This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Data-Driven Robust Taxi Dispatch Under Demand Uncertainties

Fei Miao, *Member, IEEE*, Shuo Han, *Member, IEEE*, Shan Lin, Qian Wang, John A. Stankovic, *Life Fellow, IEEE*, Abdeltawab Hendawi, Desheng Zhang, *Member, IEEE*, Tian He, and George J. Pappas, *Fellow, IEEE*

***Abstract*— In modern taxi networks, large amounts of taxi occupancy status and location data are collected from networked in-vehicle sensors in realtime. They provide knowledge of system models on passenger demand and mobility patterns for efficient taxi dispatch and coordination strategies. Such approaches face new challenges: how to deal with uncertainties of predicted customer demand while fulfilling the system's performance requirements, including minimizing taxis' total idle mileage and maintaining service fairness across the whole city; how to formulate a computationally tractable problem. To address this problem, we develop a data-driven robust taxi dispatch framework to consider spatial-temporally correlated demand uncertainties. The robust vehicle dispatch problem we formulate is concave in the uncertain demand and convex in the decision variables. Uncertainty sets of random demand vectors are constructed from data based on theories in hypothesis testing, and provide a desired probabilistic guarantee level for the performance of robust taxi dispatch solutions. We prove equivalent computationally tractable forms of the robust dispatch problem using the minimax theorem and strong duality. Evaluations on four years of taxi trip data for New York City show that by selecting a probabilistic guarantee level at 75%, the average demand–supply ratio error is reduced by 31.7%, and the average total idle driving distance is reduced by 10.13% or about 20 million miles annually, compared with nonrobust dispatch solutions.**

Manuscript received June 13, 2017; accepted September 24, 2017. Manuscript received in final form October 20, 2017. This work was supported in part by NSF through the project "CPS: Synergy: Collaborative Research: Multiple-Level Predictive Control of Mobile Cyber Physical Systems With Correlated Context" under Project CPS-1239152 and Project CNS-1239224 and in part by TerraSwarm. This paper was presented in part at the 54th IEEE Conference on Decision Control, Osaka, Japan, December 2015 [19]. Recommended by Associate Editor A. Serrani. *(Corresponding author: Fei Miao.)*

F. Miao is with the Department of Computer Science and Engineering, University of Connecticut, Storrs, CT 06269 USA (e-mail: fei.miao@uconn.edu).

S. Han is with the Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, IL 60607 USA (e-mail: hanshuo@uic.edu).

S. Lin is with the Department of Electrical and Computer Engineering, Stony Brook University, Long Island, NY 11794 USA (e-mail: shan.x.lin@stonybrook.edu).

Q. Wang is with ADVANCE.AI, Beijing 100010, China (e-mail: qianwangchina@gmail.com).

J. Stankovic and A. Hendawi are with the Department of Computer Science, University of Virginia, Charlottesville, VA 22904 USA (e-mail: stankovic@virginia.edu; hendawi@virginia.edu).

D. Zhang is with the Department of Computer Science, Rutgers University, NJ 08854 USA (e-mail: desheng.zhang@cs.rutgers.edu).

T. He is with the Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: tianhe@cs.umn.edu).

G. J. Pappas is with the Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19014 USA (e-mail: pappasg@seas.upenn.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCST.2017.2766042

***Index Terms*— computationally tractable approximation, data-driven robust optimization, demand uncertainties, probabilistic guarantee, resource allocation, taxi dispatch framework.**

I. Introduction

MODERN transportation systems are equipped with various sensing technologies for passenger and vehicle tracking, such as radio frequency identification and global positioning system (GPS). Sensing data collected from transportation systems provides us opportunities for understanding spatial-temporal patterns of passenger demand. Methods of predicting taxi-passenger demand [22], [28], travel time [3], [15], [27], and traveling speed [2], [13] according to traffic monitoring data have been developed.

Based on such rich spatial-temporal information about passenger mobility patterns and demands, many controls and coordination solutions have been designed for intelligent transportation systems. Robotic mobility-on-demand systems that minimize the number of rebalancing trips [24], [30], and smart parking systems that allocates resource based on a driver's cost function [14] have been proposed. Dispatch algorithms that aim to minimize customers' waiting time [17], [26] or to reduce cruising mile [29] have been developed. In our previous work [20], [21], we design a receding horizon control framework that incorporates predicted demand model and real-time sensing data. Considering future demand when making, the current dispatch decisions help to reduce autonomous vehicle balancing costs [30] and taxis' total idle distance [20], [21]. Strategies for resource allocation depend on the model of demand in general, and the knowledge and assumptions about the demand affect the performance of the supply providing approaches [9], [23]. These works rely on precise passenger-demand models to make dispatch decisions.

However, passenger-demand models have their intrinsic model uncertainties that result from many factors, such as weather, passenger working schedule, and city events. Algorithms that do not consider these uncertainties can lead to inefficient dispatch services, resulting in imbalanced workloads, and increased taxi idle mileage. Although robust optimization aims to minimize the worst case cost under all possible random parameters, it sacrifices average system performances [1]. For a taxi dispatch system, it is essential to address the tradeoff between the worst case and the average dispatch costs under uncertain demand. A promising yet challenging approach is a robust dispatch framework with an uncertain demand model, called an uncertainty set, that captures spatial-temporal

1063-6536 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

correlations of demand uncertainties and the robust optimal solution under this set provides a probabilistic guarantee for the dispatch cost [as defined in problem (12)].

In this paper, we consider two aspects of a robust vehicle dispatch model given a taxi-operational records data set: 1) how to formulate a robust resource allocation problem that dispatches vacant vehicles toward predicted uncertain demand and 2) how to construct spatial-temporally correlated uncertain demand sets for this robust resource allocation problem without sacrificing too much average performance of the system. We first develop the objective and constraints of a robust dispatch problem considering spatial-temporally correlated demand uncertainties. The objective of a system-level optimal dispatch solution is balancing workload of taxis in each region of the entire city with minimum total current and expected future idle cruising distance. We define an approximation of the balanced vehicle objective in this paper, such that the robust vehicle dispatch problem is concave of the uncertain demand and convex of the decision variables. We then design a data-driven algorithm for constructing uncertainty demand sets without assumptions about the true model of the demand vector. The constructing algorithm is based on hypothesis testing theories [6], [11], [25], however, how to apply these theories for spatial-temporally correlated transportation data and uncertainty sets of a robust vehicle resource allocation problem have not been explored before. To the best of our knowledge, this is the first work to design a robust vehicle dispatch model that provides a desired probabilistic guarantee using predictable and realistic demand uncertainty sets.

Furthermore, we explicitly design an algorithm to build demand uncertainty set from data according to different probabilistic guarantee level for the cost. With two types of uncertainty sets—box type and second-order-cone (SOC) type—we prove equivalent computationally tractable forms of the robust dispatch problem under these uncertainty demand models via the minimax theorem and the strong duality theorem. The robust dispatch problem formulated in this paper is convex over the decision variables and concave over the constructed uncertain sets with decision variables on the denominators. This form is not the standard form [i.e., linear programming (LP) or semidefinite programming (SDP) problems] that has already been covered by previous works [4], [6], [10]. With proofs shown in this paper, both system performance and computational tractability are guaranteed under spatial-temporal demand uncertainties. The average performance of the robust taxi dispatch solutions with SOC type of uncertain demand set is better compared with that of the box (range) type of uncertainty set in the evaluations based on data. Hence, it is critical to use a more complex type of uncertainty set, the SOC type, and the corresponding robust dispatch model we design in this paper. The contributions of this paper are as follows.

1) We develop a robust optimization model for taxi dispatch systems under spatial-temporally correlated uncertainties of predicted demand, and define an approximation of the balanced vehicle objective. The robust optimization problem of approximately balancing vacant taxis with least total idle distance is concave of the uncertain demand, convex of the decision variables, and computationally tractable under multiple types of uncertainties.
2) We design a data-driven algorithm to construct uncertainty sets that provide a desired level of probabilistic guarantee for the robust taxi dispatch solutions.
3) We prove that there exist equivalent computationally tractable convex optimization forms for the robust dispatch problem with both polytope and SOC types of uncertainty sets constructed from data.
4) Evaluations on four years of taxi trip data in New York City (NYC) show that the SOC type of uncertain set provides a smaller average dispatch cost than the polytope type. The average demand–supply ratio mismatch is reduced by 31.7%, and the average total idle distance is reduced by 10.13% or about 20 million miles annually with robust dispatch solutions under the SOC type of uncertainty set.

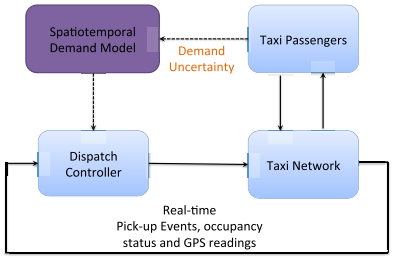


Fig. 1. Prototype of the taxi dispatch system.

The rest of this paper is organized as follows. The taxi dispatch problem is described and formulated as a robust optimization problem given a closed and convex uncertainty set in Section II. We design an algorithm for constructing uncertain demand sets based on taxi operational records data in Section III. Equivalent computationally tractable forms of the robust taxi dispatch problem given different forms of uncertainty sets are proved in Section IV. Evaluation results based on a real data set are shown in Section V. Concluding remarks are provided in Section VI.

II. Problem Formulation

The goal of taxi dispatch is to direct vacant taxis toward current and predicted future requests with a minimum total idle mileage. There are two objectives. One is sending more taxis for more requests to reduce mismatch between supply and demand across all regions in the city. The other is to reduce the total idle driving distance for picking up passengers in order to save cost. Involving predicted future demand when making current decisions benefits to increasing total profits, since drivers are able to travel to regions with better chances to pick up future passengers. In this section, we formulate a taxi dispatch problem with uncertainties in the predicted spatial-temporal patterns of demand. A typical monitoring and dispatch infrastructure is shown in Fig. 1. The dispatch center periodically collects and stores real-time information such as GPS location, occupancy status, and road conditions; dispatch solutions are sent to taxis via cellular radio.

A. Problem Description

For computational efficiency, we assume that the entire city is divided into n regions, and time of one day is discretized to

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

time slots indexed by $t = 1, 2, \ldots, K$. Taxi dispatch decision is calculated in a receding horizon process, since considering future demand when making the current dispatch decisions helps to reduce resource allocating costs [30] and taxis' total idle distance [20]. At time t, we consider the effects of current decision to the following $(t+1, \ldots, t+\tau-1)$ time slots. Only the dispatch solution for time t is implemented and solutions for remaining time slots are not materialized. When the time horizon rolls forward by one step from t to $(t+1)$, information about vehicle locations and occupancy status are observed and updated, and we calculate a new dispatch solution for $(t+1)$.

We define $r_j^k \geq 0$ as the number of total requests within region j during time k, and τ is the model predicting time horizon. We relax the integer constraint of $r_j^k \in \mathbb{N}$ to positive real, since the integer constraint will make the robust dispatch problem in this section not computationally tractable. The total number of requests at region j may have similar patterns as its neighbors, for instance, during busy hours, several downtown regions may all have peak demand. Meanwhile, demand during several consecutive time slots r^k, $k = 1, \ldots, \tau$ are temporally correlated. Typically, it is difficult to predict a deterministic value of passenger demand of a region during specific time. We define the spatial-temporally correlated uncertain demand by one closed and convex, or compact set Δ as

$$r_c = \left[(r^1)^T, (r^2)^T, \cdots, (r^\tau)^T\right]^T \in \Delta \subset \mathbb{R}_+^{\tau n}.$$

Where r_c is called the concatenated demand vector, $(r^k)^T$ means the transpose of r^k. The closed, bounded, and convex form of Δ depends on the method to construct the uncertainty set, which we will describe in detail in Section III. Since r_c depends on r^k, and r^k is one component of r_c, the uncertainty set for demand r^k at time k is defined as a closed, convex set Δ_k, and a projection of Δ

$$\Delta_k := \{r^k \ | \exists r^1, \ldots, r^{k-1}, r^{k+1}, \ldots, r^\tau, \ \text{s.t. } r_c \in \Delta\}.$$

Note that the projection of a convex set onto some of its coordinates are also convex [8, Ch. 2.3.2].

A robust dispatch model that decides the amount of vacant taxis sent between each node pair according to the demand at each node and practical constraints is described in a network flow model of Fig. 2. The edge weight of the graph represents the distance between two regions. Specifically, each region has an initial number of vacant taxis provided by real-time sensing information and an uncertain predicted demand. We define a nonnegative decision variable matrix $X^k \in \mathbb{R}_+^{n\times n}$, $X_{ij}^k \geq 0$, where X_{ij}^k is the number of vehicles dispatched from region i to j. We relax the integer constraint of $X_{ij}^k \in \mathbb{N}$ to a nonnegative real constraint, since mixed integer programming is not computationally tractable with uncertain parameters. Every time when making a resource allocation decision by solving the following robust optimization problem:

$$\min_{X^1} \max_{r^1\in\Delta_1} \min_{X^2} \max_{r^2\in\Delta_2} \ldots \min_{X^\tau} \max_{r^\tau\in\Delta_\tau}$$
$$J = \sum_{k=1}^{\tau} (J_D(X^k) + \beta J_E(X^k, r^k)) \quad \text{s.t.} \quad X^k \in \mathcal{D}_c \qquad (1)$$

where J_D is a convex cost function for allocating resources, J_E is a function concave in r^k and convex in X^k that measures the service fairness of the resource allocating strategy, and $\mathcal{D}_c$ is a convex domain of the decision variables that describes the constraints. We define specific formulations of the objective and constraint functions in the rest of this section.

