DOI: 10.1111/mice.12526

ORIGINAL ARTICLE

COMPUTER-AIDED CIVIL AND INFRASTRUCTURE ENGINEERING WILEY

Estimating dynamic origin–destination demand: A hybrid framework using license plate recognition data

Baichuan Mo[1,2] | Ruimin Li[2] | Jingchen Dai[2]

[1]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

[2]Department of Civil Engineering, Tsinghua University, Beijing, China

Correspondence
Ruimin Li, Department of Civil Engineering, Tsinghua University, Beijing 100084, China.
Email: lrmin@tsinghua.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 71361130015; Beijing Natural Science Foundation, Grant/Award Number: 8162024

Abstract

This article proposes a hybrid framework for estimating dynamic origin–destination (OD) demand that fully exploits the information available in license plate recognition (LPR) data. A Bayesian path reconstruction model is initially developed to replenish the lost information resulting from the recognition error and insufficient coverage rate of the LPR system. The link flows, initial OD demand, left-turning flows, and partial path flows are derived based on the reconstructed data. Subsequently, with the information derived, a two-step ordinary least squares (OLS) OD estimation model is formulated, which incorporates the output from the Bayesian model and coestimates the OD demand and assignment matrix. The proposed framework is qualitatively validated using the real-world LPR data collected from Langfang City, Hebei Province, China, and is quantitatively validated using the synthesized simulation data for the simplified road network of Langfang. The results show that the proposed model can estimate OD demand distribution with a mean absolute percentage error (MAPE) of about 30%. We also tested the model with different LPR coverage rates, with results showing that an LPR coverage rate of over 50% is required to obtain reasonable results.

1 | INTRODUCTION

Dynamic origin–destination (OD) demand, also known as dynamic OD matrix, is a fundamental input in many transportation assignment models for traffic management (Ukkusuri, Mathew, & Waller, 2007; Wen, Cai et al., 2018), congestion analysis (Adeli & Ghosh-Dastidar, 2004; Jiang & Adeli, 2004a, 2004b), and intelligent transportation system development (Adeli & Karim, 2005; Hooshdar & Adeli, 2004; Karim & Adeli, 2003). Traditionally, OD demand is calculated based on survey and urban land use data; however, the results of this calculation are rough and prone to modeling errors. With the introduction of new traffic detection technologies, many OD demand estimation methods that use multiple sources of traffic detection data, including GNSS, cellphone, and Bluetooth-based trajectory data, have become available (Cremer & Keller, 1981; Zhou & Mahmassani, 2006). One potential data source is license plate recognition (LPR), a major type of automatic vehicle identification technology that has been widely deployed in urban and highway transportation systems in recent years (Mo, Li, & Zhan, 2017; Nakanishi & Western, 2005; J. Yang & Sun, 2015). Unlike conventional fixed sensor detectors, an LPR system not only records the time when a vehicle passes through a stop line but also identifies the same vehicle that passes through multiple intersections by recognizing its unique license plate. Compared with decentralized probe data (e.g., GNSS and mobile phone data), LPR data have a much higher penetration rate and therefore offer a better representation of the traffic pattern of road networks (Nigro, Cipriani, & del Giudice, 2018). These unique features make LPR data a potential source for OD estimation.

Many studies have begun to examine dynamic OD demand estimation over the past few decades. These studies mostly rely on the time-dependent link counts collected by fixed sensors on estimating dynamic OD (Cremer & Keller, 1981; Larsson, Lundgren, & Peterson, 2010; Nihan & Davis, 1987; Sherali & Park, 2001; Tavana & Mahmassani, 2001). However, given that unknown OD flows usually outnumber the known link counts (Zhou & Mahmassani, 2006), these methods face an underdetermination problem. To solve such a problem, initial OD demands (e.g., historical OD demand) have been introduced to reduce the scope of local optimal solutions (Bierlaire & Crittin, 2004; Cascetta, Inaudi, & Marquis, 1993; Lundgren & Peterson, 2008; Stathopoulos & Tsekeris, 2004). Therefore, the estimation accuracy of this method is highly dependent on the selection of initial OD demands (Cipriani, Florian, Mahut, & Nigro, 2011).

LPR-data–based dynamic OD estimation studies have also employed the aforementioned framework. The difference is that new information specific to LPR data has been added. van der Zijpp (1997) proposed a Bayesian updating procedure for dynamic OD matrix estimation by fusing the partial path flow derived from LPR data with the traffic counts collected from loop detector data. Dixon and Rilett (2002) applied the Kalman filter method to estimate real-time OD demands. The link volumes, OD split proportions, link choice proportions, and travel times extracted from LPR data were used as inputs. However, these two studies were limited to a closed network (e.g., freeway network) from which route choice proportions can be observed or directly derived. Therefore, their methods cannot be applied in an urban road network with complex and unknown route choice behaviors. Zhou and Mahmassani (2006) formulated a nonlinear ordinary least squares (OLS) model to dynamically estimate OD demand distribution in urban road networks. The historical OD demand observed link flow, and link-to-link split fractions are embedded into the multiobjective function. Given that link-to-link split fractions cannot be directly observed, these variables are jointly estimated with OD demand variables, thereby complicating the model structure (Zhou & Mahmassani, 2006). These methods essentially minimize the deviation between the information derived from *raw* LPR data and the information estimated by the proposed models. Additional information, including historical OD demand, is also needed to obtain the expected solution, but the valuable underlying information in LPR data is underutilized. Despite their advantages, LPR data suffer from a low recognition rate resulting from erroneous, failed, and missed detection (Mo et al., 2017). Therefore, directly using raw LPR data may result in the omission of implied information. Based on the inherent rules of transportation systems, preprocessing LPR data can provide the information that was lost as a result of low recognition and insufficient penetration rates.

Another stream of research based on path flow reconstruction has estimated OD demand by using preprocessed LPR data. Castillo, Menéndez, and Jiménez (2008) proposed a quadratic-programming–based path flow reconstruction model by using LPR data. In this model, the unrecognized vehicles in an LPR system are managed by using a statistical method, and then the estimated path flow is aggregated to estimate the OD flow. J. Yang and Sun (2015) proposed an integrated vehicle path reconstruction method using LPR data by combining the particle filter and path flow estimator algorithms. The reconstructed path information is then used to estimate path flow. LPR data are carefully processed in the path reconstruction process based on artificial assumptions (e.g., links with large volumes are more likely to be chosen). Rao, Wu, Xia, Ou, and Kluger (2018) extended the above method to a large-scale network setting. They initially reconstructed the path flow by using a particle filter algorithm and then directly aggregated the reconstructed path flow into the OD flow. Despite further utilizing LPR data, these path-flow-reconstruction–based methods complicate the OD demand estimation problem because the path flow is disaggregated by the OD flow. Therefore, a greater number of unknown variables exist during the path flow estimation. These methods also often estimate the OD demand by simply taking the sum of the corresponding path flows, which is too arbitrary and rigid. If the path reconstruction precision is lacking, especially in cases of low LPR recognition and coverage rate, then the path can only be partially reconstructed, thereby incurring losses in the original origin and destination and resulting in serious OD estimation inaccuracies (H. Yang, Iida, & Sasaki, 1991).

The aforementioned streams of LPR-data–based dynamic OD estimation research have advantages and disadvantages. On the one hand, the former estimates OD demand by using elaborate mathematical models without exploiting LPR data and usually requires additional information (e.g., historical OD demand) to obtain reasonable results. On the other hand, the latter thoroughly processes the LPR data to obtain additional path flow information, but the bridge that connects the path and the OD flows is not well constructed.

This study aims to estimate the dynamic OD demand based on LPR data by combining the advantages of the two aforementioned methods. Specifically, we build a hybrid framework for dynamic OD demand estimation that fully exploits the information provided in LPR data. This framework does not require the utilization of information sources other than LPR data. A Bayesian path reconstruction model is initially developed to replenish the information that is lost as a result of the recognition error and insufficient penetration rate of the LPR system. Based on the reconstructed data, we derive the link flows, initial OD demand, left-turning, and partial path flows. Afterward, a two-step OLS OD estimation model is formulated by using all the derived information.

The proposed framework is qualitatively validated by using real-world LPR data collected from Langfang City, Hebei Province, China, and is quantitatively validated by using synthesized simulation data from a simplified road network of Langfang. The results show that the proposed model can estimate OD demand distribution with a mean absolute percentage error (MAPE) of about 30%. Two features distinguish this study from cutting-edge research in the literature: a hybrid framework that combines optimization-based and path-reconstruction–based OD estimation methods and deep exploitation of LPR data that derives information equivalent to a combination of traditional link count and trajectory data. The methodological contribution of this study is twofold. First, we propose a data-driven Bayesian path reconstruction model that eliminates artificial assumption of people's preference. Second, we propose a two-step OLS model for estimating dynamic OD demand, which incorporates output information from the Bayesian model and coestimates the assignment matrix.

The rest of this article is organized as follows. Section 2 presents the proposed path reconstruction method. Section 3 shows the formulation of the proposed two-step OLS model. Section 4 presents the validation design and the numerical results. Section 5 presents the conclusions and discussions.

2 | BAYESIAN PATH RECONSTRUCTION

Typical LPR systems are used for red-light violation enforcement. These systems use a camera to take pictures of vehicles that pass through a stop line. In this way, information about the passing timestamps, vehicle license plate numbers, and occupied lanes can be obtained from LPR systems. However, due to the limited recognition precision of LPR systems, the license plate of some vehicles may be incorrectly recorded as "none" if not successfully recognized. Given that those vehicles that pass through LPR stations have a corresponding record whether they have been successfully recognized or not, the link volume can be (almost) accurately extracted from the data set if this link is equipped with an LPR camera.

Nevertheless, LPR data show some limitations in providing vehicle trajectory information, which is important in OD estimation (Rao et al., 2018). Vehicles may be erroneously detected or completely undetected in LPR systems. Besides, vehicles may pass through an intersection without an LPR camera installed. These two scenarios make the vehicle trajectories directly derived from raw LPR data incomplete. Using these unclean data in OD estimation models can produce estimation errors and affect model reliability (Yu, Yang, Wu, & Ma, 2018). To solve this problem, several path reconstruction methods for LPR data have been proposed in recent years.

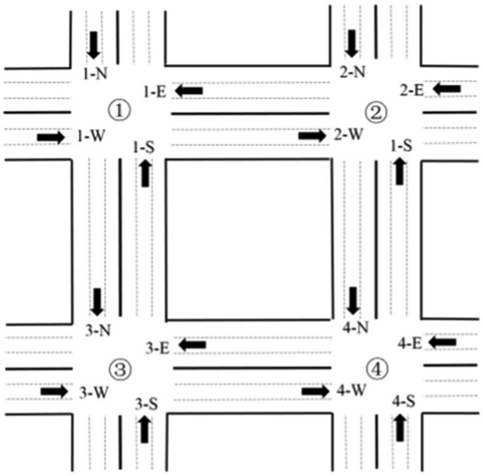


FIGURE 1 Explanatory diagram of node definition

Castillo et al. (2008) proposed a path reconstruction method based on Bayes' theorem. However, this method uses only the recognition rate of the LPR system and neglects other useful information provided in the LPR data, including travel time and vehicle-occupied lane. Feng, Sun, and Chen (2015) proposed a particle filter–based vehicle trajectory reconstruction framework using LPR and traffic count data. J. Yang and Sun (2015) developed this method further by integrating a macroscopic path flow estimator and proposing a hybrid path reconstruction model. However, both these methods need to update particle weight based on several assumptions, such as "people always prefer to take the shorter path," "those links with large volumes are more likely to be chosen," and "vehicles always arrive from an adjacent zone or with a higher traffic flow" (Feng et al., 2015; J. Yang & Sun, 2015). Although these assumptions are reasonable to some extent, we should avoid using too many artificial assumptions and focus instead on exploiting the value of the data in data-driven modeling.

In this study, we propose a data-driven Bayesian path reconstruction method that attempts to fully utilize the information in the LPR data and eliminate the man-made assumption for people's preference. This model is explained in detail in the following sections.

2.1 | Network nodes definition

We initially define a road network node in the context of the LPR data–based model. In previous studies, a node is defined as a real-world intersection. Meanwhile, we use a finer-grained node definition by combining intersection ID with arriving direction. Figure 1 presents an example of a simple road network with four LPR system detection stations installed at each intersection (labeled ①, ②, ③, and ④). At each station, four cameras are set in front of the stop line to capture the vehicles arriving from directions W, N, E, and S. Note that this graph is only used for illustration purposes.

