Transportation Research Part C 137 (2022) 103555

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Data fusion for estimating Macroscopic Fundamental Diagram in large-scale urban networks

Elham Saffari, Mehmet Yildirimoglu *, Mark Hickman

The University of Queensland, School of Civil Engineering, Brisbane, QLD 4076, Australia

ARTICLE INFO

Keywords:
Macroscopic Fundamental Diagram (MFD)
Bayesian data fusion
k-nearest neighbour (k-NN)
Probe vehicles
Loop detectors

ABSTRACT

Since the concept of the Macroscopic Fundamental Diagram (MFD) has been introduced, many studies have investigated the existence and characteristics of the MFD using empirical and simulation data. MFD is a powerful and efficient model for monitoring and managing large-scale urban networks. However, estimating the MFD for large-scale networks faces important challenges; monitoring resources are often limited in such networks. Furthermore, common sensors that are used to collect traffic data (i.e., loop detectors and probe vehicles), have limitations of their own. For instance, loop detectors are fixed sensors and cannot provide accurate density measurements. On the other hand, to estimate the MFD using probe vehicle data, the probe penetration rate must be known a priori. Given that the individual sensors cannot provide complete and accurate traffic measurements, combining the traffic data from multiple sources may improve the estimation of the MFD. The aim of this study is to combine two traffic data sources to estimate the MFD for a large-scale urban network, where the distribution of probe vehicles across the network is not necessarily homogeneous. The two traffic data sets used in this study are probe vehicle data with an unknown penetration rate, and full-scale approximate traffic data which is produced based on loop detector data. We compare the results of the fusion method with the results of a baseline method, which only uses loop detector measurements. The average flow and density estimations resulting from the Bayesian fusion method outperform the baseline method. We observe a particularly significant improvement in average density estimations, which reaffirms that loop detectors cannot accurately measure the average density.

1. Introduction

The increasing number of vehicles in today's large-scale urban networks has raised serious traffic congestion issues which makes monitoring and managing traffic a crucial task. A promising model to tackle this important challenge is the Macroscopic Fundamental Diagram (MFD) which shows the relationship between average flow and average density in large urban networks (Daganzo, 2007; Geroliminis and Daganzo, 2008). The existence and estimation of the MFD have been investigated in many studies (Johari et al., 2021). There are two main approaches for estimating the MFD: (i) data-driven approaches, and (ii) analytical approaches. Data-driven approaches typically use empirical data (Geroliminis and Daganzo, 2008; Buisson and Ladier, 2009; Shoufeng et al., 2013; Ambühl et al., 2018; Knoop et al., 2018; Loder et al., 2019; Huang et al., 2019; Ambühl et al., 2019; Mariotte et al., 2020; Ambühl et al., 2021) or simulation data (Ji et al., 2010; Knoop et al., 2014; Saberi et al., 2014; Du et al., 2016) to find the network average flow and network average density. These approaches employ either a single traffic data source (e.g., loop detector or probe vehicles)

* Corresponding author.
E-mail address: m.yildirimoglu@uq.edu.au (M. Yildirimoglu).

https://doi.org/10.1016/j.trc.2022.103555
Received 8 June 2021; Received in revised form 21 November 2021; Accepted 3 January 2022
Available online 16 February 2022

or a combination of multiple sources to estimate the MFD. On the other hand, analytical approaches estimate the MFD considering the network infrastructure and control parameters (i.e., signal settings) without relying on additional traffic data. Daganzo and Geroliminis (2008) have introduced the analytical MFD considering the principals of Variational Theory (VT) that was developed earlier by Daganzo (2005). The proposed method in this study, known as method of cuts (MoC), derives the upper bound of the MFD. Later, Leclercq and Geroliminis (2013) and Geroliminis and Boyacı (2012) further improved the method of cuts by incorporating the drivers' route choice and the effect of variability in topology and signal settings. Apart from the deterministic models presented by the aforementioned studies, Laval and Castrillón (2015) proposed a stochastic approximation approach based on MoC which extends the analytical derivation approach from corridors to networks. In a recent study, Tilg et al. (2020) compared MoC and the stochastic approximation method. The authors validated the two methods using empirical data and concluded that although the stochastic approximation approach estimates a more accurate free-flow branch, it fails to estimate the upper bound of the MFD as accurately as MoC. Note that the analytical approaches provide the demand-independent shape of the MFD, while data-driven approaches enable us to estimate the dynamic network average flow and density values, which is crucial for any real-time monitoring and control systems. This study develops a data fusion algorithm that makes use of both loop detector and probe vehicle data in order to estimate the MFD; therefore, it is an attempt in data-driven estimation with the aim of monitoring dynamic traffic conditions.

Many studies have proven the importance of the MFD in controlling large-scale urban networks. Some examples are, perimeter control in urban networks (Geroliminis et al., 2012; Keyvan-Ekbatani et al., 2015; Kouvelas et al., 2017; Keyvan-Ekbatani et al., 2019; Ingole et al., 2020; Li et al., 2021), regional route guidance (Yildirimoglu et al., 2015, 2018), pricing (Zheng et al., 2016; Gu et al., 2018), demand management (Yildirimoglu and Ramezani, 2020; Kumarage et al., 2021) and control of city-scale ride-sourcing systems (Ramezani and Nourinejad, 2018). Since collecting and processing adequate data in large-scale urban networks is costly, estimating the MFD using empirical data is not straightforward. Therefore, researchers have proposed different methodologies in order to identify the optimal amount of data along with the most important links from which to collect traffic data. For example, Keyvan-Ekbatani et al. (2013) proposed a reduced MFD concept where only a limited number of fixed sensors are employed for estimating the MFD. Ortigosa et al. (2014) and Zockaie et al. (2018) proposed a mathematical framework to find the best location and optimal amount of traffic data to minimize the estimation error between the estimated MFD and the ground-truth MFD. In a recent study, Saffari et al. (2020) applied Principal Component Analysis (PCA) to find the critical links where loop detectors should be installed. Using loop detector measurements and PCA, they succeeded to calculate link-flow and link-density values for all the links and to estimate the MFD. The model proposed by Saffari et al. (2020) is of particular interest in this study, as the proposed model will require initial estimations to be produced by this PCA-based model.

Loop detectors and probe vehicles are two common sources of traffic data that one can use to estimate MFDs. While loop detectors are appropriate tools to measure temporal variability in the network, given that they are fixed sensors, they cannot accurately capture the spatial variability. The location of a loop detector on a link or in a network could significantly change loop detector measurements (Buisson and Ladier, 2009; Courbon and Leclercq, 2011; Menendez et al., 2019). Therefore, relying on loop detector measurements for calculating the density may result in biased MFD estimations. Leclercq et al. (2014) proposed a correction method to minimize the density discrepancy when loop detectors are employed for MFD estimation purposes. However, their method does not significantly improve the results in congested parts of the network.

Being able to travel across the network, probe vehicles can however provide data from any location in the network; in other words, they capture traffic state variations while travelling across the network. Nevertheless, using probe vehicle trajectories to estimate MFDs requires simplifying assumptions. Since not all the vehicles travelling across the network are able to provide their trajectories, the average flow and average density calculated using probe vehicle trajectories are not representative of the entire traffic stream. To calculate the complete traffic measurements (i.e., flow and density representing the entire traffic stream), the proportion of probe vehicles (i.e., probe penetration rate) needs to be known a priori. For instance, Nagle and Gayah (2013) upscaled the probe vehicle measurements with an a priori penetration rate and investigated the accuracy of the resulting average flow and density estimations. Nonetheless, this approach requires a homogeneous distribution of probe vehicles across the network which is often not the case in large-scale real-world networks.

Each aforementioned traffic data source has advantages and disadvantages, and they could complement each other if they are employed in conjunction. Previous studies have shown that combining probe vehicle and loop detector data can improve the MFD estimations. Some studies (Leclercq et al., 2014; Tsubota et al., 2014; Beibei et al., 2016; Ambühl et al., 2017; Ji et al., 2018) simply used loop detector data to calculate network average flow, and employed probe vehicle data to calculate network average density. All these studies have confirmed that combining loop detector and probe vehicle data improves the accuracy of the estimated MFD compared with the approaches that rely on a single data source. Ambühl and Menendez (2016) proposed a fusion algorithm, using both loop detector and probe vehicle data simultaneously, to estimate the MFD. The proposed fusion algorithm calculates the weighted average of loop detectors and probe vehicle measurements with respect to their network coverage. The results demonstrated that if loop detector coverage stays unchanged, increasing the number of probe vehicles cannot significantly improve the estimations. Moreover, this method cannot yield an accurate MFD if loop detector coverage is low. Du et al. (2016) simulated a grid network with loop detectors emulated on the links and probe vehicles heterogeneously distributed across the network to estimate the MFD. They applied the k-means clustering method to calculate the probe penetration rate for each OD pair in the network using the existing loop detectors on the links. The authors mentioned that the number of clusters has a significant effect on the final results; thus, generalizing this method to other networks would be challenging. Lin et al. (2019) utilized a Back Propagation Neural Network (BPNN) fusion model to combine loop detector and probe vehicle data and estimate the MFD. They compared the accuracy of the estimated MFD with the result of their previous work (Lin and Xu, 2018) where they estimated the MFD by applying an Adaptive Weighted Averaging (AWA) fusion method. The results showed that the BPNN fusion model can

provide more accurate estimations than the AWA fusion model. Recently, in an empirical study, Fu et al. (2020) investigated the existence of 3D-MFDs in a large-scale urban network by fusing vehicle counts from fixed detectors and taxi GPS data. They proposed a partitioning method to divide the network into homogeneous sub-networks, and they achieved a low scatter MFD for the network.

Apart from the studies that focus on MFD estimation, there is an abundance of literature on data fusion methods. Guo et al. (2018) and Bachmann et al. (2013) have investigated different fusion methods and presented a comparison of these methods' performance. According to these studies, each method has its own strengths and drawbacks, and there is no absolute "best" fusion method. We opt for a Bayesian data fusion model to combine the two aforementioned traffic sources. Previous studies have proven that when the existing data is uncertain and imprecise, a Bayesian fusion approach could increase the accuracy of the estimations (Maskell, 2008; Castanedo, 2013). As explained earlier, loop detectors and probe vehicles both are subject to being uncertain, incomplete and imprecise. For instance, loop detectors, which are widely used in urban networks, are fixed sensor and cannot accurately capture the spatial variability on a link. Therefore, the density observations provided by them is biased. Probe vehicles, on the other hand, can provide dynamic traffic observations from across the network. However, estimating the penetration rate of probe vehicles particularly when they are not uniformly distributed within the network is challenging. Furthermore, a Bayesian framework has been reported to perform very well when the sample size is small (Muthén and Asparouhov, 2012). We later explain that the number of observations for each link in every time interval (for both traffic sources) is limited. This small number of observations is the input to the Bayesian model. Moreover, Bayesian analysis is not computationally as demanding as other fusion methods (Marcoulides, 2017); thus, it would be an efficient model for our problem, and can decrease cost of the computation.

