A Structural Equations Approach for Modeling the Endogeneity of Lane-mean Speeds

Qinglong Lu [*]

Department of Civil, Geo and Environmental Engineering,
Technical University of Munich, Germany

September 25, 2019

Abstract

Like other transportation data, lane-mean speeds are also best modeled by a system of structural equations. Several studies omit the interrelation between adjacent lane speeds, which may produce biased and inconsistent results if models are solved by ordinary least squares (OLS). The uncorrelatedness of regressors and disturbances assumption of OLS is violated since one or more independent variables are endogenous in the system. This study attempts to propose a structural equations approach to model the lane-mean speeds in multi-lane traffic with the endogeneity of adjacent lane speeds and also the downstream speeds being considered. Additionally, the equations system can serve as a prediction model for lane-mean speeds.

Several empirical analyses where data are collected from different multi-lane segments with different lengths and numbers of lanes are conducted in order to observe the performance of the equations system in different setups. The study further compares the prediction accuracy between the underlying approach and the model established by Shankar and Mannering (1998) for assessing the impact of introducing downstream speeds within the model. The findings show that more precise results are obtained generally after downstream speeds are included, emphasizing the improvements and superiority of this approach.

1 Introduction

The relationship between three fundamental variables in traffic flow theory (i.e. flow, speed and density) (Greendshield [1]) has been heavily researched since it plays an important role in demonstrating the underlying phenomenon of traffic flow which is essential in the field of traffic management and control. As one of these three variables in fundamental diagrams, traffic flow speed was also studied by many research studies. The generalized average speed which is known as time-space-mean speed (TSMS) was established by Edie [2], leading to a research interest in the average speed of vehicles. Due to the restriction of loop detectors, i.e. it reports time-mean speed (TMS) only, estimation methods for the generalized average speed were proposed by many authors. TSMS was estimated by a convex combination of the preset upper bounds and lower bounds in A. Jamshidnejad and De Shutter [3]. In order to eliminate a previously ignored potential error caused by the vehicles that are still on the given road segment at the end of a cycle, A. Jamshidnejad and De Shutter [4] introduced an iterative procedure for estimating TSMS by utilizing microscopic traffic data (i.e. individual speeds and headways of the vehicles). In addition, after the relationship between TMS and space-mean speed (SMS) was derived in Wardrop [5], estimation algorithms for TMS and SMS can also be found in other studies. For filling the gap in the literature, Rakha and Zhang [6] utilized the variance of TMS for the estimation of SMS instead

[*] E-mail: qinglong.lu@tum.de

of estimating TMS from SMS as previously done by Garber and Hoel [7] and developed a relationship between SMS variance and TMS variance as well as between SMS and travel time reliability.

With more and more data becoming available in transportation in the past years, many novel data-driven models for speed prediction have been presented. Vanajakshi and Rilett [8] did a comparison of the traffic speed prediction accuracy between support vector machines (SVM) and artificial neural networks (ANN), based on dual loop detector data collected from the freeways of San Antonio, Texas. The study showed that support vector regression (SVR) performed better when the training data were poor in quality and quantity. In De Fabritiis et al. [9], two algorithms based on ANN and Pattern-Matching respectively were designed to implement short-term speeds predictions by using real-time floating car data. A long short-term neural network (LSTM NN) (Ma et al. [10]) was applied to capture the non-linear traffic dynamic effectively and exhibited the superiority in short-term travel speed prediction with long-term temporal dependency using microwave detector data. However, when applying time series theory, neural networks and genetic algorithms so as to forecast short-term speeds, data with regular time intervals are necessary. Therefore, in order to address this issue, Ye, Szeto and Wong [11] extend these methods to adapt the data recorded at irregular intervals equipped with acceleration information at the same time.

Apart from the aforementioned studies which dealt with the given road as a whole, some studies were focused on constructing the model on the lane level that may produce more reliable results. Laval and Daganzo [12] proposed a multi-lane hybrid traffic flow model that combined a multi-lane kinematic wave module with a detailed constrained-motion model to explore the underlying lane-changing maneuvers. The cell transmission model (CTM) was also extended to simulate multi-lane traffic in more detail by Carey, Balijepalli and Watling [13].

Lane-mean speeds modeling is very important in the field of multi-lane traffic research. Some transportation data are best modeled by a system of structural equations, including lane-mean speeds (Washington et al. [14]). In Shankar and Mannering [15], a structural equations approach was used to model the lane-mean speeds and lane-speed deviations where three-stage least squares (3SLS) was applied to handle the endogeneity problem, i.e. the correlation between regressors and disturbances that means changing the value of dependent variables may lead to the changing of independent variables. The researchers conducted an empirical study based on this approach, combining the relevant data of a six-lane road (three lanes in each direction) with environmental data, temporal data and traffic flow factors. However, although the methodology solved the endogeneity problem, obtaining a best linear unbiased estimator (BLUE), its predicting accuracy was not assessed, the authors focusing on the interpretation of significant variables instead. In addition, Cheng et al. [16] proposed hierarchical models based on a Bayesian framework in order to address the same problem. However, in their empirical study, the model did not take into account the effect of trucks in traffic flow, which may result in unrealistic and unreliable results. Most importantly, neither of the two studies mentioned above accounted for the impact of downstream speed on upstream speed.

In this paper, an improved structural equations system is proposed for lane-mean speed prediction. In relation to previous literature, the underlying methodology solves the endogeneity problem by utilizing a 3SLS model. One of the main contributions of the proposed system is the fact that it considers the effect of downstream speed on upstream speed. The results of this approach are compared to those obtained by Shankar and Mannering [15]. Several tests are conducted in order to assess the feasibility and reliability of the approach.

The outline of the paper is as follows: Section 2 gives an interpretation of the model construction. The empirical analysis is provided in Section 3. Lastly, conclusions and future research questions are highlighted in Section 4.

