1-1-2016

Data-Driven Dynamic Robust Resource Allocation: Application to Efficient Transportation

Fei Miao
University of Pennsylvania, miaofei@seas.upenn.edu

Data-Driven Dynamic Robust Resource Allocation: Application to Efficient Transportation

Abstract
The transformation to smarter cities brings an array of emerging urbanization challenges. With the development of technologies such as sensor networks, storage devices, and cloud computing, we are able to collect, store, and analyze a large amount of data in real time. Modern cities have brought to life unprecedented opportunities and challenges for allocating limited resources in a data-driven way. Intelligent transportation system is one emerging research area, in which sensing data provides us opportunities for understanding spatial-temporal patterns of demand human and mobility. However, greedy or matching algorithms that only deal with known requests are far from efficient in the long run without considering demand information predicted based on data.

In this dissertation, we develop a data-driven robust resource allocation framework to consider spatial-temporally correlated demand and demand uncertainties, motivated by the problem of efficient dispatching of taxi or autonomous vehicles. We first present a receding horizon control (RHC) framework to dispatch taxis towards predicted demand; this framework incorporates both information from historical record data and real-time GPS location and occupancy status data. It also allows us to allocate resource from a globally optimal perspective in a longer time period, besides the local level greedy or matching algorithm for assigning a passenger pick-up location of each vacant vehicle. The objectives include reducing both current and anticipated future total idle driving distance and matching spatial-temporal ratio between demand and supply for service quality. We then present a robust optimization method to consider spatial-temporally correlated demand model uncertainties that can be expressed in closed convex sets. Uncertainty sets of demand vectors are constructed from data based on theories in hypothesis testing, and the sets provide a desired probabilistic guarantee level for the performance of dispatch solutions. To minimize the average resource allocation cost under demand uncertainties, we develop a general data-driven dynamic distributionally robust resource allocation model. An efficient algorithm for building demand uncertainty sets that compatible with various demand prediction methods is developed. We prove equivalent computationally tractable forms of the robust and distributionally robust resource allocation problems using strong duality. The resource allocation problem aims to balance the demand-supply ratio at different nodes of the network with minimum balancing and re-balancing cost, with decision variables on the denominator that has not been covered by previous work.

Trace-driven analysis with real taxi operational record data of San Francisco shows that the RHC framework reduces the average total idle distance of taxis by 52%, and evaluations with over 100GB of New York City taxi trip data show that robust and distributionally robust dispatch methods reduce the average total idle distance by 10% more compared with non-robust solutions. Besides increasing the service efficiency by reducing total idle driving distance, the resource allocation methods in this dissertation also reduce the demand-supply ratio mismatch error across the city.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

DATA-DRIVEN DYNAMIC ROBUST RESOURCE ALLOCATION: APPLICATION TO EFFICIENT TRANSPORTATION

Fei Miao

A DISSERTATION

in

Electrical and Systems Engineering

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2016

Supervisor of Dissertation

George J. Pappas, Joseph Moore Professor and Chair, Electrical & Systems Engineering

Graduate Group Chairperson

Alejandro Ribeiro, Rosenbluth Associate Professor, Electrical & Systems Engineering

Dissertation Committee

Victor M. Preciado, Chair of the Committee, Assistant Professor, Electrical & Systems Engineering

George J. Pappas, Professor, Electrical & Systems Engineering

Insup Lee, Professor, Computer and Information Science

John A. Stankovic, Professor, UVA Computer Science

Hamsa Balakrshnan, Associate Professor, MIT Aeronautics and Astronautics

DATA-DRIVEN DYNAMIC ROBUST RESOURCE ALLOCATION: APPLICATION TO EFFICIENT TRANSPORTATION

Dedicated to Minhui Zhang, Fujin Miao and Qian Wang

ACKNOWLEDGEMENT

I have spent almost six wonderful years at the University of Pennsylvania, and during this period I had the fortune to meet some truly amazing people.

First and foremost, I would like to express my gratitude to my adviser Prof. George J. Pappas for being the best mentor. He guided me to a completely new research field with the right research attitudes, and gave me the independence to develop critical thinking and explore various research topics. George has shown great support when I need his brilliant suggestions and inspired me to be a better researcher.

Special thanks to my thesis committee members, professors John A. Stankovic, Insup Lee, Hamsa Balakrishnan and Victor M. Preciado. They took extra effort to provide me with valuable advice during writing of this thesis. Furthermore, John has been much more than a committee member; I enjoyed working with him and I deeply appreciate the time and research vision he shared with me. I also know more about the field of cyber physical system from Insup's group meeting.

I want to extend heartfelt thanks to Prof. Miroslav Pajic and Quanyan Zhu for being great collaborators and providing valuable discussions while working on problems from the CPS security domain. I would also like to thank Shuo Han in our group for revising every detail of my paper draft patiently, helping me to improve my writing skill and suggesting interesting related work to read.

Grad school would not be such an invaluable part of my life without all the people from the GRASP Lab and the Precise Center - thank you for your friendship and all your help, it has been a privilege to work with you! Finally, for their unconditional love, patience and encouragement, I am grateful to my family: my parents and my husband.

The work presented in this thesis has been supported in part by the NSF CNS-1239483, CNS-1239108, CNS-1239226, and CPS-1239152 grants with project title: CPS: Synergy:Collaborative Research: Multiple-Level Predictive Control of Mobile Cyber Physical Systems with Correlated Context.

ABSTRACT

DATA-DRIVEN DYNAMIC ROBUST RESOURCE ALLOCATION: APPLICATION TO EFFICIENT TRANSPORTATION

Fei Miao

George J. Pappas

The transformation to smarter cities brings an array of emerging urbanization challenges. With the development of technologies such as sensor networks, storage devices, and cloud computing, we are able to collect, store, and analyze a large amount of data in real time. Modern cities have brought to life unprecedented opportunities and challenges for allocating limited resources in a data-driven way. Intelligent transportation system is one emerging research area, in which sensing data provides us opportunities for understanding spatial-temporal patterns of demand human and mobility. However, greedy or matching algorithms that only deal with known requests are far from efficient in the long run without considering demand information predicted based on data.

In this dissertation, we develop a data-driven robust resource allocation framework to consider spatial-temporally correlated demand and demand uncertainties, motivated by the problem of efficient dispatching of taxi or autonomous vehicles. We first present a receding horizon control (RHC) framework to dispatch taxis towards predicted demand; this framework incorporates both information from historical record data and real-time GPS location and occupancy status data. It also allows us to allocate resource from a globally optimal perspective in a longer time period, besides the local level greedy or matching algorithm for assigning a passenger pick-up location of each vacant vehicle. The objectives include reducing both current and anticipated future total idle driving distance and matching spatial-temporal ratio between demand and supply for service quality. We then present a robust optimization method to consider spatial-temporally correlated demand model uncertainties that can be expressed in closed convex sets. Uncertainty sets of demand vectors are constructed from data based on theories in hypothesis testing, and the sets provide a desired proba-

bilistic guarantee level for the performance of dispatch solutions. To minimize the average resource allocation cost under demand uncertainties, we develop a general data-driven dynamic distributionally robust resource allocation model. An efficient algorithm for building demand uncertainty sets that compatible with various demand prediction methods is developed. We prove equivalent computationally tractable forms of the robust and distributionally robust resource allocation problems using strong duality. The resource allocation problem aims to balance the demand-supply ratio at different nodes of the network with minimum balancing and re-balancing cost, with decision variables on the denominator that has not been covered by previous work.

Trace-driven analysis with real taxi operational record data of San Francisco shows that the RHC framework reduces the average total idle distance of taxis by 52%, and evaluations with over 100GB of New York City taxi trip data show that robust and distributionally robust dispatch methods reduce the average total idle distance by 10% more compared with non-robust solutions. Besides increasing the service efficiency by reducing total idle driving distance, the resource allocation methods in this dissertation also reduce the demand-supply ratio mismatch error across the city.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1 : Introduction

The number of cities is increasing worldwide and the transformation to smarter cities is taking place, which bring an array of emerging urbanization challenges [63, 16]. With the development of technologies such as radio-frequency identification (RFID), sensor networks, storage devices, and cloud computing, we are able to collect, store, and analyze a large amount of data efficiently [38].

Cities have grown into complex systems saturated by aging infrastructures of increasing running costs, fading control over private data, and a growing pool of interlinked socio-economic problems urging for immediate solutions. The United Nations forecasts that by 2050, over six billion people, or about 66% of the world population, will live in cities or towns [67]. Increased urbanization worldwide presents a variety of challenges related to the systems integral to any city, such as public transportation, roads and bridges, water and energy systems, and telecommunications networks.

Future cities will be highly instrumented with sensors and devices that provide almost real-time updates of various states of cities, including congestions, level of pollutions, or availability of resources. The scaling laws observed in the evolution and growth of the modern cities fundamentally have brought to life unprecedented opportunities to address these challenges in a data-driven way. In order to manage the complexity of such urban environments in a smarter way, it is inevitable that real-time control and decision be implemented based on the state of cities measured by sensors.

1.1. Opportunities with Smart Cities

Intelligent transportation system is one emerging research area, in which sensing data collected in real time provides us opportunities for understanding spatial-temporal human mobility patterns. More and more transportation systems are equipped with various sensors and wireless radios to enable better mobility service, such as intelligent highways, traffic light control, supply chain management, and autonomous fleets. The embedded sensing and control technologies in these systems significantly improve their safety and efficiency over traditional systems. Examples include traffic speed [5], travel time [41, 6], passengers' demand model of taxi network [61], and road transporta-

tion network efficiency [82].

Based on such rich spatial-temporal information about passenger mobility patterns and demand, many control solutions have been designed for intelligent transportation systems. Coverage control and coordination algorithms to allocate groups of autonomous vehicles are presented with distributed gradient descent algorithms [25]. Dispatch algorithms that aim to minimize customers' waiting time [85, 47] or to reduce cruising mile [90] have been developed. A smart parking system that assigns and reserves an optimal resource (parking space) for a driver based on the driver's cost function has been proposed, and the overall efficient utilization of parking capacity is guaranteed [35]. Although these works heavily rely on precise passenger-demand models or prior information to make dispatch decisions, they show the possibility to improve system's performance with information provided by data.

Research and development on autonomous cars is currently very active. Researchers are not only developing the technology to make autonomous cars a reality, but are also analyzing their potential impact on urban mobility. By considering average demand predicted based on either historical or streaming data when making current decisions, vehicle re-balancing and re-allocating costs are reduced for shared automated vehicles [69, 92]. Similarly, the above mentioned projects of smart parking systems [35] and coordinating algorithms for groups of vehicles towards demand [25] are both examples in the autonomous vehicle area. A case study based on Singapore data shows that autonomous car sharing could reduce the number of passenger vehicles by 60% [81]. These work provide guidelines and justification for the design of shared-vehicle mobility-on-demand systems. More work that considers different transportation system design requirements are necessary with the trend of urbanization and technology development.

Meanwhile, resource allocation schemes with various performance metrics have been designed for numerous systems in the literature [60], such as wireless networks [31], data-centers [45], power systems [19], health-care and emergency response systems [30, 32], and transportation systems [35, 81]. Resilience properties of dynamical networks are analyzed for distributed routing policies [23, 24]. Strategies for resource allocation depend on the model of demand in general, and the knowledge

(a) New York City (b) San Francisco

Figure 1: Visualization of taxi pick-up and drop-off events

and assumptions about the demand affect the performance of the supply-providing approaches [21, 68]. Based on these exiting methods designed for different service requirements, new frameworks that deal with resource allocation problems with the paradigm of smart cities can be developed.

1.2. Challenges for Data-Driven Dynamic Resource Allocation of Efficient Transportation Systems

The ultimate goal of a modern transportation system is to fulfill the mobility requirement of people and goods while minimizing various operational costs such as greenhouse gas emissions and wears of the infrastructure. In the context of urban environments, on-demand mobility, including taxicabs and other ride-sharing services, has gained popularity in recent years due to the rapidly rising expenses of car ownership in cities. Figure 1 shows a visualization of taxi pick-up and drop-off events in New York City and San Francisco. The operation of on-demand mobility services with limited service resources, however, is far from optimal — they may result to extra costs and conflicts of interests to the limited resources in cities.

Though existing works for mobility-on-demand service of autonomous vehicle consider system-level optimality [92, 69, 81], how to incorporate historical and real-time sensing data to improve dynamic resource allocation performance or how to deal with demand uncertainties has not been explicitly studied or empirically tested yet. The challenges considered in this dissertation are as follows.

1.2.1. How to Improve Global Optimality and Efficiency with Predicted Demand

Greedy strategies may increase human satisfaction myopically, while the total utilization is not optimal under conflict of interests. How to incorporate historical recording data and real-time sensing information to allocate resources from a system-level optimality perspective is critical for smart cities, since resource is limited.

Compared with transportation systems such as subway, bus, and trains, ride-sharing or taxi service is more flexible without a repeated schedule every day, and dispatch decisions should be made in real time. However, efficient coordination of taxi networks based on the current system state at a large scale is a challenging task. Traditional taxi networks in metropolitan areas heavily rely on taxi drivers' experience to look for passengers on streets to maximize individual profit. However, such self-interested, uncoordinated behaviors of drivers usually result in spatial-temporal mismatch between taxi supply and passenger demand. Greedy algorithms are widely employed by large taxi or ride-sharing service companies, such as finding the nearest vacant taxi to pick up a passenger [51], or first-come, first-served.

Considering a transportation system such as a taxi dispatch system or an on-demand ride-sharing system (e.g., uber, lyft and Sidecar), the current applied service usually assigns the driver that can reach the customer in shortest time once a request appears in the system. Though aiming to minimize each individual's waiting time, the total profit is not globally optimal and the service is not efficient– passengers at over-supplied regions have shorter average waiting time than those at under-supplied regions, and the service may lose their customer in those under-supplied regions. Meanwhile, without a system-level regulator, drivers tend to stay within areas that they think there will be more potential customers, and traverse on streets in hoping to pick up the next passenger in a short idle distance or idle time based on their own experience. Before a request enters to the system, drivers do not have ideas where to go, hence, there will be extra idle driving distance, energy consumption and unnecessary congestion or occupation of the road resources caused by the behavior of searching passengers. There has not been previous work that considers this type of real-time resource allocation problem from a system-optimal perspective, with the demand predicted

based on either existing record data or streaming data. Further more, real-time sensing data provides update of a vehicle's status such as location, speed, and vacancy, and shows the mobility pattern of both vacant and in-service vehicles. How to define the measurement of service quality or efficiency considering available information provided by data is critical for improving the performance of the system.

1.2.2. How to Consider Demand Uncertainties in Dynamic Decisions

Given a demand-related dataset, how to formulate a computationally tractable robust resource allocation problem under predicted uncertain demand is a rising questions for many smart city applications.

Previous research has shown that sensing data contains rich information about passenger and taxi mobility patterns [91, 74, 73]. Moreover, recent studies have shown that the passenger demand information can be extracted and used to reduce passengers' waiting time, taxi cruising time, or future supply re-balancing cost to serve requests [49, 75, 92]. Meanwhile, considering future demand when making the current dispatch decisions helps to reduce resource re-allocating costs [92, 84].

However, passenger-demand models have their intrinsic model uncertainties that result from many factors, such as weather, passenger working schedule, and city events etc. Algorithms that do not consider these uncertainties can lead to inefficient dispatch services, resulting in long waiting times of under-served passengers, imbalanced workloads, and increased taxi idle mileage. While robust optimization aims to minimize the worst-case cost under all possible random parameters, it sacrifices average system performances [2]. It is essential to address the trade-off between the worst-case system performance guarantee and the average dispatch cost under uncertain demand, with system performance metric such as service fairness and service allocate/re-allocate cost under practical constraints.

1.2.3. How to Efficiently Construct Spatial-Temporal Demand Uncertainty Sets

How to construct spatial-temporally correlated uncertain demand sets based on a large amount of data for robust resource allocation problems is beyond the scope of designing an accurate machine learning algorithm— we need to bridge the gap between machine learning algorithm and robust optimization methods.

It is difficult to find a very accurate demand model based on data for many applications; even such a model exists, it may be too complicated to fit the requirement of a computationally tractable robust optimization problem. Thus, building an uncertain set that includes appropriate information for robust resource allocation strategies is critical and challenging. The demand uncertainty set should include information about either the value of the distribution of the random demand to make sure that the robust solutions based on it provides the desired performance guarantee, and the computational cost of reaching such robust solutions are not too high for a large-scale system.

Many application areas need a spatial-temporal model of demand uncertainties for regulating the supply more efficiently. For instance, in the area of clean and renewable energy, an adaptive robust dispatch method has been designed for wind power systems [52] but no probabilistic guarantee of the performance is guaranteed. Motivated by portfolio management problems in financial area, data-driven robust optimization approaches have been developed for independent and identically distributed (i.i.d.) sampled random vectors in the literature [12, 27, 79, 28]. For transportation systems such as taxi systems or autonomous vehicle systems, no previous work has considered to build a spatial-temporally correlated demand uncertainty set, or formulate a robust resource allocation framework given the uncertain predicted demand yet. An efficient modeling algorithm for a large sensing dataset need to be developed, the performance improvement based on uncertainty demand sets need to be evaluated based on data.

1.3. Contributions of the Thesis

Our goal is to utilize information provided by a large amount of sensing data to optimize real-time resource allocation strategies in smart cities. From a high level perspective, we fill in the

gap between demand data to dynamic resource allocation decisions, designs both computationally tractable robust optimal resource allocation models in a real-time framework and uncertain demand modeling algorithms. With the objective of balancing demand-supply ratio for a fair service, we prove computationally tractable forms and the corresponding uncertain demand set construction process. The decision variable of the robust problem is on the denominator, which has not been covered by previous work in the literature.

Regarding to the specific example of taxi or autonomous ride-sharing car dispatch framework, both anticipated future idle driving cost and global geographical service fairness are considered, while fulfilling current, local passenger demand. To accomplish such a goal, we incorporate both system models learned from historical data and real-time taxi data into a taxi network control framework. Evaluations based on datasets of metropolitan areas in the US show that the total idle distance of all taxis is reduced by our framework, and supply is more balanced across different regions of one city.

Contributions of this dissertation are explicitly stated as the following.

1.3.1. A Receding Horizon Control Framework for Real-Time Taxi Dispatch

We design a computationally efficient moving time horizon framework for taxi dispatch with large-scale real-time information of the taxi network. Our dispatch solutions in this framework consider future costs of balancing the supply demand ratio under physical constraints. We take a receding horizon control (RHC) approach to dynamically control taxis in large-scale networks. Future demand is predicted based on either historical taxi data sets [18] or streaming data [91, 62]. The real-time GPS and occupancy information of taxis is also collected to update supply and demand information for future estimation. This design iteratively regulates the mobility of idle taxis for high performance, demonstrating the capacity of large-scale smart transportation management.

The contributions of this domain are as follows.

- To the best of our knowledge, we are the first to design an RHC framework for large-scale taxi dispatching. We consider both current and future demand, saving costs under constraints

by involving anticipated future idle driving distance for re-balancing supply.

- The framework incorporates large-scale data in real-time control. Sensing data is used to build predictive passenger demand, taxi mobility models, and serve as real-time feedback for RHC.

- Extensive trace driven analysis based on a San Francisco taxi data set shows that our approach reduces average total taxi network idle distance by 52%, and the error between local and global supply demand ratio by 45%, compared to the actual historical taxi system performance.

- Spatial-temporal context information such as disruptive passenger demand is formulated as uncertainty sets of parameters into a robust dispatch problem. This allows the RHC framework to provide more robust control solutions under uncertain contexts. The error between local and global supply demand ratio is reduced by 25% compared with the error of solutions without considering demand uncertainties.

1.3.2. Data-Driven Robust Taxi Dispatch

Though real-time sensing information corrects parts of model prediction error based on the evaluations of the receding horizon control taxi dispatch framework, demand model uncertainty is still one critical factor that affects the performance of the dispatch algorithm. To consider model uncertainty with a real-time computable resource allocation approach, we design a promising yet challenging approach — a robust dispatch framework with an uncertain demand model, called an uncertainty set, that captures spatial-temporal correlations of demand uncertainties and provides a probabilistic guarantee (as defined in problem (4.12)). Solving the robust dispatch problem with the constructed uncertainty set yields a probabilistic guarantee for the optimality of the actual dispatch cost. We have the freedom to specify a lower bound for the probability that an actual dispatch cost under the true demand vector being smaller than the optimal cost of the robust dispatch solutions. Hence, we are able to find a better solution for considering the trade-off between the average dispatch cost and the minimum cost under the worst-case scenario than previous methods that do not provide any

guarantees.

We first develop the objective and constraints of a multi-stage 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 then design a data-driven algorithm for constructing uncertainty demand sets without assumptions about the true distribution of the demand vector. The constructing algorithm is based on theories proved for independent and identically distributed (i.i.d.) sampled random vectors in the robust optimization literature [12, 27, 79]. However, how to apply these theories for spatial-temporal data and a robust resource allocation form of taxi dispatch problem based on the constructed spatial-temporally correlated uncertainty sets have not been explored before. To the best of our knowledge, this is the first work to design a robust taxi dispatch framework that provides a desired probabilistic guarantee using demand uncertainty sets built from realistic data.

With two types of uncertainty sets — one box type and one second-order-cone (SOC) type, we prove equivalent convex optimization forms of the robust dispatch problem via the strong duality theorem. The robust dispatch problem formulated in this work 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 semi-definite programming (SDP) problems) that has already been covered by previous work [8, 12, 26]. With proofs shown in this work, both system performance and computational tractability are guaranteed under spatial-temporal demand uncertainties. Based on four years of taxi trip data in New York City, we evaluate factors that affect the accuracy of the uncertainty sets, properties of each type of uncertainty sets, and trade- off between the probabilistic guarantee levels and the average dispatch costs of robust dispatch solutions.

The contributions of our work in this domain are:

- We develop a multi-stage robust optimization model for taxi dispatch systems under spatial

temporal uncertainties of predicted demand, with the weighted sum of multi- objective of balancing vacant taxi supply and reducing total idle driving distance.

- We design a data-driven algorithm to construct uncertainty sets that provide a desired level of probabilistic guarantee for the robust taxi dispatch solutions. We show that the second-order-cone type of uncertain set provides a smaller average dispatch cost than the box type via evaluations.

- We prove that there exists an equivalent computationally tractable convex optimization form for the robust dispatch problem with each type of constructed uncertainty set.

- Evaluations on four years of taxi trip data in New York City show that 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.

1.3.3. Data-Driven Dynamic Distributionally Robust Resource Allocation

The knowledge and assumptions about the demand model affect the performance of resource allocation strategies. A robust allocation scheme shows its advantage in worst-case scenarios compared with non-robust approaches [2, 54, 52]. Considering the trade-off between system's average performance and worst-case performance, robust taxi dispatch techniques with a probabilistic guarantee level for an original chance constrained problem are developed and evaluated based on a realistic dataset [57]. Stochastic programming (SP) is another approach to describe decision-making problems under uncertain parameters. However, the computational complexity of an SP problem is not polynomial of the spatial-temporal decision variables, and not scalable for dynamic resource allocation in general. Moreover, it is difficult to obtain an explicit formulation about the true distribution function of the random demand purely based on data in practice. Hence, when we are able to construct a set of distribution functions that includes the true distribution function of the random demand given a demand dataset, minimizing the expected cost over the worst-case distribution function in the set is a promising approach. Distributionally robust optimization techniques are developed under this scenario in control optimization literature [28, 36].

To minimize the average resource allocation cost under demand uncertainties, we design a data-driven distributionally robust dynamic resource allocation model under uncertain spatial-temporally correlated demand, with an application in taxi dispatch problem given demand data. An efficient algorithm for constructing an uncertain set of the distribution function based on data without assumptions about prior knowledge is proposed, by utilizing the rolling-horizon property of the distribution uncertain set. The constructing algorithm is based on theories proved for independent and identically distributed (i.i.d.) sampled random vectors in hypothesis testing and data-driven optimization literature [18, 12, 28]. We prove an equivalent computationally tractable form of the distributionally robust resource allocation problem via strong duality theorem. With proofs shown in this work, both average performance of the system and computational tractability are guaranteed under spatial-temporal demand uncertainties.

The contributions of our work in this domain are

- We design an efficient algorithm to construct distributional uncertainty set based on spatial-temporal demand data for a data-driven dynamic distributionlly robust resource allocation model.

- We derive an equivalent computationally tractable convex optimization form for a general form of resource allocation problem with each type of constructed uncertainty set. The resource allocation problem aims to balance the demand-supply ratio at different nodes of the network with minimum balancing and re-balancing cost, with decision variables on the denominator that has not been covered by previous work [8, 12, 28].

- For an example problem of fairly allocating vacant taxis according to uncertain demand at each region of the city with minimum total idle driving distance, we evaluate the average cost of the distributionally robust taxi dispatch solutions based on four years taxi trip records of New York City. Results show that the average demand-supply ratio error is reduced by 28.6%, and the average total idle driving distance is reduced by 10.05%.

1.4. Outline of the Dissertation

The dissertation is organized as follows. Chapter 2 presents a summary of the used notation and background knowledge from convex optimization, robust optimization and hypothesis testing. We present the receding horizon control framework that incorporates both historical record and real-time sensing information in Chapter 3. Chapter 4 addresses the problem of demand uncertainties with a data-driven robust taxi dispatch framework, and both the process of constructing a demand uncertainty set from data and computationally tractable robust optimization formulations are designed. Motivated by the efficient transportation problem, Chapter 5 presents a general form of distributionally robust resource allocation method and an efficient algorithm of constructing uncertainty sets. Finally, in Chapter 6, we give our concluding remarks and highlight some future work in this field.

CHAPTER 2 : Background and Notation

In this dissertation, we denote $\mathbf{1}_N$ as a length N column vector of all 1s. Superscripts of variables as in X^k, X^{k+1} denote discrete time. We denote the j-th column of matrix X^k as $X^k_{\cdot j}$. For any vector x, we denote by x^T the transpose of x, and x_i as the i-th component of x. For a random vector $y \in \mathbf{R}^n$, we denote one sample of the y as $\tilde{y}$. For a differentiable Lagrangian function $\mathcal{L}(x, y)$, we denote $\Delta_x \mathcal{L}(x, y)$ as its partial derivative over x.

2.1. Strong Duality of Convex Optimization

We briefly review the strong duality property in convex optimization literature [17], and the proofs of equivalent computationally tractable forms of the (distributionally) robust resource allocation problem are based on strong duality. In the following we describe a general standard form optimization problem and its dual, while concrete formulations of both the primal and dual problems will be defined in the following chapters of this dissertation.

Consider a standard form convex optimization problem [17]

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leqslant 0, i = 1, \ldots, m \\ & h_j(x) = 0, j = 1, \ldots, p, \end{aligned} \tag{2.1}$$

with variable $x \in \mathbb{R}^n$, and nonempty domain $x \in \mathcal{D}$. The Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ associated with the primal problem (2.1) is defined as

$$L(x, \lambda, v) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} v_j h_j(x),$$

with domain $L = \mathcal{D} \times R^m \times R^p$. We refer to the dual variables λ_i and v_i as the Lagrange multiplier associated with the ith inequality constraint $f_i(x) \leqslant 0$ and equality constraint $h_j(x) = 0$, respectively.

Then the (Lagrange) dual function $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ is defined as the minimum value of L over x:

$$g(\lambda, v) = \inf_{x \in \mathcal{D}} L(x, \lambda, v) = \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} v_j h_j(x) \right).$$

The Lagrange dual problem associated with problem (2.1) is then

$$\begin{aligned} &\text{maximize} \quad g(\lambda, v) \\ &\text{subject to} \quad \lambda \geqslant 0. \end{aligned} \tag{2.2}$$

We denote p^* as the optimal value of primal problem (2.1), and d^* as the optimal value of the dual problem (2.2). The property of weak duality always hold for the d^* and p^*, that the optimal value of the Lagrange dual problem is the best lower bound of the optimal value of the primal problem (2.1),i.e.,

$$d^* \leqslant p^*.$$

The difference $p^* - d^*$ is defined as the optimal duality gap of problem (2.1), and the gap value is always nonnegative.

2.1.1. Slater's Constraint Qualification

When the primal convex problem (2.1) satisfies that the equality constraints are affine, or $h_j(x) = 0$ is specified as the form $Ax = b$, the Slater's condition is defined as: there exists an $x \in relint\mathcal{D}$ such that

$$f_i(x) < 0, i = 1, \ldots, m, \quad Ax = b. \tag{2.3}$$

If the first l constraint functions $f_1, \ldots, f_l$ are affine, then the Slater's condition can be refined as: there exists an $x \in relint\mathcal{D}$ such that

$$f_i(x) < 0, i = 1, \ldots, l, \quad f_i(x) < 0, i = l+1, \ldots, m, \quad Ax = b. \tag{2.4}$$

When a primal convex problem (2.1) satisfies Slater's condition, strong duality holds [17], and we have

$$d^* = p^*. \tag{2.5}$$

Proof and examples of convex primal and dual problems when strong duality holds are given in book [17]; for more details about strong duality please refer to it.

It is worth noting that when the primal problem is a convex maximization problem, then the dual problem is a minimization form. The process of finding the dual form is similar as defined above.

2.2. Hypothesis Testing

We briefly review the general process of a hypothesis testing that designed for i.i.d. samples. The algorithms of building demand uncertain sets in this dissertation are based on hypothesis testing.

Hypothesis testing is a widely applied technique to examine the property of a data set [48]. A hypothesis testing starts from a given null-hypothesis H_0 that makes a claim about an unknown distribution $\mathbb{P}^*$, and we need to decide whether to accept H_0 or reject it, based on a data set $\mathcal{S}$ drawn from $\mathbb{P}^*$. The fact that a null-hypothesis is false means there is no sufficient evidence to determine its validity.

A typical test designs a statistic $T \equiv T(\mathcal{S}, H_0)$, and a threshold $\Gamma \equiv \Gamma(\alpha_h, \mathcal{S}, H_0)$, where α_h is a given significance level for data $\mathcal{S}$ on hypothesis H_0. If $T > \Gamma$, we reject H_0. T is also random since it depends on the randomly sampled data $\mathcal{S}$. The threshold Γ is the value that with a probability at most α_h, H_0 will be incorrectly rejected with respect to samples $\mathcal{S}$. Values of $\alpha =$

$1\%, 5\%, 10\%, 20\%$ are common in applications, but it can be set according to specific requirements.

