COMPUTER-AIDED CIVIL AND INFRASTRUCTURE ENGINEERING

Computer-Aided Civil and Infrastructure Engineering **29** (2014) 359–381

Surrogate-Based Optimization of Expensive-to-Evaluate Objective for Optimal Highway Toll Charges in Transportation Network

Xiqun (Michael) Chen, Lei Zhang*, Xiang He & Chenfeng Xiong

Department of Civil and Environmental Engineering, University of Maryland, College Park, MD, USA

&

Zhiheng Li

Department of Automation, Tsinghua University, Beijing, China 100084
Graduate School at Shenzhen, Tsinghua University, Shenzhen, China 518055

Abstract: *This article adopts a family of surrogate-based optimization approaches to approximate the response surface for the transportation simulation input–output mapping and search for the optimal toll charges in a transportation network. The computational effort can thus be significantly reduced for the expensive-to-evaluate optimization problem. Meanwhile, the random noise that always occurs through simulations can be addressed by this family of approaches. Both one-stage and two-stage surrogate models are tested and compared. A suboptimal exploration strategy and a global exploration strategy are incorporated and validated. A simulation-based dynamic traffic assignment model DynusT (Dynamic Urban Systems in Transportation) is utilized to evaluate the system performance in response to different link-additive toll schemes implemented on a highway in a real road transportation network. With the objective of minimizing the network-wide average travel time, the simulation results show that implementing the optimal toll predicted by the surrogate model can benefit society in multiple ways. The travelers gain from the 2.5% reduction (0.45 minutes) of the average travel time. The total reduction in the time cost during the extended peak hours would be around US$65,000 for all the 570,000 network users assuming a US$15 per hour value of time. Meanwhile, the government benefits from the 20% increase of toll revenue compared to the current situation. Thus, applying the optimized pricing scheme in real world can be an encouraging policy option to enhance the performance of the transportation system in the study region.*

1 INTRODUCTION

Travel demand management (TDM) has been recognized as an effective way to enhance the transportation system performance in the 21st century. Among all optional strategies for managing travel demand, pricing measures such as congestion pricing, high occupancy/toll (HOT), and tolling facilities were frequently discussed in previous research (Szeto et al., 2010; Ekstrom et al., 2012). There's a consensus that these measures are capable of promoting travel behavior changes as well as improving the transportation system performance.

Consider a highway corridor divided into k consecutive links by a number of entrances and exits. Both endpoints of each link are connected with other highways or urban roads. Various, time-varying toll rates are charged for vehicles at each entry or exit and determined on the basis of each link length, capacity, and expected peak and off-peak hour volumes. Such a highway corridor pricing system is called a link-additive highway toll charge system. Public highway aims to optimize toll rates to achieve total/average travel time minimization under approximate fixed demands or social welfare maximization given elastic demand functions,

*To whom correspondence should be addressed. E-mail: *lei@umd.edu*.

DOI: 10.1111/mice.12058

whereas private road owners aim to maximize total revenue by adjusting toll rates (Unnikrishnan et al., 2009). Many recent advanced achievements of intelligent transportation systems' relevant techniques were applied in the highway management and traffic delay optimization models, such as neural networks, wavelets, and chaos theory (Ghosh-Dastidar and Adeli, 2003, 2006; Karim and Adeli, 2003a, b; Jiang and Adeli, 2003, 2004a, b; Adeli and Karim, 2005; Adeli and Jiang, 2009; Sánchez et al., 2012; Gao and Zhang, 2013).

Starting from the pricing studies based on the well-known marginal cost pricing principle in the late 1960s (Beckmann, 1965), various alternative highway toll charge problems have been studied. Optimal toll charges could be obtained under deterministic user equilibrium (e.g. Yang and Huang, 1998) and stochastic user equilibrium (Yang, 1999). More equitable toll charge problems with multiple user classes were investigated by Yang and Zhang (2002) and Yang and Huang (2004). The assumptions of fixed demand (Yang and Lam, 1996), elastic demand (Yang and Bell, 1997), and link-flow restrictions have also been studied.

Dynamic network supply models are more advanced in that they can account for the capacity limit of road segments and operational improvements such as traffic signals. Meanwhile, dynamic network supply models can provide more detailed information regarding system performance than planning models (Ukkusuri et al., 2007; Nie et al., 2008). All these features make dynamic network supply models a promising approach to evaluate the transportation system performance under different pricing strategies.

In a transportation network (e.g., hundreds of traffic analysis zones [TAZs], thousands of nodes and links), random noise can not be ignored that refers to random fluctuations in the output of an experiment given the same input. For example, the percentage of completed trips of all origin-destination (OD) pairs in a specific period is usually influenced by the results of dynamic traffic assignment (DTA) (Han, 2003) and dynamic route choice behavior (Papola and Marzano, 2013). Due to the complex, dynamic, and stochastic characteristics of transportation networks, simulation models that aim to depict the time-dependent traffic states at the individual behavior level are generally time consuming with expensive-to-evaluate objectives. The computational time for one simulation run varies from less than 15 minutes to more than 60 minutes depending on the scale of the network. Examples include MATSim studies applied in Tel Aviv, Israel (Bekhor et al., 2011), Switzerland (Meister et al., 2010), and Toronto, Canada (Hatzopoulou et al., 2011), and the ongoing SimTRAVEL applications in Arizona and San Francisco County, California (Pendyala et al., 2010). Mesoscopic traffic simulation compromises between capturing detailed vehicle dynamics and the computational complexity by simulating vehicle movements at a speed determined by the local density, that is, car-following and lane-changing behaviors are not simulated (Adeli and Ghosh-Dastidar, 2004). Available mesoscopic simulators include Dynamic Urban Systems in Transportation (DynusT) (Chiu et al., 2010), DynaMIT (Ben-Akiva et al., 2002), DYNASMART (Mahmassani, 2002), Dynameq (Tian et al., 2007), etc. In this research, both microscopic and mesoscopic simulators have been tested for a real world application of a regional network in Maryland. The network is medium sized with 201 TAZs, 1,077 nodes, and 2,158 links (more detailed information can be found in Section 5). Compared to microscopic traffic simulation tools, mesoscopic approach is selected for its superior computational efficiency (less than 5 minutes for one iteration of DTA with a 2.66 GHz-quad CPU and 4 GB-Ram computer). In this sense, the sampling procedure consumes manageable simulation run time.

Travel behaviors such as departure time choice, model choice, pretrip route choice, and en-route route choice, may influence the simulation process and output. Besides, the road network simulation that embeds random seeds may cause the outputs to perform differently in separate runs even given the same input. In field applications (López and Monzón, 2010), transportation infrastructure decision makers face the problem of exploration in a high-dimensional decision variable space to find the best strategies within an affordable computational budget. Such problems with expensive-to-evaluate objectives are usually nonderivative even with implicit expression in terms of decision variables.

In the literature, transportation engineering optimization problems are characterized by computationally expensive objective functions, high-dimensional decision variables, and stochastic simulation experiments. The DTA model calibration of OD flows and all other simulation parameters was formulated as a large-scale iterative simulation-based optimization problem, which could be solved with several alternative approaches, such as Bayesian method, SNOBFIT, Box-Complex, SPSA, FDSA (Vaze et al., 2009; Flötteröd et al., 2011; Sundaram et al., 2011; Omrani and Kattan, 2012). The performance of SPSA for the calibration of large-scale traffic simulation models had also been demonstrated. For example, Balakrishna et al. (2007a) adapted the systematic traffic simulation model calibration methodology for the simultaneous calibration of all demand and supply models within a microscopic traffic simulation model using aggregate, time-varying traffic measurements; Balakrishna et al. (2007b) presented a systematic offline DTA calibration methodology that estimated all demand-and-supply inputs and parameters simultaneously. Other pioneering efforts were made to address the optimal design problem of selecting a charging cordon in a general traffic network using Genetic Algorithm (GA) (Sumalee, 2004), to develop a decision support tool for mitigating traffic congestion (Melouk et al., 2010), to optimize regional signal timing strategies using surrogate modeling (Osorio, 2010), to optimize coordinated, area-wide traffic signal control-considering drivers' rerouting behaviors using a meta-heuristic model (Teklu et al., 2007), and to develop a heuristic approach for systemwide highway project selection based on total benefit maximization under budget uncertainty (Li et al., 2010). As simulations take account of the interactions between complex travel demand patterns and network supply, it is always time consuming to evaluate the system performance. Thus, a simulation-based optimization method that requires fewer objective function evaluations will be adopted for the current study.

A comprehensive review of simulation optimization approaches for both continuous and discrete variable simulation optimization can be found in (Fu, 2002). Specifically, a surrogate-based optimization approach is one feasible alternative to solve continuous decision variable optimization problems with computationally costly objective functions. Surrogate modeling or metamodel-based simulation optimization aims to regress the response surface that characterizes the relationship between decision variable inputs and simulation outputs (Hussain et al., 2002; Queipo et al., 2005; Jakobsson et al., 2010). The surrogate simplifies simulation optimization because of its deterministic rather than stochastic relationship between the input and output (Barton and Meckesheimer, 2006). Using only an initial input data set and corresponding output values of the objective function, the response surface method (RSM) can be developed as an approximation of the expensive-to-evaluate objective. In the surrogate-based optimization approach, an unknown function that formulates the relationship between simulation input and output is approximated with a predefined parametric function whose coefficients can be determined via the experiment design. The most fundamental RSMs are linear and quadratic polynomial regression (Montgomery, 2008). On the basis of exploration and exploitation of the computationally efficient surrogate, the optimal decision variable set can be obtained.

However, the expediently implemented low-order polynomials may be heavily biased when applied in complex functions of high nonlinearity. More advanced radial basis functions (RBFs) (Björkman and Holmström, 2000; Gutmann, 2001; Regis and Shoemaker, 2005; Zhou et al., 2013) and Kriging models are capable of providing good predictions for the complex response surface. A radial basis function neural network (RBFNN) learns input–output mapping by covering the input space with basis functions that transform a vector from the input space to the output space (Adeli, 1994; Adeli and Karim, 2000; Adeli, 2001; Karim and Adeli, 2002, 2003c; Adeli and Jiang, 2003; Dharia and Adeli, 2003; Ghosh-Dastidar et al., 2008; Adeli and Jiang, 2009). A support vector regression (SVR) surrogate model provides a good compromise between prediction accuracy and robustness of other approximations (Smola and Schölkopf, 2004; Wandekokem et al., 2011; Li et al., 2013).

Forrester et al. (2006) pointed out that one smooth continuous approximation function (e.g., Gaussian RBF and ordinary Kriging) is unable to fit the discontinuous simulation output due to random noises around the true average response value. The optimization accuracy relies on how accurate the surrogate models are in capturing the performance variations.

The global optimization of an expensive-to-evaluate problem based on simulation is a great challenge. Jones et al. (1998) proposed the efficient global optimization based on Kriging basis functions, and applied the expected improvement of the surrogate to select new points. To handle noisy objective functions, Huang et al. (2006) provided the Sequential Kriging Optimization (SKO) as an extension of the efficient global optimization algorithm. Villemonteix et al. (2009) proposed the Informational Approach to Global Optimization (IAGO) that selects the infill point based on the entropy minimization. Jakobsson et al. (2010) proposed an RBF-based surrogate model for global optimization of expensive and noisy black-box functions, whereas updating infill points minimize the total model uncertainty weighting. More detailed discussions on the exploration and exploitation process can be found in Jones (2001), Forrester and Keane (2009), and Kleijnen (2009).

Inspired by the work by Forrester et al. (2008) in which common surrogate models are formulated and compared, this article targets optimizing the link-additive highway toll charge problem in a transportation network with expensive-to-evaluate objectives and random simulation noise using surrogate-based optimization. To deal with the noise in simulation outputs, we will use quadratic polynomial RSM, Gaussian RBF, and the regressing Kriging method to construct surrogate models for predicting the optima of expensive-to-evaluate functions.

In this article, we focus on the decision making of a public highway operator (such as the government or public agency) to achieve the network travelers' average travel time minimization. We assume that most commuting demands do not change departure time nor are they cancelled due to toll charges. Therefore, fixed commuting demands are assumed in extended morning peak hours for a road network that covers central and eastern Montgomery County and northwestern Prince George's County in the State of Maryland. An open source simulator DynusT is chosen as the DTA and mesoscopic vehicle simulation tool to evaluate network performance given various link-additive highway pricing rates. The computation time needed to obtain a solution from a black-box function can be considerably reduced by surrogate-based optimization models, several of which include global optima strategies that make noisy data processing and computation intensive global optimization feasible.

The rest of the article is organized as follows: Section 2 proposes the framework of surrogate-based optimization for the highway toll charge problem in the context of methodology. Section 3 presents four categories of fundamental surrogate models considering both the noise free and noisy simulation outputs. Then, suboptimal and global optimal infill strategies are introduced for each surrogate model. To test and compare surrogate-based optimization models, a toll optimization problem that is formulated by the bilevel equilibrium assignment is analyzed in Section 4. Section 5 investigates the field application of surrogate models in a link-additive toll highway in a transportation network followed by the optimization results discussion and sensitivity analysis. Finally, Section 6 concludes the article and proposes future research directions.

2 METHODOLOGY

This section presents the methodology of transportation network optimization using simulations. It is organized in two parts. First, a framework for the surrogate-based optimization using transportation simulation is developed. Second, the optimization problem that aims to minimize the mean travel time of all travelers is established. This section highlights how the methods provided in Section 3 are going to be used in the transportation network objective function approximation.

2.1 Framework

A framework of surrogate-based optimization procedures using transportation simulations is illustrated in Figure 1.

