International Journal of Intelligent Transportation Systems Research (2021) 19:572–586
https://doi.org/10.1007/s13177-021-00263-4

Exploring the Performance of Streaming-Data-Driven Traffic State Estimation Method Using Complete Trajectory Data

Garima Dahiya[1] · **Yasuo Asakura**[1]

Received: 12 January 2021 / Revised: 3 June 2021 / Accepted: 16 June 2021 / Published online: 16 July 2021

Abstract
This study aims to evaluate the performance of an extended floating car data (xFCD)-based traffic state estimation method proposed by Seo *et al.* (2015), which does not rely on any strong assumptions such as Fundamental Diagram, using high-resolution complete trajectory data, viz. Zen Traffic Data (ZTD). Traffic state estimated by this method, considering randomly sampled trajectories of ZTD as those of probe vehicles with known penetration rates, are compared with ones obtained by complete ZTD by applying Edie's generalized definitions. The variation in estimation errors and covering percentages are analyzed for varying *settings*: spatiotemporal resolution and probe penetration rates.

Keywords Traffic states estimation · Vehicle trajectory data · Big data and naturalistic datasets · Probe penetration rate · Statistical and theoretical analysis · Spatiotemporal resolution

1 Introduction

Traffic engineering studies differ from other studies in that they require extensive data from the field, which cannot be accurately generated in a laboratory. They pertain to the analysis of the traffic behavior to design facilities for safe, smooth, and economical traffic operations. Density, flow, and average speed (or simply speed, also known as velocity) are the three fundamental parameters of traffic flow. They provide information regarding the nature of traffic on a link at a macroscopic level and aid analysts in detecting any variation in the flow characteristics, which in turn aids in traffic operations and planning. However, obtaining these parameters simultaneously is difficult.

The flow (q), also known as the flow rate or volume by practitioners, is the number of vehicles that pass a given point per unit time. The density (k) is the number of vehicles per unit space at a given instance of time. The average speed (v) is the mean of the instantaneous speeds of the vehicles.

✉ Garima Dahiya
g.dahiya@plan.cv.titech.ac.jp

Yasuo Asakura
asakura@plan.cv.titech.ac.jp

[1] Department of Civil and Environmental Engineering, Tokyo Institute of Technology, 2-12-1-M1-20, O-okayama, Meguro, Tokyo 152-8552, Japan

Edie [1] proposed a generalized definition of traffic states in a time-space region A, defined as follows:

$$q(\boldsymbol{A}) = \frac{d(\boldsymbol{A})}{|\boldsymbol{A}|}, \tag{1}$$

$$k(\boldsymbol{A}) = \frac{t(\boldsymbol{A})}{|\boldsymbol{A}|}, \tag{2}$$

$$v(\boldsymbol{A}) = \frac{d(\boldsymbol{A})}{t(\boldsymbol{A})}, \tag{3}$$

where,

$d(A)$: total distance traveled by all the vehicles in region $A(vehm)$,

$t(A)$: total time all the vehicles spent in region $A(vehs)$,

$|A|$: time-space area of region $A(ms)$.

This definition can be applied to either a single lane or multiple lanes in a link. The process of the inference of traffic state variables or road segments with high spatiotemporal resolution using partially observed traffic data is referred to as Traffic State Estimation (TSE), which depends on the estimation approach, traffic flow model, and input data [2]. The estimation approach can be model-driven, data-driven or streaming-data-driven based on the input data and the assumptions made by the method on traffic dynamics. The physics-based mathematical models

of traffic flow, utilized by the model-driven estimation approach, describe the physical and theoretical aspects of traffic dynamics. TSE methods based on models developed using empirical observations are considered to have '*strong*' assumptions because these methods rely on an explicit a priori knowledge of traffic dynamics and can be vulnerable under uncertain phenomena. Although they have high explanatory power and can be integrated with traffic control operations directly, a poor physical model or poor calibration of the model may lead to poor TSE. Moreover, they are not always consistent with the detailed disaggregated mobile datasets that are recently garnering significant attention owing to recent advancements in information & communication technology. Solving boundary value problems (BVPs) can be regarded as model-driven TSE, where the boundary conditions and models are assumed to be correct. Several methods have been developed to combine mobile data with stationary data using first- or second-order models and filtering techniques such as Kalman filtering techniques (KFTs) for TSE.

This requires either an improvement of these theoretical models or the utilization of data-driven or streaming-data-driven estimation approaches [2]. Now, even though the data-driven approaches do not rely on physical traffic flow models, they rely extensively on historical data to find dependence using statistical methods or machine learning (ML). Although ML is capable of efficiently predicting nonlinear phenomena often found in the transportation field, the computation costs for training and learning can be high. Moreover, the methods can be considered black boxes, and it is difficult to obtain deductive insights. Additionally, they may fail if irregular events or long-term trends occur. Imputation methods have been developed to complement missing data and techniques such as kernel regression (KR), fuzzy c-means (FCM), k-nearest neighbors (kNN) etc., and have been used to incorporate more spatial-temporal information. Traffic flow models and the use of (statistical) dependency on historical data are considered '*strong*' assumptions.

The streaming-data-driven approaches rely on streaming data and use '*weaker*' assumptions such as conservation law (CL). They require less a priori knowledge and no historical data. They can be robust against uncertain phenomena and unpredictable incidents. The moving observer method and its variants have been used for TSE with only a random sampling assumption. In a few studies, extended floating car data (x-FCD) were used with and without the conservation law. In general, it is preferable for practical applications if accurate TSE is achievable based on '*weaker*' assumptions [2].

The objective of this study is to analyze the performance of an xFCD-based traffic state estimation method [3] using high-resolution complete trajectory data: Zen Traffic Data (ZTD). The estimation method is discussed in Section 1.1. Section 2 describes the data and methodology employed in the estimation of traffic states using ZTD. Section 3 describes the empirical analysis.

1.1 Research Objectives and the Estimation Method (Seo *et al.*, 2015)

The estimation of high-resolution traffic states is mainly beneficial for traffic control to mitigate congestion. Over the past decade, researchers have contributed to the methodologies for estimating traffic state, i.e., the density, flow, and velocity, from traffic data without any exogenous assumptions on traffic flow characteristics, such as Fundamental Diagram (FD), which renders the estimation methods robust against unpredictable or uncertain traffic phenomena. Seo *et al.* [3] proposed a streaming-data-driven estimation method for obtaining volume-related variables in predetermined time-space regions, which employed probe vehicles that could measure their positions and the distances to their leading vehicle (space headway between the probe vehicle and its leading vehicle in the same lane). The estimators for the flow, density and velocity, are formulated (using Edie's definitions in (1) – (3)) as follows:

$$q(\boldsymbol{A}) = \frac{d(\boldsymbol{A})}{|\boldsymbol{A}|} \Rightarrow q(A) = \frac{\sum_{n\in N(A)} d_n(A)}{\sum_{n\in N(A)} |a_n(A)|} \Rightarrow \widehat{q}(A) = \frac{\sum_{n\in P(A)} d_n(A)}{\sum_{n\in P(A)} |a_n(A)|}, \tag{4}$$

$$k(\boldsymbol{A}) = \frac{t(\boldsymbol{A})}{|\boldsymbol{A}|} \Rightarrow k(A) = \frac{\sum_{n\in N(A)} t_n(A)}{\sum_{n\in N(A)} |a_n(A)|} \Rightarrow \widehat{k}(A) = \frac{\sum_{n\in P(A)} t_n(A)}{\sum_{n\in P(A)} |a_n(A)|}, \tag{5}$$

$$v(\boldsymbol{A}) = \frac{d(\boldsymbol{A})}{t(\boldsymbol{A})} \Rightarrow v(A) = \frac{\sum_{n\in N(A)} d_n(A)}{\sum_{n\in N(A)} t_n(A)} \Rightarrow \widehat{v}(A) = \frac{\sum_{n\in P(A)} d_n(A)}{\sum_{n\in P(A)} t_n(A)}, \tag{6}$$

where in $\sum_{n\in N(A)} |a_n(A)|$, $a_n(A)$ represents the time-space region between vehicle n and its leading vehicle in a spatiotemporal cell, A, of a meshed spatiotemporal region, R, and,

$N(A)$: the set of all vehicles in the cell A,

$d_n(A)$: total distance traveled by vehicle n in cell A,

$t_n(A)$: total time spent by vehicle n in cell A.