Our aim in this study is to develop a data fusion method that takes advantage of both (limited number of) loop detectors and probe vehicles, which may or may not be homogeneously distributed in the network. This study builds on the premise that full-scale traffic data (i.e., covering all links in the network), albeit approximate, is available for the network, which is produced as a result of our earlier work (Saffari et al., 2020). Very briefly, the previous study identifies a small number of critical links in the network where loop detectors should be installed, and produces an approximation of true traffic variables (i.e., flow and density) for all links. Section 2 will provide further details on that previous study. In this paper, in addition to loop detector measurements from the critical links, we assume that real-time probe vehicle data with an unknown penetration rate is available. These two data sets are the inputs to our fusion algorithm. To summarize the contributions of this paper: (1) we propose a rigorous data fusion method based on Bayesian inference to estimate the MFD for a large-scale urban network given limited traffic data, (2) we investigate the effect of a heterogeneous probe vehicle distribution on the fusion algorithm and in turn, the estimated MFDs, (3) we demonstrate that taking advantage of both probe vehicle and loop detector observations significantly improve the density estimations.

The rest of the paper is organized as follows. Section 2 briefly explains the methodology to calculate the approximate full-scale traffic data. Section 3 consists of three subsections, which all together describe the proposed methodology in detail. Section 4 presents and discusses the results from the proposed methodology. Finally, Section 5 provides concluding remarks.

2. Approximate full-scale traffic data

In this study, we propose a methodology to combine two traffic data sources, namely approximate full-scale link traffic data and real-time probe vehicle data. The former is the result of our earlier study. In this section, we summarize the main steps of the methodology to build the approximate full-scale link traffic data; a detailed framework along with the evaluation of the methodology are presented in Saffari et al. (2020).

In order to produce the approximate full-scale traffic data, the following initial assumptions are made: (i) no loop detector exists in the initial network scenario, (ii) historical probe vehicle trajectories with an unknown probe penetration rate are available, (iii) probe vehicles, which provide historical trajectories, are uniformly distributed across the network.

Fig. 1 presents the flowchart of the methodology to produce the approximate full-scale link traffic data. First, using the historical probe vehicle trajectories we calculate link-flow and link-density values. The next step is to apply Principal Component Analysis (PCA), which is essentially a powerful feature selection tool with the ability of reconstructing the original data after reducing its dimensions. Using PCA, we are able to identify major traffic patterns across the network. We associate each major traffic pattern to a link in the network which we call a critical link. Loop detectors are to be installed on these links and provide flow and density measures. Note that we consider different numbers of critical links (i.e., N), ranging from 20 to 120 links on a network of 1260 links.

The other output of the PCA model is the reconstruction parameters that are used to approximate or reconstruct the original data set out of the principal components. Using the loop detector measurements (flow and density) and the reconstruction capability of PCA, we calculate the approximate full-scale traffic data. We call the resulting data set 'approximate', because in order to reconstruct the original data, we only use the measurements from a limited number of links (i.e., the critical links). Later, in this paper, we use the approximate full-scale traffic data that is reconstructed using 20, 40, 60 and 80 critical links. Note that the results from our previous analysis show that the estimation performance is very similar from 80 links to 120 links; hence, we have limited the analysis in this paper to a maximum of 80 links. In the remainder of this paper, we refer to our earlier work of Saffari et al. (2020) as the PCA-based model, and use it as a baseline for the fusion model that we develop here.

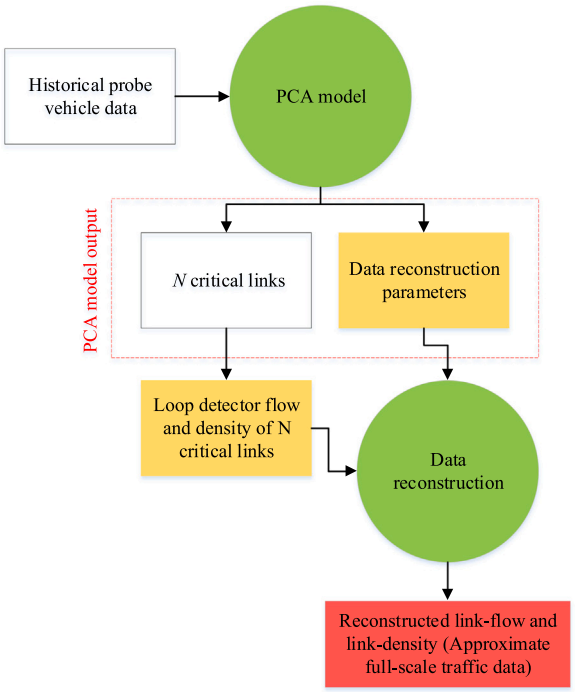


Fig. 1. Methodology to calculate the approximate full-scale traffic data.

3. Methodology

In this section, we present the proposed methodology to combine real-time probe vehicle data, with an unknown probe penetration rate, and approximate full-scale traffic data. Applying the proposed fusion method, we can calculate link-flow and link-density for all the links in the network. Fig. 2 presents the main steps of the proposed methodology. There are two main parts shown in the flowchart; (1) calculating the local penetration rates, and (2) applying the Bayesian fusion method. In the first part, we start with probe vehicle and loop detector observations on the critical links; this allows us to calculate the penetration rate for each critical link where a loop detector is installed. We then find the k-nearest critical links for each link in the network, and calculate the average penetration rate of these k links. This allows us to estimate a local penetration rate for each link, which may vary across the network. In the second part of the algorithm, we upscale probe vehicle observations, applying the estimated local penetration rates. This data is one of the inputs to the Bayesian fusion model. As shown in Fig. 2, two traffic sources (i.e., approximate full-scale traffic data and upscaled probe vehicle measurements) are combined applying the proposed Bayesian data fusion model. The output of the model is fused link-flow and link-density values which we later use to estimate the MFD for the network. We thoroughly describe each component of the flowchart in the following subsections.

3.1. Local probe penetration rate

Considering the available probe vehicle trajectories, we calculate partial link-flow $\tilde{q}_i(t)$ and partial link-density $\tilde{k}_i(t)$. To do so, we apply Edie's generalized definitions (Eq. (1)). Note that since we only use the probe vehicle trajectories to calculate link-flow and link-density (not all the vehicle trajectories), these values are partial. These values will later be upscaled based on the estimated local penetration rates.

$$\tilde{q}_i(t) = \frac{\sum_p d_{ip}(t)}{n_i l_i t} \qquad \tilde{k}_i(t) = \frac{\sum_p t_{ip}(t)}{n_i l_i t} \tag{1}$$

where $\tilde{q}_i$ and $\tilde{k}_i$ are respectively partial flow and partial density of link i, with the length l_i and number of lanes n_i, in a time interval t. Additionally, d_{ip} and t_{ip} denote the distance travelled and time spent by vehicle p on link i.

To estimate the complete link measures (i.e., representing the entire stream travelling on a link), the probe penetration rate has to be known a priori, which is one of the challenges when relying on probe vehicle observations. Network level analysis using probe vehicle measures usually assumes a homogeneous distribution of probe vehicles across the network; many studies have considered one single penetration rate in order to upscale partial probe vehicle measurements to complete measurements. This assumption may not be realistic in practice. Given the heterogeneous distribution of probe vehicles in real-world networks, finding the complete

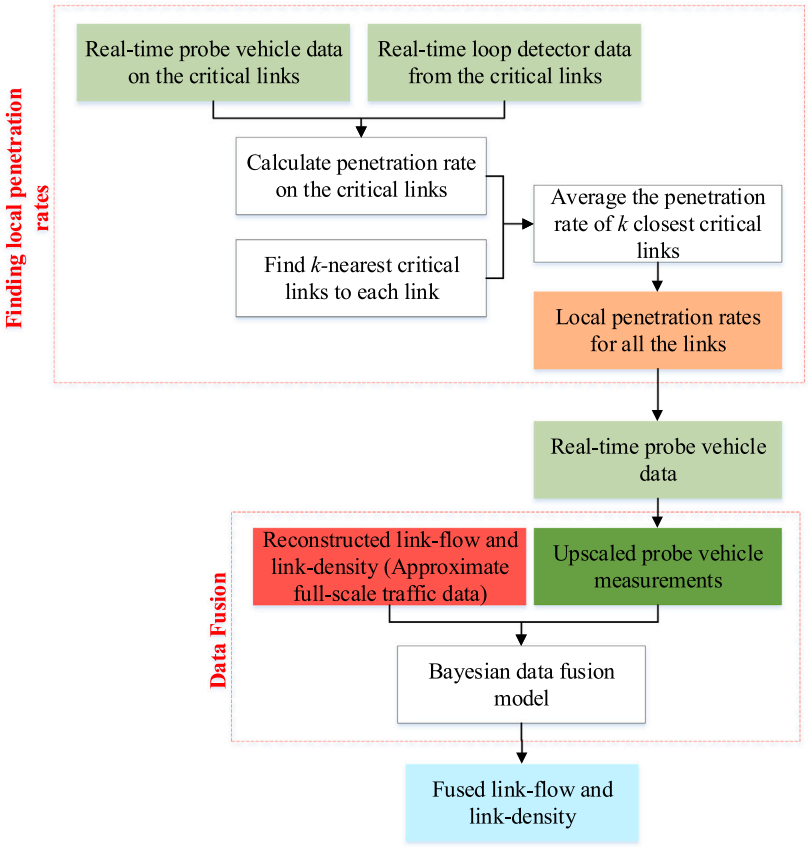


Fig. 2. Proposed methodology.

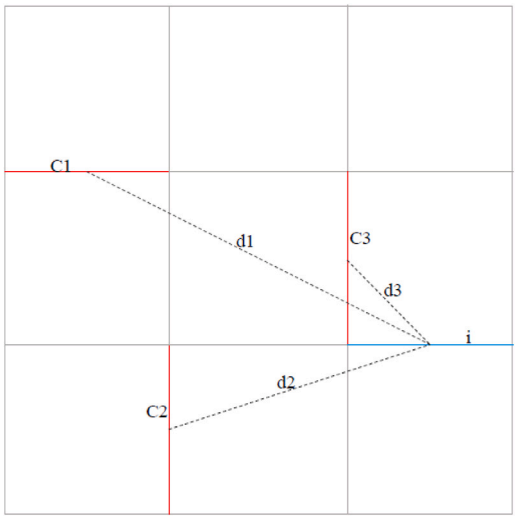


Fig. 3. Example of finding the (k = 3) closest critical links.

traffic measurements by applying a single penetration rate is unlikely to be successful. In the following paragraphs, we address the aforementioned challenges by introducing a method to estimate local penetration rates.

