Data-driven robust optimization for the itinerary planning via large-scale GPS data

Lei Wu, Mhand Hifi

Version of Record: https://www.sciencedirect.com/science/article/pii/S0950705121006997
Manuscript_0f13199200ad2df32a2e1820d791d434

Data-Driven Robust Optimization for the Itinerary Planning via Large-Scale GPS Data

Lei Wu[a,b], Mhand Hifi[c,*]

[a] *School of Public Finance and Taxation, Zhongnan University of Economics and Law, China.*
[b] *Innovation and Talent Base for Income Distribution and Public Finance, Zhongnan University of Economics and Law, China.*
[c] *EPROAD EA 4669, Université de Picardue Jules Verne, Amiens, France.*

Abstract

In this paper, we propose a data-driven robust optimization for establishing reliable itineraries through the use of GPS trajectories. The goal of the study is to provide a robust solution that is able to maximize the probability of achieving the expected travel time and minimize the delay. The designed framework can be viewed as an incremental approach, where data-driven robust optimization cooperates with a learning procedure such that both the uncertainty set and the objective function are incrementally adjusted according to the current data analysis results. In fact, two types of training models are designed in order to adapt the robust optimization model through analyzing GPS-data. The first training model tries to generate the uncertainty set for establishing the model, and the second one establishes the best parameter-settings allowing to converge towards a robust solution. Finally, a data-based simulation framework is designed for analyzing the robustness of the proposed method, where achieved solutions are tested on a simulated traffic network by using real-world orders as the comparison targets.

Keywords: Data-driven; Learning; Optimization; Robustness; Uncertainty.

1. Introduction

Decision-making with uncertainty arises in several real-world applications, like management science, logistics and finance. Among the classic approaches for decision-making

*Corresponding author
Email addresses: wu.lei@zuel.edu.cn (Lei Wu), hifi@u-picardie.fr (Mhand Hifi)

under uncertainty, stochastic programming is widely recognized as an effective model tool to provide reliable solutions. The aforementioned performance depends also on the probabilistic characterization of randomness. However, on the one hand, the distribution's probability of some information is often unavailable, like the route's travel time related to the next day or period. On the other hand, the ambiguity-based criteria is not taken into account by stochastic models (cf., (Bertsimas et al., 2018b)) while the robust optimization under uncertainty deals with the ambiguity, especially for combinatorial optimization problems (the reader can refer to (Caserta & Voβ, 2019; Su et al., 2019) and to (Tirkolaee et al., 2020) for some standard studies on robust optimization for combinatorial optimization).

Over the past decade, based on mathematical modeling and information technology, data science has made substantial progress in solving existing management issues. Deep mining / learning can effectively assist managers to reduce uncertainty in decision-making. As the size of data grow in all real-world applications, the development of data-driven models become of great significance for the robust optimization (Bertsimas & Thiele, 2006; Bertsimas et al., 2011; Hanks et al., 2019). Most studies tackling the data-driven robust optimization focus on studying the "distributionally" robust optimization problems, where uncertainty is a probability distribution (Bertsimas et al., 2018a). The principal strategy is based upon generating the uncertainty set from the observed historical data through the use of statistical analysis (for more details, the reader can be referred to Bertsimas & Sim (2004); Bertsimas & Brown (2009); Ben-Tal et al. (2009); Delage & Ye (2010). More recently, Bansak et al. (2018) proposed a data-driven-based method for improving refugee integration problem. Unlike the robust distribution optimization, their method uses historical data to provide the best optimization value, allowing for more orderly resettlement of refugees. Inspired from classical researches, this paper attempts to design a *data-driven robust optimization model*, where both the uncertainty set and the objective function can be dynamically (incrementally) adjusted according to the current data analysis results.