CHAPTER 3 : Real-Time Resource Allocation in Smart Cities: A Receding Horizon Control Approach

3.1. Introduction

Traditional taxi systems in metropolitan areas often suffer from inefficiencies due to uncoordinated actions as system capacity and customer demand change. With the pervasive deployment of networked sensors in modern vehicles, large amounts of information regarding customer demand and system status can be collected in real time. This information provides opportunities to perform various types of control and coordination for large-scale intelligent transportation systems. In this chapter, we present a receding horizon control (RHC) framework to dispatch taxis, which incorporates highly spatiotemporally correlated demand/supply models and real-time GPS location and occupancy information. The objectives include matching spatiotemporal ratio between demand and supply for service quality with minimum current and anticipated future taxi idle driving distance. Extensive trace-driven analysis with a data set containing taxi operational records in San Francisco shows that our solution reduces the average total idle distance by 52%, and reduces the supply demand ratio error across the city during one experimental time slot by 45%. Moreover, our RHC framework is compatible with a wide variety of predictive models and optimization problem formulations. This compatibility property allows us to solve robust optimization problems with corresponding demand uncertainty models that provide disruptive event information.

The rest of this chapter is organized as follows. The background of taxi monitoring system and control problems are introduced in Section 3.3. The taxi dispatch problem is formally formulated in Section 3.4, followed by the RHC framework design in Section 3.5 and a multi-level dispatch framework in Section 3.5.2. A case study with a real taxi data set from San Francisco to evaluation the RHC framework is shown in Section 3.6.

3.2. Related Work

There are three categories of research topics related to the work of this chapter: taxi dispatch systems, transportation system modeling, and multi-agent coordination and control.

A number of recent works study approaches of taxi dispatching services or allocating transportation resources in modern cities. Zhang and Pavone [92] designed an optimal rebalancing method for autonomous vehicles, which considers both global service fairness and future costs, but they didn't take idle driving distance and real-time GPS information into consideration. Truck schedule methods to reduce costs of idle cruising and missing tasks are designed in the temporal perspective in work [87], but the real-time location information is not utilized in the algorithm. Seow et.al focus on minimizing total customer waiting time by concurrently dispatching multiple taxis and allowing taxis to exchange their booking assignments [78]. A shortest time path taxi dispatch system based on real-time traffic conditions is proposed by Lee et.al [47]. In [76, 43, 75], authors aim to maximize drivers' profits by providing routing recommendations. These works give valuable results, but they only consider the current passenger requests and available taxis. Our design uses receding horizon control to consider both current and predicted future requests.

Various mobility and vehicular network modeling techniques have been proposed for transportation systems [22, 15]. Researchers have developed methods to predict travel time [34, 41] and traveling speed [5], and to characterize taxi performance features [49]. A network model is used to describe the demand and supply equilibrium in a regulated market is investigated [86]. These works provide insights to transportation system properties and suggest potential enhancement on transportation system performance. Our design takes a step further to develop dispatch methods based on available predictive data analysis.

There is a large number of works on mobility coordination and control. Different from taxi services, these works usually focus on region partition and coverage control so that coordinated agents can perform tasks in their specified regions [25, 3, 42]. Aircraft dispatch system and air traffic management in the presence of uncertainties have been addressed [9, 83], while the task models

and design objectives are different from taxi dispatching problem. Also, receding horizon control (RHC) has been widely applied for process control, task scheduling, and multi-agent transportation networks [64, 46, 50]. These works provide solid results for related mobility scheduling and control problems. However, none of these works incorporates both the real-time sensing data and historical mobility patterns into a receding horizon control design, leveraging the taxi supply based on the spatiotemporal dynamics of passenger demand.

Remark 1 *The results from this chapter have been captured in [59, 56].*

3.3. Taxi Dispatch Problem: Motivation and System

Taxi networks provide a primary transportation service in modern cities. Most street taxis respond to passengers' requests on their paths when passengers hail taxis on streets. This service model has successfully served up to 25% public passengers in metropolitan areas, such as San Francisco and New York [39, 65]. However, passenger's waiting time varies at different regions of one city and taxi service is not satisfying. In the recent years, "on demand" transportation service providers like Uber and Lyft aim to connect a passenger directly with a driver to minimize passenger's waiting time. This service model is still uncoordinated, since drivers may have to drive idly without receiving any requests, and randomly traverse to some streets in hoping to receive a request nearby based on

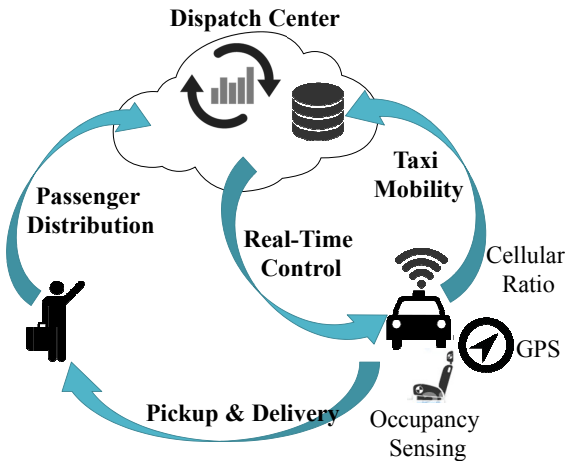


Figure 2: A prototype of the taxi dispatch system

experience.

Our goal in this work is a centralized dispatch framework to coordinate service behavior of large-scale taxi Cyber-Physical system. The development of sensing, data storage and processing technologies provide both opportunities and challenges to improve existing taxi service in metropolitan areas. Figure 14 shows a typical monitoring infrastructure, which consists of a dispatch center and a large number of geographically distributed sensing and communication components in each taxi. The sensing components include a GPS unit and a trip recorder, which provides real-time geographical coordinates and occupancy status of every taxi to the dispatch center via cellular radio. The dispatch center collects and stores data. Then, the monitoring center runs the dispatch algorithm to calculate a dispatch solution and sends decisions to taxi drivers via cellular radio. Drivers are notified over the speaker or on a special display.

Given both historical data and real-time taxi monitoring information described above, we are capable to learn spatiotemporal characteristics of passenger demand and taxi mobility patterns. This paper focuses on the dispatch approach with the model learned based on either historical data or streaming data. One design requirement is balancing spatiotemporal taxi supply across the whole city from the perspective of system performance. It is worth noting that heading to the allocated position is part of idle driving distance for a vacant taxi. Hence, there exists trade-off between the objective of matching supply and demand and reducing total idle driving distance. We aim at a scalable control framework that directs vacant taxis towards demand, while balancing between minimum current and anticipated future idle driving distances.

3.4. Taxi Dispatch Problem Formulation

Informally, the goal of our taxi dispatch system is to schedule vacant taxis towards predicted passengers both spatially and temporally with minimum total idle mileage. We use supply demand ratio of different regions within a time period as a measure of service quality, since sending more taxis for more requests is a natural system-level requirement to make customers at different locations equally served. Similar service metric of service node utilization rate has been applied in resource

allocation problems, and autonomous driving car mobility control approach [92].

The dispatch center receives real-time sensing streaming data including each taxi's GPS location and occupancy status with a time stamp periodically. The real-time data stream is then processed at the dispatch center to predict the spatiotemporal patterns of passenger demand. Based on the prediction, the dispatch center calculates a dispatch solution in real-time, and sends decisions to vacant taxis with dispatched regions to go in order to match predicted passenger demands.

Besides balancing supply and demand, another consideration in taxi dispatch is minimizing the total idle cruising distance of all taxis. A dispatch algorithm that introduces large idle distance in the future after serving current demands can decrease total profits of the taxi network in the long run. Since it is difficult to perfectly predict the future of a large-scale taxi service system in practice, we use a heuristic estimation of idle driving distance to describe anticipated future cost associated with meeting customer requests. Considering control objectives and computational efficiency, we choose a receding horizon control approach. We assume that the optimization time horizon is T, indexed by $k = 1, \ldots, T$, given demand prediction during time $[1, T]$.

3.4.1. Supply and demand in taxi dispatch

We assume that the entire area of a city is divided into n regions such as administrative sub-districts. We also assume that within a time slot k, the total number of passenger requests at the j-th region is denoted by r_j^k. We also use $r^k \triangleq [r_1^k, \ldots, r_n^k] \in \mathbb{R}^{1\times n}$ to denote the vector of all requests. These are the demands we want to meet during time $k = 1, \ldots, T$ with minimal idle driving cost. Then the total number of predicted requests in the entire city is denoted by

$$R^k = \sum_{j=1}^{n} r_j^k.$$

We assume that there are total N vacant taxis in the entire city that can be dispatched according to the real-time occupancy status of all taxis. The initial supply information consists of real-time GPS position of all available taxis, denoted by $P^0 \in \mathbb{R}^{N\times 2}$, whose i-th row $P_i^0 \in \mathbb{R}^{1\times 2}$ corresponds to

Parameters	Description
N	the total number of vacant taxis
n	the number of regions
$r^k \in \mathbb{R}^{1\times n}$	the total number of predicted requests to be served by current vacant taxis at each region
$C^k \in [0,1]^{n\times n}$	matrix that describes taxi mobility patterns during one time slot
$P^0 \in \mathbb{R}^{N\times 2}$	the initial positions of vacant taxis provided by GPS data
$W_i \in \mathbb{R}^{n\times 2}$	preferred positions of the i-th taxi at n regions
$\alpha \in \mathbb{R}^N$	the upper bound of distance each taxi can drive for balancing the supply
$\beta \in \mathbb{R}_+$	the weight factor of the objective function
$R^k \in \mathbb{R}_+$	total number of predicted requests in the city
Variables	Description
$Y^k \in \{0,1\}^{N\times n}$	the dispatch order matrix that represents the region each vacant taxi should go
$P^k \in [0,1]^{N\times n}$	predicted positions of dispatched taxis at the end of time slot k
$d_i^k \in \mathbb{R}_+$	lower bound of idle driving distance of the i-th taxi for reaching the dispatched location

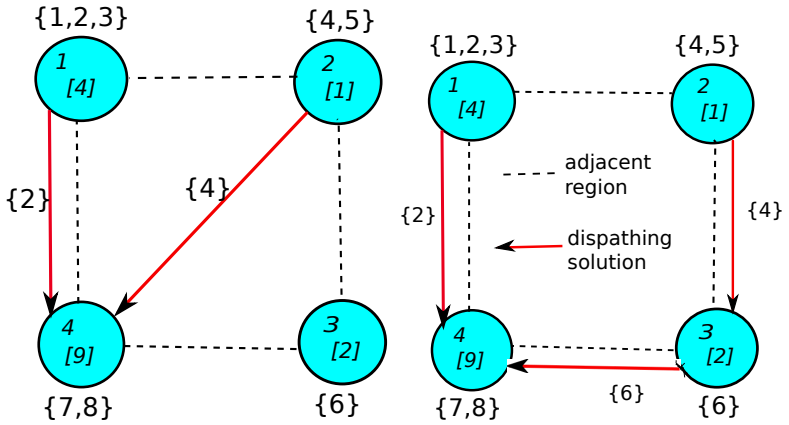
Table 1: Parameters and variables of the RHC problem (3.8).

the position of the i-th vacant taxi. While the dispatch algorithm does not make decisions for occupied taxis, information of occupied taxis affects the predicted total demand to be served by vacant taxis, and the interaction between the information of occupied taxis and our dispatch framework will be discussed in section 3.5.

The basic idea of the dispatch problem is illustrated in Figure 3. Specifically, each region has a predicted number of requests that need to be served by vacant taxis, as well as locations of all vacant taxis with IDs given by real-time sensing information. We would like to find a dispatch solution that balances the supply demand ratio, while satisfying practical constraints and not introducing large current and anticipated future idle driving distance. Once dispatch decisions are sent to vacant taxis, the dispatch center will wait for future computing a new decision problem until updating sensing information in the next period.

3.4.2. Optimal dispatch under operational constraints

The decision we want to make is the region each vacant taxi should go. With the above initial information about supply and predicted demand, we define a binary matrix $Y^k \in \{0,1\}^{N\times n}$ as the dispatch order matrix, where $Y_{ij}^k = 1$ if and only if the i-th taxi is sent to the j-th region during time



(a) A dispatch solution – taxi 2 goes to region 4, and taxi 4 goes to region 4.

(b) A dispatch solution – taxi 2 goes to region 4, taxi 4 goes to region 3, and taxi 6 goes to region 4.

Figure 3: Unbalanced supply and demand at different regions before dispatching and possible dispatch solutions. A circle represents a region, with a number of predicted requests ($[\cdot]$ inside the circle) and vacant taxis ($\{$ taxi IDs $\}$ outside the circle) before dispatching. A black dash edge means adjacent regions. A red edge with a taxi ID means sending the corresponding vacant taxi to the pointed region according to the predicted demand.

k. Then

$$Y^k \mathbf{1}_n = \mathbf{1}_N, \quad k = 1, \ldots, T$$

must be satisfied, since every taxi should be dispatched to one region during time k.

Two objectives

One design requirement is to fairly serve the customers at different regions of the city — vacant taxis should be allocated to each region according to predicted demand across the entire city during each time slot. To measure how supply matches demand at different regions, we use the metric—supply demand ratio. For region j, its supply demand ratio for time slot k is defined as the total number of vacant taxis decided by the total number of customer requests during time slot k. When the supply demand ratio of every region equals to that of the whole city, we have the following equations for

$j = 1, \ldots, n, k = 1, \ldots, T,$

$$\frac{\mathbf{1}_N^T Y_{\cdot j}^k}{r_j^k} = \frac{N}{R^k}, \iff \frac{\mathbf{1}_N^T Y_{\cdot j}^k}{N} = \frac{r_j^k}{R^k}, \tag{3.1}$$

For convenience, we rewrite equation (3.1) as the following equation about two row vectors

$$\frac{1}{N}\mathbf{1}_N^T Y^k = \frac{1}{R^k} r^k, \quad k = 1, \cdots, T. \tag{3.2}$$

However, equation (3.2) can be too strict if used as a constraint, and there may be no feasible solutions satisfying (3.2). This is because decision variables $Y^k, k = 1, \ldots, T$ are integer matrices, and taxis' driving speed is limited that they may not be able to serve the requests from any arbitrary region during time slot k. Instead, we convert the constraint (3.2) into a soft constraint by introducing a supply-demand mismatch penalty function J_E for the requirement that the supply demand ratio should be balanced across the whole city, and one objective of the dispatch problem is to minimize the following function

$$J_E = \sum_{k=1}^{T} \left\| \frac{1}{N}\mathbf{1}_N^T Y^k - \frac{1}{R^k} r^k \right\|_1. \tag{3.3}$$

The other objective is to reduce total idle driving distance of all taxis. The process of traversing from the initial location to the dispatched region will introduce an idle driving distance for a vacant taxi, and we consider to minimize such idle driving distance associated with meeting the dispatch solutions.

We begin with estimate the total idle driving distance associated with meeting the dispatch solutions. For the convenience of routing process, the dispatch center is required to send the GPS location of the destination to vacant taxis. The decision variable Y^k only provides the region each vacant taxi should go, hence we map the region ID to a specific longitude and latitude position for every taxi. In practice, there are taxi stations on roads in metropolitan areas, and we assume that each taxi has a preferred station or is randomly assigned one at every region by the dispatch system. We denote the

preferred geometry location matrix for the i-th taxi by $W_i \in \mathbb{R}^{n\times 2}$, and $[W_i]_j$, where each row of W_i is a two-dimensional geometric position on the map. The j-th row of W_i is the dispatch position sent to the i-th taxi when $Y_{ij}^k = 1$.

Once Y_i^k is chosen, then the i-th taxi will go to the location $Y_i^k W_i$, because the following equation holds

$$Y_i^k W_i = \sum_{q\neq j} Y_{iq}^k [W_i]_q + Y_{ij}^k [W_i]_j = [W_i]_j \in \mathbb{R}^{1\times 2}.$$

With a binary vector Y_i^k that $Y_{ij}^k = 1$, $Y_{iq}^k = 0$ for $q \neq j$, we have $Y_{iq}^k W_i = [0\ 0]$ for $q \neq j$. Since W_i does not need to change with time k, the preferred location of each taxi at every region in the city is stored as a matrix $\mathbf{W}$, stored in the dispatch center before the process of calculating dispatch solutions starts. When updating information of vacant taxis, matrix W_i is also updated for every current vacant taxi i.

The initial position P_i^0 is provided by GPS data. Traversing from position P_i^0 to position $Y_i^1 W_i$ for predicted demand will introduce a cost, since the taxi drives towards the dispatched locations without picking up a passenger. Hence, we consider minimizing the total idle driving distance introduced by dispatching taxis. Driving in a city is approximated as traveling on a grid road. To estimate the distance without knowing the exact path, we use the Manhattan norm or one norm between two geometric positions, which is widely applied as a heuristic distance in path planning algorithms [72]. We define $d_i^k \in \mathbb{R}$ as the estimated idle driving distance of the i-th taxi for reaching the dispatched location $Y_i^k W_i$. For $k = 1$, a lower bound of d_i^1 is given by

$$d_i^1 \geqslant \|P_i^0 - Y_i^1 W_i\|_1, \quad i = 1, \ldots, N. \tag{3.4}$$

For $k \geqslant 2$, to estimate the anticipated future idle driving distance induced by reaching dispatched position $Y_i^k W_i$ at time k, we consider the trip at the beginning of time slot k starts at the end location of time slot $k-1$. However, during time $k-1$, taxis' mobility patterns are related to pick-up and

drop-off locations of passengers, which are not directly controlled by the dispatch center. So we assume the predicted ending position for a pick-up location $Y_i^{k-1}W_i$ during time $k-1$ is related to the starting position $Y_i^{k-1}W_i$ as follows:

$$P_i^{k-1} = f^k(Y_i^{k-1}W_i), \quad f^k : \mathbb{R}^{1\times 2} \to \mathbb{R}^{1\times 2}, \tag{3.5}$$

where f^k is a convex function, called a mobility pattern function. To reach the dispatched location $Y_i^k W_i$ at the beginning of time k from position P_i^{k-1}, the approximated driving distance is

$$d_i^k \geqslant \|f^k(Y_i^{k-1}W_i) - Y_i^k W_i\|_1. \tag{3.6}$$

The process to calculate a lower bound for d_i^k is illustrated in Figure 4.

Within time slot k, the distance that every taxi can drive should be bounded by a constant vector $\alpha^k \in \mathbb{R}^N$:

$$d^k \leqslant \alpha^k.$$

Total idle driving distance of all vacant taxis though time $k = 1, \ldots, T$ to satisfy service fairness is then denoted by

$$J_D = \sum_{k=1}^{T}\sum_{i=1}^{N} d_i^k. \tag{3.7}$$

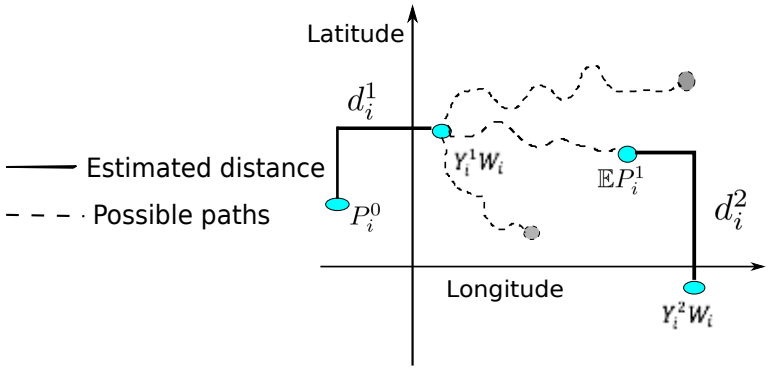


Figure 4: Illustration of the process to estimate idle driving distance to the dispatched location for the i-th taxi at $k = 2$: predict ending location of $k = 1$ denoted by $\mathbb{E}P_i^1$ in (3.9), get the distance between locations $\mathbb{E}P_i^1$ and $Y_i^2 W_i$ denoted by d_i^2 in (3.10).

It is worth noting that the idle distance we estimate here is the region-level distance to pick up predicted passengers — the distance is nonzero only when a vacant taxi is dispatched to a different region. We also require that the estimated distance is a closed form function of the locations of the original and dispatched regions, without knowledge about accurate traffic conditions or exact time to reach the dispatched region. Hence, in this work we use Manhattan norm to approximate the idle distance—it is a closed form function of the locations of the original and dispatched regions. When accessibility information of the road traffic network is considered in estimating street-level distances, for the case that a taxi may not drive on rectangular grids to pick up a passenger (for instance, when a U-turn is necessary), Lee et.al have proposed a shortest time path approach to pick up passengers in shortest time [47].

An RHC problem formulation

Since there exists a trade-off between two objectives as discussed in Section 3.3, we define a weight parameter β^k when summing up the costs related to both objectives. A list of parameters and variables is shown in Table 1. When mixed integer programming is not efficient enough for a large-scale taxi network regarding to the problem size, one standard relaxation method is replacing the constraint $Y_{ij}^k \in \{0, 1\}$ by $0 \leq Y_{ij}^k \leq 1$.

To summarize, we formulate the following problem (3.8) based on the definitions of variables,

parameters, constraints and objective function

$$
\begin{aligned}
\min_{Y^k, d^k} \quad & J = J_E + \beta J_D \\
& = \sum_{k=1}^{T} \left(\left\| \frac{1}{N} \mathbf{1}_N^T Y^k - \frac{1}{R^k} r^k \right\|_1 + \beta^k \sum_{i=1}^{N} d_i^k \right) \\
\text{s.t} \quad & d_i^1 \geqslant \|P_i^0 - Y_i^1 W_i\|_1, \quad i = 1, \ldots, N, \\
& d_i^k \geqslant \|f^k(Y_i^{k-1} W_i) - Y_i^k W_i\|_1, \\
& i = 1, \ldots, N, \quad k = 2, \ldots, T, \\
& d^k \leqslant \alpha^k, \quad k = 1, 2, \ldots, T, \\
& Y^k \mathbf{1}_n = \mathbf{1}_N, \quad k = 1, 2, \ldots, T, \\
& 0 \leqslant Y_{ij}^k \leqslant 1, i \in \{1, \ldots, N\}, j \in \{1, \ldots, n\}.
\end{aligned}
\tag{3.8}
$$

After getting an optimal solution Y^1 of problem (3.8), for the i-th taxi, we may recover binary solution through rounding by setting the largest value of Y_i^1 to 1, and the others to 0. This may violate the constraint of d_i^0, but since we set a conservative upper bound α^k, and the rounding process will return a solution that satisfies $d_i^k \leqslant \alpha^k + \epsilon$ with bounded ϵ, the dispatch solution can still be executed during time slot k.

3.4.3. Discussions on the optimal dispatch formulation

Why use supply demand ratio as a metric

An intuitive measurement of the difference between the number of vacant taxis and predicted total requests at all regions is:

$$
e = \sum_{j=1}^{n} |s_j^k - r_j^k|,
$$

where s_j^k is the total number of vacant taxis sent to the j-th region. However, when the total number of vacant taxis and requests are different in the city, this error e can be large even under the case that more vacant taxis are already allocated to busier regions and fewer vacant taxis are left to regions

with less predicted demand. We do not have an evidence whether the dispatch center already fairly allocates supply according to varying demand given the value of the above error e.

The meaning of α^k

For instance, when the length of time slot k is one hour, and α^k is the distance one taxi can traverse during 20 minutes of that hour, this constraint means a dispatch solution involves the requirement that a taxi should be able to arrive the dispatched position within 20 minutes in order to fulfill predicted requests. With traffic condition monitoring and traffic speed predicting method [5], α^k can be adjusted according to the travel time and travel speed information available for the dispatch system. This constraint also gives the dispatch system the freedom to consider the fact that drivers may be reluctant to drive idly for a long distance to serve potential customers, and a reasonable amount of distance to go according to predicted demand is acceptable. The threshold α^k is related to the length of time slot. In general, the longer a time slot is, the larger α^k can be, because of constraints like speed limit.

One example of mobility pattern function f^k

When taxi's mobility pattern during time slot k is described by a matrix $C^k \in \mathbb{R}^{n\times n}$ satisfying $\sum_{j=1}^{n} C_{ij} = 1$, where C_{ij}^k is the probability that a vacant taxi starts within region i will end within region j during time k. From the queuing-theoretical perspective such probability transition matrix approximately describes passenger's mobility [92]. Given X_i^{k-1} and the mobility pattern matrix $C^{k-1} \in [0,1]^{n\times n}$, the probability of ending at each region for taxi i is

$$p = \sum_{j=1}^{n} [C^{k-1}]_j I(Y_{ij}^{k-1} = 1) = Y_i^{k-1} C^{k-1} \in \mathbb{R}^{1\times n},$$

where the indicator function $I(Y_{ij}^{k-1} = 1) = 1$ if and only if $Y_{ij}^{k-1} = 1$, and $[C^{k-1}]_j$ is the j-th row of C^{k-1}. However, introducing a stochastic formula in the objective function will cause high computational complexity for a large-scale problem. Hence, instead of involving the probability of taking different paths in the objective function to formulate a stochastic optimization problem, we

take the expected value of the position of i-th taxi by the end of time $k-1$

$$P_i^{k-1} = \sum_{j=1}^{n} p_j [W_i]_j = pW_i = Y_i^{k-1} C^{k-1} W_i. \tag{3.9}$$

Here $P_i^{k-1} \in \mathbb{R}^{1\times 2}$ is a vector representing a predicted ending location of the i-th taxi on the map at each dimension. Then a lower bound of idle driving distance for heading to $Y_i^k W_i$ to meet demand during k is given by the distance between P_i^{k-1} defined as (3.9) and $Y_i^k W_i$.

$$d_i^k \geqslant \|(Y_i^{k-1} C^{k-1} - Y_i^k) W_i\|_1. \tag{3.10}$$

In particular, when the transition probability $C^k, k = 1, \ldots, T$ is available, we can replace the constraint about d_i^k by $d_i^k \geqslant \|(Y_i^{k-1} C^{k-1} - Y_i^k) W_i\|_1$.

It is worth noting that d_i^k is a function of Y_i^{k-1} and Y_i^k, and the estimation accuracy of idle driving distance to dispatched positions Y_i^k ($k = 2, \ldots, T$) depends on the predicting accuracy of the mobility pattern during each time slot k, or P_i^{k-1}. The distance d^1 is calculated based on real-time GPS location P^0 and dispatch position Y^1, and we use estimations $d^2, \ldots, d^T$ to measure the anticipated future idle driving distances for meeting requests.

The error of estimated C^k mainly affects the choice of idle distance d^k when the true ending region of a taxi by the end of time slot k is not as predicted based on its starting region at time slot k. This is because C^k determines the constraint for d^k ($k = 2, 3, \ldots, T$) as described by inequality (3.10). However, the system also collects real-time GPS positions to make a new decision based on the current true positions of all taxis, instead of only applying predicted locations provided by the mobility pattern matrix. According to constraint (3.4) distance d^1 is determined by GPS sensing data P^0 and dispatch decision Y^1, and only Y^1 will be executed sent to vacant taxis as the dispatch solutions after the system solving problem (3.8). From this perspective, real-time GPS and occupancy status sensing data is significant to improve the system's performance when we utilize both historical data and real-time sensing data. We also consider the effect of an inaccurate mobility pattern estimation C^k when choosing the prediction time horizon T — large prediction horizon will induce accumu-

lating prediction error in matrix C^k and the dispatch performance will even be worse. Evaluation results in Section 3.6 show how real-time sensing data helps to reduce total idle driving distance and how the mobility pattern error of different prediction horizon T affects the system's performance.

Information on road congestion and passenger destination

When road congestion information is available to the dispatch system, function in (3.5) can be generalized to include real-time congestion information. For instance, there is a high probability that a taxi stays within the same region during one time slot under congestions.

We do not assume that information of passenger's destination is available to the system when making dispatch decisions, since many passengers just hail a taxi on the street or at taxi stations instead of reserving one in advance in metropolitan areas. When the destination and travel time of all trips are provided to the dispatch center via additional software or devices as prior knowledge, the trip information is incorporated to the definition of ending position function (3.5) for problem formulation (3.8). With more accurate trip information, we get a better estimation of future idle driving distance when making dispatch decisions for $k = 1$.

Customers' satisfaction under balanced supply demand ratio

The problem we consider in this work is reaching fair service to increase global level of customers' satisfaction, which is indicated by a balanced supply demand ratio across different regions of one city, instead of minimizing each individual customer's waiting time when a request arrives at the dispatch system. Similar service fairness metric has been applied in mobility on demand systems [92], and supply demand ratio considered as an indication of utilization ratio of taxis is also one regulating objective in taxi service market [86]. For the situation that taxi i will not pick up passengers in its original region but will be dispatched to another region, the dispatch decision results from the fact that global customers' satisfaction level will be increased. For instance, when the original region of taxi i has a higher supply demand ratio than the dispatched region, going to the dispatched region will help to increase customer's satisfaction in that region. By sending taxi i to some

other region, customers' satisfaction in the dispatched region can be increased, and the value of the supply-demand cost-of-mismatch function J_E can be reduced without introducing much extra total idle driving distance J_D.

3.4.4. Robust RHC formulations

Previous work has developed multiple ways to learn passenger demand and taxi mobility patterns [5, 34, 43], and accuracy of the predicted model will affect the results of dispatch solutions. We do not have perfect knowledge of customer demand and taxi mobility models in practice, and the actual spatial-temporal profile of passenger demands can deviate from the predicted value due to random factors such as disruptive events. Hence, we discuss formulations of robust taxi dispatch problems based on (3.8).

Formulation (3.8) is one computationally tractable approach to describe the design requirements with a nominal model. One advantage of the formulation (3.8) is its flexibility to adjust the constraints and objective function according to different conditions. With prior knowledge of scheduled events that disturb the demand or mobility pattern of taxis, we are able to take the effects of the events into consideration by setting uncertainty parameters. For instance, when we have basic knowledge that total demand in the city during time k is about $\tilde{R}^k$, but each region r_j^k belongs to some uncertainty set, denoted by an entry wise inequality

$$R_1^k \preceq r^k \preceq R_2^k,$$

given $R_1^k \in \mathbb{R}^n$ and $R_2^k \in \mathbb{R}^n$. Then

$$r_j^k \in [R_{1j}^k, R_{2j}^k], j = 1, \ldots, n \tag{3.11}$$

is an uncertainty parameter instead of a fixed value as in problem (3.8). Without additional knowledge about the change of total demand in the whole city, we denote $\tilde{R}^k$ as the approximated total demand in the city under uncertain r_j^k for each region. By introducing interval uncertainty (3.11) to

r^k and fixing $\tilde{R}^k$ on the denominator, we have the following robust optimization problem (3.12)

$$\begin{aligned} \min_{Y^k, d^k} \quad & \max_{R_1^k \preceq r^k \preceq R_2^k} \sum_{k=1}^{T} \left(\left\| \frac{1}{N} \mathbf{1}_N^T Y^k - \frac{1}{\tilde{R}^k} r^k \right\|_1 + \beta^k \sum_{i=1}^{N} d_i^k \right) \\ \text{s.t.} \quad & \text{constraints of problem (3.8).} \end{aligned} \tag{3.12}$$

The robust optimization problem (3.12) is computationally tractable, and we have the following Theorem 1 to show the equivalent form to provide real-time dispatch decision.