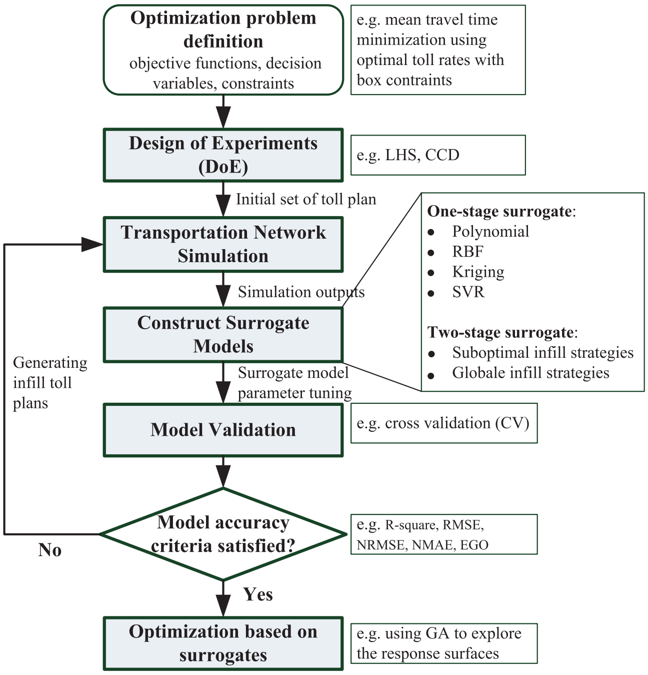

Optimization problem definition
objective functions, decision variables, constraints
e.g. mean travel time minimization using optimal toll rates with box contraints
Design of Experiments (DoE)
e.g. LHS, CCD
Initial set of toll plan
Transportation Network Simulation
One-stage surrogate:
Polynomial
RBF
Kriging
SVR
Two-stage surrogate:
Suboptimal infill strategies
Globale infill strategies
Simulation outputs
Construct Surrogate Models
Generating infill toll plans
Surrogate model parameter tuning
Model Validation
e.g. cross validation (CV)
Model accuracy criteria satisfied?
No
e.g. R-square, RMSE, NRMSE, NMAE, EGO
Yes
Optimization based on surrogates
e.g. using GA to explore the response surfaces

Fig. 1. Transportation simulation-based optimization procedure using the surrogate methodology. LHS, Latin hypercube sampling; CCD, central composite design; RBF, radial basis function; SVR, support vector regression; RMSE, root mean square error; NRMSE, normalized root mean squared error; NMAE, normalized maximum absolute error; EGO, estimated global optimum; GA, genetic algorithm.

The first step is to define the optimization problem. Given the objective functions that can be evaluated by simulation outputs, we generate an initial number of toll charges through a design of experiments (DoE), such as Latin hypercube sampling (LHS); see Appendix A.

The second step is to select a transportation simulator for the toll charge optimization problem. Run the initial set of toll charge plans for the subject highway using the chosen simulator. Based on the simulation outputs, we evaluate the objective functions. Considering the generally large number of DoE cases (that aim to obtain a more accurate surrogate), the analysis could be computationally expensive. Simulation outputs of a transportation network usually involve random noises, so it is better to run each sampling point for several repetitions and estimate the mean value of objective function evaluations, if the computational burden could be affordable.

The third step is to construct a response surface using surrogate models. In this article, we first adopt one-stage surrogate models, that is, quadratic polynomial function, Gaussian RBF, ordinary Kriging, and SVR. To find new toll plans based on the initial samples, we consider two-stage surrogate models by infilling points to the initial set using criteria such as the probability of improvement and the expectation of improvement across the response surface. Due to simulation random errors, we may also incorporate the surrogate models that are capable of dealing with noises. Once all of the cases are analyzed, proper parameters of the surrogate models can be determined. This step is regarded as the most important component in the whole framework, so we will interpret the surrogate models that are noise free and noisy, as well as suboptimal and global infill strategies in Section 3.

The fourth step is to assess and validate the assumed surrogate models by comparing an additional test set of objective function data with values estimated by the surrogates at points corresponding to the variables at which the independent objective function values are calculated. On the basis of the error observed with the validation data set, the accuracy of each surrogate model is checked using certain criteria such as correlation coefficient or coefficient of determination to determine whether the initially assumed surrogate model is appropriate; see Appendix B. If it is, the best surrogate model will be employed for the toll optimization problem to explore the optimal solutions. If the evidence shows that a certain surrogate model does not achieve the required predictive performance based on the current test data set, a proper way is to recall the two-stage surrogate models to generate infill points and run transportation simulations for new points until the accuracy criteria are reached.

Finally, we find the optimal values of the tolls using the optimized surrogate models. Though the estimated response surface may be too complex to explore its global optima using analytical techniques such as gradient descent method, we can still apply a heuristic approach, for example, GA, to seek the global optima for the estimated response surface. The computational costs of this tuning process can be neglected compared to the burden of transportation simulations.

All key components in the framework are highlighted in shadow boxes in Figure 1. We will further explain the details of each component in Section 3.

2.2 Optimization problem

From a public agency's perspective, the objective would be to minimize the total social costs of the whole network or maximize total social welfare, whereas if a road is privately operated, maximizing total toll revenue may be the main objective. In this article, the model to minimize the average travel time given fixed OD demands can be formulated

$$\min_{\mathbf{x}\in\mathbb{R}^k} \quad \mathrm{E}[f_{\mathrm{TT}}(\mathbf{z},\ \tau)] \tag{1a}$$

$$= \mathrm{E}\left[\frac{\sum_{r\in R}\sum_{s\in S}\left(d^{rs}_{\mathrm{peak}}\cdot \mathrm{TT}^{rs}_{\mathrm{peak}}(\mathbf{z}) + d^{rs}_{\mathrm{off}}\cdot \mathrm{TT}^{rs}_{\mathrm{off}}(\mathbf{z},\ \tau)\right)}{\sum_{r\in R}\sum_{s\in S}(d^{rs}_{\mathrm{peak}} + d^{rs}_{\mathrm{off}})}\right]$$

$$\text{s.t.}\quad \mathbf{z}_{\min} \le \mathbf{z} \le \mathbf{z}_{\max}, \quad \mathbf{z}\in\mathbb{R}^{k-1} \tag{1b}$$

$$0 \le \tau \le 1 \tag{1c}$$

$$\mathbf{x}^{\mathrm{T}} = [\mathbf{z}^{\mathrm{T}},\ \tau], \quad \mathbf{x}\in\mathbb{R}^k \tag{1d}$$

where $f_{\mathrm{TT}}(\mathbf{z},\ \tau)$ is the stochastic average travel time function of the network. We minimize its expectation given the same input $\mathbf{z}$ and τ to eliminate random simulation errors. The decision vector $\mathbf{x}$ includes toll rates $\mathbf{z}$ of each toll road segment and the ratio τ of off-peak-hour toll rates to the peak-hour values, so $\mathbf{x}$ is a k-dimensional decision variable vector; the origin and destination sets are $r \in R$ and $s \in S$, $(r,\ s)$ is the OD pair; $\mathrm{TT}^{rs}_{\mathrm{peak}}$ is the average travel time for trips during peak hours corresponding to the OD pair $(r,\ s)$; $\mathrm{TT}^{rs}_{\mathrm{off}}$ is the average travel time of the OD pair $(r,\ s)$ in off-peak hours; d^{rs}_{peak} and d^{rs}_{off} are the peak and off-peak demands of the OD pair $(r,\ s)$, respectively; the toll rate ratio is $0 \le \tau \le 1$; the box constraints are considered in this model, that is, $\mathbf{z}_{\min}$ and $\mathbf{z}_{\max}$, which are lower and upper boundaries for segment toll rates, respectively.

Since no *a priori* information of simulation random errors is available, such random deviations from the expected smooth response can be simplified as uniformly distributed across the feasible domain, that is, $\mathrm{Var}[f_{\mathrm{TT}}(\mathbf{x})] = \sigma^2_{\mathrm{noise}},\ \forall(\mathbf{z},\ \tau)$, which is independently distributed with the regression error variance $\hat{s}^2(\mathbf{x})$. Then the estimated response variance $\hat{s}^2_{\mathrm{TT}}(\mathbf{x})$ at $\mathbf{x}$ is given by

$$\hat{s}^2_{\mathrm{TT}}(\mathbf{x}) = \hat{s}^2(\mathbf{x}) + \hat{\sigma}^2_{\mathrm{noise}} \tag{2}$$

where $\hat{\sigma}^2_{\mathrm{noise}}$ is the estimate of the simulation noise variance.

3 OPTIMIZATION PROCEDURES

This section explains the technical steps that are necessary to apply the surrogate models to the highway toll optimization problem using transportation simulations. First, Section 3.1 clarifies how to construct one-stage surrogates using the mean value and estimation errors at an arbitrary point. Second, we explain both the suboptimal and global optimal infill strategies for two-stage surrogate models in Section 3.2. Then Section 3.3 compares the merits and scopes of applications of all surrogate models.

3.1 One-stage surrogate models

3.1.1 Quadratic polynomial function. Because transportation simulation involves multiple sources of uncertainties, for example, route choice behaviors, we suppose the scalar valued observations are noisy. Using the DoE generation approach in Appendix A, we have an initial set of sampling plan $\mathbf{X} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}]^{\mathrm{T}}$ and responses $\mathbf{y} = [y^{(1)}, y^{(2)}, \ldots, y^{(n)}]^{\mathrm{T}}$. The most commonly used RSM is the quadratic polynomial function given by

$$\hat{f}(\mathbf{x}) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i<j} \sum_{j} \beta_{ij} x_i x_j + \sum_{i=1}^{k} \beta_{ii} x_i^2, \qquad \mathbf{x} \in \mathbb{R}^k \tag{3}$$

where $\mathbf{x} = [x_1,\ x_2, \ldots,\ x_k]^{\mathrm{T}}$ is a k-dimensional point to be predicted, $\hat{f}(\mathbf{x})$ is the estimate of the real objective function $f(\mathbf{x})$, β_0 is the intercept, β_i is the linear coefficient, β_{ij} is the coefficient of interaction terms, β_{ii} is the quadratic coefficient.

3.1.2 RBF. Compared with lower order polynomial RSM, RBF surrogate models can obtain better approximations to true objective functions of high nonlinearity. RBF uses the basis function $\varphi(r)$ that only depends on the radial distance r between $\mathbf{x}$ and each sample point $\mathbf{x}^{(i)}$. It assumes that the correlation of two arbitrary sample points depends only on the distance (e.g., Euclid distance) in the decision variable space. We seek a RBF approximation to $\hat{f}$ in the form

$$\hat{f}(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\varphi} = \sum_{i=1}^{n} w_i \varphi(||\mathbf{x} - \mathbf{x}^{(i)}||) \tag{4}$$

where $\mathbf{w}$ is the weighted coefficients of RBF vector $\boldsymbol{\varphi}$, and $||\mathbf{x} - \mathbf{x}^{(i)}||$ is the Euclidean norm.

A Gaussian basis function is used in this article, that is,

$$\varphi(||\mathbf{x} - \mathbf{x}^{(i)}||) = \exp\left(-c||\mathbf{x} - \mathbf{x}^{(i)}||^2\right), \qquad c > 0 \tag{5}$$

where c is the shape parameters that can be determined by tuning the minimization of a cross-validation (CV) error estimate in the optimization step. It is worthy to note that c can be various if we define different radial basis functions (RBF network), whereas normalization of input variables is not necessary for the Gaussian basis function because there are weight parameters for each function, so a universe c will be used for the Gaussian RBF in this article.

The prediction at a new point is given by

$$\hat{f}(\mathbf{x}) = (\boldsymbol{\Phi}^{-1} y)^{\mathrm{T}} \boldsymbol{\varphi} \tag{6}$$

where $\boldsymbol{\Phi}$ denotes the so-called Gram matrix, each element of which is defined as $\boldsymbol{\Phi}_{i,j} = \varphi(||\mathbf{x}^{(j)} - \mathbf{x}^{(i)}||)$.

The prediction error at any $\mathbf{x}$ in the design space is given by (Gibbs, 1997)

$$\hat{s}^2(\mathbf{x}) = 1 - \boldsymbol{\varphi}^{\mathrm{T}} \boldsymbol{\Phi}^{-1} \boldsymbol{\varphi} \tag{7}$$

3.1.3 Ordinary Kriging methods: noise free and noise. The Kriging method predicts a response by summarizing a linear model and a high-frequency variation component that represents fluctuations around the trend. In this study, we will consider the ordinary Kriging model

$$f(\mathbf{x}) = \mu + \varepsilon(\mathbf{x}), \qquad \mathrm{E}[\varepsilon] = 0 \tag{8}$$

where μ is the mean of the objective function, and ε is the estimation error with a covariance of $\mathrm{Cov}[\varepsilon(\mathbf{x}^{(i)}),\ \varepsilon(\mathbf{x}^{(j)})] = \sigma^2 \psi(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$, σ^2 is the variance, $\psi(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ is the Kriging basis function with the following correlation form

$$\psi(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \exp\left(-\sum_{l=1}^{k} \theta_l (x_l^{(i)} - x_l^{(j)})^2\right) \tag{9}$$

where $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_k]^{\mathrm{T}}$ is a vector of scaling coefficients that allow different widths of the basis function for each dimension of the k-dimensional $\mathbf{x}$ decision variable. The element of correlation matrix based on all the observed data is $\boldsymbol{\Psi}_{i,j} = \psi(\mathbf{x}^{(i)} - \mathbf{x}^{(j)})$.

The maximum likelihood estimations of the mean value and variance at $\mathbf{x}$ based on Gaussian processes (Rasmussen and Williams, 2006) are given by

$$\hat{f}(\mathbf{x}) = \hat{\mu} + \boldsymbol{\psi}^{\mathrm{T}} \boldsymbol{\Psi}^{-1} (\mathbf{y} - \mathbf{1}\hat{\mu}) \tag{10}$$

$$\hat{s}^2(\mathbf{x}) = \hat{\sigma}^2 \left(1 - \boldsymbol{\psi}^{\mathrm{T}} \boldsymbol{\Psi}^{-1} \boldsymbol{\psi} + \frac{1 - \mathbf{1}^{\mathrm{T}} \boldsymbol{\Psi}^{-1} \boldsymbol{\psi}}{\mathbf{1}^{\mathrm{T}} \boldsymbol{\Psi}^{-1} \mathbf{1}}\right) \tag{11}$$

The aforementioned ordinary Kriging model is noise free. So the predictions at sampled points are exactly the same as observations, which may be biased when simulation noise is taken into account. As a consequence, the surrogate response surface may perform overfitting features because the estimated response surface needs to pass all sampled points in the ordinary Kriging model. A model which has been overfit will generally have poor predictive performance, as it can exaggerate minor fluctuations in the data. To get rid of this problem, a regularization constant is added into the correlation matrix to filter noise. The regressing Kriging model can be used by adding the error estimation of the observed data to the diagonal of the correlation matrix, so that the new matrix is $\tilde{\boldsymbol{\Psi}} = \boldsymbol{\Psi} + \lambda \mathbf{I}$, where λ is the regulation constant, and $\mathbf{I}$ is identity matrix. λ can be estimated by the ratio between the simulation noise variance and the estimation variance given $\mathbf{X}$ and $\mathbf{y}$, that is, $\lambda = \sigma_{\mathrm{noise}}^2 / \sigma^2$. So the optima given by the regressing Kriging model with noise can be more robust than the noise free model.