When estimating traffic states using probe vehicles, $N(A)$ is replaced by $P(A)$, which depicts the set of all probe vehicles in region A, as illustrated in Fig. 1. Furthermore, the spatiotemporal area between a probe vehicle and its leading vehicle was computed by using approximations based on the spacing measured by the probe vehicle. This method was previously verified by comparing traffic states estimated by the method, using the data obtained from the employed 20 probes equipped with mono-eye cameras and GPS loggers that drove multiple laps, and those observed by detectors at certain settings that involved two probe vehicle penetration rates and two spatiotemporal resolutions. The TSE method

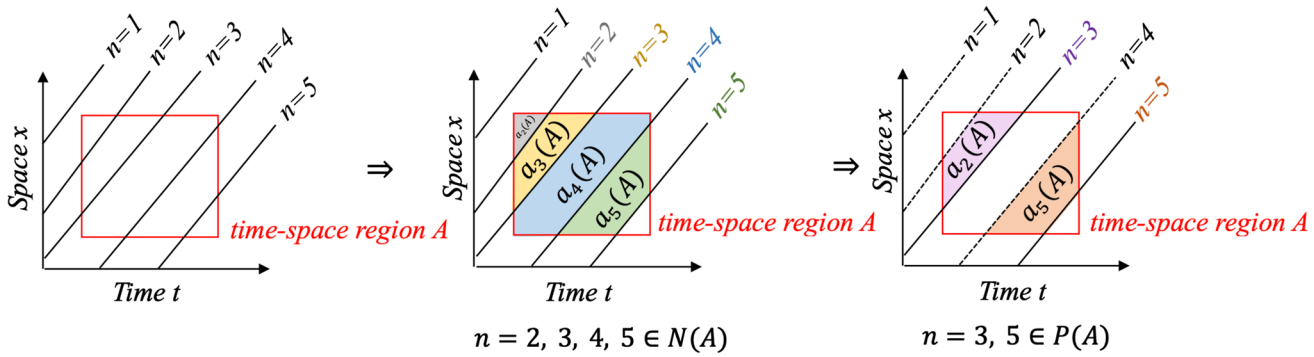

Fig. 1 Illustration of formulation of considered streaming-data-driven TSE method

under consideration relies on '*weaker*' assumptions of '*error free assumption*' (measurements by probes have no error and the driving route is identified without error) and '*random sampling assumption*' (probes are randomly distributed in traffic with unknown penetration rates during estimation and the driving behavior of probes and non-probes are similar). However, the study stated that the data acquired from the probe vehicle that was used contained biases (i.e., the random sampling assumption was not satisfied) such as differences between the driving behavior of probes and differences between the spacing measurements. It suggests that without such biases, the estimation accuracy may be improved. The spacing measurement method involved the identification of leading vehicles in the images (captured by probes), from which their apparent sizes were measured. The spacing was calculated based on the apparent size, assumed actual size, angle of view of the camera, etc. The actual body length was assumed to be the same as that of the probe vehicles (5 m). Other variables were manually measured using the images. If the assumed or measured variables contained errors, the estimated spacing contained errors, which in turn affected the calculation of $a_n(A)$. Although the assumed variables were based on common knowledge of statistics and regulations and detector data, the amount of the errors could not be determined because there is no ground truth data for the vehicle size during the experiment.

The objective of this study is to analyze the validity of the discussed probe vehicle-based traffic states estimation method using the high-resolution Zen Traffic Data (hereafter, ZTD) for different *settings*: spatial resolution (hereafter, Δx), temporal resolution (hereafter, Δt), and probe vehicle penetration rate (hereafter, $p\%$). The ZTD contains comprehensive trajectory details of 100% vehicles, which aids in identifying the leading vehicle to each vehicle in every lane. The spatiotemporal resolution considered in this analysis was not coarse; therefore, for the sake of analyzing the accuracy of the method, exact spatiotemporal coordinates of a vehicle and its leading vehicle were used to calculate a nearly accurate value of the spatiotemporal area between a vehicle and its leading vehicle ($a_n(A)$) without any approximation. With the advancement in data acquisition technologies and the advent of connected vehicles in the near future, it is expected that probe vehicles with advanced driver assistance system (ADAS), potentially capable of recording the exact spatiotemporal coordinates of the leading vehicle too, will be used to obtain data similar to ZTD. The proposed method relied on assumptions that may not always be satisfied in the real world, namely the error free and random sampling assumptions. In this study, $p\%$ vehicles are randomly selected from 100% vehicles driving on a lane for a fixed distance and time, instead of employing probe drivers, to evaluate the estimation capability of the TSE method. This satisfied the random sampling assumption of the estimation method, where the possibility of bias in the driving characteristics of the selected probes to the rest is absent.

2 Data and Methodology

Stationary data (or Eulerian data) and mobile data are two major categories of empirical traffic data available presently based on the measurement methodology [2]. Fixed sensors, such as inductive loop detectors, ultrasonic detectors, and closed circuit television cameras, can be considered as conventional that collect stationary data. Their accuracy and precision may not be reliable, for instance, because of frequent misses and/or double counting by loop detectors. The problem of missing data arises mainly because of the sparse sensor installation owing to the impracticality of installing detectors everywhere and the generally high operational costs of roadside sensors. Therefore, the limitation to them is that the amount of data they provide is not always sufficient for traffic control. Owing to recent advancements in information and communication technologies (ICTs), mobile sensors, such as on-vehicle GPS devices, call detail records (CDRs), and second generation on-board diagnostics systems (OBD-II), are relatively new. As a result of emerging connected and automated vehicles, these sensors are

increasingly used as sources of data. Vehicles with such sensors, often referred to as probe vehicles or floating cars, are a cost-effective way to collect data. The penetration rate and temporal sampling rate are two important characteristics of the probe vehicle data. Probe vehicles are capable of collecting mobile data from a wider spatiotemporal domain as compared to stationary sensors [4, 5]. The new type of mobile data, collected by probes that are equipped with advanced on-vehicle sensors, consists of more than just the positioning and speed of the vehicle trajectory; thus, it has been named extended floating car data (xFCD) [6]. However, the probe vehicle data may contain biases based on sampling and differences in the driving behavior of the probes. For instance, if the probe vehicles belong to a logistic fleet, they may travel at slower than average speeds. In addition, vehicles with recent advanced driving technologies (e.g., ADAS and connected vehicles), which may be used as probe vehicles, may exhibit different driving characteristics compared to completely manually operated vehicles and progressively change driving and traffic patterns.

Seo *et al.* [3] proposed a streaming-data-driven estimation method using only mobile data: xFCD, where each probe vehicle could measure the spacing between it and its leading vehicle (cf., Section 1.1). With continuous advancements in autonomous technologies and the massive emergence of connected vehicles, this approach may become prevalent in the near future, provided it can estimate nearly accurate traffic state. However, currently, only a few percentages of probes are expected on the highways of Japan, where the maximum size of the spatiotemporal cell can be 200 m x 300 s. Additionally, ramp metering and signal control require spatiotemporally detailed information for the target road sections. This study aims to analyze the performance of the estimation method discussed earlier at different *settings*, viz. probe penetration rate, spatial resolution, and temporal resolution. In other words, how much accuracy can be expected under finer spatiotemporal resolution and fewer probe penetration rates.

For doing so, complete and high-tech data with a high temporal sampling rate of 0.1 s, developed using image sensing technology, have been utilized, namely the Zen Traffic Data [7]. Consequently, the in-depth details of each vehicle in real complicated traffic phenomena, which has been challenging to comprehend so far, have been digitalized. It is a large-scale trajectory dataset developed by Hanshin Expressway Co. Ltd. for Ikeda Route 11, around Tsukamoto Junction (5.0 – 3.0 kp), in the inbound direction, in Japan. The section is initially an 'S'-shaped curve, which subsequently becomes a simple straight line, as shown in Fig. 2. It consists of two lanes, a merging section with a major on-ramp, two slightly curved sections, and a sag section.

It includes continuous trajectory information (and any other data affecting traffic events) of all the vehicles as described by parameters, viz. vehicle_id (vehicle ID), datetime (time with 0.1 s precision), vehicle_type (normal or large vehicle), velocity, traffic_lane (driving, passing or entrance), kilopost (distance from the starting point of the expressway route), vehicle_length (estimated vehicle length obtained from image recognition), latitude, longitude, etc. for each vehicle. For analysis, the traffic data L001_F001, which contained the details of 3,375 vehicles with vehicle IDs ranging from 0 to 3,734 from 7:00 a.m. to 8:00 a.m. for a distance of 2 km (5 kilopost to 3 kilopost) is considered. The ZTD can be considered equivalent to data obtained from all vehicles equipped with advanced driving assistance systems (ADAS), which can be utilized as a source of data for volume-related information and to verify traffic state estimation (TSE) methods.

The ZTD data is massive and can be considered appropriate for traffic studies, including traffic state estimation [8]. It can provide meaning to the validation of physical models, which are not solely data driven, and data-driven and

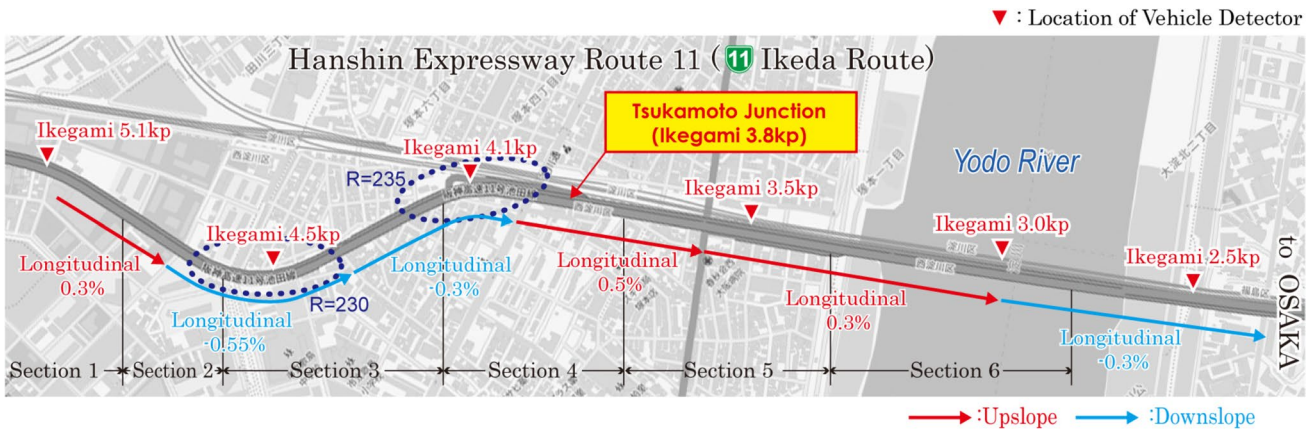

Fig. 2 Ikeda Route 11, around Tsukamoto entrance (5 – 3 kp). Source: https://zen-traffic-data.net/english/outline/

streaming-data-driven estimation models. The accuracy of the ZTD was evaluated by Seo *et al.* [9] concluding the recall rate and the precision rate to be 96.8 and 97.1%, respectively. It was observed that the detection performance was almost insensitive to traffic conditions, weather conditions, and the time of day [9]. The problems of data delay, data loss, inaccurate data, and inconsistent data, which are usually present even in data obtained from recently developed conventional vehicle-to-everything (V2X) technologies, as stated by Sun *et al.* [10], are non-existent in ZTD to a great extent.