Theorem 1 *The robust RHC problem* (3.12) *is equivalent to the following computationally efficient convex optimization problem*

$$\begin{aligned} \min_{Y^k, d^k, t^k} \quad & J' = \sum_{k=1}^{T} \left(\sum_{j=1}^{n} t_j^k + \beta^k \sum_{i=1}^{N} d_i^k \right) \\ \text{s.t} \quad & t_j^k \geq \frac{\mathbf{1}_N Y_{\cdot j}^k}{N} - \frac{R_{1j}^k}{\tilde{R}^k}, t_j^k \geq \frac{R_{1j}^k}{\tilde{R}^k} - \frac{\mathbf{1}_N Y_{\cdot j}^k}{N}, \\ & t_j^k \geq \frac{\mathbf{1}_N Y_{\cdot j}^k}{N} - \frac{R_{2j}^k}{\tilde{R}^k}, t_j^k \geq \frac{R_{2j}^k}{\tilde{R}^k} - \frac{\mathbf{1}_N Y_{\cdot j}^k}{N}, \\ & j = 1, \ldots, n, \quad , k = 1, \ldots, T, \\ & \text{constraints of problem (3.8).} \end{aligned} \tag{3.13}$$

Proof 1 *In the objective function, only the first term is related to r^k. To avoid the maximize expression over an uncertain r^k, we first optimize the term over r^k for any fixed Y^k. Let $Y_{\cdot j}^k$ represent the j-th column of Y^k, then*

$$\begin{aligned} & \max_{R_1^k \preceq r^k \preceq R_2^k} \left\| \frac{1}{N} \mathbf{1}_N^T Y^k - \frac{1}{\tilde{R}^k} r^k \right\|_1 \\ = & \max_{R_1^k \preceq r^k \preceq R_2^k} \sum_{j=1}^{n} \left| \frac{1}{N} \mathbf{1}_N^T Y_{\cdot j}^k - \frac{r_j^k}{\tilde{R}^k} \right| = \sum_{j=1}^{n} \max_{r_j^k \in [R_{1j}^k, R_{2j}^k]} \left| \frac{1}{N} \mathbf{1}_N^T Y_{\cdot j}^k - \frac{r_j^k}{\tilde{R}^k} \right|. \end{aligned} \tag{3.14}$$

The second equality holds because each r_j^k can be optimized separately in this equation. For $R_{1j}^k \leq$

$r_j^k \leq R_{2j}^k$, *we have*

$$\frac{R_{1j}^k}{\tilde{R}^k} \leq \frac{r_j^k}{\tilde{R}^k} \leq \frac{R_{2j}^k}{\tilde{R}^k}.$$

Then the problem is to maximize each absolute value in (3.14) *for* $j = 1, \ldots, n$. *Consider the following problem for* $x, a, b \in \mathbb{R}$ *to examine the character of maximization problem over an absolute value:*

$$\begin{aligned}\max_{x_0 \in [a,b]} |x - x_0| &= \begin{cases} |x-a|, & \text{if } x > (a+b)/2 \\ |x-b|, & \text{otherwise} \end{cases} \\ &= \max\{|x-a|, |x-b|\} = \max\{x-a, a-x, x-b, b-x\}.\end{aligned}$$

Similarly, for the problem related to r_j^k, *we have*

$$\max_{r_j^k \in [R_{1j}^k, R_{2j}^k]} \left| \frac{\mathbf{1}_N Y_{\cdot j}^k}{N} - \frac{r_j^k}{\tilde{R}^k} \right| = \max \left\{ \left| \frac{\mathbf{1}_N Y_{\cdot j}^k}{N} - \frac{R_{1j}^k}{\tilde{R}^k} \right|, \quad \left| \frac{\mathbf{1}_N Y_{\cdot j}^k}{N} - \frac{R_{2j}^k}{\tilde{R}^k} \right| \right\}. \tag{3.15}$$

Thus, with slack variables $t^k \in \mathbb{R}^n$, *we re-formulate the robust RHC problem as* (3.13).

Taxi mobility patterns during disruptive events can not be easily estimated (in general), however, we have knowledge such as a rough number of people are taking part in a conference or competition, or even more customer reservations because of events in the future. The uncertain set of predicted demand r^k can be constructed purely from empirical data such as confidence region of the model, or external information about disruptive events. By introducing extra knowledge besides historical data model, the dispatch system responds to such disturbances with better solutions than the those without considering model uncertainties. Comparison of results of (3.13) and problem (3.8) is shown in Section 3.6.

3.5. RHC Framework Design

Demand and taxi mobility patterns can be learned from historical data, but they are not sufficient to calculate a dispatch solution with dynamic positions of taxis when the positions of the taxis change in real time. Hence, we design an RHC framework to adjust dispatch solutions according to real-time sensing information in conjunction with the learned historical model. Real-time GPS and occupancy information then act as feedback by providing the latest taxi locations, and demand-predicting information for an on-line learning method like [91, 62]. Solving problem (3.8) or (3.12) is the key iteration step of the RHC framework to provide dispatch solutions.

RHC works by solving the cost optimization over the window $[1, T]$ at time $k = 1$. Though we get a sequence of optimal solutions in T steps – $X^1, \ldots, X^T$, we only send dispatch decisions to vacant taxis according to X^1. We summarize the complete process of dispatching taxis with both historical and real-time data as Algorithm 1, followed by a detail computational process of each iteration. The lengths of time slots for learning historical models (t_1) and updating real-time information (t_2) do not need to be the same, hence in Algorithm 1 we consider a general case for different t_1, t_2.

3.5.1. RHC Algorithm

Remark 2 *Predicted values of requests $\hat{r}(h_1)$ depend on the modeling method of the dispatch system. For instance, if the system only applies historical data set to learn each $\hat{r}(h_1)$, $\hat{r}(h_1)$ is not updated with real-time sensing data. When the system applies online training method such as [91] to update $\hat{r}(h_1)$ for each h_1, values of r, r^k are calculated based on the real-time value of $\hat{r}(h_1)$.*

Update r

We receive sensing data of both occupied and vacant taxis in real-time. Predicted requests that vacant taxis should serve during h_1 is re-estimated at the beginning of each h_1 time. To approximate the service capability when an occupied taxi turns into vacant during time h_1, we define the total number of drop off events at different regions as a vector $dp(h_1) \in \mathbb{R}^{n\times 1}$. Given $dp(h_1)$, the

Algorithm 1: RHC Algorithm for real-time taxi dispatch

Inputs: Time slot length t_1 minutes, period of sending dispatch solutions t_2 minutes (t_1/t_2 is an integer); a preferred station location table $\mathbf{W}$ for every taxi in the network; estimated request vectors $\hat{r}(h_1)$, $h_1 = 1, \ldots, 1440/t_1$, mobility patterns $\hat{f}(h_2)$, $h_2 = 1, \ldots, 1440/t_2$; prediction horizon $T \geq 1$.

Initialization: The predicted requests vector $r = \hat{r}(h_1)$ for corresponding algorithm start time h_1.

while *Time is the beginning of a* t_2 *time slot* **do**

(1) Update sensor information for initial position of vacant taxis P^0 and occupied taxis P'^0, total number of vacant taxis N, preferred dispatch location matrices W_i.

if *time is the beginning of an* h_1 *time slot* **then**

Calculate $\hat{r}(h_1)$ if the system applies an online training method; count total number of occupied taxis $n_o(h_1)$; update vector r.

end 2

Update the demand vectors r^k based on predicted demand $\hat{r}(h_1)$ and potential service ability of $n_o(h_1)$ occupied taxis; update mobility functions $f^k(\cdot)$ (for example, C^k), set up values for idle driving distance threshold α^k and objective weight β^k, $k = 1, 2, \ldots, T$.

(3) **if** *there is knowledge of demand* r^k *as an uncertainty set ahead of time* **then**

solve problem (3.13);

else

solve problem (3.8) for a certain demand model;

end 4

Send dispatch orders to vacant taxis according to the optimal solution of matrix X^1. Let $h_2 = h_2 + 1$.

end

Return: Stored sensor data and dispatch solutions.

probability that a drop off event happens at region j is

$$pd_j(h_1) = dp_j(h_1)/\mathbf{1}_n dp(h_1), \tag{3.16}$$

where $dp_j(h_1)$ is the number of drop off events at region j during h_1. We assume that an occupied taxi will pick up at least one passenger within the same region after turning vacant, and we approximate future service ability of occupied taxis at region j during time h_1 as

$$r_{oj}(h_1) = \lceil pd_j(h_1) \times n_o(h_1) \rceil, \tag{3.17}$$

where $\lceil \cdot \rceil$ is the ceiling function, $n_o(h_1)$ is the total number of current occupied taxis at the beginning of time h_1 provided by real-time sensor information of occupied taxis. Let

$$r = \hat{r}(h_1) - r_o(h_1),$$

then the estimated service capability of occupied taxis is deducted from r for time slot h_1.

Update r^k for problem (3.8)

We assume that requests are uniformly distributed during h_1. Then for each time k of length t_2, if the corresponding physical time is still in the current h_1 time slot, the request is estimated as an average part of r; else, it is estimated as an average part for time slot $h_1 + 1, h_1 + 2, \ldots$, etc. The rule of choosing r^k is

$$r^k = \begin{cases} \frac{1}{H} r, & \text{if } (k + h_2 - 1)t_2 \leq h_1 t_1 \\ \frac{1}{H} \hat{r} \left(\left\lceil \frac{(k+h_2-1)t_2}{t_1} \right\rceil \right), & \text{otherwise} \end{cases}$$

where $H = t_1/t_2$.

Update r^k for robust dispatch (3.13)

When there are disruptive events and the predicted requests number is a range $\hat{r}(h_1) \in [\hat{R}_1(h_1), \hat{R}_2(h_1)]$, similarly we set the uncertain set of r^k as the following interval for the computationally efficient form of robust dispatch problem (3.13)

$$r^k \in \begin{cases} \frac{1}{H}\left[\hat{R}_1(h_1) - r_o(h_1), \hat{R}_2(h_1) - r_o(h_1)\right], & \text{if } (k+h_2-1)t_2 \leq h_1 t_1, \\ \frac{1}{H}\left[\hat{R}_1(\left\lceil \frac{(k+h_2-1)t_2}{t_1} \right\rceil), \hat{R}_2(\left\lceil \frac{(k+h_2-1)t_2}{t_1} \right\rceil)\right], & \text{o.w.} \end{cases}$$

Spatial and temporal granularity of Algorithm 1

The main computational cost of each iteration is on step (3), and t_2 should be no shorter than the computational time of the optimization problem. We regulate parameters according to experimental results based on a given data set, since there are no closed form equations to decide optimal design values of these parameters.

For the parameters we estimate from a given GPS dataset, the method we use in the experiments (but not restricted to it) will be discussed in Section 3.6. The length of every time slot depends on the predict precision of prediction, desired control outcome, and the available computational resources. We can set a large time horizon to consider future costs in the long run. However, in practice we do not have perfect predictions, thus a large time horizon may amplify the prediction error over time. Applying real-time information to adjust taxi supply is a remedy to this problem. Modeling techniques are beyond the scope of this work. If we have perfect knowledge of customer demand and taxi mobility models, we can set a large time horizon to consider future costs in the long run. However, in practice we do not have perfect predictions, thus a large time horizon may amplify the prediction error over time. Likewise, if we choose a small look-ahead horizon, then the dispatch solution may not count on idle distance cost of the future. Applying real-time information to adjust taxi supply is a remedy to this problem. With an approximated mobility pattern matrix C^k, the dispatch solution with large T is even worse than small T.

Selection process of parameters β^k, α^k, and T

The process of choosing values of parameters for Algorithm 1 is a trial and adjusting process, by increasing/decreasing the parameter value and observing the changing trend of the dispatch cost, till a desired performance is reached or some turning point occurs that the cost is not reduced any more. For instance, objective weight β^k is related to the objective of the dispatch system, whether it is more important to reach fair service or reduce total idle distance. Some parameter is related to additional information available to the system besides real-time GPS and occupancy status data; for instance, α^k can be adjusted according to the average speed of vehicles or traffic conditions during time k as discussed in subsection 3.4.3. Adjustments of parameters such as objective weight β^k, idle distance threshold α^k, prediction horizon T when considering the effects of model accuracy, control objectives are shown in Section 3.6. A formal parameter selection method is a direction for future work.

3.5.2. Multi-level Dispatch framework

We do not assume that the customer demand is provided to the RHC framework in the previous session and only require that demand-related data is available for predicting the future service requirements. Furthermore, we do not restrict the data source of customer demand – it can be either predicted results or existing reservation records in the system. Some companies provide taxi service according to the current requests in the queue. For reservations received by the dispatch center ahead of time, the RHC framework in Algorithm 1 is compatible with this type of demand information — we then assign value of the waiting requests vector r^k, taxi mobility function f^k in (3.8) according to the reservations, and the solution is subject to customer bookings.

For customer requests received in real-time, a multi-level dispatch framework is available based on Algorithm 1. The process is as follows: run Algorithm 1 with predicted demand r^k, and send dispatch solutions to vacant taxis. When vacant taxis arrive at dispatched locations, the dispatch center updates real-time demand such as bookings that recently appear in the system. Then suboptimal dispatch or matching algorithm based on current demand such as the algorithm designed by Lee *et*

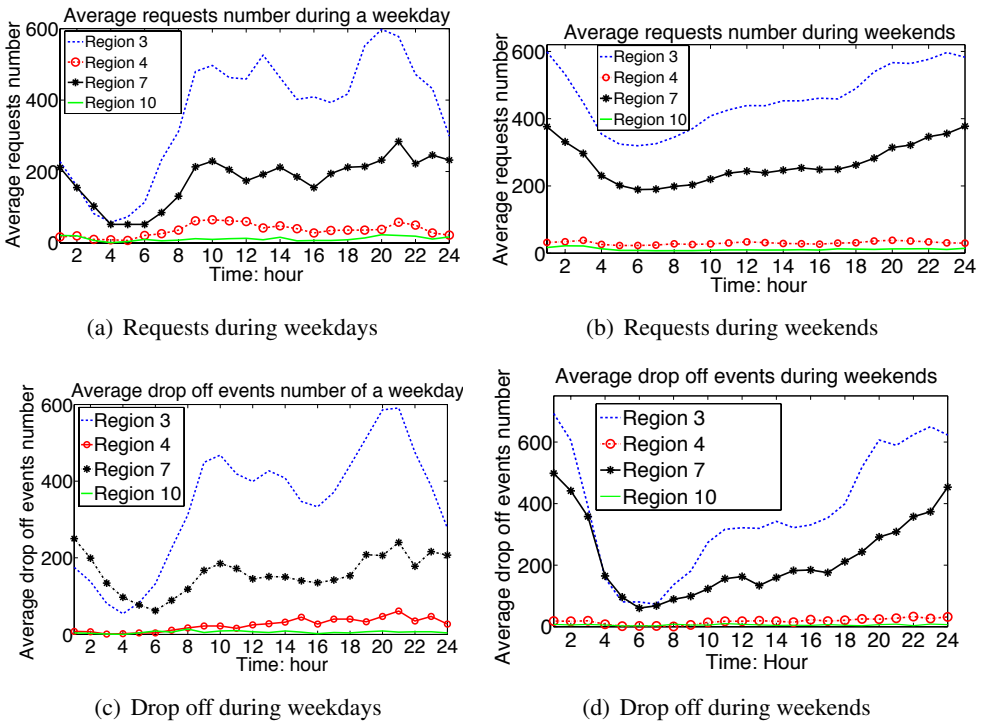
Taxicab GPS Data set			
Collection Period	Number of Taxis	Data Size	Record Number
05/17/08-06/10/08	500	$90MB$	$1,000,000$
Format			
	ID	Status	Direction
	Date and Time	Speed	GPS Coordinates

Table 2: San Francisco Data in the Evaluation Section. Giant baseball game in AT&T park on May 31, 2008 is a disruptive event we use for evaluating the robust optimization formulation.

al. [47] and Ma *et.al* [80] can be applied.

By this multi-level dispatch framework, vacant taxis are first pre-dispatched at a regional level according to predicted demand using the RHC framework. After arriving the dispatched regions, specific locations to pick up a passenger who just booked a taxi is sent to a vacant taxi. The lower level picking up decisions is a one-to-one (or multi-to one under carpooling strategies) matching between passengers and drivers. Each vacant taxi is assigned to one or multiple booking within its current region according to a heuristic or matching algorithm such as [47, 80, 1], with the benefit of real-time traffic conditions. Since previous work usually belongs to the area of heuristic, greedy dispatching algorithms or matching algorithms, we do not present or restrict a specific lower level vacant taxi allocating approach to the RHC approach designed in this dissertation.

Previous work of routing algorithms for mobility-on-demand autonomous vehicle systems [93, 81, 92, 69] or ride-sharing algorithms for taxi/autonomous vehicle systems [1] usually assumes that the trip of each request is provided to the vehicle dispatch center. Even involving a model predictive control process, the authors assume the demand trip information is given [93, 69]. In contrast, the RHC framework designed in this work does not rely on priority knowledge of the demand or mobility pattern—instead of making assumptions about the demand model, it provides an exact process of incorporating the model predicted based on historical/streaming data to calculate a system-level optimal dispatch decisions.



(a) Requests during weekdays

(b) Requests during weekends

(c) Drop off during weekdays

(d) Drop off during weekends

Figure 5: Requests at different hours during weekdays and weekends, for four selected regions. A given historical data set provides basic spatiotemporal information about customer demands, which we utilize with real-time data to dispatch taxis.

3.6. Case Study: Method Evaluation

We conduct trace-driven simulations based on a San Francisco taxi data set [74] summarized in Table 2. In this data set, a record for each individual taxi includes four entries: the geometric position (latitude and longitude), a binary indication of whether the taxi is vacant or with passengers, and the Unix epoch time. With these records, we learn the average requests and mobility patterns of taxis, which serve as the input of Algorithm 1. We note that our learning model is not restricted to the data set used in this simulation, and other models [91] and date sets can also be incorporated.

We implement Algorithm 1 in Matlab using the optimization toolbox called CVX [37]. We assume that all vacant taxis follow the dispatch solution and go to suggested regions. Inside a target region, we assume that a vacant taxi automatically picks up the nearest request recorded by the trace data, and we calculate the total idle mileage including distance across regions and inside a region by

simulation. When taxis are autonomous vehicles, this assumption will not be violated at all; for the case of human drivers, the incentive design problem that motivates the drivers to follow the dispatch suggestion is a venue of future work. The evaluation result in this work aims to show that the RHC framework we design indeed improve the system performance, and then it is valuable to implement this method in the real world.

The trace data records the change of GPS locations of a taxi in a relatively small time granularity such as every minute. Moreover, there is no additional information about traffic conditions or the exact path between two consecutive data points when they were recorded. Hence, we consider the path of each taxi as connected road segments determined by each two consecutive points of the trace data we use in this section. Assume the latitude and longitude values of two consecutive points in the trace data are $[l_{x1}, l_{y1}]$ and $[l_{x2}, l_{y2}]$, for a short road segment, the mileage distance between the two points (measured in one minute) is approximated as being proportional to the value $(|l_{x1} - l_{x2}| + |l_{y1} - l_{y2}|)$. The geometric location of a taxi is directly provided by GPS data. Hence, we calculate geographic distance directly from the data first, and then convert the result to mileage.

Experimental figures shown in Subsection 3.6.2 and 3.6.4 are average results of all weekday data from the data set 2. Results shown in Subsection 3.6.3 are based on weekend data.

3.6.1. Predicted demand based on historical data

Requests during different times of a day in different regions vary a lot, and Figure 3.5(a) and Figure 3.5(b) compare bootstrap results of requests $\hat{r}(h_1)$ on weekdays and weekends for selected regions. This shows a motivation of this work— necessary to dispatch the number of vacant taxis according to the demand from the perspective of system-level optimal performance. The detailed process is described as follows.

The original SF data set does not provide the number of pick up events, hence one intuitive way to determine a pick up (drop off) event is as follows. When the occupancy binary turns from 0 to 1 (1 to 0), it means a pick up (drop off) event has happened. Then we use the corresponding geographical position to determine which region this pick up (drop off) belongs to; use the time stamp data to

decide during which time slot this pick up (drop off) happened.

After counting the total number of pick up and drop off events during each time slot at every region, we obtain a set of vectors $r_{d'}(h_k), dp_{d'}(h_k), d' = 1, \ldots, d$, where d is the number of days for historical data . In the following analysis, each time slot h_1 is the time slot of predicting demand model chosen by the RHC framework. The SF data set includes about 24 days of data, so we use $d = 18$ for weekdays, and $d = 6$ for weekends. The bootstrap process for a given sample time number $B = 1000$ is given as follows.

(a) Randomly sample a size d dataset with replacement from the data set $\{r_1(h_1), \ldots, r_d(h_1)\}$, calculate

$$\hat{r}^1(h_1) = \frac{1}{d} \sum_{d'=1}^{d} r_{d'}(h_1), \text{ for } h_1 = 1, \ldots, 1440/h_1.$$

(b) Repeat step (a) for $(B-1)$ times, so that we have B estimates for each h_1,

$$\hat{r}^b(h_1), \quad b = 1, \ldots, B.$$

The estimated mean value of $\hat{r}(h_1)$ based on B samples is

$$\hat{r}(h_1) = \frac{1}{B} \sum_{l=1}^{B} \hat{r}^l(h_1).$$

(c) Calculate the sample variance of the B estimates of $r(h_1)$ for each h_1,

$$\hat{V}_{\hat{r}(h_1)} = \frac{1}{B} \sum_{b=1}^{B} (\hat{r}^b(h_1) - \frac{1}{B} \sum_{l=1}^{B} \hat{r}^l(h_1)). \tag{3.18}$$

To estimate the demand range for robust dispatch problem (3.13) according to historical data, we construct the uncertain set of demand r^k based on the mean and variance of the bootstrapped demand

Region ID	1	2	3	4	5	6	7	8
Transit probability	0.0032	0.0337	0.5144	0.0278	0.0132	0.0577	0.1966	0.0263
Region ID	9	10	11	12	13	14	15	16
Transit probability	0.0001	0.0050	0.0340	0.0136	0.0018	0.0082	0.0248	0.0396

Table 3: An estimation of state transition matrix by bootstrap: one row of matrix $\hat{C}(h_k)$

model. For every region j, the boundary of demand interval is defined as

$$\begin{aligned} \tilde{R}_{1,j}(h_1) &= \hat{r}_j(h_1) - \sqrt{\hat{V}_{\hat{r}(h_1),j}}, \\ \tilde{R}_{2,j}(h_1) &= \hat{r}_j(h_1) + \sqrt{\hat{V}_{\hat{r}(h_1),j}}, \end{aligned} \tag{3.19}$$

where $\hat{r}_j(h_1)$ is the average value of each step (b) and $\hat{V}_{\hat{r}(h_1),j}$ is the variance of estimated request number defined in (3.18). This one standard deviation range is used for evaluating the performance of robust dispatch framework in this work.

Estimated drop off events vectors $dp(h_1)$ are also calculated via a similar process. Figure 3.5(c) and 3.5(d) show bootstrap results of passenger drop off events $dp(h_1)$ on weekdays and weekends for selected regions.

For evaluation convenience, we partition the city map to regions with equal area. To get the longitude and latitude position $W_i \in \mathbb{R}^{n\times 2}$ of vacant taxi i, we randomly pick up a station position in the city drawn from the uniform distribution.

3.6.2. RHC with real-time sensor information

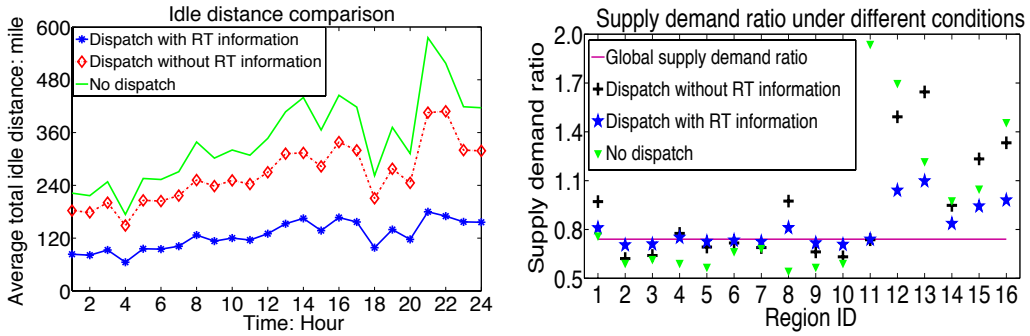
To estimate a mobility pattern matrix $\hat{C}(h_2)$, we define a matrix $T(h_2)$, where $T(h_2)_{ij}$ is the total number of passenger trajectories that starting at region i and ending at region j during time slot h_2. We also apply bootstrap process to get $\hat{T}(h_2)$, and

$$\hat{C}(h_2)_{ij} = \hat{T}(h_2)_{ij}/(\sum_j \hat{T}(h_2)_{ij}).$$

Table 3 shows one row of $\hat{C}(h_2)$ for 5:00-6:00 pm during weekdays, the transition probability for

different regions. The average cross validation error for estimated mobility matrix $\hat{C}(h_2)$ of time slot h_2, $h_2 = 1, \ldots, 24$ during weekdays is 34.8%, which is a reasonable error for estimating total idle distance in the RHC framework when real-time GPS and occupancy status data is available. With only estimated mobility pattern matrix $\hat{C}(h_2)$, the total idle distance is reduced by 17.6% compared with the original record without a dispatch method, as shown in Figure 3.6(a). We also tested the case when the dispatch algorithm is provided with the true mobility pattern matrix C^k, which is impossible in practice, and the dispatch solution reduces the total idle distance by 68% compared with the original record. When we only have estimated mobility pattern matrices instead of the true value to determine ending locations and potential total idle distance for solving problem (3.8) or (3.13), updating real-time sensing data compensates the mobility pattern error and improves the performance of the dispatch framework.

Real-time GPS and occupancy data provides latest position information of all vacant and occupied taxis. When dispatching available taxis with true initial positions, the total idle distance is reduced by 52% compared with the result without dispatch methods, as shown in Figure 3.6(a), which is compatible with the performance when both true mobility pattern matrix C^k and real-time sensing



(a) Comparison of average idle distance. Idle distance is reduced by 52% given real-time information, compared with historical data without dispatch solutions.

(b) Comparison of supply-demand ratio of the whole city and each region. With real-time GPS and occupancy data, the supply demand ratio of each region is closest to the global level. The supply demand ratio mismatch error is reduced by 45% with real-time information, compared with historical data without dispatch solutions.

Figure 6: Comparisons of average idle distance and supply-demand ratio at each region under three conditions: historical record without dispatch, dispatch without real-time data, and dispatch with real-time GPS and occupancy information.

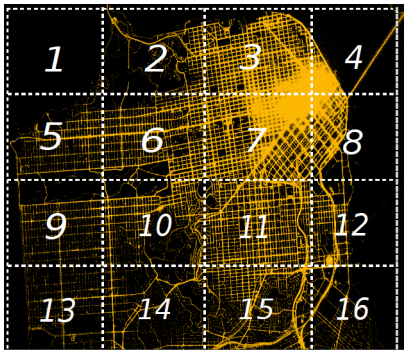


Figure 7: Heat map of passenger picking-up events in San Francisco (SF) with a region partition method. Region 3 covers several busy areas, include Financial District, Chinatown, Fisherman Wharf. Region 7 is mainly Mission District, Mission Bay, the downtown area of SF.

data are available. This is because the optimization problem (3.8) returns a solution with smaller idle distance cost given the true initial position information P^0, instead of estimated initial position provided only by mobility pattern of the previous time slot in the RHC framework. Figure 3.6(a) also shows that even applying dispatch solution calculated without real-time information is better than non dispatched result.

Based on the partition of Figure 7, Figure 3.6(b) shows that the supply demand ratio at each region of the dispatch solution with real-time information is closest to the supply demand ratio of the whole city, and the error

$$\left\| \frac{1}{N} \mathbf{1}_N^T Y^k - \frac{1}{R^k} r^k \right\|_1$$

is reduced by 45% compared with no dispatch results. Even the supply demand ratio error of dispatching without real-time information is better than no dispatch solutions. We still allocate vacant taxis to reach a nearly balanced supply demand ratio regardless of their initial positions, but idle distance is increased without real-time data, as shown in Figure 3.6(a). Based on the costs of two objectives shown in Figures 3.6(a) and 3.6(b), the total cost is higher without real-time information, mainly results from a higher idle distance.

3.6.3. *Robust taxi dispatch*

One disruptive event of the San Francisco data set is Giant baseball at AT&T park, and we choose the historical record on May 31, 2008 as an example to evaluate the robust optimization formulation (3.12). Customer request number for areas near AT&T park is affected, especially Region 7 around the ending time of the game, which increases about 40% than average value.

Figure 8 shows that with dispatch solution of the robust optimization formulation (3.12), the supply demand mismatch error $\left\| \frac{1}{N} \mathbf{1}_N^T Y^k - \frac{1}{R^k} r^k \right\|_1$ is reduced by 25% compared with the solution of (3.8) and by 46% compared with historical data without dispatch. The performance of robust dispatch solutions does not vary significantly and depends on what type of predicted uncertain demand is available when selecting the formulation of robust dispatch method. Even under solutions of (3.8), the total supply demand ratio error is reduced 28% compared historical data without dispatch. In general, we consider the factor of disruptive events in a robust RHC iteration, thus the system level supply distribution responses to the demand better under disturbance.

3.6.4. *Design parameters for Algorithm 1*

Parameters like the length of time slots, the region division function, the objective weight parameter and the prediction horizon T of Algorithm 1 affect the results of dispatching cost in practice. Opti-

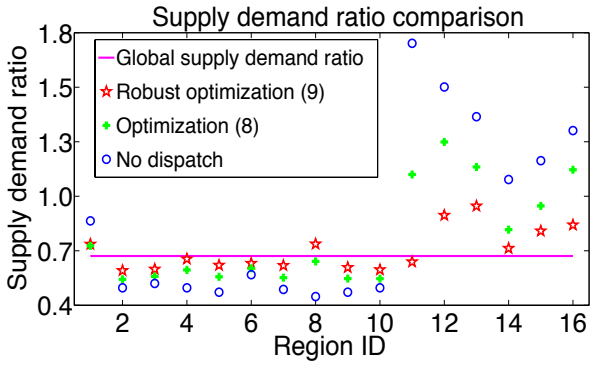


Figure 8: Comparison of supply demand ratio at each region under disruptive events, for solutions of robust optimization problems (3.12), problem (3.8) in the RHC framework, and historical data without dispatch. With the roust dispatch solutions of (3.12), the supply demand ratio mismatch error is reduced by 46%.