3.1.4 SVR. SVR is one of the most important applications of support vector machine (SVM). An overview of the basic ideas underlying SVR for regression and function estimation has been given in Smola and Schölkopf

(2004). The key attribute of SVR is that it specifies and calculates the so called ε-margin within which the sample data errors are accepted without impacts on the surrogate prediction. The prediction is determined entirely by the support vectors that lie on or outside the ε-margin (Forrester et al., 2008). In the ε-SVR, the goal is to find a surrogate that has the least ε deviation from the observations for all the training data set, and at the same time is as flat as possible. Although SVR is powerful for prediction given large, high-dimensional data sets, enough samples used in training the SVR model are usually not available for expensive-to-evaluate objective functions in transportation networks. In such situation, we wish all points as the support vectors. However, we will apply SVR as an alternative surrogate-based optimization method in this study and compare it with other aforementioned surrogate models.

3.2 Two-stage surrogate models: infill approaches

To enhance the accuracy of surrogate models based on initial samples, it requires further objective function evaluations based on certain infill or update strategies. This section incorporates the suboptimal exploration strategy that induces local optimization and the global exploration strategy that is promising to locate the global optimum.

3.2.1 Suboptimal infill strategy. The local optima search strategy can be achieved by exploration over the surrogate surface estimated using the one-stage surrogate models. In this study, we use GA to explore $\hat{f}(\mathbf{x})$ and seek its global optima. The update point is given by

$$\mathbf{x}_{\text{update}} \in \underset{\mathbf{x}_{\min} \le \mathbf{x} \le \mathbf{x}_{\max}}{\arg\min} \ \hat{f}(\mathbf{x}) \qquad (12)$$

Then the simulation output at $\mathbf{x}_{\text{update}}$ will be evaluated by an extra simulation run, that is, y_{update}. The infill strategy will be terminated when the maximal number of simulation runs is reached or the Euclidian norm of two adjacent update points is smaller than a predefined tolerance.

3.2.2 Global optimal infill strategy. Global optimization can be classified into deterministic and stochastic methods. The former one generates a deterministic sequence of points converging to a globally optimal solution, thus transportation simulation-based optimization problem may not belong to deterministic category because various sources of uncertainties lead to stochastic simulation outputs, for example, random seeds in trip generation, probabilistic route choice behaviors of travelers, and DTA. The latter one randomly generates feasible updating points to infill the initial samples using a number of heuristic algorithms for the optimal parameter tuning (Kim and Adeli, 2001; Sarma and Adeli, 2001; Baraldi et al., 2011; Chabuk et al., 2012; Song et al., 2013).

To obtain the global optima for expensive-to-evaluate functions, a series of two-stage procedures can be incorporated (the first stage is the same as DoE in one-stage models). The second stage conducts the exploitation process using estimated standard deviation information to select an infill sample with the maximum probability of improvement (PI) or expected improvement (EI), which are given by

$$\mathrm{PI}(\mathbf{x}) = \frac{1}{\sqrt{2\pi}\,\hat{s}(\mathbf{x})} \int_{-\infty}^{y_{\min}} \exp\left(-\frac{(u - \hat{f}(\mathbf{x}))^2}{2\hat{s}(\mathbf{x})^2}\right) \mathrm{d}u \qquad (13)$$

$$\mathrm{EI}(\mathbf{x}) = \frac{1}{\sqrt{2\pi}\,\hat{s}(\mathbf{x})} \int_{-\infty}^{y_{\min}} (y_{\min} - u) \exp\left(-\frac{(u - \hat{f}(\mathbf{x}))^2}{2\hat{s}(\mathbf{x})^2}\right) \mathrm{d}u \qquad (14)$$

where $\mathrm{PI}(\mathbf{x})$ and $\mathrm{EI}(\mathbf{x})$ are the PI and EI estimations at the point $\mathbf{x}$, and $y_{\min}$ denotes the smallest value of all outputs in the training data set.

3.3 Surrogate models comparison

Table 1 shows the twelve models we investigate to solve the toll optimization problem. Method 1 (M1) is the quadratic polynomial function-based surrogate, which only recalls the second-order polynomial functions with interaction terms. Methods 2 and 3 are both one-stage models that estimate the response surface only using the initial samples. The only difference between Methods 4 and 2 is that the support vector is only a part of the initial sample. Methods 5–8 are suboptimal two-stage approaches. M9 and M10 refine the global optimal infill strategy using PI and EI maximization, respectively. To deal with noisy data, M11 is the regressing Kriging model with noisy covariance matrix. The main difference between Kriging with and without noisy errors is that the estimated response surface would pass through the known points in M3, whereas M11 allows some bias to the known points to obtain a much smoother response surface. Finally, M12 is the infill Kriging with noise in a two-stage global optimization approach that updates and adds a new point which corresponds to the maximum EI. In summary, for the transportation network problem with expensive-to-evaluate objective functions, M11 and M12 that can handle simulation noise will outperform other methods. We will test and compare these methods using a small transportation network with additive toll links in Section 4 and then apply the selected methods to a real transportation network in the State of Maryland in Section 5.

Table 1
Characteristics of surrogate models

Abbr.	*Methods*	*One-stage surrogate*	*Infill surrogate*	*Global optimum*	*Simulation noise*
M1	Quadratic polynomial	√	×	×	×
M2	Gaussian RBF	√	×	×	×
M3	Kriging	√	×	×	×
M4	SVR	√	×	×	×
M5	Suboptimal updating quadratic polynomial	×	√	×	×
M6	Suboptimal updating Gaussian RBF	×	√	×	×
M7	Suboptimal updating Kriging	×	√	×	×
M8	Suboptimal updating SVR	×	√	×	×
M9	Probability of improvement infill Kriging	×	√	√	√
M10	Expected improvement infill Kriging	×	√	√	√
M11	Regressing Kriging	√	×	×	√
M12	Reinterpolating Kriging	×	√	√	√

Note: RBF, radial basis function; SVR, support vector regression; √ means a method has a property; and × means a method does not have a property.

4 NUMERICAL TEST

This section tests the surrogate models in Table 1 using a second-best social optima additive highway pricing with a fixed demand for a small network. The user equilibrium (UE) assignment is chosen because the true objective function can be exactly known through an analytical derivation, so we could validate the estimated response surfaces with the true response surface. Though UE and DTA are quite different in objective function evaluations, for example, mean travel time of all travelers, the input–output mapping could be estimated and validated through surrogate models. The UE-based toll charge problem is characterized with a deterministic and easy-to-evaluate input–output relation, whereas the DTA-based toll charge problem can be regarded as a stochastic and expensive-to-evaluate problem. The relationship between the numerical test and the Intercounty Connector (ICC) highway toll optimization case study in Section 5 is that we explore the features of different surrogate models using a small network and identify the most capable method that could be used to model the input–output mapping in a larger network.

The link-based pricing scheme is investigated as the second-best toll charging in a small road network, where tolls are charged only on a subset of selected links, which can be categorized as a mathematical program with equilibrium constraints (MPEC) (Yang and Huang, 2005). The second-best road pricing problem in this article is to choose a set of optimal toll charges to minimize the total travel time (or the average travel time due to fixed demand). The bilevel mathematical program with equilibrium constraints can be formulated. The upper-level model is

$$\min_{\mathbf{z}} \quad F(\mathbf{z},\ \mathbf{q}^*) = \sum_{a\in A} t_a(q_a^*)q_a^* \tag{15a}$$

$$\text{s.t.} \quad \mathbf{q}^* = \arg\min_{\mathbf{q}} \sum_{a\in A} \int_0^{q_a} c_a(q,\ \mathbf{z})\mathrm{d}q \tag{15b}$$

$$\mathbf{z}_{\min} \le \mathbf{z} \le \mathbf{z}_{\max} \tag{15c}$$

$$q_a = \sum_{r\in R}\sum_{s\in S}\sum_{p\in P_{rs}} f_p^{rs}\delta_{ap}^{rs}, \quad a \in A \tag{15d}$$

$$q_k = \sum_{r\in R}\sum_{s\in S}\sum_{p\in P_{rs}} f_p^{rs}\delta_{kp}^{rs}, \quad k \in K \tag{15e}$$

$$\sum_{p\in P_{rs}} f_p^{rs} = d_{rs}, \quad r \in R,\ s \in S \tag{15f}$$

$$f_p^{rs} \ge 0, \quad r \in R,\ s \in S,\ p \in P_{rs} \tag{15g}$$

where F is the total travel time function, $\mathbf{z} = [z_1, \ldots, z_k]^{\mathrm{T}}$ is the link toll vector, satisfying $k \in K$, $K \subseteq A$ is a subset of tolled links, A is the whole link set, $\mathbf{q}^* = [\ldots, q_a^*, \ldots]^{\mathrm{T}}$ is the equilibrium link flow vector, q_a^* is the equilibrium flow of link a, satisfying $a \in A$, t_a is the average travel time of link a, constraints 15b–15g are the conservation conditions, f_p^{rs} is the path flow of OD pair

Table 2
Input data and equilibrium flow for a small road network

Link	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
t_a^0	20	20	20	20	6	1	1	6
C_a	800	800	600	600	500	800	800	500
q_a^*	686	686	314	314	314	0	0	314
$t_a(q_a^*)$	21.6	21.6	20.2	20.2	6.1	1	1	6.1

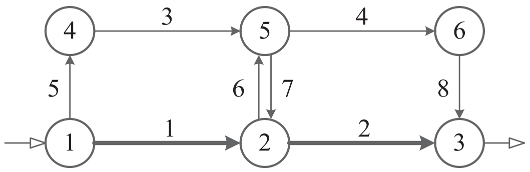


Fig. 2. Numerical network (links 1 and 2 are the tolled links).

$(r,\ s)$, δ_{ap}^{rs} is the 0–1 indicator, and R and S are origin and destination sets.

The well-known Frank–Wolfe method can be used to solve the lower level programming problem of the traffic equilibrium model (Ramadurai and Ukkusuri, 2011; Szeto et al., 2011; Aziz and Ukkusuri, 2012; Unnikrishnan and Lin, 2012). The solution of the bilevel programming problem can be obtained by using the gap function approach solved by the augmented Lagrangian algorithm (Yang el al., 2004; Meng and Wang, 2008).

Assume users are homogeneous with identical values of time, from the perspective of link-based costs, the generalized cost function c_a of link a can be expressed as follows

$$c_a(q_a,\ \mathbf{z}) = \begin{cases} z_k + \eta t_k(q_k) & a = k \in K \\ \eta t_a(q_a) & a \in A,\ a \notin K \end{cases} \tag{16}$$

where z_k and t_k are the toll charge and average travel time of link k, η is the value of time.

Consider a network depicted in Figure 2, consisting of six nodes and eight links. Links 1 and 2 are road segments subject to toll charge. The Bureau of Public Roads (BPR) link performance function is applied

$$t_a(q_a) = t_a^0 \left[1.0 + \alpha \left(\frac{q_a}{C_a} \right)^{\beta} \right], \quad a \in A \tag{17}$$

where t_a^0 is the free-flow travel time, C_a is the link capacity, parameters are $\alpha = 0.15$ and $\beta = 4$, see Table 2.

There is only one OD pair from node 1 to node 3 with a demand of $d = 1,000$ (flow units), and the value of time is $\eta = 1$. Four paths from node 1 to node 3 using links are: 1–2, 1–6–4–8, 5–3–4–8, 5–3–7–2. One of the optimal toll charges is $\mathbf{z}^* = [5.555,\ 4.045]^{\mathrm{T}}$. The minimized average travel time is $F_{\min} = 46.22$. Optimal path flows are $f_1^* = f_3^* = 686$ and $f_2^* = f_4^* = 0$. Figure 3 shows the true responses of the upper-level objective function $\hat{F}(\mathbf{z},\ \mathbf{q}^*)$ in terms of z_1 and z_2. It was an interpolation surface based on a uniform 20 × 20 grid. The simulation random errors are because the UE link flows may be not integers.

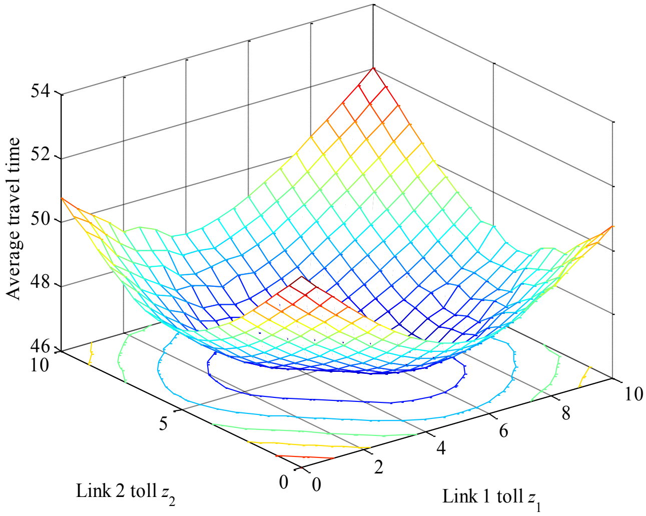


Fig. 3. The true response function of $F(\mathbf{z},\ \mathbf{q}^*)$.

To compare the surrogate-based optimization approaches shown in Table 1, we generate 10, 20, 30, and 40 initial LHS samples for the five one-stage surrogate models M1, M2, M3, M4, and M11, respectively. Then we generate other initial 8, 15, 25, and 35 LHS points for the two-stage models M5, M6, M7, M8, M9, M10, and M12, then add 2, 5, 5, and 5 infill points to the initial samples, respectively. Figure 4 shows the estimation errors by comparing the prediction values with true objective function for the uniform 20 × 20 grid. The Pearson correlation coefficient (PCC), root mean square error (RMSE), and maximum absolute error (MAE) values are plotted for each surrogate model and different sample size (see the statistical metrics listed in Appendix B). For most surrogate models except for M2 and M6, the larger the sample size is, the higher estimation accuracy of the surrogate model.

