Transportation Research Part E 173 (2023) 103108

Contents lists available at ScienceDirect

Transportation Research Part E

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

Simulation-based dynamic origin–destination matrix estimation on freeways: A Bayesian optimization approach

Jinbiao Huo [a], Chengqi Liu [a], Jingxu Chen [a], Qiang Meng [b], Jian Wang [a], Zhiyuan Liu [a,*]

[a] Jiangsu Key Laboratory of Urban ITS, Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, School of Transportation, Southeast University, Nanjing 211189, China
[b] Department of Civil and Environmental Engineering, National University of Singapore, Singapore 117576, Singapore

ARTICLE INFO

Keywords:
Dynamic OD estimation
Bayesian optimization
High-dimensional problem
Surrogate-based optimization
Freeway network

ABSTRACT

This study focuses on dynamic origin–destination demand estimation problem on freeway networks. Existing studies on this problem rely on high-coverage of traffic measurements and assumptions on travel times, exhibiting limitations in real-world applications. We formulate the problem as a bi-level programming model, where micro-simulations are incorporated to precisely model traffic flows/travel times on freeways. The bi-level programming model cannot provide explicit closed-form expressions for the objective function and its derivatives, and also intrinsically high-dimensional. Thus, it is highly challenging to find efficient solution algorithms. In this regard, a problem-specific and computationally efficient Bayesian optimization approach is designed. Herein, a novel surrogate model is proposed by embedding a *physical surrogate model* (it characterizes underlying physical mechanisms and provides global yet less precise approximations) into a *functional surrogate model* (it provides precise local approximations). The embedding provides problem-specific knowledge for the surrogate model. More importantly, it also restricts the feasible region, enabling the surrogate model to efficiently deal with high-dimensional problems. Gaussian process can be served as the functional surrogate model. Two linear physical surrogate models are proposed to capture interactions between travel demand and traffic measurements. To deal with constraints in the surrogate model, a projection-distance based acquisition function is designed. In searching for new points, the proposed acquisition function is capable of assigning unique weight of exploration to each feasible solution. The proposed approach is validated based on a freeway corridor example, which indicates its outperformance over existing dynamic origin–destination estimation methods in terms of computational efficiency and solution accuracy.

1. Introduction

The dynamic origin–destination estimation (DODE) problem refers to the estimation of time-varying travel demand between pairs of zones in a traffic system, which results in origin–destination (OD) matrices. This study focuses on DODE problems on freeway

* Corresponding author.
E-mail addresses: jinbiaoh@seu.edu.cn (J. Huo), dclily@foxmail.com (C. Liu), chenjingxu@seu.edu.cn (J. Chen), ceemq@nus.edu.sg (Q. Meng), jianw@seu.edu.cn (J. Wang), zhiyuanl@seu.edu.cn (Z. Liu).

https://doi.org/10.1016/j.tre.2023.103108
Received 26 December 2022; Received in revised form 18 February 2023; Accepted 22 March 2023
Available online 5 April 2023

networks, which received limited attention over the past few decades despite the long academic pedigree along with OD estimation (Fisk and Boyce, 1983; Cascetta et al., 1993; Ashok and Ben-Akiva, 2000; Alexander et al., 2015; Cao et al., 2021). Existing studies employed non-assignment-based methods to solve DODE problems on freeways (Chang and Wu, 1994; Lin and Chang, 2007), which estimate OD matrices through a set of system equations that characterize the relationships between OD patterns and traffic measurements. However, the methods require high coverage of traffic measurements and assume specific distributions of travel times, which are always violated in practical applications (Lin and Chang, 2005). By contrast, assignment-based methods, which solve DODE problems through dynamic network loading (DNL) models, are less affected by the coverage of traffic measurements or distributions of travel times, and therefore, are expected to provide an alternative and effective way to solve DODE problems in freeway networks.

The assignment-based methods generally formulate DODE problems as bi-level programming models (Yang et al., 1992; Shafiei et al., 2016), in which the upper level is to determine the optimal OD matrix to minimize the difference between the field and estimated traffic measurements. The lower-level employs DNL models to describe interrelations between traffic measurements and OD matrices. With the rapid development of traffic simulators, incorporating simulation-based DNL models in the lower-level is being increasingly popular, owing to the capability of simulators to model complex interactions in the dynamic and stochastic transportation systems (Rao et al., 2018; Tay and Osorio, 2022; Antoniou et al., 2015). We term these problems as simulation-based DODE problems.

Simulation-based DODE problems are typical simulation-based optimization problems, in which the objective is evaluated through traffic simulation (Amaran et al., 2016). Surrogate-based optimization algorithms (SOAs) have emerged as a mainstream to tackle these problems (Cheng et al., 2019; Wu et al., 2019; Gu and Saberi, 2021; Zheng et al., 2022; Yin et al., 2022b). The basic idea of SOA is to fit surrogate models (i.e., analytical approximate of the objective function) using data collected by evaluating the objective function at a few points (termed sample points). The surrogate model is then used to estimate the location of the optimum, and suggest new samples for model refinement. For detailed discussions on surrogate models, the readers can refer to (Forrester and Keane, 2009; Bhosekar and Ierapetritou, 2018).

A key challenge that arises in using SOAs to solve DODE problems is the high-dimensionality of the problem. DODE problems are widely recognized as high-dimensional ones (Osorio, 2019a; Dantsuji et al., 2022), which often contain hundreds or thousands of decision variables. To achieve good coverage of sampling and identify satisfactory solutions, the required number of sample points increase significantly with the problem dimension (Bull, 2011; Choffin and Ueda, 2018). Consequently, the computational cost of dealing with the sample points, including formulation of surrogate models, search of samples, and execution of simulation, is extremely high and unacceptable. To tackle this issue, extensive studies have devoted to designing SOAs for high-dimensional problems. Of all attempts, embedding problem-specific knowledge into SOAs is the most commonly used approach in the transportation discipline (Patwary et al., 2021; Tay and Osorio, 2022). Specifically, several problem-specific surrogate models have been proposed in prior literature, and successfully used in transportation problems such as congestion pricing and traffic signal control problems. However, there has been insufficient work on developing customized SOAs for DODE problems on freeways.

With these motivating considerations, this study focuses on DODE problems on freeways and aims to first, develop a simulation-based DODE model in which traffic simulators are incorporated to precisely model traffic flows, and second, design an efficient and customized SOA to solve the high-dimensional simulation-based optimization problem. To this end, a novel SOA is proposed by embedding problem-specific knowledge into the state-of-the-art Bayesian optimization (BO) algorithm. The proposed approach also offers an alternative approach of embedding problem-specific knowledge in SOAs.

1.1. Literature review

OD estimation problems stem from the 1950s when urbanization and car-ownership soared after World War II (CATS, 1959). House-hold and roadside surveys are originally used for collecting travel information. Whereas, these methods require high cost in both economic and social viewpoints (Doblas and Benitez, 2005). Alternative methods include trip distribution model-based methods and matrix estimation methods. The former suffers from impractical assumptions on the trip-distribution and fail to provide accurate estimation. The latter takes OD matrix as unknown variables and formulate mathematical optimization models to find the optimal solution. In the past 40 years, numerous works were conducted on the matrix estimation methods (Yang et al., 1992; Cascetta et al., 1993; Lundgren and Peterson, 2008; Antoniou et al., 2016; Osorio, 2019a).

Preliminary studies of matrix estimation methods focused on static OD estimation problems, which assumed that traffic conditions were stationary for a specific time period and the demand kept unchanged during this period (Yang et al., 1992; Fisk, 1988). DODE models relax the assumptions on stationary traffic conditions and demand, and incorporate traffic dynamics in the problem (Cascetta et al., 1993; Toledo and Kolechkina, 2012; Zhou and Mahmassani, 2006; Ma et al., 2020; Fu et al., 2022). In the past decades, extensive studies have been conducted in tackling DODE problems on a variety of topics, including the data sources (Ashok and Ben-Akiva, 2000; Laharotte et al., 2014; Ma and Qian, 2018; Carrese et al., 2017; Demissie and Kattan, 2022), the stochasticity of the problem (Ashok and Ben-Akiva, 2002; Tympakianaki et al., 2018), the solution algorithms (Caggiani et al., 2013; Osorio, 2019a), the online estimation (Frederix et al., 2013; Marzano et al., 2018), the transit/freight OD matrices (Zhao et al., 2020; Li and Chen, 2022; Kalahasthi et al., 2022; Peled et al., 2021) and so forth. For a brief review, we refer to Antoniou et al. (2016).

Despite the advancements over the past few decades, the DODE problems in freeway networks have received limited attention. Existing studies mainly used non-assignment-based methods to estimate OD demands (Lin and Chang, 2007), which estimate OD matrices by solving a set of system equations that characterize the relationships between OD patterns and traffic measurements. The study of DODE problem on freeway networks could date back to Bell (1991), in which the author investigated DODE problems in uncongested networks. A set of linear equations were proposed while considering the discrepancy of travel times. Comparing with other methods at that time, the model has more realistic formulation. However, it involves too many parameters to estimate when

travel times of OD pairs span more than two time intervals (Lin and Chang, 2005), making the model less effective/reliable. Chang and Wu (1994) studied the DODE problem on freeways when congestion arises. They extended the linear model proposed by Bell (1991) by incorporating traffic flows in all segments into system equations. To improve the operational efficiency, the authors assumed that the speeds of vehicles entering the freeway at the same time interval are distributed in a small range. Lin and Chang (2005) released this assumption by employing normal distributions to describe the variation of travel times, which enables the speeds of vehicles to vary in a wide range. The normal distributions are estimated using data collected from surveillance systems. Lin and Chang (2007) further generalized the system equations in (Lin and Chang, 2005) by introducing distributions of travel times for each OD pair in each segment. After (Lin and Chang, 2007), the DODE problems on freeway networks are rarely studied. In (Zhang et al., 2008) and (Marzano et al., 2018), the authors took freeway networks as scenarios to test algorithms, but did not develop customized algorithms towards DODE problems in freeways.

However, as Lin and Chang (2005) have pointed out, existing DODE methods on freeways (i.e., the non-assignment-based methods) exhibit limitations in real-world applications. First, the non-assignment-based methods require high coverage of surveillance systems. The time-varying traffic flows in all on/off-ramps, even all mainline segments, are required to establish system equations. Nonetheless, the traffic measurements are usually incomplete and have inadequate coverage. Second, existing non-assignment-based methods assume that the travel times of vehicles either span a small range or follows specific distributions. These assumptions neglect the prevalent heterogeneity of traffic flows on freeways (Qian et al., 2017; Lu et al., 2016), and may fail to characterize travel times in reality. Third, the number of parameters in system equations and difficulties in collecting complete traffic information increase significantly with the scale of networks, which hinder the deployment of non-assignment-based methods in large-scale freeway networks. Taken together, due to the incomplete collection of traffic information, intricate distribution of travel times, and large scale of networks in reality, non-assignment-based methods could not effectively deal with DODE problems on freeways in practice. By contrast, the assignment-based methods do not require high coverage of traffic measurements to establish system equations, and model travel times by applying DNL models to assign OD demand on networks, which well captures the dynamics and heterogeneity of traffic flows. Hence, the assignment-based methods are expected to provide an alternative, and more effective way to tackle DODE problems on freeway networks.

Existing DNL methodologies can be classified into two streams: analytical approach (Ge and Fukuda, 2019) and simulation-based approach (Lu et al., 2013; Ameli et al., 2020). Over decades of study, traffic simulation models have gained prominence. Microscopic traffic simulation assigns travel demand by moving individual vehicles on networks (Antoniou et al., 2015), which is able to capture the dynamic and stochastic features of traffic flows that are hard to model analytically, leading to assignment results closely tied to reality (Tympakianaki et al., 2015; Shafiei et al., 2018; Zhang et al., 2021; Dantsuji et al., 2022). As has reported in (Lin and Chang, 2007), the impact of travel time variability on the time-varying OD patterns is one of the most critical issues to be tackled for DODE problems on freeways. Therefore, to precisely model traffic flows/travel times on freeways and achieve accurate OD estimates, the microscopic traffic simulation model is incorporated into the DODE problem in this study.

A considerable number of algorithms have been developed to solve the simulation-based DODE problems. Conventionally, two types of algorithms are usually used in existing literature. One is the simultaneous perturbation stochastic approximation (SPSA) method (Cipriani et al., 2011). SPSA approximates the gradient with two successive evaluations of the objective function (Antoniou et al., 2016). Yet, the convergence rate and long run accuracy of the algorithm deteriorated significantly as the problem scale increased (Antoniou et al., 2015). To solve the problem, some recent works extend SPSA by incorporating problem-specific information into the method (Lu et al., 2015; Nigro et al., 2018; Qurashi et al., 2019). The other type of algorithm (Yang et al., 1992; Zhou et al., 2003) iterates between two basic steps: (i) running DNL based on a given OD matrix and yielding an assignment matrix; (ii) fixing the assignment matrix, optimizing the upper-level problem and updating the OD matrix. A major deficiency of the algorithm is on the efficiency (Toledo and Kolechkina, 2012). Lundgren and Peterson (2008) improved the algorithm by introducing an auxiliary solution between step (i) and (ii). Followed by (Toledo and Kolechkina, 2012; Lu et al., 2013; Shafiei et al., 2017), several improvements have been proposed on the algorithm.