Installing cameras that cover all directions is unnecessary for the model. By taking intersection ① as an example, the four nodes are defined as 1-W, 1-N, 1-E, and 1-S. Therefore, in this simple four-intersection network, a total of $4 \times 4 = 16$ nodes are defined. This fine-grained definition allows us to accurately determine the topology relationship. For example, even if a vehicle is recorded at 1-W and 4-N (with the other records being lost), we can directly guess its true path as (1-W, 2-W, 4-N) with high confidence because only those vehicles passing through intersection ② can reach node 4-N. However, if we use the previous rough node definition method that only uses intersection ID (i.e., label nodes as ①, ②, ③, and ④ in the network), then the true path of the vehicle with records ① and ④ cannot be easily guessed because path (①, ②, ④) and path (①, ③, ④) are possible candidates in the context of the rough node definition.

Based on the new definition, nodes 1-W and 2-W are topologically continuous, whereas nodes 1-W and 2-N are not because a vehicle cannot move directly from 1-W to 2-N.

2.2 | Continuous node judgment

In this subsection, we describe how we judge whether the chronological nodes extracted from raw LPR data are truly continuous. Extracting continuous nodes is the first step in the path reconstruction procedure, and the remaining discontinuous nodes are used as inputs for the path reconstruction model. The following criteria are used for the judgment:

- *Criterion 1*: The two recorded nodes must be topologically continuous.
- *Criterion 2*: The difference between the two recorded timestamps (i.e., the recorded travel time between the two nodes) must fall within the 95% confidence interval of the link travel time distribution.

The first criterion is negligible, as set out in the definition of topological continuity elaborated on in Section 2.1. For the second criterion, some vehicles may have more than one trip during a day and may park for a long time between two topologically continuous nodes. Thus, the second criterion is used to separate the topologically continuous but time-discontinuous trajectories into multiple trips. The confidence interval–based criterion is more reliable than the fixed-value threshold-based criteria. The link travel time distribution can be calculated by using the vehicle travel time information from LPR data. The kernel density estimation method (Pedregosa et al., 2011) is used to estimate the unknown distribution and to calculate the confidence level. Given that the link travel time distribution varies across time, the above procedure is conducted within predefined time intervals. Time interval is treated as the computational unit for dynamic OD estimation. A time span of 15 minutes, 30 minutes, or 1 hour is often used. In this study, a time span of 30 minutes is used to represent the length of a time interval.

After judging the continuous nodes, we obtain many raw partial trajectories by simply connecting these nodes. Note that some derived trajectories may only have one node. Afterward, we reconstruct the path between two raw partial trajectories or conclusively divide them into two separate trips.

2.3 | Path reconstruction

As mentioned above, the discontinuous nodes are used as inputs for the path reconstruction model. For a specific vehicle V, we denote the raw partial trajectories obtained in Section 2.2 as $R^{V}_{\text{partial}} = \{R^{V,1}_{\text{partial}}, R^{V,2}_{\text{partial}}, \cdots, R^{V,n_p}_{\text{partial}}\}$. The trajectory ID 1, 2, $\cdots, n_p$ is sorted by time. For trajectories $R^{V,k}_{\text{partial}}$ and $R^{V,k+1}_{\text{partial}}$, let $N^{-}_{V,k}$ be the last node of $R^{V,k}_{\text{partial}}$ and $N^{+}_{V,k+1}$ be the first node of $R^{V,k+1}_{\text{partial}}$. Then, $N^{-}_{V,k}$ and $N^{+}_{V,k+1}$ are the two discontinuous nodes that will be fed to the path reconstruction model. *Discontinuous nodes* refer to two nodes that fail to satisfy either *Criterion 1* or *2*.

$R^{V,k} = \{R^{V,k}_{1}, R^{V,k}_{2}, \cdots, R^{V,k}_{n_c}\}$ denotes the possible path candidates between $N^{-}_{V,k}$ and $N^{+}_{V,k+1}$. $R^{V,k}$ is the union of *existing paths* and *efficient paths*. Existing paths refer to all raw partial trajectories that connect $N^{-}_{V,k}$ with $N^{+}_{V,k+1}$ as obtained in Section 2.2. Efficient paths were defined by Dial (1971): A path is considered efficient if every link in this path has its initial node located closer to the path origin than to the final node of the link (i.e., a path that does not backtrack). The inefficient paths (e.g., vehicles turning away from the destination), if existing, will be included in the existing paths set. The number of candidate paths can be extremely large in an urban road network when $N^{-}_{V,k}$ is located far away from $N^{+}_{V,k+1}$, thereby incurring a high computation cost. According to Parry and Hazelton (2012), using more than six or seven routes for any given OD pair is unusual. Therefore, to improve efficiency, when the number of efficient paths exceeds six, we choose only the shortest six paths as candidates. The results of our numerical test reveal that the number of path candidates from the efficiency path slightly influences the reconstruction accuracy. Given that having too many path candidates may reduce the accuracy, people can adjust this value as appropriate. Future studies may explore for the optimal number of path candidates. The purpose of our work is to calculate the posterior probability for each candidate to be the real path conditional on the information observed in the LPR data. For the Bayesian reconstruction, we use three types of information extracted from LPR data, namely, the detected nodes set $N^{V,k}_{d} = \{N^{-}_{V,k}, N^{+}_{V,k+1}\}$, the travel time $TT_{V,k}$ between $N^{-}_{V,k}$ and $N^{+}_{V,k+1}$, and the lane $L_{V,k}$ occupied by the detected vehicle V recorded in $N^{-}_{V,k}$. These three types of information affect the posterior probability that $R^{V,k}_{i} \in R^{V,k}$ is the real path. Then, the posterior probability that $R^{V,k}_{i}$ is the real path can be formulated as $P(R_i^{V,k} \mid N_d^{V,k}, TT_{V,k}, L_{V,k})$. Based on Bayes' theorem, we expand this expression as

$$P\left(R_i^{V,k}|N_d^{V,k}, TT_{V,k}, L_{V,k}\right) = \frac{P\left(N_d^{V,k}|R_i^{V,k}, TT_{V,k}, L_{V,k}\right) \cdot P\left(R_i^{V,k}|TT_{V,k}, L_{V,k}\right)}{\sum_{j=1}^{n_c} P\left(N_d^{V,k}|R_j^{V,k}, TT_{V,k}, L_{V,k}\right) \cdot P\left(R_j^{V,k}|TT_{V,k}, L_{V,k}\right)} \tag{1}$$

where n_c is the number of path candidates. Given that the detected nodes $N_d^{V,k}$ are determined only by the path chosen by vehicle V and the recognition rate of LPR devices, $TT_{V,k}$ and $L_{V,k}$ have no effect on these nodes. Therefore, $P(N_d^{V,k}|R_i^{V,k}, TT_{V,k}, L_{V,k})$ can be simplified to $P(N_d^{V,k}|R_i^{V,k})$. $P(N_d^{V,k}|R_i^{V,k})$ must be calculated by using the recognition rate of LPR stations. Given that the recognition rate varies across time due to the changes in traffic lights and other elements, we must find the time interval τ when $N_{V,k}^-$ is recorded, and then assume that the following time-dependent computing procedures all lie within this time interval. For the quantitative analysis, we assume that recognition error is independent between users and between LPR stations (Castillo et al., 2008). Based on this assumption, $P(N_d^{V,k}|R_i^{V,k})$ can be rewritten as

$$P\left(N_d^{V,k}|R_i^{V,k}\right) = \mathrm{P}_{\mathrm{re},\tau}\left(N_{V,k}^-\right) \cdot P_{\mathrm{re},\tau}\left(N_{V,k+1}^+\right) \cdot \prod_{m=1}^{n_u}\left(1 - P_{\mathrm{re},\tau}\left(N_u^{R_i^{V,k},(m)}\right)\right) \tag{2}$$

where $P_{\mathrm{re},\tau}()$ is the LPR recognition rate of the corresponding nodes within time interval τ. This parameter can be directly calculated by dividing the number of recognized vehicles by the number of total vehicles passing the node in time interval τ. Meanwhile, $N_u^{R_i^{V,k}} = \{N_u^{R_i^{V,k},(1)}, N_u^{R_i^{V,k},(2)}, \cdots, N_u^{R_i^{V,k},(n_u)}\}$ denotes the set of unrecognized nodes of path $R_i^{V,k}$ (i.e., all nodes in path $R_i^{V,k}$, except for $N_{V,k}^-$ and $N_{V,k+1}^+$). Equation (2) reveals the probability for vehicle V to be recorded in node $N_d^{V,k}$ and not recorded in node $N_u^{R_i^{V,k}}$.

In Equation (1), $P(R_i^{V,k}|TT_{V,k}, L_{V,k})$ can be expanded as follows based on Bayes' theorem:

$$P\left(R_i^{V,k}|TT_{V,k}, L_{V,k}\right) = \frac{P\left(TT_{V,k}|R_i^{V,k}, L_{V,k}\right) \cdot P\left(R_i^{V,k}|L_{V,k}\right)}{\sum_{j=1}^{n_c} P\left(TT_{V,k}|R_j^{V,k}, L_{V,k}\right) \cdot P\left(R_j^{V,k}|L_{V,k}\right)} \tag{3}$$

where $P(R_i^{V,k}|L_{V,k})$ indicates the probability of vehicle V choosing $R_i^{V,k}$ conditional on leaving $N_{V,k}^-$ in lane $L_{V,k}$. The lane affects the probability because drivers can switch to a suitable lane in their upcoming turns. For example, if vehicle V anticipates turning left after passing $N_{V,k}^-$, then this vehicle has a high probability of switching to the leftmost lane in advance. $P(R_i^{V,k}|L_{V,k})$ can be expanded as

$$P\left(R_i^{V,k}|L_{V,k}\right) = \frac{P\left(L_{V,k}|R_i^{V,k}\right) \cdot P\left(R_i^{V,k}\right)}{\sum_{j=1}^{n_c} P\left(L_{V,k}|R_j^{V,k}\right) \cdot P\left(R_j^{V,k}\right)} \tag{4}$$

where $P(R_i^{V,k})$ denotes the prior probability for people to use $R_i^{V,k}$, and $P(L_{V,k}|R_i^{V,k})$ denotes the probability for vehicles with trajectories $R_i^{V,k}$ to leave $N_{V,k}^-$ in lane $L_{V,k}$. These probabilities can be calculated as follows by using raw trajectory data:

$$P\left(R_i^{V,k}\right) = \frac{\alpha_{R_i^{V,k}}}{\sum_{j=1}^{n_c} \alpha_{R_j^{V,k}}} \quad \text{and} \tag{5}$$

$$P\left(L_{V,k}|R_i^{V,k}\right) = \frac{\alpha_{R_i^{V,k}, L_{V,k}}}{\alpha_{R_i^{V,k}}} \tag{6}$$

where $\alpha_{R_i^{V,k}}$ is the number of vehicles using path $R_i^{V,k}$, and $\alpha_{R_i^{V,k}, L_{V,k}}$ is the number vehicles with raw trajectory $R_i^{V,k}$ that leave $N_{V,k}^-$ in lane $L_{V,k}$. These variables can be directly counted by using raw trajectory data. Therefore, Equations (5) and (6) yield

$$P\left(R_i^{V,k}|L_{V,k}\right) = \frac{\alpha_{R_i^{V,k}, L_{V,k}}}{\sum_{j=1}^{n_c} \alpha_{R_j^{V,k}, L_{V,k}}} \tag{7}$$

Before calculating Equation (1), the only remaining step is to calculate $P(TT_{V,k}|R_i^{V,k}, L_{V,k})$. $L_{V,k}$ has almost no effect on travel time $TT_{V,k}$ when the path $R_i^{V,k}$ is determined. Therefore, $P(TT_{V,k}|R_i^{V,k}, L_{V,k})$ can be simplified to $P(TT_{V,k}|R_i^{V,k})$, which can be directly derived when the travel time distribution of $R_i^{V,k}$ is known. Given that travel time may vary within the same day, its distribution must be calculated over the time interval τ. A simple method for calculating such distribution is to extract all corresponding path travel time samples from raw trajectories and then estimate the distribution by using a kernel density estimation method (Pedregosa et al., 2011). However, according to the numerical test results, the number of eligible trajectories is small and insufficient for estimating the distribution. Therefore, we calculate the path travel time by adding the travel time of all links in $R_i^{V,k}$ via Monte Carlo sampling. The link travel time distribution is presented in Section 2.2. Only the travel time distribution in time interval τ is used. We sample the travel time of each link along the path and then compute their sum to obtain a path travel time sample. This process is performed repeatedly until enough samples are generated for estimating the travel time distribution of $R_i^{V,k}$ (based on the kernel density method).