2 Methodology

The methodology used in this study consists of several steps. First, a road segment is divided into shorter cells (see Figure 1). Data on important factors (i.e. traffic data, environmental data and temporal data,

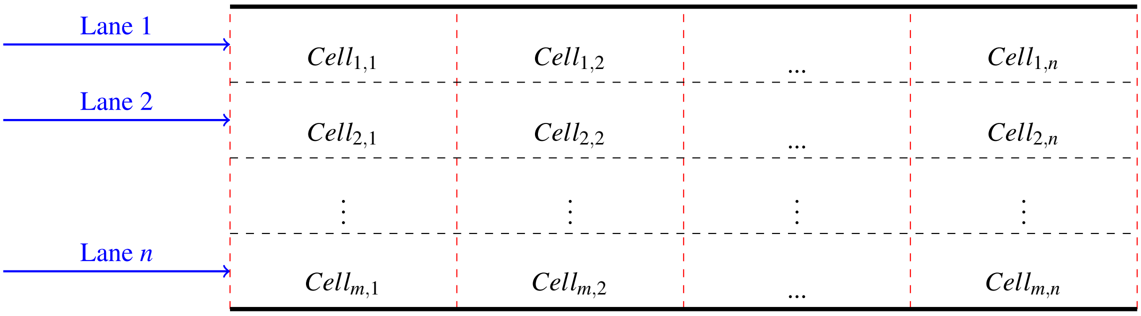

Figure 1: Cells structure of a road segment

etc.) that will serve as inputs in the estimation are collected in each cell for constructing an integrated and reliable model. Finally, the structural equations system will be employed on each cell to estimate the average vehicles speed in corresponding space.

As previously described in Shankar and Mannering [15], when using a structural equations system, a correlation between adjacent lane speeds was observed. This means that the lane-mean speed of a cell is determined not only by its own basic determinants (i.e. traffic data, environmental data and temporal data, etc.), but also by the speed of vehicles in adjacent lanes. This is because the adjacent lanes belonging to a particular segment are not isolated in space and the vehicle speeds in such lanes are interdependent (Cheng et al. [16]). As cells in the same segment are considered in a same structural equations system and their lane-mean speeds are estimated simultaneously, the lane-mean speed of a cell, which is the dependent variable in one equation, may be a regressor in other equations. Those lane-mean speeds that act as dependent variables and regressors at the same time are endogenous variables of the system.

Furthermore, it has been proven that upstream traffic flow is correlated with the downstream traffic condition (Lighthill and Whitham [17]). Thus, it is necessary to consider the impact of downstream speed in the same lane when estimating the lane-mean speed of a cell. In this study, the lane-mean speed of a cell is one of the exogenous variables in the equation with lane-mean speed in its backward cell as the dependent variable. For instance, based on Figure 1, the lane-mean speed of $Cell_{1,2}$ is an exogenous independent variable when estimating the lane-mean speed of $Cell_{1,1}$.

Hence, for lane-mean speeds, the system of equations can be written as follows (focused on segment j),

$$\begin{cases} u_{1,j} = \alpha_{1,j} + \boldsymbol{\beta}_{1,j} \cdot \boldsymbol{X}_{1,j} + \boldsymbol{\lambda}_{1,j} \cdot \boldsymbol{Z}_{1,j} + \boldsymbol{\theta}_{1,j} \cdot \boldsymbol{v}_{1,j} + \eta_{1,j} \cdot u_{1,j+1} + \varepsilon_{1,j} \\ u_{2,j} = \alpha_{2,j} + \boldsymbol{\beta}_{2,j} \cdot \boldsymbol{X}_{2,j} + \boldsymbol{\lambda}_{2,j} \cdot \boldsymbol{Z}_{2,j} + \boldsymbol{\theta}_{2,j} \cdot \boldsymbol{v}_{2,j} + \eta_{2,j} \cdot u_{2,j+1} + \varepsilon_{2,j} \\ \vdots \\ u_{i,j} = \alpha_{i,j} + \boldsymbol{\beta}_{i,j} \cdot \boldsymbol{X}_{i,j} + \boldsymbol{\lambda}_{i,j} \cdot \boldsymbol{Z}_{i,j} + \boldsymbol{\theta}_{i,j} \cdot \boldsymbol{v}_{i,j} + \eta_{i,j} \cdot u_{i,j+1} + \varepsilon_{i,j} \\ \vdots \\ u_{m,j} = \alpha_{m,j} + \boldsymbol{\beta}_{m,j} \cdot \boldsymbol{X}_{m,j} + \boldsymbol{\lambda}_{m,j} \cdot \boldsymbol{Z}_{m,j} + \boldsymbol{\theta}_{m,j} \cdot \boldsymbol{v}_{m,j} + \eta_{m,j} \cdot u_{m,j+1} + \varepsilon_{m,j} \end{cases} \tag{1}$$

where i is the lane number, j is the segment number.$u_{i,j}$ is the lane-mean speed.$\vec{X_{i,j}}$ is the vector of exogenous variables that affect the speed of $Cell_{i,j}$ except for the average vehicles speed of next segment in the same lane.$\vec{Z_{i,j}}$ is the vector of endogenous variables that varies as the dependent variable $u_{i,j}$ varies except for the mean speeds of adjacent lanes.$v_{i,j}$ is the mean speed in the crucial lane adjacent to lane i in the same segment. $\alpha_{i,j}, \vec{\beta_{i,j}}, \vec{\lambda_{i,j}}, \theta_{i,j}, \eta_{i,j}$ are estimable coefficients. $\varepsilon_{i,j}$ is disturbance term.

The crucial lane adjacent to lane i, is defined as the lane that has a slower average speed (i.e. more congested) between the two lanes adjacent to lane i. And $v_{i,j}$ is the average vehicles speed in this lane,

$$v_{i,j} = \begin{cases} \min(u_{i-1,j}, u_{i+1,j}) & \text{if } i \neq 1 \wedge i \neq m \\ u_{2,j} & \text{if } i = 1 \\ u_{m-1,j} & \text{if } i = m \end{cases} \tag{2}$$

In order to obtain unbiased and consistent results in estimating equations (1) which contain endogenous independent variables, 3SLS is appropriate. The executing procedure is as follow,

Stage 1: Getting two-stage least squares (2SLS) estimates of the model system;

Stage 2: Applying the estimates from Stage 1 to calculate the disturbances of the system;

Stage 3: Generalized least squares (GLS) is used to compute coefficients.