With encouraging results in the transportation discipline, surrogate-based optimization has been widely studied, dealing with expensive-to-evaluate simulation-based optimization problems such as congestion pricing problems (Chen et al., 2016; Gu et al., 2019; Gu and Saberi, 2021; Zheng et al., 2022), traffic signal control problems (Osorio and Bierlaire, 2013; Zheng et al., 2019) and network design problems (Chen et al., 2006; Yin et al., 2022a; Li et al., 2022). Of various surrogate-based optimization algorithms, Bayesian optimization (BO) algorithm distinguishes itself from other methods by employing surrogate models using Bayesian statistics (Frazier, 2018). The efficiency of BO stems from its capability of incorporating prior beliefs of the problem to inform the sampling of points, as well as its good balance between exploration and exploitation in the sampling. In the transportation discipline, BO has been predominantly applied to address parameter tuning (Liessner et al., 2019; Shang et al., 2019; Tang et al., 2020; Yi and Bui, 2020; Duan et al., 2022) and traffic management problems (Otsuka et al., 2019; Hickish et al., 2020; Dandl et al., 2021; Fakhrmoosavi et al., 2022; Huo et al., 2023).

Owing to the efficiency and analytical tractability of surrogate-based optimization algorithms, growing awareness has been given to applying SOAs to solve DODE problems (Osorio, 2019b; Dantsuji et al., 2022). The main challenge, as we have discussed, is to design SOAs that are feasible to deal with the *curse of dimensionality* (Osorio, 2019b). In the transportation discipline, most existing SOAs are designed to solve high-dimensional problems by embedding problem-specific knowledge into SOAs, under a framework proposed in Osorio and Bierlaire (2013). Osorio and Bierlaire proposed a novel surrogate model which is the weighted sum of two components, a physical and a functional surrogate model. For brevity, we term the novel surrogate model as the *weighted-sum surrogate model*. The physical surrogate model is a problem-specific network model. It provides global yet less precise approximations. While the functional surrogate model is a polynomial function, which is constructed purely based on samples. It provides precise local approximations.

Since problem-specific knowledge are incorporated, the *weighted-sum surrogate model* requires much less simulation runs to solve transportation problems than purely employing functional surrogate models. Under the framework, several *weighted-sum surrogate models* have been developed to solve traffic signal control problems (Osorio and Chong, 2015), congestion pricing problems (Osorio and Atasoy, 2021), parameter calibration problems (Patwary et al., 2021), and so forth. Particularly, Osorio (2019b) proposed a *weighted-sum surrogate model* to solve offline DODE problems. The efficiency of the proposed surrogate model is validated on the network of major arterials and expressways of Singapore.

Although the *weighted-sum surrogate model* has shown its effectiveness in urban networks, in the second case study of this research, using SOAs to solve DODE problems on freeways, it was found that the performance of the *weighted-sum surrogate model* in terms of convergence rate and accuracy is not satisfied. The possible reasons would be three-fold. First, existing physical surrogate models are developed towards urban networks, with necessary and customized simplifications. These simplifications give rise to inaccurate estimates of traffic mechanism on freeways. Second, the highly nonlinear and intricate relationship between OD matrices and traffic measurements cannot be fully captured by existing functional surrogate models that are formulated as linear or quadratic polynomials. Thirdly, and perhaps most significantly, the *weighted-sum surrogate model* did not actually reduce search space of high-dimensional problems. Instead, it employs physical surrogate models to adjust search directions (like soft constraints) when searching for samples. As results, the feasible set is still high-dimensional, and the efficiency and accuracy of the *weighted-sum surrogate model* rely on the accuracy of the physical surrogate models.

1.2. Objectives and contributions

In summary, despite the extensive research on DODE in urban networks, several issues have received limited attention with regards to the problem on freeway networks. First, a simulation-based dynamic OD estimation model needs to be developed, such that traffic flows/travel times can be delicately modelled and more accurate estimates of OD matrices on freeways can be achieved. Second, there has been insufficient work on designing efficient SOAs towards DODE problems on freeway networks. More specifically, exploring efficient functional/physical surrogate models as well as ways of embedding problem-specific knowledge into SOAs are necessary. This study aims to bridge these gaps. Contributions of this work are summarized as follows:

The first contribution lies in the problem formulation.

- A bi-level programming model is proposed to model simulation-based DODE problems on freeways, in which micro-simulators are served in the lower-level to delicately model travel behaviors of each individual vehicle. Comparing with existing studies, the bi-level model shows higher applicability in practical applications as it does not require high-coverage of filed measurements or assume specific distributions of travel times. For all we know, this is the first study that incorporates micro-simulators in the DODE problems on freeways.

To solve the bi-level programming model, a problem-specific and efficient Bayesian optimization approach is proposed. The second and third contributions are on the approach.

- A novel surrogate model is formulated. Gaussian process (GP), an emerging and powerful Bayesian regression model, is served as the functional surrogate model to capture the highly-nonlinear relationships between OD matrices and traffic measurements. To the best of our knowledge, this is the first work that introduces GP to solve DODE problems. Two physical surrogate models are proposed in this study. With simple linear functional forms, the physical surrogate models can capture traffic mechanisms on freeways and provide global approximations. The surrogate model is formulated by embedding the physical model into GP. Different from existing studies, the embedding can be seen as adding hard constraints, which restricts the feasible regions and leads to less required sample points. The proposed surrogate model offers an alternative approach of embedding problem-specific knowledge into SOAs.
- The third contribution is on the formulation of the acquisition function (the function to suggest where to sample next) in Bayesian optimization. A revised acquisition function is proposed to deal with the constraints in the surrogate model. In sampling new points, all solutions in feasible sets are assigned unique weights of exploration based on their distances to the boundary. The proposed acquisition function enhances the problem-solving efficiency.

The remainder of this paper is structured as follows. Section 2 formulates the dynamic OD estimation problem as a bi-level programming problem. Challenges in solving the problem are discussed. Section 3 presents the surrogate model. In Section 4, an optimization algorithm is proposed. The efficiency of the proposed methods is verified by case studies in Section 5. We conclude the paper with some remarks on the proposed improvements and recommendations for further study in Section 6.

Fig. 1. A typical freeway corridor.

2. Problem statement

Consider a freeway corridor of $N-1$ segments, as shown in Fig. 1. The traffic zones are defined at the cross-sections of the mainline as well as the on and off ramps (see Fig. 1), which are taken as the origins/destinations. The blue rectangles in Fig. 1 denote sensors for traffic measurements. The sensors, with predetermined number and locations, are set on the mainline and ramps, collect traffic measurements, and aggregate the measured data by periods of fixed length. In this paper, the most commonly available traffic measurements, traffic counts, are employed.

OD matrix defines the expected number of trips for each OD pair within given time intervals. For a region that consists of Z OD pairs, the traffic demands during T time intervals can be defined by a traffic matrix with Z columns and T rows. A simulation-based dynamic OD estimation problem aims to find the best OD matrix, such that the simulated traffic measurements can well match the field measurements.

The notations used in this paper are listed as follows.

$\boldsymbol{v}^f$	vector of field traffic counts
$\boldsymbol{v}^s$	vector of simulated traffic counts
$\boldsymbol{d}$	vector of traffic demands for each OD pair and each time interval
$\boldsymbol{p}$	vector of exogenous variables (e.g., network topology)
$\boldsymbol{a}$	vector of variables modelling travelers' behaviors (including travel speed, acceleration, deceleration, and lane selection, etc.)
T	set of indices of time intervals
K	set of links with sensors
v_{kt}^f	field traffic count at link k during time interval t
v_{kt}^s	simulated traffic count at link k during time interval t

The simulation-based dynamic OD estimation problem is mathematically formulated as a bi-level programming model (Cascetta et al., 1993; Toledo and Kolechkina, 2012).

$$\min_{\boldsymbol{d}} f\left(\boldsymbol{v}^f, \boldsymbol{v}^s(\boldsymbol{d}), \boldsymbol{d}\right) = \min_{\boldsymbol{d}} \frac{1}{|\mathrm{T}||\mathrm{K}|} \sum_{t\in \mathrm{T}} \sum_{k\in \mathrm{K}} \left| \frac{v_{kt}^f - v_{kt}^s(\boldsymbol{d})}{v_{kt}^f} \right| \tag{1}$$

subject to:

$$G\left(\boldsymbol{v}^f, \boldsymbol{v}^s(\boldsymbol{d}), \boldsymbol{d}\right) \leqslant 0 \tag{2}$$

where the simulated traffic counts $\boldsymbol{v}^s(\boldsymbol{d})$ is measured using micro-simulators, based on travelers' behaviors $\boldsymbol{a}(\boldsymbol{d})$, which is defined in the lower-level problem:

$$\boldsymbol{a}(\boldsymbol{d}) = \mathrm{argmin} T(\boldsymbol{d}, \boldsymbol{p}, \boldsymbol{a}) \tag{3}$$

subject to:

$$g(\boldsymbol{d}, \boldsymbol{p}, \boldsymbol{a}) \leqslant 0 \tag{4}$$

In the bi-level programming model, $f(\cdot)$ is the objective of the upper-level problem. $\boldsymbol{d}$ defines the OD matrix, which is the decision vector in the upper-level. G defines the constraints on $\boldsymbol{d}$. $|\mathrm{T}|$ ($|\mathrm{K}|$) denotes the cardinality of set T (K). $T(\cdot)$ defines objective of the lower-level decision makers (i.e., all travelers in the network). $\boldsymbol{a}$ is the decision vector of the lower-level problem. g defines the constraints on $\boldsymbol{a}$.

The upper-level of the bi-level programming model aims to minimize the difference between field and simulated traffic measurements. The difference is measured using mean absolute percentage errors (MAPE). The lower-level represents the simulation-based DNL. Task of the bi-level programming model is to find an optimal OD matrix $\boldsymbol{d}^*$, such that the distance function $f(\cdot)$ is minimized while assigning $\boldsymbol{d}^*$ to the network.

The lower-level problem (3)–(4) is modeled using microscopic simulators, in which the stochasticity and dynamics of traffic flows on freeways are precisely described. More specifically, the car-following behaviors, lane-changing behaviors and behaviors at the ramp-merging segments are delicately defined for each vehicle considering its surrounding traffic conditions (Cheng et al., 2021). Hence, for a given pattern of traffic demand, the simulated traffic flows on each link are the results of interactions of all the simulated travelers in the dynamic system.

Main challenge of solving the above bi-level programming model is twofold. First, for given $\boldsymbol{d}$, $\boldsymbol{v}^s$ is measured using micro-simulators. As discussed, $\boldsymbol{v}^s$ is resulted from interactions of all travelers in the dynamic system, while considering complicated traffic conditions. Thus, the relationships between $\boldsymbol{v}^s$ and $\boldsymbol{d}$ are highly-nonlinear, making the objective function most often non-convex, non-differentiable and intractable. Moreover, the objective function is expensive to evaluate due to the micro-simulators. The second challenge lies in the intrinsic high-dimensionality of the problem. The dimension of $\boldsymbol{d}$ is often in the hundreds or thousands. Whereas, for simulation-based optimizations, problems with dimension of 200 are considered as high-dimensional ones (Wang et al., 2016). Particularly, researches have proofed that, to achieve given accuracy, the number of required sample points, for some SOA, increases exponentially with the problem dimension (Bull, 2011). Hence, the computational cost of high-dimensional problems frustrates the use

of conventional SOAs. This paper aims to solve these challenges by designing computationally efficient surrogate-based optimization algorithms tailored for dynamic OD estimation problems on freeway networks.

3. A novel surrogate model

A surrogate model is defined as a mathematical model or a physical model which replaces an expensive model for analysis or optimization (Søndergaard, 2003). Surrogate models are feasible to provide a mathematically more tractable, yet less detailed formulation than the expensive model. Depending on whether problem-specific knowledge is involved, surrogate models can be categorized into two classes: the physical surrogate model and the functional surrogate model. Physical surrogate models are constructed based on the knowledge of particular physical systems, and usually provide global, yet less precise approximations. In contrast, functional surrogate models are constructed based on sample points, without consideration of problem-specific knowledge. Functional surrogate models usually provide precise approximations around the sample points, but fail to provide global approximations.

Obviously, functional surrogate models and physical surrogate models are complementary in terms of approximation. Thus, a straightforward thought of constructing efficient surrogate models is to integrate the two types of models (e.g., by linear combination) for providing global and local approximations simultaneously. In this section, a novel surrogate model is proposed towards the DODE problems on freeway networks. The surrogate model combines physical/functional surrogate models in a new way: embedding a physical surrogate model into a functional surrogate model. The combination not only embeds problem-specific knowledge into the surrogate model, but effectively restricts the feasible set, which enhances the efficiency of addressing the high-dimensional problem. We first represent the functional surrogate model, the Gaussian process regression. Then, two physical surrogate models are proposed. Finally, we introduce the combination of the surrogate model.