Travel time uncertainty is unpredictable in an urban transportation system. Es-

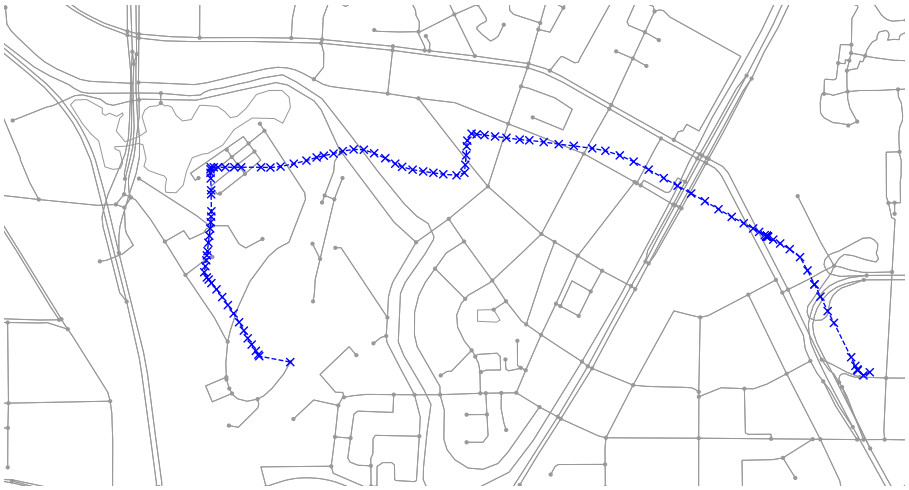


Figure 1: An instance illustrating a sequence of GPS' waypoints.

pecially in itinerary planning, travel time reliability has been considered an important criterion. GPS trajectory data analytics has offered new opportunities for understanding urban traffic networks (Xia et al., 2017). However, due to the limitations related to the satellite technology, the gap between GPS' waypoints and real-position locations is usually close to $10-70$ meters (Lee et al., 2016) (Figure 1 illustrates a sequence of GPS' waypoints collected when a taxi completes an order: GPS' waypoints fall within multiple road segments). One can observe that GPS-data contains an intractable noise and, the reliability of the probability distribution related to the travel time cannot be guaranteed. Thus, in order to make a reasonable use of data with high uncertainty, we propose a model that is able (i) to derive the uncertainty set related to the travel time and (ii) to deduce the robust decision preferences from the GPS-data. Indeed, a robust discrete optimization-based approach is introduced for characterizing (i) the decision preferences, and (ii) the travel time between two locations defined as a set of discrete scenarios (each scenario denotes a possible travel time for vehicles traveling from two different locations). The objective now is to determine a solution, where a predefined decision preference is

satisfied; that is, a decision preference satisfying a *robust criterion* (see, for instance, Aissi et al., 2009; Goli et al., 2019; Gabrel et al., 2014; Roy, 2010; Tirkolaee et al., 2018). Herein, the learning procedure combines (i) a data-mining procedure that achieves the uncertainty sets and (ii) a robust criterion used to systematically reacts with the decision preferences regarding the results induced from the current data analysis. The proposed incremental robust criterion is hereafter referred to as *win-loss robustness*, and its goal is to balance between both the "winning rate" (maximization problem) and the "expected loss" (minimization problem).

In this paper, a Data-driven Robust Optimization (DRO) model is proposed for providing reliable itineraries based on real-world taxi GPS trajectories of Chengdu City. The DRO framework is illustrated in Figure 2, where its objective is to enhance the robustness of the navigation system through the use of GPS-data, data-driven and robust optimization. First, the robust optimization model can be distinguished by its components: (i) an uncertainty set, (ii) a robust criterion and (ii) an optimization procedure. Second and last, DRO can be distinguished by two types of training models: the first type is applied to generate the traffic network related to the GPS-data, while the second one trains the robust optimization model used for providing the itineraries regarding the trained traffic network.

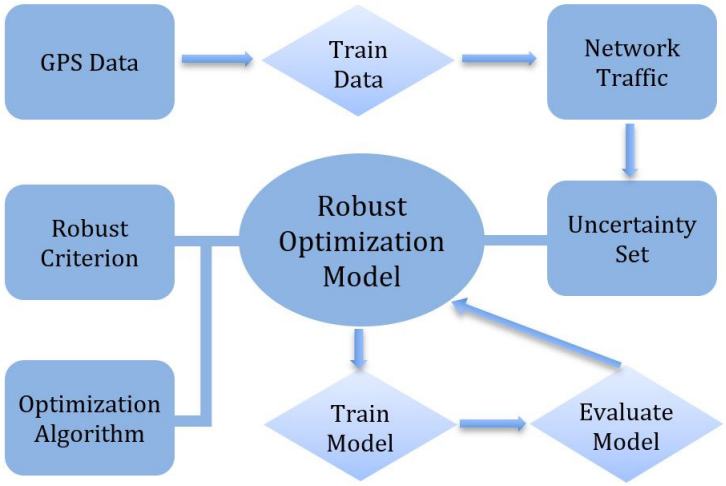


Figure 2: An overview of the data-driven robust optimization.