Refer to Zellner and Theil [18] for more detail about 3SLS.

$$\hat{\boldsymbol{Z}}_{i,j} = \hat{\boldsymbol{\beta}}_{i,j}^{(\boldsymbol{Z})} \boldsymbol{X}_{i,j} + \hat{\boldsymbol{\alpha}}_{i,j}^{(\boldsymbol{Z})}$$

$$\hat{\boldsymbol{v}}_{i,j} = \hat{\boldsymbol{\beta}}_{i,j}^{(\boldsymbol{v})} \boldsymbol{X}_{i,j} + \hat{\boldsymbol{\alpha}}_{i,j}^{(\boldsymbol{v})}$$

$$u_{i,j} = \alpha_{i,j} + \boldsymbol{\beta}_{i,j} \cdot \boldsymbol{X}_{i,j} + \boldsymbol{\lambda}_{i,j} \cdot \hat{\boldsymbol{Z}}_{i,j} + \boldsymbol{\theta}_{i,j} \cdot \hat{\boldsymbol{v}}_{i,j} + \eta_{i,j} \cdot u_{i,j+1} + \varepsilon_{i,j}$$

$$\hat{u}_{i,j}^{(2SLS)} = \hat{\alpha}_{i,j}^{(2SLS)} + \hat{\boldsymbol{\beta}}_{i,j}^{(2SLS)} \cdot \boldsymbol{X}_{i,j} + \hat{\boldsymbol{\lambda}}_{i,j}^{(2SLS)} \cdot \hat{\boldsymbol{Z}}_{i,j} + \hat{\boldsymbol{\theta}}_{i,j}^{(2SLS)} \cdot \hat{\boldsymbol{v}}_{i,j} + \hat{\eta}_{i,j}^{(2SLS)} \cdot u_{i,j+1}$$

$$\varepsilon_{i,j} = u_{i,j} - \hat{u}_{i,j}^{(2SLS)}$$

$$E(\boldsymbol{\varepsilon}_{i,j} \boldsymbol{\varepsilon}_{i,j}^{T}) = \boldsymbol{\Omega}$$

$$\boldsymbol{W}_{i,j} = (1; \boldsymbol{X}_{i,j}; \hat{\boldsymbol{Z}}_{i,j}; \hat{\boldsymbol{v}}_{i,j}; u_{i,j+1})$$

$$\boldsymbol{B}_{i,j} = (\alpha_{i,j}; \boldsymbol{\beta}_{i,j}; \boldsymbol{\lambda}_{i,j}; \boldsymbol{\theta}_{i,j}; \eta_{i,j})$$

$$\hat{\boldsymbol{B}}_{i,j}^{(3SLS)} = (\boldsymbol{W}_{i,j}^{T} \boldsymbol{\Omega}^{-1} \boldsymbol{W}_{i,j})^{-1} \boldsymbol{W}_{i,j}^{T} \boldsymbol{\Omega}^{-1} \boldsymbol{u}_{i,j}$$

$$\hat{\boldsymbol{u}}_{i,j}^{(3SLS)} = \hat{\boldsymbol{B}}_{i,j}^{(3SLS)T} \boldsymbol{W}_{i,j}$$

3 Empirical Studies

In this section, we will analyze and assess the improved structural equations system by four scenarios with empirical traffic data:

- The first scenario is to illustrate the improvement and superiority of the presented structural equations system by comparing its predicting accuracy with the approach proposed in Shankar and Mannering [15].
- The second scenario is to compare the predicting accuracy of mean speeds of different lanes that belong to same segments in order to analyze if it is reasonable to design a crucial adjacent lane for the lanes caught in the middle and also to assess the model's results in different lanes.
- The third scenario is to reveal in what transverse condition (number of lanes) can the system receive the best forecasting result, while the longitudinal condition (length of segment) is controlled for.
- The fourth scenario is the reverse of the third scenario: in what longitudinal condition can the system perform best with the transverse condition being controlled for.

3.1 Data description and variables definition

All data used in this study come from Caltrans Performance Measurement System (PeMS). In order to validate the reliability of the improved structural equations system in different conditions, several segments that have different lengths and different number of lanes are studied. As in the methodology explained above, for modeling lane-mean speeds more precise, mean speeds of downstream are also considered in the system. So, data of two consecutive stations are utilized in each empirical study. More details about the selected segments and stations are listed in Table 1. Figure 2 depicts the simplified drawing of the segments for a better explanation.

Table 1: Information of selected segments and stations

Length of Segment	Number of Lanes	Upstream Station ID	Downstream Station ID	Freeway ID
0.1 *miles* ($\pm$ 0.05 *miles*)	3	402513	414025	SR4-E
0.1 *miles* ($\pm$ 0.05 *miles*)	5	404921	404886	I80-E
0.2 *miles* ($\pm$ 0.05 *miles*)	2	1216194	1211560	SR241-N
0.2 *miles* ($\pm$ 0.05 *miles*)	3	1216161	1211387	SR241-N
0.2 *miles* ($\pm$ 0.05 *miles*)	4	401464	401489	I880-N
0.2 *miles* ($\pm$ 0.05 *miles*)	5	401079	415253	I80-E
0.3 *miles* ($\pm$ 0.05 *miles*)	3	601256	601530	SR41-N
0.3 *miles* ($\pm$ 0.05 *miles*)	5	405589	400679	I80-E
0.4 *miles* ($\pm$ 0.05 *miles*)	2	500013091	500013101	SR1-N
0.4 *miles* ($\pm$ 0.05 *miles*)	3	1111561	1122645	I8-E
0.4 *miles* ($\pm$ 0.05 *miles*)	4	718492	717977	I710-N
0.4 *miles* ($\pm$ 0.05 *miles*)	5	406056	407233	SR4-E

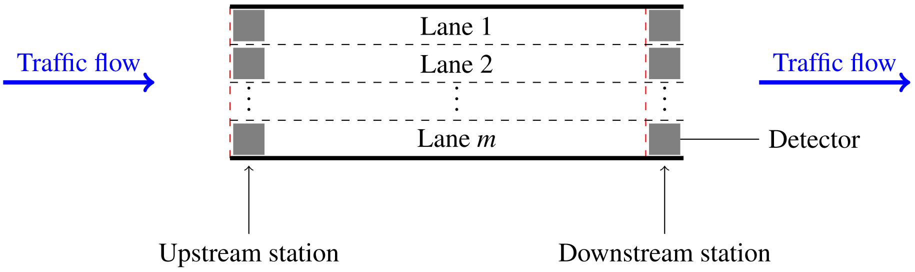

Figure 2: A segment and its attached information

Apart from the segment $\{0.4\ miles, 2\ lanes\}$ which lacks the data of Jan, 2017 and the segment $\{0.4\ miles, 5\ lanes\}$ losing the data of Jan, Feb, Dec, 2017, the traffic data from 8^{th} to 14^{th} in every month in 2017 of every concerned station and the speed data of the following station that are needed for the model are exported from the dataset of PeMS with the granularity of 5 minutes. But it should be clarified that although the data is not complete in segments $\{0.4\ miles, 2\ lanes\}$ and $\{0.4\ miles, 5\ lanes\}$, it should have no impact on the results of the coming analysis. Another point to be noted is that in Shankar and Mannering [15], the mean speeds are aggregated in a time window of an hour which may reduce the practicability of the developed approach. As a result, the observations in the tests of this paper are collected every 5 minutes. As shown in Table 1, lengths of segments are not imputed in entirely accurate numbers for the reason of easier and better indication in writing. Similarly, this will not affect the conclusion.