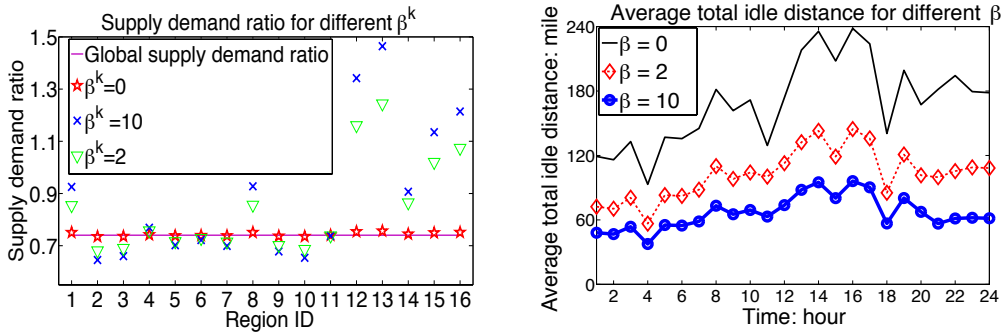
β^k	0	2	10	without dispatch
s/d error	0.645	1.998	2.049	2.664
idle distance	3.056	1.718	1.096	4. 519
total cost	0.645	5.434	13.009	47.854

Table 4: Average cost comparison for different values of β^k.

mal values of parameters for each individual data set can be different. Given a data set, we change one parameter to a larger/smaller value while keep others the same, and compare results to choose a suboptimal value of the varying parameter. We compare the cost of choosing different parameters for Algorithm 1, and explain how to adjust parameters according to experimental results based on a given historical data set with both GPS and occupancy record.

How the objective weight of (3.8) **–** β^k **affects the cost:**

The cost function includes two parts –the idle geographical distance (mileage) cost and the supply demand ratio mismatch cost. This trade-off between two parts is addressed by β^k, and the weight of idle distance increases with β^k. A larger β^k returns a solution with smaller total idle geographical distance, while a larger error between supply demand ratio, i.e., a larger $\left\| \frac{1}{N} \mathbf{1}_N^T Y^k - \frac{1}{R^k} r^k \right\|_1$ value. The two components of the cost with different β^k by Algorithm 1, and historical data without



(a) Comparison of supply-demand ratio at each region during one time slot. When β^k is smaller, we put less cost weight on idle distance that taxis are allowed to run longer to some region, and taxi supply matches the customer requests better.

(b) Average total idle distance of taxis at different hours. When β^k is larger, the idle distance cost weights more in the total cost, and the dispatch solution causes less total idle distance.

Figure 9: Comparisons of supply-demand ratio at each region and average total idle distance for different β^k values.

Algorithm 1 are shown in Table 4. The supply demand ratio mismatch is shown in the s/d error column.

We calculate the total cost as (s/d error $+\beta^k \times$ idle distance) (Use $\beta^k = 10$ for the without dispatch column). Though with $\beta^k = 0$ we can dispatch vacant taxis to make the supply demand ratio of each region closest to that of the whole city, a larger idle geographical distance cost is introduced compared with $\beta^k = 2$ and $\beta = 10$. Compare the idle distance when $\beta^k = 0$ with the data without dispatch, we get 23% reduction; compare the supply demand ratio error of $\beta^k = 10$ with the data without dispatch, we get 32%.

Average total idle distance during different hours of one day for a larger β^k is smaller, as shown in Figure 3.9(b). The supply demand ratio error at different regions of one time slot is increased with larger β^k, as shown in Figure 3.9(a).

How to set idle distance threshold α^k: Figure 10 compares the error between local supply demand ratio and global supply-demand ratio. Since we directly use geographical distance measured by the difference between longitude and latitude values of two points (GPS locations) on the map, the threshold value α^k is small — 0.1 difference in GPS data corresponds to almost 7 miles distance on the ground. When α^k increase, the error between local supply demand ratio and global supply-demand ratio decreases, since vacant taxis are more flexible to traverse further to meet demand.

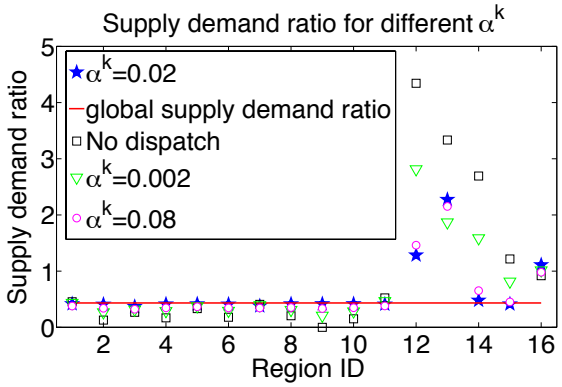


Figure 10: Comparison of supply demand ratios at each region during one time slot for different α^k. When α^k is larger, vacant taxis can traverse longer to dispatched locations and match with customer requests better.

How to choose the number of regions: In general, the dispatch solution of problem (3.8) for a vacant taxi is more accurate by dividing a city into regions of smaller area, since the dispatch is closer to road-segment level. However, we should consider other factors when deciding the number of regions, like the process of predicting requests vectors and mobility patterns based on historical data. A linear model we assume in this work is not a good prediction for future events when the region area is too small, since pick up and drop off events are more irregular in over partitioned regions. While Increasing n, we also increase the computation complexity. Note that the area of each region does not need to be the same as we divide the city in this experiment.

Figure 11 shows that the idle distance will decrease with a larger region division number, but the decreasing rate slows down; while the region number increases to a certain level, the idle distance almost keeps steady.

How to decide the prediction Horizon T: In general, when T is larger, the total idle distance to get a good supply demand ratio in future time slots should be smaller. However, when T is large enough, increasing T can not reduce the total idle distance any more, since the model prediction error compensates the advantage of considering future costs. For $T = 2$ and $T = 4$, Figure 12 shows that the average total idle distance of vacant taxis at most hours of one day decreases as T increases. For $T = 8$ the driving distance is the largest. Theoretical reasons are discussed in

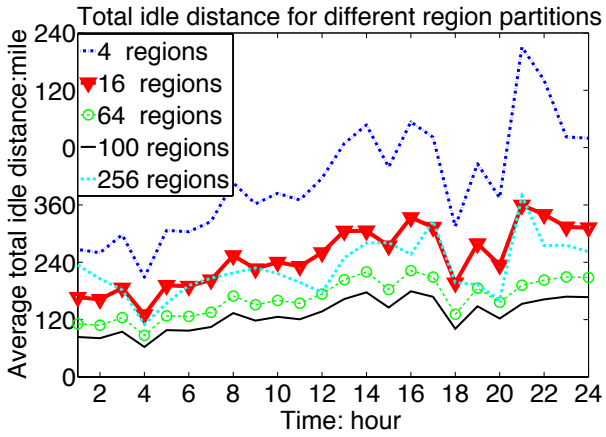


Figure 11: Average total idle distance of all taxis during one day, for different region partitions. Idle distance decreases with a larger region-division number, till the number increases to a certain level.

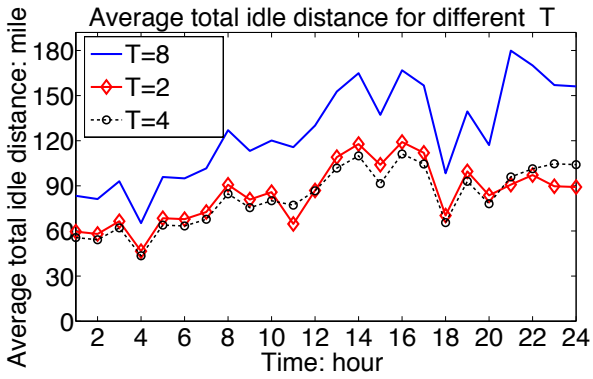
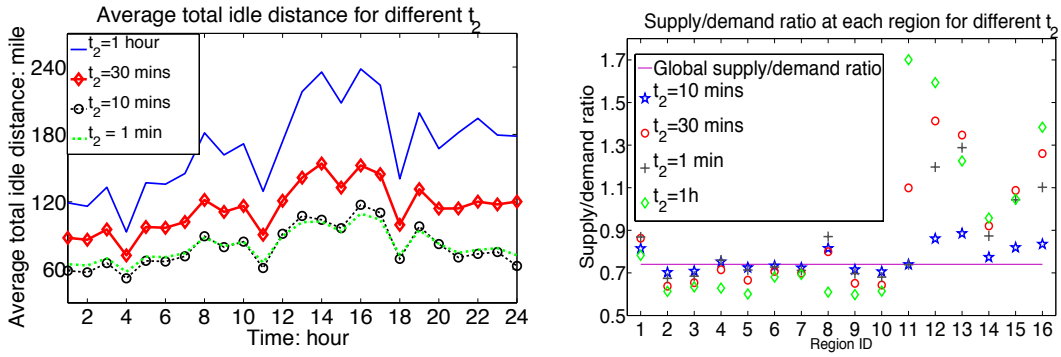


Figure 12: Average total idle distance at different time of one day compared for different prediction horizons.When $T = 4$, idle distance is decreased at most hours compared with $T = 2$. For $T = 8$ the costs are worst.

Section 3.5.

Decide the length of time slot t_2: For simplicity, we choose the time slot t_1 as one hour, to estimate requests. A smaller time slot t_2 for updating GPS information can reduce the total idle geographical distance with real-time taxi positions. However, one iteration of Algorithm 1 is required to finish in less than t_2 time, otherwise the dispatch order will not work for the latest positions of vacant taxis, and the cost will increase. Hence t_2 is constrained by the problem size and computation capability.



(a) Comparison of average total idle distance. With a smaller t_2, the cost is smaller. But when $t_2 = 1$ is too small to complete calculating problem (3.8), the dispatch result is not guaranteed to be better than $t_2 = 10$.

(b) Comparison of the supply-demand ratio at each region. For $t_2 = 30$, $t_2 = 10$ minutes and $t_2 = 1$ hour, results are similar. For $t_2 = 1$ min, the supply demand ratio is even worse at some regions, since the time slot is too short to complete one iteration

Figure 13: Comparison of average total idle distance and supply-demand ratio at each region for different t_2 – the length of time slot for updating sensor information.

Figure 3.13(a) shows that smaller t_2 returns a smaller idle distance, but when $t_2 = 1$ Algorithm 1 can not finish one step iteration in one minute, and the idle distance is not reduced. The supply demand ratio at each region does not vary much for $t_2 = 30, t_2 = 10$ minutes and $t_2 = 1$ hour, as shown in Figure 3.13(b). Comparing two parts of costs, we get that t_2 mainly affects the idle driving distance cost in practice.

CHAPTER 4 : Data-Driven Robust Resource Allocation

4.1. Introduction

Cities are known to have large concentration of resources and facilities, and billions of sensors are connected and used for efficient and effective resource management in Smart Cities [70]. They provide knowledge of system models on users' demand and spatial-temporal. Considering the specific taxi dispatch problem where large amounts of taxi occupancy status and location data are collected from networked in-vehicle sensors in real-time, a receding horizon control framework is designed for efficient resource allocation and coordination strategies in the previous chapter. Such approaches face a new challenge: how to deal with uncertainties of predicted customer demand while fulfilling the system's performance requirements, including minimizing total resource balancing cost and maintaining service fairness. Two aspects of problems exist for a robust taxi dispatch framework: (1) how to formulate a robust resource allocation problem that dispatches vacant taxis towards predicted uncertain demand given a taxi-operational records dataset, 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.

To address these problems, we develop a data-driven robust taxi dispatch framework to consider spatial-temporally correlated demand uncertainties. The robust optimization problem is concave in the uncertain demand and convex in the decision variables with decision variables on the denominators. This form is not the standard form (i.e., linear programming (LP) or semi-definite programming (SDP) problems) that has already been covered by previous work [8, 12, 26]. Box type and second order cone (SOC) type of 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 dispatch cost of robust taxi dispatch solutions. We prove equivalent computationally tractable forms of the robust dispatch problem using the minimax theorem and strong duality. Although a robust RHC formulations is designed in Chapter 3, the objective function is not concave of the uncertain parameters and can only be analytically converted to a convex optimization problem for a

special case of uncertain demand model. For a general polynomial or SOC type of demand uncertainty set, the robust dispatch model of Chapter 3 does not work, while approaches developed in this chapter are more general and include moments information about the uncertain demand.

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 non-robust dispatch solutions.

The rest of this chapter 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 4.2. The requirement of modeling uncertain demand sets are described in Section 4.3, followed by the algorithm for constructing uncertain demand sets based on taxi operational records data in Section 4.4. Equivalent computationally tractable forms of the robust taxi dispatch problem given different forms of uncertainty sets are proved in Section 4.5. Evaluation results based on a real data set are shown in Section 4.6.

Remark 3 *The results from this chapter have been captured in [54, 57].*

4.2. Problem Formulation

The goal of taxi dispatch is to direct vacant taxis towards current and future passengers with 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 expected passengers in order to save cost. Involving predicted customer demand of the future 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 passenger demand. A typical monitoring and dispatch infrastructure is shown in Figure 14. The dispatch center periodically collects and stores real-time information such as GPS

location, occupancy status and road conditions; dispatch solutions are sent to each taxi via cellular radio. An RHC framework that cooperates predicted demand model and real-time sensing data is designed in [59], where either a deterministic model or an uncertain demand model is applied to calculate a dispatch solution at each step of sliding the time window and updating the latest sensing information. However, the robust dispatch problem formulated in [59] does not provide any probabilistic guarantee as the model we design in this work. We define the problem of finding a robust dispatch in the rest of this section, which is compatible with the RHC framework of [59].

4.2.1. Problem description

We discretize time and space in problem formulation for computational efficiency. We assume that the entire city is divided into n regions, and discrete time slots are indexed by $k = 1, 2, \ldots, \tau$. Typically, it is difficult to predict a deterministic value of passenger demand of a region during specific time. With prior knowledge and data sets, we assume that the passenger demand model is described by uncertainty vectors belonging to closed and convex sets defined as

$$r^k \in \Delta_k \subset \mathbb{R}^n_+, \quad k = 1, \ldots, \tau,$$

where r^k_j is the number of total requests within region j during time k, and τ is the model predicting time horizon. Here we relax the integer constraint of $r^k_j \in \mathbb{N}$ to positive, since constructing an

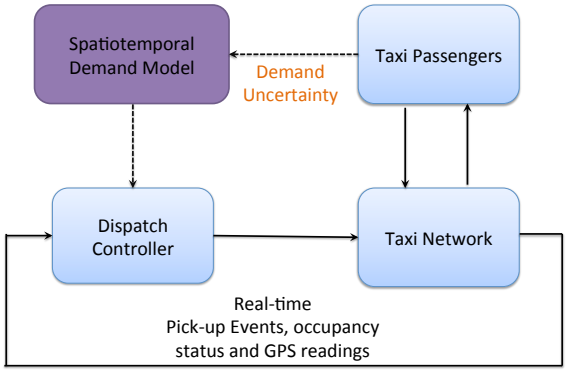


Figure 14: A prototype of the taxi dispatch system

Parameters of (4.11)	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
$C^k \in [0,1]^{n\times n}$	probability matrix for 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 (4.11)	
$X_{ij}^k \in \mathbb{R}_+$	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

Table 5: Parameters and variables of taxi dispatch problem (4.11).

uncertainty set for a continuous vector is more convenient and this relaxation provides an accurate enough demand model. The total number of requests at region j may have similar patterns as its neighbors, for instance, during busy hours, several regions locate in downtown area may all have peak demand. This type of spatial correlations of demand across each region during the same time slot k is reflected by the correlation of each element of r^k. Meanwhile, demand can also be temporal correlated, that demand during several consecutive time slots r^k, $k = 1, \ldots, \tau$ may show similar characteristics like busy hours. Hence, it is possible to describe both spatial and temporal correlations by one set Δ for uncertain demand vectors r^k, $k = 1, \ldots, \tau$. We define the concatenation of sequences $(r^1 \in \mathbb{R}^n, \ldots, r^\tau \in \mathbb{R}^n)$ as

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

and each closed, convex set Δ_k is a projection of Δ

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

Parameters of Alg. 2	Description
$r_c \in \Delta$	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
$\mathcal{U}_\epsilon$	the uncertainty set that provides $1-\epsilon$ probabilistic guarantee level for problem (4.11)
α_h	significance level of a hypothesis testing

Table 6: Parameters of Algorithm 2.

The closed and convex form of Δ depends on the method and theory applied to construct the uncertainty set, which we will describe in Section 4.3.

Considered as one type of resource allocation problem, the basic idea of a robust dispatch model that balances taxis' supply in a network flow model is described in Figure 15. The dispatch framework decides the amount of vacant taxis that should traverse between each node pair according to the demand at each node according to control requirements and practical constraints. 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 non-negative decision variable matrix $X^k \in \mathbb{R}_+^{n\times n}$, $X_{ij}^k \geq 0$, where X_{ij}^k is the number of taxis (amount of resource) dispatched from region i to region j. We relax the integer constraint of $X_{ij}^k \in \mathbb{N}$ to a non-negative constraint, since mixed integer programming is not computational efficient for a large-scale robust optimization problem. In this work we consider the following robust resource allocation problem

$$\begin{aligned} &\min_{X^1}\ \max_{r^1\in\Delta_1}\ \min_{X^2}\ \max_{r^2\in\Delta_2}\ \ldots \min_{X^\tau}\ \max_{r^\tau\in\Delta_\tau} \\ &J = \sum_{k=1}^{\tau}(J_D(X^k) + \beta J_E(X^k, r^k)) \\ &\text{s.t.}\quad X^k \in \mathcal{D}_c, \end{aligned} \tag{4.1}$$

where J_D is a convex cost function for allocating or re-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 $\mathcal{D}_c$

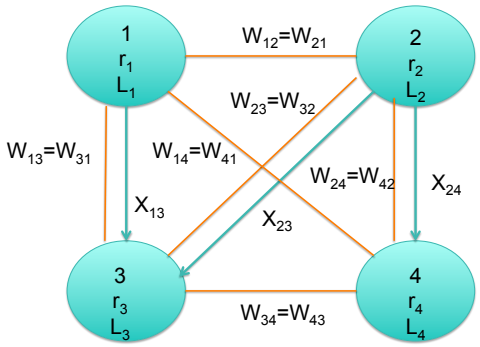


Figure 15: A network flow model of the robust taxi dispatch problem. A circle represents a region with region ID $1, 2, 3, 4$. We omit the superscript of time k since every parameter is for one time slot only. Uncertain demand is denoted by r_i, L_i is the original number of vacant taxis before dispatch at region i, and X_{ij} is a dispatch solution that sending the number of vacant taxis from region i to region j with the distance W_{ij}.

is a convex domain of the decision variables that describes the constraints of the resource allocating strategies. We define specific formulations of the objective and constraint functions for a robust taxi dispatch problem in the rest of this section.

4.2.2. Robust taxi dispatch problem formulation

Estimated cross-region idle-driving distance: When traversing from region i to region 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. 15 as $W \in \mathbb{R}^{n \times n}$, where W_{ij} is the distance between region i and region j. The across-region idle driving cost according to X^k is

$$J_D(X^k) = \sum_i \sum_j X_{ij}^k W_{ij}. \tag{4.2}$$

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

[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.

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

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

which is equivalent to

$$X_{ij}^k \geqslant 0, \quad X_{ij}^k W_{ij} \leq m X_{ij}^k, \quad \forall i, j \in \{1, \ldots, n\}. \tag{4.3}$$

To explain this, assume the constraint (4.3) holds. If $W_{ij} > m$ and $X_{ij}^k > 0$, we have

$$X_{ij}^k W_{ij} > m X_{ij}^k,$$

which contradicts to (4.3). The threshold m 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 minutes, then the value of m should be the distance one taxi can drive during 10 minutes with the current average speed on road (m can also be dependent on k, denoted as m_k if a different average speed during each time slot k can be monitored or predicted).

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 work for computational efficiency [8]. 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 $L^k \in \mathbb{R}_+^n$, 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., $N \geqslant n$, 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

$$\mathbf{1}_n^T X_{\cdot i}^k - X_{i\cdot}^k \mathbf{1}_n + L_i^k \geqslant 1, \quad i = 1, \ldots, n, \tag{4.4}$$

where $X_{\cdot i}^k$ is the i-th column of X^k and $X_{i\cdot}^k$ is the i-th row of X^k. Dispatch is an action of re-

allocating resources among regions and does not change the total number of vacant taxis N before the taxis pick up new passengers during time k

$$\sum_i (\mathbf{1}_n^T X_{\cdot i}^k - X_{i\cdot}^k \mathbf{1}_n + L_i^k) = \sum_i L_i^k = N. \tag{4.5}$$

One service metric is fairness over all regions, 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 in [92]. Meanwhile, a balanced demand-supply ratio means that regions with less demand will also get less resources, and idle driving distance will also be reduced in regions with more supply than demand if we pre-allocate possible redundant supply to those regions in need. We define the objective of minimizing demand-supply ratio mismatch between each region and the whole city as minimizing the following function

$$J_E(X^k, r^k) = \sum_i \frac{r_i^k}{(\mathbf{1}_n^T X_{\cdot i}^k - X_{i\cdot}^k \mathbf{1}_n + L_i^k)^\alpha}, \quad \alpha \to 0. \tag{4.6}$$

This is because by minimizing (4.6) under the constraints (4.4) and (4.5), we get the same optimal solution of minimizing the following demand-supply ratio mismatch function under constraints (4.4) and (4.5).

$$\sum_{k=1}^{\tau} \sum_{i=1}^{n} \left| \frac{r_i^k}{\mathbf{1}_n^T X_{\cdot i}^k - X_{i\cdot}^k \mathbf{1}_n + L_i^k} - \frac{\mathbf{1}_n^T r^k}{N} \right|. \tag{4.7}$$

It is worth noting that the function $J_E(X^k, r^k)$ defined as (4.6) is affine in r^k for any X^k, and convex in X^k for any r^k, while the mismatch function (5.18) is not concave in r^k for any X^k.

To explain how (4.6) approximates (5.18) under constraints (4.4) and (4.5), consider the following problem

$$\underset{b>0, \sum_i b_i = c}{\text{minimize}} \sum_i \frac{a_i}{b_i^\alpha}, \quad c \text{ is a constant.} \tag{4.8}$$

Substitute $b_n = c - b_1 \cdots - b_{n-1}$ into (4.8), and take partial derivatives of $\sum_i \frac{a_i}{b_i^\alpha}$ over $b_i, i = 1, \ldots, n-1$. When the minimum of (4.6) is achieved, each partial derivative should be 0, namely

$$-\alpha \frac{a_i}{b_i^{\alpha+1}} - \alpha(-1) \frac{a_n}{(c - b_1 \cdots - b_{n-1})^{\alpha+1}} = 0,$$

which is equivalent to

$$\frac{a_1}{b_1^{\alpha+1}} = \cdots = \frac{a_{n-1}}{b_{n-1}^{\alpha+1}} = \frac{a_n}{b_n^{\alpha+1}}.$$

Hence, when $\alpha \to 0$, $\alpha + 1 \to 1$, the optimal solution of minimizing J_E over X^k satisfies

$$\frac{r_j}{\mathbf{1}_n^T X_{\cdot j}^k - X_{j\cdot}^k \mathbf{1}_n + L_j^k} = \frac{\mathbf{1}_n^T r^k}{N}.$$

Therefore, with function (4.6), 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 C_{ij}^k as the probability that a taxi traverses from region i to region j and turns vacant again (after one or several drop off events) around the beginning of time $k+1$, provided it is vacant at the beginning of time k. Examples of getting C_{ij}^k based on data include but not limited to methods of describing trip patterns of taxis [59] and autonomous mobility on demand systems [92]. Then the number of vacant taxis within region j by the end of time k is $\left(\mathbf{1}_n^T X^k - (X^k \mathbf{1}_n)^T + (L^k)^T\right) C_{\cdot j}^k$, where $C_{\cdot j}^k$ is the j-th column of C^k, and

$$(L^{k+1})^T = (\mathbf{1}_n^T X^k - (X^k \mathbf{1}_n)^T + (L^k)^T) C^k. \tag{4.9}$$

Weighted-sum objective function: Since there exists a trade-off between two objectives, we define a weight parameter β of two objectives $J_D(X^k)$ in (4.2) and $J_E(X^k, r^k)$ in (4.6). Without consid-

ering model uncertainties corresponding to r^k, a convex optimization form of taxi dispatch problem is

$$\begin{aligned} \min_{X^k, L^k} \quad & J = \sum_{k=1}^{\tau} (J_D(X^k) + \beta J_E(X^k, r^k)) \\ \text{s.t} \quad & \text{(4.3), (4.4), (4.9).} \end{aligned} \tag{4.10}$$

Robust taxi dispatch problem formulation: We aim to find out a dispatch solution robust to an uncertain demand model in this work. For time $k = 1, \ldots, \tau$, uncertain demand r^k only affects the dispatch solutions of $k+1, \ldots, \tau$, similar to the multi-stage robust optimization problem in [13]. Hence, with a list of parameters and variables shown in Table 5, considering effects of current decisions to estimated future costs, a multi-stage robust taxi dispatch problem is defined as following

$$\begin{aligned} & \min_{X^1} \max_{r^1 \in \Delta_1} \min_{X^2, L^2} \max_{r^2 \in \Delta_2} \ldots \min_{X^\tau, L^\tau} \max_{r^\tau \in \Delta_\tau} \\ & J = \sum_{k=1}^{\tau} (J_D(X^k) + \beta J_E(X^k, r^k)) \\ & \quad = \sum_{k=1}^{\tau} \sum_{i} \left(\sum_{j} X_{ij}^k W_{ij} + \frac{\beta r_i^k}{(\mathbf{1}_n^T X_{\cdot i}^k - X_{i\cdot}^k \mathbf{1}_n + L_i^k)^\alpha} \right) \\ & \text{s.t.} \quad (L^{k+1})^T = (\mathbf{1}_n^T X^k - (X^k \mathbf{1}_n)^T + (L^k)^T) C^k, \\ & \quad \mathbf{1}_n^T X^k - (X^k \mathbf{1}_n)^T + (L^k)^T \geqslant 1, \\ & \quad X_{ij}^k W_{ij} \leq m X_{ij}^k, \\ & \quad X_{ij}^k \geq 0, \quad i, j \in \{1, 2, \ldots, n\}. \end{aligned} \tag{4.11}$$

After getting an optimal solution X^{1*} of (4.11), we adjust the solution by rounding methods to get an integer number of taxis to be dispatched towards corresponding regions. It does not affect the optimality of the result much in practice, since the objective function is related to the demand-supply ratio of each region. A feasible integer solution of (4.11) always exists, since $X_{ij}^k = 0, \ \forall i, j, k$ is feasible.

4.3. Constructing Uncertainty 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 future demand uncertainties benefits for minimizing worst-case demand-supply ratio mismatch error and idle distance described as shown in [59, 55]. However, the uncertainty set constructed by only using a standard deviation range [59, 55] cannot tell how possible the true real-world cost is smaller than the optimal cost. Hence, with a large amount of taxi operational records data, it is essential to construct a model that captures the spatial-temporal demand uncertainties and provides a probabilistic guarantee about the true possible values of costs by solving robust dispatch problem (4.11).

4.3.1. Samples of concatenated demand vector

Informally, we consider the concatenated demand vector r_c as a random variable. It is worth noting that we do not have additional assumptions about either the form of Δ besides closeness and convexity, or the form of marginal distribution of each element of vector r_c, or the true distribution of $\mathbb{P}^*(r_c)$.

Methods of constructing uncertainty sets in robust optimization literature is typically designed for i.i.d. sampled random vectors that utilize information from a dataset of samples to provide theoretic guarantee for the performance of robust optimization problems [12], [11], [20]. We transform the knowledge of previous work to construct an uncertainty set Δ for the random vector r_c that contains spatial-temporal relations of the demand model. We assume that one day is discretized as K time slots in total, and the demand of each region during one time slot is described as $r^k, k = 1, \ldots, K$. Then every τ discretized time slots of r^k, $k = t, \ldots, t + \tau$ are concatenated to a vector $r_c(t)$ to represent the possible temporal correlations among consecutive time slots. We define one sample of vector $r_c(t)$ of date d_l as $\tilde{r}_c(d_l, t)$, a vector calculated via aggregating total number of pick up events of all taxis at each region for time slots $t, t + 1, \ldots, t + \tau$. For instance, for consecutive time slots $(1, \ldots, \tau)$, $(2, \ldots, \tau + 1)$, ..., the sampled vectors on date d for time index $t = 1, 2, \ldots$ are

denoted as

$$\tilde{r}_c(d,1) = \begin{bmatrix} \tilde{r}^1(d) \\ \tilde{r}^2(d) \\ \vdots \\ \tilde{r}^\tau(d) \end{bmatrix}, \quad \tilde{r}_c(d,2) = \begin{bmatrix} \tilde{r}^2(d) \\ \tilde{r}^3(d) \\ \vdots \\ \tilde{r}^{\tau+1}(d) \end{bmatrix}, \ldots.$$

We consider demand vectors of different dates for the same time slot as independent samples, i.e., demand $\tilde{r}_c(d_1,t), \tilde{r}_c(d_2,t), \ldots, \tilde{r}_c(d_N,t)$ sampled from N days for time index t are independent with each other for every time index t. For convenience, we omit the time index t of $r_c(t)$ in later discussions when there is no confusion.

There are two advantages to building uncertain sets for the concatenated demand model r_c. The first one is that theories and results proposed for i.i.d. sampled dataset is applicable to design uncertainty sets based on a spatial-temporal dataset. The second one is computational efficiency, that we are able to construct an uncertain set with spatial-temporal properties for all regions during several consecutive predicting time slots by calculating a hypothesis testing one time. It is worth noting that the objective function of problem (4.11), a function concave of $r_k, k = 1, \ldots, \tau$ is still concave of the uncertain parameter r_c with the uncertainty sets constructed in this section. This property guarantees that uncertainty sets constructed in this work can be directly applied for the robust optimization problem (4.11) with $r^k, k = 1, \ldots, \tau$ as parameters.