However, the link travel time samples are not available from the LPR data when no LPR devices are installed on this link. In this case, we propose a link travel time estimation method to generate synthesized travel time samples for these *unequipped* links. The proposed method is described in Appendix A.

After this procedure, the travel time samples of all links become available and can be used to estimate the path travel time distribution by using the aforementioned Monte Carlo and kernel density estimation methods. Therefore, $P(TT_{V,k}|R_i^{V,k})$ can be obtained by directly calculating the probability when the travel time is equal to $TT_{V,k}$ based on the derived path travel time distribution. The value of $P(TT_{V,k}|R_i^{V,k})$ will never be zero in the context of kernel density estimation; this phenomenon may lead to a *reluctant* path reconstruction when none of the path candidates should be considered as the real path. To avoid this unreasonable reconstruction, we artificially set $P(TT_{V,k}|R_i^{V,k})$ to zero when $TT_{V,k}$ does not lie within the 95% confidence interval of the derived path travel time distribution.

So far, Equation (1) can be divided into different parts that are calculated individually. The reconstructed path between discontinuous nodes $N_{V,k}^-$ and $N_{V,k+1}^+$ will be the one with the highest probability as shown below:

$$R_{\text{re}}^{V,k} = \underset{R_j^{V,k} \in R^{V,k}}{\operatorname{argmax}} \; P\left(R_j^{V,k} | N_d^{V,k}, TT_{V,k}, L_{V,k}\right) \tag{8}$$

If $R_{\text{re}}^{V,k}$ is not null, then we can successfully reconstruct the path between $N_{V,k}^-$ and $N_{V,k+1}^+$. The reconstructed path is then connected with the raw trajectory before $N_{V,k}^-$ and the raw trajectory after $N_{V,k+1}^+$. The newly connected path is seen as the reconstructed trajectory before $N_{V,k+1}^+$. If $R_{\text{re}}^{V,k}$ is null (i.e., $\max_{R_i^{V,k} \in R^{V,k}} P(R_i^{V,k}|N_d^{V,k}, TT_{V,k}, L_{V,k}) = 0$), then we treat the trajectory before $N_{V,k}^-$ and the trajectory after $N_{V,k+1}^+$ as two different trips to identify the stops along a chain of trips in LPR data.

Algorithm 1. Bayesian path reconstruction model

```
For each vehicle $V$ in LPR dataset, do
    Find the raw partial trajectories set $\boldsymbol{R}_{\text{partial}}^{V} = \{R_{\text{partial}}^{V,1}, R_{\text{partial}}^{V,2}, \cdots, R_{\text{partial}}^{V,n_p}\}$.
    For $k = 1{:}\,n_p - 1$ do
        Extract $N_{V,k}^-$ and $N_{V,k+1}^+$ as two discontinuous nodes to be reconstructed
        Find the path candidates set $\boldsymbol{R}^{V,k}$ between $N_{V,k}^-$ and $N_{V,k+1}^+$
        For each $R_i^{V,k}$ in $\boldsymbol{R}^{V,k}$ do
            Calculate Eq. (1) based on Eq. (2-9)
        end for
        $R_{\text{re}}^{V,k} = \underset{R_j^{V,k} \in \boldsymbol{R}^{V,k}}{\operatorname{argmax}} P(R_j^{V,k}|\boldsymbol{N}_d^{V,k}, TT_{V,k}, L_{V,k})$
        If $R_{\text{re}}^{V,k} \neq null$ then
            return $R_{\text{re}}^{V,k}$ as the reconstructed path between $N_{V,k}^-$ and $N_{V,k+1}^+$
        else
            Treat the trajectory before $N_{V,k}^-$ and the trajectory after $N_{V,k+1}^+$ as two different trips.
        end if
    end for
end for
```

A special case must be considered here: Calculating $P(R_i^{V,k}|L_{V,k})$ necessitates counting the number of vehicles with raw trajectories $R_i^{V,k}$. When $R_i^{V,k}$ is not an existing raw trajectory (which may occur when $R_i^{V,k}$ contains unequipped links), $\alpha_{R_i^{V,k}, L_{V,k}}$ cannot be obtained given the lack of any record. In this case, Equation (7) cannot be calculated. For this special case, we need to adjust the calculation method for $P(R_i^{V,k}|L_{V,k})$. Assume that $P(R_i^{V,k}|L_{V,k}) \propto P(R_i^{V,k})$, which neglects the impact of $L_{V,k}$. Although this assumption is not realistic, as mentioned above, there is no data that we can use to account for the effect of $L_{V,k}$ in this situation. $P(R_i^{V,k})$ is the prior probability of people using path $R_i^{V,k}$. We assume it is proportional to the path capacity (assuming the prior probability is a commonly used technique in Bayesian inference problem). As path capacity is determined by the minimum capacity of the links, this leads to $P(R_i^{V,k}) \propto \min_{\text{link } a \in R_i^{V,k}} \{\frac{1}{\varphi_a}\}$, where φ_a is a parameter inversely proportional to the capacity of link a, which is calculated in Appendix A. Therefore, in the case where $R_i^{V,k}$ contains unequipped links, $P(R_i^{V,k}|L_{V,k})$ can be rewritten as

$$P\left(R_i^{V,k}|L_{V,k}\right) = \frac{\min_{\text{link } a \in R_i^{V,k}} \left\{1/\varphi_a\right\}}{\sum_{j=1}^{n_c} \min_{\text{link } a \in R_j^{V,k}} \left\{1/\varphi_a\right\}} \tag{9}$$

The aforementioned path reconstruction model is summarized in Algorithm 1.

Unlike the methods employed in previous studies, the proposed Bayesian path reconstruction model does not rely on artificial assumptions, such as "people always prefer the shorter path." Instead, people's preferences are directly embedded into this model based on their choices. For example, Equations (5) to (7) capture the people's preferences according to their path choices. These equations may generate higher probabilities for shorter paths that correspond to artificial assumptions. However, these equations also capture the special case where shorter paths are not preferable. Given that we consider different LPR device installation scenarios (e.g., equipped and unequipped links), the proposed model is adaptive to different LPR device coverage rates. However, LPR systems are expected to have a high coverage rate given the purely data-driven property of the proposed model.

3 | DYNAMIC OD DEMAND ESTIMATION

3.1 | Formulation of objective function

Before formally describing the dynamic OD estimation model, we summarize the known information to formulate the objective function. We extract four types of information from the LPR data, including link flow, initial OD matrix, left-turning flow, and partial path flow. Our objective function attempts to minimize the difference between the observed (obtained from LPR data) and estimated information.

1. Link flow

The time-dependent link flow is a typical real-world information derived from LPR data. This flow can be directly obtained by counting the number of recognized and unrecognized vehicles passing through an LPR station. $v^*_{\tau,a}$ denotes the observed flow of link a in time interval τ, $v_{\tau,a}$ denotes the corresponding estimated link flow, and $\sum_\tau \sum_a (v^*_{\tau,a} - v_{\tau,a})^2$ can be an item of the objective function in the following OLS model.

2. Initial OD matrix

The initial OD matrix is crucial in OD estimation as revealed in previous studies (Cascetta et al., 1993; Cipriani et al., 2011; Dixon & Rilett, 2005). The most commonly used initial OD demand is the historical OD obtained from travel survey data (Ashok & Ben-Akiva, 1993; Zhou & Mahmassani, 2006). However, historical OD distribution data for the study area, which is a newly developing third-tier city, are unavailable. Even if such data exist, some huge differences may be observed between these data and the current OD distribution given the high speed of motorization and urbanization in the study area. Therefore, we derive a new source of initial OD matrix from the LPR data. After reconstructing the vehicle paths in Section 2, we can leverage these paths to derive an initial OD demand. Let $\hat{q}^{rs}_\tau$ be the initial OD flow between origin r and destination s within time interval τ, and let N^{rs}_τ be the number of reconstructed paths starting from r and ending at s within the same interval. We have $\hat{q}^{rs}_\tau = \alpha_\tau \cdot N^{rs}_\tau$, where α_τ is a scaling factor that considers the unreconstructed paths. α_τ can be approximated by $\alpha_\tau = v^{\mathrm{all}}_\tau / v^{\mathrm{part}}_\tau$, where v^{part}_τ is the sum of all link flows calculated from the reconstructed paths, and v^{all}_τ is the sum of all observed link flows. This equation is similar to the path-flow-reconstruction–based OD estimation method. Therefore, $\sum_\tau \sum_{r,s} (\hat{q}^{rs}_\tau - q^{rs}_\tau)^2$ can be used as another item in the objective function.

3. Left-turning flow

Raw LPR data can provide information about the lane being occupied by vehicles at each equipped intersection. Theoretically, the turning flow (left, right, and straight) of links can also be obtained if a lane is used for only one type of turning behavior. However, for many intersections, right-turning and straight-through lanes are often the same, thereby making the right-turning flow indistinguishable from the straight-through flow. For left-turning vehicles, specific left-turning lanes are generally used. Therefore, the left-turning flow provided by LPR data are close to the ground truth values. Even if some vehicles violate the traffic rules, this small proportion of errors will not affect the results of the entire model. It is worth noting that the above analysis is based on a typical scenario in Chinese cities. One important factor to be considered when choosing turning flow is its distinguishability, that is, whether or not this flow can be inferred from the occupied lane. Let the observed left-turning flow of link a within time interval τ be $LT^*_{\tau,a}$, and let $LT_{\tau,a}$ be the corresponding estimated left-turning link flow. The item of left-turning flow in the objective function can thus be expressed as $\sum_\tau \sum_a (LT^*_{\tau,a} - LT_{\tau,a})^2$. Unlike the link flow data, the left-turning flow data contain route choice information. Therefore, these data can introduce additional constraints in OD matrix estimation and reduce the impact of underdetermined problems (Mishalani, Coifman, & Gopalakrishna, 2002).

4. Partial path flow

Previous studies that estimate OD based on traffic counts and seed matrix have insufficient information about real-world route choices. This lack of information may lead to an underdetermination problem, in which the OD estimates deviate from the true values even if the derived traffic counts are relatively accurate. Route choice information is important in mitigating such a problem (Rao et al., 2018). A typical method for adding route choice information is utilizing probe path flow data, which are generally used in mobile-phone- and GNSS-data–based research (X. Yang, Lu, & Hao, 2017). Given that probe path flow data include the *complete* trajectories of probe vehicles, they can be easily embedded into the estimation model by scaling with the penetration rate. However, in an LPR-data–based model, the reconstructed partial path flow data include the *incomplete* trajectories of vehicles. Therefore, the relationship between the reconstructed partial path flow and true path flow is not trivial. Some LPR-based studies assume that LPR stations can detect all original ODs, thus the observed path is complete (Antoniou, Ben-Akiva, & Koutsopoulos, 2004), which they admitted is a restrictive requirement in the real world. In this section, we will address how to incorporate partial path flow with incomplete trajectories. Let $f^{r's'*}_{i,\tau}$ be a reconstructed path flow within time interval τ between OD r' and s', derived from the reconstructed path data, and let $f^{rs}_{j,\tau}$ be a true path flow between OD r and s within time interval τ. We now consider the relationship between $f^{r's'*}_{i,\tau}$ and $f^{rs}_{j,\tau}$, where $\boldsymbol{R}^*_{i,\tau}$ and $\boldsymbol{R}_{j,\tau}$ denote the corresponding path trajectories. For example, let ①, ②, ③, and ④ be four consecutive nodes, path $\boldsymbol{R}^*_{i,\tau}$ (②, ③) be the reconstructed trajectory, and path $\boldsymbol{R}_{j,\tau}$ (①, ②, ③, ④) be the complete true trajectory. $\boldsymbol{R}_{j,\tau}$ may become $\boldsymbol{R}^*_{i,\tau}$ if nodes ① and ④ are misdetected. In this situation, the volume of $f^{14}_{j,\tau}$ contributes to the volume of $f^{23*}_{i,\tau}$. The contribution rate from $f^{rs}_{j,\tau}$ to $f^{r's'*}_{i,\tau}$ can be formulated as