2. Related Works

The urban transport is a highly complex system due to the variety of the transportation modes, the large amount of traffic and unpredictable factors, such as the perturbations related to weather and breakdowns. In all cases, building a reliable traffic network similar to the original road-topology remains the first step when tackling the urban transportation problem. A real-world road-topology includes two main components: the structure of the network and the traffic information related to roads. The used real-world network application is based upon the OpenStreetMap (a common tool for analyzing road networks), where the travel time is preferred to the distance. Therefore, in this paper extracting valuable information from GPS-data for measuring travel time becomes the first problem to solve. Because of the noise in GPS-data, it is often impossible to match a GPS trajectory directly to road segments. In order to overcome this point, the Hidden Markov Model (HMM) was proposed in Hummel (2006) for the map matching, which was improved later (see, for instance, Newson & Krumm, 2009; Chen et al., 2014; Yang & Gidofalvi, 2017; Wannes & Verbeke, 2018). In this study, HMM-based procedure is used for estimating road traffic information, which generates the uncertainty set of the travel time.

Determining reliable itineraries under uncertainty is equivalent to solving the Stochastic Shortest-Path Problem (SSPP) or the Robust Shortest-Path Problem (RSPP). SSPP's objective is to find a most reasonable itinerary considering the predefined travel time reliability (see, for instance, Cheng et al., 2016; Chen et al., 2016; Shao et al., 2014; Wu, 2015; Zhang et al., 2016, 2017, 2018) (recent surveys on the stochastic version of reliable itineraries can be found in Zhang et al. (2017) and Zhang et al. (2018)). Due to the uncertain traffic network, the probability distribution related to the travel time (using a stochastic optimization) cannot be evaluated to optimality. Unlike SSPP, RSPP was proposed to optimize itineraries under predefined robust criteria, especially with known partial distribution (Yu & Yang, 1998). Most of the existing studies related to RSPP are based on two phases: the construction phase providing the uncertainty sets, and the enumeration phase determining all the potential scenarios, which should be used by the

solution (more recent reviews on RSPP can be found in Goerigk & Schöbel (2016) and Kasperski & Zieliński (2016)).

In the literature, for either SSPP or RSPP, most results are obtained through the use of random graphs or random instances based on real transportation networks (Stabler et al., 2016). However, building an uncertainty set from real-world observations plays a vital role in the robust optimization problem (Bertsimas et al., 2018a). Based on live traffic data from the city of Chicago, Chassein et al. (2019) conducted a comparative study of six mainstream perspectives to model uncertainty sets for the worst-case criterion. The data considered consists of 4363 available observations, where each data point contains the traffic speed for 1257 road segments.

In addition to studying the mechanism used to build uncertainty sets for RSPP, another branch of robust optimization theory focuses on developing robust criteria based on decision preferences, such as the worst-case criterion (Yu & Yang, 1998), the min-max deviation criterion (Kouvelis & Yu, 1997) and the percentile robust criterion (Xing & Zhou, 2013). Inspired by Roy (2010), Gabrel et al. (2013) proposed a robust model for RSPP based on the expectations of the decision-maker: *the bw-robust criterion*. The paper has shown that, by adjusting the key parameters of the model, the bw-robust criterion was able to represent different decision preferences, such as risk seeking (best-case criterion), risk aversion (worst-case criterion) and min-max regret (min-max deviation criterion). Therefore, this work aims to propose a flexible robust criterion that can be driven by data.

The main contributions of the paper can be summarized as follows:

- It proposes a tractable and scalable data-driven robust optimization framework, which can drive both (i) the probability attributes from a data perspective and (ii) the risk preference from a decision perspective.
- It establishes a framework for using large-scale GPS-data that provides an uncertainty set for travel time along a real-world traffic network.
- It designs a robust criterion whose objectives are to maximize the "*winning rate*"

and to minimize the "*expected loss*", where the decision-maker can adjust the weight between both objective functions according to their own risk preferences.

- An experimental protocole for evaluating the performance of the robust data-driven approach: a simulation-based robustness analysis based on a real-world traffic network is presented. The considered GPS-data is divided into two sets −a training set and a test set−, where the solutions provided by DRO through the use of the training set will be tested on a "simulated traffic network" generated from the test set.