In the preliminary step, in order to eliminate the error caused by the data-collection equipment (detectors), Box Plots are used for deleting the possible outliers embed on the speed column. That is, the observations with the speed greater than $upper\ quartile + 1.5 \cdot IQR$ or smaller than $lower\ quartile - 1.5 \cdot IQR$ (IQR: Interquartile range) will be dropped from the dataset. All variables considered in the system are extracted from the raw dataset prepared for the ultimate estimation. Previous studies investigating the same question (Shankar and Mannering [15], Cheng et al. [16]) include geometric factors in their models, which ultimately did not present great significance in affecting the lane-mean speed. Hence, this study only concentrates on the characters of traffic factors and temporal data. The variables considered in the empirical analysis are listed in Table 2. Different from the empirical study of Shankar and Mannering [15], which focuses on interpreting the significance of each variable, the tests in this paper aim to illustrate the reliability of the structural equations system on lane-mean speed estimation in different situations. Therefore, for a better assessment of its superiority and potential shortcomings, additional tests of the performance of the model are necessary.

Table 2: Variables considered for the tests

Factor	Variables	Remark
Traffic data	Mean speed of lane	
	Downstream speed	The mean speed collected by next station
	Flow of the lane	
	Low flow indicator	Equal to 1 if flow of lane $< 75veh/h$
	Ratio of flow	Ratio of flows in current lane to crucial adjacent lane
	Truck percentage	
	Truck percentage indicator 1	Equal to 1 if truck percentage $> 60\%$ and flow $< 50veh/h$
	Truck percentage indicator 2	Equal to 1 if truck percentage $\leq 60\%$ and flow $> 200veh/h$
	Truck speed	Equal to $truck\ VMT/truck\ VHT$ [a]
	High truck flow	Equal to 1 if truck flow $> 100veh/h$
Time of year (dummy)	Spring	
	Summer	
	Autumn	
	Winter (reference)	
Time of week (dummy)	Monday	
	Tuesday	
	Wednesday	
	Thursday	
	Friday	
	Saturday	
	Sunday (reference)	
Time of day (dummy)	Early morning	From 0:00 to 6:00
	AM peak	From 7:00 to 8:00
	PM peak	From 17:00 to 19:00
	Nighttime	From 19:00 to 24:00
	Other time (reference)	

[a] truck VMT is truck Vehicle Miles Traveled and truck VHT is truck Vehicle Hours Traveled, both of which are available in PeMS.

3.2 Speed prediction accuracy

To assess the precision of estimation, mean absolute error (MAE) is introduced. It calculates the deviations between the predicted value and the corresponding observed value, i.e. fitted mean speed and observed mean speed. This study does not concern about the proportion of the speed difference in observed speed, so MAE is the only criterion for speed prediction accuracy assessment. The MAE is calculated as follows:

$$MAE = \frac{1}{n}\sum_{k=1}^{n} |u_{k,pre} - u_{k,obs}| \tag{3}$$

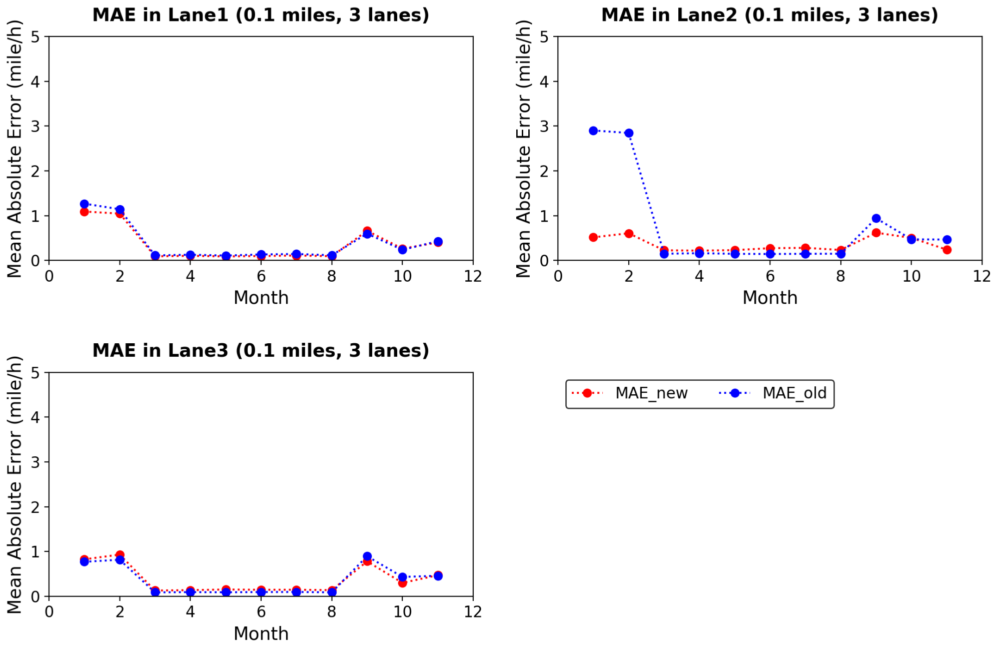

Figure 3: MAE curves of the new system and the old system in segment $\{0.1\ miles, 3\ lanes\}$

where k is the index of observation, n is the total number of observations, $u_{k,pre}$ and $u_{k,obs}$ are the predicted value of mean speed and observed mean speed of observation k respectively.

As a basic residual analysis criterion, MAE can easily reflect the effectiveness of estimation models. A relatively smaller value of MAE means the average difference between the predicted speed and observed speed of each observation is smaller, i.e. the model receives more accurate speed prediction.

3.3 Accuracy comparison between the improved and the old structural equations system

To allow the comparison to be more systematic and complete, all segments listed in Table 1 are deployed on the improved structural equations system (referred to as the new system hereafter) and the structural equations system established in Shankar and Mannering [15] (referred to as the old system hereafter). Furthermore, in the tests, every month data is utilized as test data successively while the data of other months are used to train the model, i.e. leave-one-out cross validation (LOOCV). As a method of model validation, LOOCV can mitigate the random effects in the data that may influence model analysis.