$$\Delta^{i,j,r,s}_\tau = \sum_{N_d} P\left(N_d | \boldsymbol{R}_{j,\tau}\right) \cdot P\left(\boldsymbol{R}^*_{i,\tau} | N_d\right) \tag{10}$$

where N_d is the set of detected nodes of $R_{j,\tau}$, and $\sum_{N_d}$ is the sum over all possible detected node situations. Equation (10) refers to the probability that only N_d of $R_{j,\tau}$ is recorded, and then N_d is reconstructed to $R^*_{i,\tau}$. This probability is defined as the contribution rate from $f^{rs}_{j,\tau}$ to $f^{r's'*}_{i,\tau}$. $P(N_d|R_{j,\tau})$ can be calculated by using Equation (2), whereas $P(R^*_{i,\tau}|N_d)$ can be calculated as

$$P\left(R^*_{i,\tau}|N_d\right) = \frac{P\left(N_d|R^*_{i,\tau}\right)\cdot P\left(R^*_{i,\tau}\right)}{\sum_{j=1}^{N_R} P\left(N_d|R_{j,\tau}{}^*\right)\cdot P\left(R_{j,\tau}{}^*\right)} \tag{11}$$

where $P(N_d|R_{i,\tau}{}^*)$ and $P(R_{i,\tau}{}^*)$ can be calculated using Equations (2) and (5), respectively. Through this equation, the relationship between $f^{rs}_{j,\tau}$ and $f^{r's'}_{i,\tau}{}^*$ is expressed as $f^{r's'}_{i,\tau}{}^* = \sum_{j,r,s}\Delta^{i,j,r,s}_{\tau}\cdot f^{rs}_{j,\tau}$, and the item in the objective function can be formulated as $\sum_{\tau}\sum_{i,r',s'}(f^{r's'}_{i,\tau}{}^* - \sum_{j,r,s}\Delta^{i,j,r,s}_{\tau}\cdot f^{rs}_{j,\tau})^2$. It is worth noting that using the partial path flow incorporates a good property in Bayesian path reconstruction method. Given that the path flow is reconstructed based on real-world travel data (such as travel time), the model will tend to distribute the demand over those paths that are analogous to real-world situations, thereby facilitating the OD estimation.

By now, four items with respect to link flow, initial OD matrix, left-turning flow, and partial path flow are available for the objective function, which will be integrated by different weights. The function of these four items can be explained as follows. $\hat{q}^{rs}_{\tau}$ is a rough OD reference because the reconstructed path flow may lose the original origin and destination (as discussed in Section 1). Therefore, this parameter is used to avoid an extremely unreasonable solution and to provide scale information. The ground truth observations $v^*_{\tau,a}$ and $LT^*_{\tau,a}$ can somehow offset the error in $\hat{q}^{rs}_{\tau}$. The partial path flow ($f^{r's'}_{i,\tau}{}^*$) can provide important route choice information.

3.2 | Model formulation

Previous OD estimation studies have widely used the bilevel model as their framework (Tavana, 2001; Wen, Gardner et al., 2018; Zhou & Mahmassani, 2006). The upper level of this model estimates the demand assignment matrix based on the user equilibrium assumption (or by using a traffic simulation software), whereas the lower level estimates the OD matrix based on the calculated assignment matrix. These two models are executed iteratively until convergence is reached. Using this bilevel model is practical when the user equilibrium assumption (or the traffic simulation software) corresponds well to real-world traffic situations (Duthie, Unnikrishnan, & Waller, 2011). Therefore, the network parameters (e.g., road capacity) should be precalibrated before the OD estimation. This procedure may be especially difficult to apply in new cities with a rapidly changing traffic situation and high motorization and urbanization rates. The user equilibrium assumption may also not hold in such cases (Yildirimoglu & Kahraman, 2017). Therefore, a method that avoids the precalibration procedure should be developed.

Given that we can partly capture the route choice information by using LPR data, the user equilibrium assumption can be relaxed, and consequently avoid the precalibration of network parameters. We propose a two-step OLS estimation model to infer the dynamic OD matrix based on the information aggregated in Section 3.1. This model is formulated as follows:

- P estimation step (q^{rs}_{τ} is constant):

$$\begin{aligned}\min_{p^{rs}_{j,\tau}}\ J_1 = \sum_{\tau}\Bigg(& w_1\sum_{a}(v^*_{\tau,a}-v_{\tau,a})^2 \\ & + w_2\sum_{a}\left(LT^*_{\tau,a}-LT_{\tau,a}\right)^2 + w_3 \\ & \times\sum_{i,r',s'}\left(f^{r's'*}_{i,\tau}-\sum_{j,r,s}\Delta^{i,j,r,s}_{\tau}\cdot f^{rs}_{j,\tau}\right)^2\Bigg)\Bigg)\end{aligned} \tag{12}$$

$$\text{s.t.}\begin{cases} f^{rs}_{j,\tau} = p^{rs}_{j,\tau}\ \cdot q^{rs}_{\tau},\,|\quad \forall\tau,r,s,j \\ \sum_{j} p^{rs}_{j,\tau} =\ 1,|\quad \forall\tau,r,s \\ v_{\tau,a} = \sum_{r,s,j} f^{rs}_{j,\tau}\ \cdot\delta^{rs}_{a,j},|\quad \forall\tau,a \\ LT_{\tau,a} = \sum_{r,s,j} f^{rs}_{j,\tau}\ \cdot\sigma^{rs}_{a,j},|\quad \forall\tau,a \\ f^{rs}_{j,\tau},\ v_{\tau,a},\ LT_{\tau,a},\ p^{rs}_{j,\tau}\geq 0,|\quad \forall\tau,r,s,j,a \end{cases} \tag{13}$$

- Q estimation step ($p^{rs}_{j,\tau}$ is constant):

$$\begin{aligned}\min_{q^{rs}_{\tau}}\ J_2 = \sum_{\tau}\Bigg(& w_1\sum_{a}\left(v^*_{\tau,a}-v_{\tau,a}\right)^2 \\ & + w_2\sum_{a}\left(LT^*_{\tau,a}-LT_{\tau,a}\right)^2 + w_3 \\ & \times\sum_{i,r',s'}\left(f^{r's'*}_{i,\tau}-\sum_{j,r,s}\Delta^{i,j,r,s}_{\tau}\cdot f^{rs}_{j,\tau}\right)^2 \\ & + w_4\sum_{r,s}\left(\hat{q}^{rs}_{\tau}-q^{rs}_{\tau}\right)^2\Bigg)\Bigg)\end{aligned} \tag{14}$$

$$
\text{s.t.}\begin{cases}
f_{j,\tau}^{rs} = p_{j,\tau}^{rs} \cdot q_{\tau}^{rs}, \mid \quad \forall \tau, r, s, j \\
v_{\tau,a} = \sum\limits_{r,s,j} f_{j,\tau}^{rs} \cdot \delta_{a,j}^{rs}, \mid \quad \forall \tau, a \\
LT_{\tau,a} = \sum\limits_{r,s,j} f_{j,\tau}^{rs} \cdot \sigma_{a,j}^{rs}, \mid \quad \forall \tau, a \\
f_{j,\tau}^{rs},\ v_{\tau,a},\ LT_{\tau,a}, q_{\tau}^{rs} \geq 0, \mid \quad \forall \tau, r, s, j, a
\end{cases} \qquad (15)
$$

where $p_{j,\tau}^{rs}$ is the OD assignment proportion from demand q_{τ}^{rs} to path flow $f_{j,\tau}^{rs}$, $\delta_{a,j}^{rs}$ is the typical link–path incidence, $\sigma_{a,j}^{rs}$ is the left-turning link–path incidence, $\sigma_{a,j}^{rs} = 1$ if link a and the left-turning link after a are both in path j connected by r and s, and w_i is the weight factor that can be calibrated by using the ideal point method of Zhou, Qin, and Mahmassani (2003). Several combinations of w_i are initially generated based on Latin Hypercube and are subsequently used to evaluate the OD demand. A quadratic function is then applied as the surrogate function to fit the weights and the corresponding OD estimation error. The optimal weights with the minimal OD estimation errors are selected as the final weights, that is, $w_1 = 0.2$, $w_2 = 0.5$, $w_3 = 0.5$, and $w_4 = 0.1$. The proposed model has two steps, namely, the assignment proportion $\boldsymbol{P}$ estimation step and the OD demand $\boldsymbol{Q}$ estimation step, both of which are typical quadratic programming problems that can be solved by many methods (e.g., interior point method) and are certain to converge to a globally optimal solution. The $\boldsymbol{Q}$ estimation step is similar to the aforementioned upper level model, with the difference lying in its incorporation of partial path flow information. In the $\boldsymbol{P}$ estimation step, we use the link flow, left-turning flow, and partial path flow from LPR data to infer the demand assignment proportion $p_{j,\tau}^{rs}$, which relaxes the predetermined traffic assignment assumptions. The solution method of this model is summarized in Table 1.

4 | MODEL VALIDATION

4.1 | Validation design

To validate the proposed model, we collected the real-world LPR data on November 11, 2013, in Langfang City, China. The road network of Langfang is shown in Figure 2. There are total of 100 intersections and 347 links in the road network. The red sectors (circles) means there is an LPR camera installed in this intersection direction, whereas the no circle at each intersection indicates that no LPR system is installed in this direction. This network had a device coverage rate (i.e., data penetration rate) of 53.8%, which was computed by dividing the number of existing LPR devices by the total number of devices needed to cover all intersections. The red line in Figure 3 indicates the time-varying average recognition rate of the LPR devices for every 30 minutes on November 11.

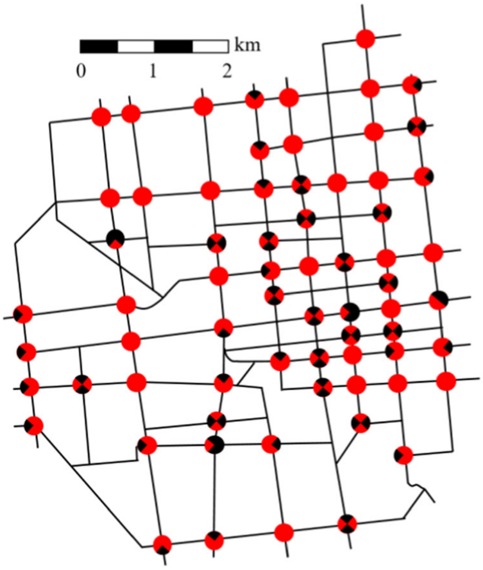


FIGURE 2 Road network of Langfang City

TABLE 1 Two-step dynamic OD estimation model solution method

Step 1:	Initialization. Starting from $k = 1$, denote $q_{\tau}^{rs} = \hat{q}_{\tau}^{rs}$, run the P estimation step to obtain the assignment proportion of first iteration $p_{j,\tau}^{rs^{(1)}}$.
Step 2:	Q estimation step. Set $p_{j,\tau}^{rs} = p_{j,\tau}^{rs^{(k)}}, \forall \tau, r, s, j$. Run the Q estimation step to obtain the dynamic OD demand $q_{\tau}^{rs^{(k)}}$.
Step 3:	P estimation step. Set $q_{\tau}^{rs} = q_{\tau}^{rs^{(k)}},\ \forall \tau, r, s$. Run the P estimation step to obtain the dynamic demand assignment matrix $p_{j,\tau}^{rs^{(k+1)}}$.
Step 4:	Convergence test. Run the Q estimation step to obtain $q_{\tau}^{rs^{(k+1)}}$ using $p_{j,\tau}^{rs^{(k+1)}}$. Calculate the average deviation between all $q_{\tau}^{rs^{(k+1)}}$ and $q_{\tau}^{rs^{(k)}}$. If the deviation is less than a predetermined threshold, i.e., $\frac{1}{M}\sum_{r,s,\tau} \frac{\lvert q_{\tau}^{rs^{(k+1)}} - q_{\tau}^{rs^{(k)}} \rvert}{q_{\tau}^{rs^{(k+1)}}} < \varepsilon$, stop and let $q_{\tau}^{rs^{(k+1)}}$ be the final estimation OD matrix; otherwise, set $k = k + 1$ and go to Step 2. Note: M is the total number of (r, s, τ) combinations, and ε is the threshold (set as 1% in this study), which can be determined by observing the convergence curve.