4.3.2. An 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. When the uncertainty parameter is included in the objective function $J(r_c, x)$ of problem (4.11), the probabilistic guarantee for the event that the true dispatch cost being smaller than the optimal

dispatch cost is described by the following chance constrained problem

$$\begin{aligned} \min_{x} \quad & M \\ \text{s.t} \quad & P_{r_c \sim \mathbb{P}^*(r_c)}(f(r_c, x) = J(r_c, x) - M \leqslant 0) \geqslant 1 - \epsilon. \end{aligned} \tag{4.12}$$

Here $x \in \mathbb{R}^n$ is the optimization variable, and $r_c \in \mathbb{R}^{\tau n}$ is an uncertain parameter. 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 of the following form of chance constrained problem

$$\begin{aligned} \min_{x} \quad & J(x) \\ \text{subject to} \quad & P_{r_c \sim \mathbb{P}^*(r_c)}(f(r_c, x) \leqslant 0) \geqslant 1 - \epsilon. \end{aligned} \tag{4.13}$$

When it is difficult to explicitly estimate $\mathbb{P}^*(r_c)$, given constraints $f(r_c, x)$ that concave in r_c for any x, we solve the following robust optimization problem such that optimal solutions of (4.14) satisfy the probabilistic guarantee of constraints for problem (4.13)

$$\min_{x} \max_{r_c \sim \Delta} \quad J(x), \quad \text{subject to} \quad f(r_c, x) \leqslant 0, \tag{4.14}$$

Then r_c of problem (4.14) can be any vector in the uncertainty set Δ instead of a random vector in problem (4.13), and we require that by solving an optimization problem with this constrained uncertain set performance of optimal solutions is guaranteed for $r_c \sim \mathbb{P}^*$. Another requirement is that the robust optimization problem is computationally tractable problem with this uncertainty set. To emphasize the probability of holding the constraint of (4.13) with the uncertainty set Δ of the robust dispatch problem, we denote the uncertainty set as $\mathcal{U}_\epsilon$ for the the process of constructing a computationally tractable uncertainty set. Hence, for a general form of constraint function $f(r_c, x)$ appeared in robust taxi dispatch problem, the uncertainty set construction problem is defined as the following:

Problem 1 *Construct an uncertainty set $\mathcal{U}_\epsilon$, given ϵ and a data set of random vectors r_c, such that*

(P1). The robust constraint (4.14) *is computationally tractable.*

(P2). The set $\mathcal{U}_\epsilon$ implies a probabilistic guarantee for the true distribution $\mathbb{P}^*(r_c)$ of a random vector r_c at level ϵ, that is, for any optimal solution $x^* \in \mathbb{R}^k$ and for any function $f(r_c, x)$ concave in r_c, we have the implication:

$$\begin{aligned} &\textit{If } f(r_c, x^*) \leq 0, \quad \textit{for } \forall r_c \in \mathcal{U}_\epsilon, \\ &\textit{then } \mathbb{P}^*_{r_c \sim \mathbb{P}^*(r_c)}(f(r_c, x^*) \leqslant 0) \geq 1 - \epsilon. \end{aligned} \tag{4.15}$$

The given probabilistic guarantee level ϵ is related to the degree of conservativeness of the robust optimization problem. The trade-off between the average cost of robust optimal solutions and the probabilistic level is shown by evaluations in Section 3.6. It is worth noting that a confidence region $\mathcal{U}_{c,\epsilon}$ of the random vector that satisfies $\mathbb{P}(r_c \in \mathcal{U}_{c,\epsilon}) \geqslant 1 - \epsilon$ does not need to be the same with the uncertainty set $\mathcal{U}_\epsilon$ satisfies (4.15) in general [7]. Instead of purely building a confidence region $\mathcal{U}_{c,\epsilon}$, we focus on the performance of the robust solutions based on the data-driven uncertainty sets.

The probabilistic guarantee considered in robust optimization literature is stronger than what we require in this work, that the above implication (4.15) should be satisfied for any feasible solution x of the robust optimization problem [12, 11]. In practice, we will apply the optimal solution of the robust dispatch problems as suggestions for taxi drivers, hence only the optimal solution will affect the performance of the dispatch framework, and we require implication and empirical test of (4.15) for optimal solutions only in this work.

4.3.3. Uncertainty Modeling

In this section, we briefly review the theories related to constructing uncertainty demand models based on a spatial-temporal dataset considered in this work. Since we do not assume that the marginal distribution for every element of vector r_c is independent with each other, we select two approaches without any assumptions about the true distribution $\mathbb{P}^*(r_c)$ in the literature [12, 27, 79]. The basic idea is to find a threshold for a hypothesis testing that is acceptable with respect to the

given dataset and a required probabilistic guarantee level, and then construct an uncertainty set based on the hypothesis testing.

Uncertainty demand sets built from marginal samples

One intuitive description about a random vector is to define a range for each element.

For instance, David and Nagaraja [27] considered the following multivariate hypothesis with given thresholds $\bar{q}_{i,0}, \underline{q}_{i,0} \in \mathbb{R}, \quad i = 1, 2, \ldots, \tau n$

$$\begin{aligned} H_{0,i} : & inf\{t : \mathbb{P}(r_{c,i} \leqslant t) \geqslant 1 - \frac{\epsilon}{\tau n}\} \geqslant \bar{q}_{i,0} \\ & inf\{t : \mathbb{P}(-r_{c,i} \leqslant t) \geqslant 1 - \frac{\epsilon}{\tau n}\} \geqslant -\underline{q}_{i,0}. \end{aligned} \tag{4.16}$$

This hypothesis is related to the bound of the $\frac{\epsilon}{\tau n}$ probability value on the random vector, and we divide ϵ by τn because r_c is a multivariate random vector that we need the hypothesis testing for each component $r_{c,i}$ holds simultaneously to provides the probabilistic guarantee described as (4.15).

Assume that we have N random samples for each component $r_{c,i}$ of r_c, ordered in increasing value as $r_{c,i}^{(1)}, r_{c,i}^{(2)}, \ldots, r_{c,i}^{(N)}$ no matter the original sample order. Then this order is also the order of the estimated value $\hat{r}_{c,i}$, i.e., $\hat{r}_{c,i}^{(1)} = r_{c,i}^{(1)}, \ldots, \hat{r}_{c,i}^{(N)}$. We define the index s by

$$s = min \left\{ k \in \mathbb{N} : \sum_{j=k}^{N} \begin{pmatrix} N \\ j \end{pmatrix} \left(\frac{\epsilon}{\tau n}\right)^{N-j} \left(1 - \frac{\epsilon}{\tau n}\right)^{j} \leqslant \frac{\alpha_h}{2\tau n} \right\}, \tag{4.17}$$

and let $s = N + 1$ if the corresponding set is empty. The test H_0 is rejected if

To construct an uncertainty set, we need an accepted hypothesis test. Hence, we set $\bar{q}_{i,0} = \hat{r}_{c,i}^{(s)}$ and $\underline{q}_{i,0} = \hat{r}_{c,i}^{(N-s+1)}$ with $\hat{r}_{c,i}^{(s)}$ and $\hat{r}_{c,i}^{(N-s+1)}$ from the sampled dataset, then $H_{0,i}$ is always accepted. The following uncertainty set is then applied in this work based on the range hypothesis testing (4.16).

Proposition 1 ([12], [27]) *If s defined by equation* (4.17) *satisfies that $N - s + 1 < s$, then, with*

probability at least $1-\alpha_h$ *over the sample, the set*

$$\mathcal{U}_{\epsilon}^{M}(r_c)=\left\{r_c\in\mathbb{R}^{\tau n}:\hat{r}_{c,i}^{(N-s+1)}\leqslant r_{c,i}\leqslant\hat{r}_{c,i}^{(s)}\right\} \tag{4.18}$$

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

The hypothesis (4.16) is tested for each component $r_{c,i}$ separately, and the uncertainty demand model also describes the range of $r_{c,i}$, $i=1,2,\ldots,\tau n$ separately provided by Proposition 1. We do not assume that the marginal distributions of $\mathbb{P}^*$ are independent, their correlations are reflected in the box uncertainty set in the sense that changing the value of n and τ result in a different index value s (4.17), and the order statistics $\hat{r}_{c,i}^{(N-s+1)}$ and $\hat{r}_{c,i}^{(s)}$ will be different. However, the model of the box type of uncertainty set formula does not directly describe the spatial-temporal correlations among components of r_c.

Uncertainty set motivated by moment hypothesis testing

Though the box type of uncertainty set reflects the spatial-temporal correlations by varying range values with different dimensions of r_c, it is not easy to tell directly from the uncertainty set (4.18) when the range of one component changes how will others be affected. To construct an uncertainty set that directly shows the spatial-temporal correlations of the demand model, we consider to apply hypothesis testing related to the first and second moments of the random vector. The following null assumptions are about the mean and covariance of the true distribution $\mathbb{P}^*(r_c)$ of random vector r_c [79]

$$H_0:\mathbb{E}^{\mathbb{P}^*}[r_c]=r_0\quad\text{and}\quad\mathbb{E}^{\mathbb{P}^*}[r_c r_c^T]-\mathbb{E}^{\mathbb{P}^*}[r_c]\mathbb{E}^{\mathbb{P}^*}[r_c^T]=\Sigma_0,$$

with test statistics T defined as $\|\hat{r}_c-r_0\|$ and $\|\hat{\Sigma}-\Sigma_0\|$. Given thresholds Γ_1^B and Γ_2^B, 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.,

$$\|\mathbb{E}^{\mathbb{P}}[\tilde{r}_c] - \hat{r}_c\|_2 > \Gamma_1^B \quad \text{or} \quad \|\mathbb{E}^{\mathbb{P}}[\tilde{r}_c\tilde{r}_c^T] - \mathbb{E}^{\mathbb{P}}[\tilde{r}_c]\mathbb{E}^{\mathbb{P}}[\tilde{r}_c^T] - \hat{\Sigma}\|_F > \Gamma_2^B,$$

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 of multiple times of experiments. The remaining problem is then to select the thresholds such that the above hypothesis testing holds given the dataset. In the following Section **??**, the detailed steps of calculating the thresholds Γ_1^B and Γ_2^B at a desired significance value α_h and probabilistic guarantee level ϵ based on the given dataset is described[2].

The uncertainty set derived based on the moment hypothesis testing is defined in the following proposition.

Proposition 2 ([12], [79]) *With probability at least $1-\alpha_h$ with respect to the sampling, the following uncertainty set $\mathcal{U}_\epsilon^{CS}(r_c)$ implies a probabilistic guarantee level of ϵ for $\mathbb{P}^*(r_c)$*

$$\begin{aligned}\mathcal{U}_\epsilon^{CS}(r_c) =& \{r_c \geqslant \mathbf{0}, \hat{r}_c + y + C^T w : \exists y, w \in \mathbb{R}^{n\tau} \ s.t. \\ & \|y\|_2 \leqslant \Gamma_1^B, \|w\|_2 \leqslant \sqrt{\frac{1-\epsilon}{\epsilon}}\},\end{aligned} \tag{4.19}$$

where $C^T C = \hat{\Sigma} + \Gamma_2^B \mathbf{I}$ is a Cholesky decomposition.

By testing the properties of both first and second moments of the dataset, the uncertainty set (4.19) reflects the spatial-temporal correlations of the demand model directly compared with the box type (4.18). 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 (4.19). More properties of each type of uncertainty set and application level problems, such as how to choose the number of samples N for the hypothesis testing with high dimensional r_c will be discussed in evaluations of Section 3.6.

[2] Bootstrapped thresholds and theoretic bounds proposed by work [48] are compared in [12]. The bootstrapped thresholds result in a smaller uncertainty set in general, hence reduces the ambiguity in $\mathbb{P}^*$. In this work, we apply the bootstrapped thresholds Γ_1^B and Γ_2^B based on the dataset.

4.4. Algorithm For Constructing Uncertain Demand Sets

Given a dataset, the algorithm for constructing uncertainty sets includes three main steps—getting a sample set of r_c from the original dataset 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. In this section, we explain each step, summarize the process in Algorithm 2, and discuss factors to consider for choosing parameters of the algorithm. Numerical examples are shown in Section 3.6.

4.4.1. Aggregating demand and partition the sample set

The first step is to transform the original dataset of taxi operational records to a dataset of sampled vector $\tilde{r}_c(d,t)$ of different dates for each index t. For instance, assume we choose the length of each time slot as one hour, and the dataset records all trip information of taxis during each day. According to the start time and GPS coordinate of the pick-up position of each trip, we aggregate the total number of pick up events during one hour at each region to get samples of $r^k, k \in \{1,2,\ldots,\tau\}$ and the concatenated demand vector r_c. It is computationally efficient to process the original data for obtaining a sample set of r_c in general, though the amount of available taxi trips or trajectory information is large – the time complexity is $O(N_{record})$ of the number of total records N_{record}. By only passing through the raw data once, we are able to group each pick up and drop off events to a specific discretized time slot and region.

We assume that the dataset contains independent samples of the random vector r_c, and we do not impose any prior knowledge of the true distribution $\mathbb{P}^*(r_c)$. It is always possible to describe the support of the distribution of the entire dataset, even when all samples contained in the dataset do not follow the same distribution, as explained in Figure 16. When there is prior knowledge or categorical information such that the dataset can be partitioned into several subsets according to some feature space, we get a more accurate uncertainty set according to each sub-dataset to provide the same probabilistic guarantee level compared with the uncertainty set from the entire dataset.

Clustering algorithms with categorical information [44] is applicable for dataset partition when in-

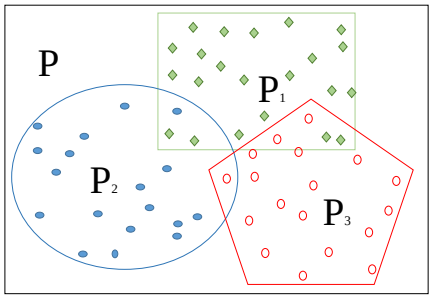


Figure 16: Intuition for partitioning the whole dataset. When the data set includes data from three distributions P_1, P_2, P_3, without prior knowledge, we can build a larger uncertainty set that describes the range of all samples in the dataset. The problem is that the uncertainty set is not accurate enough.

formation besides pick up events is available in the dataset, such as weather or traffic conditions. It is worth noting that if the uncertainty sets are built for a categorical information set $\mathcal{I} = \{\mathcal{I}_1, \mathcal{I}_2, \dots\}$, then for the robust dispatch problems, we require the same set of categories is available in real-time, hence we apply the uncertainty set built for $\mathcal{I}_1$ to find solutions when the current situation is considered as $\mathcal{I}_1$. For instance, when there is additional information like weather or traffic condition for each trip provided by the taxi operational records, these types of information can be used as categorical information for clustering. The dataset applied in the evaluations of Section 3.6 does not have additional categorical information of trips that available for a clustering algorithm such as [44], hence, we partition the dataset as demand during weekdays and demand during weekends. Even with this simple and intuitive partition process, we shrink the area of an uncertainty for the same probabilistic guarantee level. Then during weekdays (weekends) we use uncertainty sets built from weekdays (weekends) data to calculate robust dispatch solutions.

4.4.2. Algorithm

The uncertainty sets designed in this work require an accepted null hypothesis testing. Given original operational records 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 for each subset of sampled demand vectors. With a threshold of the test statistics calculated via the given dataset, we then apply the formula (4.18) for

constructing a box type of uncertainty set, and the formula (4.19) for an SOC type of uncertainty set, respectively. The following Algorithm 2 describes the complete process for constructing uncertain demand sets based on the original dataset.

Algorithm 2: Algorithm for 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 $\mathcal{S}(I_p) \subset \mathcal{S}$, $p = 1, \ldots, P$ as the subset partitioned according to either prior knowledge or categorical information I_p. Denote the partitioned sample subset for each time index t as $\mathcal{S}(t, I_p)$.

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 $\mathcal{S}^j(t, I_p) = \{\tilde{r}_c(d_1, t, I_p), \ldots, \tilde{r}_c(d_N, t, I_p)\}$ 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 (4.18) and (4.19), respectively, for each t and I_p. **Output: Uncertainty sets for problem** (4.11)

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 this step. The estimations of this step are considered as the true mean and covariance for calculating Γ_1^B and Γ_2^B in the following repeated sampling process. For step $2.(2)$, the process for the box type of uncertainty sets is: calculate index s that satisfies (4.17) with the given ϵ, sort each component of sampled vectors $r_c(d_l, t, I_p)$, and get the order statistics $r_{c,i}^{(N-s+1)}(j, t, I_p)$, $r_{c,i}^{(s)}(j, t, I_p)$ of the j-th 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 $r_{c,i}^{(N-s+1)}(j,t,I_p)$ for the i-th component of each t and I_p. 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.

Remark 4 *The process of constructing uncertainty sets only requires that the hypothesis test is accepted for i.i.d. samples of the random vector. We accept the hypothesis test when there is not enough evidence to reject it, which does not mean the claim of H_0 is true. This property is very important for constructing the uncertainty demand set of the robust dispatch problem, since the true distribution function of a demand model can be complex and we only have datasets of taxi operational records instead of ground truth knowledge of the distribution function. Hence, even without enough knowledge of the true, high-dimensional demand model, based on the dataset and an accepted hypothesis test, we are able to construct an uncertainty set with probabilistic guarantee for the robust taxi dispatch problem.*

It is worth noting that the above Algorithm 2 provides a valid estimation of uncertain sets based on hypothesis testing and bootstrapped thresholds for the robust resource allocation problem when the sampled data set is consistent with the real world scenario. For demand missed in the dataset, for instance, some customer might leave the request queue after waiting for a long time and the operational records did not show the event of picking up the customer, we are not able to get the exact rate of missed customers. However, missed requests are only part of the historical requests, and this type of events is also random – for instance, even for the same time length of waiting, some customers were more patient and finally got a taxi. By constructing an uncertainty set to describe the demand model based on occurred records of the original dataset, we involve the effect of random missing events better than only applying a deterministic model from this perspective.

In summary, to construct a spatial-temporal uncertain demand model for the robust taxi dispatch (4.11), 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 dataset 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 apply theories proved for i.i.d. samples of random vectors in the literature [12] [27] [79] to construct a box type and an SOC type of uncertainty sets. The key advantage of the data-driven approach we propose is that we do not rely on prior knowledge of the true distribution of the random demand vector to provide a desired probabilistic guarantee of robust solutions. Furthermore, theories proved for i.i.d. datasets are applicable to construct uncertainty sets that reflect the spatial-temporal correlations of the demand model.

4.5. Computationally Tractable Formulations

We build equivalent computationally tractable formulations of problem (4.11) with different definitions of uncertain sets built in Section 4.3 in this section, and show that the robust taxi dispatch problem in this work can be solved efficiently. Computational tractability of a robust linear programming problem for ellipsoid uncertainties are discussed in [8]. The process is to reformulate constraints of the original problem to equivalent convex constraints that must hold given the uncertainty set. The objective function of problem (4.11) is concave of the uncertain parameters r^k, convex of the decision variables X^k, L^k with the decision variables on the denominators, hence, not standard forms of linear programming (LP) or semi-definite programming (SDP) problems that already covered by previous work [8, 12]. Hence, we prove one equivalent computationally tractable form of problem (4.11) for each uncertainty set constructed in Section 4.3.

Only the J_E components of objective functions in (4.11) 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 (4.18) is a special form of polytope, hence, we first prove an equivalent standard form of convex optimization problem for (4.11) for a polytope uncertainty set as the following.

Theorem 2 *(Next step dispatch) If the uncertainty set of problem* (4.11) *when* $\tau = 1$ *is defined as the following polytope*

$$\Delta := \{r \geq 0, Ar \leq b\},$$

and we omit the superscripts k *for variables and parameters without confusion. Then problem* (4.11) *with* $\tau = 1$ *is equivalent to the following convex optimization problem*

$$\begin{aligned}
\underset{X\geq 0,\lambda\geq 0}{\textit{minimize}} \quad & \sum_i \sum_j X_{ij} W_{ij} + b^T \lambda \\
\textit{subject to} \quad & A^T \lambda - \beta \begin{bmatrix} \frac{1}{(\mathbf{1}_n^T X_{\cdot 1} - X_{1\cdot}\mathbf{1}_n + L_1)^\alpha} \\ \vdots \\ \frac{1}{(\mathbf{1}_n^T X_{\cdot n} - X_{n\cdot}\mathbf{1}_n + L_n)^\alpha} \end{bmatrix} \geq 0, \\
& \mathbf{1}_n^T X - X\mathbf{1}_n + L^T \geqslant 1, \\
& X_{ij} W_{ij} \leq m X_{ij}, \\
& X_{ij} \geq 0, \quad \forall i, j \in \{1, \ldots, n\}.
\end{aligned} \tag{4.20}$$

Proof 2 *See Appendix A.1.1.*

For the multi-stage robust optimization problem (4.11), we prove that the order of minimize and maximum is exchangeable in the following theorem, and equivalent computationally tractable forms are proved based on this theorem.

Lemma 1 *(Exchange the order of minimize and maximum) Assume that the definition of the uncertainty set* Δ *satisfies that the domain of each* r^k *is a compact set, then the multi-stage robust dispatch problem defined as* (4.11) *is equivalent to the following robust dispatch problem*

$$\begin{aligned}
\underset{X^k, L^k}{\textit{min.}} \ \underset{r^k \in \Delta_k}{\textit{max}} \quad & J = \sum_{k=1}^{\tau} (J_D(X^k) + \beta J_E(X^k, r^k)) \\
\textit{s.t.} \quad & \textit{constraints of } (4.11), k = 1, \ldots, \tau.
\end{aligned} \tag{4.21}$$

Here L^1 *is the initial number of vacant taxis within each region before dispatch provided by sensor*

information, not a decision variable, and we omit the time index of $L^k, k = 2, \ldots, \tau$ *in minimization for notation convenience.*

Proof 3 *See Appendix A.1.2.*

For the multi-stage robust optimization problem (4.11), the computationally tractable convex form depends on the definition of uncertainty sets. For a multi-stage robust optimization problem that minimax theorem does not hold, an approximated semidefinite programming form for calculating time dependent control input of linear dynamical systems affected by uncertainty is proposed in [13]. When conditions of Lemma 1 hold, equivalent convex optimization forms of problem (4.11) are derived based on problem (4.21).

The box type uncertainty set (4.18) 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 (4.11) 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 2 *If the uncertain set for* $r^k, k = 1, \ldots, \tau$ *describes each demand vector* r^k *separately as a polytope with the form*

$$\Delta_k := \{r^k \geq 0, A_k r^k \leq b_k\}, \quad k = 1, \ldots, \tau, \tag{4.22}$$

problem (4.11) *is equivalent to the following convex optimization problem*

$$\begin{aligned} \underset{X^k, \lambda^k, L^k \geq 0}{\textit{min.}} \quad & \sum_{k=1}^{\tau} \Big(\sum_i \sum_j X_{ij}^k W_{ij} + b_k^T \lambda^k \Big) \\ \textit{subject to} \quad & A_k^T \lambda^k - \beta \begin{bmatrix} \frac{1}{(\mathbf{1}_n^T X_{\cdot 1}^k - X_{1\cdot}^k \mathbf{1}_n + L_1^k)^\alpha} \\ \vdots \\ \frac{1}{(\mathbf{1}_n^T X_{\cdot n}^k - X_{n\cdot}^k \mathbf{1}_n + L_n^k)^\alpha} \end{bmatrix} \geq 0, \\ & \textit{other constraints of } (4.11), k = 1, \ldots, \tau. \end{aligned} \tag{4.23}$$

Proof 4 *See Appendix A.1.3.*

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 (4.11).

Theorem 3 *When Δ is defined as the following polytope*

$$\Delta := \{(\Delta_1, \ldots, \Delta_\tau) | A_1 r^1 + \cdots + A_\tau r^\tau \leq b, r^k \geq 0\}, \tag{4.24}$$

problem (4.11) *is equivalent to the following convex optimization problem*

$$\begin{aligned} \underset{X^k, L^k, \lambda \geq 0}{\textit{min.}} \quad & \sum_{k=1}^{\tau} (\sum_i \sum_j X_{ij}^k W_{ij}) + b^T \lambda \\ \textit{subject to} \quad & A_k^T \lambda - \beta \begin{bmatrix} \frac{1}{(\mathbf{1}_n^T X_{\cdot 1}^k - X_{1\cdot}^k \mathbf{1}_n + L_1^k)^\alpha} \\ \vdots \\ \frac{1}{(\mathbf{1}_n^T X_{\cdot n}^k - X_{n\cdot}^k \mathbf{1}_n + L_n^k)^\alpha} \end{bmatrix} \geq 0, \\ & \textit{constraints of } (4.11), \quad k = 1, \ldots, \tau. \end{aligned} \tag{4.25}$$

Proof 5 *See Appendix A.1.3.*

With an uncertain demand model defined as (4.19) for concatenated $r^1, \ldots, r^\tau$, the following theorem derive the equivalent computationally tractable form of problem (4.11).

Theorem 4 *When the uncertainty set for $r^1, \ldots, r^\tau$ is defined as the SOC form of* (4.19)*, problem* (4.11) *is equivalent to the following convex optimization problem* (4.26)*.*

$$\begin{aligned} \underset{X^k, L^k, z}{\textit{min.}} \quad & \sum_{k=1}^{\tau} \sum_i \sum_j X_{ij}^k W_{ij} + \beta \left(\hat{r}_c^T z + \Gamma_1^B \|z\|_2 + \sqrt{\frac{1}{\epsilon} - 1} \|Cz\|_2 \right) \\ \textit{subject to} \quad & c_l(X) \leqslant z, \\ & \textit{constraints of } (4.11), \quad k = 1, \ldots, \tau, \end{aligned} \tag{4.26}$$

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

Proof 6 *See Appendix A.1.4.*

It is worth noting that any optimal solution for problem (4.10) has a special form between any pair of regions (i, q).

Proposition 3 *Assume $X^{1*}, \ldots, X^{\tau*}$ is an optimal solution of* (4.10)*, then any X^{k*} satisfies that for any pair of (p, q), at least one value of the two elements X_{qi}^{k*} and X_{iq}^{k*} is* 0*.*

Proof 7 *We prove by contradiction. Assume that one optimal solution has the form X^k such that $X_{qi}^k > 0$ and $X_{iq}^k > 0$. Without loss of generality, we assume that $X_{qi}^k \geq X_{iq}^k$, and let*

$$X_{qi}^{k*} = X_{qi}^k - X_{iq}^k, X_{iq}^{k*} = 0,$$

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

$$\mathbf{1}_n^T X_{\cdot i}^k - X_{i\cdot}^k \mathbf{1}_n + L_i^k = \mathbf{1}_n^T X_{\cdot i}^{k*} - X_{i\cdot}^{k*} \mathbf{1}_n + L_i^k,$$

because

$$\begin{aligned}
&\sum_j X_{ji}^k - \sum_l X_{il}^k = X_{qi}^k - X_{iq}^k + \sum_{j \neq q} X_{ij}^k - \sum_{l \neq q} X_{qi}^k \\
=&X_{qi}^{k*} + 0 + \sum_{j \neq q} X_{ij}^{k*} - \sum_{l \neq q} X_{qi}^{k*} = \sum_j X_{ji}^{k*} - \sum_l X_{il}^{k*}.
\end{aligned}$$

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 (4.11)*.*

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

$$X_{qi}^k > X_{qi}^{k*},\ X_{qi}^k W_{qi} + X_{iq}^k W_{iq} > X_{qi}^{k*} W_{qi} + X_{iq}^{k*} W_{iq}.$$

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 (4.11) is computationally tractable and solved efficiently.

4.6. Data-Driven Evaluations

4.6.1. A Motivation Example

We first conduct simulations based on a San Francisco taxi data set [74]. Information for each individual taxi includes three components: the Unix epoch time, the geometric position (latitude and longitude), and a binary indicator of whether the taxi is vacant or with passengers. We show the motivation to find robust dispatch solutions with model uncertainties, and compare the optimal cost of robust dispatch (4.11) with convex optimization form (4.10) in this section.

Estimate uncertainty sets for demand r^k:

A boxplot of total number of requests (pick up events) during one hour ($5:00-6:00$ pm) in different regions is shown in Figure 17. The mean and standard deviation of the model are calculated

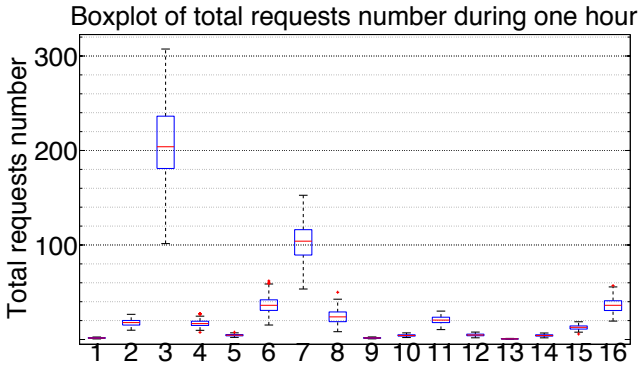


Figure 17: Boxplot of total number of equests at each region during one hour. The red line in the middle shows the median value of all samples, the box shows the distribution of data, with range first quartile and third quartile.

via bootstrap [18]. Figure 17 shows a motivation of this work — a robust dispatch algorithm to balance the number of taxis according to the demand from the perspective of system-level optimal performance.

How vacant taxis are balanced across regions with different α values: Figure 18 shows mismatch between supply and demand defined as (5.18) for different optimal solutions of minimizing J_E defined in (4.6) for $\alpha \in (0, 1]$. With α closer to 0, the optimal value of (5.18) is smaller. We choose $\alpha = 0.1$ for calculating optimal solutions of (4.11) and (4.10) in this section.

Compare robust solutions with non-robust solutions: We compare the cost distribution of 200 Monte-Carlo simulations based on the data set of robust optimization solutions (4.11) and convex optimization solutions (4.10) in Figure 19. The customer demand models applied in the two algorithms are different. For the objective function (4.10), the nominated demand prediction r^k is a deterministic value, for instance — the average or mean of the bootstrap model which is constructed based on the historical data set. For the robust problem formulation (4.11) considered in this work, the uncertainty set is a box defined according to the mean and covariance matrix of the bootstrap model.

Figure 19 shows that the robust dispatch solutions result in 35.5% fewer experiments with a cost

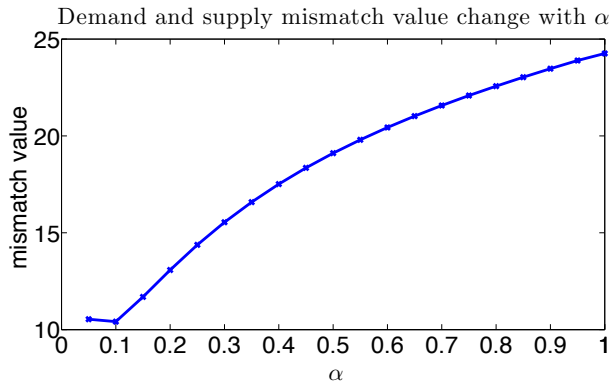


Figure 18: Comparison of demand and supply mismatch values defined as (5.18) with different solutions for minimizing J_E defined in (4.6) with α in range $(0,\ 1]$. The value of function (5.18) under an optimal solution of J_E is smaller with an α closer to 0, which means the dispatch solution tends to be more balanced throughout the entire city.