Network	
N	a set of nodes, $N = \{i \mid i = 1, \ldots, n\}$;
E	a set of edges (road segments), $E = \{e_l \mid l = 1, \ldots, m\}$ or $\{(i,j) \mid i \neq j \in N\}$;
o, d	an origin-destination pair (of nodes) in N;
p_{od}	a path, a set of edges connecting a pair of nodes (o, d) in G.
Robust optimization	
$\tilde{t}_{ij}$	a random variable describing the travel time of e_{ij};
$\hat{t}_{ij}$	a set of scenarios used to simulate $\tilde{t}_{ij}$;
$\hat{t}_{ij}^{k}$	the observed travel time of e_{ij} in the kth scenario of $\hat{t}_{ij}$;
$\hat{T}$	the uncertainty set used to simulate $\tilde{t}_{ij}$, $\forall\ (i,j) \in E$;
x_{ij}	a binary variable that indicates whether e_{ij} is selected in the path;
b	the best accepted value;
w	the worst accepted value.
Learning model	
$cord_t^u$	the GPS coordinates of the sample u at time t, $cord_t^u = \{latitude, longitude, time\}$;
$cord^u$	a sequence of GPS coordinates of the sample u, $cord^u = \{cord_t^u \mid t = 1, 2, \ldots\}$;
$\hat{t}_{(o,d)}^u$	the observed travel time of path p_{od} for the sample u;
S_{train}	the training set used to train the uncertainty set and the robust optimization model;
S_{test}	the test set used to analyze the performance of DRO;
G_{sim}	simulated traffic network used to analyze the performance of DRO;
OD_{real}	an origin-destination pair (of nodes) generated from real GPS-data.

Table 1: Notations used for the rest of the paper.

3. A New Robust Criterion for the RSPP

Let $G = (N, E)$ be a network, where N represents a set of n nodes and E denotes a set of m edges. Under the assumption that the travel time of each edge is uncertain, the objective of RSPP is to determine a most reliable path linking a predefined origin-destination (OD) of G (Table 1 reports the notations used for the rest of the paper).

3.1. The Worst-Case Criterion

To simulate the uncertainty related to the travel time, the data-driven model applies a discrete scenario-based optimization approach. Formally, for each edge e_{ij}, the travel time from i to j is described as a random variable $\tilde{t}_{ij}$ whose probability distribution is unknown. Let $\hat{t}_{ij}$ be a set of S scenarios $\hat{t}_{ij}^k,\ k = 1, \ldots, S$, where each scenario represents an available observation of the travel time $\tilde{t}_{ij}$. $\hat{t}_{ij}$ can be considered a set of samples used to approximately represent $\tilde{t}_{ij}$. Therefore, a feasible path p_{od} connecting a pair of nodes (o, d) in G is associated with S scenarios. We note that the kth scenario related to p_{od} depends only on $\hat{t}_{ij}^k$, for all edges $e_{ij} \in p_{od}$.

A typical strategy available in the literature opts to minimize the worst observation over all considered scenarios. The aforementioned strategy is usually referred as the worst-case criterion, which is often used to find an absolute robust solution. Formally, the standard linear program for RSPP (noted P$_w$), based on the worst-case criterion, can be written as follows:

$$(\mathrm{P}_w) \qquad \min\ r \tag{1}$$

$$\text{s.t.} \quad \sum_{(i,j)\in E} \hat{t}_{ij}^k x_{ij} \le r, \quad k = 1, \ldots, S,$$

$$\sum_{(i,j)\in E} x_{ij} - \sum_{(j,i)\in E} x_{ji} = \begin{cases} 1, & i = o, \\ 0, & i \in N\backslash\{o, d\}, \\ -1, & i = d, \end{cases} \tag{2}$$

$$r \ge 0,\ x_{ij} \in \{0, 1\}, \forall\ (i, j) \in E,$$

where the objective function (equation 1) consists of minimizing the worst observation on p_{od} over S scenarios. Such a solution may represent a solution preferred by the decision-maker who is reluctant to take risks (P_w is an NP-hard optimization problem as proven in Yu & Yang (1998)).

3.2. The bw*-robust Criterion*

In order to establish a more flexible decision recommendation mechanism, Gabrel et al. (2013) considered the *bw*-robust criterion to achieve robust paths: b is the best expected travel time and w is the travel times that decision-maker cannot accept. Formally, the linear program for RSPP (noted by P_{bw}), using the *bw*-robust criterion, can be written as follows:

$$(\mathrm{P}_{bw}) \qquad \max \sum_{k \in S} y_k \tag{3}$$

$$\text{s.t.} \quad \textit{Constraints } (2),$$

$$\sum_{(i,j) \in E} \hat{t}_{ij}^{k} x_{ij} - (1 - y_k) w \leq b y_k, \quad k = 1, \ldots, S, \tag{4}$$

$$x_{ij} \in \{0,1\}, \forall\ (i,j) \in E,\ y_k \in \{0,1\}, k = 1, \ldots, S.$$

In P_{bw}, the objective function (equation 3) maximizes the total number of scenarios that are better than the expected travel time b while ensuring that no considered scenarios exceed the worst expectation w. The robustness of the solutions provided by P_{bw} depends on both parameters b and w. However, choosing a reasonable assignment of the pair (b, w) is inefficient, because constraints (inequalities 4) provide a solution respecting the worst value of w for all scenarios. Hence, for some w the problem P_{bw} cannot be solved in reasonable runtime.