3.1. The functional surrogate model: Gaussian process

Let $(\boldsymbol{d}_i, f(\boldsymbol{d}_i))$ denote a sample point in the dynamic OD estimation problem. For conciseness, we use f_i to represent $f(\boldsymbol{d}_i)$. For the sample point set $\{(\boldsymbol{d}_1, f_1), \ldots, (\boldsymbol{d}_n, f_n)\}$, GP assumes that the vector $\boldsymbol{F}_n = [f_1, \ldots, f_n]^T$ follows an n-dimensional multivariate Gaussian distribution

$$\boldsymbol{F}_n \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{n}}, \mathbf{K}_{\mathrm{n}}) \tag{5}$$

where $\boldsymbol{\mu}_n$ is an $n \times 1$ matrix with entries $E[f_i]$, and $\boldsymbol{K}_n$ is an $n \times n$ matrix with entries $\mathrm{Cov}(f_i, f_j)$. $E[f_i], \forall i \in \{1, 2, \ldots, n\}$ denotes the expected value of f_i. The expected value does not affect the optimization process, therefore, $\boldsymbol{\mu}_n$ is usually set to be a zero vector. In GP, the covariance between $f(\boldsymbol{d}_i)$ and $f(\boldsymbol{d}_j)$, $\forall i, j \in \{1, \ldots, n\}$ is measured by kernel functions. In this paper, one of the most widely used kernel functions, the radial basis function (RBF) kernel is used.

$$\mathrm{Cov}(f_i, f_j) = \kappa(\boldsymbol{d}_i, \boldsymbol{d}_j) = \exp\left(-\frac{||\boldsymbol{d}_i - \boldsymbol{d}_j||_2^2}{2l^2}\right) \tag{6}$$

In Eq. (6), the parameter is estimated using the maximum likelihood estimation.

$$l = \underset{l>0}{\mathrm{argmax}} \quad -\frac{1}{2}\boldsymbol{F}_n^T \boldsymbol{K}_n^{-1} \boldsymbol{F}_n - \frac{1}{2}\ln|\boldsymbol{K}_n| - \frac{n}{2}\ln(2\pi) \tag{7}$$

Based on the definition of GP, for an unsampled point, $(\boldsymbol{d}', f(\boldsymbol{d}'))$, the objective function follows the following distribution.

$$\begin{bmatrix} \boldsymbol{F}_n \\ f' \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_{\mathrm{n}} \\ \mu_* \end{bmatrix}, \begin{bmatrix} \mathbf{K_n} & \mathbf{K}_* \\ \mathbf{K}_*^{\mathbf{T}} & \mathbf{K}_{**} \end{bmatrix} \right) \tag{8}$$

where $\boldsymbol{K}_* = [\mathrm{Cov}(f_1, f'), \mathrm{Cov}(f_2, f'), \ldots, \mathrm{Cov}(f_n, f')]^T$, $\boldsymbol{K}_{**} = \mathrm{Cov}(f', f')$. Given $\boldsymbol{d}'$, the posterior distribution of $f(\boldsymbol{d}')$ is derived:

$$f(\boldsymbol{d}')|\boldsymbol{d}' \sim \mathcal{N}(\mathbf{K}_*^{\mathrm{T}} \mathbf{K}_{\mathrm{n}}^{-1}(\mathbf{F}_{\mathrm{n}} - \boldsymbol{\mu}_{\mathrm{n}}) + \mu_*, \quad \mathbf{K}_{**} - \mathbf{K}_*^{\mathrm{T}} \mathbf{K}_{\mathrm{n}}^{-1} \mathbf{K}_*) \tag{9}$$

In the normal distribution, $\boldsymbol{K}_*^T \boldsymbol{K}_n^{-1}(\boldsymbol{Y}_n - \boldsymbol{\mu}_n) + \mu_*$ defines the estimated value of $f(\boldsymbol{d}')$ given $\boldsymbol{d}'$. $\boldsymbol{K}_{**} - \boldsymbol{K}_*^T \boldsymbol{K}_n^{-1} \boldsymbol{K}_*$ denotes the variance of estimation, which measures the uncertainty information at $\boldsymbol{d}'$.

It can be seen from Eq. (6) and Eq. (9) that GP provides more accurate estimations at points (i.e., the posterior variance is smaller) which are closer to the sample points. For regions where sampled points are distributed sparsely, the GP cannot approximate the objective function well.

3.2. Physical surrogate models

In this section, two physical surrogate models are formulated. With the physical surrogate models, we aim to provide global and analytical approximations for the relationship between OD matrices and traffic flows.

3.2.1. The fixed-speed model (M1)

The fixed speed model is inspired by observations on freeway traffic flows: vehicles on freeway generally travel at steady speeds. Hence, we assume that (*i*) all vehicles are moved in the network with a fixed speed, v_f; (*ii*) for each traffic demand, the departure time of vehicles is uniformly distributed. This way, the stochasticity of driver behaviors is neglected and thus the traffic flow can be seen as the linear combination of traffic demands. Considering the freeway corridor as shown in Fig. 2, c_i $i \in \{0, ..., N-1\}$ defines the distance on the mainline from zone i to zone j, b_i $i \in \{0, ..., N\}$ denotes the distance from zone $i \in \{0, ..., N\}$ to the mainline, particularly, $b_0 = b_N = 0$. In a typical corridor, all sensors are categorized into 3 classes: the sensors on the off-ramp (D_1), the sensors on the on-ramp (D_2), and the sensors on the mainline (D_3). Let Δ_{k-on} (resp. Δ_{k-off}) denote the distance from the sensor on the on-ramp (resp. off-ramp) of zone k to the mainline. Let $d_{sr}^{(t_1,t_2)}$ denote the travel demand during t_1 and t_2 from zone s to zone r. $v_i^{(t_1,t_2)} i \in \{1, 2, 3\}$ denote the traffic counts collected by sensor D_i from t_1 to t_2.

Without loss of generality, traffic flows in one direction are considered. Depending on the sensor types, the collected traffic flows are categorized into 3 classes:

i) The traffic flows obtained from sensor D_1 are composed of traffic demands with origin of zone $i \in \{0, 1, ..., k-1\}$ and destination of zone k.
ii) The traffic flows obtained from sensor D_2 are composed of traffic demands with origin of zone k and destination of zone $j \in \{k+1, ..., N\}$.
iii) The traffic flows obtained from sensor D_3 are composed of traffic demands with origins of zone $i \in \{0, 1, ..., k\}$ and destinations of zone $j \in \{k+1, ..., N\}$.The three types of traffic flows can be represented as follows:

$$v_1^{(t,t+\Delta t)} = \sum_{i=0}^{k-1} d_{ik}^{\left(t-\Delta_{ik}^1, t+\Delta t-\Delta_{ik}^1\right)} \tag{10}$$

$$v_2^{(t,t+\Delta t)} = \sum_{j=k+1}^{N} d_{kj}^{\left(t-\Delta_{kj}^2, t+\Delta t-\Delta_{kj}^2\right)} \tag{11}$$

$$v_3^{(t,t+\Delta t)} = \sum_{i=0}^{k} \sum_{j=k+1}^{N} d_{ij}^{\left(t-\Delta_{kij}^3, t+\Delta t-\Delta_{kij}^3\right)} \tag{12}$$

where $\Delta_{ki}^1 = \frac{\left(\Delta_{k-on}+b_i+\sum_{m=i}^{k-1} c_m\right)}{v_f}$, $\Delta_{kj}^2 = \frac{\left(\Delta_{k-off}+b_j+\sum_{m=k}^{j} c_m\right)}{v_f}$, $\Delta_{kij}^3 = \frac{\left(\Delta_c+b_i+\sum_{m=i}^{k-1} c_m\right)}{v_f}$. Since traffic demands vary with a fixed time interval, $d_{sr}^{(t_1,t_2)}$ is actually a linear combination of traffic demands. Therefore, Eqs. (10)–(12) are all linear equations. For brevity, the relationship between the traffic counts and the traffic demands is described as:

M1:

$$\boldsymbol{P}\boldsymbol{d} = \boldsymbol{v}^f \tag{13}$$

where $\boldsymbol{v}^f$ is the field counts, $\boldsymbol{d}$ is the vector of traffic demands. The matrix $\boldsymbol{P}$ defines the proportions of traffic demands contributing to the field traffic counts. $\boldsymbol{P}$ is defined by the network topology and the fixed speed v_f, according to Eqs. (10)–(12). Hence, $\boldsymbol{P}$ is a constant matrix. Regarding the limited traffic surveillance facilities in networks, $\boldsymbol{P}$ is not of full rank.

M1 defines the linear relationship between the traffic demands and the field traffic counts. Compared to the complex simulation process, the formula of M1 is much simpler and easier to analyze. According to Eq. (13), the variables $\boldsymbol{d}$ can be divided into two sets: the basic variables $\boldsymbol{d}_B$ and the non-basic variables $\boldsymbol{d}_N$. Let $\boldsymbol{P}_B$ and $\boldsymbol{P}_N$ denote the coefficient matrices corresponding to $\boldsymbol{d}_B$ and $\boldsymbol{d}_N$ respectively. We have the following relationship:

$$\boldsymbol{d}_B = \boldsymbol{P}_B^{-1}\boldsymbol{v}^f - \boldsymbol{P}_B^{-1}\boldsymbol{P}_N\boldsymbol{d}_N \tag{14}$$

Eq. (14) is exactly a linear approximation of the inherently nonlinear relationships among traffic demands. Fig. 3 intuitively illustrates the linear approximation using a two-variable example. The orange surface is the underlying real function of traffic flow with respect to traffic demands, denoted as $v = f(d_1, d_2)$. The green plane is a linear approximation of the underlying function (i.e., the relationship defined in M1). Given an observation on the traffic flow, v_{det}, the intersection between $v = v_{\text{det}}$ and $v = f(d_1, d_2)$ (resp., $v + k_1 d_1 + k_2 d_2 = 0$) defines the real (resp. approximate) relationship between d_1 and d_2. As shown in Fig. 3, Eq. (14) analytically approximates the nonlinear relationship (i.e., the orange curve) by a linear equation (i.e., the green line).

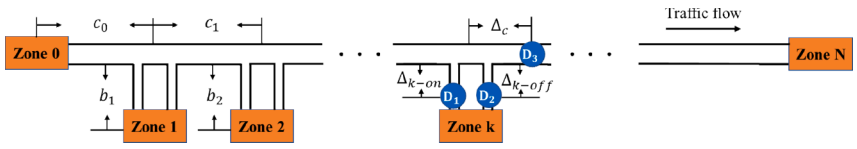


Fig. 2. Three classes of sensor on a typical freeway corridor.

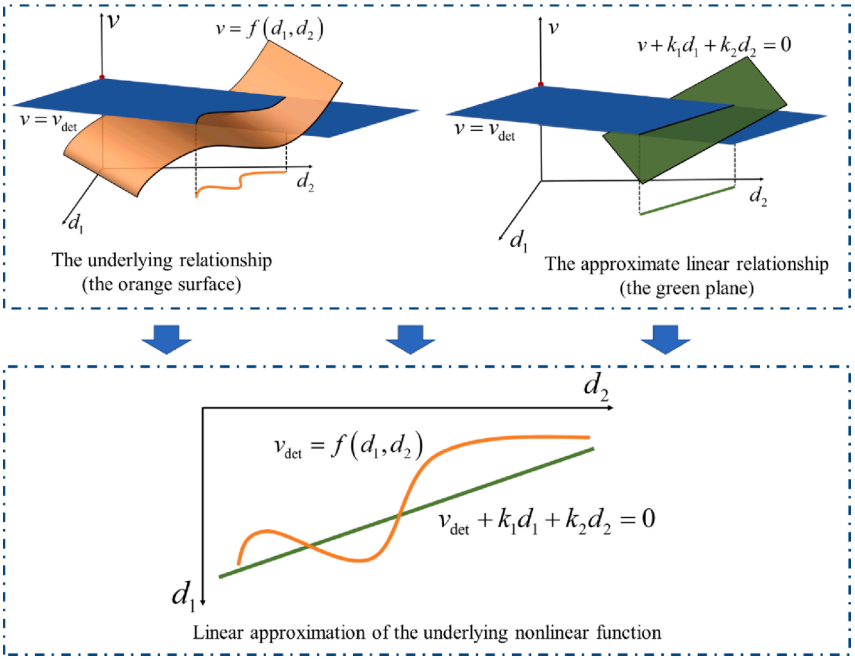


Fig. 3. An intuitive example of the linear approximation.

3.2.2. *The steady-speed model (M2)*

The drawback of M1 lies in the assumption that all vehicles travel at a fixed speed. As shown in Fig. 3, the linear approximation provided by M1 can only roughly describe the trend of the underlying non-linear function. Thus, M2 is proposed by relaxing the fixed speed assumption in M1. Assume that all vehicles travel at a steady speed and the travel time from traffic zones to sensors follows normal distributions. This assumption has been reported in previous literature (Lin and Chang, 2005), and supported by (Hollander and Liu, 2008). Thus, we have:

M2

$$\boldsymbol{P}\boldsymbol{d} = \boldsymbol{v} + \boldsymbol{\varepsilon} \tag{15}$$

$$\boldsymbol{\varepsilon} \sim \text{Multivariate normal}(0, \Sigma) \tag{16}$$

where Σ is the covariance matrix. Traffic counts collected from different sensors are assumed to be independent. Hence, Σ is a diagonal matrix.

$$\Sigma = \phi \boldsymbol{E} \phi^T \tag{17}$$

where $\boldsymbol{E}$ is a unit matrix, $\phi = [\sigma_1, ..., \sigma_K]^T$. $\sigma_i^2, i \in \{1, ..., K\}$ defines the variance of traffic counts given $\boldsymbol{d}$.