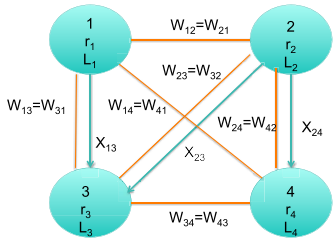


Fig. 2. Network flow model of the robust taxi dispatch problem. A circle represents a region with region ID 1, 2, 3, and 4. We omit the superscript of time k since every parameter is for one time slot only. Uncertain demand is denoted by r_i, L_i is the original number of vacant taxis before dispatch at region i, and X_{ij} is a dispatch solution that sending the number of vacant taxis from region i to j with the distance W_{ij}.

B. Robust Taxi Dispatch Problem Formulation

1) Estimated Cross-Region Idle Driving Distance: When traversing from region i to j, taxi drivers take the cost of cruising on the road without picking up a passenger till the target region. Hence, we consider to minimize this kind of idle driving distance while dispatching taxis. We define the weight matrix of the network in Fig. 2 as $W \in \mathbb{R}^{n\times n}$, where W_{ij} is the distance between region i and j. The across-region idle driving cost according to X^k is

$$J_D(X^k) = \sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}^k W_{ij}. \qquad (2)$$

We assume that the region division method is time-invariant in this paper, and W is a constant matrix for the optimization problem formulation – for instance, the value of W_{ij} represents the length of the shortest path on streets from the center of region i to the center of region j.[1]

The distance every taxi can drive should be bounded by a threshold parameter $m^k \in \mathbb{R}^+$ during limited time

$$X_{ij}^k = 0 \text{ if } W_{ij} > m^k$$

which is equivalent to

$$X_{ij}^k \geq 0, \quad X_{ij}^k W_{ij} \leq m^k X_{ij}^k, \quad \forall i, j \in \{1, \ldots, n\}. \qquad (3)$$

To explain this, assume the constraint (3) holds. If $W_{ij} > m^k$ and $X_{ij}^k > 0$, we have $X_{ij}^k W_{ij} > m^k X_{ij}^k$, which contradicts to (3). The threshold m^k is related to the length of time slot and traffic conditions on streets. For instance, with an estimated average speed of cars in one city during time $k = 1, \ldots, \tau$, and idle driving time to reach a dispatched region is required to be less than 10 min, then the value of m^k should be the distance one taxi can drive during 10 min with the current average speed on road.

[1]For control algorithms with a dynamic region division method, the distance matrix can be generalized to a time-dependent matrix W^k as well.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

2) Metric of Serving Quality: We design the metric of service quality as a function $J_E(X^k, r^k)$ concave in r^k and convex in X^k in this paper for computational efficiency [4]. Besides vacant taxis traverse to region j according to matrix X^k, we define $L_j^k \in \mathbb{R}_+$ as the number of vacant taxis at region j before dispatching at the beginning of time k, and $L^1 \in \mathbb{R}_+^n$ is provided by real-time sensing information. We assume that the total number of vacant taxis is greater than the number of regions, i.e., $N^k \geq n$, and each region should have at least one vacant taxi after dispatch. Then the total number of vacant taxis at region i during time k satisfies that

$$\sum_{j=1}^{n} X_{ji}^k - \sum_{j=1}^{n} X_{ij}^k + L_i^k > 0 \tag{4}$$

$$\sum_{i=1}^{n} \left(\sum_{j=1}^{n} X_{ji}^k - \sum_{j=1}^{n} X_{ij}^k + L_i^k \right) = \sum_{i=1}^{n} L_i^k = N^k. \tag{5}$$

One service metric is fairness, or that the demand–supply ratio of each region equals to that of the whole city. A balanced distribution of vacant taxis is an indication of good system performance from the perspective that a customer's expected waiting time is short as shown by a queuing theoretic model [30]. Meanwhile, a balanced demand–supply ratio means that regions with less demand will get less resources, and idle driving distance will be reduced in regions with more supply than demand, if we preallocate possible redundant supply to those regions in need. We aim to minimize the mismatch value or the total difference between local region demand–supply ratio and the global demand–supply ratio of the whole city, similarly as the objective defined in [20] and [21]

$$\sum_{k=1}^{\tau} \sum_{i=1}^{n} \left| \frac{r_i^k}{\sum_{j=1}^{n} X_{ji}^k - \sum_{j=1}^{n} X_{ij}^k + L_i^k} - \frac{\sum_{j=1}^{n} r_j^k}{N^k} \right|. \tag{6}$$

However, the function (6) is not concave in r^k for any X^k. It is worth noting that we need a function $J_E(X^k, r^k)$ concave in r^k for any X^k, and convex in X^k for any r^k, to make sure the robust optimization problem is computationally tractable. Hence, we define

$$J_E(X^k, r^k) = \sum_{i=1}^{n} \frac{r_i^k}{\left(\sum_{j=1}^{n} X_{ji}^k - \sum_{j=1}^{n} X_{ij}^k + L_i^k \right)^{\alpha}}, \quad \alpha > 0 \tag{7}$$

as a service fairness metric to minimize. This is because we approximately minimize (6) by minimizing (7) under the constraints (4) and (5) with an α value chosen according to the desired approximation level, and the following Lemma explains this approximation.

Lemma 1: Given deterministic demand vectors $(r^1, \ldots, r^\tau)$ and an initial number of vacant vehicles before dispatch $(L^1, \ldots, L^\tau)$ that satisfy $r_i^k \geq 1$, $L_i^k \geq 0$, $\sum_{i=1}^{n} L_i^k = N^k$, for any $\epsilon_0 > 0$, any $i \in \{1, \ldots, n\}$, $k = 1, \ldots, \tau$, there exists an $\alpha > 0$, such that the optimal solution $(X^k)^*$ by minimizing (7) under constraints (4) and (5) satisfies

$$\sum_{k=1}^{\tau} \sum_{i=1}^{n} \left| \frac{r_i^k}{\sum_{j=1}^{n} (X_{ji}^k)^* - \sum_{j=1}^{n} (X_{ij}^k)^* + L_i^k} - \frac{\sum_{j=1}^{n} r_j^k}{N^k} \right| < n\tau\epsilon_0. \tag{8}$$

Proof: See Appendix A. ∎

According to the proof, we can always choose α to be small enough (or close enough to 0) in order to obtain a desired level of approximation ϵ_0. Hence, in the experiments of Section V, we numerically choose $\alpha = 0.1$ based on simulation results. Therefore, with function (7), we map the objective of balancing supply according to demand across every region in the city to a computationally tractable function that concave in the uncertain parameters and convex in the decision variables for a robust optimization problem.

The number of initial vacant taxis L_j^{k+1} depends on the number of vacant taxis at each region after dispatch during time k and the mobility patterns of passengers during time k, while we do not directly control the latter. We define P_{ij}^k as the probability that a taxi traverses from region i to j and turns vacant again (after one or several drop-off events) at the beginning of time $(k+1)$, provided it is vacant at the beginning of k. Methods of getting P_{ij}^k based on data include but not limited to modeling trip patterns of taxis [21] and autonomous mobility on demand systems [30]. Then, the number of vacant taxis within each region j by the end of time k satisfies

$$(L^{k+1})^T = \left(\mathbf{1}_n^T X^k - (X^k \mathbf{1}_n)^T + (L^k)^T \right) P^k. \tag{9}$$

3) Weighted-Sum Objective Function: Since there exists a tradeoff between two objectives, we define a weighted-sum with parameter $\beta > 0$ of the two objectives $J_D(X^k)$ defined in (2) and $J_E(X^k, r^k)$ defined in (7) as the objective function. Let $X^{1:\tau}$ and $L^{2:\tau}$ represent decision variables $(X^1, \ldots, X^\tau)$ and $(L^2, \ldots, L^\tau)$. Without considering model uncertainties corresponding to r^k, a convex optimization form of taxi dispatch problem is

$$\min_{X^{1:\tau}, L^{2:\tau}} \quad J = \sum_{k=1}^{\tau} (J_D(X^k) + \beta J_E(X^k, r^k))$$
$$\text{s.t. (3), (4), (9).} \tag{10}$$

4) Robust Taxi Dispatch Problem Formulation: We aim to find out a dispatch solution robust to an uncertain demand model in this paper. For time $k = 1, \ldots, \tau$, uncertain demand r^k only affects the dispatch solutions of time $(k, k+1, \ldots, \tau)$, and dispatch solution at $k + \tau$ is related to uncertain demand at $(k+1, \ldots, \tau)$, similar to the multistage robust optimization problem in [7]. However, the control laws considered in [7] are polynomial in past-observed uncertainties; in this paper, we do not restrict the decision variables to be any forms of previous-observed uncertain demands. The dispatch decisions are numerical optimal solution of a robust optimization problem. With a list of parameters and variables shown in Table I, considering both the current and future dispatch costs when making the current decisions, we define a robust taxi dispatch

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE I
PARAMETERS AND VARIABLES OF TAXI DISPATCH PROBLEM (11)

Parameters of (11)	Description
n	the number of regions
τ	model predicting time horizon
$r^k \in \Delta_k$	the uncertain total number of requests at each region during time k
$W \in \mathbb{R}^{n\times n}$	weight matrix, W_{ij} is the distance from region i to region j
$P^k \in [0,1]^{n\times n}$	probability matrix that describes taxi mobility patterns during one time slot
$L^1 \in \mathbb{N}^n$	the initial number of vacant taxis at each region provided by GPS and occupancy status data
$m \in \mathbb{R}^+$	the upper bound of distance each taxi can drive idly for picking up a passenger
$\alpha \in \mathbb{R}_+$	the power on the denominator of the cost function
$\beta \in \mathbb{R}_+$	the weight factor of the objective function
Variables of (11)	
$X^k_{ij} \in \mathbb{R}_+$	the number of taxis dispatched from region i to region j during time k
$L^k \in \mathbb{R}^n_+$	the number of vacant taxis at each region before dispatching at the beginning of time k
Parameters of Algorithm 1	
$r_c \in \Delta$	the uncertain concatenated demand vector of τ consecutive time slots
$\tilde{r}_c(d_l, t, I_p)$	one sample of $r_c(t)$ according to sub-dataset I_p, records of date d_l
α_h	significance level of a hypothesis testing

problem as the following:

$$\min_{X^1}\ \max_{r^1\in\Delta_1}\ \min_{X^2,L^2}\ \max_{r^2\in\Delta_2}\ \ldots\ \min_{X^\tau,L^\tau}\ \max_{r^\tau\in\Delta_\tau}$$

$$J=\sum_{k=1}^{\tau}\sum_{i=1}^{n}\left(\sum_{j=1}^{n}X^k_{ij}W_{ij}+\frac{\beta r^k_i}{\left(\sum_{j=1}^{n}X^k_{ji}-\sum_{j=1}^{n}X^k_{ij}+L^k_i\right)^{\alpha}}\right)$$

$$\begin{aligned}\text{s.t.}\quad & (L^{k+1})^T=(\mathbf{1}_n^TX^k-(X^k\mathbf{1}_n)^T+(L^k)^T)P^k,\\ & \mathbf{1}_n^TX^k-(X^k\mathbf{1}_n)^T+(L^k)^T\geq\mathbf{1}_n^T,\\ & X^k_{ij}W_{ij}\leq mX^k_{ij},\\ & X^k_{ij}\geq 0,\quad i,j\in\{1,2,\ldots,n\}.\end{aligned}\tag{11}$$

After getting an optimal solution $(X^1)^*$ of (11), we adjust the solution by rounding methods to get an integer number of taxis to be dispatched toward corresponding regions. It does not affect the optimality of the result much in practice, since the objective or cost function is related to the demand–supply ratio of each region. A feasible integer solution of (11) always exists, since $X^k_{ij}=0,\ \forall i,j,k$ is feasible. Although we cannot provide any theoretical guarantee on the suboptimality of the rounded integer solution, in the numerical experiments the costs under integer solution after rounding and the original real value optimal solution are comparable.

III. ALGORITHM FOR CONSTRUCTING UNCERTAIN DEMAND SETS

With many factors affecting taxi demand during different time within different areas of a city, explicitly describing the model is a strict requirement and errors of the model will affect the performance of dispatch frameworks. Considering the future demand and demand uncertainties, benefits for minimizing the worst case demand–supply ratio mismatch error and idle distance [20], [21]. It is then essential to construct a model that captures the spatial-temporal demand uncertainties and provides a probabilistic guarantee about the vehicle resource allocation cost. We construct demand uncertainty sets via Algorithm 1—getting a sample set of r_c from the original data set and partition the sample set, bootstrapping a threshold for the test statistics according to the requirement of the probability guarantee, and calculating the model of uncertainty sets based on the thresholds.