The test results show that except for the segment $\{0.1\ miles, 3\ lanes\}$, in other segments researched in this paper, the new system performs better than the old system, namely it gets more beneficial speed prediction accuracy. And even though in segment $\{0.1\ miles, 3\ lanes\}$ where the old system gets a preferred result, actually the MAE of speed in both systems are very close, which can be seen in Figure 3. When the tested data are collected in winter, the new system even shows better performance. So, from the result, we can draw a conclusion that the improved structural equations system has superiority in speed prediction compared to the previous system overall. What should be noticed is that the main difference between these two systems is whether downstream speeds are considered when modeling the lane-mean speeds. Obviously, downstream speed plays a positive role in modeling lane-mean speed as an important independent variable in the model. From the forecast results of segments $\{0.2\ miles, 2\ lanes\}$, $\{0.2\ miles, 3\ lanes\}$ $\{0.2\ miles, 4\ lanes\}$, $\{0.2\ miles, 5\ lanes\}$ provided in Figure 4, Figure 5, Figure 6 and Figure 7 respectively, it can be observed with more lanes in the segment (length is controlled) both systems are relatively less precise in general.

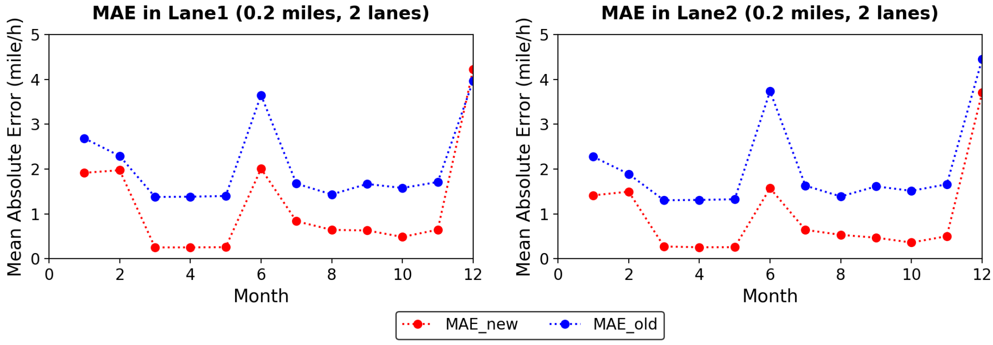

Figure 4: MAE curves of the new system and the old system in segment $\{0.2\ miles, 2\ lanes\}$

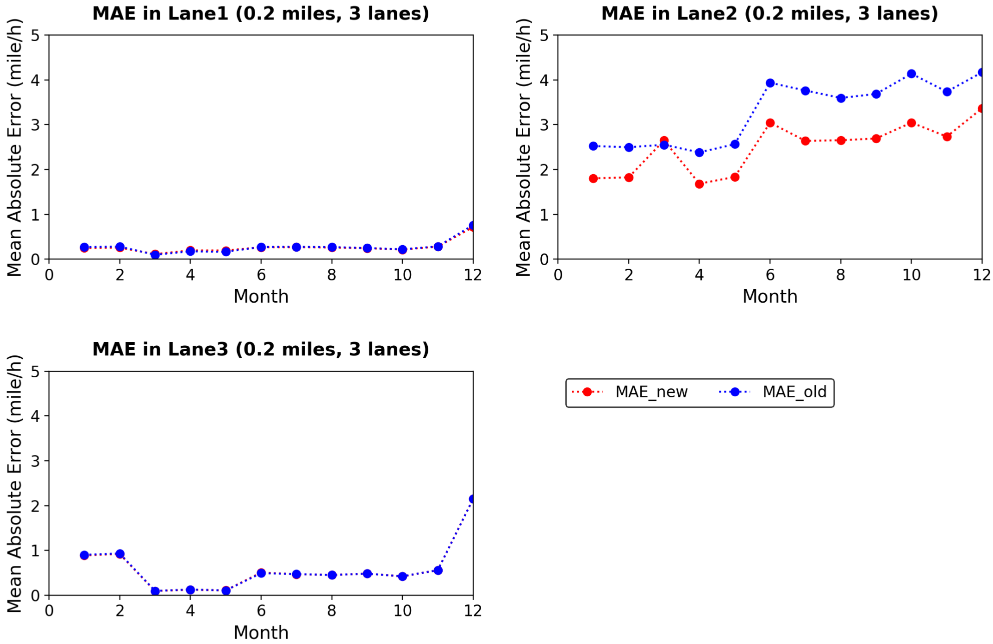

Figure 5: MAE curves of the new system and the old system in segment $\{0.2\ miles, 3\ lanes\}$

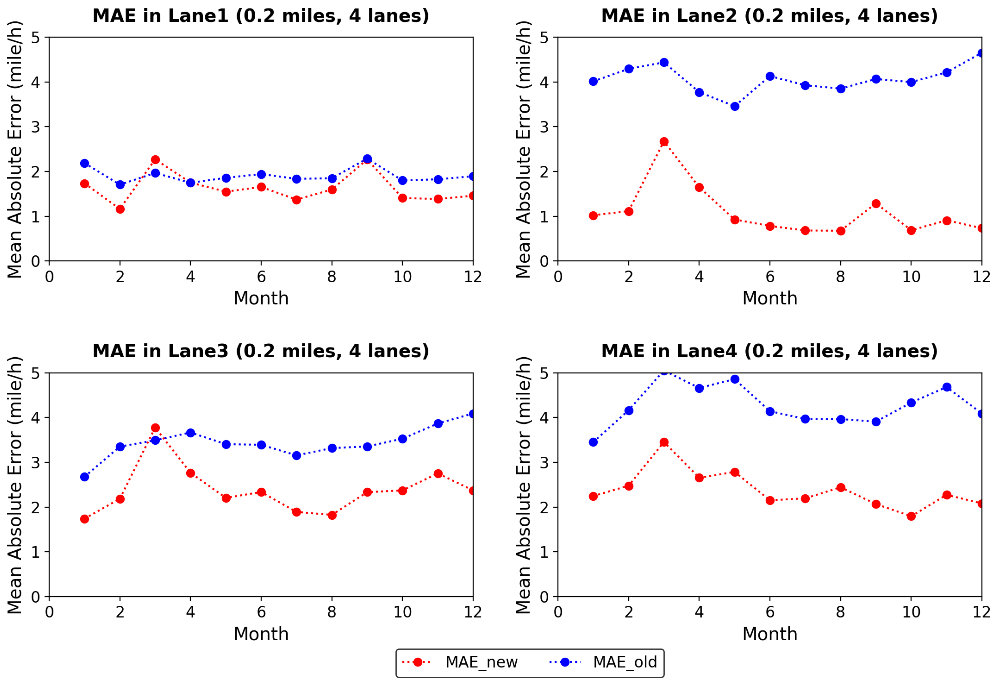

Figure 6: MAE curves of the new system and the old system in segment $\{0.2\ miles, 4\ lanes\}$