3.3. The win-loss Robust Criterion

The proposed *win-loss robust criterion* is inspired from the *bw*-robust criterion. The new criterion allows the decision-maker to adjust the weight between both “win" and “loss". In other words, its aim is to maximize the “winning rate" and minimize the

"expected loss". A linear program related to the win-loss robust criterion, denoted by P_{new}, can be stated as follows:

$$(\mathrm{P}_{new}) \qquad \max\ (1-\alpha) \times \sum_{k \in S} y_k - \alpha \times \theta \tag{5}$$

$$\text{s.t.} \quad \textit{Constraints}\ (2),$$

$$\sum_{(i,j) \in E} \hat{t}_{ij}^{k} x_{ij} \leq w + \theta, \qquad k = 1, \ldots, S, \tag{6}$$

$$\sum_{(i,j) \in E} \hat{t}_{ij}^{k} x_{ij} - (1 - y_k) M \leq b, \quad k = 1, \ldots, S, \tag{7}$$

$$\theta \geq 0,\ x_{ij} \in \{0,1\}, \forall\ (i,j) \in E,\ y_k \in \{0,1\}, \forall\ k = 1, \ldots, S. \tag{8}$$

In the objective function (equation 5), the term $\sum_{k \in S} y_k$ represents the total number of scenarios, where the travel time meets the best expected value b, and θ measures the delay related to the worst expected value w. Differently stated, the greater the value of $\sum_{k \in S} y_k$, the higher the chance of winning, and the smaller the value of θ, the lower the loss. The parameter α adjusts the weight between maximizing the win and minimizing the loss. In our study, the value of w is reached by solving P_w, and that of b may be estimated by using the empirical probability distribution of travel time related to each segment in the network (see Section 5.2).

We have already mentioned that P_w (and P_{bw}) is (are) NP-hard. One can observe that by setting $\alpha = 1$, and b and w to the optimal objective value of P_w, the optimal solution of P_{new} is the same as the optimal solution using the worst-case (its objective value is equal to 0). Consequently, the win-loss criterion leads to an optimization problem at least as difficult to solve as the problem induced by the worst-case criterion. Therefore, the robust shortest-path problem combined with the win-loss criterion is also NP-hard.

4. Lagrangian Relaxation-based Algorithm for the RSPP

As shown in Section 3.3, RSPP using the win-loss criterion remains NP-hard. According to the large number of vehicle to dispatch during critical hours, the computational time remains a critical factor when using the win-loss criterion to path planning. Further, in the tailored learning model, a large number of P_{bw} values must be solved in

order to reach the best parameter settings. Based on the special structure of RSPP, Lagrangian relaxation is widely used to design solution procedures for the shortest-path problem under uncertainty (see, for instance, Xing & Zhou, 2011, 2013; Yang & Zhou, 2014; Zeng et al., 2015; Zhang et al., 2017). In what follows, we introduce the Lagrangian relaxation-based algorithm for the RSPP, where the win-loss criterion is considered.

Let $\lambda = \{\lambda_1, \ldots, \lambda_S\}$ be a set of nonnegative Lagrangian multipliers associated to constraints (7); the following Lagrangian relaxation is established:

$$Z_d(\lambda) = \max\ (1-\alpha) \times \sum_{k=1}^{S} y_k - \alpha \times \theta + \sum_{k=1}^{S} \lambda_k \left(b - \sum_{(i,j)\in E} \hat{t}_{ij}^k x_{ij} + (1-y_k)M \right) \quad (9)$$

$$\text{s.t.} \quad (2),\ (6),\ (8),$$

$$\lambda_k \geq 0, \forall\ k = 1, \ldots, S.$$

$Z_d(\lambda)$ is defined in the primary space of the solution (x, y, θ). For a given $\bar{\lambda} \geq 0$, an upper bound of P_{new} can be computed by solving $Z_d(\bar{\lambda})$. By gathering all variables in the objective function (equation 9), we deduce that