A. Uncertainty Set With Probabilistic Guarantee

For convenience, we concisely denote all the variables of the taxi dispatch problem as x. Assume that we do not have knowledge about the true distribution $\mathbb{P}^*(r_c)$ of the random demand vector r_c. With the objective function $J(r_c,x)$ of problem (11), the probabilistic guarantee for the event that the true dispatch cost being smaller than the optimal dispatch cost is defined as the following chance constrained problem:

$$\begin{aligned}\min_{x}\quad & M\\ \text{s.t.}\quad & P_{r_c\sim\mathbb{P}^*(r_c)}(f(r_c,x)=J(r_c,x)-M\leq 0)\geq 1-\epsilon.\end{aligned}\tag{12}$$

The constraint f and objective function J are concave in r_c for any x, and convex in x for any r_c. Without loss of generality about the objective and constraint functions, equivalently we aim to find solutions for

$$\begin{aligned}\min_{x}\quad & J(r_c,x)\\ \text{s.t.}\quad & P_{r_c\sim\mathbb{P}^*(r_c)}(f(r_c,x)\leq 0)\geq 1-\epsilon.\end{aligned}\tag{13}$$

When it is difficult to explicitly estimate $\mathbb{P}^*(r_c)$, we solve the following robust problem such that its optimal solutions satisfy the probabilistic guarantee requirement for (13):

$$\min_{x}\ \max_{r_c\sim\Delta}\quad J(r_c,x),\quad \text{s.t.}\quad f(r_c,x)\leq 0.\tag{14}$$

Then r_c of problem (14) can be any vector in the uncertainty set Δ instead of a random vector in (13). The uncertainty set that keeps the optimal solution of (14) satisfying the constraints of problem (13) is defined as the following.

Problem 1: Construct an uncertainty set $\Delta, r_c\in\Delta$, given $0<\epsilon<1$ and samples of random vectors r_c, such that

(P1). The robust constraint (14) is computationally tractable.

(P2). The set Δ implies a probabilistic guarantee for the true distribution $\mathbb{P}^*(r_c)$ of a random vector r_c at level ϵ, that is, for

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

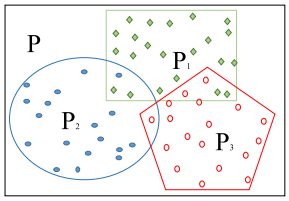


Fig. 3. Intuition for partitioning the whole data set. When the data set includes data from three distributions P_1, P_2, P_3, without prior knowledge, we can build a larger uncertainty set that describes the range of all samples in the data set. The problem is that the uncertainty set is not accurate enough.

any optimal solution $x^* \in \mathbb{R}^k$ and for any function $f(r_c, x)$ concave in r_c, we have the implication as follows:

$$\begin{aligned} &\text{If } f(r_c, x^*) \le 0, \quad \text{for } \forall r_c \in \Delta \\ &\text{then } \mathbb{P}^*_{r_c \sim \mathbb{P}^*(r_c)}(f(r_c, x^*) \le 0) \ge 1 - \epsilon. \end{aligned} \tag{15}$$

The given probabilistic guarantee level ϵ is related to the degree of conservativeness of the robust optimization problem.

B. Aggregating Demand and Partition the Sample Set

Every τ discretized time slots of demand $(r^t, \ldots, r^{t+\tau})$ are concatenated to a vector $r_c(t)$. The first step is to transform the original taxi operational data to a data set of sampled vector $\tilde{r}_c(d, t)$ of different dates d for each index t. For instance, assume we choose the length of each time slot as 1 h, and the data set records all trip information of taxis during each day. According to the start time and GPS coordinate of each pick-up event, we aggregate the total number of pick up events during 1 h at each region to get samples $\tilde{r}_c(d, t)$.

It is always possible to describe the support of the distribution of all samples contained in the data set even they do not follow the same distribution, as explained in Fig. 3. When there is prior knowledge or categorical information to partition the data set into several subsets, we get a more accurate uncertainty set for each subdata set to provide the same probabilistic guarantee level compared with the uncertainty set from the entire data set. Clustering algorithms with categorical information [16] is applicable for data set partition when information besides pick-up events is available, such as weekdays/weekends and weather or traffic conditions. It is worth noting that if the uncertainty sets are built for a categorical information set $\mathcal{I} = \{\mathcal{I}_1, \mathcal{I}_2, \ldots\}$, then for the robust dispatch problems, we require the same set of categories is available in realtime, hence we apply the uncertainty set of $\mathcal{I}_1$ to find solutions when the current situation is considered as $\mathcal{I}_1$.

C. Uncertainty Modeling

The basic idea to define an uncertainty set is to find a threshold for a hypothesis testing that is acceptable with respect to the given data set and a required probabilistic guarantee level, and the formula of an uncertainty set is related to the threshold value of an acceptable hypothesis testing. Given the original data, the null hypothesis H_0, α_h, and the test statistics T, we need to find a threshold that accepts H_0 at significance value α_h for each subset of sampled demand vectors. Since we do not assume that the marginal distribution for every element of vector r_c is independent with each other, we apply two models without any assumptions about the true distribution $\mathbb{P}^*(r_c)$ in the robust optimization literature [6], [11], [25] on the spatial-temporally correlated demand data.

1) Box Type of Uncertainty Demand Sets Built From Marginal Samples: One intuitive description about a random vector is to define a range for each element. For instance, consider the following multivariate hypothesis holds simultaneously for $i = 1, 2, \ldots, \tau n$ with given thresholds $\bar{q}_{i,0}, \underline{q}_{i,0} \in \mathbb{R}$ [11]:

$$\begin{aligned} &H_{0,i} : \inf\left\{t : \mathbb{P}(r_{c,i} \le t) \ge 1 - \frac{\epsilon}{\tau n}\right\} \ge \bar{q}_{i,0} \\ &\inf\left\{t : \mathbb{P}(-r_{c,i} \le t) \ge 1 - \frac{\epsilon}{\tau n}\right\} \ge -\underline{q}_{i,0}. \end{aligned} \tag{16}$$

Assume that we have N_B random samples for each component $r_{c,i}$ of r_c, ordered in increasing value as $r_{c,i}^{(1)}, r_{c,i}^{(2)}, \ldots, r_{c,i}^{(N_B)}$ no matter what is the original sampling order. We define the index s by

$$s = \min\left\{k \in \mathbb{N} : \sum_{j=k}^{N_B} \binom{N_B}{j} \left(\frac{\epsilon}{\tau n}\right)^{N_B - j} \left(1 - \frac{\epsilon}{\tau n}\right)^j \le \frac{\alpha_h}{2\tau n}\right\} \tag{17}$$

and let $s = N_B + 1$ if the corresponding set is empty. The test H_0 is rejected if $r_{c,i}^{(s)} \ge \bar{q}_{i,0}$ or $-r_{c,i}^{(N_B - s + 1)} \ge -\underline{q}_{i,0}$. To construct an uncertainty set, we need an accepted hypothesis test. Hence, we set $\bar{q}_{i,0} = r_{c,i}^{(s)}$ and $\underline{q}_{i,0} = r_{c,i}^{(N_B - s + 1)}$. The following uncertainty set is then applied in this paper based on the range hypothesis testing (16).

Proposition 1 ([6], [11]): If s defined by (17) satisfies that $N_B - s + 1 < s$, then, with probability at least $1 - \alpha_h$ over the sample, the set

$$\mathcal{U}_\epsilon^M(r_c) = \left\{r_c \ge \mathbf{0} | r_{c,i}^{(N_B - s + 1)} \le r_{c,i} \le r_{c,i}^{(s)}\right\} \tag{18}$$

implies a probabilistic guarantee for $\mathbb{P}^*(r_c)$ at level ϵ.

2) SOC Type of Uncertainty Set Motivated by Moment Hypothesis Testing: It is not easy to tell directly from the uncertainty set (18) when the range of one component changes how will others be affected. To directly show the spatial-temporal correlations of the demand, we also apply hypothesis testing related to both the first and second moments of the true distribution $\mathbb{P}^*(r_c)$ of the random vector [25]

$$H_0 : \mathbb{E}^{\mathbb{P}^*}[r_c] = r_0 \text{ and } \mathbb{E}^{\mathbb{P}^*}\left[r_c r_c^T\right] - \mathbb{E}^{\mathbb{P}^*}[r_c]\mathbb{E}^{\mathbb{P}^*}\left[r_c^T\right] = \Sigma_0 \tag{19}$$

where r_0 and Σ_0 are the (unknown) true mean and covariance of r_c, $\mathbb{E}^{\mathbb{P}^*}[r_c]$ and $\mathbb{E}^{\mathbb{P}^*}[r_c r_c^T]$ are the estimated mean and covariance from data. Without knowledge of r_0 and Σ_0, H_0 is rejected when the difference among the estimation of mean or covariance according to multiple times of samples is greater than the threshold, i.e., $\|\mathbb{E}^{\mathbb{P}}[\tilde{r}_c] - \hat{r}_c\|_2 > \Gamma_1^B$ or $\|\mathbb{E}^{\mathbb{P}}[\tilde{r}_c \tilde{r}_c^T] - \mathbb{E}^{\mathbb{P}}[\tilde{r}_c]\mathbb{E}^{\mathbb{P}}[\tilde{r}_c^T] - \hat{\Sigma}\|_F > \Gamma_2^B$, where $\mathbb{E}^{\mathbb{P}}[\tilde{r}]$ is the estimated mean value of one experiment, $\hat{r}_c$ and $\hat{\Sigma}$ are the estimated mean and covariance from multiple experiments,

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Algorithm 1 Constructing Uncertain Demand Sets

Input: A dataset of taxi operational records
1. Demand aggregating and sample set partition
Aggregate demand to get a sample set $\mathcal{S}$ of the random demand vector r_c from the original dataset. Partition the sample set $\mathcal{S}$ and denote a subset $\mathcal{S}(t, I_p) \subset \mathcal{S}$, $p = 1, \ldots, P$ as the subset partitioned for each time index t according to either prior knowledge or categorical information I_p.
2. Bootstrapping thresholds for test statistics
for each subset $\mathcal{S}(t, I_p)$ **do**
Initialization: Testing statistics T, a null-hypothesis H_0, the probabilistic guarantee level ϵ, a significance level $0 < \alpha_h < 1$, the number of bootstrap time $N_B \in \mathbb{Z}_+$.
Estimate the mean $\hat{r}_c(t, I_p)$ **and covariance** $\hat{\Sigma}(t, I_p)$ **for vector** r_c **based on subset** $\mathcal{S}(t, I_p)$**.**
for $j = 1, \ldots, N_B$ **do**
(1). Re-sample $\mathcal{S}^j(t, I_p) = \{\tilde{r}_c(d_1, t, I_p), \ldots, \tilde{r}_c(d_{N_B}, t, I_p)\}$ data points from $\mathcal{S}(t, I_p)$ with replacement for each t.
(2). Get the value of the test statistics based on $\mathcal{S}^j(t, I_p)$.
end for
(3). Get the thresholds of the α significance level for H_0.
end for
3. Calculate the model of uncertainty sets
Get the box type and the SOC type of uncertainty sets according to (18) and (20), respectively, for each t and I_p.
Output: Uncertainty sets for problem (11)

Γ_1^B and Γ_2^B are the thresholds. The remaining problem is then to find the values of the thresholds such that hypothesis testing (19) holds given the data set. The uncertainty set derived based on the moment hypothesis testing is defined in the following proposition.

Proposition 2 ([6], [25]): With probability at least $1 - \alpha_h$ with respect to the sampling, the following uncertainty set $\mathcal{U}_\epsilon^{\text{CS}}(r_c)$ implies a probabilistic guarantee level of ϵ for $\mathbb{P}^*(r_c)$:

$$\mathcal{U}_\epsilon^{\text{CS}}(r_c) = \left\{ {}_c \geq \mathbf{0} | r_c = \hat{r}_c + y + C^T w : \exists y, w \in \mathbb{R}^{n\tau} \right. \\ \left. \text{s.t. } \|y\|_2 \leq \Gamma_1^B, \|w\|_2 \leq \sqrt{\frac{1-\epsilon}{\epsilon}} \right\} \quad (20)$$

where $C^T C = \hat{\Sigma} + \Gamma_2^B \mathbf{I}$ is a Cholesky decomposition.
When one component of r_c increases or decreases, we have an intuition how it affects the value of other components of r_c by the expression (20).

D. Algorithm

With a threshold of the test statistics calculated via the given data set, we then apply the formula (18) for constructing a box type of uncertainty set, and the formula (20) for an SOC type of uncertainty set, respectively. The following Algorithm 1 describes the complete process for constructing uncertain demand sets based on the original data set.