In order to account for a heterogeneous probe vehicle distribution with an unknown penetration rate, we propose a methodology to calculate local probe penetration rates (i.e., variable penetration rate) across the network. We assume that probe penetration rate

is a local parameter; in other words, it is locally uniform. While the penetration rate can vary across the network due to different travel patterns of probe vehicles, we assume that it will remain approximately constant in local areas and it will vary slowly from one link to another. To calculate the local penetration rates, we rely on the loop detectors that are installed on the critical links. Obviously, these are the only links in the network where we collect both probe vehicle and loop detector data. Dividing the probe vehicle flows ($\tilde{q}_i$), i.e. partial link flows, on critical link i by the loop detector flows (q_i^l), i.e. full link flows, on critical link i, we can approximate the probe penetration rate on each critical link, $\rho_i = \sum_t \tilde{q}_i(t) / \sum_t q_i^l(t)$. Given the penetration rate of each critical link is known, and based on the assumption that penetration rate is a local parameter, we can approximate the penetration rate of each link in the network considering its nearby or local critical links.

To identify the local critical links associated to each link, we apply k-nearest neighbour (k-NN) algorithm which basically finds the closest critical links for every link. We explain this on a simple grid network with three critical links (or three links with loop detectors) as shown in Fig. 3. In this network, $C1$, $C2$ and $C3$ are the critical links, and the task is to estimate the penetration rate for link i. The first step is to calculate the euclidean distance between the midpoint of link i and all the critical links (i.e., d_1, d_2, d_3) and sort them in ascending order (i.e., $d_3 < d_2 < d_1$). Penetration rate of link i can be approximated by averaging the penetration rate of k nearest critical links. The same procedure is applied for all the links to find their local penetration rates. We further discuss how to find the value of k for our case study in Section 4.

The procedure above allows us to estimate a penetration rate for each link in the network, ρ_i. We upscale the partial link flows and densities, $\tilde{q}_i(t)$ and $\tilde{k}_i(t)$, to the complete probe-based link flows and densities, $q_p^i(t)$ and $k_p^i(t)$, using the following relations; $q_p^i(t) = \tilde{q}_i(t) \cdot \rho_i$ and $k_p^i(t) = \tilde{k}_i(t) \cdot \rho_i$ for $\forall i, t$. These complete probe-based measurements will compose one of the two data sources for the data fusion algorithm that is presented in the next subsection.

3.2. Bayesian data fusion

The data fusion method that we adopt in this paper is based on Bayesian inference. This method uses Bayes theorem which combines prior information (or belief) about a parameter with the evidence from the information contained in a sample (or observed data). Bayesian inference has been long used in the literature for different purposes. The reason for its popularity is that it is informative and easy to interpret since it returns a probability of possible values for the unknown parameter given the observed data. Therefore, it has a natural way to incorporates the idea of confidence for our estimates. Moreover, Bayesian allows us to update our prior knowledge as we observe new data. For example, if we apply a Bayesian data fusion model for our network, this could be the prior for the real-time data that we observe in the future.

According to Bayes theorem:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \tag{2}$$

where $P(X|\theta)$ is the probability of the data given the prior (likelihood), $P(\theta)$ is the prior information, $P(X)$ is the probability of the data and finally, $P(\theta|X)$ is the posterior, or in other words, the updated belief after the data (or evidence) is observed. Basically, once we observe the data, we are able to update the prior information considering the new information that we obtain.

As mentioned, in our problem, there are two sets of data, real-time probe vehicle and approximate full-scale traffic data, that we aim to combine by applying a Bayesian data fusion model. Let $\mathbf{q}_a^{i,t}, \mathbf{q}_p^{i,t}, \mathbf{k}_a^{i,t}$ and $\mathbf{k}_p^{i,t}$ denote approximate flow, flow based on probe vehicles, approximate density and density based on probe vehicles of link i in time of day t, respectively.

$$\begin{aligned}
\mathbf{q}_a^{i,t} &= \{q_a^{1i}(t), q_a^{2i}(t), \dots, q_a^{n_a i}(t)\} \\
\mathbf{q}_p^{i,t} &= \{q_p^{1i}(t), q_p^{2i}(t), \dots, q_p^{n_p i}(t)\} \\
\mathbf{k}_a^{i,t} &= \{k_a^{1i}(t), k_a^{2i}(t), \dots, k_a^{n_a i}(t)\} \\
\mathbf{k}_p^{i,t} &= \{k_p^{1i}(t), k_p^{2i}(t), \dots, k_p^{n_p i}(t)\}
\end{aligned} \tag{3}$$

where n_a and n_p are the number of approximate traffic data observations and the number of probe vehicle observations, respectively. Let us assume M days of traffic measurements are available (i.e., M replications simulated in Aimsun). Thus, the number of observations on each link in each time of day interval can vary between zero and M ($0 \le n_a, n_p \le M$).

We can re-write Eq. (2) for our problem as follows:

$$P(\mu_q^{i,t}|\mathbf{q}_a^{i,t}\mathbf{q}_p^{i,t}) = \frac{P(\mathbf{q}_a^{i,t}, \mathbf{q}_p^{i,t}|\mu_q^{i,t})P(\mu_q^{i,t})}{P(\mathbf{q}_a^{i,t}, \mathbf{q}_p^{i,t})} \qquad P(\mu_k^{i,t}|\mathbf{k}_a^{i,t}, \mathbf{k}_p^{i,t}) = \frac{P(\mathbf{k}_a^{i,t}, \mathbf{k}_p^{i,t}|\mu_k^{i,t})P(\mu_k^{i,t})}{P(\mathbf{k}_a^{i,t}, \mathbf{k}_p^{i,t})} \tag{4}$$

where $P(\mu_q^{i,t}|\mathbf{q}_a^{i,t}\mathbf{q}_p^{i,t})$ and $P(\mu_k^{i,t}|\mathbf{k}_a^{i,t}, \mathbf{k}_p^{i,t})$ denote the posterior distributions defining the probability of fused flow and fused density of link i in time t given the two defined data sets, respectively. $P(\mathbf{q}_a^{i,t}, \mathbf{q}_p^{i,t}|\mu_q^{i,t})$ and $P(\mathbf{k}_a^{i,t}, \mathbf{k}_p^{i,t}|\mu_k^{i,t})$ are the likelihood, that is, conditional probability of the data (flow and density values) given the value of $\mu_q^{i,t}$ and $\mu_k^{i,t}$. Lastly, $P(\mu_q^{i,t})$ and $P(\mu_k^{i,t})$ denote prior distributions of fused link-flow and fused-link density, respectively. We omit i and t in the notation in the following equations for the sake of brevity.

Note that to avoid duplication, we write the equations only for flow calculations. The same formula will be applied for density to calculate the fused density values for the network links. Here, we assume that $P(\mu_q)$, $P(\mathbf{q}_a|\mu_q)$ and $P(\mathbf{q}_p|\mu_q)$ follow normal distributions,

$$P(\mu_q) \propto \exp\left\{-\frac{1}{2\sigma_0^2}(\mu_q - \mu_0)^2\right\} \propto N(\mu_q|\mu_0, {\sigma_0}^2) \tag{5}$$

$$P(\mathbf{q}_a|\mu_q) \propto \exp\left\{-\frac{1}{2\sigma_a^2}\sum_{r=1}^{n_a}(q_a^r-\mu_q)^2\right\} \propto N(q_a^r|\mu_q,{\sigma_a}^2) \tag{6}$$

$$P(\mathbf{q}_p|\mu_q) \propto \exp\left\{-\frac{1}{2\sigma_p^2}\sum_{s=1}^{n_p}(q_p^s-\mu_q)^2\right\} \propto N(q_p^s|\mu_q,{\sigma_p}^2) \tag{7}$$

where μ_0 and σ_0^2 denote the mean and variance of the prior distribution, respectively. ${\sigma_a}^2$ and ${\sigma_p}^2$ denote the variance of reconstructed flow and real-time flow observations, respectively.

Plugging in Eqs. (5)–(7) into Eq. (4), the probability of fused flow given the observed flow data can be calculated using the formula below:

$$P(\mu_q|\mathbf{q}_a,\mathbf{q}_p) \propto \exp\left\{-\frac{1}{2}\left[\sum_{r=1}^{n_a}\left(\frac{q_a^r-\mu_q}{\sigma_a}\right)^2+\sum_{s=1}^{n_p}\left(\frac{q_p^s-\mu_q}{\sigma_p}\right)^2+\left(\frac{\mu_q-\mu_0}{\sigma_0}\right)^2\right]\right\} \tag{8}$$

When the prior and the observed data both follow normal distributions, the posterior (in this case, the probability of the fused data given the observed data) also follows a normal distribution with mean and variance of μ_f and σ_f^2 as follows:

$$P(\mu_q|\mathbf{q}_a,\mathbf{q}_p) \propto \exp\left\{\frac{1}{2\sigma_f^2}(\mu_q-\mu_f)^2\right\} \propto N(\mu_f,{\sigma_f}^2) \tag{9}$$

Combining Eqs. (8) and (9) (see the Appendix for derivation of Eq. (10)), we can find the mean of fused link-flow (μ_f) which is given as:

$$\mu_f = \frac{1}{\frac{n_a}{\sigma_a^2}+\frac{n_p}{\sigma_p^2}+\frac{1}{\sigma_0^2}}\left(\frac{\sum_{r=1}^{n_a}q_a^r}{\sigma_a^2}+\frac{\sum_{s=1}^{n_p}q_p^s}{\sigma_p^2}+\frac{\mu_0}{\sigma_0^2}\right) \tag{10}$$

As mentioned earlier, we omit the link i and time t from the equations for simplification purposes. In other words, by applying Eq. (10), we calculate the fused flow value on link i in time interval t. Therefore, to find link flow values for all links in every time interval, Eq. (10) needs to be applied $N_{tot}\times T$ times, where N_{tot} is the total number of links in the network and T is the total number of time intervals. Once link-flow and link-density values are calculated, we can find the MFD parameters of network average flow, $Q(t)$, and network average density, $K(t)$, using the following formulas:

$$Q(t)=\frac{\sum q_i(t)l_i}{\sum l_i} \qquad K(t)=\frac{\sum k_i(t)l_i}{\sum l_i} \tag{11}$$

where $q_i(t)$ and $k_i(t)$ are fused link-flow and density measurements from link i in time interval t (i.e., μ_f calculated in Eq. (10)), respectively. The length of each link is denoted by l_i.