Table 3 shows the results of the 12 models in terms of five measures of effectiveness (MoEs) under the largest test sample size, that is, 40 evaluations of the

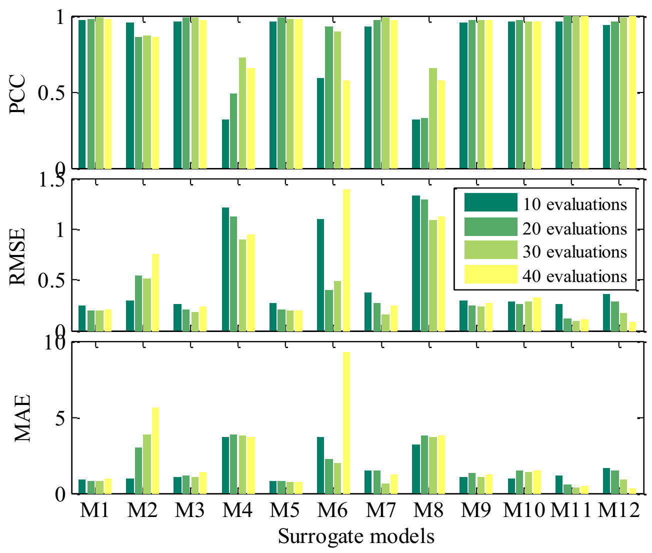


Fig. 4. Validation of surrogate models. MAE, maximum absolute error; RMSE, root mean square error; PCC, Pearson correlation coefficient.

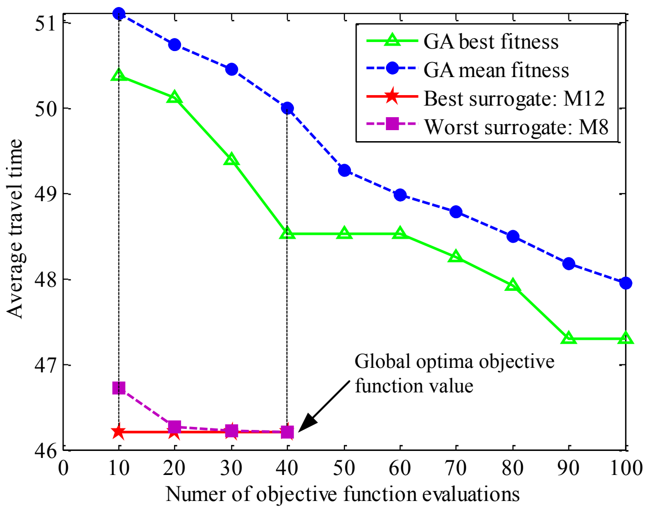


Fig. 5. An illustration of the higher computational efficiency of the surrogate models compared to genetic algorithm (GA).

objective function. The minimum values of each column are highlighted in bold. Results show that the best model with the smallest errors of RMSE, MAE, NRMSE, and NMAE is M11, and the second-best model is M12. The smallest EGO (estimated global optimum) = $\hat{F}(\hat{\mathbf{z}}^*, \mathbf{q}^*)$ is 44.33 obtained by M6, however, its true response is 48.68 that is larger than the estimated minimum. From other MoEs of M6, we can see its overall prediction accuracy is poor. In the last column, seven models finally converge to the global optima.

To demonstrate how surrogate models can intelligently mimic simulation-based objective function evaluation and reduce computational times. The novelty in this work is the computational time savings. We compare the convergence and efficiency of surrogate models with GA (the population size is 10, generations are 10, other parameters are default value given by MATLAB R2010a GA Toolbox). Figure 5 quantifies the computational savings obtained from this method using the number of objective function evaluations. We can see that the best surrogate model (M12) can find the global optima only using 10 evaluations.

In summary, the two best one-state surrogate models are M1 and M3 (M1 performs better mainly because the true response surface is less complex shown in Figure 3), and the two best two-stage surrogate models are M11 and M12. In the following sections, we will apply these four models in a transportation network.

Table 3
The estimation accuracy comparison of surrogate model, under 40 times of objective function evaluations

Method	*RMSE*	*MAE*	*NRMSE*	*NMAE*	*EGO*	$f(\hat{\mathbf{x}}^*)$
M1	0.21	0.95	0.44%	0.68	46.15	**46.22**
M2	0.76	5.62	1.57%	4.03	46.20	46.24
M3	0.24	1.44	0.50%	1.03	46.12	**46.22**
M4	0.94	3.67	1.95%	2.63	46.45	46.25
M5	0.19	0.71	0.40%	0.51	46.22	46.23
M6	1.30	7.21	2.69%	5.17	**44.33**	48.68
M7	0.24	1.20	0.50%	0.86	46.22	**46.22**
M8	1.13	3.76	2.34%	2.69	47.20	46.23
M9	0.36	1.41	0.75%	1.01	46.22	**46.22**
M10	0.23	1.09	0.48%	0.78	46.22	**46.22**
M11	**0.10**	**0.45**	**0.21%**	**0.32**	46.09	**46.22**
M12	0.16	0.57	0.33%	0.41	46.21	**46.22**

Note: RMSE, root mean square error; MAE, maximum absolute error; NRMSE, normalized root mean squared error; NMAE, normalized maximum absolute error; EGO, estimated global optimum.

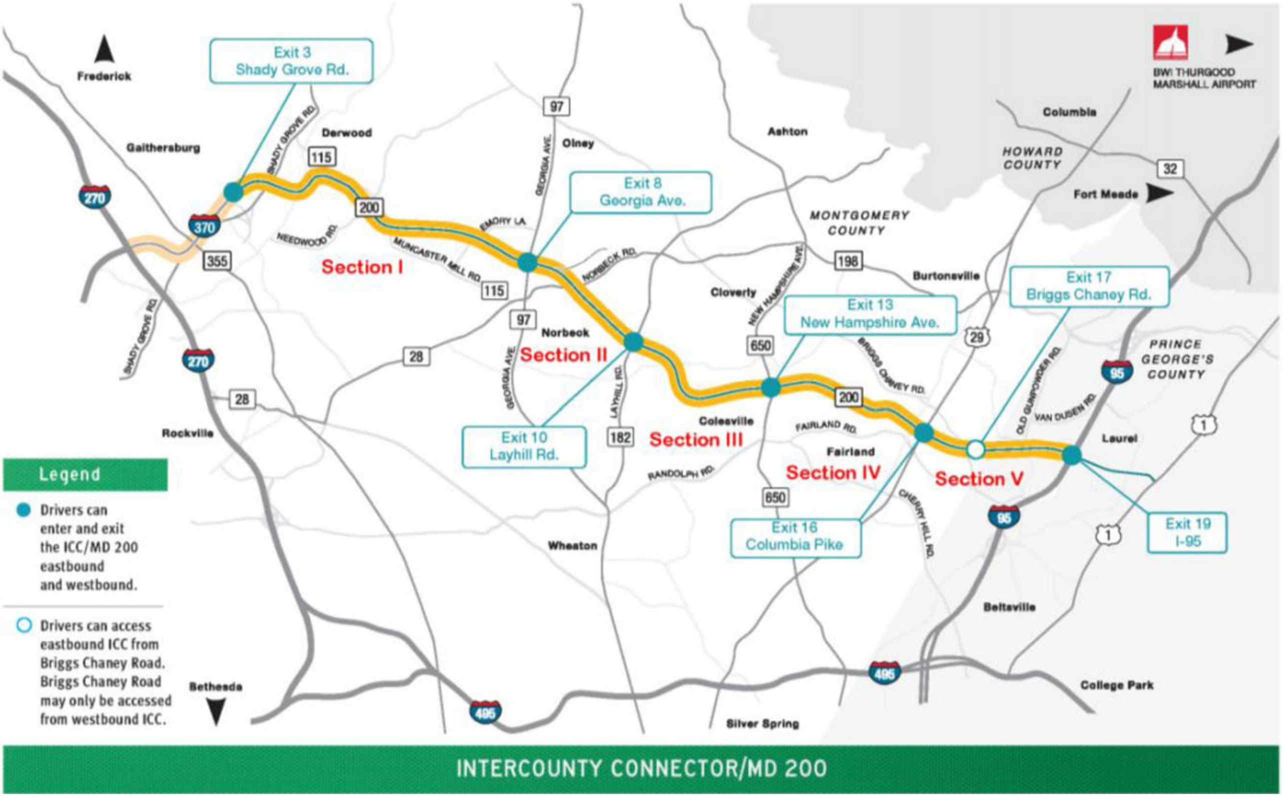


Fig. 6. The Intercounty Connector (ICC) (in thick line) and regional network. (*Source:* http://www.mdta.maryland.gov/ICC/Toll_Rates.html)

5 ICC TOLL CHARGES OPTIMIZATION

5.1 Road network

The ICC is probably the most significant and high-profile highway project in Maryland since the completion of the existing interstate freeway system several decades ago. It links existing and proposed development areas between the I-270/I-370 and I-95/US-1 corridors within central and eastern Montgomery County and northwestern Prince George's County. The simulation model covers the central and eastern Montgomery County and the northwestern Prince George's County of the State of Maryland. Before the construction of ICC, there was no freeway connecting the areas lying northwest and northeast of the capital beltway. The traffic between these two areas usually travels through I-495, which contributes to the severe congestion on I-495 during peak hours. The ICC was constructed aiming at promoting development of the surrounding areas as well as alleviating congestion on I-495. The ICC is a toll facility with different toll rates for its five segments, and the toll rate for each segment is variable along time. Vehicles with E-ZPass, an electronic toll transponder, are charged directly when they travel through ICC. If a vehicle without E-ZPass uses ICC, a US$1 video-processing fee is added to the total price and a bill sent to the vehicle registration address.

To test the effectiveness of applying a simulation-based optimization method to improve the transportation system performance, a case study on optimizing the toll scheme of the ICC in Maryland has been conducted. A simulation model for the regional network is developed to evaluate the system performance. All freeways and arterial roads within the region in Figure 6 are included in the transportation network, which is relatively large with 201 TAZs, 1,077 nodes, and 2,158 links. In our case study, actuated signal timings were coded into DynusT for all intersections that have signals applied in real world in the network. The simplified current pricing scheme for two-axle vehicles with E-ZPass during different time periods as well as the proposed limit of the toll rates is summarized in Table 4. The objective of this problem is minimizing the average travel time for all travelers in the entire network.

5.2 Open source DTA simulator: DynusT

DTA models fill the gap between static travel forecasting models and microscopic traffic simulation models,

Table 4
Selected design parameters and baseline values

Parameter (unit)	*Symbol*	*Baseline*	*Lower bound*	*Upper bound*
Toll charge in peak period, Segment I (US$) between I-370 and MD 97	z_1	1.45	0	3.00
Toll charge in peak period, Segment II (US$) between MD 97 and MD 182	z_2	0.60	0	1.50
Toll charge in peak period, Segment III (US$) between MD 182 and MD 650	z_3	0.75	0	1.50
Toll charge in peak period, Segment IV (US$) between MD 650 and US 29	z_4	0.65	0	1.50
Toll charge in peak period, Segment V (US$) between US 29 and I-95	z_5	0.70	0	1.50
Off-peak/peak toll charge ratio	τ	80%	0	100%

and enable modeling traffic dynamics at a relatively large scale within a reasonable amount of time. In the DTA framework, UE condition is only applied to travelers departing at the same time between the same OD pairs. Time-dependent shortest paths for travelers are computed based on time-varying link travel times when they arrive at the various links along a route.

DynusT is a simulation-based DTA model, which adopts the dynamic interactions between the network supply and user demand. DynusT performs well regarding its computational efficiency. However, it is essentially a route choice model. Some important aspects of travel demand analysis such as trip generation, mode choice, and departure time choice are not enabled in DynusT. Time-varying link travel time needed for DTA in DynusT is retrieved from the Anisotropic Mesoscopic Simulation (AMS) model (Chiu, et al., 2010), which is a vehicle-based mesoscopic traffic simulation approach that explicitly considers the anisotropic property of traffic flow in the vehicle state update at each simulation step. DynusT applies a gap function vehicle-based (GFV) solution algorithm to solve the DTA problem (Chiu and Bustillos, 2009). For each iteration and each OD-departure time combination, the number of vehicles to be updated with a new path is dependent on the relative gap function value, and vehicles with longer travel time are prioritized for path updating. Compared with the widely used successive average method, GFV can avoid over adjustments of flow and thus lead to more consistent and robust assignment results. Meanwhile, DynusT adopts a method of isochronal vehicle assignment which divides analysis periods into epochs and sequentially performs vehicle assignment in each epoch (Nava and Chiu, 2012). This significantly improves the model scalability regardless of the total analysis period. In the newly released 2012 version, DynusT has been fully parallelized in simulation, time-dependent shortest path, and assignment algorithms, and therefore boosts the computational speed dramatically. However, the current simulator does not address capacity drop due to congestion.

Although other models of both microscopic and mesoscopic traffic simulation are widely available (e.g., DynaMIT, DYNASMART, and Dynameq for mesoscopic models; TransModeler, VISSIM, and AIMSUN for microscopic models), and some of them may possess some desirable features, DynusT is selected in this study due to its advantage in computation time.

To simplify the optimization problem, the DynusT model only simulates travels during the extended morning peak (5–10 AM) including the entire morning peak (6–9 AM) and 2 hours during the off-peak periods. Meanwhile, off-peak pricing for all five segments is assumed to be a certain percentage of the peak pricing for the corresponding segments. In this way, there are six decision variables for the optimization problem, which are the peak pricing for each of the five segments and the rate of off-peak pricing to peak pricing.