We do not restrict the method of estimating mean $\hat{r}_c(t, I_p)$ and covariance $\hat{\Sigma}(t, I_p)$ matrices of a subset $\mathcal{S}(t, I_p)$ in step 2, and bootstrap is one method. For step 2.(2), the process for the box type of uncertainty sets is: calculate index s that satisfies (17) with the given ϵ, sort each component of sampled vectors $r_c(d_l, t, I_p)$, and get the order statistics $r_{c,i}^{(N_B-s+1)}(j, t, I_p)$, $r_{c,i}^{(s)}(j, t, I_p)$ of the jth sample set $\mathcal{S}^j(t, I_p)$. For the SOC type, we calculate the mean and covariance of the samples of the vector according to the subset $\mathcal{S}^j(t, I_p)$ as $\hat{r}_c(j, t, I_p)$ and $\hat{\Sigma}(j, t, I_p)$, respectively.

In step 2.(3), the α_h level thresholds for the box type of uncertainty sets are the $\lceil N_B(1-\alpha_h) \rceil$th largest value of the upper bound $r_{c,i}^{(s)}(j, t, I_p)$ and the $\lceil N_B \alpha_h \rceil$th largest value of the lower bound $r_{c,i}^{(N_B-s+1)}(j, t, I_p)$ for the ith component. For the SOC type of uncertainty sets, we calculate the mean and covariance of $r_c(t, I_p)$ for the N_B times bootstrap as $\hat{r}_c(t, I_p)$ and $\hat{\Sigma}(t, I_p)$, and get $\Gamma_1(j, t, I_p) = \|\hat{r}_c(j, t, I_p) - \hat{r}_c(t, I_p)\|_2$, $\Gamma_2(j, t, I_p) = \|\hat{\Sigma}(j, t, I_p) - \hat{\Sigma}(t, I_p)\|_2$. Denote the $\lceil N_B(1-\alpha_h) \rceil$th largest value of $\Gamma_1(j, t, I_p)$ and $\Gamma_2(j, t, I_p)$as $\Gamma_1^B(t, I_p)$ and $\Gamma_2^B(t, I_p)$, respectively.

In summary, to construct a spatial-temporal uncertain demand model for problem (11), in this section, we consider the taxi operational record of each day as one independent and identically distributed (i.i.d.) sample for the concatenated demand vector r_c. By partitioning the entire data set to several subsets according to categorical information such as weekdays and weekends, we are able to build uncertainty sets for each subset of data without additional assumptions about the true distribution of the spatial-temporal demand profile. Then we design Algorithm 1 to construct a box type and an SOC type of uncertainty sets based on data that provide a desired probabilistic guarantee of robust solutions.

IV. Computationally Tractable Formulations

We build equivalent computationally tractable formulations of problem (11) with different definitions of uncertain sets calculated by Algorithm 1 in this section. Hence, the robust taxi dispatch problem considered in this paper can be solved efficiently. Computational tractability of a robust LP problem for ellipsoid uncertainty sets is discussed in [4]. The process is to reformulate constraints of the original problem to its equivalent convex constraints that must hold given the uncertainty set. The objective function of problem (11) is concave of the uncertain parameters r^k, convex of the decision variables X^k, L^k with the decision variables on the denominators, not standard forms of LP or SDP problems that already covered by previous works [4], [6]. Hence, we prove one equivalent computationally tractable form of problem (11) for each uncertainty set constructed in Section III.

Only the J_E components of objective functions in (11) include uncertain parameters, and the decision variables of the function are in the denominator of the function J_E. The box-type uncertainty set defined as (18) is a special form of polytope; hence, we first prove an equivalent standard form of convex optimization problem for (11) for a polytope uncertainty set as the following.

Theorem 1 (Next Step Dispatch): If the uncertainty set of problem (11) when $\tau = 1$ is defined as the nonempty polytope $\Delta := \{r \geq 0, Ar \leq b\}$, and we omit the superscripts k

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

for variables and parameters without confusion. Then problem (11) with $\tau = 1$ is equivalent to the following convex optimization problem:

$$\begin{aligned}\min_{X\geq 0,\lambda\geq 0}\quad & \sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}W_{ij} + b^T\lambda \\ \text{s.t.}\quad & A^T\lambda - \beta\begin{bmatrix} \frac{1}{\left(\sum_{j=1}^{n} X_{j1} - \sum_{j=1}^{n} X_{1j} + L_1\right)^{\alpha}} \\ \vdots \\ \frac{1}{\left(\sum_{j=1}^{n} X_{jn} - \sum_{j=1}^{n} X_{nj} + L_n\right)^{\alpha}} \end{bmatrix} \geq 0 \\ & \mathbf{1}_n^T X - X\mathbf{1}_n + L^T \geq 1, \\ & X_{ij}W_{ij} \leq mX_{ij}, \\ & X_{ij} \geq 0, \quad \forall i, j \in \{1,\ldots,n\}. \end{aligned} \tag{21}$$

□

Proof: See Appendix B. ■

To directly use the demand uncertainty set that describes the spatial-temporal correlation of $(r^1,\ldots,r^\tau)$ like (18) and (20) for the concatenated demand r_c in problem (11), we first consider to group the maximization over each r^k together to save the process of projection $r_c \in \Delta$ for individual $r^k \in \Delta_k$. Furthermore, we can find the dual (a minimizing problem) of the maximizing cost problem over $r_c \in \Delta$, and then numerically efficiently solve (11) that minimizes the total cost during time $(1,2,\ldots,\tau)$ under uncertain demand r_c. Hence, we first prove that the minimax equality holds for the maximin problem over each pair of k and $k+1$ for problem (11), and (11) is equivalent to the robust optimization problem shown in the following lemma.

Lemma 2 (Minimax Equality): Given the assumption that the definition of the uncertainty sets $r_c \in \Delta$ and $r^k \in \Delta_k$ is compact (closed and convex), the robust dispatch problem (11) is equivalent to the following robust dispatch problem:

$$\begin{aligned}\min_{X^{1:\tau},L^{2:\tau}}\ \max_{r_c\in\Delta}\quad & J = \sum_{k=1}^{\tau}(J_D(X^k) + \beta J_E(X^k, r^k)) \\ \text{s.t.}\quad & \text{constraints of (11)}, k = 1,\ldots,\tau. \end{aligned} \tag{22}$$

□

Proof: See Appendix C. ■

For the robust optimization problem (11), the computationally tractable convex form depends on the definition of uncertainty sets.When conditions of Lemma 2 hold, equivalent convex optimization forms of problem (11) are derived based on problem (22). For a multistage robust optimization problem that restricts the near-optimal control input of linear dynamical systems to be a certain degree of polynomial of previous-observed uncertainties, an approximated SDP method for calculating the time dependent control input is proposed in [7]. The method does not require minimax equality holds for the robust optimal control problem.

The box-type uncertainty set (18) is a special form of polytope that the uncertain demand model during different time of a day is described separately. The process of converting problem (11) to an equivalent computationally tractable convex form is similar to that of the one-stage robust optimization problem. The result is described as the following lemma.

Lemma 3: If the uncertain set for $r^k, k = 1,\ldots,\tau$ describes each demand vector r^k separately as a nonempty polytope with the form

$$\Delta_k := \{r^k \geq 0, A_k r^k \leq b_k\}, \quad k = 1,\ldots,\tau \tag{23}$$

problem (11) is equivalent to the following convex optimization problem:

$$\begin{aligned}\min_{X^k,\lambda^k,L^k\geq 0}\quad & \sum_{k=1}^{\tau}\left(\sum_{i}^{n}\sum_{j}^{n} X_{ij}^k W_{ij} + b_k^T\lambda^k\right) \\ \text{s.t.}\quad & A_k^T\lambda^k - \beta\begin{bmatrix} \frac{1}{\left(\sum_{j=1}^{n} X_{j1}^k - \sum_{j=1}^{n} X_{1j}^k + L_1^k\right)^{\alpha}} \\ \vdots \\ \frac{1}{\left(\sum_{j=1}^{n} X_{jn}^k - \sum_{j=1}^{n} X_{nj}^k + L_n^k\right)^{\alpha}} \end{bmatrix} \geq 0 \\ & \text{constraints of (11)}, \quad k = 1,\ldots,\tau. \end{aligned} \tag{24}$$

□

Proof: See Appendix D1. ■

For a more general case that the uncertainty sets for $r^1,\ldots,r^\tau$ are temporally correlated, the following theorem and proof describe the equivalent computationally tractable convex form of (11).

Theorem 2: When Δ is defined as the following nonempty polytope set:

$$\Delta := \{(\Delta_1,\ldots,\Delta_\tau) | A_1 r^1 + \cdots + A_\tau r^\tau \leq b, r^k \geq 0\} \tag{25}$$

problem (11) is equivalent to the following convex optimization problem:

$$\begin{aligned}\min_{X^k,L^k,\lambda\geq 0}\quad & \sum_{k=1}^{\tau}\left(\sum_{i}^{n}\sum_{j}^{n} X_{ij}^k W_{ij}\right) + b^T\lambda \\ \text{s.t.}\quad & A_k^T\lambda - \beta\begin{bmatrix} \frac{1}{\left(\sum_{j=1}^{n} X_{j1}^k - \sum_{j=1}^{n} X_{1j}^k + L_1^k\right)^{\alpha}} \\ \vdots \\ \frac{1}{\left(\sum_{j=1}^{n} X_{jn}^k - \sum_{j=1}^{n} X_{nj}^k + L_n^k\right)^{\alpha}} \end{bmatrix} \geq 0 \\ & \text{constraints of (11)}, \quad k = 1,\ldots,\tau. \end{aligned} \tag{26}$$

□

Proof: See Appendix D2. ■

With an uncertain demand model defined as (20) for concatenated $r^1,\ldots,r^\tau$, the following theorem derives the equivalent computationally tractable form of problem (11).

Theorem 3: When the uncertainty set for $r^1,\ldots,r^\tau$ is defined as the SOC form of (20), problem (11) is equivalent to the following convex optimization problem (27):

$$\begin{aligned}\min_{X^k,L^k,z}\quad & \sum_{k=1}^{\tau}\sum_{i}^{n}\sum_{j}^{n} X_{ij}^k W_{ij} \\ & + \beta\left(\hat{r}_c^T z + \Gamma_1^B\|z\|_2 + \sqrt{\frac{1}{\epsilon} - 1}\|Cz\|_2\right) \\ \text{s.t.}\quad & c_l(X) \leq z, \\ & \text{constraints of (11)}, \quad k = 1,\ldots,\tau \end{aligned} \tag{27}$$

where $c_l(X) \in \mathbb{R}^{\tau n}$ is the concatenation of $c(X^1),\ldots,c(X^\tau)$.

Proof: See Appendix E. ■

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE II
NYC DATA IN THE EVALUATION SECTION

Taxi Trip Data set			Format		
Collection Period	Data Size	Record Number	ID	Trip Time	Trip Location
01/01/2010-12/31/2013	$100GB$	about 7 million	Date	Start and end time	GPS coordinates of start and end

It is worth noting that any optimal solution for problem (10) has a special form between any pair of regions (i, q).

Proposition 3: Assume $X^{1*}, \ldots, X^{\tau*}$ is an optimal solution of (10), then any X^{k*} satisfies that for any pair of (p, q), at least one value of the two elements X^{k*}_{qi} and X^{k*}_{iq} is 0.

Proof: We prove by contradiction. Assume that one optimal solution has the form X^k such that $X^k_{qi} > 0$ and $X^k_{iq} > 0$. Without loss of generality, we assume that $X^k_{qi} \geq X^k_{iq}$, and let

$$X^{k*}_{qi} = X^k_{qi} - X^k_{iq}, \quad X^{k*}_{iq} = 0$$

other elements of X^{k*} equal to X^k. Then

$$\begin{aligned}\sum_{j=1}^{n} X^k_{ji} - \sum_{j=1}^{n} X^k_{ij} &= X^k_{qi} - X^k_{iq} + \sum_{j\neq q} X^k_{ji} - \sum_{j\neq q} X^k_{ij} \\ &= X^{k*}_{qi} + 0 + \sum_{j\neq q} X^{k*}_{ji} - \sum_{j\neq q} X^{k*}_{ij} \\ &= \sum_{j=1}^{n} X^{k*}_{ji} - \sum_{j=1}^{n} X^{k*}_{ij} \\ \sum_{j=1}^{n} X^k_{ji} - \sum_{j=1}^{n} X^k_{ij} + L^k_i &= \sum_{j=1}^{n} X^{k*}_{ji} - \sum_{j=1}^{n} X^{k*}_{ij} + L^k_i.\end{aligned}$$

Hence, we have $J_E(X^k, r^k) = J_E(X^{k*}, r^k)$. All constraints are satisfied and X^{k*} is also a feasible solution for (11).

Next, we compare $J_D(X^k)$ and $J_D(X^{k*})$. With $X^k_{qi} > X^k_{iq} > 0$, and $X^{k*}_{qi} = X^k_{qi} - X^k_{iq} \geq 0$, we have

$$X^k_{qi} > X^{k*}_{qi}, \; X^k_{qi} W_{qi} + X^k_{iq} W_{iq} > X^{k*}_{qi} W_{qi} + X^{k*}_{iq} W_{iq}.$$

Thus the partial cost $J_D(X^k) > J_D(X^{k*})$, which contradicts with the assumption that X^k is an optimal solution. To summarize, we show that an optimal solution cannot have $X^k_{qi} > 0, X^k_{iq} > 0$ at the same time, and at least one of X^{k*}_{qi} and X^{k*}_{iq} should be 0. ■

With equivalent convex optimization forms under different uncertainty sets, robust taxi dispatch problem (11) is computationally tractable and solved efficiently.