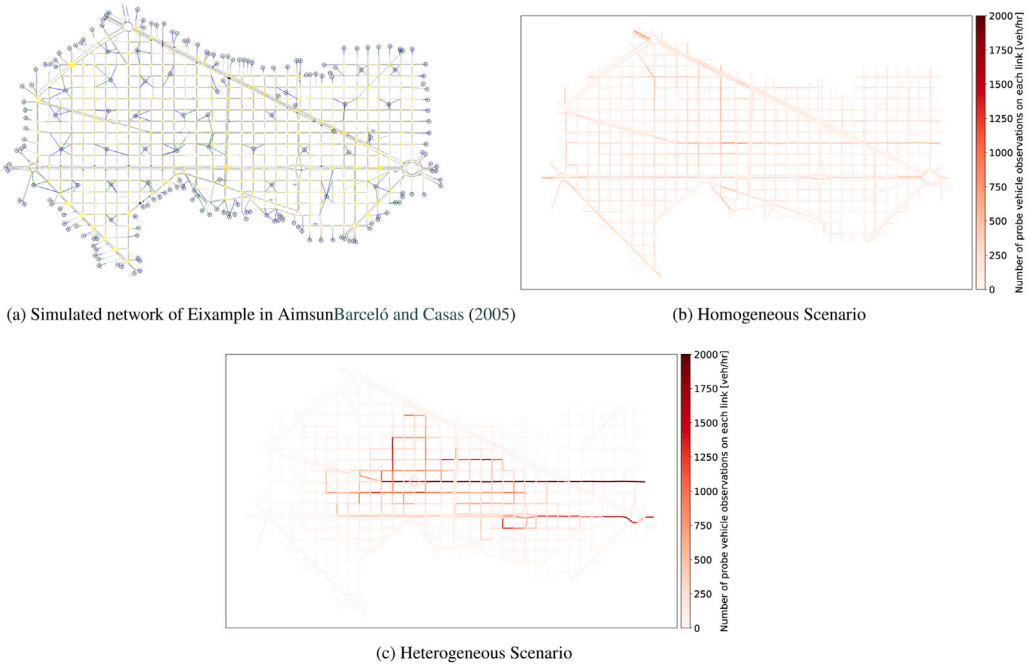
4. Results and discussions

4.1. Network and scenario description

This section consists of three subsections. In the first subsection, we investigate the optimum number of nearest neighbours (i.e., the value of k in k-NN) to find the local penetration rates for every link in the network. In the second subsection, we present the estimated MFDs resulting from the proposed Bayesian data fusion method considering the optimum k. Finally, in the third subsection, we explore the effect of homogeneity/heterogeneity in probe vehicle distribution on the final results.

The network of the study is a large-scale urban network of Eixample district in Barcelona, Spain, which is modelled in Aimsun, a well-known traffic simulation package; see Fig. 4(a) (Barceló and Casas, 2005). The network consists of 1260 links and 712 signalized and un-signalized intersections.

In order to generate probe vehicles and extract their trajectories, we use Aimsun API (Application Programming Interface) in a micro-simulation environment. The simulation period represents a 90 minute morning peak time. We consider seven replications of the explained micro-simulation model, representing '7 days' (i.e. the maximum number of observations is $M = 7$). We investigate two scenarios; homogeneous and heterogeneous probe vehicle sampling. In the homogeneous scenario, we randomly sample 10% of the vehicles from the entire network as probe vehicles. In the heterogeneous scenario, we only sample the vehicles travelling between specific OD pairs. We select six OD pairs with the largest number of vehicles travelling between them. The total number of vehicles travelling between the selected OD pairs is roughly 10% of the total number of vehicles in the network. The reason for choosing the six particular OD pairs in the first scenario is that we need the least number of OD pairs that produce roughly the same percentage of vehicles. Having fewer OD pairs to sample the probe vehicles means we only sample from limited parts of the network, which leads us to a heterogeneous sample. In fact, this resulted in possibly the most heterogeneous scenario that one might expect, which creates exceptionally challenging settings for our algorithm. Figs. 4(b) and 4(c) show the number of probe observations on the links reflecting the homogeneous and heterogeneous probe vehicle distribution across the network, respectively. As shown in the figures, vehicles are uniformly distributed in the homogeneous scenario, while observations across the network vary significantly in the heterogeneous scenario. We further investigate these two scenarios in relation to the estimation quality of the MFD in Section 4.5.



(a) Simulated network of Eixample in AimsunBarceló and Casas (2005)

(b) Homogeneous Scenario

(c) Heterogeneous Scenario

Fig. 4. Probe vehicle distribution in the simulated network.

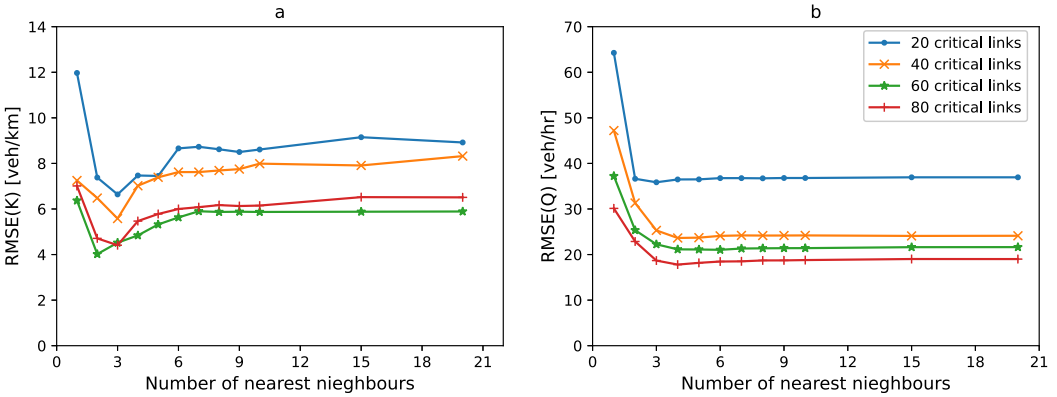


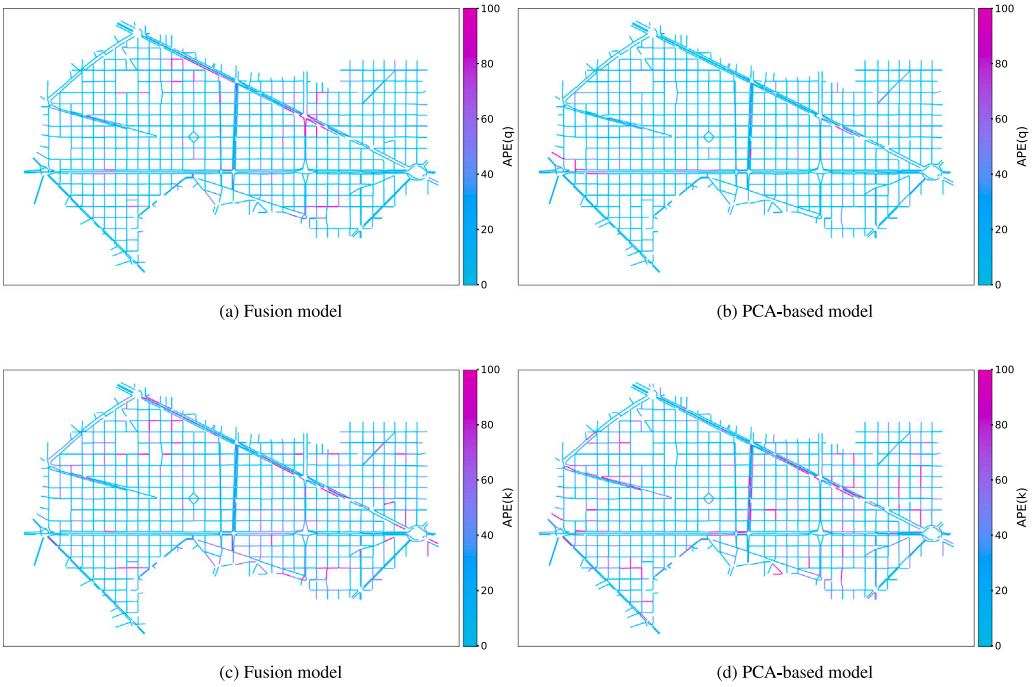
Fig. 5. (a) Estimation error of average density, (b) Estimation error of average flow with respect to k.

4.2. Determining the optimum k

As described in Section 3, each link in the network is matched with k-nearest critical links, and the average penetration rate in those k links is used to upscale the probe measurements. Nonetheless, the value of k is unknown and has to be determined after a careful analysis. For this purpose, we first find the local penetration rates for each link with respect to varying values of k and upscale real-time probe vehicle observations subsequently. Applying the fusion algorithm for each value of k, we calculate Root Mean Square Error for network average flow (RMSE(Q)) and network average density (RMSE(K)) for T time intervals during simulation (Eq. (12)).

$$\mathrm{RMSE}(Q) = \sqrt{\frac{\sum_{t=1}^{T}(\hat{Q}(t) - Q(t))^2}{T}} \qquad \mathrm{RMSE}(K) = \sqrt{\frac{\sum_{t=1}^{T}(\hat{K}(t) - K(t))^2}{T}} \tag{12}$$

where $Q(t)$ and $K(t)$ stand for the estimated average network flow and the estimated average network density in time interval t (see Eq. (11)), respectively; and $\hat{Q}(t)$ and $\hat{K}(t)$ are the ground-truth flow and density in time interval t, respectively. To calculate the ground-truth, we use all vehicle trajectories and apply Edie's generalized definition (Eq. (1)). Once we have the complete link-flow and link-density, applying Eq. (11), we can aggregate the link-level values and calculate the ground-truth MFD.



(a) Fusion model (b) PCA-based model

(c) Fusion model (d) PCA-based model

Fig. 6. Link-flow, (a) and (b), and link-density, (c) and (d), estimations.