5.3 Initial LHS training set

As shown in Table 5, we obtain the optimized sample plan (see Appendix A), that is, $\mathbf{X}$, including 64 initial LHS points plus three chosen inputs: the minimal toll plan $\mathbf{x} = \mathbf{x}_{\min} = \mathbf{0}$, the maximal toll plan $\mathbf{x} = \mathbf{x}_{\max}$, and the baseline inputs $\mathbf{x} = \mathbf{x}_{\text{baseline}}$, as the DoE for the toll charge strategy optimization. To achieve the convergence, 10 iterations of DTA and simulations were run for each toll plan, which generally takes 50 minutes, and the relative gaps for the DTA were found to be below 3% for all experiments. However, to save computation effort as much as possible, every toll plan was evaluated by the simulation-based DTA only once but including 10 iterations, despite the existence of noise in the route choice results. For each simulation run, the simulator obtains valid results when the convergence

Table 5
Space-filling Latin hypercube sampling of parameters for DoE

Sample	z_1	z_2	z_3	z_4	z_5	τ	**y**	*Sample*	z_1	z_2	z_3	z_4	z_5	τ	**y**
	US$	*US$*	*US$*	*US$*	*US$*	%	*Minutes*		*US$*	*US$*	*US$*	*US$*	*US$*	%	*Minutes*
1	1.24	1.29	0.07	1.31	0.95	0%	17.21	33	2.24	0.36	0.86	0.93	0.24	52%	18.47
2	2.57	0.76	1.10	0.83	0.81	92%	18.16	34	0.95	1.40	0.81	0.14	0.71	56%	19.23
3	0.90	0.98	0.95	1.43	0.10	71%	17.47	35	1.71	0.57	0.10	0.12	0.29	86%	18.38
4	2.52	0.38	1.29	1.05	1.26	70%	17.79	36	0.38	1.45	1.02	0.43	1.21	75%	17.52
5	2.67	0.95	1.48	0.55	1.43	60%	17.52	37	1.14	1.21	0.93	0.74	1.00	46%	18.52
6	1.81	1.24	0.24	0.71	0.93	33%	17.25	38	0.19	0.17	0.05	1.12	0.98	38%	17.65
7	0.43	1.07	0.43	0.76	0.31	29%	17.85	39	0.57	0.64	1.50	1.07	0.88	41%	17.62
8	1.48	0.00	1.26	0.05	0.90	35%	17.68	40	1.52	1.00	0.29	1.14	0.40	21%	17.65
9	1.86	0.40	1.31	0.38	0.43	59%	17.46	41	2.33	0.12	1.43	0.33	1.38	89%	16.77
10	1.76	0.86	0.83	0.50	0.69	30%	17.30	42	0.10	0.88	0.17	0.95	1.14	17%	17.99
11	2.48	1.17	0.57	0.36	0.07	63%	17.46	43	1.38	0.48	0.71	0.21	1.29	90%	18.49
12	0.86	0.24	0.19	0.48	0.05	32%	18.20	44	2.10	0.93	0.40	0.90	0.74	95%	18.37
13	0.52	0.55	0.88	0.45	1.24	54%	17.22	45	0.24	0.05	1.33	1.10	0.17	65%	18.10
14	0.62	1.38	1.38	0.79	1.12	16%	17.67	46	1.33	0.60	0.74	1.45	0.67	62%	17.34
15	0.71	1.48	1.21	0.67	0.50	67%	17.32	47	2.38	0.14	1.45	1.19	0.45	81%	17.12
16	3.00	1.10	0.36	0.31	0.62	24%	17.80	48	1.57	0.90	0.76	1.33	1.45	19%	18.16
17	0.05	0.02	0.14	0.02	0.38	10%	18.48	49	0.00	0.50	0.67	1.40	0.26	57%	18.53
18	2.29	0.07	1.14	1.24	0.57	25%	16.95	50	0.76	0.45	0.79	0.81	0.55	40%	17.56
19	2.71	0.81	0.60	0.62	1.50	87%	17.14	51	1.43	0.43	0.38	0.98	1.02	68%	17.29
20	1.19	0.71	0.50	0.00	1.19	37%	17.20	52	1.95	1.05	1.07	1.00	1.48	84%	18.20
21	2.43	1.43	1.12	0.40	0.76	3%	17.33	53	1.00	1.12	0.00	0.07	0.21	14%	18.30
22	1.29	0.79	0.26	0.29	0.14	49%	17.63	54	0.33	0.19	1.24	0.86	0.86	94%	17.38
23	0.48	1.33	1.19	1.50	1.36	51%	18.11	55	2.05	1.50	0.45	1.38	0.83	44%	17.43
24	0.29	0.21	1.05	0.60	0.33	2%	17.42	56	2.00	1.26	0.64	0.10	1.31	48%	17.42
25	0.81	1.02	0.12	0.57	0.48	100%	17.75	57	0.14	1.36	1.36	0.19	0.00	97%	18.23
26	2.86	0.52	1.17	0.17	0.79	22%	17.17	58	1.05	1.31	0.69	0.26	1.05	8%	18.38
27	2.81	0.74	0.52	0.88	0.12	78%	17.49	59	0.67	0.83	0.33	1.48	1.07	76%	17.20
28	2.62	0.26	0.55	1.17	1.40	43%	18.40	60	2.90	0.67	0.48	1.29	0.02	13%	17.00
29	2.76	0.33	0.98	1.02	1.10	11%	17.39	61	2.19	0.31	0.21	1.21	0.36	79%	17.11
30	1.67	0.62	1.00	0.52	1.33	6%	18.62	62	1.90	1.14	0.90	0.64	0.52	83%	17.57
31	1.10	0.10	0.02	1.26	0.60	5%	17.77	63	2.95	0.69	1.40	1.36	0.19	27%	16.98
32	1.62	1.19	0.31	0.24	0.64	73%	17.66	64	2.14	0.29	0.62	0.69	1.17	98%	17.76
Lower bound	0.00	0.00	0.00	0.00	0.00	0%	18.67	Upper bound	3.00	1.50	1.50	1.50	1.50	100%	18.05
Baseline	1.45	0.6	0.75	0.65	0.7	80%	18.15								

is achieved after several times of assignments and vehicular platoon simulations. The average output value is $\bar{y} = 17.74$ minutes, the minimum and maximum outputs are $y_{\min} = 16.77$ minutes and $y_{\max} = 19.23$ minutes, respectively.

To characterize baseline, we run 10 baseline repetitions that produce 17.96, 17.94, 17.71, 18.22, 18.30, 17.83, 18.04, 17.78, 19.31, and 18.44 minutes, respectively. The average baseline output is 18.15 minutes, and the standard deviation is 0.47 minutes, which could be an estimate of the simulation noise standard deviation, that is, $\hat{\sigma}_{\text{noise}} = 0.47$ minutes. To estimate the mean and standard deviation of stochastic simulation outputs of the optimal variables, we will also run 10 repetitions, then compare the estimated mean objective function value and standard deviation with the baseline to see how much improvement can be achieved after optimization.

5.4 Results

As the simulation of the ICC network costs about 50–60 minutes for each sample shown in Table 5, the surrogate models help reduce tremendous computation time compared to a traditional scenario study, which needs to evaluate all possible solutions through the entire feasible domain.

Using the four models we chose in Section 4, Table 6 shows the evaluation of response surfaces using the leave-one-out CV based on the first five MoEs (see Appendix B). We can see that the ordinary

Table 6
Leave-one-out cross-validation results

Methods	*RMSE*	*MAE*	*NRMSE*	*NMAE*	*EGO*	$\mathrm{E}[f_{\mathrm{TT}}(\hat{\mathbf{x}}^*)]$
M1	0.64	1.63	3.63%	3.18	**15.69**	–
M3	0.52	1.16	2.91%	2.26	17.14	–
M11	**0.52**	**1.41**	**2.92%**	**2.74**	17.51	–
M12	0.53	1.48	2.99%	2.90	17.36	**17.70**

Note: RMSE, root mean square error; MAE, maximum absolute error; NRMSE, normalized root mean squared error; NMAE, normalized maximum absolute error; EGO, estimated global optimum; dashes indicate not available.

Kriging (M3), regressing Kriging (M11), and the infill Kriging (M12) approaches produce smaller estimation errors than the quadratic polynomial RSM model (M1), which does not regress the response surface very well (larger RMSE, MAE, and NMAE) though its EGO is extremely small. However, the simulation outputs of the samples should be treated as random variables with simulation errors instead of an accurate number. In this case, the aforementioned five MoEs are not suitable for evaluating the performance of these response surface models any more. Thus, a new method analyzing the effectiveness of estimated confidence interval was proposed to evaluate model performance for the cases with significant simulation errors. The model was identified as the best model in approximating the real response surface in the ICC simulation, and was applied to search for the optimal toll rate in minimizing network average travel time.

Figure 7 applies the leave-one-out CV to predict the response of each sample point using other points. Four surrogate methods (M1, M3, M11, and M12) are compared. Figure 7a shows the estimated mean of average travel time values $\hat{\mathbf{f}}_{\mathrm{TT}}(\mathbf{x})$ at the initial 67 points as shown in Table 5 using M1. The regressing error $\hat{s}^2(\mathbf{x})$ is given by Equation (7), so the total estimation standard error given $\hat{\sigma}_{\mathrm{noise}} = 0.47$ is formulated by

$$\hat{s}_{\mathrm{TT}}(\mathbf{x}) = \sqrt{\hat{s}^2(\mathbf{x}) + \hat{\sigma}^2_{\mathrm{noise}}} \qquad (18)$$

The estimated mean values with one standard error upper and lower bounds are given by $\hat{\mathbf{f}}_{\mathrm{TT}}(\mathbf{x}) \pm \hat{\mathbf{s}}_{\mathrm{TT}}(\mathbf{x})$ for the training set $\mathbf{X}$. As a comparison, we plot the random observations $\mathbf{y}$ as well. The sample points in $\mathbf{X}$ are sorted according to estimated mean values of the average travel time in a descending order. The $\hat{\mathbf{f}}_{\mathrm{TT}}(\mathbf{x})$ of M1 decreases quickly and the estimated response of the last point adds bias to the observations.

Figures 7b and c show the CV results of M3 and M11 for the 67 initial sample points. The difference between them is very small, which is also shown in Table 6. Figure 7d shows the results of the Kriging model (M12) for 97 points (including the initial samples and 30 infill points using the EI maximization criterion). The estimated standard deviation is smaller in the Kriging, which means the optima after 30 infill points would have a smaller variance. Note that all random observations are within two standard deviations from the mean accounting for about 95% confidence level. The results show two advantages of M12: (1) M12 shows much narrower estimation bounds than M3, M11, or M1, which indicates a higher predicting concentration; (2) for 30 infill points, the mean value of the objective function is smaller than M3 and M11.

Therefore, based on the overall regression performance (indicated by RMSE, MAE, NRMSE, NMAE, and EGO) and the prediction error bounds, we find the infill Kriging with noise (M12) gives the best solution to the problem. At the end of the 30th update, the estimated best solution is $\hat{\mathbf{x}}^* = [\text{US\$2.28, US\$0.15, US\$1.29, US\$1.31, US\$0.24, 69\%}]^{\mathrm{T}}$, the estimated global mean value of the minimized objective function is $\hat{f}^*_{\mathrm{TT}} = \hat{f}_{\mathrm{TT}}(\hat{\mathbf{x}}^*) = 17.36$ minutes. The ratio of the off-peak to peak pricing is reduced, not suggesting that tolls in the off-peak are increased, because the peak-hour toll charge rates of the optima increase for highway Segments I, III, and IV, but the rates decrease for highway Segments II and V. We calculate the total revenue for both off-peak and peak periods in this article. Based on an additional 10 runs of simulation under the same best input (they are 17.96, 17.94, 17.71, 18.22, 18.30, 17.83, 18.04, 17.78, 19.31, 18.44, 17.70 minutes), we find the mean value of simulation outputs is 17.70 minutes that is 2.5% less than with the mean value of the baseline average travel time. We compare the *F* statistic and *p* value based on the analysis of variance (ANOVA) for the baseline and optimized toll charge plans. Results show the *F* statistic is 4.68 and the corresponding *p* value is 0.04 that is close to zero indicating mean travel times are significantly different.

The simulation generates statistics of the network performance every minute. As link volume fluctuates significantly at the 1-minute interval, Figure 8 shows 10-minute moving average traffic flow volume of the 10 additive toll links in the network. The left column shows westbound ICC, and the right column shows eastbound ICC. Two subfigures in each row represent westbound

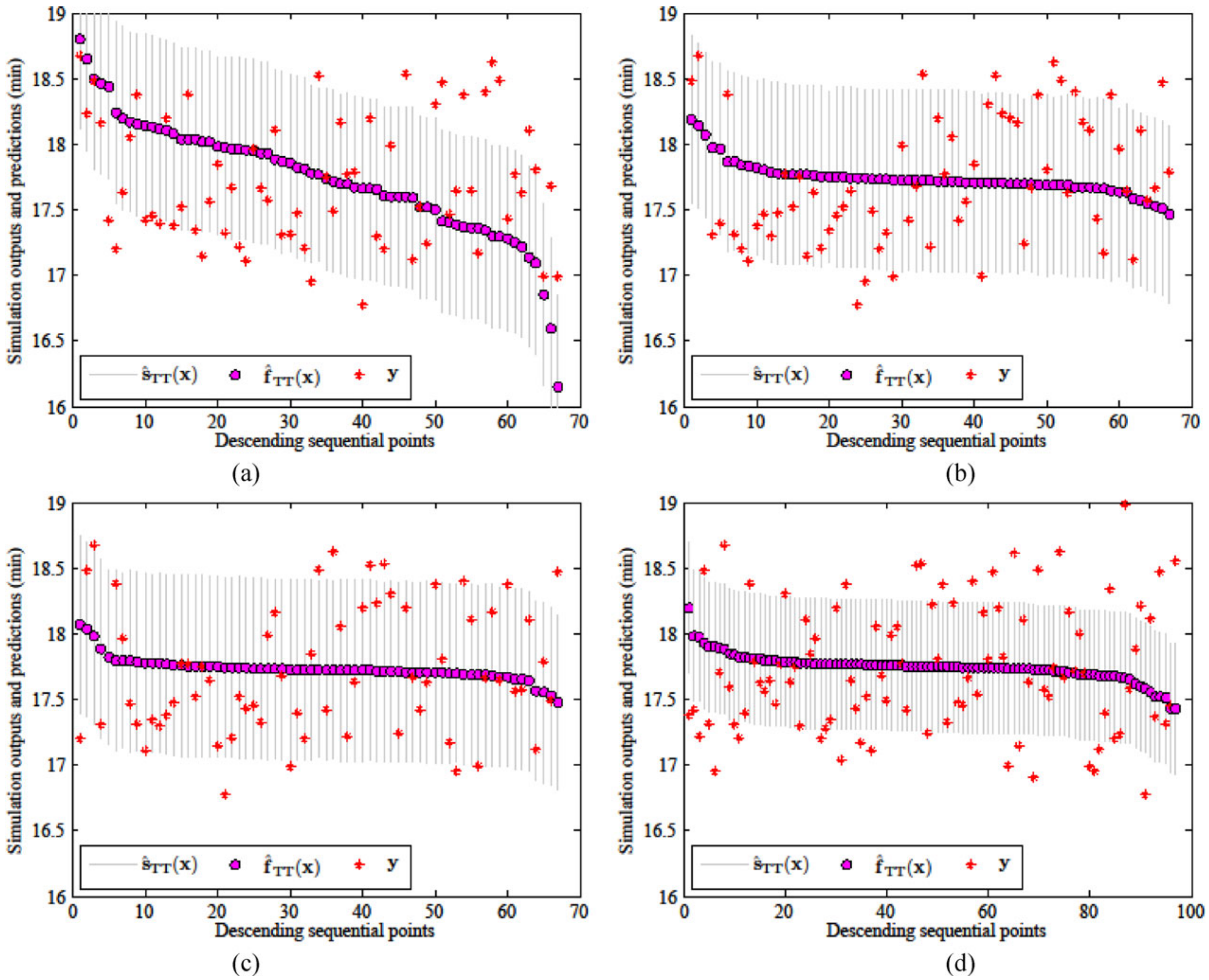


Fig. 7. Prediction accuracy of the leave-one-out cross-validation: (a) M1; (b) M3; (c) M11; (d) M12.