V. DATA-DRIVEN EVALUATIONS

We conduct data-driven evaluations based on four years of taxi trip data of NYC [12]. A summary of this data set is shown in Table II. In this data set, every record represents an individual taxi trip, which includes the GPS coordinators of pick-up and drop-off locations, and the date and time (with precision of seconds) of pick-up and drop-off locations. The dispatch solutions based on different granularities of equal-area region partitions have been compared in [20], and other region partition methods are discussed in [18]. In the following experiments, we use equal-area grid partition since it is a baseline, and compare the robust and nonrobust solutions based on the same region partition method. One partition example given the map of Manhattan area is shown in Fig. 4, where we visualize the density of taxi passenger demand with the data we use for large-scale data-driven evaluations. The lighter the region, the higher the daily demand density, and the middle regions typically have higher density than the uptown and downtown regions. We construct uncertainty sets according to Algorithm 1, discuss factors that affect modeling of the uncertainty set, and compare optimal costs of the robust dispatch formulation (11) and the nonrobust optimization form (10) in this section.



Fig. 4. Map of Manhattan area in NYC.

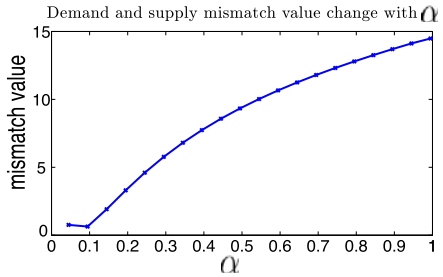


Fig. 5. Comparison of demand and supply mismatch values defined as (6) with different solutions for minimizing J_E defined in (7) with α in range (0, 1]. The value of function (6) under an optimal solution of J_E is smaller with an α closer to 0, which means the dispatch solution tends to be more balanced throughout the entire city.

How Vacant Taxis Are Balanced Across Regions With Different α Values: Fig. 5 shows mismatch between supply and demand defined as (6) for different optimal solutions of minimizing J_E defined in (7) for $\alpha \in (0, 1]$. With α closer to 0, the optimal value of (6) is smaller. We choose $\alpha = 0.1$ for calculating optimal solutions of (10) and (11) in this section.

A. Box Type of Uncertainty Set

For all box type of uncertainty sets shown in this section with the model described in Section III-C1, we set the confidence level of hypothesis testings as $\alpha_h = 10\%$, bootstrap time as $N_b = 1000$, and the number of randomly sampled data (with replacement) for each time of bootstrap as $N_B = 10\,000$.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

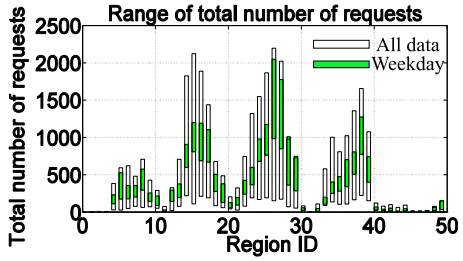


Fig. 6. Comparison of box type of uncertainty sets constructed from all data and those constructed only based on trip records of weekdays. When keeping all parameters the same, by applying data of weekdays, the range of uncertainty set for each $r_{c,i}$ is smaller than that based on the whole data set.

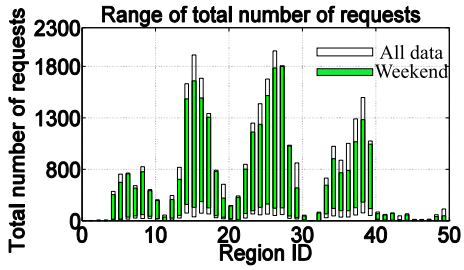


Fig. 7. Comparison of box type of uncertainty sets constructed from all data and uncertainty sets constructed only based on trip records of weekends.

1) Partitioned Data Set Compared With Nonpartitioned Data Set: We show the effects of partitioning the trip record data set by weekdays and weekends in Figs. 6 and 7. The whole city is partitioned into 50 regions, the prediction time horizon is $\tau = 4$, where one time instant means 1 h, $\epsilon = 0.3$, and every $r_c \in \mathbb{R}^{200\times 1}$. Figs. 6 and 7 show the lower and upper bounds of each region during one time slot of (18). By applying data of weekdays and weekends separately, the range $[r_{c,i}^{(s)},\ r_{c,i}^{(N_B-s+1)}]$ of each component is reduced. To get a measurement of the uncertainty level, we defined the sum of range of every component for r_c as

$$U(r_c) = \sum_{i=1}^{\tau n} \left(r_{c,i}^{(s)} - r_{c,i}^{(N_B-s+1)}\right).$$

For the box type of uncertainty sets, when values of the dimension of r_c, i.e., τn, α_h, and ϵ are fixed, a smaller $U(r_c)$ means a smaller area of the uncertainty set, or a more accurate model. We denote $U(r_c)$ calculated via records of weekdays and weekends as $U_{wd}(r_c)$ and $U_{wn}(r_c)$, respectively, compared with $U(r_c)$ constructed from the complete data set, we have $[U(r_c) - U_{wd}(r_c)/U(\hat{r}_c)] = 52\%$, $[U(r_c) - U_{wn}(r_c)/U(\hat{r}_c)] = 28\%$. This result shows that when by constructing an uncertainty set for each subset of partitioned data, we reduce the range of uncertainty sets to provide the same level of probabilistic guarantee for the robust dispatch problem. This is because samples contained in each subset of data do not follow the same distribution and can be categorized as two clusters.

2) Choose an Appropriate N_B for High-Dimensional r_c: It is worth noting that the index s affects the range selection for every component $r_{c,i}$; hence, for different values of $\alpha_h, \epsilon, \tau, n$, we should adjust the number of samples N to get an accurate estimation of the marginal range. As shown in Table III, N needs to be large enough for a large τn value, or s is too close to N and the upper and lower bounds $r_{c,i}^{(N_B-s+1)}$, $r_{c,i}^{(s)}$ cover almost the whole range of samples. Hence, the box-type uncertainty set is not a good choice for large τn value, though the computational cost of solving problem (26) is smaller than that of (27) with the same size of τn.

TABLE III

VALUE OF INDEX s FOR THE BOX-TYPE UNCERTAINTY SET (17). FOR LARGE τn, N_B NEEDS TO BE LARGE, OR s IS TOO CLOSE TO N_B THAT THE RANGE COVERS VALUES OF ALMOST ALL SAMPLES

N_B	α_h	ϵ	n	τ	s
10000	0.1	0.2	50	2	9992
10000	0.1	0.5	50	2	9970
10000	0.3	0.2	50	2	9991
10000	0.1	0.2	1000	2	9999
10000	0.1	0.5	1000	2	9999

TABLE IV

COMPARING THRESHOLDS WITH AND WITHOUT DISCRIMINATING WEEKDAYS AND WEEKENDS DATA. WHEN Γ_1^B OR Γ_2^B IS SMALLER, THE VOLUME OF THE UNCERTAINTY SET IS SMALLER. HERE $n = 1000$, $\tau = 3$, $N_B = 1000$, $\epsilon = 0.3$, AND $\alpha_h = 0.2$

Data type	Weekdays	Weekends	Non partitioned
Γ_1^B	10.53	13.84	17.96
Γ_2^B	2576.94	2923.35	3864.47

B. SOC Type of Uncertainty Set

The SOC type of uncertainty set is a high-dimensional convex set that is not able to be plotted. The bootstrapped thresholds for the hypothesis testing to construct the SOC uncertainty sets based on partitioned and nonpartitioned data are summarized in Table IV. Similarly as the box type of uncertainty sets, when we separate the data set and construct an uncertainty demand model for weekdays and weekends, respectively, the sets are smaller compared to the uncertain demand model for all dates. When α and ϵ values are fixed, with smaller Γ_1^B and Γ_2^B, the demand model $\mathcal{U}_\epsilon^{\text{CS}}$ is more accurate to guarantee that with at least probability $1-\epsilon$, the constraints of the robust dispatch problems are satisfied. Numerical results of this conclusion are shown in Table IV.

1) How n and τ Affect the Accuracy of Uncertainty Sets: For a box type of uncertainty set, when τn is a large value, the bootstrap sample number N_B should be large enough such that index s is not too close to N. Without a large enough sample set, we choose to construct an SOC type of uncertainty set (such as $\tau n = 1000$ and $N_B = 10\,000$ in Table V). Since SOC captures more information about the second moment properties of the random vector compared with the box-type uncertainty set, some uncorrelated components of r_c will be reflected by the estimated covariance matrix, and the volume of the uncertainty set will be reduced. We show the value of Γ_1^B and Γ_2^B with different dimensions of r_c or τn values in Table V. When increasing the value of τn, values of Γ_1^B and Γ_2^B are reduced, which means the uncertainty set is smaller. However, it is not helpful to reduce the granularity of region

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE V

COMPARING THRESHOLDS OF SOC UNCERTAINTY SETS FOR DIFFERENT DIMENSIONS r_c, BY CHANGING EITHER THE REGION PARTITION NUMBER n OR THE PREDICTION TIME HORIZON τ

	Γ_1^B	Γ_2^B
$n = 50, \tau = 1$	42.37	1.52×10^5
$n = 50, \tau = 3$	52.68	4.29×10^4
$n = 50, \tau = 6$	107.35	8.23×10^5
$n = 10, \tau = 3$	71.35	3.56×10^5
$n = 1000, \tau = 3$	10.53	2576.94

partition to a smaller than street level, since we construct the model for a robust dispatch framework and a too large n is not computationally efficient for the dispatch algorithm.

C. Compare Robust Solutions With Nonrobust Solutions

In the experiments, the idle geographical distance of one taxi between a drop-off event and the following pick-up event is approximated as one norm distance between the 2-D geographical coordinates (provided as longitude and latitude values of GPS data in the data set) of the two points. Then the corresponding idle miles on ground are converted from the geographical distance according to the geographical coordinates of NYC. To test the quality of the uncertainty sets applied in the robust dispatch problems, we use the idea of cross-validation from machine learning. The data set is separated as a training set for building the uncertain demand model, and a testing set for comparing the results of the dispatch solutions. The customer demand models applied in the robust and nonrobust optimization problems are different. For the nonrobust dispatch problem, the demand prediction r^k is a deterministic vector. For instance, in this paper, we use the average or mean of the bootstrapped value of the training data set. The nonrobust dispatch solution for each time k is calculated by solving the convex optimization form of dispatch problem formulated in [20] and [21], with deterministic demand model. For all the experiments, we let $\beta = 10$ and $\alpha = 0.1$ in problem (11) to calculate the optimal solutions.

In the robust dispatch problem, the penalty function directly includes the uncertain demand that r^k is for violating a balanced demand–supply ratio requirement. For each testing data r^k, we denote the demand–supply ratio mismatch error of a dispatch solution as (6). We then compare the value of (6) of robust dispatch solutions with the SOC type of uncertainty set constructed in this paper with the value of (6) of nonrobust solutions of testing samples. The distribution of values is shown in Fig. 8. The average demand–supply ratio error is reduced by 31.7% with robust solutions. We compare the cost distribution of total idle distance in Fig. 9. It shows that the average total idle distance is reduced by 10.13%. For all testing, the robust dispatch solutions result in no idle distance greater than 0.8×10^5, and nonrobust solutions has 48% of samples with idle distance greater than 0.8×10^5. The cost of robust dispatch (11) is a weighted sum of both the demand–supply ratio error and estimated total idle driving distance, and the average cost is reduced by 11.8% with robust solutions. It is worth noting that the cost is calculated based on the integer vehicle dispatch solution after rounding the real value optimal solution of (11), and the cost is only 1% higher than the optimal cost of (11). The performance of the system is improved when the true demand deviates from the average historical value considering model uncertainty information in the robust dispatch process. It is worth noting that the number of total idle distance shown in this figure is the direct calculation result of the robust dispatch problem. When we convert the number to an estimated value of corresponding miles in one year, the result is a total reduction of 20 million miles in NYC.

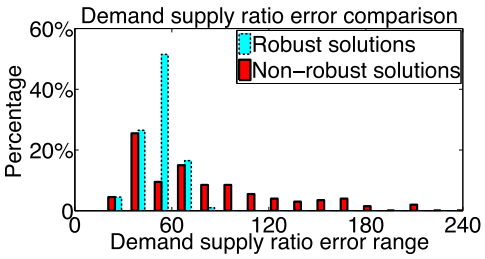


Fig. 8. Demand–supply ratio error distribution of the robust optimization solutions with the SOC type of uncertain demand set ($\epsilon = 0.25$, or probabilistic guarantee level 75%) and nonrobust optimization solutions. The demand–supply ratio error of robust solutions is smaller than that of the nonrobust solutions that the average demand–supply ratio error is reduced by 31.7%.