2.1 Data Preparation

On the distance of 2 km (5 kilopost to 3 kilopost i.e., 5000 to 3000 m) lane changing is prohibited for the distance between 4200 and 3400 m, and a merging to the driving lane from outside the entrance lane occurs at 3.8 kilopost (Tsukamoto junction) as depicted in Fig. 3. The color bar in the figure aids in understanding the speed profiles of all the vehicles at all space-time locations for both, the driving and the passing lanes for 1 h (7:00 a.m. to 8:00 a.m.) on the said 2 km distance. In this lane change prohibited distance, two sections: 4250 to 3850 m (400 m long section) and 3750 to 3450 m (300 m long section) are specifically considered for this analysis that have minimum lane-changing behavior and maintain the conservation of vehicles throughout each section. However, from 7:00 a.m. to 8:00 a.m., some lane changing behavior was still observed in both these sections: 106 out of 3405 vehicles (3.1%) changed lanes on the 300 m section and 389 out of 3391 vehicles (11.5%) changed lanes on the 400 m section. The percentage of vehicles showing differences in driving behavior was not high; therefore, these were excluded from the analysis to assume homogenous driving behavior among drivers. Resultingly, the number of considered vehicles that drove on 300 m section (lane 1) (R_1), 300 m section (lane 2) (R_2), 400 m section (lane 1) (R_3) and 400 m section (lane 2) (R_4) for one hour (7:00 a.m. to 8:00 a.m.) without changing lanes were 1400, 1735, 1182 and 1715 respectively. Using voluminous ZTD, it was possible to identify the sequential order of vehicles driving in each lane of each section for one morning peak hour for 2 kms and which was maintained throughout the section. Hence, the leading vehicle to each vehicle was identified along with their trajectories in their respective time-space regions (R_i). This serves as an essential ingredient in estimating traffic states by the estimation method using ZTD.

Each time-space area is divided into meshes of varying spatiotemporal resolutions i.e., each time-space region (R_i) subject to the traffic state estimation is divided into multiple discrete, identical, and rectangular time-space regions that can be horizontal or vertical depending on the combination of spatial and temporal resolutions as per Fig. 4. Any rule

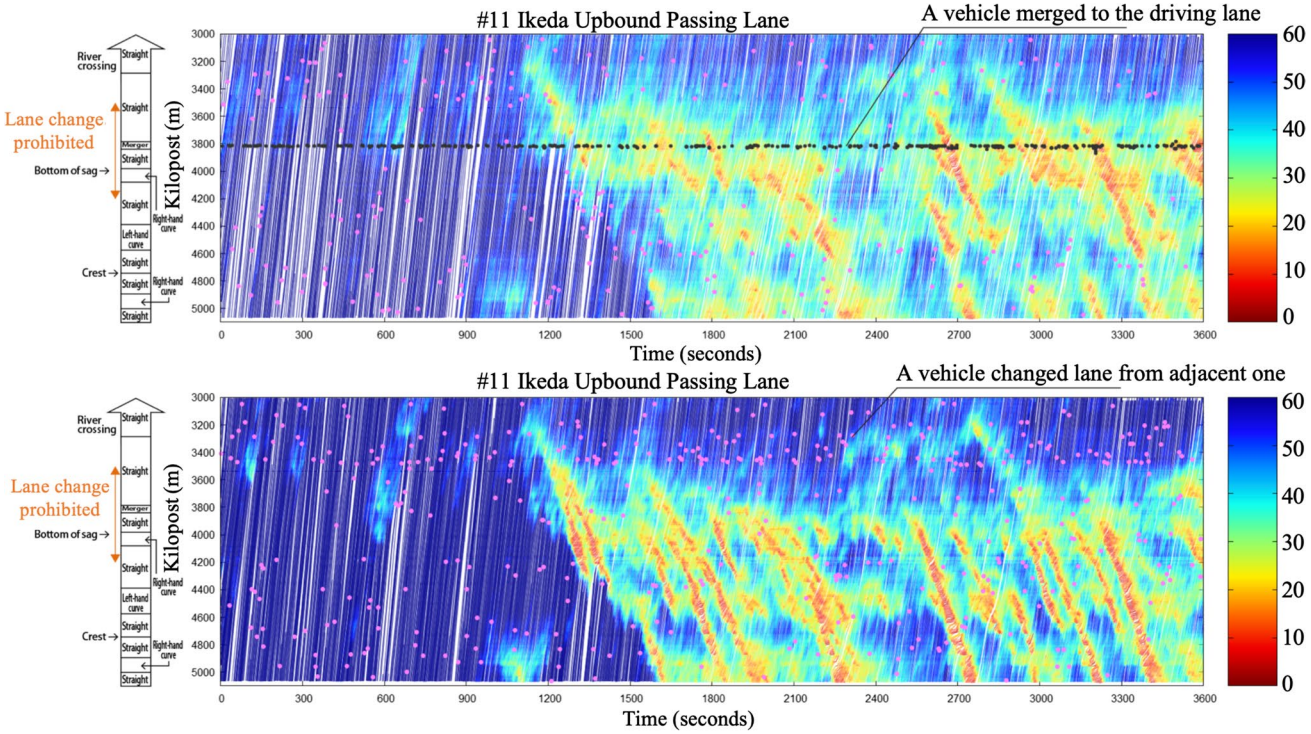

Fig. 3 Time-space diagram per lane (Hanshin Expressway Route 11 Ikeda Line (Osaka Bound)). Source: https://zen-traffic-data.net/english/outline/

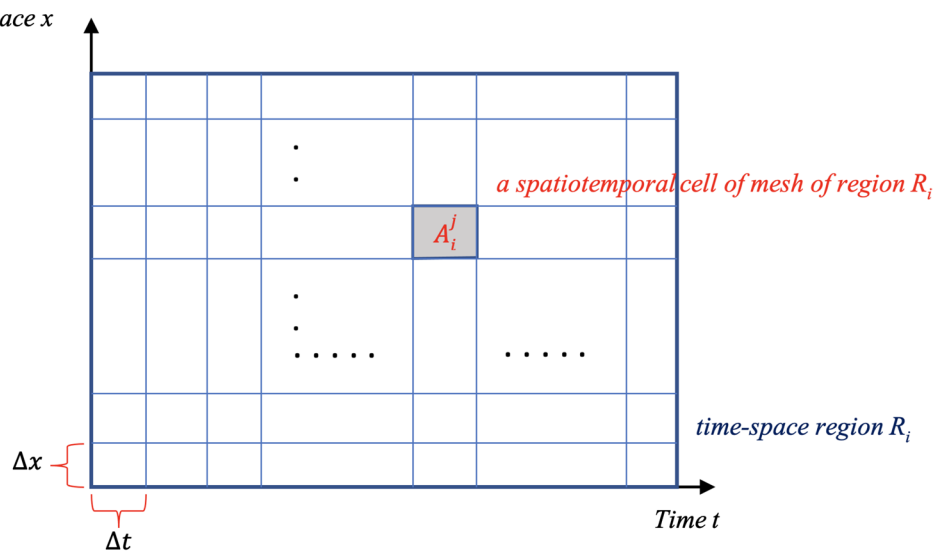

Fig. 4 Time-space area divided into Eulerian rectangles

can be used to divide the time-space region of the traffic flow. The simplest rules are employed in this study, where the traffic flow is divided into Eulerian rectangles of identical sizes. These are familiar coordinates in current traffic flow data, where fixed-point detectors are installed at a certain time and space resolution or interval. The coordinates can be represented as follows:

$$A_i^j = \left\{(t,x)|t_i \le t \le t_{i+1}, x_j \le x \le x_{j+1}\right\} \; i \ge 0, j \ge 0\,, \tag{7}$$

$$t_{i+1} = t_i + \Delta t, \tag{8}$$

$$x_{j+1} = x_j + \Delta x, \tag{9}$$

where,

i, j: non-negative indices for time and space,

(t_0, x_0): coordinates of the predetermined origin,

(t_i, x_j): coordinates of the upper-left corner of region A_i^j,

Δt : predetermined time resolution i.e., $\Delta t = \{15s, 30s, 60s, 120s, 300s\}$,

Δx : predetermined space resolution i.e., $\Delta x = \{25m, 50m, 100m, 150m, 300m\}$ for R_1 and R_2, and $\Delta x = \{25m, 50m, 100m, 200m, 400m\}$ for R_3 and R_4,

The value of x varies as $3450 \le x \le 3750$ for R_1 and R_2 and $3850 \le x \le 4250$ for R_3 and R_4, and t varies as $25,200,000ms \le t \le 28,800,000ms$ (7:00 a.m. to 8:00 a.m.). Corresponding to each R_i, there are 25 combinations of Δt and Δx (25 meshes), where each cell of each mesh is identified by cell A_i^j (hereafter, A).