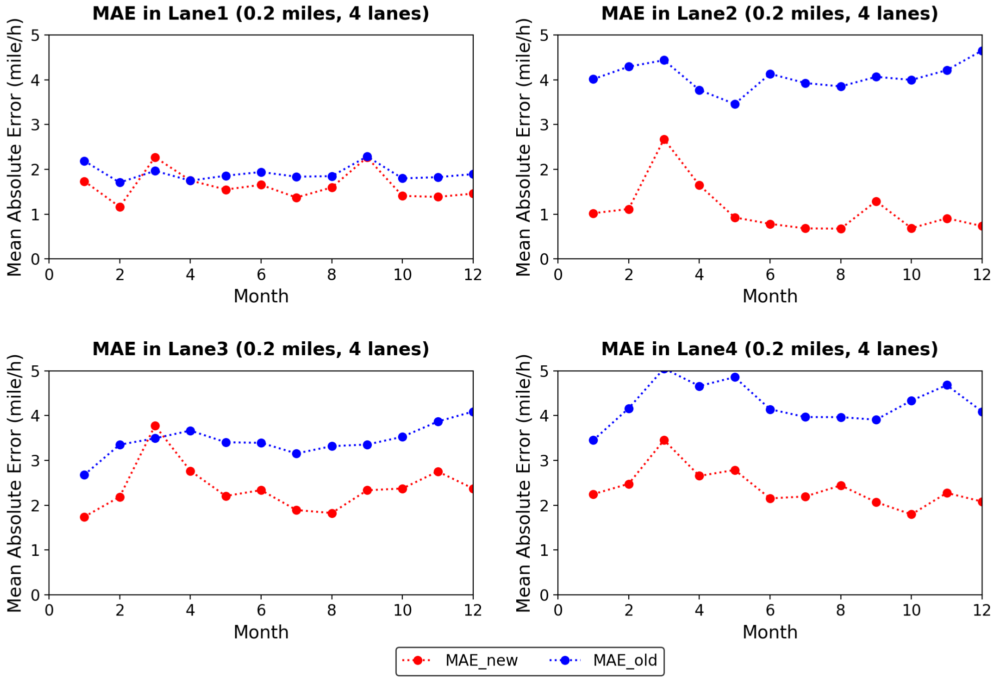

Figure 7: MAE curves of the new system and the old system in segment $\{0.2\ miles, 5\ lanes\}$

3.4 Accuracy comparison between different lanes in same segments

Different from the old system in which mean speeds in all adjacent lanes are included as endogenous variables in an equation, a creatively defined crucial adjacent lane is introduced in the improved system to indicate the dominant adjacent lane for middle lanes and reduce computational demand as well. Regarding this process, the old system is more persuasive and thoughtful. However, in some cases, drivers only take care of the traffic situation of the lane concerned by them, e.g. the left lane is noticed when drivers want to overtake the vehicle forward, nevertheless, when drivers want to change to the off-ramp, the right lane is critical. In this study, the adjacent lane which has a lower average vehicle speed in the time window (5 minutes) is defined as the crucial adjacent lane and only its mean speed is contained by the model. Note that the crucial adjacent lane may change in different time steps depending on the lane-mean speeds of adjacent lanes.

The result shows that only in the segment $\{0.2\ \mathit{miles}, 3\ \mathit{lanes}\}$ and $\{0.4\ \mathit{miles}, 4\ \mathit{lanes}\}$ the errors of middle lanes are bigger than the errors of side lanes as shown in Figure 8. In other segments, either the errors of different lanes are at same the level (for instance, as shown in Figure 9) or are uncorrelated with the lane positions (examples are in Figure 10). There are no particular underlying correlations between errors and lane positions. Additionally, the accuracy of prediction may relate to the data quality and quantity. So, the introduction of the crucial adjacent lane does not have an obvious negative effect on the lane-mean speed prediction and hence it is reasonable. Furthermore, we also notice that in the two-lane segments (i.e. $\{0.2\ \mathit{miles}, 2\ \mathit{lanes}\}$ and $\{0.4\ \mathit{miles}, 2\ \mathit{lanes}\}$) MAE curves of different lanes are very close to each other (shown in Figure 11), which means the new system is stable in speed prediction in two-lane segments.