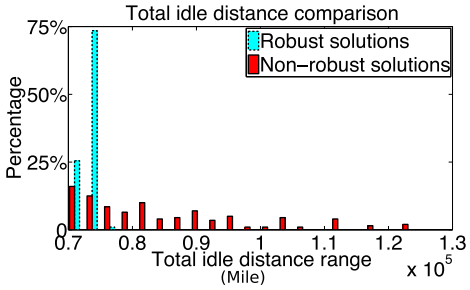


Fig. 9. Total idle distance comparison of robust optimization solutions with the SOC type of uncertain demand set ($\epsilon = 0.25$, or probabilistic guarantee level 75%) and nonrobust optimization solutions. The average total idle distance is reduced by 10.13%. For all samples used in testing, the robust dispatch solutions result in no idle distance greater than 0.8×10^5, and nonrobust solutions has 48% of samples with idle distance greater than 0.8×10^5. The number of total idle distance shown in this figure is the direct calculation result of the robust dispatch problem, and we convert the number to an estimated value of corresponding miles in one year, the result is a total reduction of 20 million miles in NYC.

1) Check Whether the Probabilistic Level ϵ Is Guaranteed: Theoretically, the optimal solution of the robust dispatch problems with the uncertainty set should guarantee that with at least the probability $(1 - \epsilon)$, when the system applies the robust dispatch solutions, the actual dispatch cost under a true demand is smaller than the optimal cost of the robust dispatch problem. Figs. 10 and 11 show the cross-validation testing result that the probabilistic guarantee level is reached for both box type and SOC type of uncertainty sets via solving (26) and (27), respectively. Comparing these two figures, one key insight is that the robust dispatch solution with an SOC type uncertainty set provides a tighter bound on the probabilistic guarantee level that can be reached under the

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

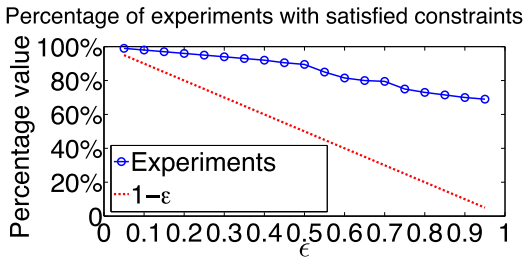


Fig. 10. Percentage of tests that have a smaller true dispatch cost than the optimal cost of the robust dispatch problem with the box-type uncertainty set constructed from data. When $1-\epsilon$ decreases, the percentage value also decreases, but always greater than $1-\epsilon$.

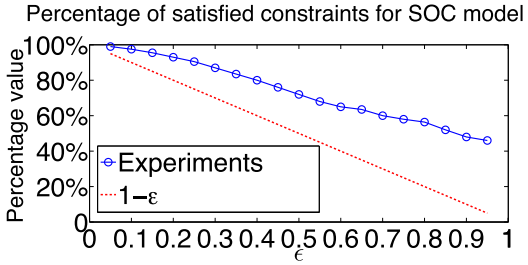


Fig. 11. Percentage of tests that have a smaller true dispatch cost than the optimal cost of the robust dispatch problem with the SOC type of uncertainty set. When $1-\epsilon$ decreases, the percentage value also decreases, but always greater than $1-\epsilon$. The true percentage value is closer to the value of $1-\epsilon$ compared with the solution given a box-type uncertainty set.

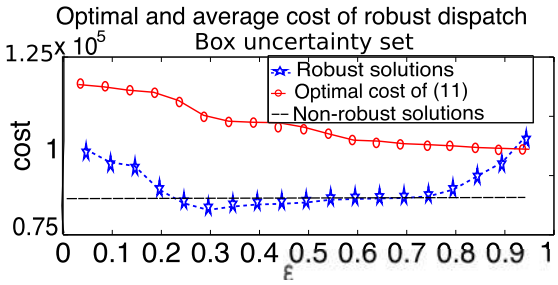


Fig. 12. Comparison of the optimal cost of the robust dispatch problem with box type of uncertainty set and the average cost when applying the robust solutions for the test subset of sampled r_c. When $\epsilon = 0.3$ the average cost is the smallest.

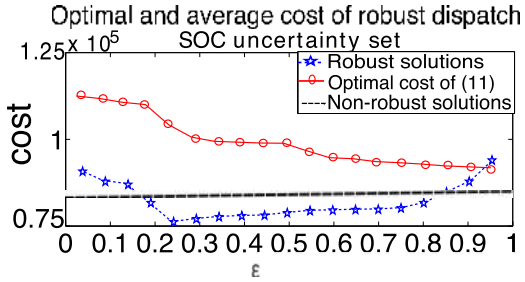


Fig. 13. Comparison of the optimal cost of the robust dispatch problem with SOC type of uncertainty set and the average cost when applying the robust solutions for the test subset of sampled r_c. When $\epsilon = 0.25$ the average cost is the smallest.

true random demand compared with solutions of the box-type uncertainty set. It shows the advantage of considering second-order moment information of the random vector, though the computational cost is higher to solve problem (27) than to solve problem (26).

2) How Probabilistic Guarantee Level Affects the Average Cost: There exists a tradeoff between the probabilistic guarantee level and the average cost with respect to a random vector r_c. Selecting a value for ϵ is case by case, depending on whether a performance guarantee for the worst case scenario is more important or the average performance is more important. For a high probabilistic guarantee level or a large $1-\epsilon$ value, the average cost may not be good enough since we minimize a worst case that rarely happens in the real world. When $(1-\epsilon)$ is relatively small, the average cost can also be large since many possible values of the random vector are not considered.

We compare the optimal cost of robust solutions and the average cost of empirical tests for two types of uncertainty sets via solving (26) and (27) in Fig. 12 and 13, respectively. The optimal cost shows that the result of minimized worst case scenario for all possible r_c included in the uncertainty set, and the average cost shows the empirical testing cost when we applying the optimal solution to dispatch taxis under random testing data of demand r_c. The horizontal line shows the average cost of nonrobust solutions that are not related to ϵ. The ϵ values that provide the best average costs are not exactly the same for different types of uncertainty sets according to the experiments. For the box type of uncertainty set in Fig. 12, $\epsilon = 0.3$ provides the smallest average experimental cost; and for SOC type of uncertainty set in Fig. 13, $\epsilon = 0.25$ provides the smallest average cost. The minimum average cost of an SOC robust dispatch solution is smaller than that of a box type. It indicates that the second-order moment information of the random variable should be included for modeling the uncertainty set and calculating robust dispatch solutions, though its computational cost is higher.

VI. Conclusion

In this paper, we develop a multistage robust optimization model considering demand model uncertainties in taxi dispatch problems. We model spatial-temporal correlations of the uncertainty demand by partitioning the entire data set according to categorical information, and applying theories without assumptions on the true distribution of the random demand vector. We prove that an equivalent computationally tractable form exist with the constructed polytope and SOC types of uncertainty sets, and the robust taxi dispatch solutions are applicable for a large-scale transportation system. A robust dispatch formulation that purely minimizes the worst case cost under all possible demand usually sacrifices the average system performance. The robust dispatch method we design allows any probabilistic guarantee level for a minimum cost solution, considering the tradeoff between the worst case cost and the average performance. Evaluations show that under the robust dispatch framework we design, the average demand–supply ratio mismatch error is reduced by 31.7%, and the average total idle driving distance is reduced by 10.13% or about 20 million miles in total in one year. In the future, we will enhance problem formulation considering more uncertain characteristics of taxi network model, like traffic conditions.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

APPENDIX

A. Proof of Lemma 1

Proof: We first consider the problem of minimizing $\sum_{i=1}^{n}|(r_i^k/\sum_{j=1}^{n}X_{ji}^k-\sum_{j=1}^{n}X_{ij}^k+L_i^k)-(\sum_{j=1}^{n}r_j^k/N^k)|$ for one time slot k

$$\sum_{j=1}^{n}X_{ji}^k-\sum_{j=1}^{n}X_{ij}^k+L_i^k=b_i^k,\quad i=1,\ldots,n. \tag{28}$$

Given a vector L^k that satisfies $L_i^k\geq 0$, $\sum_{i=1}^{k}=N^k$, we have $\sum_{i=1}^{n}b_i^k=\sum_{i=1}^{n}L_i^k=N^k$, since balancing vacant vehicles does not change the total number of vacant vehicles in the city.

To explain how (7) approximates (6) under constraints (4) and (5), consider the following problem given $r_1^k,\ldots,r_n^k$, $N^k=c$:

$$\underset{b_i^k>0,\sum_i b_i^k=c}{\text{minimize}}\ \sum_i\frac{r_i^k}{\left(b_i^k\right)^{\alpha}},\quad c \text{ is a constant.} \tag{29}$$

We substitute $b_n^k=c-b_1^k\cdots-b_{n-1}^k$ into (29), and take partial derivatives of $\sum_i(r_i^k/(b_i^k)^{\alpha})$ over $b_i^k, i=1,\ldots,n-1$. When the minimum of (7) is achieved, each partial derivative should be 0, $-\alpha(r_i^k/(b^k)_i^{\alpha+1})-\alpha(-1)(r_n^k/(c-b_1^k\cdots-b_{n-1}^k)^{\alpha+1})=0$, which is equivalent to $(r_1^k/(b_1^k)^{\alpha+1})=\cdots=(r_{n-1}^k/(b_{n-1}^k)^{\alpha+1})=(r_n^k/(b_n^k)^{\alpha+1})$.

Let $(r_1^k/(b_1^k)^{\alpha+1})=\cdots=(r_{n-1}^k/(b_{n-1}^k)^{\alpha+1})=(r_n^k/(b_n^k)^{\alpha+1})=c_0$, $\gamma=(1/\alpha+1)$, when $\alpha>0$, $0<\gamma<1$. Assume that $\sum_{i=1}^{n}r_i^k=a$, then

$$\left(r_1^k\right)^{\gamma}=b_1^kc_0,\quad\ldots,\quad\left(r_n^k\right)^{\gamma}=b_n^kc_0$$

$$\sum_{i=1}^{n}\left(r_i^k\right)^{\gamma}=\left(b_1^k+\cdots+b_n^k\right)c_0=cc_0\Rightarrow c_0=\frac{1}{c}\sum_{i=1}^{n}\sum_{i=1}^{n}\left(r_i^k\right)^{\gamma}$$

$$\left(r_i^k\right)^{\gamma}=\frac{b_i}{c}\sum_{j=1}^{n}\left(r_j^k\right)^{\gamma},\quad\frac{r_i^k}{b_i^k}=\frac{\left(a_i^k\right)^{1-\gamma}}{c}\sum_{j=1}^{n}\left(r_j^k\right)^{\gamma}.$$

We would like to prove that for any $\epsilon_0>0$, any $i\in\{1,\ldots,n\}$, there exists a $0<\gamma<1$, such that

$$\left|\frac{\left(r_i^k\right)^{1-\gamma}}{c}\sum_{j=1}^{n}\left(r_j^k\right)^{\gamma}-\frac{a}{c}\right|<\epsilon_0. \tag{30}$$

To prove (30), it is worth noting that for any given values of $r_i^k\geq 1, i=1,\ldots,n, c>0$, function $f_i(\gamma)=[(r_i^k)^{1-\gamma}/c]\sum_{j=1}^{n}(r_j^k)^{\gamma}$ is a continuous function of γ, and $f_i(\gamma=1)=(a/c)$ for any i. Then, for any $\epsilon_0>0$ and any (i,k), there exists a $\delta_i^k>0$, such that

$$|\gamma-1|<\delta_i^k\Rightarrow\left|\frac{\left(r_i^k\right)^{1-\gamma}}{c}\sum_{j=1}^{n}\left(r_j^k\right)^{\gamma}-\frac{a}{c}\right|<\epsilon_0.$$

Then let $\delta=\min\{\delta_1^1,\delta_2^1,\ldots,\delta_n^k\}$ (when ϵ_0 is small, δ indicates a small range, so $0<\delta<1$), then for any γ in the range $1-\delta<\gamma<1$, the inequality (30) holds for all k. Without loss of generality, let $\gamma=1-0.5\delta$, $\alpha=(2/2-\delta)$, then the optimal solution of problem (29) is $b_i^k=(cr_i^k/(r_i^k)^{0.5\delta}\sum_{j=1}^{n}(r_j^k)^{1-0.5\delta})$, $i=1,\ldots,n$.

It is worth noting that given any values of $b_1^k>0,\ldots,b_n^k>0$, $L_1^k\geq 0,\ldots,L_n^k\geq 0$ that satisfies $\sum_{i=1}^{n}b_i^k=\sum_{i=1}^{n}L_i^k$, the equation set (28) has a feasible solution for $n\times n$ variables of the matrix X^k. This can be checked by vectorizing matrix X^k to a vector $Y^k\in\mathbb{R}^{n^2}$ and transforming equation set (28) to a new equation set of Y^k. We get a homogeneous equation set with n equations and $n\times n$ variables of Y^k, which always has a feasible solution. Hence, we plug in the values of $b_i^k=(cr_i^k/(a_i^k)^{0.5\delta}\sum_{j=1}^{n}(r_j^k)^{1-0.5\delta})$ to (28) to get values of X_{ij}^k. When a solution violates the nonnegative constraint of X_{ij}^k, just compare the value of X_{ij}^k and X_{ji}^k, without loss of generality we assume that $X_{ij}^k>X_{ji}^k$, then let the final feasible solution be $(X_{ij}^k)'=X_{ij}^k-X_{ji}^k$, $(X_{ji}^k)'=0$, the equation set (28) still holds and we have a nonnegative optimal solution of X_{ij}^k, X_{ji}^k that keeps the inequality (8) hold. It is worth noting that we may have multiple optimal solutions of X_{ij}^k by minimizing (7) under constraints (4) and (5), with $\alpha=(2/2-\delta)$. However, these optimal solutions will result in different values of the other term (2) about the total idle distance in the objective function of (11), and only solutions of problem (7) that also satisfy other constraints such as (3) can be feasible solutions of problem (11). Hence, we use (7) as an service fairness metric term of the objective function for problem (11), and approximately minimize the difference between local and global demand–supply ratios by minimizing (7).