Based on the sampled points in GP, we employ Bayesian linear regression to estimate the coefficient matrix $\boldsymbol{P}$ and the covariance matrix Σ. Details on the linear regression are enclosed in the Appendix. Let $\widehat{\boldsymbol{P}}$ and $\widehat{\Sigma}$ denote the estimation of $\boldsymbol{P}$ and Σ, respectively. $\widehat{\phi}$ defines the estimation of the standard variance vector, $\widehat{\phi} = [\widehat{\sigma}_1, ..., \widehat{\sigma}_K]^T$. Given the field traffic counts, $\boldsymbol{v}^f$, we have

$$\widehat{\boldsymbol{P}}\boldsymbol{d} \sim \text{Multivariate normal}\left(\boldsymbol{v}^f, \widehat{\Sigma}\right) \tag{18}$$

Eq. (18) defines the distribution of traffic demand $\boldsymbol{d}$. For a normal distribution, the values less than two standard deviation away from the mean account for 95.45 % of the set; while three standard deviations from the mean account for 99.73 %. Hence, the solution of $\boldsymbol{d}$ is restricted in the region:

$$\boldsymbol{v}^f - r\widehat{\phi} \leqslant \widehat{\boldsymbol{P}}\boldsymbol{d} \leqslant \boldsymbol{v}^f + r\widehat{\phi} \tag{19}$$

where $r > 0$ defines the confidence bound of M2.

Compared with M1, M2 introduces a disturbance term to represent the error of traffic count caused by the variation of travel time. Instead of linear approximation, M2 approximates the nonlinear relationship using a restricted region. Fig. 4 represents the restricted region using the two-variable example. The real nonlinear function (i.e., the orange surface) is restricted in a region, which is bounded by two linear functions (i.e., the green planes). Given an observation of traffic flow, the relationship between the two variables can be determined. The nonlinear relationship between the two variables is approximated using the green region, which is bounded by linear

equations.

Even though there have been other studies that established linear relationships between OD demands and traffic flows, existing linear models are usually formulated purely based on assumptions of travel speeds/times. However, the proposed physical surrogate model is constructed using Bayesian statistics, by integrating assumptions on travel speed/times and outputs of simulation. This way, the physical surrogate model could be updated in each iteration based on simulation outputs, through which the physical surrogate model can adjust its parameters and provide accurate global approximations.

3.3. Combing two surrogate models

As aforementioned, Gaussian process provides comparatively accurate approximations around sample points, whereas, the performance of GP is not guaranteed in unsampled regions. In contrast, global, yet less precise approximations are provided by the physical surrogate model. That is, solutions defined by M1/M2 can provide reasonable solutions on the DODE problem, whereas, the physical surrogate model cannot assess which point is better. In terms of combing functional/physical surrogate models, the most widely used approach is the *weighted sum surrogate model* Osorio and Bierlaire (2013). M3 shows the surrogate model that combines GP and the linear model in the way of weighted sum.

M3:

$$f(\boldsymbol{d}) = \beta_{GP} f_{GP}(\boldsymbol{d}) + \beta_{LM} f_{LM}(\boldsymbol{d}) \tag{20}$$

where $f(\boldsymbol{d})$ is the surrogate model with respect to $\boldsymbol{d}$. β_{GP} and β_{LM} are weight coefficients of the two components. $f_{GP}(\boldsymbol{d})$ and $f_{LM}(\boldsymbol{d})$ denote the approximate functions derived by GP and the physical surrogate model. In (20), $f_{LM}(\boldsymbol{d})$ can be seen as soft constraints in M3. That is, the physical surrogate model revises the approximate function formulated by GP. As results, problem-specific knowledge information could be considered in searching for new points.

This study combines the two types of surrogate models in a new way: embedding the physical surrogate model into the functional surrogate model. The surrogate model is mathematically formulated as follows:

M*:

$$f_{GP}(\boldsymbol{d}) \tag{21}$$

subject to:

$$g_{LM}(\boldsymbol{d}) = 0 \tag{22}$$

$$h_{LM}(\boldsymbol{d}) < 0 \tag{23}$$

where (22)–(23) are the constraints derived from the physical surrogate model. Except for providing global approximations, the physical surrogate model also restricts the feasible region. Given the restricted region, the functional surrogate model is employed to provide accurate local approximations for assessing which point is better. Different from M3, the physical surrogate model in M* can be seen as hard constraints, which are expected to reduce the feasible region and lead to less required sample points.

A direct way of embedding M1 into GP is adopting Eq. (14) as the equality constraints (22) in M*. This way, the problem dimension is reduced (i.e., the basic variables are represented using the non-basic ones). Whereas, this leads to a severely narrow feasible set. It is revealed in practice that the surrogate is prone to be stuck in local optimum. Hence, in this study, we formulate the surrogate model by embedding M2 into Gaussian process. Given the field traffic counts $\boldsymbol{v}^f$ and the sample points set $\{(\boldsymbol{d}_1, f_1), \ldots, (\boldsymbol{d}_n, f_n)\}$, the surrogate model is thus formulated as:

$$f(\boldsymbol{d})|\boldsymbol{d} \sim N(\boldsymbol{K}_*^T \boldsymbol{K}_n^{-1}(\boldsymbol{F}_n - \boldsymbol{\mu}_n) + \mu_*,\ \boldsymbol{K}_{**} - \boldsymbol{K}_*^T \boldsymbol{K}_n^{-1} \boldsymbol{K}_*) \tag{24}$$

subject to:

$$\boldsymbol{v}^f - r\hat{\phi} \leqslant \hat{\boldsymbol{P}}\boldsymbol{d} \leqslant \boldsymbol{v}^f + r\hat{\phi} \tag{25}$$

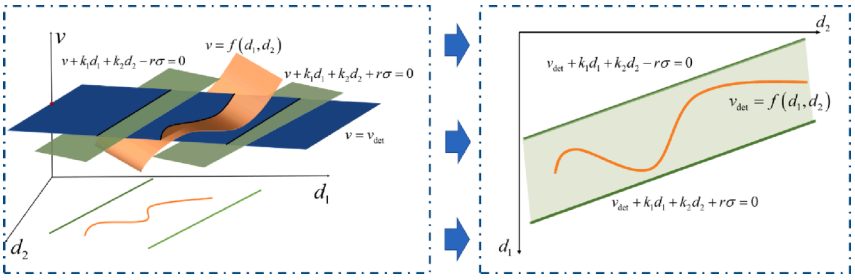


Fig. 4. Approximating the nonlinear relationship by using a restricted region.

where $r > 0$ defines the confidence bound of the physical surrogate model.

4. The optimization algorithm

As aforementioned, surrogate-based optimization algorithms iterate between the surrogate model formulation step (Section 3) and the new sample points selection step. In the second step, acquisition functions are used to suggest new sample points for the refinement of surrogates. We first formulate the acquisition function, which is established considering the constraints. Then, details on the optimization algorithm are presented.

4.1. A projection-distance based acquisition function

Acquisition functions are designed based on surrogate models to suggest where additional evaluations may help improve the estimate of objective functions. Acquisition functions are generally formulated as a tradeoff between high expected performance and high uncertainty. One of the most widely used acquisition functions is the Gaussian Process Upper Confidence Bound (GP-UCB). GP-UCB may be first proposed in Srinivas et al. (2009), in which the regret bound of adopting GP-UCB is derived for common kernels. GP-UCB is formulated as the linear combination of the estimated value (i.e., the expectation in (9)) and the uncertainty (i.e., the variance in (9)):

$$\boldsymbol{d}_{n+1} = \underset{\boldsymbol{d}}{\operatorname{argmax}} \mu_n(\boldsymbol{d}) + \gamma(n+1)\sigma_n(\boldsymbol{d}) \tag{26}$$

where $\mu_n(\boldsymbol{d}) = \boldsymbol{K}_*^T \boldsymbol{K}_n^{-1}(\boldsymbol{Y}_n - \boldsymbol{\mu}_n) + \mu_*, \sigma_n(\boldsymbol{d}) = \boldsymbol{K}_{**} - \boldsymbol{K}_*^T \boldsymbol{K}_n^{-1} \boldsymbol{K}_*$. $\gamma(n+1) > 0$ is a trade-off parameter, depending on the current iteration. To accommodate constraints derived from physical surrogate models (i.e., Eq. (25)), the constraints are incorporated into the optimization of GP-UCB.

$$\boldsymbol{d}_{n+1} = \underset{\boldsymbol{d}}{\operatorname{argmax}} \mu_n(\boldsymbol{d}) + \gamma(n+1)\sigma_n(\boldsymbol{d}) \tag{27}$$

subject to:

$$\boldsymbol{v}^f - r\widehat{\phi} \leqslant \widehat{\boldsymbol{P}}\boldsymbol{d} \leqslant \boldsymbol{v}^f + r\widehat{\phi} \tag{28}$$

Since the physical surrogate model is constructed in a linear form, the constraints are also linear, making them easy to handle using existing nonlinear optimization techniques, e.g., Lagrange methods (Huang et al., 2022).

Many studies have dedicated to the design of $\gamma(n+1)$ to achieve quick convergence. For example, Contal et al. formulated $\gamma(n+1)$ such that the objective function is contained by the high confidence region for all iterations (Contal et al., 2013). Berk et al. (2020) set $\gamma(n+1)$ by sampling from a gamma distribution whose parameters are associated with the current iteration. Note that in all the previous settings, the tradeoff parameter is only concerned with the current iteration rounds. In other words, in an arbitrary iteration, the weights of uncertainty are equally considered for all points in the feasible region. However, when it comes to optimization problems with constraints, these settings of $\gamma(n+1)$ do not seem to make sense. As few samples are selected outside of the feasible region, the variance of estimation for points around boundaries is extremely high. Thus, points around boundaries have higher acquisition function values and higher probability of being selected than interior points. This excessive exploration around boundaries leads to large amounts of redundant sample points around boundaries and waste of computational resources.

To compensate the drawbacks of the tradeoff parameter in the widely used acquisition functions, we proposed a projection-distance weighting factor (PWF). PWF is a continuously differentiable function defined in the feasible set, ranging from 0 to 1. For each point in the feasible region, PWF provides a unique weight of uncertainty based on projection distance. Let $h(\boldsymbol{d})$ denote the minimal projection distance from point $\boldsymbol{d}$ to boundaries of the feasible region. PWF is defined as a piecewise function:

$$PWF(h(\boldsymbol{d})) = \begin{cases} -\dfrac{1}{\rho^2}h(\boldsymbol{d})^2 + \dfrac{2}{\rho}h(\boldsymbol{d}) & 0 \leqslant h(\boldsymbol{d}) \leqslant \rho \\ 1 & h(\boldsymbol{d}) \geqslant \rho \end{cases} \tag{29}$$

In (29), $\rho > 0$ defines the maximal distance of adopting the reduced weight, that is, the weights of uncertainty are reduced only for points away from boundaries less than ρ. With PWF, the constrained GP-UCB acquisition function is reformulated as:

$$\min_{\boldsymbol{d}} H(\boldsymbol{d}) = \min_{\boldsymbol{d}} (\mu_n(\boldsymbol{d}) - PWF(h(\boldsymbol{d}))\gamma(n)\sigma_n(\boldsymbol{d})) \tag{30}$$

subject to:

$$\boldsymbol{v}^f - r\widehat{\phi} \leqslant \widehat{\boldsymbol{P}}\boldsymbol{d} \leqslant \boldsymbol{v}^f + r\widehat{\phi} \tag{31}$$

4.2. Algorithm for solving dynamic OD estimation problems on freeway networks

In this section, the solving algorithm is represented in detail.

Algorithm for solving the dynamic OD estimation problem on the freeway network

Step 0: *Initialization.* Randomly sample n points in the feasible region. Evaluate the points through simulation. Let $\Theta_n = \{(\boldsymbol{d}_1, f_1), \ldots, (\boldsymbol{d}_n, f_n)\}$ denote the sampled point set.

Step 1: *Formulating the surrogate model.*

a) Adopt Bayesian linear regression to estimate the parameters in M2. Update the feasible region according to (19)
b) Formulate the Gaussian process regression model according to (9)

Step 2: *Sampling new points.*

a) Solve the optimization problem (30)–(31) and obtain a new point $\boldsymbol{d}_{n+1}$.
b) Evaluate the new point through simulation, calculate its associated objective function value f_{n+1}.

Step 3: *Dataset updating.* Update the sampled point set: $\Theta_n = \Theta_n \cup \{(\boldsymbol{d}_{n+1}, f_{n+1})\}, n = n + 1$

Step 4: *The criticality step.* If the predetermined accuracy level is met, then switch to **Step 5**, or switch to **Step 1**.

Step 5: *Searching for the optimal solution.* Given Θ_n, formulate Gaussian process regression model. Let $\mu_n(\boldsymbol{d})$ denote $\mathrm{E}[f(\boldsymbol{d})|\Theta_n, \boldsymbol{d}]$. Solve $\boldsymbol{d}_{opt} = \underset{\boldsymbol{d}}{\operatorname{argmin}} \mu_n(\boldsymbol{d})$. **Return** $\boldsymbol{d}_{opt}$.