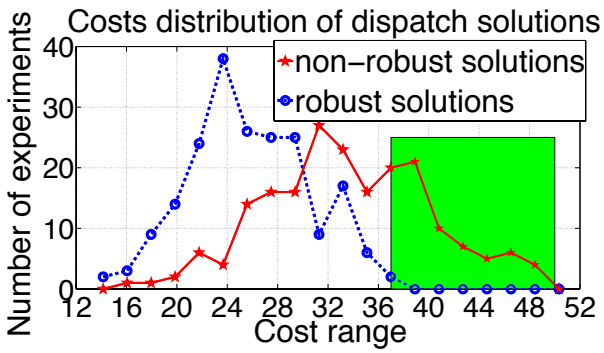


Figure 19: Cost distribution comparison of robust optimization (4.11) solutions in this work and non-robust optimization (4.10) solutions. The lines show the number of experiments with cost falling in intervals $[12,\ 14]$, $(14, 16]$, $\ldots, (48, 50]$ of two methods applying Monte-Carlo experiments based on the historical data set. Robust optimization solutions in this work has a shorter tail than non-robust solutions.

greater than 37, compared with non-robust solutions. It means the cost distribution of the robust optimization (4.11) in this work has a shorter tail than that of the deterministic convex optimization formulation (4.10). With model uncertainty information in decision making, system performance is improved compared with solutions only based on the nominal demand model.

4.6.2. Evaluations based on a 100GB dataset

We then conduct data-driven evaluations based on four years of taxi trip data of New York City [29]. A summary of this data set is shown in Table 7. 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.

One region partition example according to the map of Manhattan of New York City is shown in Figure 20 where we visualize the density of taxi passenger demand with the data we used for our large-scale data-driven evaluation. The lighter the region, the higher the daily demand density. As we can see in the figure, the middle regions typically have higher density than the uptown and downtown regions in Manhattan. We construct uncertainty sets according to Algorithm 2, discuss factors that affect modeling of the uncertainty set, and compare optimal costs of the robust dispatch formulation (4.11) and the non-robust optimization form (4.10) in this section.

Taxi Trip Data		
Collecting Period	Data Size	Record Number
01/01/2010-12/31/2013	$100GB$	700 million
Data Format		
Trip Information	Trip Time	Trip Locations
Start and end points	Date/hour:minute:second	GPS coordinates

Table 7: New York city data used in this evaluation section.

4.6.3. Box type of uncertainty set

For all box type of uncertainty sets shown in this subsection with the model described in Subsection 4.3.3, we set the confidence level of hypothesis testings as $\alpha_h = 10\%$, bootstrap time as $N_b = 1000$, number of randomly sampled data (with replacement) for each time of bootstrap as $N = 10000$.

Partitioned dataset compared with non-partitioned dataset: We show the effects of partitioning the trip record dataset by weekdays and weekends in Figure 4.21(a) and 4.21(b). The whole city is partitioned into 50 regions and the prediction time horizon $\tau = 4$, $\epsilon = 0.3$, and every $r_c \in \mathbb{R}^{200\times 1}$. Figures 4.21(a) and 4.21(b) show the lower and upper bounds of each region during one time slot of (4.18). By applying data of weekdays and weekends separately, the range $[\hat{r}_{c,i}^{(s)},\ \hat{r}_{c,i}^{(N-s+1)}]$ of each component is reduced. To get a measurement of the uncertainty level, we defined the sum of

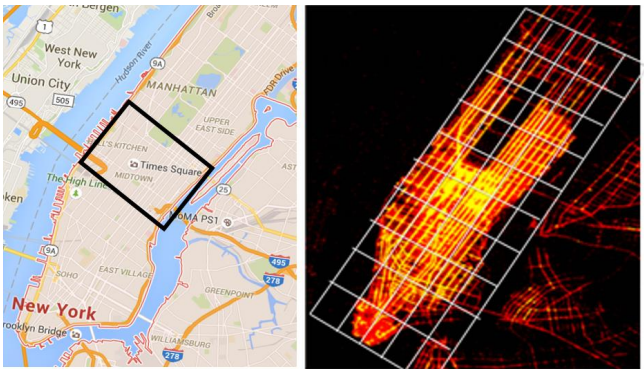
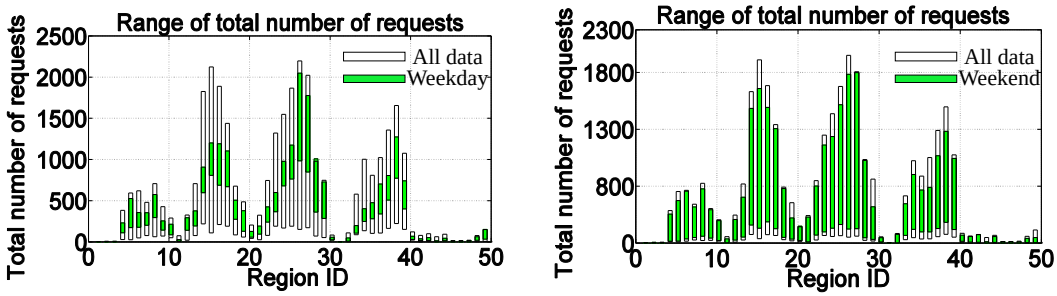


Figure 20: Map of Manhattan area in New York City.



(a) Comparison of box type of uncertainty sets constructed from all data and those constructed only based on trip records of weekdays.

(b) Comparison of box type of uncertainty sets constructed from all data and uncertainty sets constructed only based on trip records of weekends.

Figure 21: Comparison of box type of uncertainty sets constructed from all data and those constructed only based on trip records of weekdays and weekends. When keeping all parameters the same, by applying data of weekdays or weekends only, the range of uncertainty set for each $r_{c,i}$ is smaller than that based on the whole dataset.

range of every component for $\hat{r}_c$ as the following

$$U(\hat{r}_c) = \sum_{i=1}^{\tau n} (\hat{r}_{c,i}^{(s)} - \hat{r}_{c,i}^{(N-s+1)}).$$

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 $U(\hat{r}_c)$ means a smaller area of the uncertainty set, or a more accurate model. We denote $U(\hat{r}_c)$ calculated via records of weekdays and weekends as $U_{wd}(\hat{r}_c)$ and $U_{wn}(\hat{r}_c)$ respectively, compared with $U(\hat{r}_c)$ constructed from the complete dataset, we have

$$\frac{U(\hat{r}_c) - U_{wd}(\hat{r}_c)}{U(\hat{r}_c)} = 52\%, \quad \frac{U(\hat{r}_c) - U_{wn}(\hat{r}_c)}{U(\hat{r}_c)} = 28\%.$$

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.

Choose an appropriate N 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

N	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 8: Value of index s for the box type uncertainty set (4.17). For large τn, N need to be large, or s is too close to N that the range covers values of almost all samples.

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

Table 9: 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 $n = 1000$, $\tau = 3$, $N = 1000$, $\epsilon = 0.3$, $\alpha_h = 0.2$.

Table 11, N need to be large enough for a large τn value, or s is too close to N and the upper and lower bounds $\hat{r}_{c,i}^{(N-s+1)}$, $\hat{r}_{c,i}^{(s)}$ 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 (4.25) is smaller than that of (4.26) with the same size of τn.

4.6.4. 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 non-partitioned data are summarized in Table 9. Similarly as the box type of uncertainty sets, when we separate the dataset 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 $\mathcal{U}_\epsilon^{CS}$ 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 9.

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 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

	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

Table 10: 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.

set (such as $\tau n = 1000, N = 10000$ in Table 11). 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 11. 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 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.

4.6.5. Compare robust solutions with non-robust solutions

For testing the quality of the uncertainty sets applied in the robust dispatch problems, we use the idea of cross-validation from machine learning. The dataset 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 non-robust optimization problems are different. For the non-robust dispatch problem, the demand prediction r^k is a deterministic value. For instance, in this work we use the average or mean of the bootstrapped value of the training dataset.

In the experiments, the idle geographical distance of one taxi between a drop-off event of one passenger and the following pick-up event is approximately as the one norm distance between the 2D geographical coordinates (provided as longitude and latitude values of GPS data in the trip dataset) of the two points. Then the corresponding idle miles on ground is converted from the geographical distance according to the geographical coordinates of New York City.

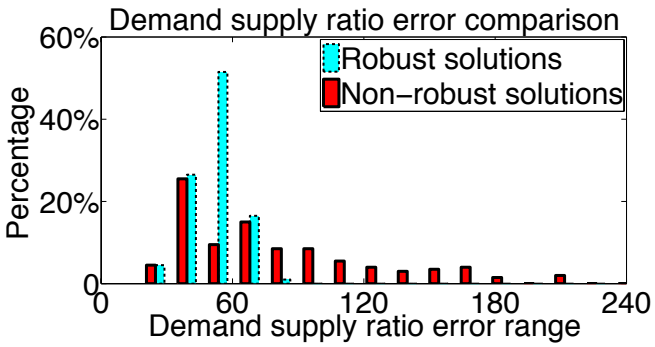


Figure 22: Demand-supply ratio error distribution of the robust optimization solutions with the SOC type of uncertain demand set ($\epsilon = 0.25$, or probabilistic guarantee level 75%) and non-robust optimization solutions. The demand-supply ratio error of robust solutions is smaller than that of the non-robust solutions, that the average demand-supply ratio error is reduced by 31.7%.

In the robust dispatch problem, the part that directly includes the uncertain demand r^k is the penalty function 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 (5.18). We then compare the value of (5.18) of robust dispatch solutions with the SOC type of uncertainty set constructed in this work with the value of (5.18) of non-robust solutions of testing samples. The distribution of values are shown in Figure 22. The average demand-supply ratio error is reduced by 31.7% with robust solutions.

We compare the cost distribution of total idle distance in Figure 23. It shows 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 non-robust solutions has 48% of samples with idle distance greater than 0.8×10^5. The cost of robust dispatch (4.11) 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 means that 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.

Check whether the probabilistic level ϵ is guaranteed: Theoretically, the optimal solution of the

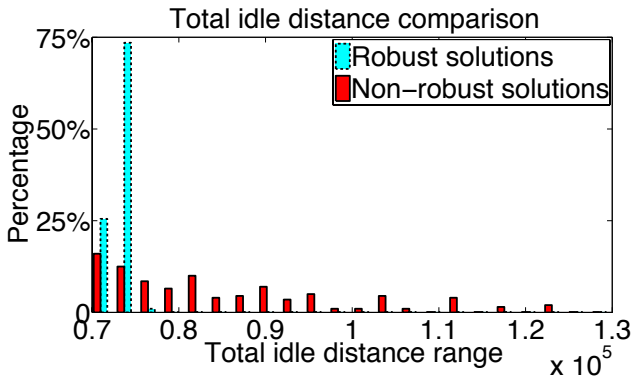
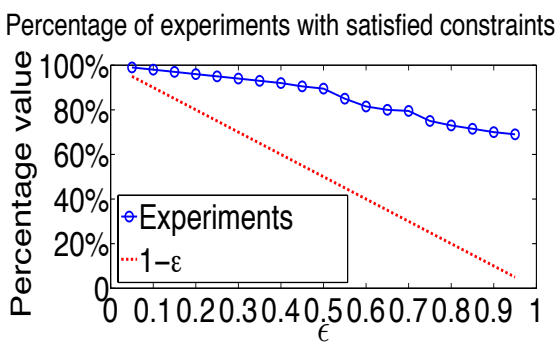
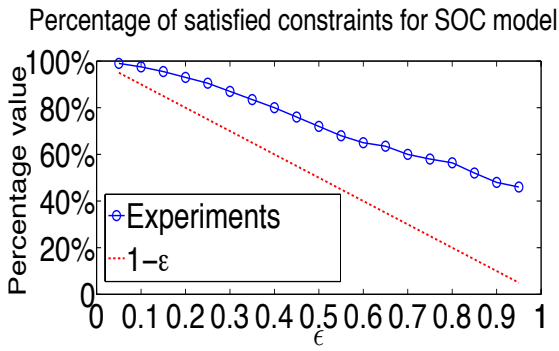


Figure 23: Total idle distance comparison of robust optimization solutions with the SOC type of uncertain demand set ($\epsilon = 0.25$, or probabilistic guarantee level 75%) and non-robust optimization solutions. The average total idle distance is reduced by 10.13%. For all samples used in testing, the robust dispatch solutions result in no idle distance greater than 0.8×10^5, and non-robust 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.

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. Figures 4.24(a) and 4.24(b) show the cross-validation testing result that the probabilistic guarantee level is reached for both

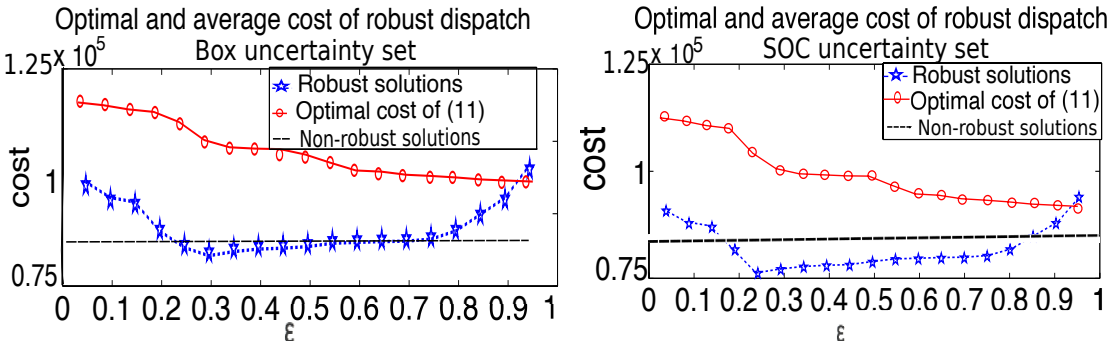


(a) Comparison result with the box type of uncertainty set.



(b) Comparison result with the SOC type of uncertainty set. The true percentage value is closer to the value of $1 - \epsilon$ compared with the solution given a box type uncertainty set.

Figure 24: The percentage of tests that have a smaller true dispatch cost than the optimal cost of the robust dispatch problem with the box and SOC types of uncertainty sets constructed from data. When $1 - \epsilon$ decreases, the percentage value also decreases, but always greater than $1 - \epsilon$.



(a) Comparison result with box type of uncertainty set. When $\epsilon = 0.3$ the average cost is the smallest.

(b) Comparison result with SOC type of uncertainty set. When $\epsilon = 0.25$ the average cost is the smallest.

Figure 25: Comparisons of the optimal cost of the robust dispatch problem with box and SOC types of uncertainty sets and the average cost when applying the robust solutions for the test subset of sampled r_c.

box type and SOC type of uncertainty sets via solving (4.25) and (4.26), 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 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 (4.26) than to solve problem (4.25).

How probabilistic guarantee level affects the average cost: There exists a trade-off 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 whether the average cost 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 the $1 - \epsilon$ value 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 average cost of empirical tests for two types of uncertainty sets via solving (4.25) and (4.26) in Figure 4.25(a) and 4.25(b), respectively. The optimal cost of the robust dispatch framework shows that the result of minimized worst case scenario

for all possible r_c included in the uncertainty set, and the average cost of empirical tests show the real world scenario when we applying the optimal solution to dispatch taxis under random demand r_c. The horizontal line shows the average cost of non-robust solutions since this cost is 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 shown in Figure 4.25(a), $\epsilon = 0.3$ provides the smallest average experimental cost, and for SOC type of uncertainty set shown in Figure 4.25(b), $\epsilon = 0.25$ provides the smallest average experimental 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 for the dataset we use in this section, though its computational cost is higher.

CHAPTER 5 : Data-Driven Dynamic Distributionally Robust Resource Allocation

5.1. Introduction

With the transformation to smarter cities and the development of technologies, a large amount of data is collected from networked sensors in real-time [40, 71]. This paradigm provides both opportunities and challenges for improving systems' performance in the city. Considering the trade-off between system's average performance and worst-case performance, robust taxi dispatch techniques with a probabilistic guarantee level for an original chance constrained problem are developed and evaluated based on a realistic dataset in Chapter 4. However, we do not know the average service performance before running empirical testing by the robust dispatch methods developed in Chapter 4. Hence, motivated by the taxi dispatch problem under demand uncertainties, in this chapter, we consider a general form of data-driven dynamic resource allocation problem that takes the optimal average resource allocation cost or payoff under uncertain distributions of the demand as the control goal of the decisions.

We develop a data-driven distributionally robust resource allocation framework to consider spatial-temporally correlated uncertainties, motivated by the problem of taxi dispatch under demand uncertainties. The optimal resource allocation problem has an objective function that is concave in the uncertain demand and convex in the decision variables, with decision variables on the denominator that has not been covered by the optimization literature. The form of objective function is related to the demand-supply ratio, since the demand-supply ratio or supply-demand ratio is one critical factor that affects the utility or price of resources discussed in previous work such as virtual machine allocations of cloud computing [4, 89], bandwidth providing strategy of video-on-demand systems [66, 88], and power systems [33, 52].

We then design an efficient algorithm for constructing uncertain distribution sets of random demand vectors based on theories in hypothesis testing and distributionally robust optimization literature. This construction process is compatible with various machine learning methods. We prove equivalent computationally tractable forms of the distributionally robust resource allocation problem with

the constructed distributional uncertainty set using strong duality.

With a taxi dispatch problem aiming to balance demand-supply ratio at each region of the city with minimum idle driving distance, we evaluate the performance of the distributionally robust resource allocation framework. Based on four years of taxi trip data for New York City, we show that the average demand-supply ratio error is reduced by 28.6%, and the average total idle driving distance is reduced by 10.05%.

The rest of the chapter is organized as follows. The distributionally robust resource allocation problem motivated by a taxi dispatch problem under demand uncertainties is described in Section 5.2. An efficient algorithm for constructing distributional uncertainty sets based on spatial-temporal data is designed in Section 5.3, and generalized to more learning methods in Subsection 5.3.3. An equivalent computationally tractable form of the general distributionally robust resource allocation problem is proved in Section 5.4. With an example of taxi dispatch problem, evaluations based on a real data set are shown in Section 5.5.

Remark 5 *Some parts of the work presented in this chapter have been captured in [58].*

5.2. Dynamic Distributionally Robust Resource Allocation

The robust allocation scheme designed in Chapter 4 shows its advantage in worst-case scenarios compared with non-robust approaches with the example of efficient transportation resource allocation. However, the robust solutions do not provide a value for the average cost before we test the performance empirically. In this section, we propose a dynamic distributionally robust resource allocation model motivated by the multi-stage taxi dispatch problem under demand uncertainties. We first briefly review the robust taxi dispatch problem with an objective of fairly allocating resources with minimum idle driving distance [54, 57]. For the sake of generality, we then define a form of distributionally robust resource allocation problem that covers the taxi dispatch problem formulated in [54, 57].

The resource allocating solutions we consider in this work are calculated in a receding horizon

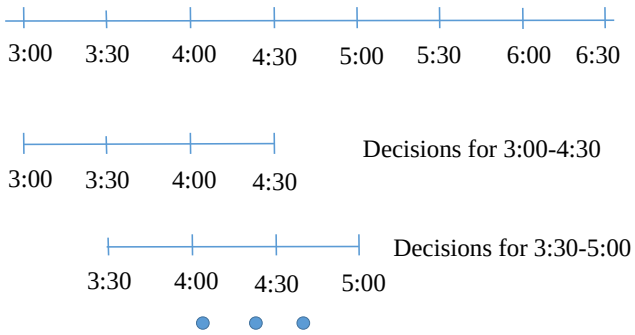


Figure 26: Concept of receding time horizon with 30-minute time periods and $\tau = 3$.

framework. With a time window of τ time slots for $k = 1, 2, \ldots, \tau$, the effect of current decisions to the future allocating cost is involved. The idea of receding time horizon is explained in Figure 26. Only the solution of $k = 1$ is implemented, while the allocating solutions for remaining time slots are not materialized. When the time horizon rolls forward by one time step, information about uncertain demand is first updated, and available resources are observed, provided to solve a new resource allocation problem for the current time window. Examples of receding time horizon format of resource allocation frameworks include economic dispatch of power systems [52], taxi dispatch systems [59], etc.

5.2.1. Problem Formulation

We assume that there are n regions (nodes) to be served, with $r_j^k \geqslant 0$ as the predicted total amount of demand (number of passengers for a taxi dispatch system) within region j during time window k, $j = 1, \ldots, n$, $k = 1, \ldots, \tau$. We define $r^k \in \mathbb{R}^n$ as a random demand vector instead of a deterministic value, and demand during every τ consecutive time slots also have spatial-temporal correlations. Hence, we define the concatenation of demand sequences $(r^1 \in \mathbb{R}^n, \ldots, r^\tau \in \mathbb{R}^n)$ as

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

We assume that F^* is the true distribution function for the random vector r_c, i.e., $r_c \sim F^*$.

We consider a single type of resource allocation problem under the above demand model. We denote

by a nonnegative matrix X^k the matrix of resource allocation decisions at time k, where

$$X^k \in \mathbb{R}_+^{n\times n}, \quad X_{ij}^k \geq 0,$$

and X_{ij}^k is the amount of resource (number of taxis for a taxi dispatch problem) sent from region i to region j (or node i to node j) at time k according to demand or service requirements. For notational convenience, we define a concatenation of decision variables as

$$X^{[1,\tau]} = [X^1\ X^2\ \ldots\ X^\tau].$$

With an objective function $J(X^{[1,\tau]}, r_c)$ related to the random demand r_c, a stochastic optimization form of resource allocation problem is defined as the following

$$\begin{aligned} &\min_{X^k}\ \mathbb{E}_{r_c\sim F^*}\left[J(X^{[1,\tau]}, r_c)\right] \\ &\text{s.t} \quad X^1,\ldots,X^\tau \in \mathcal{D}_c. \end{aligned} \tag{5.1}$$

However, in many application problems we only limited knowledge about the true distribution function F^*. Moreover, problem (5.1) is computationally demanding, not suitable for a large-scale dynamic resource allocation framework in general. With historical or streaming data (or prior knowledge if there is any), we assume that we are able to construct a set of distribution functions $\mathcal{F}$ such that $F^* \in \mathcal{F}$. Then the uncertainty information about demand r_c is described through $\mathcal{F}$. In this work, we propose the following form of distributionally robust resource allocation problem as a robust form of problem (5.1) to minimize the worst-case expected cost

$$\begin{aligned} &\min_{X^k}\ \max_{F\in\mathcal{F}}\quad \mathbb{E}\left[J(X^{[1,\tau]}, r_c)\right] \\ &\qquad\text{s.t} \quad X^1,\ldots,X^\tau \in \mathcal{D}_c. \end{aligned} \tag{5.2}$$

Then by solving (5.2), the average resource allocation cost is guaranteed to be smaller than the optimal solution of (5.2), since we minimize the expected cost for the worst-case distribution function included in $\mathcal{F}$.

Specifically, to define the form of $J(X^{[1,\tau]}, r_c)$, we first introduce an example of fair resource allocation problem— a taxi dispatch problem. We take the definitions of objective and constraint functions of the robust taxi problem defined in [54], and a distributionally robust taxi dispatch problem considered in this work has the following form

$$\begin{aligned}
\min_{X^k, L^k F \in \mathcal{F}} \max \mathbb{E} \left[J = \sum_{k=1}^{\tau} (J_D(X^k) + \beta J_E(X^{[1,\tau]}, r^k)) \right] \\
\text{s.t.} \quad (L^{k+1})^T = (\mathbf{1}_n^T X^k - (X^k \mathbf{1}_n)^T + (L^k)^T) C^k, \\
\mathbf{1}_n^T X^k - (X^k \mathbf{1}_n)^T + (L^k)^T \geqslant 1, \\
X_{ij}^k W_{ij} \leq m X_{ij}^k, \\
X_{ij}^k \geq 0, \quad i, j \in \{1, 2, \ldots, n\},
\end{aligned} \tag{5.3}$$

where

$$\begin{aligned}
J_D(X^k) &= \sum_i \sum_j X_{ij}^k W_{ij}, \\
J_E(X^{[1,\tau]}, r^k) &= \sum_i \frac{r_i^k}{(\mathbf{1}_n^T X_{\cdot i}^k - X_{i\cdot}^k \mathbf{1}_n + L_i^k)^\alpha}.
\end{aligned}$$

Here $J_D(X^k)$ measures the resource balancing and re-balancing cost, $J_E(X^{[1,\tau]}, r^k)$ is a penalty function for violating service fairness that relates to the demand-supply ratio of each region, and L^{k+1} is the amount of available resources at time $k+1$ (released resource after serving tasks during time k) before allocating resources as X^{k+1}.

The above distributionally robust taxi dispatch problem cannot be immediately translated into an LP or SDP form. The fairness requirement is encoded in an objective function that has decision variables on the denominator. Motivated by it, we consider a general form of function $J_E(X^{[1,\tau]}, r^k)$ as a metric to be minimized and a measurement of how resource is allocated to serve demand according to the requirements. We define

$$s : \mathbb{R}^{n \times \tau n} \to \mathbb{R}_+^{\tau n}$$

as a function of the decision variables $X^{[1,\tau]}$, and

$$[s(X^{[1,\tau]})]_{(k-1)n+i} > 0$$

is the $((k-1)n+i)$-th component of $s(X^{[1,\tau]})$ such that $\frac{1}{[s(X^{[1,\tau]})]_{(k-1)n+i}}$ is convex of $X^k, k = 1,\ldots,\tau$. And $J_E(X^{[1,\tau]}, r^k)$ takes the following form with constants $a_{ik} > 0, i = 1,\ldots,n$, $k = 1,\ldots,\tau$

$$J_E(X^{[1,\tau]}, r^k) = \sum_i \left(\frac{a_{ik} r_i^k}{[s(X^{[1,\tau]})]_{(k-1)n+i}} \right). \tag{5.4}$$

Here, $J_E(X^{[1,\tau]}, r^k)$ is a function concave (linear) in r^k and convex in $X^k, k = 1,\ldots,\tau$ that measures how demand is matched with the resource allocating strategy, and $J_E(X^{[1,\tau]}, r^k)$ has the decision variables on the denominator. Assume that $J_D(X^k)$ is a convex cost function for allocating or re-allocating resources, and $\mathcal{D}_c$ is a convex domain of the decision variables that describes the constraints of the resource allocating strategies. Then a distributionally robust resource allocation problem considered in this work is

$$\begin{aligned} \min_{X^k} \max_{F \in \mathcal{F}} \mathbb{E} & \left[\sum_{k=1}^{\tau} \left(J_D(X^k) + \beta \sum_i \frac{a_{ik} r_i^k}{[s(X^{[1,\tau]})]_{(k-1)n+i}} \right) \right] \\ \text{s.t} \quad & X^1, \ldots, X^\tau \in \mathcal{D}_c, \end{aligned} \tag{5.5}$$

5.2.2. Forms of Objective Function

Problem (5.3) is one example of fair resource allocation covered by the general form of problem defined in (5.5), where $a_{ik} = 1$, and

$$[s(X^{[1,\tau]})]_{(k-1)n+i} = (\mathbf{1}_n^T X_{\cdot i}^k - X_{i\cdot}^k \mathbf{1}_n + L_i^k)^\alpha$$

is related to the total number of available resources that can provide service within region i during time k. The power α is a constant parameter designed according to the objective. For instance, with

$\alpha \to 0$ in (5.3), as explained in Chapter 4, a surrogate function for balanced demand-supply ratio at each region is part of the objective function.

For other fair resource allocation problems with a metric that demand-supply ratio at each region should be as close to the global level as possible, we can use a similar form of objective function. For instance, when the total amount of resource is limited and fixed (smaller than the total number of demand), it is impossible to satisfy the demand of all users at the same time. Under this scenario, the most efficient way to fairly allocate a single type of resource is to use all [45]. Then for a fair single resource allocation, let the function $s(X^{[1,\tau]})$ be

$$[s(X^{[1,\tau]})]_{(k-1)n+i} = ([S(X^{[1,\tau]})]_i^k)^\alpha,$$

where $[S(X^{[1,\tau]})]_i^k$ is the total amount of resource available within region i during time k (but may not have the exact form of $(\mathbf{1}_n^T X_{\cdot i}^k - X_{i\cdot}^k \mathbf{1}_n + L_i^k)$ in taxi dispatch problem (5.3)), and

$$N^k = \sum_i [S(X^{[1,\tau]})]_i^k$$

is the total amount of available resource during time k. Then problem (5.5) is a distributionally robust form of fair resource allocation problem given uncertain demand r^k and limited total resources N^k, $k = 1, 2, \ldots, \tau$.

For queuing models, the average number of waiting customers in the queue is related to the demand-supply ratio or supply-demand ratio for a stable queue [53, 14]. It also indicates that considering a balanced demand-supply ratio is to consider balance the average number of waiting customers intuitively.

Region priorities: Taking into account service priority of different regions in one city involves simply adjusting the value of $a_i k$. In problem (5.3), $a_{ik=1}, i = 1, \ldots, n,\ k = 1, \ldots, \tau$, and the resource allocation strategy aims to provide fair service for each region. We can give a higher priority to regions with important events or assign weight, or values of a_{ik} according to price incentives.

5.3. Efficient Distributional Set Construction Algorithm

We design an algorithm for constructing the distributional set $\mathcal{F}$ of problem (5.5), with spatial-temporal data that provides information about the true distribution function F^* of the demand vector r_c. Delage and Ye propose a model of distributional set and prove a confidence region for the mean and the covariance matrix of a random vector [28]. While applying the theoretical bound of the distributional set is too conservative in practice, with a large enough dataset, constructing $\mathcal{F}$ via a bootstrap method [18] and hypothesis testing results good empirical performance in portfolio management problems [28, 12]. How to model spatial-temporally correlated demand uncertainties based on thresholds of accepted hypothesis testing is first analyzed in work [57]. Considering the computational cost of building a distributional set for each time window of one day, we modify the bootstrapped uncertainty set construction algorithm and develop a more efficient algorithm in this section.

To describe the demand changing trend at different time of one day, we assume that one day is discretized as K time slots in total, and the demand of each region during one time slot is described as $r^h, h = 1, \ldots, K$. We denote one sample of vector

$$r_c(t) = [(r^t)^T, (r^{(t+1)})T, \ldots, (r^{(t+\tau)})T]^T$$

at date d_l as $\tilde{r}_c(d_l, t)$, a vector of aggregated total number of demand at each region for time slots $h = t, t+1, \ldots, t+\tau$. We define the distribution uncertainty set for a random demand vector $r_c(t)$ as $\mathcal{F}(t), t = 1, 2, \ldots, K$. Demand sampled from N days

$$\tilde{r}_c(d_1, t), \tilde{r}_c(d_2, t), \ldots, \tilde{r}_c(d_N, t)$$

for time index t are independent with each other for every time index t. Hence, for each time index t, we aim to construct a distributional set $\mathcal{F}(t)$ that describes possible distribution function of $r_c(t)$ based on the support, mean and covariance values of a random vector of a given dataset.