It is worth noting that when ϵ_0 is small and γ_0 is close to 1, α is close to 0. ■

B. Proof of Theorem 1

Proof: To find the equivalent form of the minimax problem (11) when $\tau=1$ (here we only have variable X, not $X^2,\ldots,X^{\tau}$), the main step is to find the dual problem of the maximization over r for any fixed X and L. No constraint of problem (11) is a function of r, when considering the maximization problem with variable r and already fixed X and L, the constraints do not affect the values of r. Hence, to find the equivalent minimization form of the maximization problem, we do not include constraints irrelevant to r and only consider the objective function part. For any fixed X and L, the maximum part of problem (11) is equivalent to

$$\max_{r\in\Delta}\ J_D(X)+\beta J_E(X,r)=J_D(X)+c^T(X)r$$

$$[c(X)]_i=\beta\frac{1}{\left(\sum_{j=1}^{n}X_{ji}-\sum_{j=1}^{n}X_{ij}+L_i\right)^{\alpha}}$$

$$J_D(X)=\sum_i\sum_j X_{ij}W_{ij}. \tag{31}$$

Here $J_E(X,r)$ is affine (also concave) of r for any fixed value of (X,L), since with (X,L) fixed, function $[c(X)]_i$ also has a fixed value. And $J_E(X,r)$ a convex function of (X,L) for any fixed value of r. The function of power $\frac{1}{x^{\alpha}}$ is convex on scalar $x>0$ when $\alpha>0$ [8, Ch. 3.1.5]. Consider a concatenated matrix $[X,L]\in\mathbb{R}^{n\times(n+1)}$ with the last column as vector $L\in\mathbb{R}^n$, and a matrix $A^i\in\mathbb{R}^{n\times(n+1)}$ with $A_{ji}^i=1$, $j=1,\ldots,n$, $A_{ij}^i=-1$, $j=1,\ldots,n$,

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

$A^i_{i,(n+1)} = 1$. Then $\sum_{j=1}^n X_{ji} - \sum_{j=1}^n X_{ij} + L_i = TrA^i[X,L] = \sum_{i=1}^n \sum_{j=1}^{n+1} A^i_{ij}[X,L]_{ij}$, and $[c(X)]_i = (1/(TrA^i[X,L])^\alpha)$ is a composition of function $(1/x^\alpha)$ with affine mapping $TrA^i[X,L] : \mathbb{R}^{n\times(n+1)} \to R_+$, trace of the multiplication of matrices $[X,L]$ and A^i. Because composition with an affine mapping is an operation that preserves convexity [8, Ch. 3.2.2], $[c(X)]_i$ is a convex function of X and L. Finally, $J_E(X,r) = \sum_{i=1}^n \beta r_i [c(X)]_i$, $\beta r_i \geq 0$ is a nonnegative weighted sum of convex functions $[c(X)]_i$, an operation that preserves convexity [8, Ch. 3.2.1]. Hence, $J_E(X,r)$ is a convex function of X and L.

The Lagrangian of problem (31) with the Lagrangian multipliers $\lambda \geq 0, \upsilon \geq 0$ is $\mathcal{L}(X,r,\lambda,\upsilon) = J_D(X) + b^T\lambda - (A^T\lambda - c(X) - \upsilon)^T r$, where $(A^T\lambda - c(X) - \upsilon)^T r$ is a linear function of r, and the upper bound exists only when $A^T\lambda - c(X) - \upsilon = 0$. The objective function of the dual problem is

$$g(X,\lambda,\upsilon) = \sup_{r\in\Delta} \mathcal{L}(X,r,\lambda,\upsilon) = \begin{cases} J_D(X) + b^T\lambda & \text{if } A^T\lambda - c(X) - \upsilon = 0 \\ \infty & m \text{ otherwise.} \end{cases}$$

With $\upsilon \geq 0$, the constraint $A^T\lambda - c(X) - \upsilon = 0$ is equivalent to $A^T\lambda - c(X) \geq 0$. Strong duality holds for problem of (31) since it satisfies the refined Slater's condition for affine inequality constraints [8, Ch. 5.2.3]—the primal problem is convex, $c^T(X)r$ is affine of r, and by the definition of the uncertainty set, the nonempty affine inequality constraint of r is feasible. The primal convex problem is feasible with affine inequality constraints. The dual problem of (31) is

$$\min_{\lambda\geq 0} \quad J_D(X) + b^T\lambda \quad \text{s.t.} \quad A^T\lambda - c(X) \geq 0. \tag{32}$$

The minimization problem (32) is the dual problem of (31) with the same optimal cost for any fixed value of X and L, and problem (11) is to minimize the same objective $J_D(X) + b^T\lambda$ also over X (when $T = 1$, L is the number of initial empty vehicles at each region measured by GPS data, so L is a provided parameter in this case. When $\tau \geq 2$, L^k, $k = 2,\ldots,\tau$ are variables) together with the constraints about X. The constraint $A^T - c(X) \geq 0$ is convex of X, since $[c(X)]_i$ is convex of X for $i = 1,\ldots,n$, and the constraint of $A^T - c(X) \geq 0$ is equivalent to n inequalities between convex functions and a scalar 0, which are convex constraint inequalities. Grouping the minimization objective and all the constraints of problem (11), we get problem (21) as the equivalent convex optimization form of problem (11). ■

C. Proof of Lemma 2

Proof: Now consider the maximin problem over stage k and $k+1$, $1 \leq k \leq \tau - 1$ of problem (11)

$$\max_{r^k\in\Delta_k} \min_{X^{k+1},L^{k+1}} \quad J = \sum_{k=1}^{\tau} (J_D(X^k) + \beta J_E(X^k, r^k)) \quad \text{s.t. constraints of (11).} \tag{33}$$

The domain of problem (33) satisfies that $X^{k+1}, L^{k+1}, \lambda$ is compact, and the domain of r^k is compact. The objective function is a closed function convex over X^{k+1}, L^{k+1} and concave over r^k. According to Proposition 2.6.9 with condition (1) of [5], when the objective and constraint functions are convex of the decision variables, concave of the uncertain parameters, and the domain of decision variables and uncertain parameters are compact, the set of saddle points for the maximin problem at time k and $k+1$, i.e., $\max_{r^k\in\Delta_k} \min_{X^{k+1},L^{k+1}} J$ with the objective function and constraints of problem (33) is nonempty. The minimax equality holds for problem (33) at time k and $k+1$

$$\max_{r^k\in\Delta_k} \min_{X^{k+1},L^{k+1}} J = \min_{X^{k+1},L^{k+1}} \max_{r^k\in\Delta_k} J.$$

Repeat the above proof process from $k = \tau - 1$ backward to $k = 1$, we get a minimax form of robust optimization problem $\min_{X^{1:\tau},L^{2:\tau}} \max_{r^1\in\Delta_1,\ldots,r^\tau\in\Delta_\tau} J = \min_{X^{1:\tau},L^{2:\tau}} \max_{r_c\in\Delta} J$. ■

D. Proof of Lemma 3 and Theorem 2

1) Proof of Lemma 3:

Proof: With the polytope form of uncertainty set (23), the domain of each r^k is closed and convex, i.e., is compact, and Lemma 2 holds. Considering the maximizing part of problem (22)

$$\max_{r^1\in\Delta_1,\ldots,r^\tau\in\Delta_\tau} \quad J, \quad \text{s.t. constraints of (11)} \tag{34}$$

the Lagrangian of (34) with multipliers $\lambda^k \geq 0, \upsilon^k \geq 0$ is $\mathcal{L}(X^k, r^k, \lambda^k, \upsilon^k) = \sum_{k=1}^{\tau} (J_D(X^k) + b_k^T\lambda^k - (A_k^T\lambda^k - c(X^k) - \upsilon^k)^T r^k)$. Hence, based on the proof of Theorem 1, we take partial derivative of the Lagrangian for every $r^k \in \Delta_k$. The inequality constraint of $r^k \in \Delta_k$ defined as (23) is affine of r^k and feasible (nonempty), $c^T(X^k)r^k$ is affine of r^k, and problem (34) is convex with feasible affine inequality constraints. Hence, refined Slater's condition for affine constraints is satisfied and strong duality holds for problem (34). An equivalent form of (11) under uncertainty set (23) is defined as (24). ■

2) Proof of Theorem 2:

Proof: With uncertain set defined as (25), the domain of each r^k is compact and Lemma 2 holds. We consider the equivalent problem (22) of (11), and first derive the Lagrangian of the maximum part of the objective function (22) with constraint $\lambda \geq 0, \upsilon_k \geq 0$

$$\mathcal{L}(X^k, r^k, \lambda, \upsilon_k) = b^T\lambda - \sum_{k=1}^{\tau} \left(\left(A_k^T\lambda - c(X^k) - \upsilon_k\right)^T r^k - J_D(X^k) \right). \tag{35}$$

Similarly as the proof of Theorem 1, we take the partial derivative of (35) over each r^k, the objective function of the dual problem is

$$\sup_{r^k\in\Delta_k} \mathcal{L}(X^k, r^k, \lambda, \upsilon_k) = \sum_{k=1}^{\tau} J_D(X^k) + b^T\lambda$$

when $A_k^T\lambda - c(X^k) - \upsilon_k = 0$.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Since the inequality constraint of the uncertainty set defined as (25) is affine of each r^k and feasible (nonempty uncertainty set), $c^T(X^k)r^k$ is affine of r^k, and problem (34) is convex with feasible affine inequality constraints, refined Slater's condition with affine inequality constraints is satisfied. Then strong duality holds, problem (26) is a equivalent to the computationally tractable convex optimization form (11) under uncertain set (25). ■

E. Proof of Theorem 3

Proof: Under the definition of uncertainty set (20) for concatenated r^k, the domain of each r^k is compact, and problem (11) is equivalent to (22). We now consider the dual form for the objective function $\sum_{k=1}^{\tau} J_E(X^k, r^k)$ that relates to r^k. By the definition of inner product, we have $\sum_{k=1}^{\tau} c^T(X^k)r^k = c_l^T(X)r_c$, $c_l(X) = [c^T(X^1) \ \dots \ c^T(X^\tau)]^T$. When the uncertainty set of r_c is an SOC defined as (20), problem (22) is equivalent to

$$\begin{aligned}\min_{X^k, L^k} \max_{r_c \geq 0} \quad & \left(c_l^T(X) r_c + \sum_{k=1}^{\tau} \sum_i \sum_j X_{ij}^k W_{ij} \right) \\ \text{s.t.} \quad & r_c = \hat{r}_c + y + C^T w \\ & \|y\|_2 \leq \Gamma_1^B, \|w\|_2 \leq \sqrt{\frac{1}{\epsilon} - 1} \\ & \text{constraints of (11).}\end{aligned} \tag{36}$$

We first consider the following minimax problem related to the uncertainty set:

$$\begin{aligned}\max_{r_c \geq 0} \quad & c_l^T(X) r_c \\ \text{s.t.} \quad & r_c = \hat{r}_c + y + C^T w.\end{aligned} \tag{37}$$

The constraints of problem (37) have a feasible solution $r_c = \hat{r}_c$, $y = 0$ and $w = 0$, such that $\|y\|_2 < \Gamma_1^B$, $\|w\|_2 < ((1/\epsilon) - 1)^{1/2}$, and $c_l^T(X)r_c$ is affine of r_c. Hence, Slater's condition is satisfied and strong duality holds.

To get the dual form of problem (37), we start from the following Lagrangian with $v \geq 0$, $\mathcal{L}(X, r_c, z, v) = c_l^T(X)r_c + z^T(\hat{r}_c + y + C^T w - r_c) + v^T r_c$. By taking the partial derivative of the above Lagrangian over r_c, we get the supreme value of the Lagrangian as

$$\sup_{r_c} \mathcal{L}(X, r_c, z, v) = \begin{cases} z^T(\hat{r}_c + y + C^T w) & \text{if } c_l(X) \leq z \\ \infty & \text{o.w.} \end{cases}$$

Then with the norm bound of y and w, we have

$$\begin{aligned}\sup_{\|y\|_2 \leq \Gamma_1^B, \|w\|_2 \leq \sqrt{\frac{1}{\epsilon}-1}} & (z^T(\hat{r}_c + y + C^T w)) \\ & = \hat{r}_c^T z + \Gamma_1^B \|z\|_2 + \sqrt{\frac{1}{\epsilon} - 1} \|Cz\|_2.\end{aligned}$$

Hence, the objective function of the dual problem for (37) is

$$\begin{aligned} g(X, r_c, z) &= \sup_{r_c \in \mathcal{U}_\epsilon^{\text{CS}}} \mathcal{L}(X, r_c, z) \\ &= \begin{cases} \hat{r}_c^T z + \Gamma_1^B \|z\|_2 + \sqrt{\frac{1}{\epsilon} - 1} \|Cz\|_2, & \text{if } c_l(X) \leq z \\ \infty & \text{o.w.} \end{cases}\end{aligned}$$

Together with the objective function $J_D(X^k)$ and other constraints that do not directly involve r_c, an equivalent convex form of (11) given the uncertainty set (20) is shown as (27). ■

REFERENCES

[1] S. Ali, A. A. Maciejewski, H. J. Siegel, and J.-K. Kim, "Measuring the robustness of a resource allocation," *IEEE Trans. Parallel Distrib. Syst.*, vol. 15, no. 7, pp. 630–641, Jul. 2004.