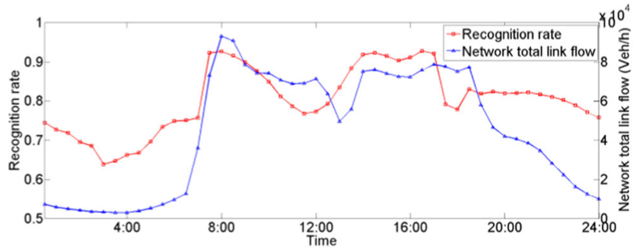


FIGURE 3 Recognition rate of LPR devices on November 11

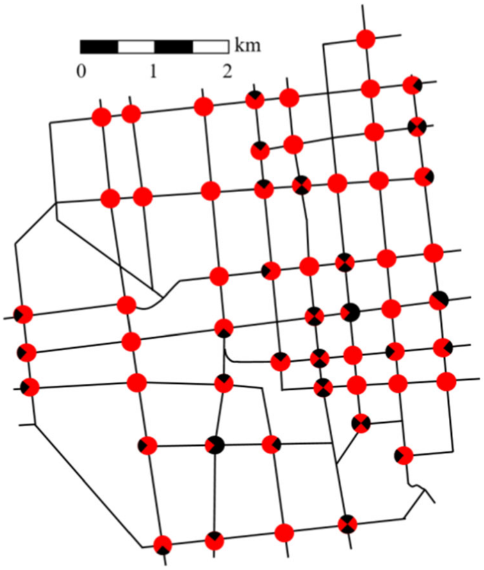


FIGURE 4 Simplified road network of Langfang for the simulation

Given that the LPR system is affected by lighting, the recognition rate was relatively high at day and low at night. The slight drop in recognition rate around 12:00 may be ascribed to the intense illumination at noon. The average recognition rate for the whole day was 80.3%. The blue line in the figure denotes the time-varying network link flow, which is computed as the sum of all data-available link flows recorded by LPR devices. This flow also provides an overview of the total number of recognized vehicles. Given that Langfang is a third-tier city in China, its traffic volume is not too high, and some congested links are observed during the morning and evening peak hours (Li, Liu, & Zhang, 2018). However, most of these links become uncongested during off-peak hours.

Given the extreme difficulty in obtaining the ground truth OD demand in an urban road network, we cannot quantitatively validate the estimation results by using real-world LPR data. Therefore, the real-world LPR data are only used for the qualitative validation (Section 4.3.1). To quantitatively evaluate the proposed model, we generated a synthetic data set based on a simplified Langfang road network (Figure 4). OD estimation models are often built in the literature by using synthetic data to do the validation (Rao et al., 2018; Zhou & Mahmassani, 2007). Compared with the true network, the simplified network omits some low-level roads. This network contains 68 intersections and 237 links, although its LPR camera layout is the same as the real-world layout. However, by ignoring some low-level roads, a slight increase is observed in the coverage rate of the simplified network (57.7%). Based on the real-world link flow, we generated and fine-tuned the *synthetic OD demand* by using the OD estimator in a traffic simulation program (i.e., TransModeler). The synthetic OD demand was used by the Lang Fang Bureau of Traffic Management and was verified to be practical for real-world analysis. We loaded the synthetic OD demand into the simplified network by using TransModeler and set the route choice criteria of drivers as purely dynamic. Specifically, these drivers are allowed to change their routes based on their time-dependent expected travel time. They are also assumed to know the perfect information. We obtained the complete trajectories of each vehicle after running the program for a whole simulation day. Afterward, we discarded the information of unequipped roads, and randomly discarded some records for equipped roads based on the recognition rate for each LPR station. The remaining vehicle records were treated as the *synthetic LPR data set* that will be used to run the proposed OD estimation model. Although some previous studies assume that the link volume is perfectly reliable (Dixon & Rilett, 2002; van der Zijpp & Hamerslag, 1994), to be conservative, we added 5% and 10% random errors to the derived link flow and left-turning flow, respectively. The synthesization was repeated 10 times. All following estimation results for synthesized data are shown in average.

4.2 | Validation of path reconstruction

Given the importance of the path reconstruction results in the OD estimation, we evaluate the performance of the proposed Bayesian path reconstruction model with real-world and synthetic data. Two evaluation indicators are used. The first indicator, *completeness of reconstruction*, is an amount-based indicator that indicates how much flow can be reconstructed. The second indicator, *accuracy of reconstruction*, indicates whether the reconstructed trajectories are the same as the complete trajectories.

4.2.1 | Validation using real-world LPR data

The real-world LPR data can reflect actual route choice behaviors. Using these data in the validation can also justify the applicability of the proposed path reconstruction model in real-world situations.

The completeness of the path reconstruction is shown in Figure 5. In this figure, the red line denotes the ground truth link flow that can be directly derived from the LPR data, the blue line denotes the reconstructed link flow that can be computed as the sum of the reconstructed path flow that

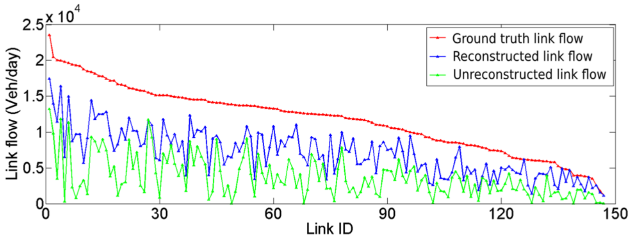


FIGURE 5 Completeness of path reconstruction using real-world LPR data

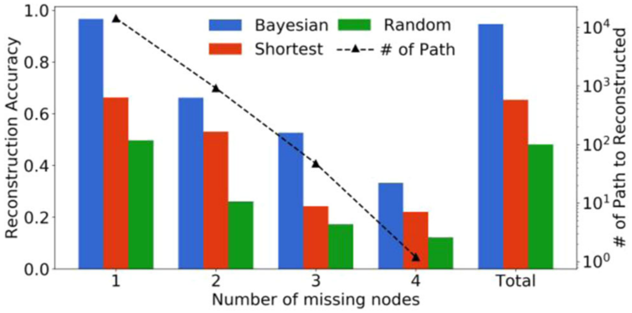


FIGURE 6 Accuracy of path reconstruction using real-world LPR data

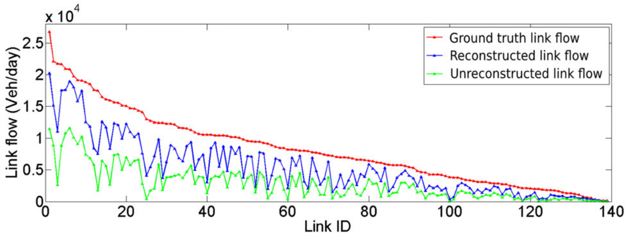


FIGURE 7 Completeness of path reconstruction using synthetic LPR data

contains the specific link, and the green line denotes the unreconstructed link flow that can be computed as the sum of the raw unreconstructed path flow that contains a specific link. We sort the link ID by the ground truth link flow. As shown in Figure 5, the reconstructed link flow is closer to the ground truth compared with the unreconstructed link flow. The average ratios of the unreconstructed and reconstructed link flows to the ground truth link flow are 0.312 and 0.604, respectively, which suggest that approximately 29.2% of the additional flow information has been filled after the reconstruction. The variance of the reconstructed link flow curve is also highly consistent with that of the ground truth curve, which implies that the reconstructed information is more harmonious than the unreconstructed one in terms of missing data rate. Therefore, the scaling method can produce better results by using the reconstructed data (e.g., the scaling of penetration rate of initial OD demand, see Section 3.1).

The accuracy of the reconstruction is evaluated afterward. However, given that the true complete trajectories of real-world LPR data are unavailable, we generate a test data set based on real-world data. We regard the existing raw trajectories derived from LPR data as *dummy complete* trajectories that may lack several nodes compared with the true complete trajectories. Nevertheless, these trajectories can still reflect some route choice behaviors. Based on the recognition rate of the LPR system, we randomly discard several nodes from the dummy complete trajectories and then generate the real-world test data set for estimating the reconstruction accuracy. This process is similar to our procedure for generating the synthetic simulation data set. We run the reconstruction model by using the real-world test data set, repair the discarded nodes, and compare the reconstructed trajectories with the dummy complete trajectories. A shortest path reconstruction method and a random draw method are set as benchmarks to compare with the Bayesian model. The shortest path method means we use the shortest distance path as the reconstructed path. The random draw method means every path candidate is equally likely to be chosen as the reconstructed path. We run these models for 30 replications and average the results to avoid random errors. The proposed model demonstrates a stable performance despite showing a small variance. The accuracy of the reconstruction results is shown in Figure 6, where the right axis shows the number of paths to reconstruct and the left axis shows the successful reconstruction accuracy. A reconstruction is considered successful if the reconstruction path is exactly the same as the initial path. Our proposed model demonstrates a superior performance across all scenarios compared with the benchmarks. In line with our intuition, the number of missing nodes increases along with a decreasing reconstruction accuracy. Given that the test data set has been generated from raw trajectories that are always short and incomplete, the maximum number of missing nodes is 4. Most path reconstruction problems have only one missing node that leads to an artificial impression of high reconstruction accuracy. The real-world test data achieve a reconstruction accuracy of 96.0%. The following validation using synthetic data will properly show the model performance.

4.2.2 | Validation using synthetic LPR data

Given that the complete trajectory of each vehicle is known in the synthetic data set, we can directly evaluate the path reconstruction performance of the proposed model.

As discussed in Section 4.2.1, we initially evaluate the completeness of path reconstruction (Figure 7), and the results are similar to those obtained by using real-world LPR data. Compared with the unreconstructed link flow, the reconstructed link flow is closer to the ground truth and is more consistent with respect to its change trend. The average ratios of the unreconstructed and reconstructed link flows to the ground truth link flow are 0.337 and 0.642, respectively, thereby suggesting

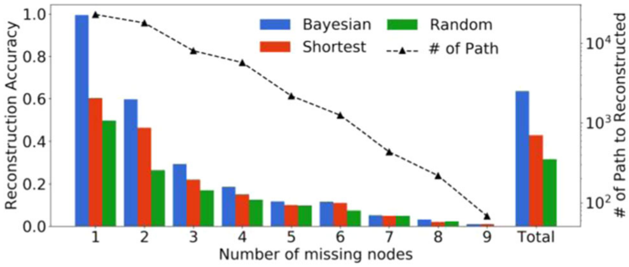


FIGURE 8 Accuracy of path reconstruction using synthetic LPR data

that approximately 30.5% of the additional flow information has been repaired after the reconstruction.

Given that the true complete trajectories are known, we can directly test the reconstruction accuracy of the proposed model by using synthetic LPR data. The validation results are shown in Figure 8. The maximum number of missing nodes increases to 9, which shows a more diverse situation in path reconstruction problems. Meanwhile, the reconstruction accuracy decreases along with an increasing number of missing nodes. Our model outperforms both benchmarks when the number of missing nodes is less than 5. When this number increases, the number of possible path candidates exponentially increases, thereby making the true path untraceable. However, given the small number of paths with more than 5 missing nodes, the total reconstruction accuracy can reach as high as 62.3%, and this value objectively reflects the performance of our proposed Bayesian path reconstruction model. Future studies may explore more advanced methods for solving the path reconstruction problem in terms of high missing nodes scenarios.

4.3 | Validation of OD demand estimation

We divide each day into 48 time intervals with 30 minutes each given that most trips can be completed within 30 minutes. The time interval is seen as the minimum time unit for OD estimation. This is actually a quasi-dynamic modeling framework, which is widely used in previous study (Bierlaire & Crittin, 2004; Zhou & Mahmassani, 2006). Future research can be done to explore the pure-dynamic framework.