5. Numerical experiments

Numerical experiments are conducted for validating the proposed models and algorithms in this section. The capability of the physical surrogate model on providing global approximations is first validated in Section 5.1 on a freeway corridor. Then, the performance of the Bayesian optimization approach is evaluated on a large-scale network. Simulation evaluations in this section are conducted using the popular microscopic traffic simulator SUMO (Krajzewicz et al., 2002). The methods are coded in Python and implemented on a personal computer with Intel Core (TM) Duo 2.3 GHz CPU, 16 GB RAM.

5.1. Validation of the physical surrogate model

The physical surrogate model is validated on a freeway corridor in Gansu, China (Fig. 5). The freeway corridor consists of 3 traffic zones (i.e., 2 mainline cross-sections and 1 ramp) and 6 ODs (i.e., two directions of traffic flow are considered). The yellow rectangles in Fig. 5 denote road sensors, from which traffic flows are collected with an interval of 5-minute. Assume that the traffic demands vary with an interval of 15-minute. In this experiment, traffic demands within one interval are analyzed.

We first evaluate the capability of the physical surrogate model to capture the underlying interactions between traffic demands and traffic flows. 5 initial points are uniformly drawn from the feasible region and evaluated by the simulator, which are used for estimating parameters of the physical surrogate model (i.e., $\widehat{\boldsymbol{P}}$ and $\widehat{\Sigma}$ in (18)). Then 60 points are generated as the validation set. The points

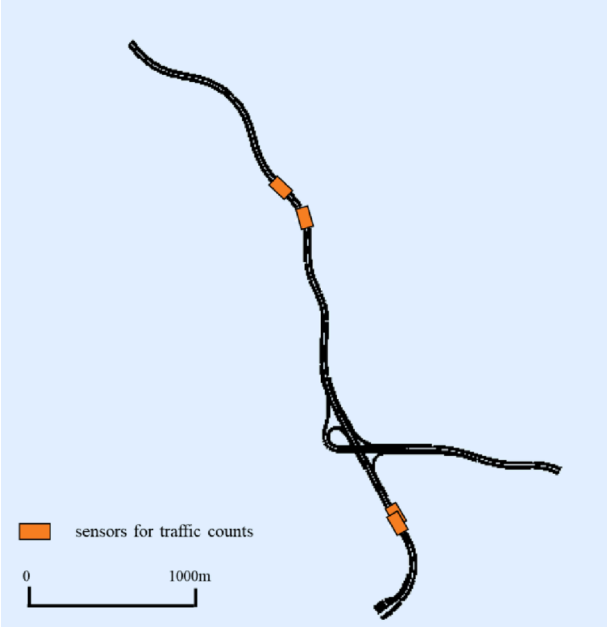


Fig. 5. The simulation network of a freeway corridor.

are obtained by assuming that all OD's have common values (i.e., all elements of the OD matrix have the same value). Fig. 6 displays, for each sensor, the simulated traffic flows and the estimated traffic flows versus the common OD value. The simulated traffic flows are represented using green circles. Since traffic flows are assumed to follow norm distributions in M2, the expectation of estimated traffic flows (the blue dash lines) as well as the 3σ confidence bounds (the blue solid lines) are shown in Fig. 6. As shown in Fig. 6, all simulated traffic flows fall into the 99.7 % confidence intervals defined by the physical surrogate model. For all demand levels, the physical surrogate model yields an accurate estimation on traffic flows.

We then evaluate the performance of the surrogate model to approximate the objective function (i.e., Eq. (1)). Since field data is not accessible, the field traffic counts are generated, based on a randomly generated OD matrix, through simulation. The above-mentioned model parameters and validation data set are used to estimating the objective function value. For each point in the validation set, 20 sets of traffic flows are sampled from the distribution given by M2, which are then used for calculating objective function values. Fig. 7 shows the comparison of the estimated objective function values (i.e., the blue points) and the exact calculated ones (i.e., the red pluses). The physical surrogate model yields accurate approximations on the objective function. Moreover, the dark blue points in Fig. 7 denote points which satisfy the constraints defined by the physical surrogate model and the field traffic counts (i.e., the constraint (19)) when the parameter r takes 10. As shown in Fig. 7, these points are concentrated, with the minimal objective function values. This represents the capability of the physical surrogate model to identify solutions with good performance. Meanwhile, the physical surrogate model can effectively restrict the feasible set.

5.2. Experiments on a large-scale freeway network

The proposed approaches are validated on a large-scale network in this section. The study area, with a length of 40.2 km, is a freeway corridor in Taipei, China, see Fig. 8 (a). It consists of 6 traffic zones (i.e., 3 ramps and 3 mainline cross-sections) for a total of 30 OD pairs in two directions. In this study, the dynamic OD demand on the network in 2 hours is analyzed. Assume that the OD demand varies with an interval of 15-minute. Thus, the dimension of the dynamic OD estimation problem is 240. There are 28 sensors in the corridor (i.e., the yellow rectangles in Fig. 8 (b)). Traffic flows are obtained from the sensors with an interval of 5-minute. The field data is not accessible, and hence a synthetic OD matrix is generated and set as the "true" OD. The "true" OD matrix is constructed to account for the time-varying demand conditions (i.e., from uncongested to congested scenarios and vice versa). The average travel demand of all OD pairs in each time interval is represented in Fig. 8 (c). Based on the "true" OD, field traffic counts are generated through simulation. It was revealed in practice that sensors being deployed at different locations of the same segment would construct "similar" inequalities in M* (i.e., inequalities have the same structure and slightly different parameters). As results, multiple sensors on the same segment would make the surrogate model ill-conditioned. To avoid this issue, for segments with multiple sensors deployed, one sensor is randomly chosen to construct physical surrogate models.

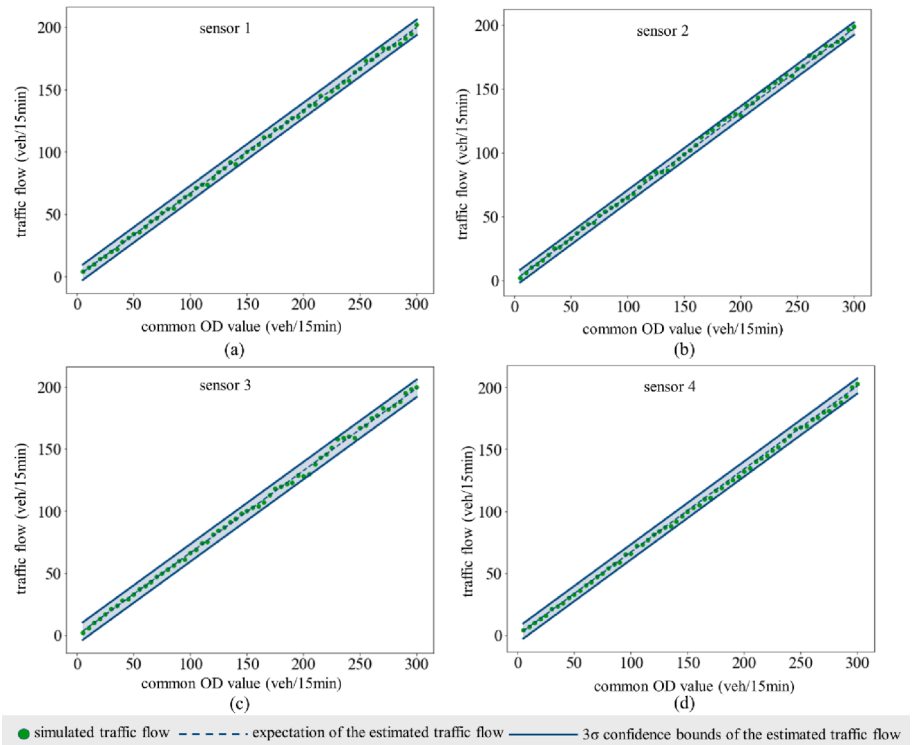
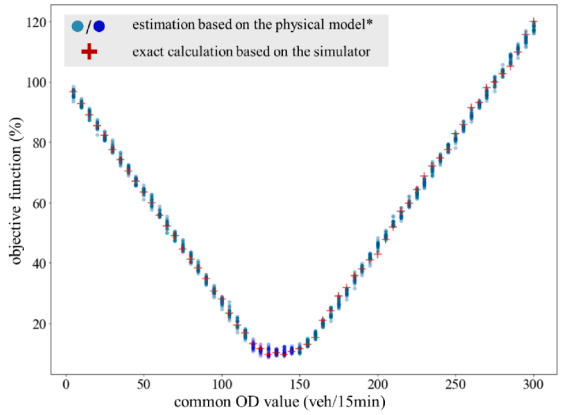
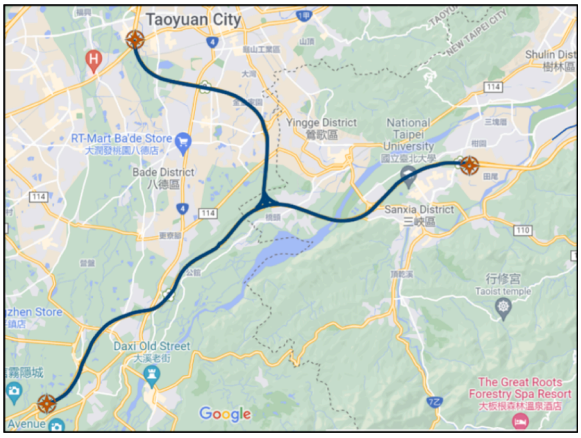


Fig. 6. Comparison of the simulated traffic flow and the traffic flow estimated by the physical surrogate model, each plot for one sensor.

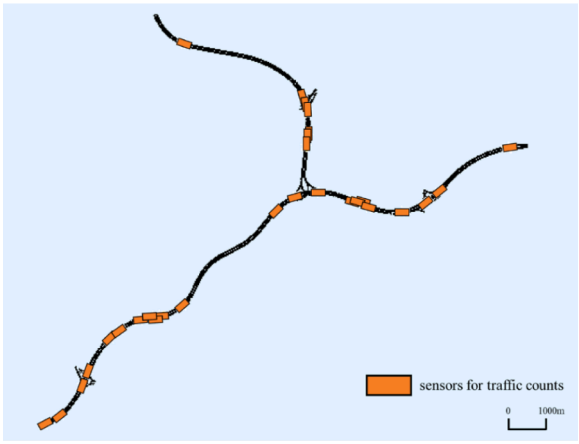


* The dark blue points denote the solutions that satisfy the constraints defined by the physical surrogate model and detected traffic flows (i.e., Eq.(19)) when r takes 10

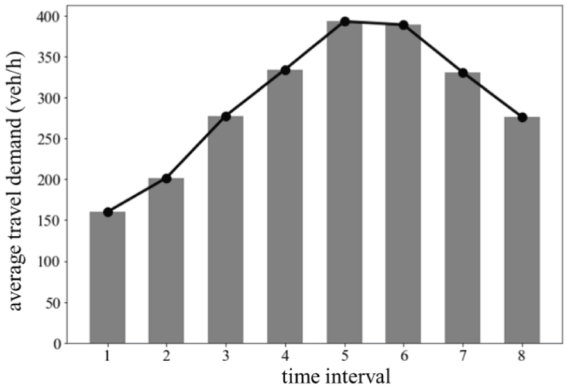
Fig. 7. Comparison of the simulation-based estimates and the physical surrogate model-based approximations of the objective function.



(a)



(b)



(c)

Fig. 8. The freeway network: (a) map of the freeway network. (b) The simulation network. (c) The average travel demand of all OD pairs in each time interval.

We now benchmark the performance of the proposed method to that of three other methods: the stochastic perturbation simultaneous approximation (SPSA) algorithm, the Bayesian optimization algorithm and the approach proposed in Osorio (2019b). The SPSA parameters are set based on standard guidelines in Spall (2005). The trade-off parameter $\gamma(n+1)$ in Eq. (26) is set to be 1. The parameter r in (25) is set to be 2. The effects of r will be discussed later.

The performance of different methods is compared under the same computational budget. The experiment is designed as follows: one point (i.e., an OD demand matrix) is sampled randomly in the feasible region and set as the initial point. We terminate an algorithm after 24 hours. The performance of the methods as a function of computational time is shown in Fig. 9. The x-axis displays the CPU time and the y-axis displays the objective function estimate of the current iterate (i.e., of the best point identified so far by the method).

As shown in Fig. 9, the proposed method outperforms the benchmark methods under the same computational budget. The proposed method identified points with good performance (i.e., the points with corresponding objective function values less than 10 %) at the first few iterations. This demonstrates that the physical model can provide global approximations on the objective function and restricts the feasible region effectively. After the first few iterations, there are also slight decreases on the objective function estimate. These decreases are mainly resulted from the functional surrogate model, the Gaussian process. Similar decreases can also be found in the performance of BO (the blue line with crosses). This indicates that GP is able to capture the complex relationships between OD matrices and traffic counts, and suggests better solutions based on samples.

In Fig. 9, the model of Osorio (2019b) shows similar decreasing trends with the proposed method at the first few iterations. Whereas, after that, the objective function estimate keeps flat. This indicates that the physical surrogate model of the model may fail to accurately model traffic flows on freeways, despite its superior performance on urban networks. The comparison between the model and SPSA complements the results of Osorio (2019b) on freeways. The reasons can be analyzed as follows: the physical surrogate model in Osorio (2019b) is originally designed for urban networks, in which the travel demand is proportionally assigned to different routes. Link flows are calculated as the sum of the expected route demand for routes containing the link. Nonetheless, when the model is applied to freeways, where route choices are not involved, the link flows are the sum of total travel demands passing through the link, which results in extremely high traffic flows for all links on freeways. Since the physical surrogate model is less suitable for freeways, the model of Osorio (2019b) shows comparatively lower accuracy than SPSA. Comparison between the proposed method and benchmark methods illustrated the effectiveness of the proposed physical surrogate model and GP.

