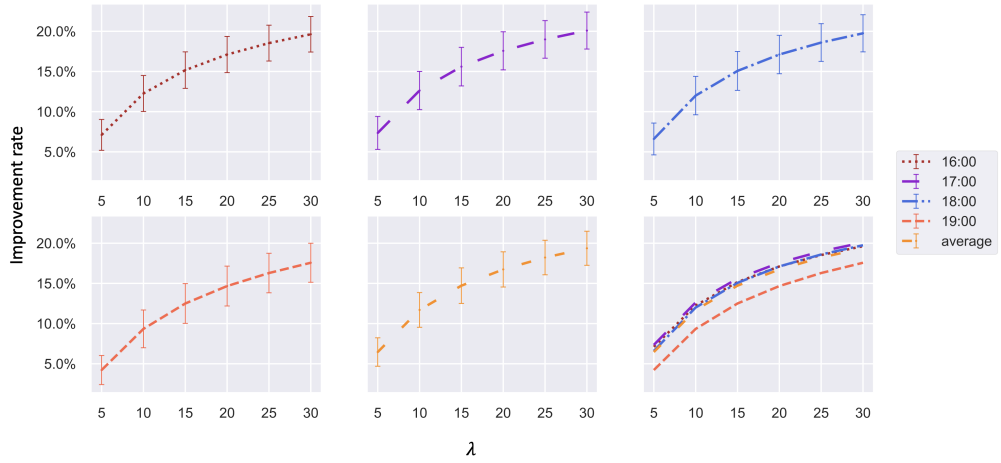


Figure 15: Improvement rates on different λ in Midtown (error bar represents the standard deviation).

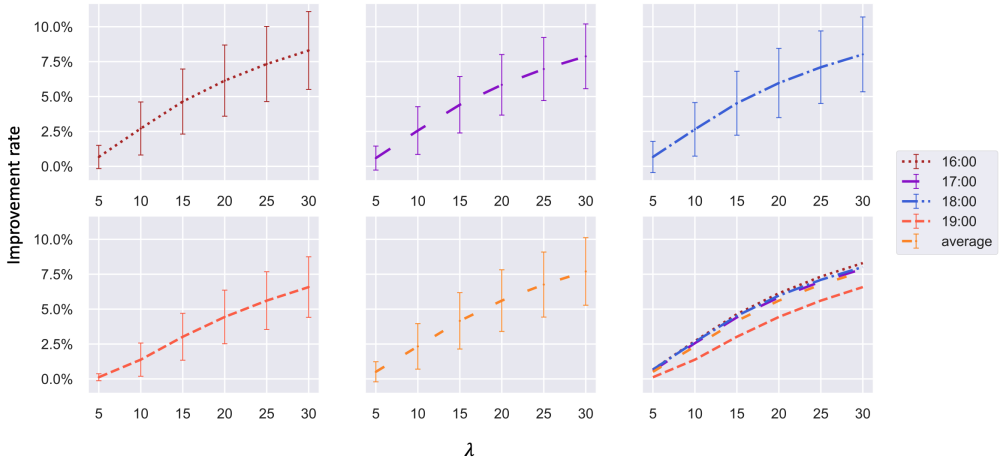


Figure 16: Improvement rates on different λ in Central Park (error bar represents the standard deviation).

In general, higher traffic demands encourage a larger improvement rate for both areas. Re-routing 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 17 and Figure 18.

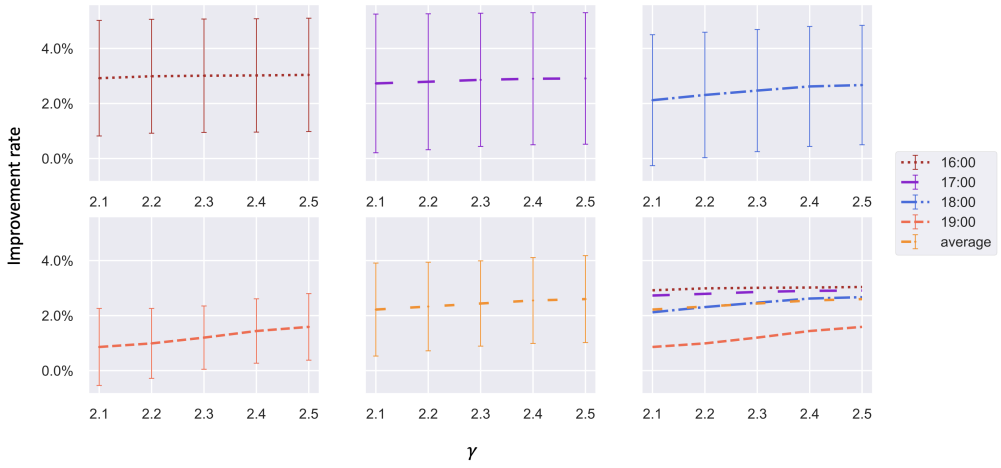


Figure 17: Improvement rates on different γ in Midtown (error bar represents the standard deviation).

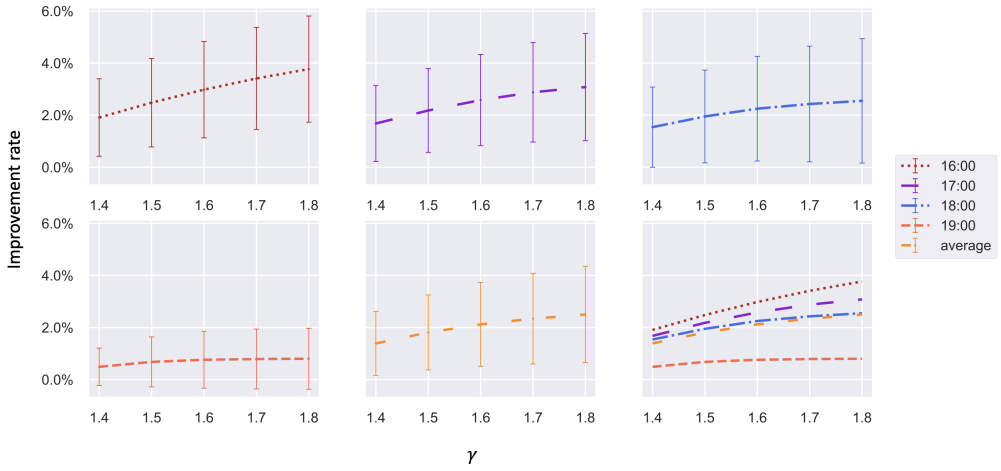


Figure 18: Improvement rates on different γ in Central Park (error bar represents the standard deviation).

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.