Two definitions of OD have been adopted in the literature. The first definition, which describes OD as a road network node, agrees with the transportation network model and can be conveniently used for modeling purposes. The second definition, which describes OD as a block of land (e.g., a community or traffic analysis zone), is in line with reality and can be conveniently used for transportation planning purposes. In this study, we first define ODs as network nodes for modeling. In the result-display section, we aggregate the node-based OD flow into block-based OD flow to show the estimation results. In this way, we can not only easily process the model but also export the applicable OD demand for transportation planning. Because we know how the TransModeler assigns block-based OD to the nearby intersections, we can do the inverse transformation to aggregate node-based OD to block-based OD. Thus, the transform between node-based and block-based OD will not produce error to the estimation accuracy in our synthetic validation.

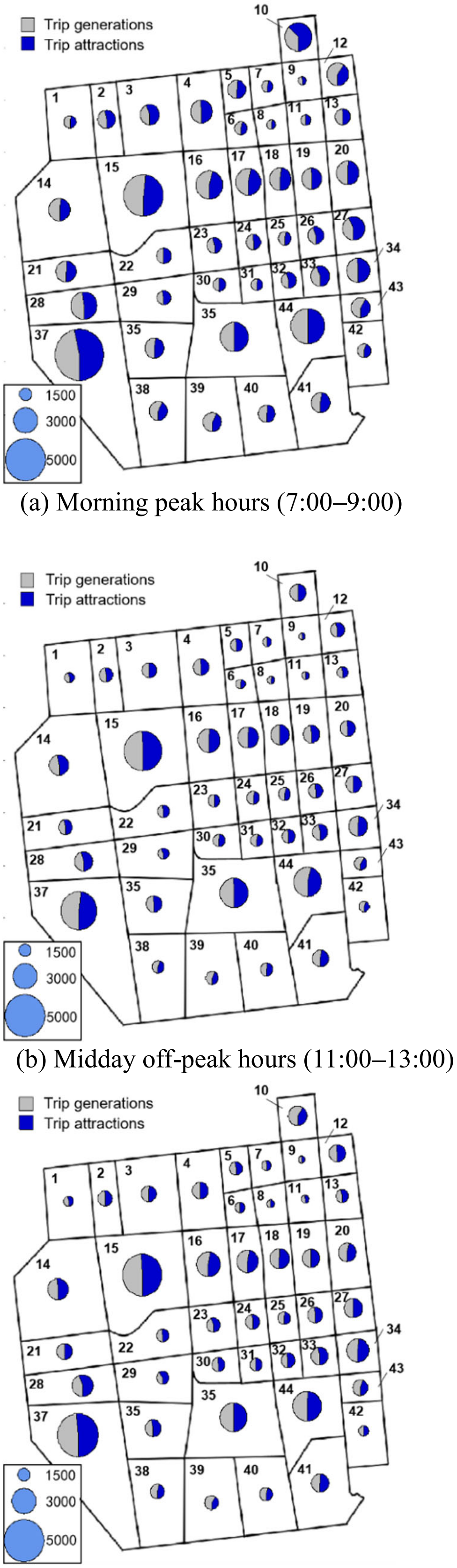
We divide the urban area into 44 blocks according to the traffic analysis zones, and each of these blocks can be considered a block-based origin or destination.

4.3.1 | Qualitative validation using real-world LPR data

Given that information on the ground truth OD demand is unavailable, we use real-world LPR data for the *qualitative* model validation. That is, we will check whether the estimated OD patterns from real-world LPR data are consistent with the context of the corresponding land use, urban design, and traffic conditions. The estimated OD demand distribution during morning peak (7:00–9:00), midday off-peak (11:00–13:00), and evening peak hours (17:00–19:00) are plotted in Figure 9. The areas of circles in this figure denote the OD flow (veh/h) of a specific block, the gray part denotes trip generation, and the blue part denotes trip attraction. The number in each block represents the block ID. Take block 10 as an example. The OD distribution of this block shows obvious commute characteristics. Specifically, during morning peak hours, the trip attractions in this block is higher than the trip generations, but the opposite is observed during evening peak hours because block 10 acts as the main artery in an economic development zone where many people work but few people live. Meanwhile, blocks 15 to 20 show a higher OD demand compared with the other blocks. These zones are located in the central area of a city with a large population and mixed residential and commercial areas. Therefore, these blocks have a relatively high traffic flow and a balanced trip generation and attraction. The OD flow in block 37 is also relatively high because of its location close to the freeway entrance. Therefore, many vehicles pass through this block every day to enter or leave the city.

In terms of temporal dimension, the total level of OD demand during midday off-peak hours is less than that during morning and evening peak hours, and this observation is in line with our intuition. The commute characteristic is not obvious in many blocks because Langfang is a third-tier city in China. For the convenience of citizens, no apparent boundary can be observed between the living and working areas from an urban planning perspective.

In sum, the OD demand estimated by real-world LPR data is in line with our expectations and qualitatively highlights the effectiveness of our proposed OD estimation model.



(a) Morning peak hours (7:00–9:00)

(b) Midday off-peak hours (11:00–13:00)

(c) Evening peak hours (17:00–19:00)

FIGURE 9 Estimated OD demand distribution using real-world LPR data

4.3.2 | Quantitative validation using synthetic LPR data

Given that the complete true trajectories of all vehicles are known, we quantitatively validate the performance of our proposed model by using synthetic LPR data. We use the following benchmark models for the comparison:

- Benchmark 1: The naïve trajectory count (NTC) model

The NTC model (Ashok, 1996) directly extracts OD demand from raw LPR data. The OD demand between origin r and destination s can be formulated as $q_\tau^{rs} = \alpha_\tau \cdot x_\tau^{rs}$, where x_τ^{rs} is the number trajectories with origin r and destination s, and α_τ is the expansion factor that accounts for the unrecognized vehicles. In practice, a vehicle may have several trajectories in a single day. Therefore, a maximum dwell time of 30 minutes is used for this benchmark model to split the trajectories. α_τ is set equal to its value calculated in Section 3.1.

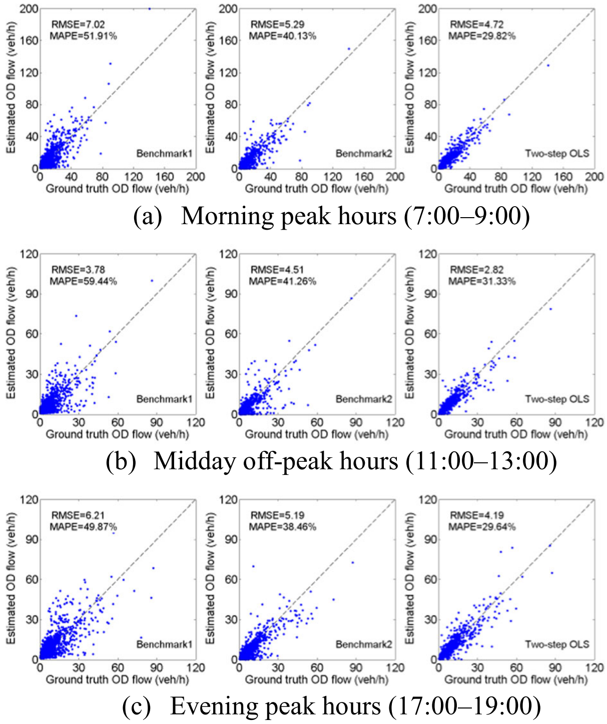
- Benchmark 2: Bilevel OD estimation model

The bilevel OD estimation model (Lundgren & Peterson, 2008; Tavana, 2001; Tavana & Mahmassani, 2001) is often used when the network link flow is available. The upper level minimizes the deviation between the overserved and estimated data, whereas the lower level estimates the path choice fraction under the user equilibrium constraint given a specific OD matrix. Previous studies have mostly used link flow and historical OD demand in their objective functions. Therefore, we follow the same framework and include these two items in the bilevel OD benchmark model. The historical OD matrix for this model is set equal to the initial OD demand derived in Section 3.1. The formulation of this model is illustrated in Appendix B.

We choose the above benchmark models for three reasons. First, benchmark 1 is a model-free method typically used in the industry. Many data analysts without traffic knowledge use the naïve method to count the rough OD demand, which in turn can provide a lower bound for accuracy. Second, the bilevel model that uses link flow and historical OD demand as inputs is a classical formulation that has been used as a benchmark in recent studies (Antoniou et al., 2016; Walpen, Mancinelli, & Lotito, 2015). Third, although the bilevel model has other advanced extensions, these extensions require additional information as inputs that may not be directly available in LPR data.

We evaluate the estimation accuracy of the proposed model based on root mean square error (RMSE) and MAPE, which are computed as

$$RMSE = \sqrt{\frac{\sum_{r,s,r\neq s}\left(q_\tau^{rs} - \tilde{q}_\tau^{rs}\right)^2}{N_\mathrm{b} \times \left(N_\mathrm{b} - 1\right)}} \quad \text{and} \qquad (16)$$



(a) Morning peak hours (7:00–9:00)

(b) Midday off-peak hours (11:00–13:00)

(c) Evening peak hours (17:00–19:00)

FIGURE 10 Estimation results of the benchmark and two-step OLS models using synthetic LPR data

$$MAPE = \frac{\left(\sum_{r,s,r\neq s} \left|q_\tau^{rs} - \tilde{q}_\tau^{rs}\right| / \overline{q}_\tau\right)}{N_{\mathrm{b}} \times \left(N_{\mathrm{b}} - 1\right)} \cdot 100\% \qquad (17)$$

where q_τ^{rs} is the estimated OD flow from block r to s within time interval τ, $\tilde{q}_\tau^{rs}$ is the corresponding true OD flow (synthetic OD flow), $\overline{q}_\tau$ is the average true OD flow within time interval τ, N_{b} is the number of blocks that can be regarded as ODs, and $N_{\mathrm{b}} \times (N_{\mathrm{b}} - 1)$ is the total number of OD pairs.

A series of numerical tests is performed to test the convergence of the proposed two-step OLS model. The test results reveal that around 40 iterations are enough to make the algorithm reach convergence. The estimation results for the morning peak (7:00–9:00), midday off-peak (11:00–13:00), and evening peak hours (17:00–19:00) are presented in Figure 10. The left two graphs present the results of the benchmark models, whereas the right graph presents the results of the proposed two-step OLS model. These figures show that the two-step OLS model obtains better results than both benchmark models for all three periods and demonstrates slight differences in its estimation accuracy across these periods. Specifically, the results for the morning and evening peak hours are better than those for the midday off-peak hours. Some extreme values located far from the dashed line can also be observed in the graph. However, these poor estimates, which are generally for OD pairs with long distance, are reasonable because a farther OD distance indicates that less information can be obtained from the LPR devices, thereby leading to poor estimation results.

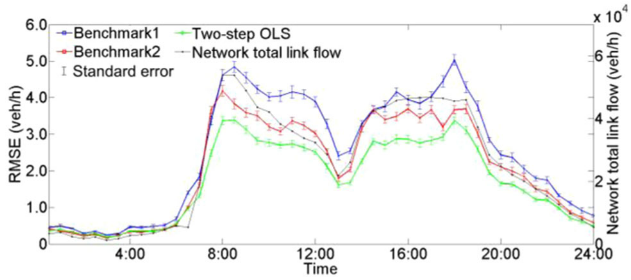


FIGURE 11 RMSE of the benchmark and two-step OLS models

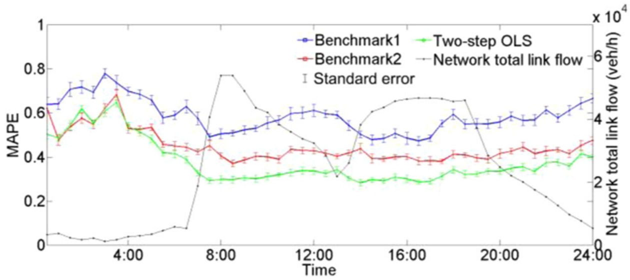


FIGURE 12 MAPE of benchmark models and the two-step OLS model

The RMSE and MAPE across different time intervals are plotted in Figures 11 and 12 along with the standard error of 10 replications. The proposed two-step OLS model obtains a lower RMSE and MAPE compared with the benchmark models for nearly all 48 time intervals. In most time intervals, the curve of the proposed model is twice greater than the standard deviation of the baseline models, thereby indicating that this model has a significantly superior performance compared with the benchmarks. Meanwhile, the bilevel model (benchmark 2) outperforms the NTC model (benchmark 1). These benchmark models have average RMSEs of 2.54 veh/30 min and 3.27 veh/30 min, respectively, which are higher than that of the proposed two-step OLS model (2.05 veh/30 min). In terms of MAPE, the weighted (by the quantity of OD flow) average MAPEs of the first and second benchmark models are 55.93% and 41.78%, respectively, which are both higher than that of the two-step OLS model (32.73%). Therefore, we report a decrease of around 10% in the relative error.