Fig. 5 presents the estimation errors for average flow and average density with respect to changing values of k in four scenarios where the number of critical links varies from 20 to 80. Real-time probe vehicle trajectories are from the heterogeneous scenario, as presented in Fig. 4(c). Note that the value of k influences the estimated local penetration rate for each link, and essentially the upscaled real-time probe vehicles data and the final MFD estimations. As we can see in Fig. 5, adding more neighbours up to three improves both flow and density estimations in almost all critical link scenarios; however, after that, RMSE(K) increases and RMSE(Q) stays flat. Thus, we opt for three nearest neighbours since adding more neighbours will not produce better estimations. Note that the results presented in the following subsections is based on $k = 3$, the optimum number of closest critical links to find the local penetration rates.

4.3. Bayesian data fusion results

In this subsection, we investigate the performance of the proposed fusion algorithm based on a heterogeneous probe vehicle distribution (see Fig. 4(c)) and different subsets of critical links. Note that the number of critical links shows the number of links with loop detectors placed on them. We explore subsets of 20, 40, 60 and 80 links that represent approximately 2%, 3%, 5% and 7% of the links, respectively. For each subset of critical links, we first apply k-NN to find the three nearest critical links for all the links in the network. Then, the penetration rate on each link is calculated by averaging the penetration rate of the three nearest critical links. Incorporating the penetration rates, we can upscale partial probe vehicle observations to complete traffic measurements (link-flow and link-density). Note that this is obviously an approximation of the complete traffic measurements. The upscaled real-time probe vehicle data set is one of the two inputs to the Bayesian fusion algorithm. The second input, as explained before, is the approximate full-scale traffic data. To produce this data, we apply the PCA method, as explained in Section 2. The next step is to apply the Bayesian data fusion method (Eq. (11)) and combine the two aforementioned data sets and calculate the fused link-flow and link-density values. Fig. 6 illustrates the absolute percentage errors in the link-level flow and density estimations applying the fusion model (see Figs. 6(a) and 6(c)) and the PCA-based model (see Figs. 6(b) and 6(d)). The estimation error of each link is calculated using the following equation:

$$\mathrm{APE}_i(q) = \frac{|\hat{q}_i - \bar{q}_i|}{\hat{q}_i} \times 100 \qquad \mathrm{APE}_i(k) = \frac{|\hat{k}_i - \bar{k}_i|}{\hat{k}_i} \times 100 \tag{13}$$

where $\bar{q}_i$ and $\bar{k}_i$ denote the averages of the estimated flow and density values on link i throughout the simulation (i.e., the average of 1 min measurements across the 90 min simulation), respectively; $\hat{q}_i$ and $\hat{k}_i$ denote the averages of the ground-truth flow and density values on link i throughout the same period, respectively; $\mathrm{APE}_i(q)$ and $\mathrm{APE}_i(k)$ are the absolute percentage errors in flow and density estimations on link i, respectively.

Comparing Figs. 6(c) and 6(d), we can clearly observe that the fusion model outperforms the PCA-based model when estimating link-density values. Nevertheless, given there are many links in the network, comparing the performance of the two models is quite

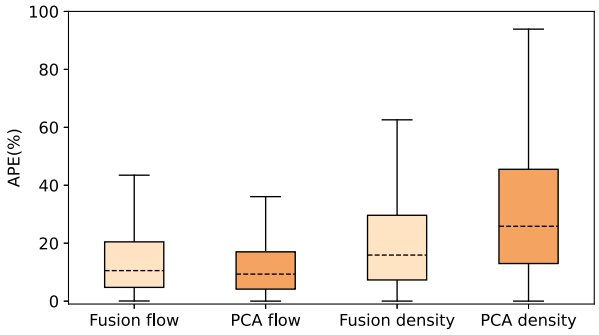


Fig. 7. Box-plots of link level estimation errors.

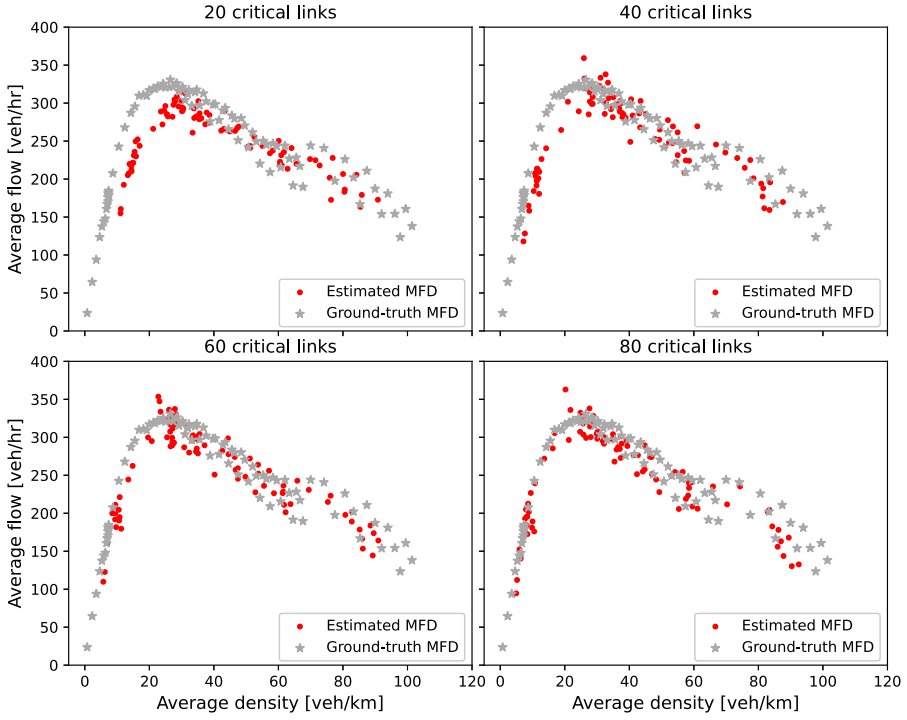


Fig. 8. Estimated MFDs based on different subsets of critical links.

Challenging. Particularly, the difference between the two models in terms of the link-flow estimations is not as apparent, please see Figs. 6(a) and 6(b). To further analyse the results, we plot the link level estimations in the form of box-plots shown in Fig. 7. While the error range of estimated fused link flow values is greater than the error range of the PCA-based model, the median errors of the two models are very close. This shows that the performance of the fusion and the PCA-based model in estimating the link-flow values is quite similar. Moreover, the box-plots confirm the conclusion regarding the density estimations; that is, the fusion model provides more accurate link-density estimations in comparison to the PCA-based model.

The next step after finding the link level estimations is to use Eq. (11) to find the network average flow and network average density and estimate the MFDs using this fusion method. Fig. 8 illustrates the estimated MFDs with respect to different subsets of critical links along with the ground-truth MFD which is calculated from the trajectories of all vehicles in the network. One expected observation is that having more critical links, which subsequently means having more loop detectors in the network, results in a better fit and less scatter. Having more loop detectors spread out in the network leads to better local penetration rate estimations which essentially improves the MFD estimations.

To evaluate our estimations, we compare the estimated network average flow and network average density with the ground-truth values by applying Eq. (12). The results of this calculation are presented in Table 1. This table also compares the estimation errors of the Bayesian fusion method and the PCA-based method (see Section 2) applied in our earlier study.

As we can see in Table 1, the Bayesian data fusion method improves the average flow and the average density estimations in most of the scenarios. Although we do not see a significant difference between RMSE(Q) from the fusion and the PCA-based method, we

Table 1
Estimation errors with respect to different subsets of critical links.

	Bayesian data fusion		PCA-based method		Percentage improvement (%)	
# Critical links	RMSE(K) [veh/km]	RMSE(Q) [veh/hr]	RMSE(K) [veh/km]	RMSE(Q) [veh/hr]	RMSE(K)	RMSE(Q)
20	6.65	35.86	12.30	37.00	46	3
40	5.58	25.30	7.27	27.19	23	7
60	4.52	22.25	7.23	22.66	37	2
80	4.41	18.68	5.96	18.75	26	0

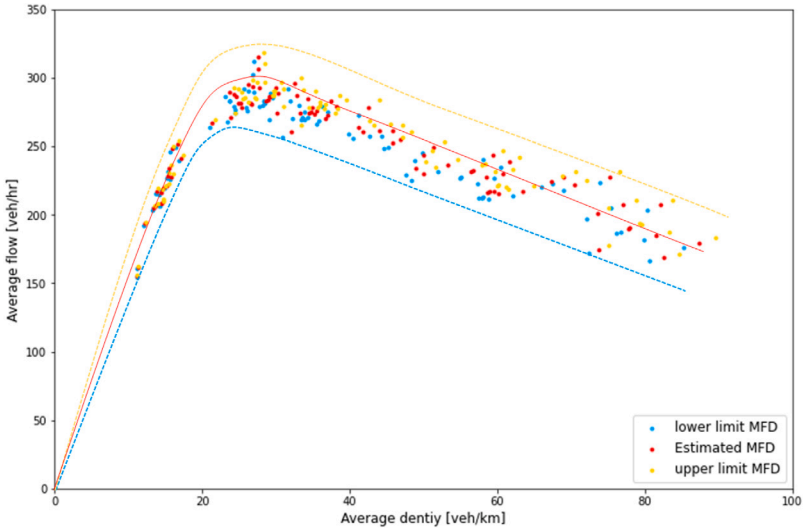


Fig. 9. Lower and upper limit of the MFD wit respect to 95% credible interval.

clearly observe a great improvement in all RMSE(K) values when applying the fusion model. The improvement in RMSE(K) ranges from 23% to 46%, whereas the improvement in RMSE(Q) is at most 7%. Note that, as we explained in Section 2, loop detector observations form the basis of the approximate full-scale traffic data; real-time measurements from the loop detectors are fed to the PCA-based algorithm to identify the major patterns and to approximate full-scale traffic flow and density measurements. While the resulting flow estimations are fairly accurate and up to par with the fusion algorithm, the density estimations are significantly worse. We discussed in Section 1 that loop detector observations cannot provide accurate density estimations since they are fixed sensors, and the location of loop detectors on a link can significantly change the estimations. Therefore, capturing different traffic patterns with only loop detector observations is challenging. This is why incorporating probe vehicle trajectories applying the fusion method significantly improves the unreliable loop detector density measurements. Note that loop detectors are valuable since they can provide the total number of vehicles travelling through a particular link and in turn enables the estimation of the probe penetration rate. Moreover, they can provide fairly accurate flow measurements, whereas probe vehicles can only measure partial flow and density. Additionally, one possible reason that we do not see a considerable improvement in flow estimations, when incorporating probe vehicle observations, is that loop detector measurements do not significantly differ from the probe vehicle measurements. In other words, adding probe vehicle measurements to the loop detector measurements may not provide more information about the traffic state on the links.