For notational convenience, we omit t for the following problem definition. Based on the distributional set designed in [28] and the bootstrap algorithm for calculating the support (range), mean and covariance values [12], the problem of constructing a distributional set is defined as

Problem 2 *Given a dataset of r_c, find the values of $\hat{r}_c, \hat{\Sigma}_c$, γ_1^B and γ_2^B, with probability at least $1-\alpha$ with respect to the samples, the following distributional set $\mathcal{F}$ is true for r_c based on the given dataset*

$$\begin{aligned}&\mathcal{F}(\hat{r}_{c,l}, \hat{r}_{c,h}, \hat{r}_c, \hat{\Sigma}_c, \gamma_1^B, \gamma_2^B)\\ =&\{r_c \in [\hat{r}_{c,l}, \hat{r}_{c,h}] : (\mathbb{E}[r_c] - \hat{r}_c)^T \hat{\Sigma}_c^{-1} (\mathbb{E}[r_c] - \hat{r}_c) \leqslant \gamma_1^B,\\ &\mathbb{E}[(r_c - \hat{r}_c)(r_c - \hat{r}_c)^T] \leq \gamma_2^B \hat{\Sigma}_c\}\end{aligned} \tag{5.6}$$

where $supp(r_c) \subset [\hat{r}_{c,l}, \hat{r}_{c,h}]$ is the support of r_c, $\hat{r}_{c,l}$ and $\hat{r}_{c,h}$ is the lower bound and higher bound of each component of the demand vector, respectively.

Problem 2 is related to a hypothesis testing H_0 given a dataset of random vector r_c: given mean μ_0 and covariance Σ_0, test statistics γ_1, γ_2, with probability at least $1-\alpha$, the random vector r_c satisfies that

$$\begin{aligned}H_0 :&(\tilde{r}_c - \mu_0)^T \Sigma_0^{-1} (\tilde{r}_c - \mu_0) \leqslant \gamma_1,\\ &(\tilde{r}_c - \mu_0)(\tilde{r}_c - \mu_0)^T \preceq \gamma_2 \Sigma_0.\end{aligned} \tag{5.7}$$

Since we do not have prior knowledge about the support, the true mean, covariance, and threshold values γ_1, γ_2 of the test statistics, constructing set $\mathcal{F}$ based on data is an inverse process of a hypothesis testing. We then design Algorithm 3 to calculate the bootstrapped estimations of $\hat{r}_{c,l}, \hat{r}_{c,h}, \hat{r}_c, \hat{\Sigma}_c, \gamma_1^B, \gamma_2^B$ for every $r_c(t), t = 1, 2, \ldots, K$, that makes H_0 defined in (5.7) acceptable and consistent with data.

5.3.1. Reducing Computational Complexity

The computational cost of constructing a distributional set with bootstrapped method for spatial-temporal data considered in this work is higher than that of the return model of financial assets in

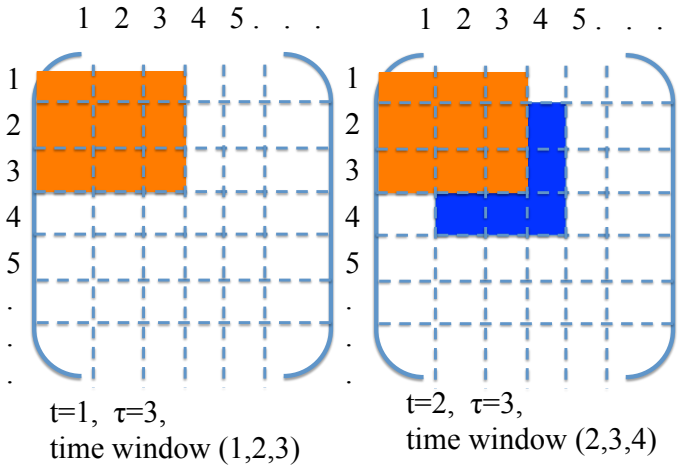


Figure 27: The idea of calculating $\hat{\Sigma} \in \mathbb{R}^{Kn \times Kn}$ when receding time horizon. For example, when index moves from $t = 1$ to $t = 2$, only the blocks of components in matrix $\hat{\Sigma}$ shown in blue are new and necessary for calculating $\hat{\Sigma}_c(t), t = 2$, and we only calculate these blocks of variance and covariance matrices, store them in the corresponding positions of matrix $\hat{\Sigma}$ for the future computing process.

the literature [12, 28]. This is because $\mathcal{F}(t)$ is a function of time index t, the dimension of $\hat{r}_c, \hat{\Sigma}_c$ is decided by the number of regions n and prediction horizon τ, which can be large for applications rising in smart cities, such as taxi or autonomous driving car dispatch problems and bicycle re-balancing problems.

However, the mean and covariance matrices for $t, t+1, \ldots, t+\tau$ have overlapping components: for instance, $\hat{r}_c(t)$ and $\hat{r}_c(t+1)$ both include estimated mean values of demand during time $(t+1, t+2, \ldots, t+\tau)$. Hence, instead of always repeating the process of calculating a mean and covariance value for τ time slots together for each index t, the key idea of reducing computational cost of constructing $\mathcal{F}(t), t = 1, \ldots, K$ is to calculate the mean and covariance of each pair of time slots of the whole day only once. Then pick up the corresponding components needed to construct $\hat{r}_c(t)$ and $\hat{\Sigma}_c(t)$ for each index t.

Specifically, we define the whole day demand vector as $r = [(r^1)^T, (r^2)^T, \ldots, (r^K)^T]^T \in \mathbb{R}^{Kn}$, i.e., a concatenated demand vector that includes the total number of requests within each region at each time slot of one day. And we denote $\hat{r}$ as the estimated mean of the random vector r. To get all

covariance component for each index t, the process is: at $t = 1$, calculate the covariance of $r_c(1)$, store it as $\bar{\Sigma}_{[1:n,1:n]}$; and every time when rolling the time horizon from t to $t+1$, only calculate the covariance between τ pairs of $(r^{t+\tau}, r^{t+k})$ and store the result as

$$\bar{\Sigma}_{[(t-1+\tau)n:(t+\tau)n,(t-1+k)n:(t+k)n]} = \bar{\Sigma}_{[(t-1+k)n:(t+k)n,(t-1+\tau)n:(t+\tau)n]} = \text{cov}(r^{t+\tau}, r^{t+k}) \quad (5.8)$$

for $k = 1, 2, \ldots, \tau$, where $[(t-1+k)n : (t+k)n]$ means components from the $(t-1+k)n$-th to the $(t+k)n$-th row or column in the matrix. This process of calculating $\hat{\Sigma}$ is explained in Figure 27.

Remark 6 *The computational complexity of repeating the process of calculating $\hat{r}_c(t)$, $\hat{\Sigma}_c(t)$ for each index t is $O(BN_BKn^2\tau^2)$, while the computational complexity of calculating $\hat{r}, \hat{\Sigma}$ for the whole day first and picking up the corresponding components for each index t is $O(BN_BKn^2\tau)$.*

5.3.2. Algorithm

Then we have the following Algorithm 3 that describes the complete process of constructing distributional sets. For instance, given a taxi trajectory or trip data, we count the total number of pick up events during one hour at each region as $r^k, k \in \{1, 2, \ldots, \tau\}$ according to the start time and GPS coordinate of the pick-up position of each trip. If the given dataset is the arriving time of each customer at different service nodes of a network, then the total number of customer appeared in every service node during each hour or every 30 minutes is a vector r^k, and we concatenate τ time slots of r^k as one vector r_c. The motivation of partitioning or clustering the entire dataset to several subsets is explained in the uncertainty set constructing algorithm of work [57]. We denote $I_p, p = 1, 2, \ldots, P$ as the categorical information index for data partition. A partition category can be high demand season or low demand season of one year, normal days or holidays/special event days of one year, different weather conditions or a combination of different contexts, etc. It depends on information available to the process of constructing distributional sets.

For step $3(1)$, the process of picking components from the mean and covariance matrices of the

Algorithm 3: Algorithm for constructing distributional sets

Input: A dataset of spatial-temporal demand

1. Demand aggregating and sample set partition

Aggregate demand to get a sample set $\mathcal{S}$ of demand for the whole day r (denote $\mathcal{S}(t)$ as a sample set for $r_c(t)$) from the original data. Partition $\mathcal{S}(\mathcal{S}(t))$ and denote $\mathcal{S}(I_p) \subset \mathcal{S}$ ($\mathcal{S}(t, I_p) \subset \mathcal{S}(t)$), $p = 1, \ldots, P$ as the subset partitioned according to categorical information I_p.

2. Bootstrapping mean and covariance matrix

Initialization: a significance level $0 < \alpha_h < 1$, the number of bootstrap time $N_B \in \mathbb{Z}_+$.

for $j = 1, \ldots, N_B$ **do**

Re-sample $\mathcal{S}^j(I_p) = \{\tilde{r}(d_1, I_p), \ldots, \tilde{r}(d_N, I_p)\}$ from $\mathcal{S}(I_p)$ with replacement. Get the mean and covariance matrix of the whole day demand vector of sample set $\mathcal{S}^j(I_p)$, denoted as $\bar{r}^j(I_p)$ and $\bar{\Sigma}^j(I_p)$ (calculated as (5.8)), respectively.

end for

Get the bootstrapped mean covariance, and support of the whole day demand vector ($i = 1, \ldots, Kn$)

$\hat{r}(I_p) = \frac{1}{B} \sum_{j=1}^{B} \bar{r}^j(I_p)$, $\hat{\Sigma}(I_p) = \frac{1}{B} \sum_{j=1}^{B} \bar{\Sigma}^j(I_p)$,

$\hat{r}_{i,l}(I_p) = min_d \tilde{r}_i(d, I_p)$, $\hat{r}_{i,h}(I_p) = max_d \tilde{r}_i(d, I_p)$, for all samples $\tilde{r}(d, I_p)$ in the subset $\mathcal{S}(I_p)$.

3. Bootstrapping γ_1^B and γ_2 for each time index t

for each subset $\mathcal{S}^j(t, I_p)$ **do**

for $j = 1, \ldots, N_B$ **do**

(1) Get the mean and covariance vector for time index t of the bootstrapped estimation, and the j-th re-sample, from the mean and covariance matrix of the whole day demand vector in step 2: $\hat{r}_c(t, I_p)$, $\hat{\Sigma}_c^j(t, I_p)$, $\bar{r}_c^j(t, I_p)$, $\bar{\Sigma}_c^j(t, I_p)$.

(2). Get values of $\gamma_1^j(t, I_p)$ and $\gamma_2^j(t, I_p)$ according to (5.9) and (5.10), respectively.

end for

(3). Get 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.

end for

3. Calculate the model of distributionally uncertainty sets

Get the model of set defined as (5.6) for every t and I_p.

Output: Distributionally uncertainty sets for problem (5.5)

whole day demand is

$$\hat{r}_c(t, I_p) = \hat{r}_{[tn:(t+\tau)n]}(I_p), \quad \bar{r}_c^j(t, I_p) = \bar{r}^j_{[tn:(t+\tau)n]}(I_p),$$
$$\hat{\Sigma}_c^j(t, I_p) = \hat{\Sigma}^j_{[tn:(t+\tau)n, tn:(t+\tau)n]}(I_p), \quad \bar{\Sigma}_c^j(t, I_p) = \bar{\Sigma}^j_{[tn:(t+\tau)n, tn:(t+\tau)n]}(I_p),$$

where $[tn : (t + \tau)n]$ means components from the tn-th to the $(t + \tau)n$-th row or column of a vector/matrix.

For the j-th re-sampled subset $\mathcal{S}^j(t, I_p)$, the mean and covariance matrices are $\mathbb{E}[r_c] = \bar{r}_c^j(t, I_p)$ and $\mathbb{E}[r_c r_c^T] = \bar{\Sigma}_c^j(t, I_p)$, respectively. For step $3(2)$, according to the definition of $\mathcal{F}$ in (5.6), we get $\gamma_1^j(t, I_p)$ by the following equation

$$\gamma_1^j(t, I_p) = [\bar{r}_c^j(t, I_p) - \hat{r}_c(t, I_p)]^T \hat{\Sigma}_c^{-1}(t, I_p)[\bar{r}_c^j(t, I_p) - \hat{r}_c(t, I_p)]. \tag{5.9}$$

According to definition (5.6), the left part of the inequality related to γ_2^B satisfies that

$$\mathbb{E}[(r_c - \hat{r}_c)(r_c - \hat{r}_c)^T] = \mathbb{E}[r_c r_c^T] - \hat{r}_c \mathbb{E}[r_c^T] - \mathbb{E}[r_c]\hat{r}_c^T + \hat{r}_c \hat{r}_c^T = \bar{\Sigma}_c - \hat{r}_c \hat{r}_c^T.$$

Then we get γ_2^j for index (t, I_p) by solving the convex optimization problem

$$\begin{aligned} \min_{\gamma_2} \quad & \gamma_2 \\ \text{s.t} \quad & \bar{\Sigma}_c^j(t, I_p) - [\hat{r}_c(t, I_p)][\hat{r}_c(t, I_p)]^T \leq \gamma_2 \hat{\Sigma}_c(t, I_p) \end{aligned} \tag{5.10}$$

5.3.3. Constructing Uncertainty Sets for a General Demand Prediction Model

The above Algorithm 3 considers to construct an uncertainty set of the concatenated demand vector r_c, and the estimated demand $\hat{r}_c(t)$ for each index t is the average value of bootstrapped samples. It is worth noting that besides directly building an uncertainty set for r_c, Algorithm 3 is also compatible with a general modeling method, that we can follow a similar process to build an uncertainty set for the estimation residual.

We do not restrict the learning or modeling method to predict demand, and assume that $f_r : \mathcal{I}_{[k-l,k]} \to \mathbf{R}^n$ is a function that mapping sensing data available to the system by time k (from time $(k-l)$ to time k) to predicted demand at time $k+1$

$$\hat{r}_{k+1} = f_r(I_{[k-l,k]}), \quad r_{k+1} = \hat{r}_{k+1} + \delta_{k+1}. \tag{5.11}$$

Here $\delta_{k+1} \in \mathbb{R}^n$ is the estimation residual that measures the difference between the true demand and the estimated value. One example of model (5.11) is time series function [77]. The available data $\mathcal{I}_{[k-l,k]}$ can be either purely historical data stored in the system, or purely on-line/streaming or real-time vehicle state monitoring data, or both.

Then for each sample $\tilde{r}_{k+1}$ of r_{k+1}, a corresponding sample of residual is

$$\tilde{\delta}_{k+1} = \tilde{r}_{k+1} - \hat{r}_{k+1}.$$

For a subset of samples $\mathcal{S}(k+1) = \{\tilde{r}_{k+1}\}$, there will be one estimated value for $\hat{r}_{k+1}$, and the corresponding mean and covariance values for the residual δ_{k+1}.

When constructing uncertainty set of r_k with prediction function f_r by the bootstrapped process Algorithm 3, every step is the same, except one step — we use the estimation equation (5.11) instead of the mean value of all samples for the estimated $\hat{r}_{k+1}$. It is worth noting that even for an on-line learning algorithm such as the short-term time horizon demand prediction approach using streaming data [62], the uncertainty set construction Algorithm 3 can be run off-line. Then the predicted demand (5.11) is a sum of estimation based on streaming data and residual quantified by a closed convex set calculated via historical data.

Similarly, to build an uncertainty set for the concatenated demand vector r_c based on prediction method f_r, we only need to calculate the estimated concatenated demand $\hat{r}_c(t+1)$ as

$$\hat{r}_c(t+1) = f_r(I_{c,[t-l,t]}), \quad r_c(t+1) = \hat{r}_c(t+1) + \delta_{t+1}, \tag{5.12}$$

where $I_{c,[t-l,t]}$ is the available data related to r_c by time t (from time (t_l) to time (t)).

5.4. Computationally Tractable Form

In this section, we present the main theorem of this work–equivalent computationally tractable form of the distributionally robust resource allocation (5.5). By the definition of the objective function and constraints, only $J_E(r^k, X^{[1,\tau]})$ part of problem (5.5) is related to the random demand r_c. Hence, in the equivalent computationally tractable problem, the form of $J_D(X^k)$ keeps the same and the process of converting problem (5.5) to a convex optimization problem is mainly about finding an equivalent form for the $J_E(r^k, X^{[1,\tau]})$ part. The objective function of the resource allocation problem defined in this work is convex over the decision variables and concave (linear) over the constructed uncertain sets, with decision variables on the denominators. This form is not a linear programming (LP) or a semi-definite programming (SDP) examined by previous work [8, 12, 26].

The following theorem shows an equivalent convex optimization form for problem (5.5) with the objective function defined as (5.4) in this work.

Theorem 5 *The distributionally robust resource allocation problem defined in* (5.5) *has the following equivalent convex optimization form*

$$
\begin{aligned}
\textit{min.} \quad & \beta(v+t) + \sum_{k=1}^{\tau} J_D(X^k) \\
\textit{s.t} \quad & \begin{bmatrix} v + (y_1^+)^T \hat{r}_{c,l} - (y_1^-)^T \hat{r}_{c,h} & \frac{1}{2}(q - y - y_1)^T \\ \frac{1}{2}(q - y - y_1) & Q \end{bmatrix} \succeq 0 \\
& t \geqslant (\gamma_2^B \hat{\Sigma}_c + \hat{r}_c \hat{r}_c^T) \cdot Q + \hat{r}_c^T q + \sqrt{\gamma_1^B} \|\hat{\Sigma}_c^{1/2}(q + 2Q\hat{r}_c)\|_2 \\
& \frac{a_{ik}}{[s(X^{[1,\tau]})]_{(k-1)n+i}} \leqslant y_{(k-1)n+i} \\
& y_1 = y_1^+ - y_1^-, y_1^+, y_1^-, y \geqslant 0, Q \succeq 0 \\
& X^1, \ldots, X^\tau \in \mathcal{D}_c.
\end{aligned}
\tag{5.13}
$$

Proof 8 *We have* $\frac{a_{ik}}{[s(X^{[1,\tau]})]_{(k-1)n+i}} > 0$ *and* $r_c \geqslant 0$ *by the definitions of* J_E *in* (5.4) *and the demand model, then for any vector* $y \in \mathbb{R}^{\tau n}$ *that satisfies*

$$0 < \frac{a_{ik}}{[s(X^{[1,\tau]})]_{(k-1)n+i}} \leqslant y_{(k-1)n+i},$$

we also have

$$0 \leqslant \sum_{k=1}^{\tau} \sum_{i} \frac{a_{ik} r_i^k}{[s(X^{[1,\tau]})]_{(k-1)n+i}} \leqslant y^T r_c,$$

and the second inequality strictly holds when all

$$\frac{a_{ik} r_i^k}{[s(X^{[1,\tau]})]_{(k-1)n+i}} = y_{(k-1)n+i}, \quad i = 1, \ldots, n, \quad k = 1, \ldots, \tau >$$

The constraints of problem (5.5) *are independent of* r_c*, hence, for any* r_c*, the following minimization problem*

$$\begin{aligned} \underset{X^k}{\textit{min.}} \quad & \beta \sum_{k=1}^{\tau} \sum_{i} \frac{a_{ik} r_i^k}{[s(X^{[1,\tau]})]_{(k-1)n+i}} + \sum_{k=1}^{\tau} J_D(X^k) \\ \textit{s.t} \quad & X^1, \ldots, X^\tau \in \mathcal{D}_c \end{aligned}$$

is equivalent to

$$\begin{aligned} \underset{X^k}{\textit{min.}} \quad & \beta y^T r_c + \sum_{k=1}^{\tau} J_D(X^k) \\ \textit{s.t} \quad & \frac{a_{ik}}{[s(X^{[1,\tau]})]_{(k-1)n+i}} \leqslant y_{(k-1)n+i}, \\ & X^1, \ldots, X^\tau \in \mathcal{D}_c \end{aligned} \tag{5.14}$$

In the following proof, we use the objective function of problem (5.14)*. In particular, only the part of* $y^T r_c$ *is related to* r_c*, and we first consider the following maximum problem*

$$\underset{r_c \sim F, F \in \mathcal{F}}{\textit{max}} \mathbb{E}[y^T r_c] \tag{5.15}$$

By the definition of problem (5.5) *and problem* (5.14)*, only the objective function includes the random vector r_c, and is concave of r_c, convex of $X^k, k = 1, \ldots, \tau$. The distributional set $\mathcal{F}$ constructed by Algorithm 3, the domain of y, X^k, $k = 1, \ldots, \tau$ are convex, closed, and bounded sets. Hence, problem* (5.15) *satisfies the conditions of Lemma 1 in [28], and the maximum expectation value of $y^T r_c$ for any possible $r_c \sim F, F \in \mathcal{F}$ equals to the optimal value of the problem*

$$\begin{aligned} \underset{Q,q,v,t}{\textit{min.}} \quad & v + t \\ \textit{s.t} \quad & v \geqslant y^T r_c - r_c^T Q r_c - r_c^T q, \quad \forall r_c \in [\hat{r}_{c,l}, \hat{r}_{c,h}] \\ & t \geqslant (\gamma_2^B \hat{\Sigma}_c + \hat{r}_c \hat{r}_c^T) \cdot Q + \hat{r}_c^T q + \sqrt{\gamma_1^B} \|\hat{\Sigma}_c^{1/2}(q + 2Q\hat{r}_c)\|_2 \\ & Q \succeq 0. \end{aligned} \tag{5.16}$$

Hence, we first analytically find the optimal value of problem (5.16)*. Note that the first constraint about v is equivalent to $v \geqslant f(r_c^*, y)$, where $f(r_c^*, y)$ is the optimal value of the following problem*

$$\begin{aligned} \underset{r_c}{\textit{max.}} \quad & y^T r_c - r_c^T Q r_c - r_c^T q \\ \textit{s.t} \quad & \hat{r}_{c,l} \leqslant r_c \leqslant \hat{r}_{c,h}. \end{aligned} \tag{5.17}$$

For a positive semidefinite Q, the optimal solution of problem (5.17) *exists. The Lagrangian of* (5.17) *under the constraint $y_1^+, y_1^- \geqslant 0$ is*

$$\mathcal{L}(r_c, y_1^+, y_1^-) = y^T r_c - r_c^T Q r_c - r_c^T q + (y_1^+ - y_1^-)^T r_c - (y_1^+)^T \hat{r}_{c,l} + (y_1^-)^T \hat{r}_{c,h}.$$

When $Q \succeq 0$, the supreme value of the Lagrangian is calculated via taking the partial derivative over r_c, let $\Delta_{r_c} \mathcal{L} = 0$, and

$$\begin{aligned} \underset{r_c}{\textit{sup}} \mathcal{L}(r_c, y_1^+, y_1^-) =& \frac{1}{4}(q - y - y_1)^T Q^{-1}(q - y - y_1) - (y_1^+)^T \hat{r}_{c,l} + (y_1^-)^T \hat{r}_{c,h}, \\ y_1 =& y_1^+ - y_1^-, y_1^+, y_1^- \geqslant 0. \end{aligned}$$

Then the first inequality constraint of problem (5.16) *for any* $\hat{r}_{c,l} \leqslant r_c \leqslant \hat{r}_{c,h}$ *is equivalent to*

$$v \geqslant \frac{1}{4}(q - y - y_1)^T Q^{-1}(q - y - y_1) - (y_1^+)^T \hat{r}_{c,l} + (y_1^-)^T \hat{r}_{c,h}.$$

By Schur complement, the above constraint is

$$\begin{bmatrix} v + (y_1^+)^T \hat{r}_{c,l} - (y_1^-)^T \hat{r}_{c,h} & \frac{1}{2}(q - y - y_1)^T \\ \frac{1}{2}(q - y - y_1) & Q \end{bmatrix} \succeq 0$$

Together with other constraints, the equivalent convex optimization form of problem (5.5) *is problem* (5.13).

Specifically, with the constraints of problem (5.3) to represent the constraint $X^1, \ldots, X^\tau \in \mathcal{D}_c$ in (5.13), and

$$\frac{a_{ik}}{[s(X^{[1,\tau]})]_{(k-1)n+i}} = \frac{r_i^k}{(\mathbf{1}_n^T X_{\cdot i}^k - X_{i\cdot}^k \mathbf{1}_n + L_i^k)^\alpha},$$

we have a computationally tractable form for the distributionally robust taxi dispatch problem (5.3).

5.5. Evaluations with Taxi Trip Data

With taxi dispatch problem as one example of resource allocation problem, we evaluate the performance of the distributionally robust dispatch framework (5.3) considered in this work based on four years of taxi trip data in New York City [29]. Information for every record includes the GPS coordinators of locations, and the date and time (with precision of seconds) of pick up and drop off locations, as summarized in Table 7. We construct distributional uncertainty sets according to Algorithm 3, compare the average dispatch cost of the distributionally robust dispatch method (5.3) with the robust dispatch model and non-robust dispatch method introduced in [57] in this section.

How does the number of samples affect the accuracy of distributional set: We partition the map of New York City shown in Figure 20 into different number of equal-area grids and count the total number of pick-up events within each region as the total demand. Then we compare the values

		Γ_1^B	Γ_2^B
$N_B = 10$	$n = 50, \tau = 2$	0.739	5.24
$N_B = 100$	$n = 50, \tau = 2$	0.368	2.47
$N_B = 1000$	$n = 50, \tau = 3$	0.013	1.56
$N_B = 5000$	$n = 50, \tau = 6$	0.012	1.49

Table 11: Comparing thresholds γ_1^B and γ_2^B for different N_B and dimensions of r_c

of γ_1^B and γ_2^B resulting from Algorithm 3. The set construction Algorithm 3 captures information about the support, the first and second moments of the random vector. We show the value of γ_1^B and γ_2^B with different sample numbers N_B in Algorithm 3 and the dimensions of r_c or τn values in table 11. When the value of N_B is increased, values of γ_1^B and γ_2^B are reduced, which means the volume of the distributional set is smaller. For a large enough N_B, the value of τn does not affect the results of γ_1^B and γ_2^B much.

Compare different types of robust solutions and non-robust solutions: To compare the average dispatch cost of different methods, we use the idea of cross-validation from machine learning. All data is separated as a training subset for constructing the uncertain distribution set and a testing subset for comparing the true costs of different dispatch solutions for each time of testing. The cost of each dispatch solution, such as the distributionally robust method (5.3) or the robust dispatch model of [57] is a weighted sum of both the demand-supply ratio mismatch error and estimated total idle driving distance. For each testing example r^k, we denote the demand-supply ratio mismatch error of a dispatch solution as the following:

$$\sum_{k=1}^{\tau}\sum_{i=1}^{n}\left|\frac{r_i^k}{\mathbf{1}_n^T X_{\cdot i}^k - X_{i\cdot}^k \mathbf{1}_n + L_i^k} - \frac{\mathbf{1}_n^T r^k}{N}\right|. \tag{5.18}$$

The idle distance of each taxi between two trips with passengers is approximated as the distance between one drop-off event and the following-up pick-up event. We use bootstrapped mean value of the training dataset as predicted demand for the non-robust dispatch framework in the experiments.

We compare the average costs of cross-validation tests for the distributionally robust dispatch solutions via solving (5.13), two types of uncertainty sets of the robust dispatch methods designed

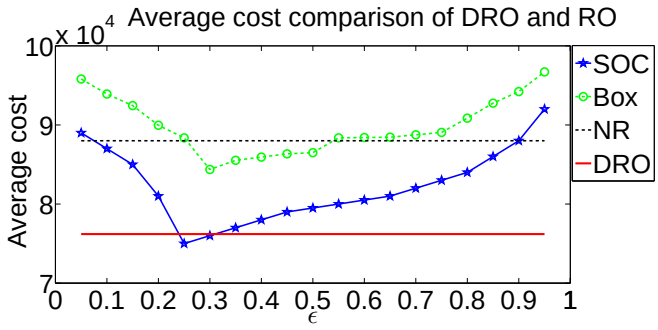


Figure 28: The average cost of empirical tests for the distributionally robust dispatch solutions via solving (5.13), two types of uncertainty sets of the robust dispatch methods designed in [57] and non-robust dispatch solutions. The line "DRO" represents the average cost of the distributionally robust dispatch solutions via solving problem (5.13).

in [57] and non-robust dispatch solutions in Figure 28. The average costs show the real world scenario when we applying the optimal solution of each method to dispatch taxis under all testing samples of the random demand r_c.

The minimum average cost of an SOC robust dispatch solution is close to the average cost of the distributionally robust dispatch solutions of (5.13). They both use the first and second moments information of the random demand vector. In particular, the average demand-supply ratio mismatch error is reduced by 28.6%, and the average total idle driving distance is reduced by 10.05%, the weighted-sum cost of the two components is reduced by 10.98% compared with non-robust dispatch solutions.

Comparing these methods, we know the average cost under true demand should be no greater than the optimal cost of problem (5.13) but not the guarantee for a single worst-case example. Robust dispatch solutions with the box type of uncertainty set and the SOC type of uncertainty set provide a desired level of probabilistic guarantee — the probability that an actual dispatch cost under the true demand vector being smaller than the optimal cost of the robust dispatch solutions is greater than $(1-\epsilon)$. However, they do not directly minimize the average performance of the solutions and we need to tune the value of ϵ and test the average cost. The horizontal lines show the average cost of distributionally robust solutions and non-robust solutions, since these costs are irrelevant to ϵ. The average cost of solutions of (5.13) is always smaller than costs of robust dispatch solutions

based on the box type uncertainty set. It indicates that the second order moment information of the random variable should be included for modeling the uncertainty of the demand model and calculating dispatch solutions.

Either the distributionally robust dispatch framework (5.3) or the SOC robust dispatch framework designed in [57] has its advantage, and does not provide full information about both the average performance and the out-of-sample or worst-case performance together by only solving an optimization problem. In practice, we choose a method according to the type of guarantee (average performance or worst-case scenario) we want to provide.

CHAPTER 6 : Conclusion and Future Work

In this chapter we outline the contributions of this dissertation and present potential future work.

6.1. Thesis Summary and Contributions

The objective of the dissertation has been to investigate the problem of data-driven dynamic resource allocation under demand uncertainties. We have focused on two domains, the receding horizon control framework that incorporates both historical and real-time sensing data to control decisions, and the robust/distributionally robust resource allocation models with uncertain demand sets constructed from data. Furthermore, our goal has been to balance supply according to the demand at different regions (nodes) of a network system in order to increase service efficiency. Applications in taxi dispatch system based on real-world data has shown that the approaches designed in this dissertation can improve performance of the taxi system by reducing total idle distance and increasing service fairness level.