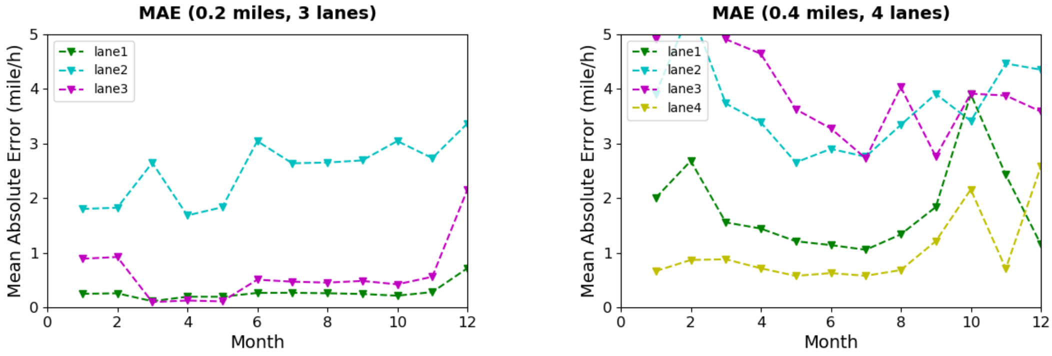

Figure 8: Segments where errors of middle lanes bigger than side lanes

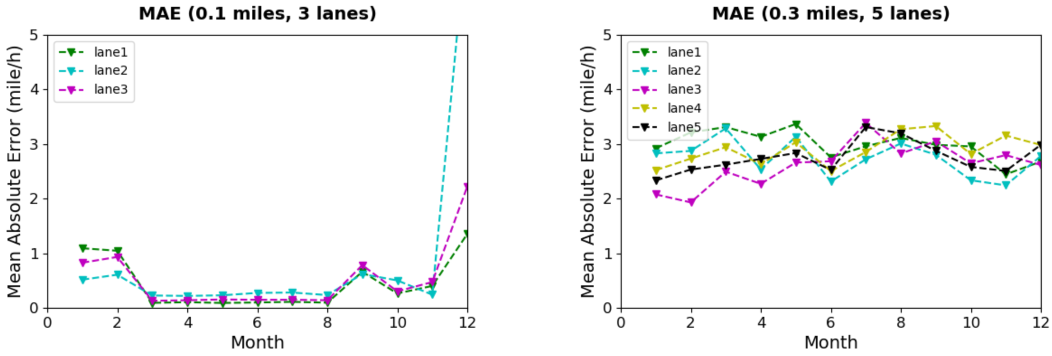

Figure 9: Segments where errors of different lanes are at the same level

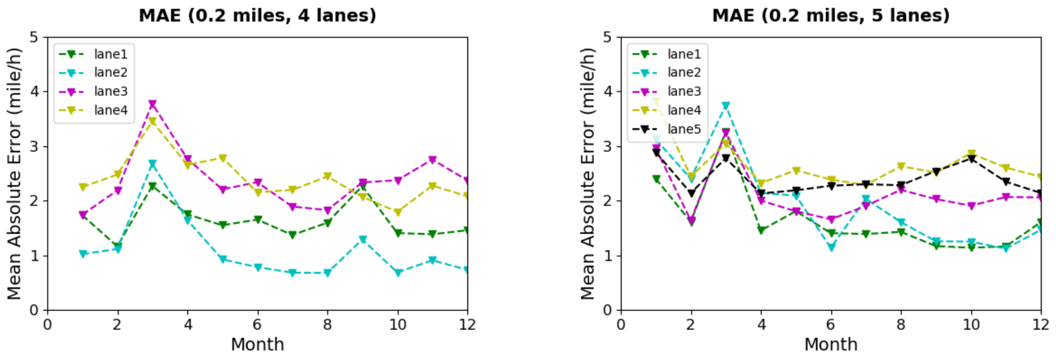

Figure 10: Segments where errors of different lanes are uncorrelated with lanes positions

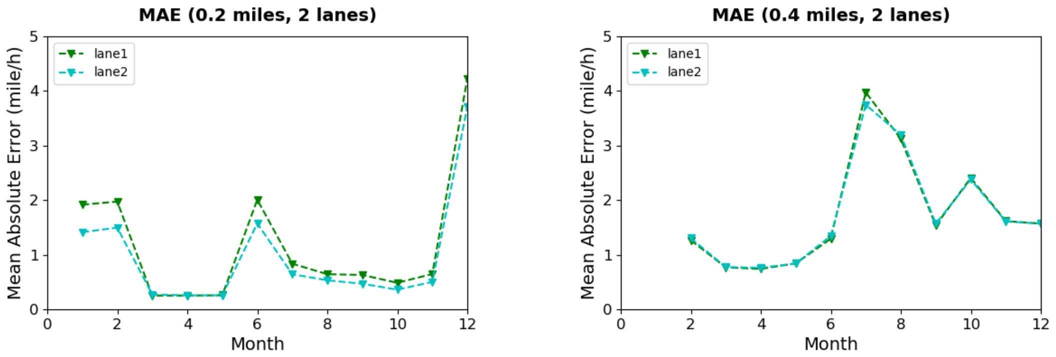

Figure 11: MAE curves of different lanes in two-lane segments

3.5 Accuracy comparison among different number-of-lane segments (length is controlled)

In order to uncover in what latitudinal condition (number of lanes) can the improved system perform best, two segment series are studied in this subsection, and in each of them, all segments share the same length. For easier reference in the subsequent text, they will be named as 0.2-mile series and 0.4-mile series respectively. Figure 12 depicts the MAE curves of all segments in 0.2-mile series and 0.4-mile series (zooming into a range of MAE less than 5). As it can be seen in the figures, overall, segments with fewer lanes receive better results. The segments with 4 lanes and 5 lanes are having higher MAE curves than the segments with 2 and 3 lanes. But what also attracts attention is that the curves of 2-lane segments are fluctuating more than those of 3-lane segments (i.e. predictions accuracy are more stable in 3-lane segments) and 3-lane segments get better results sometimes. Thus, it is clear that the improved structural equations system is more reliable in 3-lane segments.

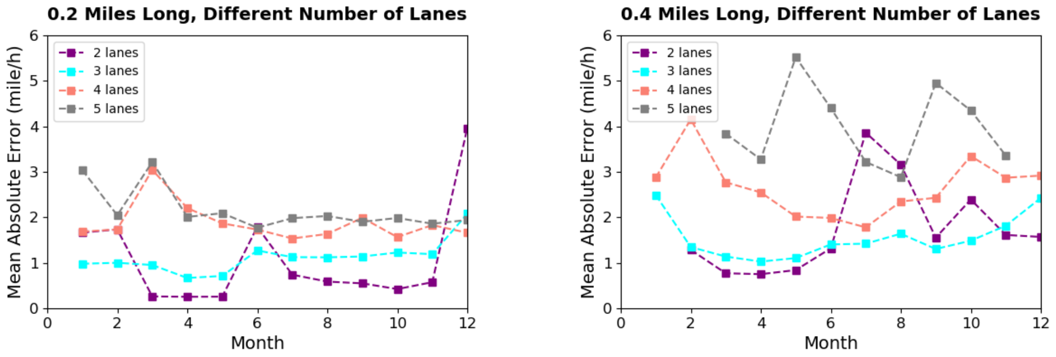

Figure 12: 0.2-mile series and 0.4-mile series

3.6 Accuracy comparison among different length segments (number of lanes is controlled)

Similarly, for exploring in what longitudinal condition (length) can the system predict lane-mean speed most precise, two segments series are studied in this subsection, and in each of them, all segments have the same number of lanes. They will be indicated as 3-lane series and 5-lane series in the following text. The MAE curves of these segments are plotted in Figure 13. Segment $\{0.1\ miles, 3\ lanes\}$ has the lowest MAE curve in 3-lane series, while $\{0.1\ miles, 5\ lanes\}$ has the highest one in 5-lane series. And in $\{0.1\ miles, 3\ lanes\}$ when the data of December act as the test dataset, it reaches an unusual point compared with other months. So, combining results from both plots shown in Figure 13, it can be concluded that the system is more reliable in 0.2-mile segments.