The RMSE changes along with the total link flow in the network because this parameter is a quantity-related absolute error index. The RMSE difference between the benchmark models and the proposed model is relatively small at night and relatively high during the day. Meanwhile, MAPE is a relative error index that remains relatively harmonious for the daytime time intervals, thereby suggesting that the proposed model remains stable across different traffic flow scenarios. Interestingly, the MAPE of the proposed model is relatively large late at night (0:00–6:00) because of the low recognition rate of LPR devices during this time (Figure 4). These devices

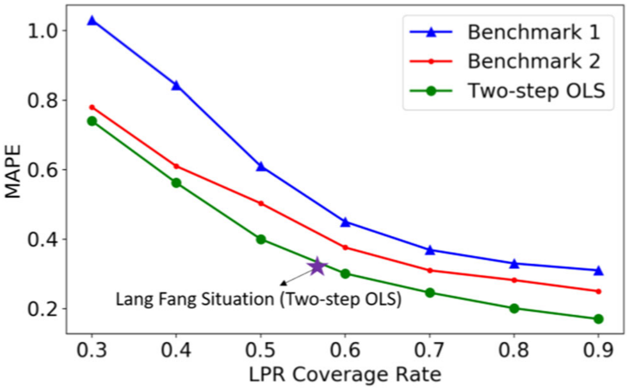


FIGURE 13 MAPE of different LPR coverage rates

can provide less information for running the model, which in turn leads to a relatively high MAPE.

In terms of MAPE, the proposed model outperforms the bilevel model by only about 5% to 10%. However, the bilevel model requires a precalibration of the network parameters (e.g., link impedance function). In this study, we provide the bilevel model with a reasonable set of network parameters that are derived from a simulation software. These parameters may require great effort to obtain in real-world situations. However, the proposed two-step OLS model does not require a precalibration of the network parameters, thereby making this model highly practical in developing cities with high motorization and urbanization rates.

4.3.3 | Impact of different LPR coverage rates

Many studies have shown that the estimation quality of the LPR data-based method is dependent on the LPR coverage rates (Rao et al., 2018; Zhou & Mahmassani, 2006). Testing of different LPR coverage rates can demonstrate (1) whether the method can be applied to a large-scale network with a low-density LPR system and (2) the minimal requirement for data sources. To generate synthetic LPR data with different coverage rates, we randomly discard the LPR records of some intersections, yielding a data set of seven different coverage rates: 90%, 80%, 70%, 60%, 50%, 40%, and 30%. The recognition rate of all LPR devices is set as the average value of those in the Lang Fang case, that is, 80.3%. Figure 13 presents the results at different coverage rates, where one point represents the average MAPE from 8:00 to 22:00. We find that the MAPE of all methods decreases with an increase in the coverage rates. The change is relatively smooth when the coverage rate is between 90% and 60%. However, when the sampling rate is less than 50%, the MAPE becomes larger than 40% and increases rapidly. Thus, to achieve a reliable OD demand estimation, the LPR coverage rate should be above 50%. This corresponds with previous research (Rao et al., 2018). Furthermore, in all testing scenarios, the proposed method can outperform two benchmark models.

5 | CONCLUSION AND DISCUSSION

This article proposes a hybrid framework for dynamic OD demand estimation that fully exploits the information available in LPR data. The LPR data used in this study contain information regarding the passing time, location, and lanes occupied by vehicles. Information sources other than LPR data are not required in this model. A Bayesian path reconstruction model is initially developed to replenish the information loss resulting from the recognition error and insufficient penetration rate of the LPR system. Based on the reconstructed data, we derive the link flows, initial OD demand, left-turning flows, and partial path flows to increase the amount of available information from LPR data, given that some information is not directly available in the raw data set. Afterward, we formulate a two-step OLS OD estimation model by using all the above information. The proposed framework is qualitatively validated by using real-world LPR data collected from Langfang City, China, and quantitatively validated by using synthesized simulation data in a simplified road network in Langfang. Results show the proposed model can estimate well the OD demand distribution and shows significant improvements in estimation accuracy compared with the NTC benchmark model. However, when compared with the bilevel OD estimation model, despite showing only slight improvements in accuracy, the proposed two-step OLS model does not require the precalibration of road network parameters, thereby making this model more practical in cities where road network parameters are difficult to calibrate.

Although this study obtains promising results, some of its limitations warrant further exploration in future work. First, this study shows that to obtain reliable OD demand estimation results, the LPR coverage rate should be above 50% (when the LPR recognition rate is 80%). However, installing a large number of LPR devices in large cities can be cost inefficient and impractical, and some cities may have lower LPR recognition rates compared with others. Therefore, our method may be more practical in developing cities where LPR systems have a relatively high coverage rate and where historical OD demand is not available. For large-scale cities with low LPR coverage rates, two possible solutions are applicable: (1) If the LPR devices are uniformly distributed, the network may be simplified by eliminating some low-level roads without LPR devices, thus improving the coverage rate. This approach has been used in other studies (e.g., Osorio, 2019). (2) If the LPR devices are concentrated in some specific regions (instead of uniformly distributed), the model can be applied in the areas with high coverage rates. For areas with low coverage rates, the first approach can be applied to obtain a rough estimation.

Second, traffic volume may also influence the OD estimation results as revealed in previous studies (Frederix, Viti, & Tampère, 2013; Shafiei, Saberi, Zockaie, & Sarvi, 2017).

Such influence mainly comes from the assumed traffic assignment criteria or from the relationship between OD flow and link flow. Given that no preassumed traffic assignment criteria are used in this work, our model may be able to avoid the impact of traffic congestion. The estimated MAPE presented in Figure 12 can verify this inference to some extent. The estimated MAPE is relatively harmonious at daytime despite the dynamic changes in traffic volumes during this period. To systematically prove this property of the proposed model, future studies can adjust the scale of the input OD matrix in their traffic simulation software and then test different scenarios.

Third, privacy poses an obstacle in the application of the proposed model in traffic planning. Not all cities have accessible LPR data. Given that the proposed model does not require the "true" license plate number of vehicles, any identical index for vehicles (e.g., hash from the original license plate number) can be used. So, the privacy issues may be relieved if the hashed license plate number can be provided.

In the field of OD estimation, the classic volume-based methods depend primarily on loop-detector data. LPR data contain not only link volume information but also partial path information. This gives LPR data the potential to become the mainstream data for OD estimation in the future. Although the coverage rate of loop detectors in the United States may be higher than that of LPR, in China and other Asian countries (e.g., Singapore), LPR is becoming more and more popular and has a higher penetration rate due to its use in law enforcement (Mo et al., 2017). There is also much evidence of increasing LPR deployment in the United States (Lum et al., 2019). Therefore, we expect that LPR data can replace previous data sources in the future.

Furthermore, LPR is an automatic vehicle identification (AVI) technique. The proposed method is a general framework for all AVI data sources (e.g. radio frequency identification [RFID] and near field communication [NFC]). AVI will be the major monitoring system for developing smart cities and connected autonomous vehicles (CAV) and is thus expected to become popular in the future. Our method may play a more significant role at that time.

ACKNOWLEDGMENTS

This research was supported by the National Natural Science Foundation of China (71361130015) and the Beijing Natural Science Foundation (8162024). The authors thank the Bureau of Traffic Management of Langfang for providing the data. The authors also thank Prof. Xuesong Zhou and the 11 anonymous reviewers for their useful comments and suggestions to help improve this article. The first author also thanks Mr. Zhiyong Liu for the relevant discussions.

REFERENCES

Adeli, H., & Ghosh-Dastidar, S. (2004). Mesoscopic-wavelet freeway work zone flow and congestion feature extraction model. *Journal of Transportation Engineering*, *130*(1), 94–103.

Adeli, H., & Karim, A. (2005). *Wavelets in intelligent transportation systems*. Chichester, UK: John Wiley & Sons, Inc.

Antoniou, C., Barceló, J., Breen, M., Bullejos, M., Casas, J., Cipriani, E., … Marzano, V. (2016). Towards a generic benchmarking platform for origin–destination flows estimation/updating algorithms: Design, demonstration and validation. *Transportation Research Part C: Emerging Technologies*, *66*, 79–98.

Antoniou, C., Ben-Akiva, M., & Koutsopoulos, H. N. (2004). Incorporating automated vehicle identification data into origin-destination estimation. *Transportation Research Record*, *1882*(1), 37–44.

Ashok, K. (1996). *Estimation and prediction of time-dependent origin-destination flows*. Cambridge, MA: Massachusetts Institute of Technology.

Ashok, K., & Ben-Akiva, M. E. (1993). Dynamic origin-destination matrix estimation and prediction for real-time traffic management systems. *International Symposium on the Theory of Traffic Flow and Transportation* (12th: 1993: Berkeley, CA). Transportation and traffic theory.

Bierlaire, M., & Crittin, F. (2004). An efficient algorithm for real-time estimation and prediction of dynamic OD tables. *Operations Research*, *52*(1), 116–127.

Cascetta, E., Inaudi, D., & Marquis, G. (1993). Dynamic estimators of origin-destination matrices using traffic counts. *Transportation Science*, *27*(4), 363–373.

Castillo, E., Menéndez, J. M., & Jiménez, P. (2008). Trip matrix and path flow reconstruction and estimation based on plate scanning and link observations. *Transportation Research Part B*, *42*(5), 455–481.

Cipriani, E., Florian, M., Mahut, M., & Nigro, M. (2011). A gradient approximation approach for adjusting temporal origin–destination matrices. *Transportation Research Part C: Emerging Technologies*, *19*(2), 270–282.

Cremer, M., & Keller, H. (1981). Dynamic identification of flows from traffic counts at complex intersections. *Proceedings of 8th International Symposium on Transportation and Traffic Theory*. Toronto, Canada: University of Toronto Press, pp. 199–209.

Dial, R. B. (1971). A probabilistic multipath traffic assignment model which obviates path enumeration. *Transportation Research*, *5*(2), 83–111.

Dixon, M. P., & Rilett, L. R. (2002). Real-time OD estimation using automatic vehicle identification and traffic count data. *Computer-Aided Civil and Infrastructure Engineering*, *17*(1), 7–21.

Dixon, M. P., & Rilett, L. R. (2005). Population origin–destination estimation using automatic vehicle identification and volume data. *Journal of Transportation Engineering*, *131*(2), 75–82.

Duthie, J. C., Unnikrishnan, A., & Waller, S. T. (2011). Influence of demand uncertainty and correlations on traffic predictions and decisions. *Computer-Aided Civil and Infrastructure Engineering*, *26*(1), 16–29.

Feng, Y., Sun, J., & Chen, P. (2015). Vehicle trajectory reconstruction using automatic vehicle identification and traffic count data. *Journal of Advanced Transportation*, *49*(2), 174–194.

Frederix, R., Viti, F., & Tampère, C. M. J. (2013). Dynamic origin–destination estimation in congested networks: Theoretical findings and implications in practice. *Transportmetrica A: Transport Science*, *9*(6), 494–513.

Hooshdar, S., & Adeli, H. (2004). Toward intelligent variable message signs in freeway work zones: Neural network model. *Journal of Transportation Engineering*, *130*(1), 83–93.

Jiang, X., & Adeli, H. (2004a). Clustering-neural network models for freeway work zone capacity estimation. *International Journal of Neural Systems*, *14*(3), 147–163.

Jiang, X., & Adeli, H. (2004b). Object-oriented model for freeway work zone capacity and queue delay estimation. *Computer-Aided Civil and Infrastructure Engineering*, *19*(2), 144–156.

Karim, A., & Adeli, H. (2003). Fast automatic incident detection on urban and rural freeways using wavelet energy algorithm. *Journal of Transportation Engineering*, *129*(1), 57–68.

Larsson, T., Lundgren, J. T., & Peterson, A. (2010). Allocation of link flow detectors for origin-destination matrix estimation—A comparative study. *Computer-Aided Civil and Infrastructure Engineering*, *25*(2), 116–131.