2.2 Traffic States Estimation Using Zen Traffic Data

First, the traffic state, which at a macroscopic level is a set of the following variables: flow q, density k, and average speed v, is computed using Edie's definitions for each cell (A) of each mesh (for every combination of Δx and Δt) corresponding to every R_i. Under every *setting*, the trajectory information from the ZTD of all vehicles driving through a cell is used to compute q (veh/s), k (veh/m), and v (m/s). Assuming the ZTD as a source of ground truth, these values are used to make comparison with the traffic state computed using the estimation method. Using the described methodology, Fig. 5 shows the traffic flow computed for $\Delta x = 25m$ and $\Delta t = 30s$ on R_1 (for instance). For all time-space regions R_i, the traffic flow ranges from 0.3 to 0.5 veh/s (18–30 veh/min) in a majority of the meshed cells (A), and at a few positions and times on the sections the traffic flow is over 0.6 veh/s (36 veh/min) (reaching values of flow at critical density), which mostly occurs on lane 2 and before 7:20 a.m.

Estimating the traffic state using the estimation method requires random sampling of $p\%$ vehicles (hereafter referred to as probe vehicles) from the total number of vehicles driving through each time-space region R_i. Each vehicle (with its trajectory data) is chosen entirely by chance by utilizing the pseudo-random decimal numbers (real numbers between 0 and 1) generated by the *RAND* function in *MS Excel* and has an equal probability of being selected as an element of the random sample, in alignment to the probability theory and statistics. The selection isn't based on any uniform pattern, such as the selection of a vehicle after every fixed number of vehicles or in every fixed unit of time. These $p\%$ selected vehicles are a part of the actual traffic and not deployed for analysis. For instance, Fig. 6 depicts the traffic trajectories of 5% randomly selected vehicles from 100% of vehicles driving in region R_1 (1400).

For varying $p\%$ values, the traffic states are estimated using the Eqs. (5) – (7) (right) for each cell (A) of each mesh (for every combination of Δx and Δt) corresponding to every R_i, through which at least one probe vehicles pass.

The topmost rows in Fig. 7(a), (b) & (c) correspond to the traffic state, k (veh/m), q (veh/s), and v (m/s), respectively, estimated using Edie's definitions and the ZTD of 100 % vehicles for R_1 ($\Delta x = 50m, \Delta t = 60s$). The following rows in Fig. 7(a), (b) & (c) illustrate the traffic states, k (veh/m), q (veh/s) and v (m/s), respectively, estimated from the

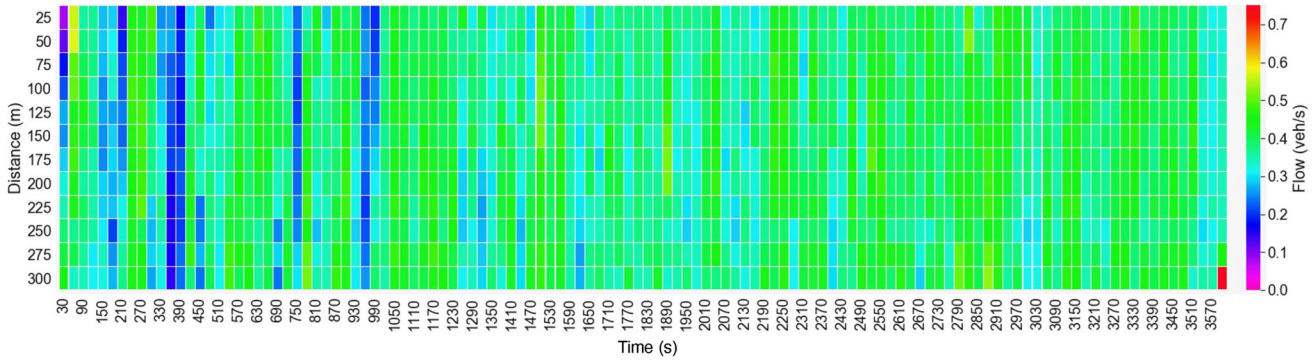

Fig. 5 Traffic flow in R_1 from Edie's definitions and the 100% ZTD ($\Delta x = 25m, \Delta t = 30s$)

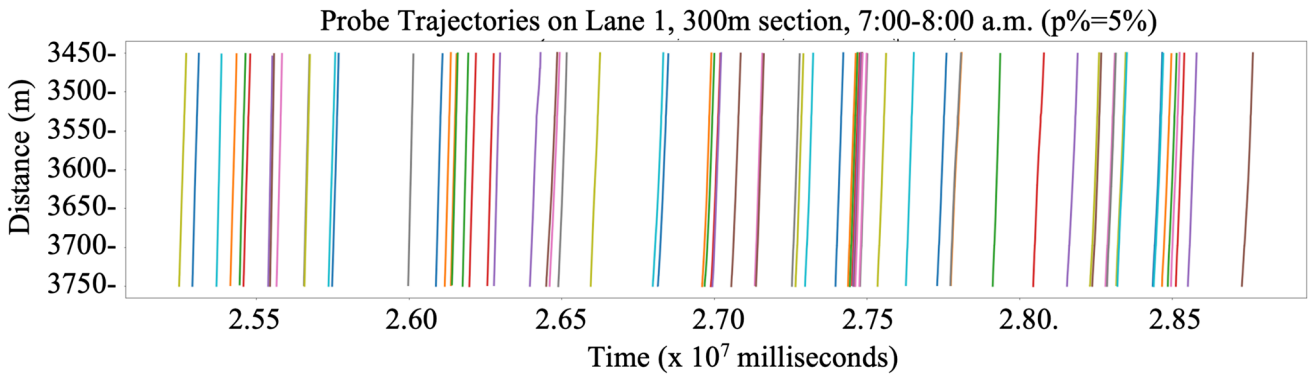

Fig. 6 Traffic trajectories of randomly selected 5% probe vehicles from 100 % vehicles on R_1: Lane 1, 300 m Sec. (7:00–8:00 am) using ZTD

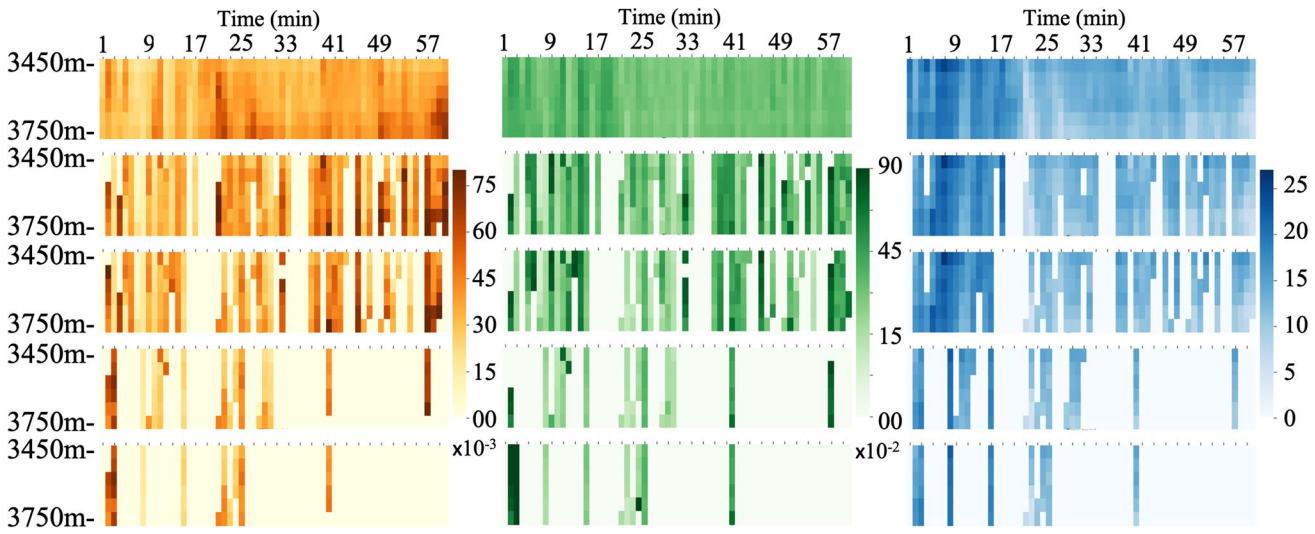

Fig. 7 (**a**) (left) Density (veh/m), (**b**) (middle) Flow (veh/s), (**c**) (right) Speed (m/s) estimated from Edie's definitions (topmost rows) and estimation method for R_1 ($\Delta x = 50\ m,\ \Delta t = 60\ s,\ p\%\ varying\ from\ 5\%\ to\ 0.5\%$) (bottom four rows in order)

estimation method for $p\% = 5\%, 3\%, 1\%$ *and* 0.5%, in this order for R_1 ($\Delta x = 50m, \Delta t = 60s$). For computing the spatiotemporal area ($a_n(A)$) between a probe vehicle (n) and its leading vehicle in the same lane, identified using the ZTD, the exact spatiotemporal coordinates of their trajectories at a 0.1 s pitch are used. For doing so, Gauss's area formula, described by Meister [11] and by Carl Friedrich Gauss in 1795 was implemented in Python. It is also known as the Surveyor's formula [12] and is considered as a special case of Green's theorem (first presented by Cauchy [13]). Let the set of spatiotemporal coordinates of vehicle n and its leading vehicle enclosed within the time-space region of cell A, which form a polygon in the clockwise or anticlockwise direction in the spatiotemporal plane, be represented as $\{(t_1, x_1), (t_2, x_2), \ldots, (t_N, x_N)\}$. The area $a_n(A)$ is derived as follows:

$$a_n(A) = \frac{1}{2}\left|\sum_{i=1}^{N-1} t_i x_{i+1} + t_N x_1 - \sum_{i=1}^{N-1} t_{i+1} x_i + t_1 x_N\right|. \tag{10}$$