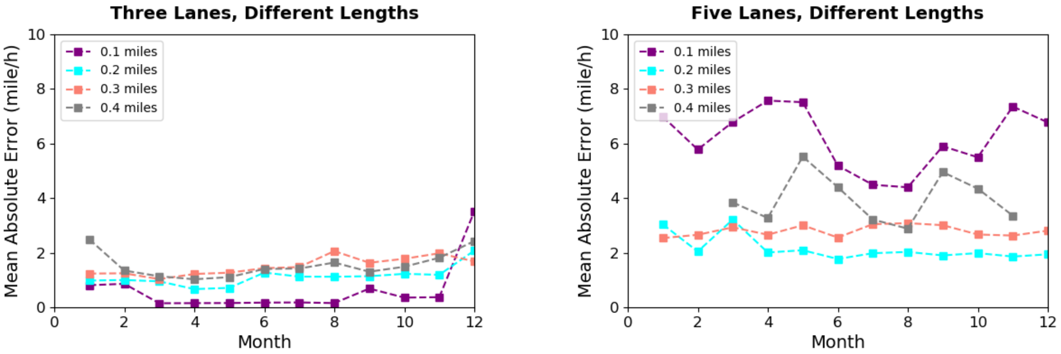

Figure 13: 3-lane series and 5-lane series

4 Conclusions

In this paper, an improved structural equations system is proposed for predicting the lane-mean speed of several segments which have different spatial parameters, while addressing the endogeneity issue at the same time. It is an improved system based on the approach established by Shankar and Mannering [15]. In the improved system, in order to measure the impact of downstream speeds when modeling lane-mean speeds, downstream mean speed is included as an exogenous variable in every equation. In addition, unlike the old system in which both adjacent lanes for the middle lanes are considered in equations, we prefer to find out which adjacent lane affect the current lane most. Regarding this aspect, a crucial adjacent lane is introduced and is defined as the lane which has a smaller average vehicle speed in two adjacent lanes for middle lanes. In the model, only the mean speed of the crucial adjacent lane is considered and it is an endogenous variable. Similar to the old system, 3SLS is used to solve the problem caused by endogenous variables (violate a key assumption of OLS) and for achieving BLUE.

For evaluation and assessment purposes, several empirical tests are conducted. The results of four scenarios are analyzed, which lead to the following conclusions. First of all, by comparing the prediction accuracy of the improved system with the old one, it is shown that the improved system basically generates smaller errors in all segments tested in this study. Hence, it can be concluded that the improved system has superiority in forecasting lane-mean speed compared to the old system. This means that the introduction of downstream speeds can improve the accuracy of lane-mean speed prediction. Secondly, by comparing the MAE of different lanes in same segments, it is shown that there is no underlying correlation between the prediction accuracy and the lane positions and, therefore, defining a crucial adjacent lane for each lane will not decrease the model precision. The comparison between segments that have a different number of lanes but are of the same length suggests that the system is more reliable in 3-lane segments. The last scenario is a reverse version of the third one, in which we discuss in which longitudinal condition (length) can the system perform best. The analysis finds that in 0.2-mile segments the improved system can receive more stable and accurate results.

All in all, compared with the previous approach for simultaneously estimating the lane-mean speeds and addressing the endogeneity issue, the system developed in this study shows better performance, reliability and practicality. However, the improved system still cannot integrate all factors that may affect the estimation, e.g. environment factors, etc. The lane-changing phenomenon in traffic stream in each cell and the effects of on-ramp stream and off-ramp stream on the speed of vehicles running on the main road are not considered in the presented model that should be improved in future works. What's more, the relationship between inputs and the output is not analyzed enough in the paper. The significance of each variable is not shown because the emphasis of this paper is to prove the superiority of the improved system and explore its best applying situations. But it may make it difficult to uncover the underlying correlation between the factors and mean speed and also be unconvinced by readers. Another point needed to be concerned is more studies for researching traffic safety and accidents based on lane-mean speeds should be carried out.

References

[1] B. Greenshields, W. Channing, H. Miller, *et al.*, "A study of traffic capacity," in *Highway research board proceedings*, National Research Council (USA), Highway Research Board, vol. 1935, 1935.

[2] L. C. Edie, *Discussion of traffic stream measurements and definitions*. Port of New York Authority, 1963.

[3] A. Jamshidnejad and B. De Schutter, "An iterative procedure for estimating the generalized average speed using microscopic point measurements," in *2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, IEEE, 2015, pp. 38–44.

[4] A. Jamshidnejad and B. De Schutter, "Estimation of the generalised average traffic speed based on microscopic measurements," *Transportmetrica A: transport science*, vol. 11, no. 6, pp. 525–546, 2015.

[5] J. G. Wardrop, "Road paper. some theoretical aspects of road traffic research.," *Proceedings of the institution of civil engineers*, vol. 1, no. 3, pp. 325–362, 1952.

[6] H. Rakha and W. Zhang, "Estimating traffic stream space mean speed and reliability from dual-and single-loop detectors," *Transportation Research Record*, vol. 1925, no. 1, pp. 38–47, 2005.

[7] N. J. Garber and L. A. Hoel, *Traffic and highway engineering*. Cengage Learning, 2014.

[8] L. Vanajakshi and L. R. Rilett, "A comparison of the performance of artificial neural networks and support vector machines for the prediction of traffic speed," in *IEEE Intelligent Vehicles Symposium, 2004*, IEEE, 2004, pp. 194–199.

[9] C. De Fabritiis, R. Ragona, and G. Valenti, "Traffic estimation and prediction based on real time floating car data," in *2008 11th International IEEE Conference on Intelligent Transportation Systems*, IEEE, 2008, pp. 197–203.

[10] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data," *Transportation Research Part C: Emerging Technologies*, vol. 54, pp. 187–197, 2015.

[11] Q. Ye, W. Y. Szeto, and S. C. Wong, "Short-term traffic speed forecasting based on data recorded at irregular intervals," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 4, pp. 1727–1737, 2012.

[12] J. A. Laval and C. F. Daganzo, "Multi-lane hybrid traffic flow model: Quantifying the impacts of lane-changing maneuvers on traffic flow," 2004.

[13] M. Carey, C. Balijepalli, and D. Watling, "Extending the cell transmission model to multiple lanes and lane-changing," *Networks and Spatial Economics*, vol. 15, no. 3, pp. 507–535, 2015.

[14] S. P. Washington, M. G. Karlaftis, and F. Mannering, *Statistical and econometric methods for transportation data analysis*. Chapman and Hall/CRC, 2010.

[15] V. Shankar and F. Mannering, “Modeling the endogeneity of lane-mean speeds and lane-speed deviations: A structural equations approach,” *Transportation Research Part A: Policy and Practice*, vol. 32, no. 5, pp. 311–322, 1998.

[16] W. Cheng, G. S. Gill, T. Sakrani, D. Ralls, and X. Jia, “Modeling the endogeneity of lane-mean speeds and lane-speed deviations using a bayesian structural equations approach with spatial correlation,” *Transportation Research Part A: Policy and Practice*, vol. 116, pp. 220–231, 2018.

[17] M. J. Lighthill and G. B. Whitham, “On kinematic waves ii. a theory of traffic flow on long crowded roads,” *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 229, no. 1178, pp. 317–345, 1955.

[18] A. Zellner and H. Theil, “Three-stage least squares: Simultaneous estimation of simultaneous equations,” in *Henri Theil’s Contributions to Economics and Econometrics*, Springer, 1992, pp. 147–178.