As mentioned earlier, Bayesian inference inherently provides information with respect to the uncertainty of the estimations. In addition to the point estimates, Bayesian inference produces a credible interval within which estimations fall with a particular probability. This is in fact one of the advantages of Bayesian inference since it returns a distribution (the posterior distribution) rather than just a single value. This makes interpreting the confidence of estimated values much easier. Here, we present the 95% credible interval for average flow and average density. Note that the Bayesian data fusion model is applied for each link and each time interval separately, which returns a posterior distribution for each value. To calculate the lower and upper limit in the MFD, we first find the lower and upper limits for each link from its associated posterior distribution. Then, applying Eq. (11), we calculate the network average flow and density. Fig. 9 presents the uncertainty in the estimated MFD using 20 critical links. To better interpret the results, we draw the lower and upper envelopes. The figure clearly shows that our estimations become more uncertain when the network moves to the congested state. In the free flow branch, we see only a slight difference between the lower and upper bound of the MFDs. The reason is that the variance of the observations is greater in the congested branch, which causes uncertainty in the data and consequently in the estimations.

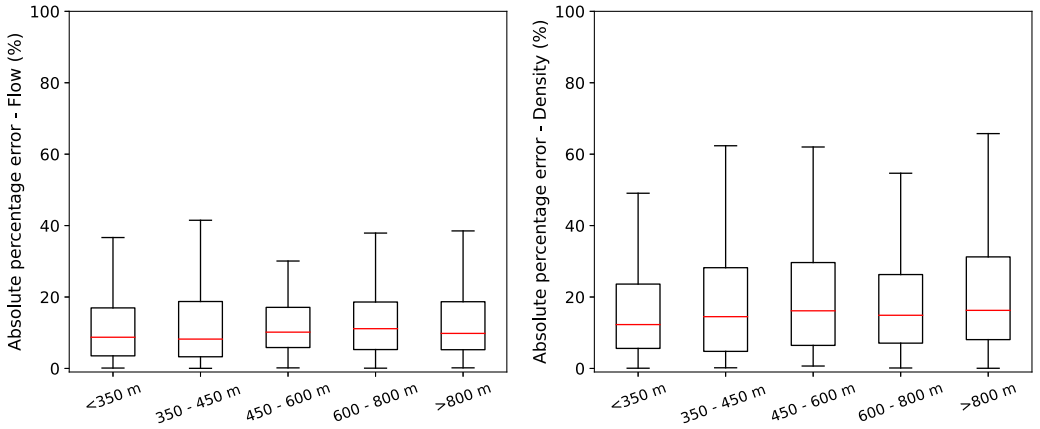


Fig. 10. Estimation error with respect to average distance between each link and its nearest critical links.

4.4. Investigating the sensitivity of the estimations to the distance between the links and the critical links

To investigate the sensitivity of the estimations to the distance between the links and their nearest critical links, we consider the scenario where we have 20 critical links. Since this scenario has the least number of critical links, it is more likely to exhibit a high degree of variation in the distance between links and critical links. We calculate the absolute percentage error for average flow and average density of each link over the simulation period. Fig. 10 presents the results of this calculation with respect to average distance between each link and its nearest three closest critical links. To evaluate the sensitivity of the estimation error, we divide the average distance into five groups so that we roughly have the same number of observations in each group. As shown in Fig. 10, for both flow and density, we do not observe a specific trend in the estimation errors. There seems to be no apparent relation between percentage error and the distance between links in relation to the flow estimation. On the other hand, average density estimation error increases with increasing distance up to 600 m, and decreases when the average distance is between 600 m to 800 m and beyond 800 m. In summary, a slight increase in estimation errors can be observed for the links that have a longer average distance to their nearest critical link. Nevertheless, the accuracy of the estimations is not highly dependent on this average distance parameter.

4.5. Comparing homogeneous and heterogeneous probe vehicle distributions

The results presented in Section 4.3 show that the proposed model performs well in a heterogeneous probe vehicle distribution scenario. The aim in this subsection is to explore the sensitivity of the proposed model to the changing features of the probe vehicle distribution. In particular, we investigate the impact of homogeneous and heterogeneous probe vehicle distribution on the final estimation results. As explained earlier, we sample a portion of vehicles as the probe vehicles using two sampling methods, homogeneous and heterogeneous sampling. Fig. 4(b) shows the homogeneous sampling method where distribution of probe vehicles across the network is homogeneous (uniform), while Fig. 4(c) presents heterogeneous sampling method. As shown in Fig. 4(c), probe vehicles do not exist on all links in the network, nor is the probe penetration rate similar across the network.

As mentioned earlier, the results presented in the previous subsection are based on a heterogeneous probe vehicle subset where the total number of probe vehicles is roughly 10% of all vehicles in the network. Here, to be able to compare the two scenarios, we uniformly sample 10% of the total vehicles and use them as probe vehicles. Then, we apply the same procedure that is explained in Section 3 in order to find the local penetration rates and prepare the real-time probe vehicle data for the fusion model.

While comparing the homogeneous and heterogeneous scenarios, we also investigate the effect of applying local penetration rates across the network. For this purpose, we apply the proposed methodology for both scenarios, assuming only one single penetration rate in the whole network (i.e., uniform network penetration rate). This uniform penetration rate, which is calculated by averaging the penetration rate of all critical links, is applied to upscale the partial probe vehicle observations in the entire network. Once the real-time probe vehicle data is upscaled, we apply the Bayesian data fusion model to find the fused flow and density values. This analysis will expose the added value resulting from the estimation of local penetration rates. Clearly, in the homogeneous scenario, the penetration rates are not expected to significantly fluctuate across the network. Hence, the estimation of local penetration rates may not be crucial. Nonetheless, this (homogeneous) scenario will serve as a baseline where we do not expect a significant change as a result of local penetration rates.

To evaluate the results, we consider scenarios with different number of critical links (i.e., 20, 40, 60 and 80 critical links), and use Eq. (12) to calculate RMSE(Q) and RMSE(K). Figs. 11 and 12 illustrate density estimation error and flow estimation error with respect to different critical links scenarios, respectively. Each figure compares the estimation error of three methods, that is, (1) PCA-based method, (2) fusion method with local penetration rate and (3) fusion method with a uniform network penetration rate. There are some important insights regarding the figures which we discuss in the following paragraphs.

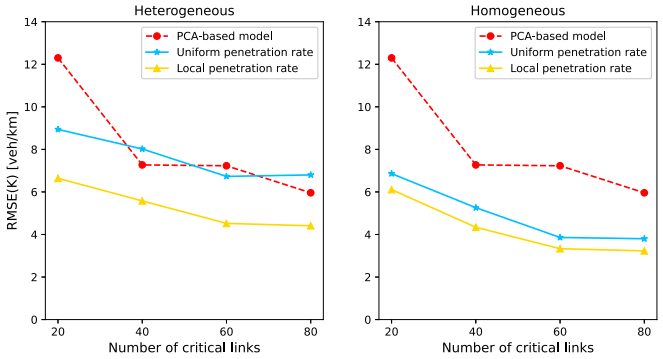


Fig. 11. Density estimation error with respect to number of critical links. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

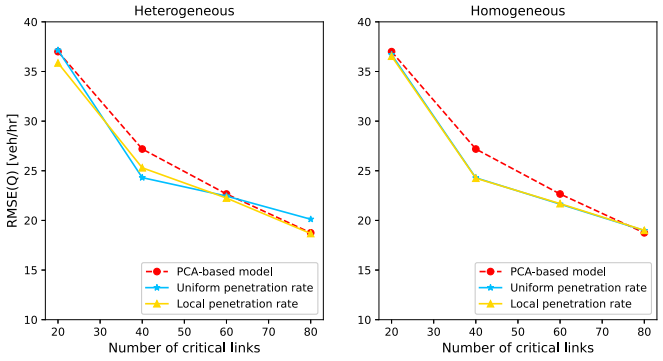


Fig. 12. Flow estimation error with respect to number of critical links.

The most evident observation is that flow and density estimation errors across all three methods tend to decrease when the number of critical links increases. Having more critical links means more links in the network with loop detectors placed on them. This provides us more traffic data from different parts of the network which eventually improves the accuracy of the estimated MFD. Moreover, as Figs. 11 and 12 depict, the fusion method, regardless of how the probe vehicles are distributed across the network (i.e., homogeneous or heterogeneous) and regardless of how the penetration rate is estimated (i.e., uniform or local), outperforms the PCA-based method in almost all cases. As mentioned before, the fusion method yields a significantly smaller average density estimation error in comparison to the PCA-based method (see Fig. 11). When estimating the average flow (see Fig. 12), however, the fusion method has a similar performance compared to the PCA-based method.

Comparing the density estimation in the homogeneous or heterogeneous scenarios (left and right plots in Fig. 11), we can see that when we only consider one single penetration rate to upscale the partial probe vehicle measurements (i.e., blue curves), the estimation errors in the heterogeneous scenario are higher than in the homogeneous scenario. This indicates that when the distribution of the probe vehicles across the network is heterogeneous, assuming a single penetration rate for the entire network leads to worse estimations. The proposed k-NN method enables the estimation of the local penetration rate, and significantly improves the density estimations in the heterogeneous probe vehicle scenario, please see the yellow curve in the left plot in Fig. 11. Albeit rather limited, the local penetration rates can also improve the density estimation in the homogeneous probe vehicle scenario, please see the difference between yellow and blue curves in the right plot in Fig. 11. On the other hand, Fig. 12 shows that there is not a considerable difference across the three methods in terms of the flow estimations in neither heterogeneous or homogeneous probe vehicle scenarios. This analysis allows us to conclude that the proposed local penetration rate estimation method is crucial for the estimation of average density values; particularly in realistic heterogeneous probe vehicle scenarios, this approach leads to a significant improvement in the density estimations.

Furthermore, Fig. 13 shows the estimated MFDs resulting from the fusion method with local penetration rates and uniform penetration rate considering different number of critical links. The results are also shown in a numerical fashion in Table 2 to help interpreting the MFDs derived from each methods. Note that the heterogeneous probe vehicle data set is used to estimate the following MFDs. The ground-truth MFD is plotted along with the estimated MFDs for comparison purposes. The figures confirm the results that are presented in the previous paragraph; that is, when probe vehicles are heterogeneously distributed, applying local penetration rates results in a more accurate MFD with less scatter especially in the congested part of the MFD.

Table 2
Average flow and density RMSEs applying the local and uniform penetration rates.