and eastbound toll segment, that is, Segments I, II, III, IV, and V from top to bottom. Overall, the time series of toll link volumes is changed under the optimized toll compared to the baseline, whereas the total flow passing through each toll link during the simulation period did not change a lot. In 8 of the 10 additive toll links, traffic volume is significantly higher during the fifth hour, that is, 9–10 AM (240–300 minutes) under the optimal toll.

Figure 9 illustrates the average travel time per vehicle at each instance, over time in minutes for the entire network. The average travel time was estimated by DynusT based on the experienced travel time of vehicles that finished their trips for every minute. It clearly shows that network average travel time is reduced in the optimal toll case. The most significant reduction occurs around 8:50 AM to 9:30 AM. This period contains the end of morning peak and the adjacent half hour after the morning peak, which is the most congested period with the longest travel time. Thus, the optimal toll rate successfully helps alleviate congestion throughout the network.

Figure 10a shows the comparison of toll revenue dynamics between the optimal toll and the baseline. The curve of cumulative toll revenue collected along time in Figure 10b shows that the optimized toll case generates a toll revenue of around US$62,000 (the value of time is US$15 per hour), which is a 20% increase compared to the current toll case. During the first hour of the simulation period, toll revenue collected under both cases is almost the same. During the three peak hours from 6 to 9 AM, as traffic flow of all toll links is very close, the increased toll revenue of the optimal toll case mainly comes from the increased peak toll rates. The peak/off-peak ratio of the optimal toll case (69%) is smaller than that of the current toll case (80%), and the average off-peak toll rates of the two cases are about on the same level. The increase of toll revenue during the last hour of the simulation period under the optimal toll mainly

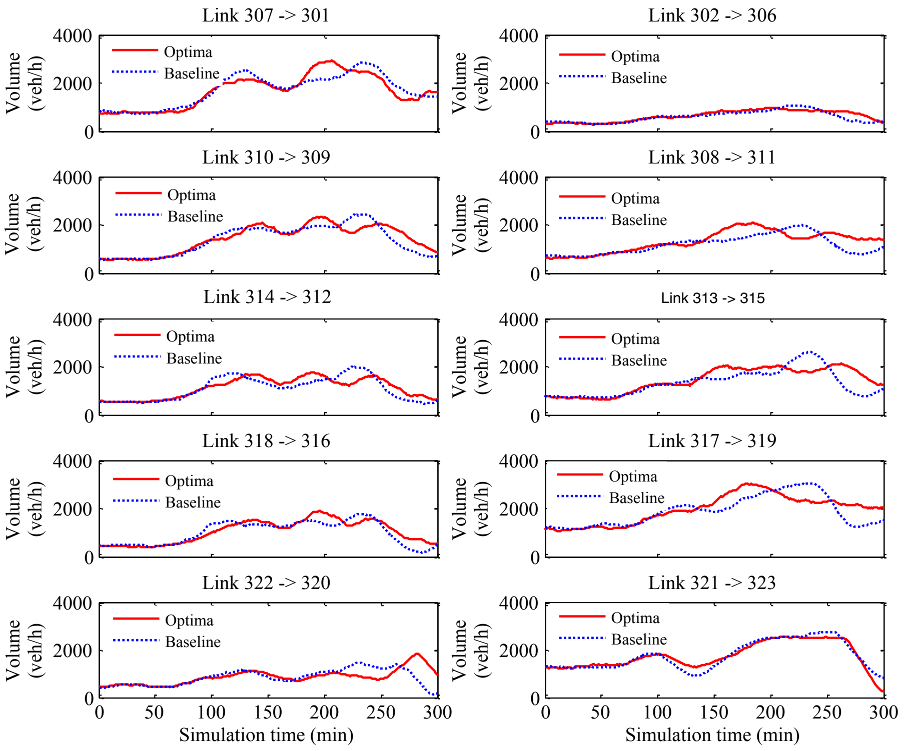


Fig. 8. Comparisons of traffic flow volumes of additive toll links between the baseline and optimized toll rates.

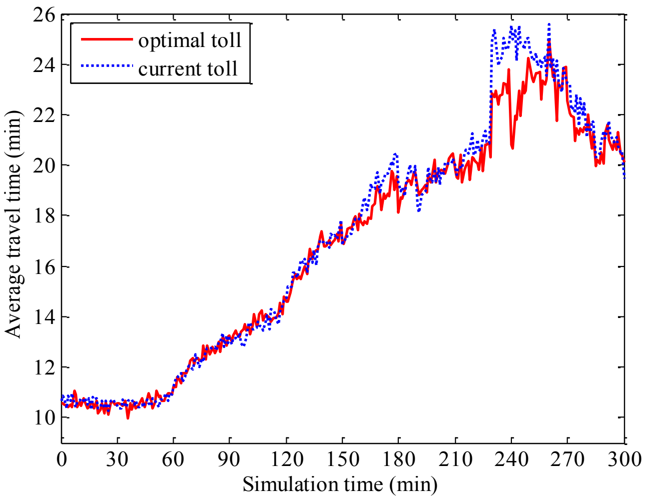
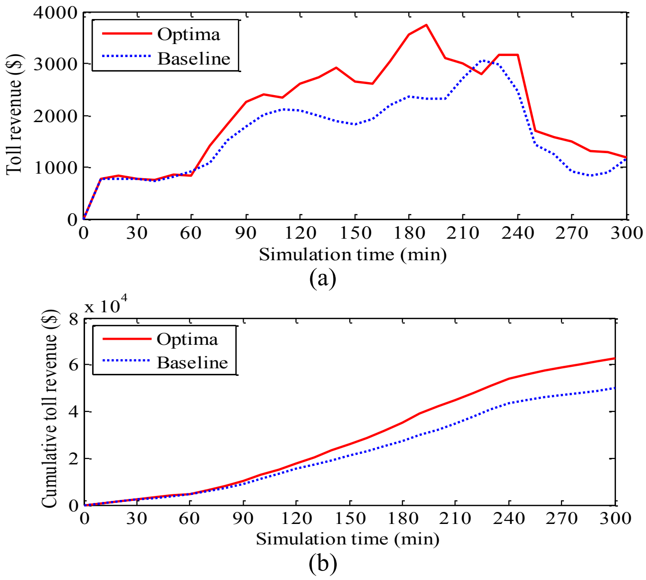


Fig. 9. Comparisons of the overall performance of the road network between the baseline and optimized toll rates.

comes from the increase of link volumes. Figure 11 compares the total flow throughput at the network exit for the optimal and the baseline solutions. The optima increase the throughput capacity in peak hours (especially from 180 to 240 minutes).

In this article, we only suggest change the toll rates of one highway, the influence of which on the whole



(a)

(b)

Fig. 10. Comparison of toll revenue collected on additive toll links between the baseline and optimized toll rates. (a) Toll revenue; (b) cumulative toll revenue.

transportation network should be small. However, the small improvement in the mean travel time of all users of the transportation network (2.5% reduction) can not be neglected because there are more than US$10,000

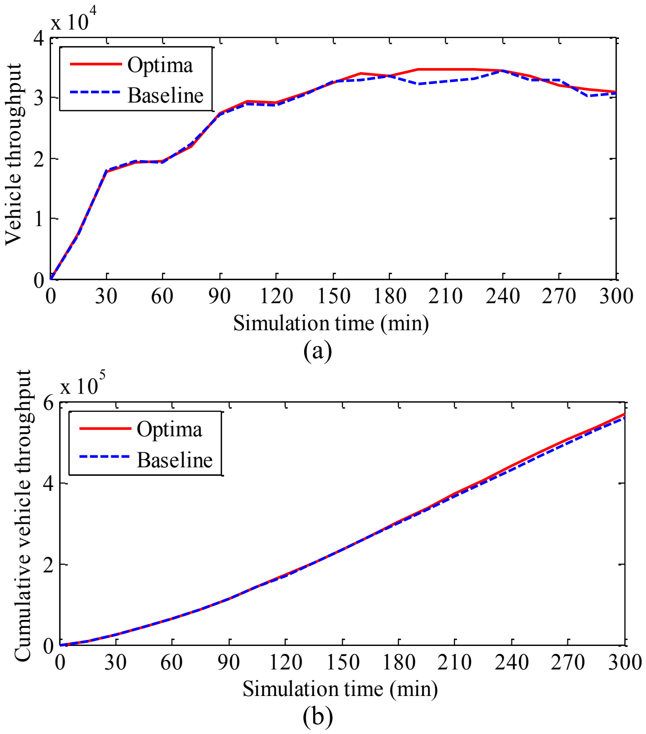


Fig. 11. Comparison of the vehicle throughout between the baseline and optimized toll rates. (a) Vehicle throughput; (b) cumulative vehicle throughput.

savings when we multiply the saving time by the number of users and the value of time (US$15 per hour) for the 5-hour simulation. If we consider the whole 24 hours of each day, and even consider a 1-year effect, such a small improvement in mean travel time in the extended peak hours would mean a huge savings from an operational and policy standpoint.

Overall, the simulation results show that implementing the optimal toll predicted by the Kriging model considering simulation noise can benefit society in multiple ways. Travelers gain from the reduction of travel cost and the government benefits from the increase of toll revenue, whereas there is hardly any cost associated with the change of toll rates. Thus, adjusting the current toll rate to the optimal toll rate should be an encouraging policy option to enhance transportation system performance in the study region.

5.5 Sensitivity analysis

To explore the sensitivity for the baseline $\boldsymbol{x}_{\text{baseline}}$ and the optima $\hat{x}^*$ in the Kriging approach considering noise data (M12), we provide a joint contour plot of the baseline in Figure 12a. Each tile shows a contour of the estimated surrogate function $\hat{f}_{\text{TT}}(\boldsymbol{x})$ (the average travel time) versus two of the six variables, with the remaining four variables held at the baseline value. This helps visualize how the surrogate values change around the $\boldsymbol{x}_{\text{baseline}}$. The baseline values and the ranges of each dimension can be found in Table 5. Take the upper-left contour plot of the average travel time surrogate function as an example, it is a conditional function of $\hat{f}_{\text{TT}}(z_1, z_2|z_3, z_4, z_5, \tau)$ in terms z_1 and z_2 given $z_3 =$ US$0.75, $z_4 =$ US$0.65, $z_5 =$ US$0.70, $\tau = 80\%$. The warmer colors of the joint contour plots indicate the longer average travel time, whereas the cooler colors show shorter travel time values (see color figure online).

Analogically, the joint contour plots in Figure 12b show the sensitivity analysis around the optima $\hat{x}^*$, which is denoted by the squares. The main difference from Figure 12a is that the values around $\hat{x}^*$ are toward much cooler colors (closer to cyan and blue) than the baseline sensitivity (closer to orange and red) (see color

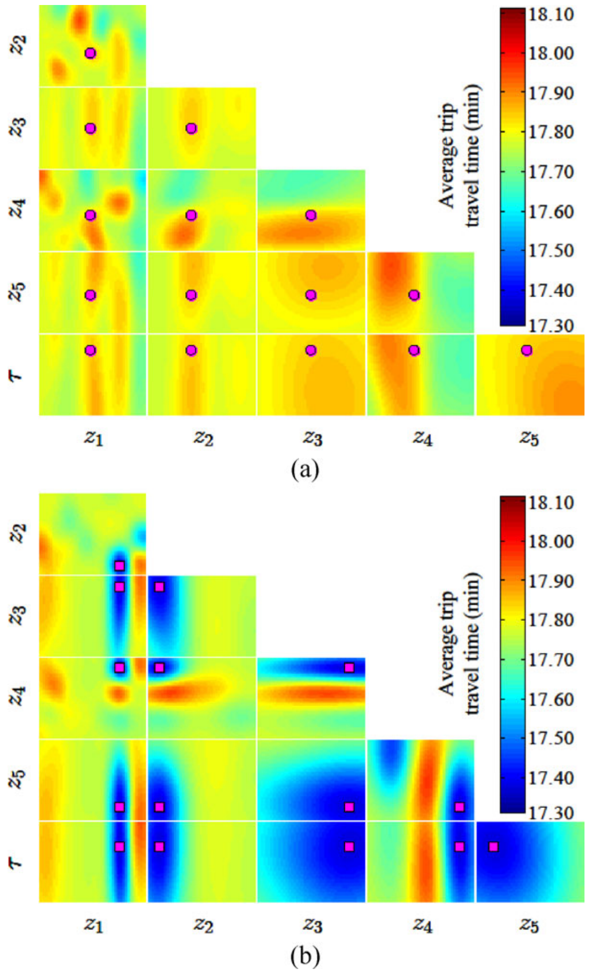


Fig. 12. Sensitivity analysis of the (a) baseline and (b) optima.

figure online). It also validates that the optimal solution performs better than the baseline inputs.