Alternatively,

$$a_n(A) = \frac{1}{2}\left|\sum_{i=1}^{N} t_i (x_{i+1} - x_{i-1})\right| = \frac{1}{2}\left|\sum_{i=1}^{N} x_i (t_{i+1} - t_{i-1})\right|, \tag{11}$$

$$a_n(A) = \frac{1}{2}\left|\sum_{i=1}^{N} (t_i x_{i+1} - t_{i+1} x_i)\right|, \tag{12}$$

$$a_n(A) = \frac{1}{2}\left|\sum_{i=1}^{N} (t_{i+1} + t_i)(x_{i+1} - x_i)\right| = \frac{1}{2}\left|\sum_{i=1}^{N} det\begin{pmatrix} t_i & t_{i+1} \\ x_i & x_{i+1} \end{pmatrix}\right| \tag{13}$$

For any cell through which no probe vehicle passes, the values of the allocated traffic state equal zero, as illustrated in Fig. 8. This is a type of missing data that is different from missing data caused by randomness, attrition, or unobserved original data; rather, it is an intentional missing as part of extracting only $p\%$ data for this analysis.

3 Empirical Analysis

A fixed combination of Δx, Δt and $p\%$ is referred to as a *setting*. The total number of such *settings* equals 100 (*ref.* Section 3.1). The traffic states obtained under each *setting*, for each cell A of all spatiotemporal regions R_i, using the estimation method are compared with traffic state obtained using ZTD of all the vehicles driving through cell A on a one-to-one basis. To yield the least biased comparison for cells through which no probe passed, the analysis strategy used is a direct approach: *Deletion Method (Listwise Deletion)*. It is a complete-case analysis, where only the cells with observed probes are considered from both datasets. The $p\%$ probe vehicles are selected randomly; therefore, the cells with no probes do not occur in any systematic order, which could lead to a bias. Its advantages are simplicity and comparability across analyses. The reasons for not considering value-allocating methods for assigning values to the cells through which no probe drives (such as the mean imputation method, using information from related cells, or a hybrid of both methods) are discussed. The objective of this analysis is to study the accuracy of the estimation method for different $p\%$ values. For this method to be applicable in actual scenario, it is important to check the accuracy by not deliberately adding any biases. When $p\%$ is very less then technique-filled cells, for higher spatiotemporal resolutions, will be much larger than the cells with method-estimated data. This would not reflect true errors during comparison. When we fill the empty cells with values from other *settings*, it will give an amalgamation of values and it will not reflect the true variation in error over spatiotemporal resolutions and $p\%$.

To visualize the performance of the estimation method, flow-density (q-k) diagrams were plotted for all combinations of spatiotemporal resolutions and probe vehicle penetration rates combined for all four regions R_i. A few of them

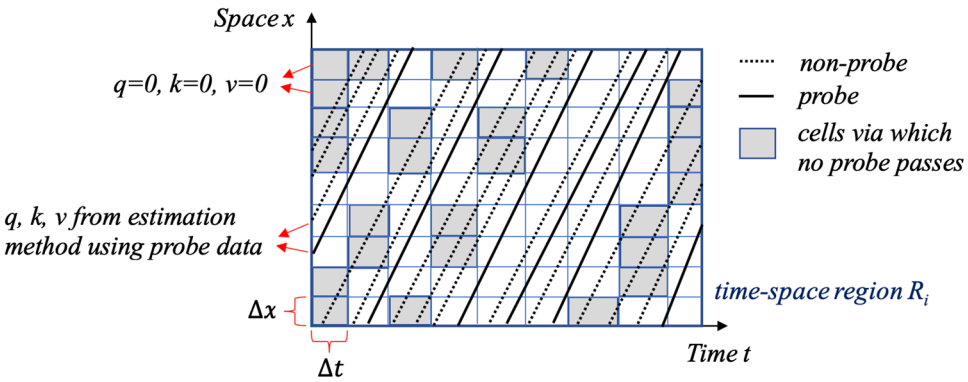

Fig. 8 *Observed cells*: via which at least 1 probe vehicle passes

are illustrated in Fig. 9. The *q-k* plots suggest that the estimation method is able to capture the robust behavior of actual traffic dynamics when the traffic is in the free flow regime. However, for densities beyond the density around critical density the performance of the estimation method appears degraded. Although there exists a cloud of incorrect estimations beyond the critical density, it coexists with the correct estimations to some extent. This implies that existence of a density greater than the critical density in a spatiotemporal cell A is not the sole reason for the diversion of predictions made by the estimation method from actual traffic states in that cell A. For an extensive evaluation, statistical analysis was conducted as discussed in the following section.

3.1 Statistical Error Analysis

To analyze the numeric differences in the traffic states estimated by the probe vehicle-based estimation method (E_i) and those obtained from the ZTD of 100% vehicles (O_i) driving in a spatiotemporal cell A of region R_i, the percent

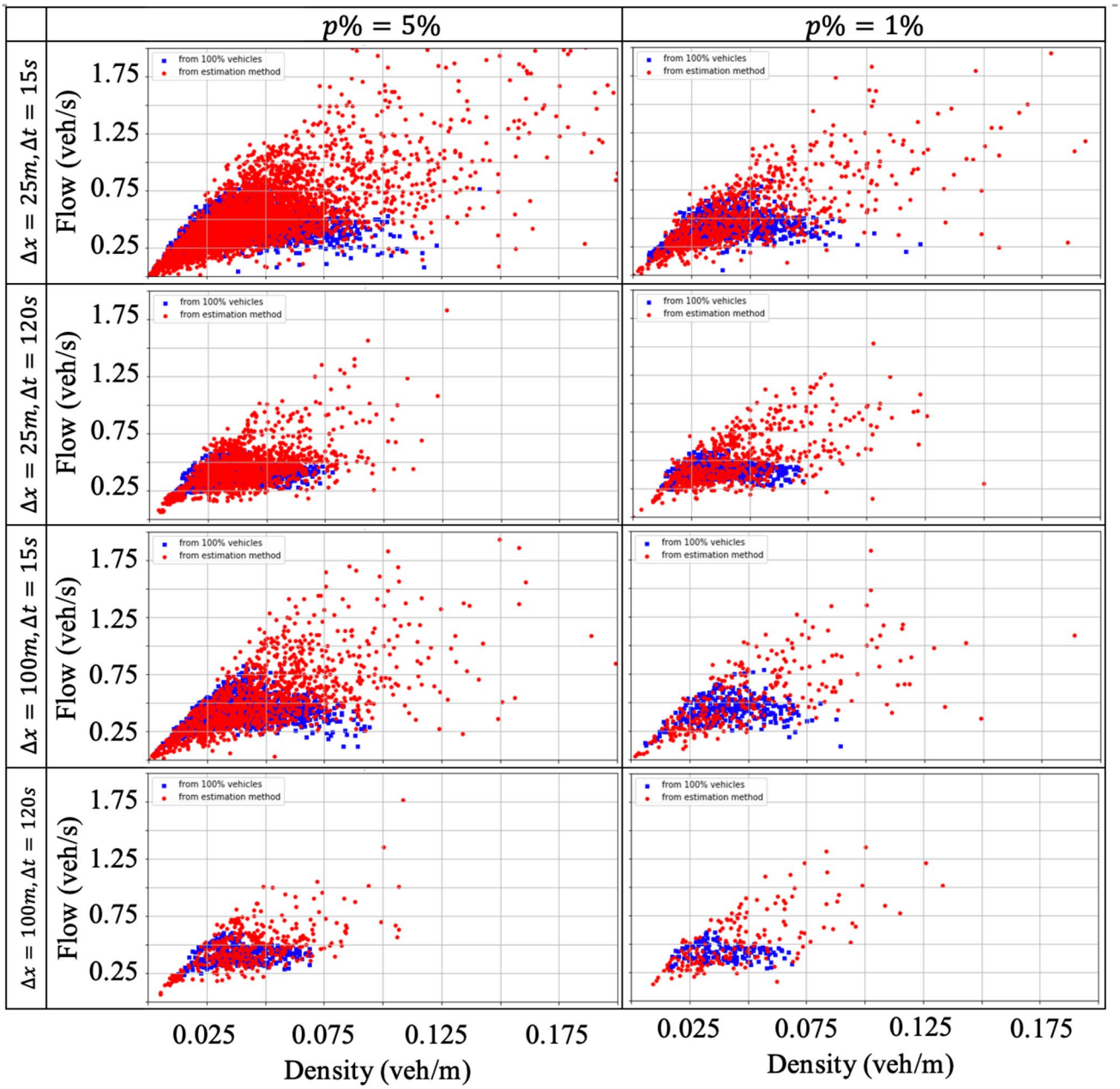

Fig. 9 *q-k* plot for traffic state estimated from Edie's definitions using the ZTD of 100% vehicles (Blue) and from the estimation method (Red) for a few different *settings*

error is calculated for each considered cell A under all 100 *settings* as per (14). Furthermore, the mean absolute percentage errors (MAPEs) and root mean square errors (RMSEs) are also calculated (n: number of cells considered) as per (15) and (16).