	Uniform penetration rate		Local penetration rate	
#Critical links	RMSE(K) [veh/km]	RMSE(Q) [veh/hr]	RMSE(K) [veh/km]	RMSE(Q) [veh/hr]
20	8.94	37.15	6.64	35.85
40	8.02	24.31	5.58	25.30
60	6.73	22.45	4.52	22.23
80	7.17	20.12	4.41	18.68

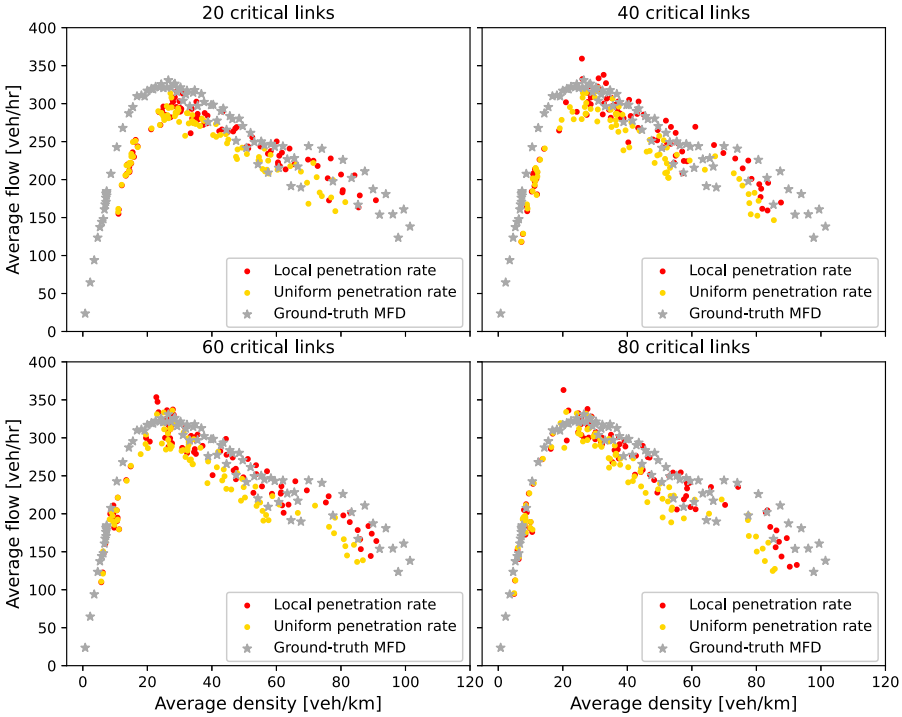


Fig. 13. Estimated MFDs based on different subsets of critical links.

5. Conclusion

The aim of this paper is to estimate the Macroscopic Fundamental Diagram (MFD) for a large-scale urban network using multi-sensor data. While the MFD is a powerful tool to control and manage urban networks, due to lack of sufficient traffic data for such large networks, estimating the MFD is not straightforward. Moreover, loop detectors and probe vehicles that are common sources of traffic data have limitations. Therefore, combining multiple traffic sources would not only provide a more complete data set, but each source can complement the limitations of the other source. For this purpose, in this paper we propose a Bayesian data fusion method to combine two traffic data sources (i.e., loop detectors and probe vehicles). We assume that real-time probe vehicle data with an unknown probe penetration rate is available for our network, and distribution of probe vehicles across the network is non-uniform. The second traffic data source is a full-scale approximation of traffic measurements (i.e., flow and density) using a PCA-based model. Note that this data was produced in our earlier study where we used Principal Component analysis (PCA) to calculate full-scale link-flow and density in every time interval. We also applied this PCA-based method to find the critical links in the network where loop detectors should be placed.

Given that the probe penetration rate is often not uniform across the network, we apply a local penetration rate where we assume probe penetration rate is only locally uniform. To calculate the local penetration rates, we use the k-nearest neighbour algorithm and find the k nearest critical links to each link in the network. Using the loop detector observations along with probe vehicle observations, we can calculate probe penetration rate on each critical link. Then, the probe penetration rate of each link can be approximated by averaging the penetration rates of its k nearest critical links. Having the penetration rate of each link, we are able to upscale the partial probe vehicle observations to complete observations and apply the Bayesian data fusion afterwards. The results show that the fusion model can significantly decrease the average density estimation error when compared with the PCA-based model. However, it performs as well as the PCA-based model in estimating the average flow.

Note that the objective of this paper is to test the proposed fusion method on the entire range of the MFD, from free-flow conditions with low average density values to very congested settings with close to jam density values. This requires us to consider

only the loading phase of the simulation where the average density in the network (almost) monotonically increases over time. We also acknowledge that the unloading phase of the simulation might reveal hysteresis patterns. Nevertheless, this requires changes in the defined scenario and in turn, re-training of the PCA model as well as the fusion model. Note that our estimation algorithm does not presume any arbitrary shape or function (e.g., bell-shaped or 3rd degree polynomial) with respect to the MFD. Instead, it retrieves data from loop detectors and probe vehicles, and derives the average flow and density values that arise from such observations through Bayesian inference. Therefore, we expect the algorithm to perform well even in the existence of hysteresis patterns, once it is retrained with relevant observations.

In this study, we relax the assumption of uniform probe vehicle distribution by using the vehicles travelling between specific O-D pairs as the probe vehicles. The proposed data fusion method is practical for large scale-urban networks. With advances in technology collecting and processing data has become less challenging; albeit for such large networks it is still complex and time-consuming. The findings presented in this paper show that even with limited available traffic data the Bayesian fusion algorithm provides promising results. In the future studies, it is important to investigate the effects of different probe vehicle sets on the final results of the fusion algorithm. Moreover, utilizing other traffic data sources (i.e., vehicle spacing data) for estimating the average flow and average density is another direction for our future research.

CRediT authorship contribution statement

Elham Saffari: Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Mehmet Yildirimoglu:** Conceptualization, Methodology, Writing – review & editing, Supervision. **Mark Hickman:** Writing – review & editing, Supervision.

Acknowledgement

This research is partly funded by iMOVE CRC and supported by the Cooperative Research Centres program, an Australian Government initiative. The authors would like to thank Dr Zili Li for his assistance with early stages of developing Bayesian fusion algorithm.

Appendix

Based on Bayes inference,

$$P(\mu_q|\mathbf{q}_a, q_p) \propto P(\mathbf{q}_a, \mathbf{q}_p|\mu_q)P(\mu_q) \tag{14}$$

Assuming the likelihood and the prior are following normal distributions,

$$P(\mu_q|\mathbf{q}_a, \mathbf{q}_p) \propto \exp\left\{-\frac{1}{2{\sigma_a}^2}\sum_{r=1}^{n_a}(q_a^r-\mu_q)^2\right\}\times\exp\left\{-\frac{1}{2{\sigma_p}^2}\sum_{s=1}^{n_p}(q_p^s-\mu_q)^2\right\}\times\exp\left\{-\frac{1}{2{\sigma_0}^2}(\mu_q-\mu_0)^2\right\} \tag{15}$$

The right-hand side of Eq. (15) can be expanded to

$$\exp\left\{\left[-\frac{1}{2{\sigma_a}^2}\sum_{r=1}^{n_a}((q_a^r)^2+\mu_q^2-2q_a^r\mu_q)\right]+\left[-\frac{1}{2{\sigma_p}^2}\sum_{s=1}^{n_p}((q_p^s)^2+\mu_q^2-2q_p^s\mu_q)\right]+\left[-\frac{1}{2{\sigma_0}^2}(\mu_q^2+\mu_0^2-2\mu_q\mu_0)\right]\right\} \tag{16}$$

Since the product of three normal distributions is itself a normal distribution, we can assume $P(\mu_q|\mathbf{q}_a, \mathbf{q}_p) \sim N(\mu_f, {\sigma_f}^2)$, and re-write the above equation in the form of,

$$\begin{aligned} P(\mu_q|\mathbf{q}_a, q_p) &\propto \exp\left[\frac{-\mu_q^2}{2}\left(\frac{1}{\sigma_0^2}+\frac{n_a}{\sigma_a^2}+\frac{n_p}{\sigma_p^2}\right)+\mu_q\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum q_a}{\sigma_a^2}+\frac{\sum q_p}{\sigma_p^2}\right)-\left(\frac{\mu_0^2}{2\sigma_0^2}+\frac{\sum q_a}{2\sigma_a^2}+\frac{\sum q_p}{2\sigma_p^2}\right)\right] \\ &\overset{\text{def}}{=} \exp\left\{\frac{-1}{2\sigma_f^2}(\mu_q-\mu_f)^2\right\} = \exp\left[\frac{-1}{2\sigma_f^2}(\mu_q^2+\mu_f^2-2\mu_q\mu_f)\right] \end{aligned} \tag{17}$$

Matching coefficients of μ_q^2 from Eq. (17), we can find σ_f^2 as,

$$\frac{1}{\sigma_f^2} = \frac{1}{\sigma_0^2}+\frac{n_a}{\sigma_a^2}+\frac{n_p}{\sigma_p^2} \tag{18}$$

$$\sigma_f^2 = \frac{1}{\frac{1}{\sigma_0^2}+\frac{n_a}{\sigma_a^2}+\frac{n_p}{\sigma_p^2}} \tag{19}$$

Matching coefficients of μ_q from Eq. (17), we can find μ_f as,

$$\frac{\mu_0}{\sigma_0^2}+\frac{\sum q_a}{\sigma_a^2}+\frac{\sum q_p}{\sigma_p^2} = \frac{\mu_f}{\sigma_f^2} \tag{20}$$

Plugging in the value of σ_f^2 from (19) into (20), we can find μ_f as,

$$\mu_f = \frac{1}{\frac{1}{\sigma_0^2}+\frac{n_a}{\sigma_a^2}+\frac{n_p}{\sigma_p^2}}\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum q_a}{\sigma_a^2}+\frac{\sum q_p}{\sigma_p^2}\right) \tag{21}$$

References

Ambühl, L., Loder, A., Bliemer, M.C., Menendez, M., Axhausen, K.W., 2018. Introducing a re-sampling methodology for the estimation of empirical macroscopic fundamental diagrams. Transp. Res. Rec. 2672, 239–248.

Ambühl, L., Loder, A., Leclercq, L., Menendez, M., 2021. Disentangling the city traffic rhythms: A longitudinal analysis of MFD patterns over a year. Transp. Res. C 126, 103065.

Ambühl, L., Loder, A., Menendez, M., Axhausen, K.W., 2017. Empirical macroscopic fundamental diagrams: New insights from loop detector and floating car data. In: TRB 96th Annual Meeting Compendium of Papers. Transportation Research Board, pp. 17–03331.

Ambühl, L., Loder, A., Zheng, N., Axhausen, K.W., Menendez, M., 2019. Approximative network partitioning for MFDs from stationary sensor data. Transportation Research Record 2673, 94–103.