6 DISCUSSIONS AND CONCLUSIONS

6.1 Conclusions

The primary contributions of this article include: (1) a systematic framework of transportation simulation-based optimization is proposed to solve the highway toll optimization with expensive-to-evaluate objective functions and obvious simulation random errors that cannot be neglected; (2) the main novelty of computational time savings can be achieved by using the efficient surrogate-based optimization approach that can intelligently mimic the simulation input–output mapping; (3) the application of these optimization procedures are successfully tested in a transportation network using a mesoscopic simulation-based DTA. The computational burden is heavy and is usually not affordable when we use a transportation simulation tool to solve an optimization problem with multidimensional decision variables.

Building surrogate models is an intuitive way to deal with optimization problems with no-closed-form objective functions, especially when the evaluation of objective functions is very computationally expensive. This article adopts a family of surrogate-based optimization approaches to model response surface of transportation simulation input–output with expensive-to-evaluate objectives and random noise. Both one-stage and two-stage surrogate models are tested and compared. A suboptimal exploration strategy that induces local optimization and a global exploration strategy to locate the global optimum are incorporated and validated. This article tested 12 response surface models with a small-scale numerical study. Five MoEs were proposed to compare the performance of these models. Seven of the 12 models were successful in finding the exact optimal toll rate for the case study and these models were always associated with smaller estimation errors.

A simulation-based DTA model DynusT is utilized to evaluate the system performance in response to different toll charges in a real transportation network. This article utilizes response surface models to solve the problem of minimizing average travel time by adjusting variable toll rates of five links in the ICC network in the State of Maryland. Four of the seven promising models were tested using the ICC road network. It shows that with only 97 samples, the Kriging model considering simulation noise could produce highly reliable estimates of simulation outputs over the entire feasible domain, and thus successfully help find the optimal toll rate.

The predicted optimal toll rate obtained from the Kriging model was then evaluated through simulation. The predicted output was relatively consistent with simulation outputs. The average travel time under optimal toll was 17.70 minutes, which is 2.5% shorter (0.45 minutes shorter) than the average travel time under the current tolls. The total vehicle throughput for the simulation period is around 570,000. Assuming the value of time is US$15 per hour for the network users, the reduction in average travel time equals a total of US$65,000 savings in travel cost for the extended morning peak period (5 hours) for each day. Another interesting finding is that the implementation of the optimal toll rate led to a 20% increase of toll revenue.

6.2 Limitations and future studies

Although the surrogate-based optimization framework and a family of models are revisited, implemented, and compared using a bilevel toll optimization example of a small road network and in a transportation network with expensive-to-evaluate and noisy objective functions, we recommend the following further research:

1. **Multiple objective functions**. The objective of the current research is to reduce travel time, which is mainly from the government's perspective. Some other objective functions may be taken into account, such as minimization of average travel time, total travel time and/or delay of network travelers, total revenue maximization of private infrastructure operators, maximization of total social welfare from the perspective of government, and maximization of road network reliability indicators. Taking advantage of the simulation, optimization of combination of policies, such as travel demand policies (e.g., high occupancy vehicle [HOV], HOT, bus only lane) combined with traffic control policies (e.g., ramp meter, bus priority signal) can also be analyzed.

   In general, though the optima of each objective may be different, multiple objective solutions based on Pareto improvement would be beneficial to policy makers. The algorithms to solve multiobjective optimization problems based on the surrogate approximation approaches will be discussed in the coming report.
2. **Behavior adjustments: elastic demands**. When we model and simulate traveler behavior adjustments to road pricing, for example, departure time choice, route choice, and even mode choice, the assumption of fixed demands (deterministic OD tables for single occupancy vehicle [SOV], HOV, and truck in the current study) will be released

to allow elastic demands. Behavioral adjustments (Zhang et al., 2012) will be integrated into open source simulation tools, for example, DynusT, to better model the real responses of travelers' dynamics in a transportation network.

3. **Advanced exploitation of surrogates**. There are two potential ways to make full use of the surrogate approximations. The first is to explicitly estimate the probability distribution of the optimum location, which allows an information-based global optimization strategy, and the other one can be applying the Gaussian Process (GP) to obtain a better infill strategy that guarantees maximum *a posteriori* (MAP) for prediction.
4. **Heterogeneous travelers: value of time distribution**. The value of time is assumed to be uniform across homogeneous travelers in this article. However, heterogeneous characteristics of travelers, for example, multiple driver classes with different values of time, time-varying tolls, may have direct influences on the route choice behavior. This interesting issue of incorporating the value of time distribution for network travelers will be discussed in the future.

ACKNOWLEDGMENTS

This work was funded by the National Science Foundation CAREER Award Project under Grant 1150925 "Reliability as an Emergent Property of Transportation Networks," the U.S. Federal Highway Administration Exploratory Advanced Research Program, and the Maryland State Highway Administration.

REFERENCES

Adeli, H. (1994), ed. *Advances in Design Optimization*, Chapman & Hall, London.

Adeli, H. (2001), Neural networks in civil engineering: 1989–2000, *Computer-Aided Civil and Infrastructure Engineering*, **16**(2), 126–42.

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

Adeli, H. & Jiang, X. (2003), Neuro-fuzzy logic model for freeway work zone capacity estimation, *Journal of Transportation Engineering*, **129**(5), 484–93.

Adeli, H. & Jiang, X. (2009), *Intelligent Infrastructure: Neural Networks, Wavelets, and Chaos Theory for Intelligent Transportation Systems and Smart Structures*, CRC Press, Taylor & Francis, Boca Raton, FL.

Adeli, H. & Karim, A. (2000), Fuzzy-wavelet RBFNN model for freeway incident detection, *Journal of Transportation Engineering*, **126**(6), 464–71.

Adeli, H. & Karim, A. (2005), *Wavelets in Intelligent Transportation Systems*, John Wiley and Sons, West Sussex, United Kingdom.

Aziz, H. M. & Ukkusuri, S. V. (2012), Integration of environmental objectives in a system optimal dynamic traffic assignment model, *Computer-Aided Civil and Infrastructure Engineering*, **27**(7), 494–511.

Balakrishna, R., Antoniou, C., Ben-Akiva, M., Koutsopoulos, H. N. & Wen, Y. (2007a). Calibration of microscopic traffic simulation models: methods and application, *Transportation Research Record*, **1999**, 198–207.

Balakrishna, R., Ben-Akiva, M. & Koutsopoulos, H. N. (2007b). Offline calibration of dynamic traffic assignment: simultaneous demand-and-supply estimation, *Transportation Research Record*, **2003**, 50–8.

Baraldi, P., Canesi, R., Zio, E., Seraoui, R. & Chevalier, R. (2011), Genetic algorithm-based wrapper approach for grouping condition monitoring signals of nuclear power plant components, *Integrated Computer-Aided Engineering*, **18**(3), 221–34.

Barton, R. R. & Meckesheimer, M. (2006), *Metamodel-Based Simulation Optimization, Handbooks in Operations Research and Management Science*, Elsevier B.V, Amsterdam.

Beckmann, M. J. (1965). On optimal tolls for highways, tunnels and bridges, in L. Edie, R. Herman, and R. Rothery (eds.), *Vehicular Traffic Science*, American Elsevier, New York, pp. 331–41.

Bekhor, S., Dobler, C., & Axhausen, K. W. (2011), Integration of activity-based with agent-based models: an example from the Tel Aviv model and MATSim, in the *90th Annual Meeting of Transportation Research Board*, Washington, DC.

Ben-Akiva, M., Bierlaire, M., Koutsopoulos, H. N. & Mishalani, R. (2002), Real-time simulation of traffic demand-supply interactions within DynaMIT, in M. Gendreau and P. Marcotte (eds.), *Transportation and Network Analysis, Current Trends*, Kluwer, pp. 19–36.

Björkman, M. & Holmström, K. (2000), Global optimization of costly nonconvex functions using radial basis functions, *Optimization and Engineering*, **1**(4), 373–97.

Chabuk, T., Reggia, J. A., Lohn, J. & Linden, D. (2012), Causally-guided evolutionary optimization and its application to antenna array design, *Integrated Computer-Aided Engineering*, **19**(2), 111–24.

Chiu, Y.-C. & Bustillos, B. (2009), A gap function vehicle-based solution procedure for consistent and robust simulation-based dynamic traffic assignment, in *The 88th Annual Meeting of Transportation Research Board*, Washington DC.

Chiu, Y.-C., Zhou, L. & Song, H. (2010), Development and calibration of the anisotropic mesoscopic simulation model for uninterrupted flow facilities, *Transportation Research Part B*, **44**(1), 152–74.

Dharia, A. & Adeli, H. (2003), Neural network model for rapid forecasting of freeway link travel time, *Engineering Applications of Artificial Intelligence*, **16**(7), 607–13.

Ekstrom, J., Sumalee, A. & Lo, H. K. (2012), Optimizing toll locations and levels using a mixed integer linear approximation approach, *Transportation Research Part B*, **46**(7), 834–54.

Flötteröd, G., Bierlaire, M. & Nagel, K. (2011), Bayesian demand calibration for dynamic traffic simulations, *Transportation Science*, **45**(4), 541–61.

Forrester, A., Sobester, A. & Keane, A. (2008), *Engineering Design via Surrogate Modelling: A Practical Guide*, Wiley, Chichester, United Kingdom.

Forrester, A. I. & Keane, A. J. (2009), Recent advances in surrogate-based optimization, *Progress in Aerospace Sciences*, **45**(1), 50–79.

Forrester, A. I., Keane, A. J. & Bressloff, N. W. (2006), Design and analysis of "noisy" computer experiments, *AIAA Journal*, **44**(10), 2331–9.

Fu, M. C. (2002), Optimization for simulation: theory vs. practice, *INFORMS Journal on Computing*, **14**(3), 192–215.

Gao, H. & Zhang, X. (2013), A Markov-based road maintenance optimization model considering user costs, *Computer-Aided Civil and Infrastructure Engineering*, **28**(6), 451–64.

Ghosh-Dastidar, S. & Adeli, H. (2003), Wavelet-clustering-neural network model for freeway incident detection, *Computer-Aided Civil and Infrastructure Engineering*, **18**(5), 325–38.

Ghosh-Dastidar, S. & Adeli, H. (2006), Neural network-wavelet micro-simulation model for delay and queue length estimation at freeway work zones, *Journal of Transportation Engineering, ASCE*, **132**(4), 331–41.

Ghosh-Dastidar, S., Adeli, H. & Dadmehr, N. (2008), Principal component analysis-enhanced cosine radial basis function neural network for robust epilepsy and seizure detection, *IEEE Transactions on Biomedical Engineering*, **55**(2), 512–8.

Gibbs, M. N. (1997), Bayesian Gaussian processes for regression and classification. Doctoral dissertation, University of Cambridge.

Gutmann, H.-M. (2001), A radial basis function method for global optimization, *Journal of Global Optimization*, **19**(3), 201–27.

Han, S. (2003), Dynamic traffic modelling and dynamic stochastic user equilibrium assignment for general road networks, *Transportation Research Part B*, **37**(3), 225–49.

Hatzopoulou, M., Hao, J. Y. & Miller, E. J. (2011). Simulating the impacts of household travel on greenhouse gas emissions, urban air quality, and population exposure, *Transportation*, **38**, 871–87.

Huang, D., Allen, T. T., Notz, W. I. & Zeng, N. (2006), Global optimization of stochastic black-box systems via sequential Kriging meta-models, *Journal of Global Optimization*, **34**(3), 441–66.

Hussain, M. F., Barton, R. R. & Joshi, S. B. (2002), Metamodeling: radial basis functions, versus polynomials, *European Journal of Operational Research*, **138**(1), 142–54.

Jakobsson, S., Patriksson, M., Rudholm, J. & Wojciechowski, A. (2010), A method for simulation based optimization using radial basis functions, *Optimization and Engineering*, **11**(4), 501–32.

Jiang, X. & Adeli, H. (2003), Freeway work zone traffic delay and cost optimization model, *Journal of Transportation Engineering, ASCE*, **129**(3), 230–41.

Jiang, X. & Adeli, H. (2004a), Object-oriented model for freeway work zone capacity and queue delay estimation, *Computer-Aided Civil and Infrastructure Engineering*, **19**(2), 144–56.

Jiang, X. & Adeli, H. (2004b), Wavelet packet-autocorrelation function method for traffic flow pattern analysis, *Computer-Aided Civil and Infrastructure Engineering*, **19**(6), 324–37.

Jones, D. R. (2001), A taxonomy of global optimization methods based on response surfaces, *Journal of Global Optimization*, **21**(4), 345–83.

Jones, D. R., Schonlau, M. & Welch, W. J. (1998), Efficient global optimization of expensive black-box functions, *Journal of Global Optimization*, **13**(4), 455–92.

Karim, A. & Adeli, H. (2002), Comparison of fuzzy-wavelet radial basis function neural network freeway incident detection model with California algorithm, *Journal of Transportation Engineering*, **28**(1), 21–30.

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

Karim, A. and Adeli, H. (2003b), CBR model for freeway work zone traffic management, *Journal of Transportation Engineering, ASCE*, **129**(2), 134–45.

Karim, A. & Adeli, H. (2003c), Radial basis function neural network for work zone capacity and queue estimation, *Journal of Transportation Engineering*, **129**(5), 494–503.

Kim, H. & Adeli, H. (2001), Discrete cost optimization of composite floors using a floating point genetic algorithm, *Engineering Optimization*, **33**(4), 485–501.

Kleijnen, J. P. C. (2009), Kriging metamodeling in simulation: a review, *European Journal of Operational Research*, **192**(3), 707–16.

Li, D., Xu, L., Goodman, E., Xu, Y. & Wu, Y. (2013), Integrating a statistical background-foreground extraction algorithm and SVM classifier for pedestrian detection and tracking, *Integrated Computer-Aided Engineering*, **20**(3), 201–16.