$$Percent\ error(\delta) = \left|\frac{E_i - O_i}{O_i}\right|.100\% \tag{14}$$

$$Mean\ absolute\ percentage\ error = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{E_i - O_i}{O_i}\right| \tag{15}$$

$$Root\ mean\ square\ error = \sqrt{\frac{\sum_{i=1}^{n}(E_i - O_i)^2}{n}} \tag{16}$$

Additionally, the number of probes driving through each spatiotemporal cell was recorded for all combinations of considered spatial resolution, temporal resolution, and probe vehicle penetration rate. Intuitively, as the spatial resolution and/or temporal resolution becomes more coarser, or the probe vehicle penetration rate increases, the average number of probes in each cell is expected to increase. However, to determine the precise numerical value, Table 1 details the average number of probes observed in the cells through which at least one probe vehicle passed under a few of the different *settings*, averaged over all four regions (R_i). Under the considered *settings*, the higher values of the average number of probes observed in the cells ranges from 6.22 to 6.75 for $\Delta t = 300s$ and $p\% = 5\%$. The value of Δx is not influencing the averages as such. The second reason that can be considered for the deviation of estimated traffic states from the actual ones is the average number of probe vehicles in the spatiotemporal area under consideration. Figure 10 (a) illustrates that with an increase in average number of probe vehicles in a spatiotemporal area results in a drastic decrease in the MAPE in the estimated density and flow. When the average number of probes is 1 in a cell A, the MAPE in the estimated density and flow is as high as around 140%. At the same time number of probes in a cell is not influencing the errors in the estimated speed very much. Similarly, Fig. 10 (b), (c) and (d) show the depletion in RMSE in estimated k, q, and v with an increase in the average number of probe vehicles in the spatiotemporal area. When the average number of probes is as high as around 6 or 7, the MAPE in estimated k, q and v are as low as around 20% for k and q and less than 10% for v. Also, the RMSE in k, q and v will be around 0.01 veh/m, 0.09 veh/s (5.4 veh/min) and 0.75 to 1.25 m/s, when the average number of probe vehicles in a spatiotemporal region is around 6 or 7.

Primarily, high vehicular density and/or low availability of probes driving through a spatiotemporal region leads to a substandard performance of the estimation method in

Table 1 Average number of probe vehicles in each cell A at a few different *settings*

Δx	25m					100m					200m					300m				
Δt	15s	30s	60s	120s	300s	15s	30s	60s	120s	300s	15s	30s	60s	120s	300s	15s	30s	60s	120s	300s
p%=5%	1.176	1.345	1.764	2.743	6.217	1.251	1.434	1.878	2.876	6.351	1.318	1.505	1.974	3.018	6.354	1.556	1.753	2.258	3.250	6.917
p%=3%	1.081	1.162	1.397	1.920	3.965	1.115	1.211	1.452	1.991	4.032	1.168	1.265	1.467	2.041	3.978	1.244	1.397	1.667	2.265	4.348
p%=1%	1.037	1.071	1.141	1.261	1.784	1.047	1.080	1.157	1.270	1.805	1.056	1.080	1.119	1.218	1.743	1.092	1.170	1.235	1.400	1.941
p%=0.5%	1.004	1.016	1.031	1.031	1.210	1.012	1.015	1.035	1.037	1.225	1.000	1.000	1.000	1.000	1.148	1.073	1.074	1.105	1.125	1.417

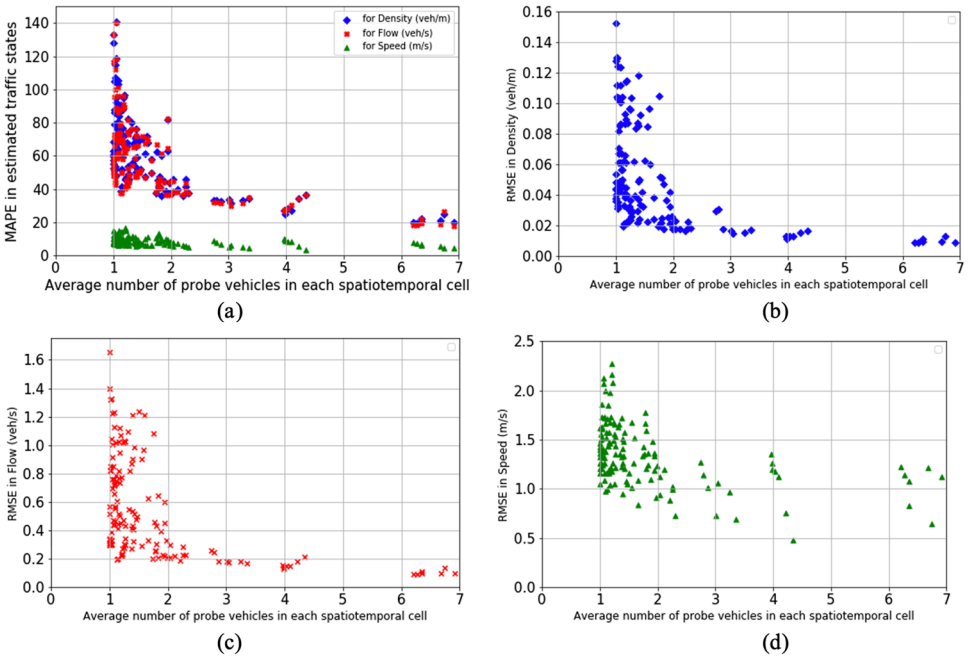

Fig. 10 Variation in (**a**) MAPE,(**b**) RMSE in Density (veh/m), (**c**) RMSE in Flow (veh/s), and (**d**) RMSE in Speed(m/s), with the variation in the average number of probes in *A*

replicating the actual behavior of traffic flow and estimating traffic state. When the probe penetration rate drops below 3% and the temporal resolution becomes finer than 2 min, the average number of probe vehicles in the considered spatiotemporal regions falls below 2 and the MAPE in the estimated density and flow rises over 40%. The variation in MAPE under all the different *settings* can be more clearly visualized in Fig. 11. Under all *settings*, the MAPE for k, q, and v went as low as around 20%, 18%, and 4.5%, respectively. The variation in Δx did not significantly affect the average number of probe vehicles that drove through the considered spatiotemporal cells of fixed Δt and $p\%$ and in turn did not affect much the variation in MAPE in k, q, and v. However, for a fixed Δx and $p\%$, Δt exhibits a monotonically increasing non-linear relationship with the average number of probes observed driving through the cells of the spatiotemporal mesh. This implies that as Δt becomes coarser, the MAPE is expected to decrease. Likewise, to Δt, a drop in $p\%$ leads to a decrease in the average number of probes; however, this drop is gradual for a Δt and steep when Δt is greater than 120s. This, by its nature, has a direct effect on the propagation of MAPE i.e., for a fixed Δx and Δt, a drop in $p\%$ results in an escalation in MAPE. A similar trend was observed with the variation in RMSE of the estimated traffic state being predominantly affected by Δt and $p\%$ (Table 2). The method estimates v with much lower MAPE and RMSE, as compared to the k and q, irrespective of the observation *settings* and the average number of probe vehicles in the spatiotemporal cell. To analyze the effect of the employed random sampling method on the stability of estimation, the estimation method was evaluated for different series of randomly sampled probe vehicles from the complete ZTD at the same *settings*. It implied that the variation in MAPE in k, q, and v at different *settings* was similar, except for a very fine temporal resolution (say Δt=15s) and a low probe percentage (such as 1%). Under such *settings*, the estimation performance was unstable but definitively poor. This instability in estimation using the TSE method can be lessened by considering relatively larger $p\%$ or setting the temporal resolution to be coarser than 15s. Overall, this justifies the reliability of the employed *random sampling procedure* for evaluating the estimation capability of the considered estimation method.