The specific contributions of this dissertation are the following:

A Receding Horizon Control Framework for Real-Time Taxi Dispatch

With the development of data sensing, storage and processing technologies, the service efficiency of modern transportation systems can be increased by utilizing the model information provided by data to make resource allocation decisions. However, existing approaches and platforms usually apply greedy algorithms and transportation service such as taxis are far from optimal. Hence, we propose an RHC framework for the taxi dispatch problem. This method utilizes both historical and real-time GPS and occupancy data to build demand models, and applies predicted models and sensing data to decide dispatch locations for vacant taxis considering both current and anticipated future demand and service costs. From a system-level perspective, we compute suboptimal dispatch solutions for reaching a globally balanced supply demand ratio with least associated cruising distance under practical constraints. Demand model uncertainties under disruptive events are considered in the decision making process via robust dispatch problem formulations.

By applying the RHC framework on a data set containing taxi operational records in San Francisco, we show how to regulate parameters such as objective weight, idle distance threshold, and prediction horizon in the framework design process according to experiments. Evaluation results based on a SF dataset support system level performance improvements of our RHC framework, that the total idle driving distance is reduced by 52% compared with the original historical record (without any dispatch algorithm).

Data-Driven Robust Taxi Dispatch under Demand Uncertainties

Large amounts of sensing data provide opportunities to better regulate resource supply to meet the demand. However, We develop a multi-stage 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 trade-off 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.

A General Form of Data-Driven Distributionally Robust Resource Allocation

The robust resource allocation framework provides a probabilistic guarantee for system's performance under the worst-case scenario. However, the robust solutions do not provide a value for the average cost before we test the performance empirically. Motivated by the problem of minimizing the worst-case expected cost of taxi dispatch under demand uncertainties, we design a data-driven

distributionally robust resource allocation model. Then we design an efficient algorithm to construct an uncertain distribution set given a spatial-temporal historical demand dataset, by applying theories in hypothesis testing literature. The resource allocation problem we consider is concave in the random demand variable, convex in the decision variables and has decision variables on the denominator. We prove that an equivalent computationally tractable form exists based on strong duality and theories in distributionally robust optimization literature.

Evaluations show that by solving the computationally tractable form of distributionally robust dispatch problem, the average demand-supply ratio mismatch error is reduced by 28.6%, and the average total idle driving distance is reduced by 10.05%, compared with non-robust dispatch solutions. In the future, we will design different resource allocation strategies in transportation systems considering other objectives and constraints.

6.2. Future Work

A Data-Driven Dynamic Hierarchical Resource Allocation Framework for Efficient Mobility

The problem of optimizing on-demand mobility services can be viewed as a resource allocation problem, where the available resources are the empty vehicles under dispatch. In Chapter 3 we have presented a receding horizon control framework for proactive planning of vehicle dispatch based on robust optimal control theory. A multi-level algorithm that solves a centralized optimization problem in a higher level first and runs heuristic algorithm in the lower level is introduced in Section 3.6. The simulation process of experiments in Chapter 3 also applies this multi-level idea. In Chapter 4, our framework explicitly takes account of model uncertainties, which are quantified from historical data via statistical methods, and the framework ensures that the resulting resource allocation is robust to those model uncertainties. Numerical experiments on a realistic data set of taxi operational records in New York City have shown that our framework significantly outperforms naive proactive planning that does not incorporate model uncertainties.

While our previous resource allocation framework focuses on high-level planning of the distribution of vehicles, the framework does not address how each vehicle should be routed from an optimal

control perspective. Also, the framework requires a centralized authority to collect all available information and make decisions for every vehicle in real time, and such an approach may not scale to transportation networks with a large number of areas. To address theses open issues, a promising approach is a hierarchical resource allocation framework that consists of both high-level planning and low-level distributed control of the vehicles, with the strategic goal being to further bridge the gaps between our previous proactive planning framework and practical implementation. Two desired features of the formal mathematical model of the hierarchical framework include accommodation to multi-modal transportation and scalability to large networks.

Since on-demand mobility is not the exclusive mode of transportation in cities, the framework needs to take account of other co-existing modes to resolve any potential conflicts of road utilization (for instance,buses and private cars) and ensures that on-demand mobility stays minimally disruptive. The high-level planner not only needs to provide target areas for on-demand vehicles but also appropriate routing suggestions considering both mobility demand and the operation of other modes of transportation.

For a large transportation network, the amount of data collected by each vehicle can be prohibitive to transmit to the central planner in real time. Each low-level local controller needs to intelligently determine what information should be communicated with the central planner based on collected sensor information as well as limitation of the communication network. The lower-level controllers for individual vehicles should be designed to handle local information such as road conditions and communicate with the centralized high-level planner.

Design incentive mechanisms for real-time ridesharing and desirable social behavior under traffic congestion

Future cities will be highly instrumented with sensors and devices providing an almost real-time update of its various states, including traffic congestion and availability of resources. While the taxicabs may follow the dispatch commands from their companies, other ridesharing services such as Uber are operated under a different business model so that directly sending dispatch commands

becomes impractical. In those cases, a common solution is to offer monetary incentives (such as surge pricing currently implemented by Uber) so that drivers may be willing to relocate to the areas with higher demand. The current implementation of monetary incentives suffers from the fact that it only reacts to current demand and supply. As a result, the implementation often exhibits high volatility (e.g., Uber's surge pricing can often change rapidly within a few minutes) and may fail to achieve the desired re-balancing of demand and supply.

Hence, a better potential approach is a real-time ridesharing framework considering motivation strategies to motivate drivers and passengers follow the suggestions of ridesharing pairs designed by the control system. The scheme is proactive to future demand and can achieve similar performance to direct dispatch, reduce congestion and energy consumption by motivating resource sharing especially under the case that people are not willing to execute system-level optimal strategies due to short-term conflicting with personal interest.

APPENDIX

A.1. Appendix

A.1.1. Proof of Theorem 2

Proof 9 *For any fixed X, the maximum part of the objective function is equivalent to*

$$\begin{aligned} \max_{r\in\Delta} \quad & J_D(X) + \beta J_E(X,r) = J_D(X) + c^T(X)r \\ [c(X)]_i =& \beta \frac{1}{(\mathbf{1}_n^T X_{\cdot i} - X_{i\cdot}\mathbf{1}_n + L_i)^\alpha}, \quad J_D(X) = \sum_i \sum_j X_{ij} W_{ij}. \end{aligned} \tag{A.1}$$

The Lagrangian of problem (A.1) *with the Lagrangian multipliers $\lambda \geq 0, v \geq 0$ is*

$$\mathcal{L}(X, r, \lambda, v) = J_D(X) + b^T\lambda - (A^T\lambda - c(X) - v)^T r,$$

where $(A^T\lambda - c(X) - v)^T r$ is a linear function of r, and the upper bound exists only when

$$A^T\lambda - c(X) - v = 0.$$

The objective function of the dual problem is

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

With $v \geq 0$, the constraint $A^T\lambda - c(X) - v = 0$ is equivalent to $A^T\lambda - c(X) \geq 0$. Strong duality holds for problem (A.1) *since it satisfies the Slater's condition—the primal problem is convex and*

$c^T(X)r$ *is affine of* r. *The dual problem of* (A.1) *is*

$$\begin{aligned} \underset{\lambda \geq 0}{\textit{minimize}} \quad & J_D(X) + b^T\lambda \\ \textit{subject to} \quad & A^T\lambda - c(X) \geq 0. \end{aligned} \tag{A.2}$$

Hence, problem (4.11) *with* $\tau = 1$ *can be solved as the convex optimization problem defined in* (4.20).

A.1.2. Proof of Lemma 1

Proof 10 *Now consider the minimax problem over stage* $k+1$ *and* k, $1 \leqslant k \leqslant \tau - 1$ *of problem* (4.11)

$$\begin{aligned} \underset{r^k \in \Delta_k}{\textit{max}} \ \underset{X^{k+1}, L^{k+1}}{\textit{min}} \quad & J = \sum_{k=1}^{\tau} (J_D(X^k) + \beta J_E(X^k, r^k)) \\ \textit{s.t.} \quad & \textit{constraints of } (4.11). \end{aligned} \tag{A.3}$$

The domain of problem (A.3) *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 [10], 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 of* (A.3) *is nonempty. It means there exists an optimal minimax solution that is also optimal for the maximin problem, and we can exchange the order of max and min without changing such an optimal solution, i.e.,*

$$\underset{r^k \in \Delta_k}{\textit{max}} \ \underset{X^{k+1}, L^{k+1}}{\textit{min}} \ J = \underset{X^{k+1}, L^{k+1}}{\textit{min}} \ \underset{r^k \in \Delta_k}{\textit{max}} \ J.$$

A.1.3. Proof of Lemma 2 and Theorem 3

Proof of Lemma 2

Proof 11 *With the polytope form of uncertainty set* (4.22)*, the domain of each* r^k *is closed and convex, i.e., is compact, and Lemma 1 holds. Considering the maximizing part of problem* (4.21)

$$\max_{r^k \in \Delta_k} \quad J, \quad \textit{s.t.} \quad \textit{constraints of } (4.11), \tag{A.4}$$

the Lagrangian of (A.4) *with multipliers* $\lambda^k \geq 0, v^k \geq 0$ *is*

$$\begin{aligned} &\mathcal{L}(X^k, r^k, \lambda^k, v^k) \\ =&\sum_{k=1}^{\tau}(J_D(X^k) + b_k^T\lambda^k - (A_k^T\lambda^k - c(X^k) - v^k)^T r^k), \end{aligned} \tag{A.5}$$

Hence, based on the proof of Theorem 2, we take partial derivative of the Lagrangian (A.5) *for every* $r^k \in \Delta_k$*. The inequality constraint of* $r^k \in \Delta_k$ *defined as* (4.22) *is affine of* r^k*,* $c^T(X^k)r^k$ *is affine of* r^k*, and problem* (A.4) *is convex. Hence Slater's condition is satisfied and strong duality holds for problem* (A.4)*. An equivalent form of* (4.11) *under uncertainty set* (4.22) *is defined as* (5.13)*.*

Proof of Theorem 3

Proof 12 *With uncertain set defined as* (4.24)*, the domain of each* r^k *is compact and Lemma 1 holds. We consider the equivalent problem* (4.21) *of problem* (4.11)*, and first derive the Lagrangian of the maximum part of the objective function* (4.21) *with constraint* $\lambda \geq 0, v_k \geq 0$

$$\begin{aligned} &\mathcal{L}(X^k, r^k, \lambda, v_k) \\ =&b^T\lambda - \sum_{k=1}^{\tau}((A_k^T\lambda - c(X^k) - v_k)^T r^k - J_D(X^k)), \end{aligned} \tag{A.6}$$

Similarly as the proof of Theorem 2, we take the partial derivative of (A.6) *over each* r^k*, the objective function of the dual problem is*

$$g(X^k, L^k, \lambda, r^k) = \sup_{r^k \in \Delta_k} \mathcal{L}(X^k, r^k, \lambda, v_k)$$
$$= \begin{cases} \infty \quad \textit{if } \exists k \textit{ s.t. } A_k^T \lambda - c(X^k) - v_k \neq 0, \\ \sum_{k=1}^{\tau} J_D(X^k) + b^T \lambda \quad \textit{o.w.} \end{cases}$$

Since Slater's condition is satisfied and strong duality holds, problem (4.25) *is a equivalent to the computationally tractable convex optimization form* (4.11) *under uncertain set* (4.24)*.*

A.1.4. Proof of Theorem 4

Proof 13 *Under the definition of uncertainty set* (4.19) *for concatenated* r^k*, the domain of each* r^k *is compact, and problem* (4.11) *is equivalent to* (4.21)*. 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* (4.19)*, problem* (4.21) *is equivalent to*

$$\begin{aligned} \underset{X^k, L^k}{\textit{min.}} \ \underset{r_c \geqslant 0}{\textit{max}} \quad & \left(c_l^T(X) r_c + \sum_{k=1}^{\tau} \sum_i \sum_j X_{ij}^k W_{ij} \right) \\ \textit{subject to} \quad & r_c = \hat{r}_c + y + C^T w, \\ & \|y\|_2 \leqslant \Gamma_1^B, \|w\|_2 \leqslant \sqrt{\frac{1}{\epsilon} - 1}, \\ & \textit{constraints of } (4.11). \end{aligned} \tag{A.7}$$

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

$$\begin{aligned} \underset{r_c \geqslant 0}{max} \quad & c_l^T(X) r_c \\ \textit{subject to} \quad & r_c = \hat{r}_c + y + C^T w, \\ & \|y\|_2 \leqslant \Gamma_1^B, \|w\|_2 \leqslant \sqrt{\frac{1}{\epsilon} - 1}. \end{aligned} \tag{A.8}$$

The constraints of problem (A.8) *has a feasible solution* $r_c = \hat{r}_c$ *such that* $\|y\|_2 < \Gamma_1^B, \|w\|_2 < \sqrt{\frac{1}{\epsilon} - 1}$*, 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 (A.8)*, we start from the following Lagrangian with* $v \geqslant 0$

$$\mathcal{L}(X, r_c, z, v) = c_l^T(X) r_c + z^T(\hat{r}_c + y + C^T w - r_c) + v^T r_c.$$

By taking the partial derivative of the above Lagrangian over r_c*, we get the supreme value of the Lagrangian as*

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

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

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

Hence, the objective function of the dual problem for (A.8) *is*

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

Together with the objective function $J_D(X^k)$ *and other constraints that do not directly involve* r_c*, an equivalent convex form of* (4.11) *given the uncertainty set* (4.19) *is shown as* (4.26)*.*

Bibliography

[1] N. Agatz, A. Erera, M. Savelsbergh, and X. Wang. Optimization for dynamic ride-sharing: A review. *European Journal of Operational Research*, 223(2):295 – 303, 2012.

[2] S. Ali, A. Maciejewski, H. Siegel, and J.-K. Kim. Measuring the robustness of a resource allocation. *IEEE Transactions on Parallel and Distributed Systems*, 15(7):630–641, July 2004.

[3] I. Amundson and X. D. Koutsoukos. A survey on localization for mobile wireless sensor networks. In *Mobile Entity Localization and Tracking in GPS-less Environnments*, pages 235–254. Springer Berlin Heidelberg, 2009.

[4] B. An, V. Lesser, D. Irwin, and M. Zink. Automated negotiation with decommitment for dynamic resource allocation in cloud computing. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1*, AAMAS '10, pages 981–988, Richland, SC, 2010. International Foundation for Autonomous Agents and Multiagent Systems.

[5] M. Asif, J. Dauwels, C. Goh, A. Oran, E. Fathi, M. Xu, M. Dhanya, N. Mitrovic, and P. Jaillet. Spatiotemporal patterns in large-scale traffic speed prediction. *IEEE Transactions on Intelligent Transportation Systems*, 15(2):797–804, 2014.

[6] R. K. Balan, K. X. Nguyen, and L. Jiang. Real-time trip information service for a large taxi fleet. In *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services*, ACM MobiSys '11, pages 99–112, 2011.

[7] A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton University Press, 2009.

[8] A. Ben-Tal and A. Nemirovski. Robust convex optimization. *Mathematics of Operations Research*, 23(4):769–805, 1998.

[9] M. E. Berge and C. A. Hopperstad. Demand driven dispatch: A method for dynamic aircraft capacity assignment, models and algorithms. *Operations Research*, 41(1):153–168, 1993.

[10] D. Bertsekas, A. Nedi, A. Ozdaglar, et al. *Convex Analysis and Optimization*. Athena Scientific, 2003.

[11] D. Bertsimas and D. B. Brown. Constructing uncertainty sets for robust linear optimization. *Oper. Res.*, 57(6):1483–1495, 2009.

[12] D. Bertsimas, V. Bupta, and N. Kallus. Data-driven robust optimization. *Operations Research*, (arXiv: 1401.0212), 2015.

[13] D. Bertsimas, D. Iancu, and P. Parrilo. A hierarchy of near-optimal policies for multistage adaptive optimization. *IEEE Transactions on Automatic Control*, 56(12):2809–2824, Dec 2011.

[14] B. S. Blanchard, W. J. Fabrycky, and W. J. Fabrycky. *Systems engineering and analysis*, volume 4. Prentice Hall New Jersey;, 1990.

[15] S. Blandin, D. Work, P. Goatin, B. Piccoli, and A. Bayen. A general phase transition model for vehicular traffic. *SIAM Journal on Applied Mathematics*, 71(1):107–127, 2011.

[16] B. Bowerman, J. Braverman, J. Taylor, H. Todosow, and U. Von Wimmersperg. The vision of a smart city. In *2nd International Life Extension Technology Workshop, Paris*, volume 28, 2000.

[17] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[18] E. Bradley. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1-26), 1979.

[19] R. E. Brown. Impact of smart grid on distribution system design. In *Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century, 2008 IEEE*, pages 1–4, July 2008.

[20] G. C. Calafiore and L. E. Ghaoui. On distributionally robust chance-constrained linear programs. *Journal of Optimization Theory and Applications*, 130(1):1–22, 2006.

[21] G. Chasparis, M. Maggio, E. Bini, and K.-E. Årzén. Design and implementation of distributed resource management for time sensitive applications. *Automatica*, 2015. Accepted for publication.

[22] D. R. Choffnes and F. E. Bustamante. An integrated mobility and traffic model for vehicular wireless networks. In *Proceedings of the 2Nd ACM International Workshop on Vehicular Ad Hoc Networks*, pages 69–78, 2005.

[23] G. Como, K. Savla, D. Acemoglu, M. A. Dahleh, and E. Frazzoli. Robust distributed routing in dynamical networks–part i: Locally responsive policies and weak resilience. *IEEE Transactions on Automatic Control*, 58(2):317–332, Feb 2013.

[24] G. Como, K. Savla, D. Acemoglu, M. A. Dahleh, and E. Frazzoli. Robust distributed routing in dynamical networks–part ii: Strong resilience, equilibrium selection and cascaded failures. *IEEE Transactions on Automatic Control*, 58(2):333–348, Feb 2013.

[25] J. Cortes, S. Martinez, T. Karatas, and F. Bullo. Coverage control for mobile sensing networks. *IEEE Transactions on Robotics and Automation*, 20(2):243–255, 2004.

[26] F. A. Cuzzola, J. C. Geromel, and M. Morari. An improved approach for constrained robust model predictive control. *Automatica*, 38(7):1183–1189, 2002.

[27] H. David and H. Nagaraja. *Order statistics*. Wiley Online Library, 1970.

[28] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

[29] B. Donovan and D. B. Work. Using coarse gps data to quantify city-scale transportation system resilience to extreme events. In *presented at the 2015 Transportation Research Board Annual Meeting*, July 2015.

[30] H.-G. Eichler, S. X. Kong, W. C. Gerth, P. Mavros, and B. Jnsson. Use of cost-effectiveness analysis in health-care resource allocation decision-making: How are cost-effectiveness thresholds expected to emerge? *Value in Health*, 7(5):518–528, 2004.

[31] A. Eryilmaz and R. Srikant. Fair resource allocation in wireless networks using queue-length-based scheduling and congestion control. In *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, volume 3, pages 1794–1803, 2005.

[32] F. Fiedrich, F. Gehbauer, and U. Rickers. Optimized resource allocation for emergency response after earthquake disasters. *Safety Science*, 35(13):41 – 57, 2000.

[33] M. D. Galus, R. A. Waraich, F. Noembrini, K. Steurs, G. Georges, K. Boulouchos, K. W. Axhausen, and G. Andersson. Integrating power systems, transport systems and vehicle technology for electric mobility impact assessment and efficient control. *IEEE Transactions on Smart Grid*, 3(2):934–949, June 2012.

[34] R. Ganti, M. Srivatsa, and T. Abdelzaher. On limits of travel time predictions: Insights from a new york city case study. In *IEEE 34th ICDCS*, pages 166–175, June 2014.

[35] Y. Geng and C. Cassandras. New "smart parking" system based on resource allocation and reservations. *IEEE Transactions on Intelligent Transportation Systems*, 14(3):1129–1139, 2014.

[36] J. Goh and M. Sim. Distributionally robust optimization and its tractable approximations. *Operations Research*, 58(4-part-1):902–917, 2010.

[37] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, 2014.

[38] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami. Internet of things (iot): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, 29(7):1645 – 1660, 2013.

[39] Hara Associates Inc. and Corey, Canapary& Galanis. Taxi user surveys. `http://www.sfmta.com/sites/default/files/Draft%20SF%20UserSurvey%2055%20WEB%20version04042013.pdf`, 2013.

[40] M. Hasan, E. Hossain, and D. Niyato. Random access for machine-to-machine communication in lte-advanced networks: issues and approaches. *IEEE Communications Magazine*, 51(6):86–93, June 2013.

[41] J. Herrera, D. Work, R. Herring, X. Ban, Q. Jacobson, and A. Bayen. Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. *Transportation Research Part C*, 18(4):568–583, 2010.

[42] A. Howard, M. Matarić, and G. Sukhatme. Mobile sensor network deployment using potential fields: A distributed, scalable solution to the area coverage problem. In *Distributed Autonomous Robotic Systems 5*. Springer Japan, 2002.

[43] Y. Huang and J. W. Powell. Detecting regions of disequilibrium in taxi services under uncertainty. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, number 10, pages 139–148, 2012.

[44] Z. Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3):283–304, Sept. 1998.

[45] C. Joe-Wong, S. Sen, T. Lan, and M. Chiang. Multi-resource allocation: Fairness-efficiency tradeoffs in a unifying framework. In *INFOCOM, 2012 Proceedings IEEE*, pages 1206–1214, March 2012.

[46] K.-D. Kim. Collision free autonomous ground traffic: A model predictive control approach. In *Proceedings of the ACM/IEEE 4th International Conference on Cyber-Physical Systems*, number 10, pages 51–60, 2013.

[47] D.-H. Lee, R. Cheu, and S. Teo. Taxi dispatch system based on current demands and real-time traffic conditions. *Transportation Research Record:Journal of the Transportation Research Board*, 8(1882):193–200, 2004.

[48] E. Lehmann and J. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics, 2010.

[49] B. Li, D. Zhang, L. Sun, C. Chen, S. Li, G. Qi, and Q. Yang. Hunting or waiting? discovering passenger-finding strategies from a large-scale real-world taxi dataset. In *IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pages 63–38, 2011.

[50] W. Li and C. Cassandras. A cooperative receding horizon controller for multivehicle uncertain environments. *IEEE Transactions on Automatic Control*, 51(2):242–257, 2006.

[51] Z. Liao. Real-time taxi dispatching using global positioning systems. *Commun. ACM*, 46(5):81–83, May 2003.

[52] A. Lorca and A. Sun. Adaptive robust optimization with dynamic uncertainty sets for multi-period economic dispatch under significant wind. In *Power Energy Society General Meeting, 2015 IEEE*, pages 1–1, July 2015.

[53] J. Medhi. *Stochastic models in queueing theory*. Academic Press, 2002.

[54] F. Miao, S. Han, S. Lin, and G. J. Pappas. Robust taxi dispatch under model uncertainties. In *54th IEEE Conference on Decision and Control (CDC)*, pages 2816–2821, Dec 2015.

[55] F. Miao, S. Han, S. Lin, and G. J. Pappas. Robust taxi dispatch under model uncertainties. In *54th IEEE Conference on Decision and Control (CDC)*, pages 2816–2821, Dec 2015.

[56] F. Miao, S. Han, S. Lin, J. A. Stankovic, D. Zhang, S. Munir, H. Huang, T. He, and G. J. Pappas. Taxi dispatch with real-time sensing data in metropolitan areas: A receding horizon control approach. *IEEE Transactions on Automation Science and Engineering*, 13(2):463–478, April 2016.

[57] F. Miao, S. Han, S. Lin, Q. Wang, J. Stankovic, A. Hendawi, D. Zhang, T. He, and G. J. Pappas. Data-driven robust taxi dispatch under demand uncertainties. *submitted, preprint can be found at http://www.seas.upenn.edu/ miaofei/taxi-journal.pdf*.

[58] F. Miao, S. Han, and G. J. Pappas. Data-driven distributionally robust taxi dispatch under

model uncertainties. In *55th IEEE Conference on Decision and Control (CDC)*, pages 2816–2821, Dec 2015.

[59] F. Miao, S. Lin, S. Munir, J. A. Stankovic, H. Huang, D. Zhang, T. He, and P. G. J. Taxi dispatch with real-time sensing data in metropolitan areas — a receding horizon control approach. In *6th International Conference of Cyber-Physical Systems*, 2015.

[60] E. S. Mills. An aggregative model of resource allocation in a metropolitan area. *The American Economic Review*, 57(2):197–210, 1967.

[61] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas. Predicting taxi-passenger demand using streaming data. *IEEE Transactions on Intelligent Transportation Systems*, 14(3):1393–1402, Sept 2013.

[62] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas. Predicting taxi x2013;passenger demand using streaming data. *IEEE Transactions on Intelligent Transportation Systems*, 14(3):1393–1402, Sept 2013.

[63] M. Naphade, G. Banavar, C. Harrison, J. Paraszczak, and R. Morris. Smarter cities and their innovation challenges. *Computer*, 44(6):32–39, June 2011.

[64] R. Negenborn, B. D. Schutter, and J. Hellendoorn. Multi-agent model predictive control for transportation networks: Serial versus parallel schemes. *Engineering Applications of Artificial Intelligence*, 21(3):353 – 366, 2008.

[65] New York City Taxi and Limousine Commission. Taxi of tomorrow survey results. `http://www.nyc.gov/html/tlc/downloads/pdf/tot_survey_results_02_10_11.pdf`, 2011.

[66] D. Niu, H. Xu, B. Li, and S. Zhao. Quality-assured cloud bandwidth auto-scaling for video-on-demand applications. In *INFOCOM, 2012 Proceedings IEEE*, pages 460–468, March 2012.

[67] T. P. D. of the Department of Economic and S. A. of the United Nations. 2014 revision of world

urbanization prospects. http://esa.un.org/unpd/wup/Publications/Files/WUP2014-PressRelease.pdf, 2014.

[68] A. Pantoja and N. Quijano. A population dynamics approach for the dispatch of distributed generators. *IEEE Transactions on Industrial Electronics*, 58(10):4559–4567, Oct 2011.

[69] M. Pavone, S. L. Smith, E. Frazzoli, and D. Rus. Robotic load balancing for mobility-on-demand systems. *Int. J. Rob. Res.*, 31(7):839–854, June 2012.

[70] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos. Sensing as a service model for smart cities supported by internet of things. *Transactions on Emerging Telecommunications Technologies*, 25(1):81–93, 2014.

[71] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos. Sensing as a service model for smart cities supported by internet of things. *Transactions on Emerging Telecommunications Technologies*, 25(1):81–93, 2014.

[72] C. Petres, Y. Pailhas, P. Patron, Y. Petillot, J. Evans, and D. Lane. Path planning for autonomous underwater vehicles. *IEEE Transactions on Robotics*, 23(2):331–341, 2007.

[73] S. Phithakkitnukoon, M. Veloso, C. Bento, A. Biderman, and C. Ratti. Taxi-aware map: Identifying and predicting vacant taxis in the city. In *the First International Joint Conference on Ambient Intelligence*, pages 86–95, 2010.

[74] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser. A parsimonious model of mobile partitioned networks with clustering. In *First International Communication Systems and Networks and Workshops (COMSNETS)*, pages 1–10, 2009.

[75] J. W. Powell, Y. Huang, F. Bastani, and M. Ji. Towards reducing taxicab cruising time using spatio-temporal profitability maps. In *Proceedings of the 12th International Conference on Advances in Spatial and Temporal Databases*, number 19, pages 242–260, 2011.

[76] M. Qu, H. Zhu, J. Liu, G. Liu, and H. Xiong. A cost-effective recommender system for taxi

drivers. In *20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 45–54, 2014.

[77] G. C. Reinsel. *Elements of Multivariate Time Series Analysis*. Springer-Verlag New York, New York, NY, USA, 1997.

[78] K.-T. Seow, N. H. Dang, and D.-H. Lee. A collaborative multiagent taxi-dispatch system. *IEEE Transactions on Automation Science and Engineering*, 7(3):607–616, 2010.

[79] J. Shawe-Taylor and N. Cristianini. Estimating the moments of a random estimating the moments of a random vector with applications. In *GRETSI Conference*, pages 47–52, 2003.

[80] O. W. Shuo Ma, Yu Zheng. T-share: A large-scale dynamic taxi ridesharing service. ICDE 2013, April 2013.

[81] K. Spieser, T. Kyle Ballantyne, Z. Rick, E. Frazzoli, D. Morton, and M. Pavone. Toward a systematic approach to the design and evaluation of automated mobility-on-demand systems a case study in singapore. In *the IEEE/RSJ International Conference on Intelligent Robots and Systems*.

[82] H. Terelius and K. H. Johansson. An efficiency measure for road transportation networks with application to two case studies. In *CDC*, pages 5149–5155, 2015.

[83] C. Tomlin, G. Pappas, and S. Sastry. Conflict resolution for air traffic management: a study in multiagent hybrid systems. *IEEE Transactions on Automatic Control*, 43(4):509–521, 1998.

[84] S. K. Verma and H. T. Vo. A predictive taxi dispatching system for improved user satisfaction and taxi utilization. In *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, pages 175–182. IEEE, 2015.

[85] K. I. Wong and M. G. H. Bell. The optimal dispatching of taxis under congestion: A rolling horizon approach. *Journal of Advanced Transportation*, 40:203–220, 2006.

[86] f. Yang, S. Wong, and K. Wong. Demand supply equilibrium of taxi services in a network under competition and regulation. *Transportation Research Part B: Methodological*, 36:799–819, 2002.

[87] J. Yang, P. Jaillet, and H. Mahmassani. Real-time multivehicle truckload pickup and delivery problems. *Transportation Science*, 38(2):135–148, 2004.

[88] X. Yin, A. Jindal, V. Sekar, and B. Sinopoli. A control-theoretic approach for dynamic adaptive video streaming over http. *SIGCOMM Comput. Commun. Rev.*, 45(4):325–338, Aug. 2015.

[89] S. Zaman and D. Grosu. Combinatorial auction-based allocation of virtual machine instances in clouds. *Journal of Parallel and Distributed Computing*, 73(4):495 – 508, 2013.

[90] D. Zhang, T. He, S. Lin, S. Munir, and J. Stankovic. Online cruising mile reduction in large-scale taxicab networks. *IEEE Transactions on Parallel and Distributed Systems*, 26(11):3122–3135, Nov 2015.

[91] D. Zhang, T. He, S. Lin, S. Munir, and J. A. Stankovic. Dmodel: Online taxicab demand model from big sensor data in. In *IEEE International Congress on Big Data (BigData Congress)*, pages 152–159, 2014.

[92] R. Zhang and M. Pavone. Control of robotic mobility-on-demand systems: a queueing-theoretical perspective. In *Proceedings of Robotics: Science and Systems*, July 2014.

[93] R. Zhang, F. Rossi, and M. Pavone. Model predictive control of autonomous mobility-on-demand systems. *arXiv preprint arXiv:1509.03985*, 2015.