Fig. 10 considers performance of the solutions derived by each method. Each plot displays, for all links with sensors, the field traffic flows (i.e., the traffic flow obtained through the synthetic OD matrix) along the x-axis and the traffic flows recorded in the simulator along the y-axis. The red line in each plot denotes the diagonal line ($y = x$).The results show that the OD matrix derived by the proposed method provides a comparatively accurate fit to the field traffic flows.

The effects of the confidence bound on the performance of the surrogate model are then evaluated. We consider three values of r in (25), $r \in \{1, 2, 3\}$. For each r, we run the algorithm for a common initial point and terminate the algorithm after 24 hours. Note that we did not evaluate the case when r takes 0. Because in the case, M2 becomes M1, and this leads to a severely narrow or sometimes empty feasible set, and prohibits the sampling of new samples. Fig. 11 shows the performance of each method as a function of computational time. Fig. 11 shows that the surrogate model reaches the best performance when r takes 1. Meanwhile, the surrogate model has similar

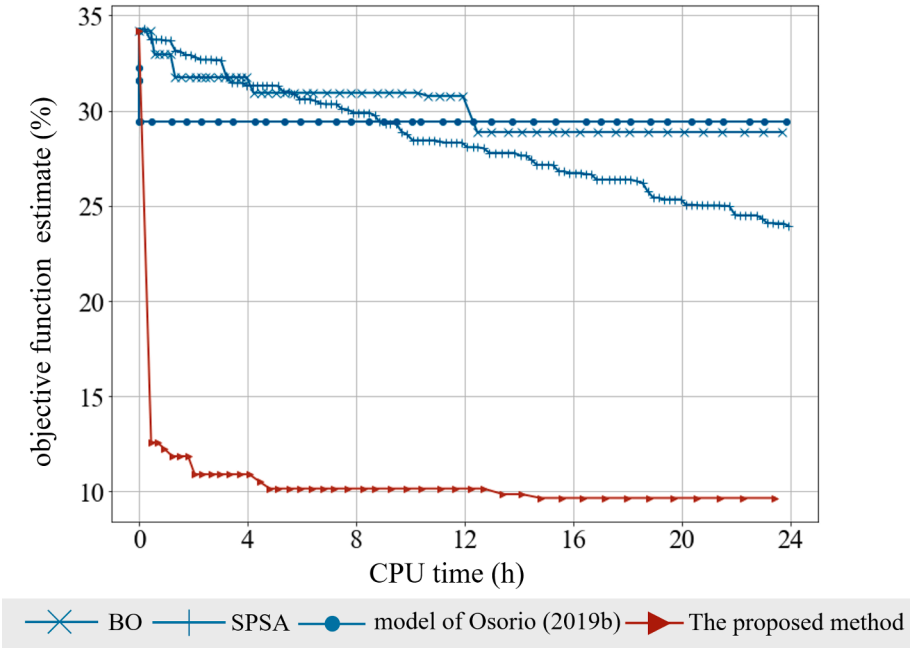


Fig. 9. Performance of each method as a function of the CPU time.

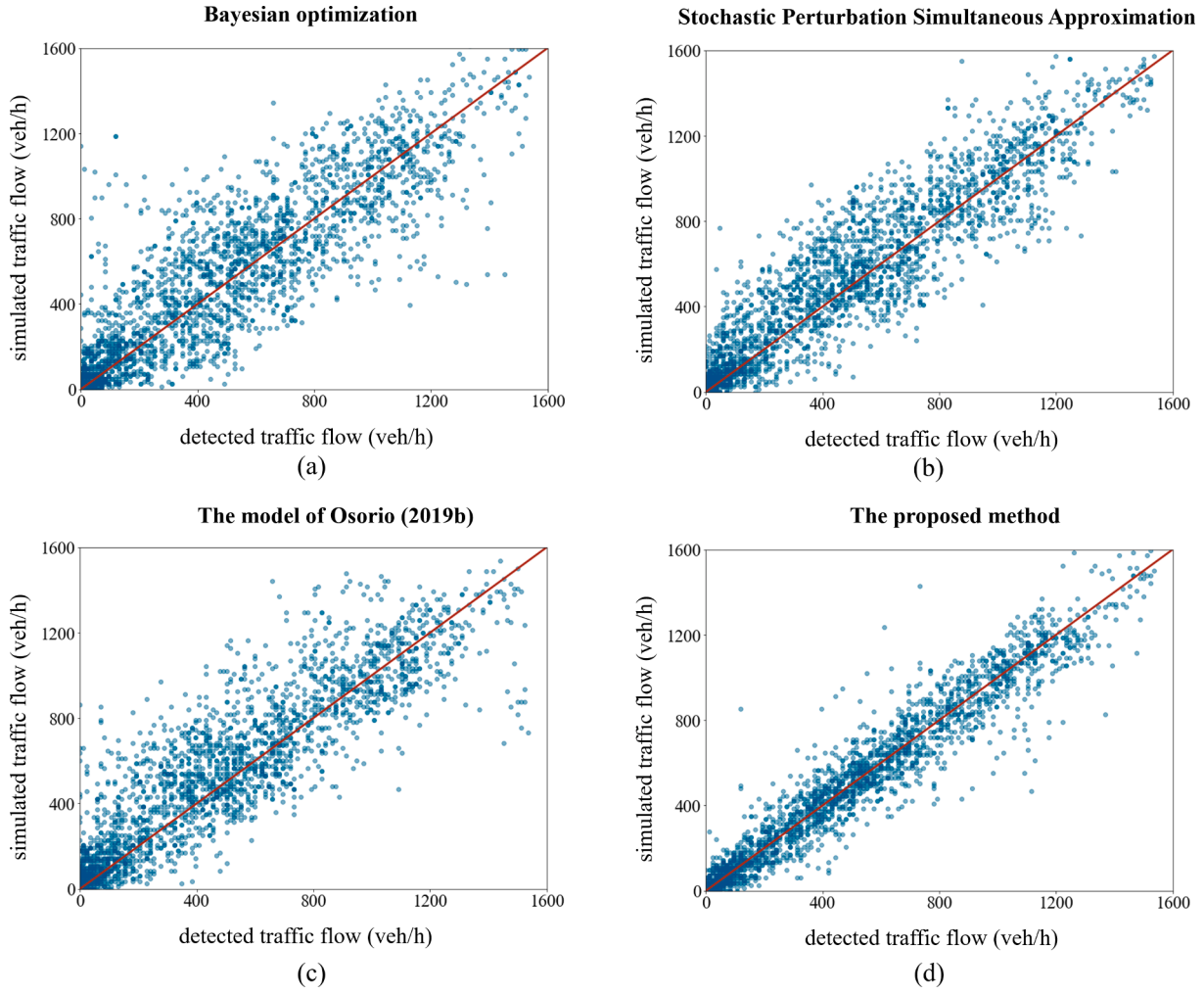


Fig. 10. Fit to traffic flows of the solutions derived by each method.

performance when r takes 1 and 2, which shows robustness of the surrogate model on r. This demonstrates that a comparatively hard restriction on the feasible region is suggested when solving the high-dimensional problem.

Fig. 12 validates the effectiveness of the proposed approach to formulate surrogate models, i.e., embedding the physical surrogate model into the functional surrogate model. The performance of the proposed surrogate model is compared with surrogate models that are formulated by weighted sum (i.e., M3 in Section 3.3). Let $\beta = \beta_{LM}/\beta_{GP}$ denote the relative weight of the two components. We consider three values of β, $\beta \in \{0.3, 0.6, 0.9\}$. $f_{LM}(\boldsymbol{d}')$ in M3 is derived by substituting (13) into (1). Fig. 12 compares the performance of each method as a function of computational time. The results indicate that while solving DODE problems on freeways, embedding outperforms the weighted sum method in designing surrogate models. This demonstrates the effectiveness of the embedding approach in designing surrogate models for high-dimensional problems.

6. Conclusion

This paper focused on simulation-based dynamic OD estimation problems on freeway networks. Surrogate-based optimization algorithms were considered in this study. To cope with the high-dimensionality of the problem, a tailored Bayesian optimization approach was proposed, in which a novel surrogate model was formulated by embedding a physical surrogate model into a functional surrogate model. More specifically, two physical surrogate models were proposed to capture the interrelations between traffic demands and traffic counts. An emerging Bayesian regression model, Gaussian process, is served as the functional surrogate model. Moreover, to deal with constraints in the surrogate model, a projection-distance based acquisition function was proposed to enhance the efficiency of the algorithm.

Effectiveness of the proposed model and algorithm was verified based on real-world freeway corridors. Results indicate that:

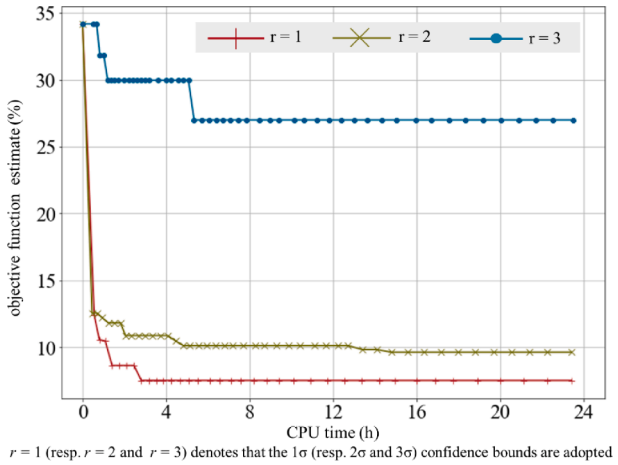


Fig. 11. Performance of the proposed surrogate model as a function of the CPU time, for different confidence bounds.

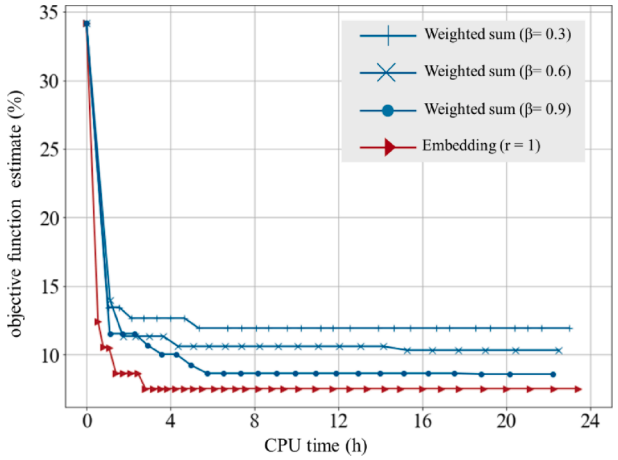


Fig. 12. Comparison of the two approaches (i.e., weighted sum and embedding) to combine the functional surrogate model and the physical surrogate model.

(i) The proposed physical surrogate model can well capture the interrelations between traffic demands and traffic counts, and provide global approximations. Given field traffic counts, the physical surrogate model can effectively identify solutions with good performance. Thus, the physical surrogate model is capable of effectively restricting the feasible set.

(ii) Under the same computational budget, the proposed method outperforms the benchmark methods in terms of accuracy of the solution. Owing to the effectiveness of the physical surrogate model, the proposed method can identify points with good performance at the first few iterations. Since GP is able to model complex and highly-nonlinear relationships, the proposed method can further identify better points after the first few iterations.

(iii) The proposed method achieves the best performance when the parameter r takes 1. A comparatively hard restriction on the feasible region is suggested when solving the high-dimensional problem.

(iv) We also compared the effectiveness of ways to combine functional surrogate models and physical surrogate models. Experimental results show that the proposed embedding approach outperforms the existing weighted sum approach.

While the proposed method is developed specifically for dynamic OD estimation problems, it provides a general framework to solve high-dimensional black-box optimizations in different domains: i) it constructs problem-specific physical surrogate models based on domain knowledge. The physical surrogate model should be analytical, and could globally describe the complex system; ii) it incorporates the physical surrogate model into existing surrogate-based optimization frameworks (e.g., the Bayesian optimization algorithm), such that problem-specific knowledge could be used to restrict feasible regions of high-dimension problems and inform the sampling of solutions.

There are also limitations of the proposed method. First, the physical surrogate model is constructed by assuming that all vehicles travel at a steady speed and the travel time from traffic zones to sensors follows normal distributions. This assumption may be violated

in practical applications, as traffic conditions in reality could be more complex, for example, working zones and speed limits are usually involved in freeways. These conditions could bring biased estimation of OD matrices. Second, the proposed method requires a microscopic traffic simulation model, which involves high computational costs and detailed data requirements (e.g., vehicle types, driver behaviors, road network geometry and topology). These requirements may limit the applicability of the proposed method in practice, especially for local authorities with limited resources and access to detailed traffic data. Therefore, developing a lightweight version of the proposed method or using simplified traffic simulation models that could still capture essential travel behaviors is an important direction for future research (Simon et al., 2022).