[2] M. Asif *et al.*, "Spatiotemporal patterns in large-scale traffic speed prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 2, pp. 794–804, Apr. 2014.

[3] R. K. Balan, K. X. Nguyen, and L. Jiang, "Real-time trip information service for a large taxi fleet," in *Proc. ACM 9th Int. Conf. Mobile Syst., Appl., Services (MobiSys)*, 2011, pp. 99–112.

[4] A. Ben-Tal and A. Nemirovski, "Robust convex optimization," *Math. Oper. Res.*, vol. 23, no. 4, pp. 769–805, 1998.

[5] D. Bertsekas *et al.*, *Convex Analysis and Optimization*. Belmont, MA, USA: Athena Scientific, 2003.

[6] D. Bertsimas, V. Gupta, and N. Kallus. (Nov. 2014). "Data-driven robust optimization." [Online]. Available: https://arxiv.org/abs/1401.0212

[7] D. Bertsimas, D. A. Iancu, and P. A. Parrilo, "A hierarchy of near-optimal policies for multistage adaptive optimization," *IEEE Trans. Autom. Control*, vol. 56, no. 12, pp. 2809–2824, Dec. 2011.

[8] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge Univ. Press, 2004.

[9] G. C. Chasparis, M. Maggio, E. Bini, and K.-E. Årzén, "Design and implementation of distributed resource management for time-sensitive applications," *Automatica*, vol. 64, pp. 44–53, Feb. 2015,

[10] F. A. Cuzzola, J. C. Geromel, and M. Morari, "An improved approach for constrained robust model predictive control," *Automatica*, vol. 38, no. 7, pp. 1183–1189, 2002.

[11] H. A. David and H. N. Nagaraja, *Order Statistics*. New York, NY, USA: Wiley, 1970.

[12] B. Donovan and D. B. Work, "Using coarse GPS data to quantify city-scale transportation system resilience to extreme events," presented at the Transp. Res. Board Annu. Meet., Jul. 2015.

[13] R. Ganti, M. Srivatsa, and T. Abdelzaher, "On limits of travel time predictions: Insights from a New York city case study," in *Proc. IEEE 34th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jun./Jul. 2014, pp. 166–175.

[14] Y. Geng and C. G. Cassandras, "New 'smart parking' system based on resource allocation and reservations," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, pp. 1129–1139, Sep. 2014.

[15] J. C. Herrera, D. B. Work, R. Herring, X. Ban, Q. Jacobson, and A. M. Bayen, "Evaluation of traffic data obtained via GPS-enabled mobile phones: The *Mobile Century* field experiment," *Transp. Res. C, Emerg. Technol.*, vol. 18, no. 4, pp. 568–583, 2010.

[16] Z. Huang, "Extensions to the K-means algorithm for clustering large data sets with categorical values," *Data Mining Knowl. Discovery*, vol. 2, no. 3, pp. 283–304, Sep. 1998.

[17] D.-H. Lee, H. Wang, R. Cheu, and S. Teo, "Taxi dispatch system based on current demands and real-time traffic conditions," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1882, pp. 193–200, Jan. 2004.

[18] F. Miao, S. Han, A. M. Hendawi, M. E. Khalefa, J. A. Stankovic, and G. J. Pappas, "Data-driven distributionally robust vehicle balancing using dynamic region partitions," in *Proc. 8th Int. Conf. Cyber-Phys. Syst. (ICCPS)*, 2017, pp. 261–271.

[19] F. Miao, S. Han, S. Lin, and G. J. Pappas, "Robust taxi dispatch under model uncertainties," in *Proc. 54th IEEE Conf. Decision Control (CDC)*, Dec. 2015, pp. 2816–2821.

[20] F. Miao *et al.*, "Taxi dispatch with real-time sensing data in metropolitan areas: A receding horizon control approach," *IEEE Trans. Autom. Sci. Eng.*, vol. 13, no. 2, pp. 463–478, Apr. 2016.

[21] F. Miao *et al.*, "Taxi dispatch with real-time sensing data in metropolitan areas—A receding horizon control approach," in *Proc. 6th ACM/IEEE Int. Conf. Cyber-Phys. Syst.*, 2015, pp. 100–109.

[22] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas, "Predicting taxi–passenger demand using streaming data," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, pp. 1393–1402, Sep. 2013.

[23] A. Pantoja and N. Quijano, "A population dynamics approach for the dispatch of distributed generators," *IEEE Trans. Ind. Electron.*, vol. 58, no. 10, pp. 4559–4567, Oct. 2011.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[24] M. Pavone, S. L. Smith, E. Frazzoli, and D. Rus, "Robotic load balancing for mobility-on-demand systems," *Int. J. Robot. Res.*, vol. 31, no. 7, pp. 839–854, Jun. 2012.
[25] J. Shawe-Taylor and N. Cristianini, "Estimating the moments of a random vector with applications," in *Proc. GRETSI Conf.*, 2003, pp. 47–52.
[26] K. I. Wong and M. G. H. Bell, "The optimal dispatching of taxis under congestion: A rolling horizon approach," *J. Adv. Transp.*, vol. 40, no. 2, pp. 203–220, 2006.
[27] D. B. Work, S. Blandin, O.-P. Tossavainen, B. Piccoli, and A. M. Bayen, "A traffic model for velocity data assimilation," *Appl. Res. Math. Exp.*, vol. 2010, no. 1, pp. 1–35, Jan. 2010.
[28] D. Zhang, T. He, S. Lin, S. Munir, and J. A. Stankovic, "Dmodel: Online taxicab demand model from big sensor data in a roving sensor network," in *Proc. IEEE Int. Congr. Big Data (BigData Congress)*, Jun./Jul. 2014, pp. 152–159.
[29] D. Zhang, T. He, S. Lin, S. Munir, and J. A. Stankovic, "Online cruising mile reduction in large-scale taxicab networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 11, pp. 3122–3135, Nov. 2015.
[30] R. Zhang and M. Pavone, "Control of robotic mobility-on-demand systems: A queueing-theoretical perspective," *Int. J. Robot. Res.*, vol. 35, nos. 1–3, pp. 186–203, Jan. 2016.



Fei Miao (S'13–M'16) received the B.Sc. degree in automation from Shanghai Jiao Tong University, Shanghai, China, in 2010, and the M.A. degree in statistics and the Ph.D. degree in electrical and systems engineering from the University of Pennsylvania, Philadelphia, PA, USA, in 2015 and 2016, respectively.

She was a Post-Doctoral Researcher with the Department of Electrical and Systems Engineering, GRASP Laboratory, PRECISE Center, University of Pennsylvania. She is currently an Assistant Professor with the Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA. Her current research interests include data-driven real-time optimization and control of cyber-physical systems under model uncertainties, and resilient and secure control cyber-physical systems.

Dr. Miao was a Best Paper Award Finalist at the 6th ACM/IEEE International Conference on Cyber-Physical Systems in 2015 and the Charles Hallac and Sarah Keil Wolf Award for Best Doctoral Dissertation.



Shuo Han (S'08–M'14) received the B.E. and M.E. degrees in electronic engineering from Tsinghua University, Beijing, China, in 2003 and 2006, respectively, and the Ph.D. degree in electrical engineering from the California Institute of Technology, Pasadena, CA, USA, in 2013.

He was a Post-Doctoral Researcher with the Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, USA. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, IL, USA. His current research interests include developing rigorous frameworks for data-driven decision making that enable reliable and efficient operations of networked cyber-physical systems, including many smart city applications such as power and transportation networks.

Dr. Han was a finalist for the Best Student Paper Award at the 2013 American Control Conference.



Shan Lin received the Ph.D. degree in computer science at the University of Virginia, Charlottesville, VA, USA, in 2010, where his Ph.D. dissertation is on taming networking challenges with feedback control.

He is currently an Assistant Professor with the Electrical and Computer Engineering Department, Stony Brook University, Stony Brook, NY, USA, where he is involved in wireless network protocols, interoperable medical devices, smart transportation systems, and intelligent sensing systems. His current research interest includes networked systems, with an emphasis on feedback control-based design for cyber-physical systems and sensor systems.



Qian Wang received the bachelor's and master's degrees in computer and information science from the University of Pennsylvania, Philadelphia, PA, USA.

He was involved in scalable and high-performance data platforms at Two Sigma Investments, New York, NY, USA, a quantitative hedge fund headquartered. He has been serving product and strategy leadership roles in multiple financial technology firms in Asia, including Advance.AI, a fintech company mainly serving South East Asia, and China Chengxin Credit Information, one of the leading credit data service providers in China.



John A. Stankovic (LF'14) received the Ph.D. degree from Brown University, Providence, RI, USA.

He is currently the BP America Professor with the Computer Science Department, University of Virginia, Charlottesville, VA, USA. He has an H-index of 105 and over 40 000 citations.

Prof. Stankovic is a fellow of the ACM. He received an Honorary Doctorate from the University of York. He was a recipient of the IEEE Real-Time Systems Technical Committee's Award for Outstanding Technical Contributions and Leadership, the IEEE Technical Committee on Distributed Processing's Distinguished Achievement Award, and seven best paper awards, including one for ACM SenSys 2006.



Abdeltawab Hendawi received the M.Sc. and Ph.D. degrees in computer science and engineering from the University of Minnesota, Minneapolis, MN, USA, in 2012 and 2015, respectively.

He is currently a Post-Doctoral Research Associate in computer science with the University of Virginia, Charlottesville, VA, USA. His current research interests include data management and analysis with more focus on smart cities related applications.

Dr. Hendawi was a recipient of the Best Paper Award at the ACM SIGSPATIAL MobiGIS 2012, the Best Poster Award at the U-Spatial Symposium 2013, the Best Demo Paper Awards at the ACM SIGSPATIAL 2014 and at the IEEE MDM 2015, and the Hobby Post-Doctoral Research Fellowship 2015–2018 from the Computer Science Department, University of Virginia. His vision for smart location-aware services in smart cities also won a highly recognized Award, Blue Sky Ideas Award 2016.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



Desheng Zhang (F'09–M'10) is an Assistant Professor with the Department of Computer Science, Rutgers University, New Brunswick, NJ, USA. He is broadly concentrated on bridging cyber-physical systems and big urban data by technical integration of communication, computation, and control in data-intensive urban systems. He is focused on the life cycle of big data-driven urban systems, from multisource data collection to streaming-data processing, heterogeneous data management, model abstraction, visualization, privacy, service design, and deployment in complex urban settings. He is currently interested in real-time interactions among heterogeneous urban systems including cellphone, smartcard, taxi, bus, truck, subway, bike, personal vehicle, electric vehicle, and road networks.



Tian He is currently an Associate Professor with the Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA. He has authored or co-authored over 200 papers in premier network journals and conferences with over 17 000 citations (H-Index 52). His current research interests include wireless sensor networks, cyber-physical systems, intelligent transportation systems, real-time embedded systems, and distributed systems.

Dr. He was a recipient of the NSF CAREER Award, the George W. Taylor Distinguished Research Award, the McKnight Land-Grant Professorship, and many best paper awards in networking.



George J. Pappas (S'90–M'91–SM'04–F'09) received the Ph.D. degree in electrical engineering and computer sciences from the University of California, Berkeley, CA, USA, in 1998.

He is currently the Joseph Moore Professor and Chair with the Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, USA. He also holds a secondary appointment with the Department of Computer and Information Sciences and the Department of Mechanical Engineering and Applied Mechanics, University of Pennsylvania, Philadelphia, PA, USA. He is a member of the GRASP Laboratory, Philadelphia, PA, USA, and the PRECISE Center, University of Pennsylvania. He had previously served as the Deputy Dean for Research with the School of Engineering and Applied Science, University of Pennsylvania. His current research interests include control theory and, in particular, hybrid systems, embedded systems, cyber-physical systems, and hierarchical and distributed control systems, with applications to unmanned aerial vehicles, distributed robotics, green buildings, and bimolecular networks.

Dr. Pappas was a recipient of the Antonio Ruberti Young Researcher Prize, the George S. Axelby Award, the Hugo Schuck Best Paper Award, the George H. Heilmeier Award, the National Science Foundation PECASE award, and numerous best student papers awards at ACC, CDC, and ICCPS.