Ambühl, L., Menendez, M., 2016. Data fusion algorithm for macroscopic fundamental diagram estimation. Transp. Res. C 71, 184–197.

Bachmann, C., Abdulhai, B., Roorda, M.J., Moshiri, B., 2013. A comparative assessment of multi-sensor data fusion techniques for freeway traffic speed estimation using microsimulation modeling. Transp. Res. C 26, 33–48.

Barceló, J., Casas, J., 2005. Dynamic network simulation with AIMSUN. In: Simulation Approaches in Transportation Analysis. Springer, pp. 57–98.

Beibei, J., van Zuylen, H., Shoufeng, L., 2016. Determining the macroscopic fundamental diagram on the basis of mixed and incomplete traffic data. In: TRB 95th Annual Meeting Compendium of Papers.

Buisson, C., Ladier, C., 2009. Exploring the impact of homogeneity of traffic measurements on the existence of macroscopic fundamental diagrams. Transp. Res. Rec. 2124, 127–136.

Castanedo, F., 2013. A review of data fusion techniques. Sci. World J. 2013.

Courbon, T., Leclercq, L., 2011. Cross-comparison of macroscopic fundamental diagram estimation methods. Procedia Soc. Behav. Sci. 20, 417–426.

Daganzo, C.F., 2005. A variational formulation of kinematic waves: Basic theory and complex boundary conditions. Transp. Res. B 39, 187–196.

Daganzo, C.F., 2007. Urban gridlock: Macroscopic modeling and mitigation approaches. Transp. Res. B 41, 49–62.

Daganzo, C.F., Geroliminis, N., 2008. An analytical approximation for the macroscopic fundamental diagram of urban traffic. Transp. Res. B 42, 771–781.

Du, J., Rakha, H., Gayah, V.V., 2016. Deriving macroscopic fundamental diagrams from probe data: Issues and proposed solutions. Transp. Res. C 66, 136–149.

Fu, H., Wang, Y., Tang, X., Zheng, N., Geroliminis, N., 2020. Empirical analysis of large-scale multimodal traffic with multi-sensor data. Transp. Res. C 118, 102725.

Geroliminis, N., Boyacı, B., 2012. The effect of variability of urban systems characteristics in the network capacity. Transp. Res. B 46, 1607–1623.

Geroliminis, N., Daganzo, C.F., 2008. Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. Transp. Res. B 42, 759–770.

Geroliminis, N., Haddad, J., Ramezani, M., 2012. Optimal perimeter control for two urban regions with macroscopic fundamental diagrams: A model predictive approach. IEEE Trans. Intell. Transp. Syst. 14, 348–359.

Gu, Z., Shafiei, S., Liu, Z., Saberi, M., 2018. Optimal distance-and time-dependent area-based pricing with the network fundamental diagram. Transp. Res. C 95, 1–28.

Guo, F., Polak, J.W., Krishnan, R., et al., 2018. Predictor fusion for short-term traffic forecasting. Transp. Res. C, Emerg. Technol. 92, 90–100.

Huang, C., Zheng, N., Zhang, J., 2019. Investigation of bimodal macroscopic fundamental diagrams in large-scale urban networks: Empirical study with GPS data for shenzhen city. Transp. Res. Rec. 2673, 114–128.

Ingole, D., Mariotte, G., Leclercq, L., 2020. Perimeter gating control and citywide dynamic user equilibrium: A macroscopic modeling framework. Transp. Res. C 111, 22–49.

Ji, Y., Daamen, W., Hoogendoorn, S., Hoogendoorn-Lanser, S., Qian, X., 2010. Investigating the shape of the macroscopic fundamental diagram using simulation data. Transp. Res. Rec. 2161, 40–48.

Ji, Y., Xu, M., Li, J., van Zuylen, H.J., 2018. Determining the macroscopic fundamental diagram from mixed and partial traffic data. Promet Traffic Transp. 30, 267–279.

Johari, M., Keyvan-Ekbatani, M., Leclercq, L., Ngoduy, D., Mahmassani, H.S., 2021. Macroscopic network-level traffic models: Bridging fifty years of development toward the next era. Transp. Res. C 131, 103334. http://dx.doi.org/10.1016/j.trc.2021.103334, URL: https://www.sciencedirect.com/science/article/pii/S0968090X21003375.

Keyvan-Ekbatani, M., Gao, X., Gayah, V.V., Knoop, V.L., 2019. Traffic-responsive signals combined with perimeter control: Investigating the benefits. Transp. B Transp Dyn. 7, 1402–1425.

Keyvan-Ekbatani, M., Papageorgiou, M., Papamichail, I., 2013. Urban congestion gating control based on reduced operational network fundamental diagrams. Transp. Res. C 33, 74–87.

Keyvan-Ekbatani, M., Yildirimoglu, M., Geroliminis, N., Papageorgiou, M., 2015. Multiple concentric gating traffic control in large-scale urban networks. IEEE Trans. Intell. Transp. Syst. 16, 2141–2154.

Knoop, V.L., De Jong, D., Hoogendoom, S.P., 2014. Influence of road layout on network fundamental diagram. Transp. Res. Rec. 2421, 22–30.

Knoop, V.L., van Erp, P.B., Leclercq, L., Hoogendoorn, S.P., 2018. Empirical MFDs using google traffic data. In: 2018 21st International Conference on Intelligent Transportation Systems, ITSC. IEEE, pp. 3832–3839.

Kouvelas, A., Saeedmanesh, M., Geroliminis, N., 2017. Enhancing model-based feedback perimeter control with data-driven online adaptive optimization. Transp. Res. B 96, 26–45.

Kumarage, S., Yildirimoglu, M., Ramezani, M., Zheng, Z., 2021. Schedule-constrained demand management in two-region urban networks. Transp. Sci. 55, 857–882.

Laval, J.A., Castrillón, F., 2015. Stochastic approximations for the macroscopic fundamental diagram of urban networks. Transp. Res. Procedia 7, 615–630.

Leclercq, L., Chiabaut, N., Trinquier, B., 2014. Macroscopic fundamental diagrams: A cross-comparison of estimation methods. Transp. Res. B 62, 1–12.

Leclercq, L., Geroliminis, N., 2013. Estimating MFDs in simple networks with route choice. Procedia Soc. Behav. Sci. 80, 99–118.

Li, Y., Yildirimoglu, M., Ramezani, M., 2021. Robust perimeter control with cordon queues and heterogeneous transfer flows. Transp. Res. C 126, 103043.

Lin, X., Xu, J.M., 2018. Macroscopic fundamental diagram estimation fusion method of road networks based on adaptive weighted average. J. Transp. Syst. Eng. Inf. Technol. 18, 102–109.

Lin, X., Xu, J., Cao, C., 2019. Simulation and comparison of two fusion methods for macroscopic fundamental diagram estimation. Arch. Transp. 51.

Loder, A., Ambühl, L., Menendez, M., Axhausen, K.W., 2019. Understanding traffic capacity of urban networks. Sci. Rep. 9, 1–10.

Marcoulides, K.M., 2017. A Bayesian Synthesis Approach to Data Fusion using Augmented Data-Dependent Priors. Arizona State University.

Mariotte, G., Leclercq, L., Batista, S., Krug, J., Paipuri, M., 2020. Calibration and validation of multi-reservoir MFD models: A case study in lyon. Transp. Res. B 136, 62–86.

Maskell, S., 2008. A Bayesian approach to fusing uncertain, imprecise and conflicting information. Inf. Fusion 9, 259–277.

Menendez, M., Ambühl, L., Loder, A., Zheng, N., Axhausen, K.W., 2019. Approximative network partitioning for MFDs from stationary sensor data.

Muthén, B., Asparouhov, T., 2012. Bayesian structural equation modeling: A more flexible representation of substantive theory. Psychol. Methods 17, 313.

Nagle, A.S., Gayah, V.V., 2013. A method to estimate the macroscopic fundamental diagram using limited mobile probe data. In: 16th International IEEE Conference on Intelligent Transportation Systems, ITSC 2013. IEEE, pp. 1987–1992.

Ortigosa, J., Menendez, M., Tapia, H., 2014. Study on the number and location of measurement points for an MFD perimeter control scheme: A case study of Zurich. EURO J. Transp. Logist. 3, 245–266.
Ramezani, M., Nourinejad, M., 2018. Dynamic modeling and control of taxi services in large-scale urban networks: A macroscopic approach. Transp. Res. C 94, 203–219.
Saberi, M., Mahmassani, H.S., Hou, T., Zockaie, A., 2014. Estimating network fundamental diagram using three-dimensional vehicle trajectories: Extending Edie's definitions of traffic flow variables to networks. Transp. Res. Rec. 2422, 12–20.
Saffari, E., Yildirimoglu, M., Hickman, M., 2020. A methodology for identifying critical links and estimating macroscopic fundamental diagram in large-scale urban networks. Transp. Res. C 119, 102743.
Shoufeng, L., Jie, W., van Zuylen, H., Ximin, L., 2013. Deriving the macroscopic fundamental diagram for an urban area using counted flows and taxi GPS. In: 16th International IEEE Conference on Intelligent Transportation Systems, ITSC 2013. IEEE, pp. 184–188.
Tilg, G., Amini, S., Busch, F., 2020. Evaluation of analytical approximation methods for the macroscopic fundamental diagram. Transp. Res. C 114, 1–19.
Tsubota, T., Bhaskar, A., Chung, E., 2014. Macroscopic fundamental diagram for Brisbane, Australia: Empirical findings on network partitioning and incident detection. Transp. Res. Rec. 2421, 12–21.
Yildirimoglu, M., Ramezani, M., 2020. Demand management with limited cooperation among travellers: A doubly dynamic approach. Transp. Res. B 132, 267–284.
Yildirimoglu, M., Ramezani, M., Geroliminis, N., 2015. Equilibrium analysis and route guidance in large-scale networks with MFD dynamics. Transp. Res. Procedia 9, 185–204.
Yildirimoglu, M., Sirmatel, I.I., Geroliminis, N., 2018. Hierarchical control of heterogeneous large-scale urban road networks via path assignment and regional route guidance. Transp. Res. B 118, 106–123.
Zheng, N., Rérat, G., Geroliminis, N., 2016. Time-dependent area-based pricing for multimodal systems with heterogeneous users in an agent-based environment. Transp. Res. C 62, 133–148.
Zockaie, A., Saberi, M., Saedi, R., 2018. A resource allocation problem to estimate network fundamental diagram in heterogeneous networks: Optimal locating of fixed measurement points and sampling of probe trajectories. Transp. Res. C 86, 245–262.