A natural progression of this work is to apply the proposed approaches to solve other high-dimensional simulation-based transportation problems. One possible challenge would be designing customized and suitable physical surrogate models for these problems. In the transportation discipline, a considerable amount of analytical traffic models (Huang et al., 2021, Cheng et al., 2022, Jiang and Nielsen, 2022, Zhu et al., 2022) have been established towards different problems in various scenarios. These analytical models are usually elegant and efficient, and have great potential to be employed as physical surrogate models in the proposed framework. We believe that integrating existing analytical traffic models into our proposed method and designing tailored physical surrogate models for different transportation problems would be significant extensions of this study. We also noticed that traffic measurements with flaws could affect the performance of the proposed method. Therefore, incorporating data preprocessing techniques in conjunction with our proposed method to solve dynamic OD estimation problems could be a meaningful extension in further research.

7. Author statement

The authors confirm contribution to the paper as follows; study conception and design: J. Huo, Z. Liu, C. Liu, J. Wang, Q. Meng; data preparation: J. Huo, J. Chen; analysis and interpretation of results: J. Huo, Q. Cheng, J. Chen, J. Wang; draft manuscript preparation: J. Huo, J. Chen, Z. Liu, Q. Cheng. All authors reviewed the results and approved the final version of the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This study is supported by the National Key Research and Development Program of China (No. 2021YFB1600100), the Key Project (No. 52131203) and the Youth Program (No. 52102375) of National Natural Science Foundation of China, the Youth Program (No. BK20210247) of Natural Science Foundation of Jiangsu Province, China, and the Fundamental Research Funds for the Central Universities, China (No. 2242022R40025).

Appendix A. Bayesian linear regression for constructing the steady-speed model (M2)

This section represents the method of constructing M2 using sample points. The relationship between OD matrices and traffic counts on a link during a specific time period is taken as an example to illustrate the regression process.

Given input demand $\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n$, and the corresponding observed traffic counts on a link during a specific time period, $v_1, v_2, \ldots, v_n$, assume that the input variables and the observed traffic counts follow the following relationship:

$$v_i = \boldsymbol{\beta}^T \boldsymbol{d}_i + \varepsilon_i \tag{32}$$

where $\boldsymbol{\beta}$ is a vector of coefficients to be estimated from the data, ε_i is a random variable, and follows a norm distribution:

$$\varepsilon_1, \quad \varepsilon_2, \ \ldots, \varepsilon_n \sim i.d.d.\ \mathcal{N}\left(0, \sigma^2\right) \tag{33}$$

Given (32) and (33), we aim to estimate $\boldsymbol{\beta}$ and σ based on the sample points. Let $\boldsymbol{D} = (d_1, d_2, \ldots, d_n)^T, \boldsymbol{V} = (v_1, v_2, \ldots, v_n)^T$. $\boldsymbol{V}$ follows a multivariate norm distribution:

$$V \sim \mathcal{MVN}\left(\boldsymbol{D}\boldsymbol{\beta}, \sigma^2 \mathrm{I}\right) \tag{34}$$

In Bayesian linear regression, $\boldsymbol{\beta}$ and σ are assumed to have the following prior distribution:

$$\boldsymbol{\beta} \sim \mathcal{MVN}(\boldsymbol{\beta}_0, \Sigma_0) \tag{35}$$

$$1/\sigma^2 \sim \Gamma\left(v_0/2, v_0\sigma_0^2/2\right) \tag{36}$$

In the prior distributions, $\boldsymbol{\beta}_0$ is determined in the same way of determining $\boldsymbol{P}$ in (13). σ_0^2 is set as $(V - \boldsymbol{D}\boldsymbol{\beta}_0)^T (V - \boldsymbol{D}\boldsymbol{\beta}_0)/(n-1)$, and Σ_0 is set as $\left(\boldsymbol{D}^T\boldsymbol{D}\right)^{-1}\sigma_0^2$ The full conditional posterior distributions of $\boldsymbol{\beta}$ and σ are:

$$p(\boldsymbol{\beta}|\boldsymbol{D},\boldsymbol{V},\sigma^2)\propto p(\boldsymbol{V}|\boldsymbol{D},\boldsymbol{\beta},\sigma^2)p(\boldsymbol{\beta}) \tag{37}$$

$$p(\sigma^2|\boldsymbol{D},\boldsymbol{V},\boldsymbol{\beta})\propto p(\boldsymbol{V}|\boldsymbol{D},\boldsymbol{\beta},\sigma^2)p(\sigma^2) \tag{38}$$

where $p(\boldsymbol{V}|\boldsymbol{D},\boldsymbol{\beta},\sigma^2)$ is given by (34). Substituting (35) and (36) into (37) and (38), respectively, we obtain the posterior distributions of $\boldsymbol{\beta}$ and σ:

$$\boldsymbol{\beta}|\boldsymbol{D},\boldsymbol{V},\sigma^2 \sim \mathscr{MVN}(\mathrm{E}_\beta, \mathrm{VAR}_\beta) \tag{39}$$

$$E_\beta = \left(\Sigma_0^{-1} + \boldsymbol{D}^T\boldsymbol{D}/\sigma^2\right)^{-1}\left(\Sigma_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{D}^T\boldsymbol{V}/\sigma^2\right) \quad VAR_\beta = \left(\Sigma_0^{-1} + \boldsymbol{D}^T\boldsymbol{D}/\sigma^2\right)^{-1}$$

$$1/\sigma^2|\boldsymbol{V},\boldsymbol{D},\boldsymbol{\beta} \sim \Gamma\left([v_0+n]/2, \left[v_0\sigma_0^2 + SSR(\boldsymbol{\beta})\right]/2\right) \tag{40}$$

where $SSR(\boldsymbol{\beta}) = (\boldsymbol{V}-\boldsymbol{D}\boldsymbol{\beta})^T(\boldsymbol{V}-\boldsymbol{D}\boldsymbol{\beta})$. Given, (39) and (40), the full conditional posterior distribution of $\boldsymbol{\beta}$ and $1/\sigma^2$ can be approximated using a Gipps sampler (Hoff, 2009).

References

Alexander, L., Jiang, S., Murga, M., González, M.C., 2015. Origin–destination trips by purpose and time of day inferred from mobile phone data. Transport. Res. Part C: Emerg. Technol. 58, 240–250.

Amaran, S., Sahinidis, N.V., Sharda, B., Bury, S.J., 2016. Simulation optimization: a review of algorithms and applications. Ann. Operat. Res. 240 (1), 351–380.

Ameli, M., Lebacque, J.P., Leclercq, L., 2020. Simulation-based dynamic traffic assignment: meta-heuristic solution methods with parallel computing. Comput. Aided Civ. Inf. Eng. 35 (10), 1047–1062.

Antoniou, C., Azevedo, C.L., Lu, L., Pereira, F., Ben-Akiva, M., 2015. W-SPSA in practice: approximation of weight matrices and calibration of traffic simulation models. Transp. Res. Proc. 7, 233–253.

Antoniou, C., Barceló, J., Breen, M., Bullejos, M., Casas, J., Cipriani, E., Ciuffo, B., Djukic, T., Hoogendoorn, S., Marzano, V., 2016. Towards a generic benchmarking platform for origin–destination flows estimation/updating algorithms: design, demonstration and validation. Transport. Res. Part C: Emerg. Technol. 66, 79–98.

Ashok, K., Ben-Akiva, M.E., 2000. Alternative approaches for real-time estimation and prediction of time-dependent origin–destination flows. Transp. Sci. 34 (1), 21–36.

Ashok, K., Ben-Akiva, M.E., 2002. Estimation and prediction of time-dependent origin-destination flows with a stochastic mapping to path flows and link flows. Transp. Sci. 36 (2), 184–198.

Bell, M.G., 1991. The real time estimation of origin-destination flows in the presence of platoon dispersion. Transp. Res. B: Methodol. 25 (2–3), 115–125.

Berk, J., Gupta, S., Rana, S., Venkatesh, S., 2020. Randomised gaussian process upper confidence bound for Bayesian optimisation. arXiv preprint arXiv:2006.04296.

Bhosekar, A., Ierapetritou, M., 2018. Advances in surrogate based modeling, feasibility analysis, and optimization: a review. Comput. Chem. Eng. 108, 250–267.

Bull, A.D., 2011. Convergence rates of efficient global optimization algorithms. J. Mach. Learn. Res. 12 (10).

Caggiani, L., Ottomanelli, M., Sassanelli, D., 2013. A fixed point approach to origin–destination matrices estimation using uncertain data and fuzzy programming on congested networks. Transport. Res. Part C: Emerg. Technol. 28, 130–141.

Cao, Y., Tang, K., Sun, J., Ji, Y., 2021. Day-to-day dynamic origin–destination flow estimation using connected vehicle trajectories and automatic vehicle identification data. Transport. Res. Part C: Emerg. Technol. 129, 103241.

Carrese, S., Cipriani, E., Mannini, L., Nigro, M., 2017. Dynamic demand estimation and prediction for traffic urban networks adopting new data sources. Transport. Res. Part C: Emerg. Technol. 81, 83–98.

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

Cats, W., 1959. Chicago Area Transportation Study. Chicago Area Transportation Study Chicago.

Chang, G.-L., Wu, J., 1994. Recursive estimation of time-varying origin-destination flows from traffic counts in freeway corridors. Transp. Res. B: Methodol. 28 (2), 141–160.

Chen, A., Subprasom, K., Ji, Z., 2006. A simulation-based multi-objective genetic algorithm (SMOGA) procedure for BOT network design problem. Optim. Eng. 7 (3), 225–247.

Chen, X.M., Xiong, C., He, X., Zhu, Z., Zhang, L., 2016. Time-of-day vehicle mileage fees for congestion mitigation and revenue generation: a simulation-based optimization method and its real-world application. Transport. Res. Part C: Emerg. Technol. 63, 71–95.

Cheng, Q., Wang, S., Liu, Z., Yuan, Y., 2019. Surrogate-based simulation optimization approach for day-to-day dynamics model calibration with real data. Transport. Res. Part C: Emerg. Technol. 105, 422–438.

Cheng, Q., Liu, Z., Lin, Y., Zhou, X.S., 2021. An s-shaped three-parameter (S3) traffic stream model with consistent car following relationship. Transp. Res. B: Methodol. 153, 246–271.

Cheng, Q., Liu, Z., Guo, J., Wu, X., Pendyala, R., Belezamo, B., Zhou, X.S., 2022. Estimating key traffic state parameters through parsimonious spatial queue models. Transport. Res. Part C: Emerg. Technol. 137, 103596.

Choffin, B., Ueda, N., 2018. Scaling Bayesian optimization up to higher dimensions: a review and comparison of recent algorithms. In: 2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, pp. 1–6.

Cipriani, E., Florian, M., Mahut, M., Nigro, M., 2011. A gradient approximation approach for adjusting temporal origin–destination matrices. Transport. Res. Part C: Emerg. Technol. 19 (2), 270–282.

Contal, E., Buffoni, D., Robicquet, A., Vayatis, N., 2013. Parallel Gaussian process optimization with upper confidence bound and pure exploration. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 225–240.

Dandl, F., Engelhardt, R., Hyland, M., Tilg, G., Bogenberger, K., Mahmassani, H.S., 2021. Regulating mobility-on-demand services: tri-level model and Bayesian optimization solution approach. Transport. Res. Part C: Emerg. Technol. 125, 103075.

Dantsuji, T., Hoang, N.H., Zheng, N., Vu, H.L., 2022. A novel metamodel-based framework for large-scale dynamic origin–destination demand calibration. Transport. Res. Part C: Emerg. Technol. 136, 103545.

Demissie, M.G., Kattan, L., 2022. Estimation of truck origin-destination flows using GPS data. Transport. Res. Part E: Logist. Transport. Rev. 159, 102621.

Doblas, J., Benitez, F.G., 2005. An approach to estimating and updating origin–destination matrices based upon traffic counts preserving the prior structure of a survey matrix. Transp. Res. B: Methodol. 39 (7), 565–591.

Duan, J., Gao, F., He, Y., 2022. Test scenario generation and optimization technology for intelligent driving systems. IEEE Intell. Transp. Syst. Mag. 14 (1).

Fakhrmoosavi, F., Kamjoo, E., Kavianipour, M., Zockaie, A., Talebpour, A., Mittal, A., 2022. A stochastic framework using Bayesian optimization algorithm to assess the network-level societal impacts of connected and autonomous vehicles. Transport. Res. Part C: Emerg. Technol. 139, 103663.