The selection of the $p\%$ of probes is random; therefore, it is possible that a probe belongs to a logistic fleet, which may lead to a slower than average traveling speed. This will lead to a biased traffic state estimation in the time-space cells via which such a probe vehicle passes. This bias, in general, can be ignored for this analysis when $p\%$ is not very small; however, when $p\%$ is very small such as 1 or 0.5%, the MAPE and RMSE, calculated between traffic states obtained from Edie's generalized definitions and the estimation method, may be affected. In such a case, the differences between the individual sampled probe and others in a particular cell, that could be due to driver's and/or vehicular condition, may lead to a lower accuracy. Additionally, certain vehicle_ids that were changing lanes in the 'lane change prohibited' area were excluded beforehand from the ZTD, which could have led to a false recognition of the leading vehicle to a probe vehicle. The accuracy of the estimation method in estimating traffic states positively correlates with the number of probe vehicles

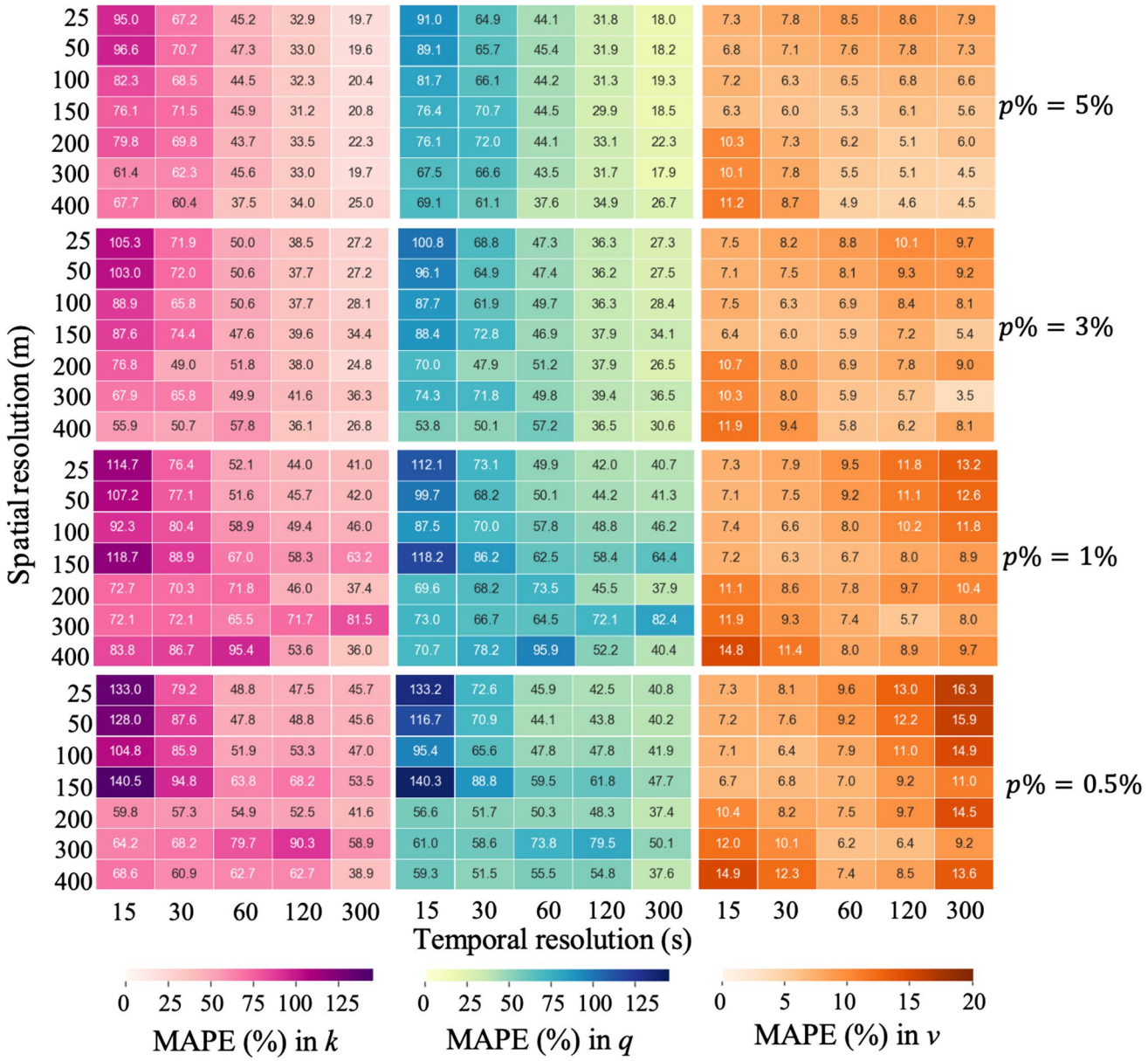

Fig. 11 Variation in MAPEs in Density (veh/m) (left), Flow (veh/s) (middle) and Velocity (or speed) (m/s) (right) over varying *settings* for R_1, R_2, R_3 and R_4 combined

in the time-space region. According to the available $p\%$ or the required accuracy, the practitioners can choose the desired spatiotemporal resolution *settings*. The accuracy depends on the *settings*: mainly, temporal resolution (Δt), and probe penetration rate ($p\%$), but indirectly. This analysis provides an insight into various combinations of *settings*, expected probe vehicles in spatiotemporal cells, and the corresponding expected accuracy. Another important factor to be considered when employing a set of *settings* in estimating traffic states is the covering percentage ($c\%$), which is discussed in the following section.

3.2 Covering Percentage

The covering percentage ($c\%$) is the percentage of cells through which probe vehicles pass given a fixed *setting* over a region R_i. It is intuitive that the $c\%$ has a positive correlation with the probe vehicle penetration rate i.e., the number of probes and the size of the cell in R_i, which was corroborated by the inferences from the analysis. Unlike the accuracy of the estimation method on which Δx has a low to negligible effect, Δx has a positive correlation with the $c\%$. In fact, in terms of the difference in

Table 2 RMSE obtained by comparing traffic state estimated using the estimation method and from the ZTD of 100% vehicles for a fewdifferent *settings*

Spatiotemporal resolution		RMSE in Density (veh/m)				RMSE in Flow (veh/s)				RMSE in Speed (m/s)			
		Probe penetration rate (p%)											
Δx	Δt	5%	3%	1%	0.50%	5%	3%	1%	0.50%	5%	3%	1%	0.50%
25m	15s	0.114	0.123	0.124	0.152	1.119	1.232	1.323	1.657	1.672	1.720	1.608	1.326
	30s	0.084	0.086	0.085	0.104	0.819	0.816	0.848	0.827	1.351	1.468	1.382	1.498
	60s	0.052	0.049	0.040	0.044	0.459	0.434	0.446	0.387	1.350	1.409	1.437	1.521
	120s	0.029	0.025	0.024	0.031	0.259	0.223	0.277	0.302	1.269	1.478	1.654	1.857
	300s	0.009	0.013	0.022	0.023	0.087	0.153	0.249	0.233	1.224	1.351	1.772	2.272
100m	15s	0.096	0.104	0.104	0.130	1.029	1.020	0.913	1.322	1.555	1.573	1.245	1.262
	30s	0.092	0.087	0.101	0.130	0.985	0.772	0.761	0.817	1.203	1.179	1.163	1.264
	60s	0.037	0.042	0.038	0.035	0.392	0.475	0.513	0.326	1.095	1.079	1.158	1.208
	120s	0.018	0.017	0.028	0.035	0.181	0.207	0.340	0.325	1.008	1.234	1.434	1.615
	300s	0.010	0.013	0.025	0.023	0.098	0.148	0.302	0.234	1.080	1.176	1.589	2.074

c% brought about by unit change in a *setting*, the factors that affect the c% in order of decreasing dominance are p%, Δt, and Δx. The variation in c% over different *settings* for R_1 (for instance) is shown in Fig. 12. However, to be able to retrieve the estimates of traffic states in complete spatiotemporal domain is always desirable i.e., to have a higher c%. The c% is positively related to the p%, implying that the c% increases as the average number of probe vehicles driving through the spatiotemporal cells in the mesh of the time-space region R_i increases. However, a higher covering percentage does not imply a high accuracy by an estimation method for obtaining traffic states. For instance, the traffic states of a very large spatiotemporal area estimated using trajectory data from a single probe may lead to a high covering percentage, but with lower accuracy. Hence, for a combination of finer Δt (finer than 2 min) and a lower p% i.e., below 3%, a compromise is made with both accuracy and the c%.