Li, R., Liu, Z., & Zhang, R. (2018). Studying the benefits of carpooling in an urban area using automatic vehicle identification data. *Transportation Research Part C: Emerging Technologies*, *93*, 367–380.

Lum, C., Koper, C. S., Willis, J., Happeny, S., Vovak, H., & Nichols, J. (2019). The rapid diffusion of license plate readers in US law enforcement agencies. *Policing: An International Journal*, *42*(3), 376–393.

Lundgren, J. T., & Peterson, A. (2008). A heuristic for the bilevel origin–destination-matrix estimation problem. *Transportation Research Part B: Methodological*, *42*(4), 339–354.

Mishalani, R. G., Coifman, B., & Gopalakrishna, D. (2002) Evaluating real-time origin-destination flow estimation using remote sensing-based surveillance data. In K. C. P. Wang, S. Madanat, S. Nambisan, & G. Spring (Eds.), *Applications of Advanced Technologies in Transportation: Proceedings of the Seventh International Conference* (pp. 640–647). https://doi.org/10.1061/9780784406328

Mo, B., Li, R., & Zhan, X. (2017). Speed profile estimation using license plate recognition data. *Transportation Research Part C Emerging Technologies*, *82*, 358–378.

Nakanishi, Y., & Western, J. (2005). Ensuring the security of transportation facilities: Evaluation of advanced vehicle identification technologies. *Transportation Research Record: Journal of the Transportation Research Board*, *1938*, 9–16.

Nigro, M., Cipriani, E., & del Giudice, A. (2018). Exploiting floating car data for time-dependent origin–destination matrices estimation. *Journal of Intelligent Transportation Systems*, *22*(2), 159–174.

Nihan, N. L., & Davis, G. A. (1987). Recursive estimation of origin-destination matrices from input/output counts. *Transportation Research Part B: Methodological*, *21*(2), 149–163.

Osorio, C. (2019). Dynamic origin-destination matrix calibration for large-scale network simulators. *Transportation Research Part C: Emerging Technologies*, *98*, 186–206.

Parry, K., & Hazelton, M. L. (2012). Estimation of origin–destination matrices from link counts and sporadic routing data. *Transportation Research Part B: Methodological*, *46*(1), 175–188.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rao, W., Wu, Y.-J., Xia, J., Ou, J., & Kluger, R. (2018). Origin-destination pattern estimation based on trajectory reconstruction using automatic license plate recognition data. *Transportation Research Part C: Emerging Technologies*, *95*, 29–46.

Shafiei, S., Saberi, M., Zockaie, A., & Sarvi, M. (2017). Sensitivity-based linear approximation method to estimate time-dependent origin–destination demand in congested networks. *Transportation Research Record: Journal of the Transportation Research Board*, *2669*, 72–79.

Sherali, H. D., & Park, T. (2001). Estimation of dynamic origin–destination trip tables for a general network. *Transportation Research Part B: Methodological*, *35*(3), 217–235.

Stathopoulos, A., & Tsekeris, T. (2004). Hybrid meta-heuristic algorithm for the simultaneous optimization of the O–D trip matrix estimation. *Computer-Aided Civil and Infrastructure Engineering*, *19*(6), 421–435.

Tavana, H. (2001). *Internally-consistent estimation of dynamic network origin-destination flows from intelligent transportation systems data using bi-level optimization* (Doctoral dissertation). The University of Texas at Austin.

Tavana, H., & Mahmassani, H. S. B. T. (2001). *Estimation of dynamic origin-destination flows from sensor data using bi-level optimization method*. Paper No.: 01–3241, Transportation Research Board Meeting.

Ukkusuri, S. V., Mathew, T. V., & Waller, S. T. (2007). Robust transportation network design under demand uncertainty. *Computer-Aided Civil and Infrastructure Engineering*, *22*(1), 6–18.

van der Zijpp, N. (1997). Dynamic origin-destination matrix estimation from traffic counts and automated vehicle identification data. *Transportation Research Record: Journal of the Transportation Research Board*, *1607*(1), 87–94. https://doi.org/10.3141/1607-13

van der Zijpp, N. J., & Hamerslag, R. (1994). Improved Kalman filtering approach for estimating origin-destination matrices for freeway corridors. *Transportation Research Record*, *1443*, 54–64.

Walpen, J., Mancinelli, E. M., & Lotito, P. A. (2015). A heuristic for the OD matrix adjustment problem in a congested transport network. *European Journal of Operational Research*, *242*(3), 807–819.

Wen, T., Cai, C., Gardner, L., Dixit, V., Waller, S. T., & Chen, F. (2018). A strategic user equilibrium for independently distributed origin-destination demands. *Computer-Aided Civil and Infrastructure Engineering*, *33*(4), 316–332.

Wen, T., Gardner, L., Dixit, V., Waller, S. T., Cai, C., & Chen, F. (2018). Two methods to calibrate the total travel demand and variability for a regional traffic network. *Computer-Aided Civil and Infrastructure Engineering*, *33*(4), 282–299.

Yang, H., Iida, Y., & Sasaki, T. (1991). An analysis of the reliability of an origin-destination trip matrix estimated from traffic counts. *Transportation Research Part B: Methodological*, *25*(5), 351–363.

Yang, J., & Sun, J. (2015). Vehicle path reconstruction using automatic vehicle identification data: An integrated particle filter and path flow estimator. *Transportation Research Part C: Emerging Technologies*, *58*, 107–126.

Yang, X., Lu, Y., & Hao, W. (2017). Origin-destination estimation using probe vehicle trajectory and link counts. *Journal of Advanced Transportation*, *2017*, 1–18.

Yildirimoglu, M., & Kahraman, O. (2017). How far is traffic from user equilibrium? *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)* Yokohama, 2017, pp. 1–6.

Yu, H., Yang, S., Wu, Z., & Ma, X. (2018). Vehicle trajectory reconstruction from automatic license plate reader data. *International Journal of Distributed Sensor Networks*, *14*(2), 1550147718755637.

Zhou, X., & Mahmassani, H. S. (2006). Dynamic origin-destination demand estimation using automatic vehicle identification data.

IEEE Transactions on Intelligent Transportation Systems, *7*(1), 105–114.

Zhou, X., & Mahmassani, H. S. (2007). A structural state space model for real-time traffic origin–destination demand estimation and prediction in a day-to-day learning framework. *Transportation Research Part B Methodological*, *41*(8), 823–840.

Zhou, X., Qin, X., & Mahmassani, H. (2003). Dynamic origin-destination demand estimation with multiday link traffic counts for planning applications. *Transportation Research Record: Journal of the Transportation Research Board*, 1831, 30–38.

How to cite this article: Mo B, Li R, Dai J. Estimating dynamic origin–destination demand: A hybrid framework using license plate recognition data. *Comput Aided Civ Inf.* 2020;35:734–752. https://doi.org/10.1111/mice.12526

APPENDIX A

Link travel time estimation method for unequipped links

Based on the link impedance function proposed by the Bureau of Public Roads (BPR), we model the link travel time as

$$y_a = FT_a \times \left(1 + \alpha_a \left(\frac{x_a}{C_a}\right)^{\beta}\right) \tag{A1}$$

where y_a is the travel time of link a, x_a is the corresponding link flow, FT_a is the free-flow travel time of link a, C_a is the effective capacity of link a, α_a is the link-dependent parameter, and β is the global parameter. We assume that the free-flow travel time is proportional to the length of the link and treat y_a as a random variable. Then, Equation (A1) can be rewritten as

$$y_a = \gamma_a l_a \times \left(1 + \varphi_a {x_a}^{\beta}\right) + \varepsilon_a \tag{A2}$$

where γ_a is the parameter that describes the relationship between link length l_a and free-flow travel time FT_a, φ_a is the parameter that contains α_a and C_a, and ε_a is the random error that is assumed to be normally distributed with an expectation of 0 and a standard deviation of σ_a (i.e., $\varepsilon \sim N(0, {\sigma_a}^2)$). Based on the numerical test results obtained using real-world data, link travel time and link flow show a basic linear relationship (i.e., $\beta = 1$). This may be because the research area is a developing city with a moderate traffic volume (Li et al., 2018). Thus, we set $\beta = 1$, and this parameter can be adjusted to tune real-world traffic situations.

Algorithm 2. Unequipped link travel time estimation model

```
For each link a that belongs to unequipped link set do
    find the nearest $n_{eq}$ equipped links with the same road hierarchy
    For each link b belonging to the extracted $n_{eq}$ equipped links do
        Estimate $\varphi_b$ and $\gamma_b$ using the least square estimation method. Estimate $\sigma_b$ using the standard deviation
        of travel time samples
    end for
    $\varphi_a = (\sum_b \varphi_b)/n_{eq}$
    $\gamma_a = (\sum_b \gamma_b)/n_{eq}$
    $\sigma_a = (\sum_b \sigma_b)/n_{eq}$
   Input $x_a = (\sum_b x_b)/n_{eq}$ to generate enough travel time samples for unequipped links using Eq. (A2)
```

Furthermore, the different forms of link impedance function can also be applied given the traffic conditions. The relationship between link travel time and link flow can be plotted based on LPR data by trying to fit the relationship with different β values and then choosing the best one.

For equipped links (with LPR data records), φ_a and γ_a can be estimated by using travel time and link flow samples via a simple least square estimation (LSE) method. σ_a can also be estimated by using the standard deviation of the travel time samples. The unequipped link travel time estimation model is shown in Algorithm 2. According to the numerical test results, φ_a, γ_a, and σ_a of link a and link flow x_a are similar to those of the surrounding links within the same road hierarchy. Therefore, we can use the average value of the parameters of the equipped links to estimate the parameters of the unequipped links. $n_{eq} = 6$ is used empirically in this study. The value of this parameter can be adjusted depending on the road network conditions. Equipped links with insufficient travel time samples (e.g., in the midnight time interval) can also be supplied with synthesized samples by using Algorithm 2.

APPENDIX B

Formulation of benchmark 2

The bilevel benchmark model can be formulated as

- upper level model ($p^{rs}_{j,\tau}$ is constant):

$$\min_{q^{rs}_{\tau}} J_3 = \sum_{\tau} \left(w'_1 \sum_{a} (v^*_{\tau,a} - v_{\tau,a})^2 + w'_2 \sum_{r,s} (\hat{q}^{rs}_{\tau} - q^{rs}_{\tau})^2 \right) \tag{B1}$$

$$\text{s.t.} \begin{cases} f^{rs}_{j,\tau} = p^{rs}_{j,\tau} \cdot q^{rs}_{\tau}, \mid \quad \forall \tau, r, s, j \\ v_{\tau,a} = \sum\limits_{r,s,j} f^{rs}_{j,\tau} \cdot \delta^{rs}_{a,j}, \mid \quad \forall \tau, a \\ f^{rs}_{j,\tau},\ v_{\tau,a},\ q^{rs}_{\tau} \geq 0, \mid \quad \forall \tau, r, s, j, a \end{cases} \tag{B2}$$

- lower level model (q^{rs}_{τ} is constant):

$$\min_{p^{rs}_{j,\tau}} J_4 = \sum_{\tau,a} \int_0^{v_{\tau,a}} c_{\tau,a}(u)\ \mathrm{d}u \tag{B3}$$

$$\text{s.t.}\begin{cases} f_{j,\tau}^{rs} = p_{j,\tau}^{rs} \cdot q_{\tau}^{rs}, \mid \quad \forall \tau, r, s, j \\ v_{\tau,a} = \sum_{r,s,j} f_{j,\tau}^{rs} \cdot \delta_{a,j}^{rs}, \mid \quad \forall \tau, a \\ \sum_{j} p_{j,\tau}^{rs} = 1, \mid \quad \forall \tau, r, s \\ f_{j,\tau}^{rs}, v_{\tau,a}, p_{j,\tau}^{rs} \geq 0, \mid \quad \forall \tau, r, s, j, a \end{cases}, \tag{B4}$$

where $c_{\tau,a}$ is the link impedance function for link a within time interval τ. We use the BPR function form for $c_{\tau,a}$ in this study. All other variables have the same notations as the formulation presented in Section 3.2. $\hat{q}_{\tau}^{rs}$ is the set equal to the initial OD matrix obtained in Section 3.1. It is worth noting that the parameters for $c_{\tau,a}$ must be thoroughly calibrated before running this model. This model is solved by using the iterative algorithm proposed by Zhou and Mahmassani (2006).