Fisk, C., 1988. On combining maximum entropy trip matrix estimation with user optimal assignment. Transp. Res. B: Methodol. 22 (1), 69–73.
Fisk, C.S., Boyce, D.E., 1983. A note on trip matrix estimation from link traffic count data. Transp. Res. B: Methodol. 17 (3), 245–250.
Forrester, A.I., Keane, A.J., 2009. Recent advances in surrogate-based optimization. Prog. Aerosp. Sci. 45 (1–3), 50–79.
Frazier, P.I., 2018. A tutorial on Bayesian optimization. arXiv preprint arXiv:1807.02811.
Frederix, R., Viti, F., Tampère, C.M., 2013. Dynamic origin–destination estimation in congested networks: theoretical findings and implications in practice. Transportmet. A: Transp. Sci. 9 (6), 494–513.
Fu, H., Lam, W.H., Shao, H., Kattan, L., Salari, M., 2022. Optimization of multi-type traffic sensor locations for estimation of multi-period origin-destination demands with covariance effects. Transport. Res. Part E: Logist. Transport. Rev. 157, 102555.
Ge, Q., Fukuda, D., 2019. A macroscopic dynamic network loading model for multiple-reservoir system. Transp. Res. B: Methodol. 126, 502–527.
Gu, Z., Saberi, M., 2021. Simulation-based optimization of toll pricing in large-scale urban networks using the network fundamental diagram: a cross-comparison of methods. Transport. Res. Part C: Emerg. Technol. 122, 102894.
Gu, Z., Waller, S.T., Saberi, M., 2019. Surrogate-based toll optimization in a large-scale heterogeneously congested network. Comput. Aided Civ. Inf. Eng. 34 (8), 638–653.
Hickish, B., Fletcher, D.I., Harrison, R.F., 2020. Investigating Bayesian Optimization for rail network optimization. Int. J. Rail Transport. 8 (4), 307–323.
Hoff, P.D., 2009. A First Course in Bayesian Statistical Methods. Springer.
Hollander, Y., Liu, R., 2008. Estimation of the distribution of travel times by repeated simulation. Transport. Res. Part C: Emerg. Technol. 16 (2), 212–231.
Huang, D., Xing, J., Liu, Z., An, Q., 2021. A multi-stage stochastic optimization approach to the stop-skipping and bus lane reservation schemes. Transportmet. A: Transp. Sci. 17 (4), 1272–1304.
Huang, D., Wang, Y., Jia, S., Liu, Z., Wang, S., 2022. A Lagrangian relaxation approach for the electric bus charging scheduling optimisation problem. Transportmet. A: Transp. Sci. 1–24.
Huo, J., Liu, Z., Chen, J., Cheng, Q., Meng, Q., 2023. Bayesian optimization for congestion pricing problems: a general framework and its instability. Transp. Res. B: Methodol. 169, 1–28.
Jiang, Y., Nielsen, O.A., 2022. Urban Multimodal Traffic Assignment, Vol. 1. Elsevier, pp. 100027.
Kalahasthi, L., Holguín-Veras, J., Yushimito, W.F., 2022. A freight origin-destination synthesis model with mode choice. Transport. Res. Part E: Logist. Transport. Rev. 157, 102595.
Krajzewicz, D., Hertkorn, G., Rössel, C., Wagner, P., 2002. SUMO (Simulation of Urban MObility)-an open-source traffic simulation. In: Proceedings of the 4th middle East Symposium on Simulation and Modelling (MESM20002).), pp. 183–187.
Laharotte, P.-A., Billot, R., Come, E., Oukhellou, L., Nantes, A., El Faouzi, N.-E., 2014. Spatiotemporal analysis of bluetooth data: Application to a large urban network. IEEE Trans. Intell. Transp. Syst. 16 (3), 1439–1448.
Li, G., Chen, A., 2022. Frequency-based path flow estimator for transit origin-destination trip matrices incorporating automatic passenger count and automatic fare collection data. Transport. Res. Part E: Logist. Transport. Rev. 163, 102754.
Li, Z., Tian, Y., Sun, J., Lu, X., Kan, Y., 2022. Simulation-based optimization of large-scale dedicated bus lanes allocation: using efficient machine learning models as surrogates. Transport. Res. Part C: Emerg. Technol. 143, 103827.
Liessner, R., Lorenz, A., Schmitt, J., Dietermann, A.M., Baker, B., 2019. Simultaneous electric powertrain hardware and energy management optimization of a hybrid electric vehicle using deep reinforcement learning and Bayesian optimization. In: 2019 IEEE Vehicle Power and Propulsion Conference (VPPC). IEEE, pp. 1–6.
Lin, P.-W., Chang, G.-L., 2005. Robust model for estimating freeway dynamic origin-destination matrix. Transp. Res. Rec. 1923 (1), 110–118.
Lin, P.-W., Chang, G.-L., 2007. A generalized model and solution algorithm for estimation of the dynamic freeway origin–destination matrix. Transp. Res. B: Methodol. 41 (5), 554–572.
Lu, Z., Meng, Q., Gomes, G., 2016. Estimating link travel time functions for heterogeneous traffic flows on freeways. J. Adv. Transp. 50 (8), 1683–1698.
Lu, L., Xu, Y., Antoniou, C., Ben-Akiva, M., 2015. An enhanced SPSA algorithm for the calibration of Dynamic Traffic Assignment models. Transport. Res. Part C: Emerg. Technol. 51, 149–166.
Lu, C.-C., Zhou, X., Zhang, K., 2013. Dynamic origin–destination demand flow estimation under congested traffic conditions. Transport. Res. Part C: Emerg. Technol. 34, 16–37.
Lundgren, J.T., Peterson, A., 2008. A heuristic for the bilevel origin–destination-matrix estimation problem. Transp. Res. B: Methodol. 42 (4), 339–354.
Ma, W., Pi, X., Qian, S., 2020. Estimating multi-class dynamic origin-destination demand through a forward-backward algorithm on computational graphs. Transport. Res. Part C: Emerg. Technol. 119, 102747.
Ma, W., Qian, Z.S., 2018. Estimating multi-year 24/7 origin-destination demand using high-granular multi-source traffic data. Transport. Res. Part C: Emerg. Technol. 96, 96–121.
Marzano, V., Papola, A., Simonelli, F., Papageorgiou, M., 2018. A Kalman filter for quasi-dynamic od flow estimation/updating. IEEE Trans. Intell. Transp. Syst. 19 (11), 3604–3612.
Nigro, M., Cipriani, E., Del Giudice, A., 2018. Exploiting floating car data for time-dependent origin–destination matrices estimation. J. Intell. Transp. Syst. 22 (2), 159–174.
Osorio, C., 2019a. Dynamic origin-destination matrix calibration for large-scale network simulators. Transport. Res. Part C: Emerg. Technol. 98, 186–206.
Osorio, C., 2019b. High-dimensional offline origin-destination (OD) demand calibration for stochastic traffic simulators of large-scale road networks. Transp. Res. B: Methodol. 124, 18–43.
Osorio, C., Atasoy, B., 2021. Efficient simulation-based toll optimization for large-scale networks. Transp. Sci. 55 (5), 1010–1024.
Osorio, C., Bierlaire, M., 2013. A simulation-based optimization framework for urban transportation problems. Oper. Res. 61 (6), 1333–1345.
Osorio, C., Chong, L., 2015. A computationally efficient simulation-based optimization algorithm for large-scale urban transportation problems. Transp. Sci. 49 (3), 623–636.
Otsuka, T., Shimizu, H., Iwata, T., Naya, F., Sawada, H., Ueda, N., 2019. Bayesian optimization for crowd traffic control using multi-agent simulation. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC). IEEE, pp. 1981–1988.
Patwary, A.U.Z., Huang, W., Lo, H.K., 2021. Metamodel-based calibration of large-scale multimodal microscopic traffic simulation. Transport. Res. Part C: Emerg. Technol. 124, 102859.
Peled, I., Lee, K., Jiang, Y., Dauwels, J., Pereira, F.C., 2021. On the quality requirements of demand prediction for dynamic public transport. Commun. Transport. Res. 1, 100008.
Qian, Z.S., Li, J., Li, X., Zhang, M., Wang, H., 2017. Modeling heterogeneous traffic flow: a pragmatic approach. Transp. Res. B: Methodol. 99, 183–204.
Qurashi, M., Ma, T., Chaniotakis, E., Antoniou, C., 2019. PC–SPSA: employing dimensionality reduction to limit SPSA search noise in DTA model calibration. IEEE Trans. Intell. Transp. Syst. 21 (4), 1635–1645.
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. Transport. Res. Part C: Emerg. Technol. 95, 29–46.
Shafiei, S., Saberi, M., Sarvi, M., 2016. Application of an exact gradient method to estimate dynamic origin-destination demand for melbourne network. In: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC). IEEE, pp. 1945–1950.
Shafiei, S., Saberi, M., Zockaie, A., Sarvi, M., 2017. Sensitivity-based linear approximation method to estimate time-dependent origin–destination demand in congested networks. Transp. Res. Rec. 2669 (1), 72–79.
Shafiei, S., Gu, Z., Saberi, M., 2018. Calibration and validation of a simulation-based dynamic traffic assignment model for a large-scale congested network. Simul. Model. Pract. Theory 86, 169–186.
Shang, Q., Tan, D., Gao, S., Feng, L., 2019. A hybrid method for traffic incident duration prediction using BOA-optimized random forest combined with neighborhood components analysis. J. Adv. Transp.

Simon, X.Z., Cheng, Q., Wu, X., Li, P., Belezamo, B., Lu, J., Abbasi, M., 2022. A meso-to-macro cross-resolution performance approach for connecting polynomial arrival queue model to volume-delay function with inflow demand-to-capacity ratio. Multimodal Transport. 1 (2), 100017.
Søndergaard, J., 2003. Optimization using surrogate models-by the Space Mapping technique.
Spall, J.C., 2005. Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control. John Wiley & Sons.
Srinivas, N., Krause, A., Kakade, S.M., Seeger, M., 2009. Gaussian process optimization in the bandit setting: no regret and experimental design. arXiv preprint arXiv: 0912.3995.
Tang, J., Zheng, L., Han, C., Liu, F., Cai, J., 2020. Traffic incident clearance time prediction and influencing factor analysis using extreme gradient boosting model. J. Adv. Transp.
Tay, T., Osorio, C., 2022. Bayesian optimization techniques for high-dimensional simulation-based transportation problems. Transp. Res. B: Methodol. 164, 210–243.
Toledo, T., Kolechkina, T., 2012. Estimation of dynamic origin–destination matrices using linear assignment matrix approximations. IEEE Trans. Intell. Transp. Syst. 14 (2), 618–626.
Tympakianaki, A., Koutsopoulos, H.N., Jenelius, E., 2015. c-SPSA: cluster-wise simultaneous perturbation stochastic approximation algorithm and its application to dynamic origin–destination matrix estimation. Transport. Res. Part C: Emerg. Technol. 55, 231–245.
Tympakianaki, A., Koutsopoulos, H.N., Jenelius, E., 2018. Robust SPSA algorithms for dynamic OD matrix estimation. Proc. Comput. Sci. 130, 57–64.
Wang, W., Wan, H., Chang, K.-H., 2016. Randomized block coordinate descendant STRONG for large-scale stochastic optimization. In: 2016 Winter Simulation Conference (WSC). IEEE, pp. 614-625.
Wu, W., Liu, R., Jin, W., Ma, C., 2019. Simulation-based robust optimization of limited-stop bus service with vehicle overtaking and dynamics: A response surface methodology. Transport. Res. Part E: Logist. Transport. Rev. 130, 61–81.
Yang, H., Sasaki, T., Iida, Y., Asakura, Y., 1992. Estimation of origin-destination matrices from link traffic counts on congested networks. Transp. Res. B: Methodol. 26 (6), 417–434.
Yi, H., Bui, K.-H.-N., 2020. An automated hyperparameter search-based deep learning model for highway traffic prediction. IEEE Trans. Intell. Transp. Syst. 22 (9), 5486–5495.
Yin, R., Liu, X., Zheng, N., Liu, Z., 2022a. Simulation-based analysis of second-best multimodal network capacity. Transport. Res. Part C: Emerg. Technol. 145, 103925.
Yin, R., Liu, Z., Zheng, N., 2022b. A simulation-based model for continuous network design problem using Bayesian optimization. IEEE Trans. Intell. Transp. Syst. 23 (11), 20352–20367.
Zhang, H.M., Nie, Y., Qian, Z., 2008. Estimating time-dependent freeway origin-destination demands with different data coverage sensitivity analysis. Transp. Res. Record (2047), 91–99.
Zhang, H., Seshadri, R., Prakash, A.A., Antoniou, C., Pereira, F.C., Ben-Akiva, M., 2021. Improving the accuracy and efficiency of online calibration for simulation-based Dynamic Traffic Assignment. Transport. Res. Part C: Emerg. Technol. 128, 103195.
Zhao, D., Balusu, S.K., Sheela, P.V., Li, X., Pinjari, A.R., Eluru, N., 2020. Weight-categorized truck flow estimation: a data-fusion approach and a Florida case study. Transport. Res. Part E: Logist. Transport. Rev. 136, 101890.
Zheng, L., Xue, X., Xu, C., Ran, B., 2019. A stochastic simulation-based optimization method for equitable and efficient network-wide signal timing under uncertainties. Transp. Res. B: Methodol. 122, 287–308.
Zheng, L., Liu, P., Huang, H., Ran, B., He, Z., 2022. Time-of-day pricing for toll roads under traffic demand uncertainties: a distributionally robust simulation-based optimization method. Transport. Res. Part C: Emerg. Technol. 144, 103894.
Zhou, X., Mahmassani, H.S., 2006. Dynamic origin-destination demand estimation using automatic vehicle identification data. IEEE Trans. Intell. Transp. Syst. 7 (1), 105–114.
Zhou, X., Qin, X., Mahmassani, H.S., 2003. Dynamic origin-destination demand estimation with multiday link traffic counts for planning applications. Transp. Res. Rec. 1831 (1), 30–38.
Zhu, J., Tasic, I., Qu, X., 2022. Flow-level coordination of connected and autonomous vehicles in multilane freeway ramp merging areas. Multimodal Transport. 1 (1), 100005.