4 Conclusions and Scope for Future Research

This study aimed to evaluate the performance of an xFCD-based traffic state estimation method, unconfined by any exogenous assumptions such as FD, proposed by Seo *et al.* [3]. This was conducted at finer spatiotemporal resolutions and varying probe vehicle penetration rates using high resolution complete trajectory data, viz. the Zen Traffic Data. The initial challenge in validating this estimation method lies in the identification of the leading vehicle to a probe vehicle. This was meticulously performed with the aid of the ZTD, which enabled the identification of the exact trajectories of the leading and the probe vehicles. The detailed resolution of the ZTD played a critical role in evaluating the actual performance without any approximations based on the spacing measurements calculated using assumptions. The exact spatiotemporal coordinates of vehicles were utilized in reckoning the spatiotemporal area between a probe vehicle and its leading vehicle. In spite that currently probe

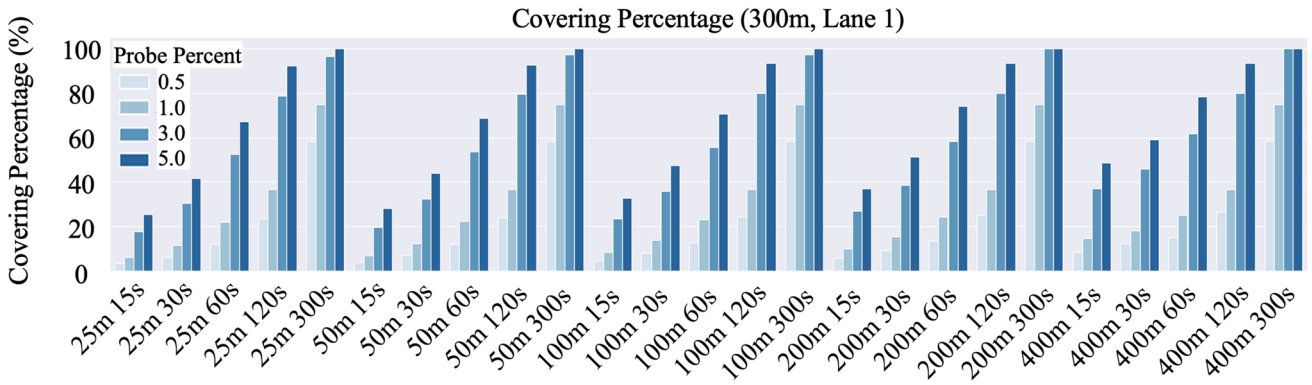

Fig. 12 Variation in covering percentage in R_1 over varying *settings*

vehicles (even with ADAS) are impotent in collecting data comparable to that of the ZTD yet inspecting the estimation method using the ZTD elucidated the application of the estimation method at desired *settings* using probe vehicles that are proficient at providing information regarding the spacing between it and its leading vehicle. The importance of this result lies in the utilization of detailed ZTD in estimating the traffic state using the discussed estimation method, while using other conventional datasets failed to provide the same degree of accuracy. The ZTD is more reliable than other conventional datasets in deducing inferences from the performance or accuracy evaluation of estimation methods. This is because reactive traffic controls such as ramp metering and signal control require spatiotemporally detailed information (e.g., on the speed, flow, and queue length) at target road sections [14]. A unique vehicle ID has been allocated to each vehicle that traveled on the expressway, which is observed and maintained throughout a target section and target time duration. There is a continuity of data at 0.1 s time step with no loss. This complete information for 100% vehicles is impossible to be acquired with conventional datasets such as loop detectors or probe vehicles, such as a GPS-equipped probe vehicle or moving observer methods. Moreover, the ZTD is more advantageous than the currently popular NGSIM dataset because the latter can only cover smaller segments of the highway and is limited in both spatial and temporal terms.

The q-k plots for the estimated traffic states along with the actual traffic states suggested that in the free-flow regime, the estimation method was able to reproduce the scatter, present in the q-k plots of the actual traffic states, in the estimated states without the assumption of stationarity. As the density increases further, the performance of the estimation method deteriorates. The statistical analysis suggested that the MAPE and RMSE scores for the estimated density and flow are inversely related to the number of probes in a spatiotemporal region, which is predominantly affected by only the temporal resolution and the probe vehicle penetration rate. Specifically, the MAPE for q and k can be as high as 140% for the finest spatiotemporal resolution among the considered *settings* if the average number of probes in the cells of a spatiotemporal mesh is 1. Whereas, when the average number of probes in the cells of a spatiotemporal mesh is around 6 or 7 the MAPE values can be lower than 20% for k and q and around 10% for v. Concurrently, the RMSE in k, q, and v curtails to 0.01 veh/m, 0.09 veh/s (5.4 veh/min), and 0.75 to 1.25 m/s, respectively. When the probe penetration rate falls below 3% and the temporal resolution is finer than 2 min, the MAPE in estimated k and q rises over 40%. Nevertheless, under all the *settings* considered for this analysis, the MAPE for k, q, and v went as low as around 20%, 18%, and 4.5%, respectively. The method estimates v with much lower MAPE and RMSE values, irrespective of the observation *settings*, as compared to k and q. The accuracy of the estimates depends on two *settings*: temporal resolution (Δt), and probe penetration rate (p%), but indirectly. This analysis provides an insight into the various combinations of *settings*, expected probe vehicles in spatiotemporal cells, the corresponding covering percentage, and the expected accuracy. It is always desirable to be able to retrieve the estimates of traffic states in a complete spatiotemporal domain i.e., to have a higher c%. However, for a combination of a finer Δt i.e., finer than 2 min and a lower p% i.e., below 3%, a compromise is made with both accuracy and the c%. Additionally, the consideration of appropriate value of Δx may be ignored in terms of accuracy yet Δx has a positive correlation with the c%. Thus, according to the available p% or the required accuracy and c%, practitioners could select the desired and appropriate spatiotemporal resolution *settings*. In actual, few percentage of GPS probes are expected in the actual highways of Japan (where the maximum cell size for traffic control is Δx =200 m and Δt =300 s), and the *settings* considered in this analysis aided in visualizing expected errors in the estimation results using this method at finer Δx and Δt and a lower p%. With few percentages of prove vehicles, the method can estimate traffic states at coarser resolutions with 100% coverage when the expressway is not in the congested state. This low resolution is sometimes useful for planning purposes and for potential area-wide traffic management.

Scope for future research includes the furtherance of the estimation method, keeping in mind the scope of improvement required for spatiotemporal regions with a higher vehicular density. Furthermore, at an age of near ubiquitous sensor (e.g., cell phone) penetration, and with the massive emergence of connected vehicles, the validation result suggests that the approach might become prevalent in the near future for transportation planning purposes with a probe vehicle penetration rate of several percentages. Namely, for traffic management and control purposes, the proposed method may require a higher penetration rate; considering the possible widespread implementation of ADAS in the future, such a high penetration rate might be realized.

Acknowledgements The Zen Traffic Data were provided by Hanshin Expressway Co. Ltd. This study was financially supported by the Japan Society for the Promotion of Science (KAKENHI 17H01297). The authors would like to express their sincere appreciation for the help.

References

1. Edie, L.C.: Discussion of traffic stream measurements and definitions. In: Almond J. (ed.) Proceedings of the 2nd international symposium on the theory of traffic flow, pp. 139–154 (1963)
2. Seo, T., Bayen, A.M., Kusakabe, T., Asakura, Y.: Traffic state estimation on highway: A comprehensive survey. Annu. Rev. Control **43**, 128–151 (2017)

3. Seo, T., Kusakabe, T., Asakura, Y.: Estimation of flow and density using probe vehicles with spacing measurement equipment. Transp. Res. Part **C53**, 134–150 (2015)
4. Herrera, J.C., Work, D.B., Herring, R., Ban, X.J., Jacobson, Q., Bayen, A.M.: Evaluation of traffic data obtained via GPS-enabled mobile phones: The mobile century field experiment. Transp. Res. Part C: Emerg. Technol. **18**(4), 568–583 (2010)
5. Zito, R., D'Este, G., Taylor, M.A.P.: Global positioning systems in the time domain: How useful a tool for intelligent vehicle-highway systems? Transp. Res. Part C: Emerg. Technol. **3**(4), 193–209 (1995)
6. Huber, W., Lädke, M., Ogger, R.: Extended floating-car data for the acquisition of traffic information. In: Proceedings of the 6th world congress on intelligent transport systems, pp. 1–9 (1999)
7. Zen Traffic Data source: https://zen-traffic-data.net
8. Dahiya, G., Asakura, Y.: Evaluation of probe vehicle-based traffic states estimation method using Zen Traffic Data. In: Proceedings of the 18th Intelligent Transportation Systems (ITS) Symposium, Matsuyama, Japan (2020)
9. Seo, T., Tago, Y., Shinkai, N., Nakanishi, M., Tanabe, J., Ushirogochi, D., Kanomori, S., Abe, A., Komada, T., Yoshimura, S., Ishihara, M., Nakanishi, W.: Evaluation of large-scale complete vehicle trajectories dataset on two kilometers highway segment for one hour duration: Zen Traffic Data. In: Proceedings of International Symposium on Transportation Data & Modeling (ISTDM 2021, Ann Arbor, Michigan, U.S.A.) (forthcoming).
10. Sun, D., Zhao, H., Yue, H., Zhao, M., Cheng, S., Han, W.: ST TD outlier detection. IET Intell. Transp. Syst. **11**(4), 203–211 (2017)
11. Meister, A.L.F.: Generalia de genesi figurarum planarum et inde pendentibus earum affectionibus. Nov. Com. Gött. (in Latin), pp. 1–144 (1769)
12. Braden, B.: The Surveyor's Area Formula. Coll. Math. J. **17**(4), 326–337 (1986)
13. Cauchy, A.: Sur les intégrales qui s'étendent à tous les points d'une courbe fermée. (On integrals that extend over all of the points of a closed curve), Compt. Rendus. 23, pp. 251–255 (1846)
14. Dahiya, G., Asakura, Y., Nakanishi, W.: A study of speed-density functional relations for varying spatiotemporal resolution using Zen Traffic Data. In: Proceedings of the IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), pp. 1–8, Rhodes, Greece (2020)

Garima Dahiya PhD student at Tokyo Institute of Technology, Tokyo, Japan. MSc. degree in Applied Mathematics from Indian Institute of Technology, Mandi, India. Key interests: Traffic State Estimation, Applied Mathematics, Data Analysis, ML.

Yasuo Asakura Professor at Tokyo Institute of Technology, Tokyo, Japan. Research interests: Traffic Engineering and Transportation Planning, Network Analysis, Travel Behavior Analysis.