Li, Z., Madanu, S., Zhou, B., Wang, Y. & Abbas, M. (2010), A heuristic approach for selecting highway investment alternatives, *Computer-Aided Civil and Infrastructure Engineering*, **25**(6), 427–39.

López, E. & Monzón, A. (2010), Integration of sustainability issues in strategic transportation planning: a multicriteria model for the assessment of transport infrastructure plans, *Computer-Aided Civil and Infrastructure Engineering*, **25**(6), 440–51.

Mahmassani, H. S. (2002), Dynamic network traffic assignment and simulation methodology for advanced system management applications, in the *81st Annual Meeting of Transportation Research Board*, Washington DC.

Meister, K., Balmer, M., Ciari, F., Horni, A., Rieser, M., Waraich, R. A. & Axhausen, K. W. (2010), Large-scale agent-based travel demand optimization applied to Switzerland, including mode choice, in the *12th World Conference on Transport Research*, Lisbon, Portugal.

Melouk, S. H., Keskin, B. B., Armbrester, C. & M. Anderson. (2010), A simulation optimization-based decision support tool for mitigating traffic congestion, *Journal of the Operational Research Society*, **62**(11), 1971–82.

Meng, Q. & Wang, X. (2008), Sensitivity analysis of logit-based stochastic user equilibrium network flows with entry–exit toll schemes, *Computer-Aided Civil and Infrastructure Engineering*, **23**(2), 138–56.

Montgomery, D. C. (2008), *Design and Analysis of Experiments*, Wiley, New York.

Morris, M. D. & Mitchell, T. J. (1995), Exploratory designs for computational experiments, *Journal of Statistical Planning and Inference*, **43**(3), 381–402.

Nava, E. & Chiu, Y.-C. (2012), A temporal domain decomposition algorithmic scheme for large-scale dynamic traffic assignment, *International Journal of Transportation Science and Technology*, **1**(1), 1–24.

Nie, Y., Ma, J. & Zhang, H. M. (2008), A polymorphic dynamic network loading model, *Computer-Aided Civil and Infrastructure Engineering*, **23**(2), 86–103.

Omrani, R. & Kattan, L. (2012). Demand and supply calibration of dynamic traffic assignment models, *Transportation Research Record*, **2283**, 100–12.

Osorio, C. (2010), Mitigating network congestion: analytical models, optimization methods and their applications. Ph.D. dissertation, Ecole Polytechnique Fédérale de Lausanne, Lausanne.

Papola, A. & Marzano, V. (2013), A network generalized extreme value model for route choice allowing implicit route enumeration, *Computer-Aided Civil and Infrastructure Engineering*, **28**(8), 560–80.

Pendyala, R. M., Chiu, Y. C., Waddell, P., Hickman, M., Konduri, K. C. & Sana, B. (2010), The design of an integrated model of the urban continuum—location choices, activity-travel behavior, and dynamic traffic patterns, in the *12th World Congress of Transport Research*, Lisbon, Portugal.

Queipo, N. V., Haftka, R. T., Shyy, W., Goel, T., Vaidyanathan, R. & Tucker, P. K. (2005), Surrogate-based analysis and optimization, *Progress in Aerospace Sciences*, **41**(1), 1–28.

Ramadurai, G. & Ukkusuri, S. (2011), B–dynamic: an efficient algorithm for dynamic user equilibrium assignment in activity–travel networks, *Computer-Aided Civil and Infrastructure Engineering*, **26**(4), 254–69.

Rasmussen, C. E. & Williams, C. K. I. (2006), *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA.

Regis, R. G. & Shoemaker, C. A. (2005), Constrained global optimization of expensive black box functions using radial basis functions, *Journal of Global Optimization*, **31**(1), 153–71.

Sánchez, A., Nunes, E. O. & Conci, A. (2012), Using adaptive background subtraction into a multi-level model for traffic surveillance, *Integrated Computer-Aided Engineering*, **19**(3), 239–56.

Sarma, K. C. & Adeli, H. (2001), Bi-level parallel genetic algorithms for optimization of large steel structures, *Computer-Aided Civil and Infrastructure Engineering*, **16**(5), 295–304.

Smola, A. J. & Schölkopf, B. (2004), A tutorial on support vector regression, *Statistics and Computing*, **14**(3), 199–222.

Song, M., Yin, M., Chen, X., Zhang, L. & Li, M. (2013), A simulation-based approach for sustainable transportation systems evaluation and optimization: theory, systematic framework and applications, in *Proceedings of the 13th COTA International Conference of Transportation Professionals*, Shenzhen, China.

Sumalee, A. (2004), Optimal road user charging cordon design: a heuristic optimization approach, *Computer-Aided Civil and Infrastructure Engineering*, **19**(5), 377–92.

Sundaram, S., Koutsopoulos, H. N., Ben-Akiva, M., Antoniou, C. & Balakrishna, R. (2011). Simulation-based dynamic traffic assignment for short-term planning applications, *Simulation Modelling Practice and Theory*, **19**(1), 450–62.

Szeto, W. Y., Jaber, X. & O'Mahony, M. (2010), Time-dependent discrete network design frameworks considering land use, *Computer-Aided Civil and Infrastructure Engineering*, **25**(6), 411–26.

Szeto, W. Y., Solayappan, M. & Jiang, Y. (2011), Reliability-based transit assignment for congested stochastic transit networks, *Computer-Aided Civil and Infrastructure Engineering*, **26**(4), 311–26.

Teklu, F., Sumalee, A. & Watling, D. (2007), A genetic algorithm approach for optimizing traffic control signals considering routing, *Computer-Aided Civil and Infrastructure Engineering*, **22**(1), 31–43.

Tian, X., Hahut, M., Jha, M., & Florian, M. A. (2007), Dynameq application to evaluating the impact of freeway reconstruction, in the *86th Annual Meeting of Transportation Research Board*, Washington DC.

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

Unnikrishnan, A. & Lin, D.-Y. (2012), User equilibrium with recourse: continuous network design problem, *Computer-Aided Civil and Infrastructure Engineering*, **27**(7), 512–24.

Unnikrishnan, A., Valsaraj, V., Damnjanovic, I. & Waller, S. T. (2009), Design and management strategies for mixed public private transportation networks: a meta-heuristic approach, *Computer-Aided Civil and Infrastructure Engineering*, **24**(4), 266–79.

Vaze, V., Antoniou, C., Wen, Y. & Ben-Akiva, M. (2009). Calibration of dynamic traffic assignment models with point-to-point traffic surveillance, *Transportation Research Record*, **2090**, 1–9.

Villemonteix, J., Vazquez, E. & Walter, E. (2009), An informational approach to the global optimization of expensive-to-evaluate functions, *Journal of Global Optimization*, **44**(4), 509–34.

Wandekokem, E. D., Mendel, E., Fabris, F., Valentim, M., Batista, R. J., Varejao, F. M. & Rauber, T. W. (2011), Diagnosing multiple faults in oil rig motor pumps using support vector machine classifier ensembles, *Integrated Computer-Aided Engineering*, **18**(1), 61–74.

Yang, H. (1999), System optimum, stochastic user equilibrium and optimal link tolls, *Transportation Science*, **33**(4), 354–60.

Yang, H. & Bell, M. G. H. (1997), Traffic restraint, road pricing and network equilibrium, *Transportation Research Part B*, **31**(4), 303–14.

Yang, H. & Huang, H. J. (1998), Principle of marginal-cost pricing: how does it work in a general network? *Transportation Research Part A*, **32**, 45–54.

Yang, H. & Huang, H. J. (2004), The multiclass, multicriteria traffic network equilibrium and system optimum problem, *Transportation Research Part B*, **38**(1), 1–15.

Yang, H. & Huang, H. J. (2005), *Mathematical and Economic Theory of Road Pricing*, Elsevier, Oxford.

Yang, H. & Lam, W. H. K. (1996), Optimal road tolls under conditions of queuing and congestion, *Transportation Research Part A*, **30**(5), 319–32.

Yang, H. & Zhang, X. (2002), Multi-class network toll design problem with social and spatial equity constraints, *Journal of Transportation Engineering*, **128**(5), 420–8.

Yang, H., Zhang, X. & Meng, Q. (2004), Modeling private highways in networks with entry–exit based toll charges, *Transportation Research Part B*, **38**(3), 191–213.

Zhang, L., Chang, G. L., Zhu, S., Xiong, C., Du, L., Mollanejad, M. & Mahapatra, S. (2012), Integrating an agent-based travel behavior model with large-scale microscopic traffic simulation for corridor-level and subarea transportation operations and planning applications, *Journal of Urban Planning and Development*, **139**(2), 94–103.

Zhou, L., Yan, G. & Ou, J. (2013), Response surface method based on radial basis functions for modeling large-scale structures in model updating, *Computer-Aided Civil and Infrastructure Engineering*, **28**(3), 210–26.

APPENDIX A: LATIN HYPERCUBE SAMPLING (LHS)

A space-filling DoE is useful when only few runs of simulation can be afforded within the computational budget. In LHS, each design variable is stratified into an equal number of intervals. We use LHS to generate initial samples to fit surrogate models. Different from classic designs such as 2^k or 3^k fractional factorial designs and central composite designs (CCD) in Montgomery (2008), each dimension of the decision variable is split into a relatively large number of equal-size bins, in which subsamples are uniformly generated. The advantage of LHS is that the mapping of high-dimensional design inputs into each dimension is uniformly distributed without overlap. Thus, such a property makes LHS one of the space-filling types of DoE.

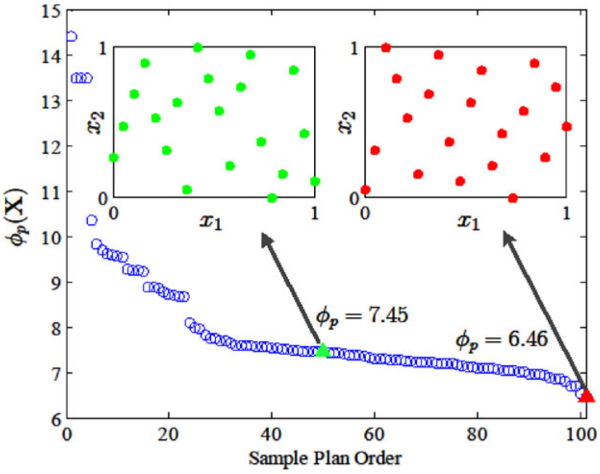


Fig. A.1. An illustration of 100 LHS DoEs based on the spacing filling criterion $\phi_p(\mathbf{X})$.

In this article, we recall the maximin design defined by Forrester et al. (2008). The ranking criterion function proposed by Morris and Mitchell (1995) is

$$X^* = \arg\min_X \phi_p(X) = \arg\min_X \left(\sum_{j=1}^{m} J_j d_j^{-p}\right)^{1/p} \quad \text{(A.1)}$$

where $d = ||\mathbf{x}^{(i)} - \mathbf{x}^{(j)}|| = \sqrt{(\mathbf{x}^{(i)} - \mathbf{x}^{(j)})^{\mathrm{T}}(\mathbf{x}^{(i)} - \mathbf{x}^{(j)})}$, and values of m distances are sorted in an ascending order, that is, $d_1 \leq d_2 \leq, \ldots, \leq d_m$. Let $J_1, J_2, \ldots, J_m$ be the number of $d_1, d_2, \ldots, d_m$, respectively. To illustrate the concept of LHS, we generate 100 different plans. Each one contains 20 points. Then we calculate the ϕ_p values of 100 different plans. Figure A.1 shows the evolutionary process of the sampling plan space-filling values, two of which are zoomed in and compared. The LHS plan with a lower value of ϕ_p distributes more uniformly in the feasible domain.

APPENDIX B: MEASURES OF EFFECTIVENESS (MOE)

To test the accuracy of the surrogate models, we apply the CV approach in this article. The sample data are divided into n_{CV} subsets (n_{CV}-fold CV) of approximately equal size, if we let n_{CV} equal the sample size, that is, $n_{\mathrm{CV}} = n$, the leave-one-out CV can be conducted. The overall performance of the surrogate models is evaluated using five accuracy measures:

1. Root Mean Square Error (RMSE), which provides a global error measure over the entire design domain

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}^{(i)}) - \hat{f}(\mathbf{x}^{(i)}))^2} \quad \text{(B.1)}$$

2. Maximum Absolute Error (MAE), which is indicative of local deviations

$$\mathrm{MAE} = \max_{1\leq i\leq n} |f(\mathbf{x}^{(i)}) - \hat{f}(\mathbf{x}^{(i)})| \quad \text{(B.2)}$$

3. Normalized Root Mean Squared Error (NRMSE)

$$\mathrm{NRMSE} = \sqrt{\frac{\sum_{i=1}^{n}[f(\mathbf{x}^{(i)}) - \hat{f}(\mathbf{x}^{(i)})]^2}{\sum_{i=1}^{n}(f(\mathbf{x}^{(i)}))^2}} \quad \text{(B.3)}$$

4. Normalized Maximum Absolute Error (NMAE)

$$\mathrm{NMAE} = \frac{\max_{1\leq i\leq n}|f(\mathbf{x}^{(i)}) - \hat{f}(\mathbf{x}^{(i)})|}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}^{(i)}) - \bar{f})^2}}, \quad \bar{f} = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(\mathbf{x}^{(i)}) \quad \text{(B.4)}$$

5. Estimated Global Optimum (EGO)

$$\mathrm{EGO} = \hat{f}(\hat{\mathbf{x}}^*) \quad \text{(B.5)}$$

where $\mathbf{x}^* = \arg\min f(\mathbf{x})$, $\hat{\mathbf{x}}^* = \arg\min \hat{f}(\mathbf{x})$.

6. Pearson correlation coefficient (PCC)

$$r^2 = \left(\frac{N \sum f\hat{f} - \sum f \sum \hat{f}}{\sqrt{[N \sum f^2 - (\sum f)^2][N \sum \hat{f}^2 - (\sum \hat{f})^2]}} \right)^2 \tag{B.6}$$

where N is the number of independent set of objective function data to be compared, f denotes objective function values from the independent test set, and $\hat{f}$ are the corresponding surrogate model estimations. If $r^2 = 1$, the surrogate is exactly predicting the test data, whereas $r^2 = 0$ indicates no correlation between the surrogate and the objective function.