$$Z_d(\lambda) = \max \sum_{k=1}^{S} (1 - \alpha - \lambda_k M) y_k - \sum_{k=1}^{S} \sum_{(i,j)\in E} \lambda_k \hat{t}_{ij}^k x_{ij} - \alpha \times \theta + \sum_{k=1}^{S} \lambda_k (b + M)$$

$$\text{s.t.} \quad (2),\ (6),\ (8),$$

$$\lambda_k \geq 0, \forall\ k \in S.$$

The dual function of $Z_d(\lambda)$, namely Z_d, is solved through the use of the sub-gradient method. Then, the optimal Lagrange multiplier vector is denoted by λ^*. Let Ω be the finite set of feasible solutions of $Z_d(\lambda)$, where

$$\Omega = \left\{ (x^t, y^t, \theta^t) \quad \text{subject to} \quad (2),\ (6),\ (8),\ t = 1, \ldots, T \right\}.$$

Therefore, the Lagrangian dual problem can be written as follows:

$$(\mathrm{LR}_d) \qquad Z_d = \min\ u$$

$$\text{s.t.} \quad u \geq (1-\alpha) \sum_{k=1}^{S} y_k^t - \alpha \times \theta^t + \lambda^t g^t, \quad t = 1, \ldots, T,$$

where $g^t = (g_1^t, \ldots, g_S^t)$ is an S-vector such that, $\forall\ k = 1, \ldots, S$,

$$g_k^t = b - \sum_{(i,j)\in E} \hat{t}_{ij}^k x_{ij}^t + (1 - y_k)M.$$

The S-vector v is called a sub-gradient of $\mathrm{LR}_d(\lambda)$ at λ' when the following condition is satisfied:

$$LR_d(\lambda) \geq LR_d(\lambda') + v(\lambda - \lambda'), \quad \forall\ \lambda \geq 0.$$

The S-vector g^t is a sub-gradient of x^t at λ', where x^t is the optimal solution of $Z_d(\lambda')$. Finally, the value of λ is updated according to the solution procedure proposed in Guignard (2003).

Algorithm 1 Lagrangian Relaxation-based Algorithm for P_{new} (LRA)

Require: An instance of P_{new}.

Ensure: A local optimal solution: $p_{od}^\star$.

1: **Initialization**: Set $t = 0$, $UB = +\infty$, $LB = 0$, $\lambda^t = (0, \ldots, 0)$;
2: **while** $\left(t < Iteration_{max}\right)$ **do**
3: Compute the optimal solution p_{od}^t of $Z_d(\lambda^t)$ by using Dijkstra's algorithm;
4: Update UB and $p_{od}^\star$ according to p_{od}^t, and set LB to the objective value of $p_{od}^\star$;
5: **if** $\left(LB<UB\right)$ **then**
6: Compute λ^{t+1} by using the sub-gradient method, and set $t = t + 1$;
7: **else**
8: Go to Step 11;
9: **end if**
10: **end while**
11: **return** $p_{od}^\star$.

Algorithm 1 describes the main steps of LRA used to solve P_{new}. The main loop (Steps from 2 to 11) always updates both current upper bound (UB) and lower bound (LB). At each iteration of LRA, UB denotes the objective value of P_{new} (related to a feasible solution) while LB corresponds to the objective value of $Z_d(\lambda)$ for a given λ. The main loop stops when the maximum number of iterations (noted $Iteration_{max}$) is performed. For the t^{th} iteration ($\forall\ t = 0, \ldots, Iteration_{max}$), LRA applies Dijkstra's

algorithm for solving $Z_d(\lambda^t)$ (Step 3). It is worth noting that if the worst observation related to p_{od}^t exceeds w, the value of θ is then adjusted until the constraint (6) is satisfied. According to the provided (feasible) solution p_{od}^t, LRA updates both current best upper and lower bounds (Step 4). If the optimality is not proved, the Lagrangian multipliers are computed by using the sub-gradient method (Step 6). Finally, the algorithm returns the best solution provided so far (Step 11).

5. Data-driven Robust Optimization Model

We have already mentioned that DRO is composed of two training models (see Figure 2) to derive itineraries according to GPS-data. Both training models serve to transform GPS tracks as a direct input of the robust optimization model P_{new}. In fact, the first model applies the hidden Markov model to generate the current traffic network information according to both GPS-data and the real-world transportation network. Based on the provided traffic network, the second model is used for determining the best parameter settings of P_{new}, i.e., the dimension of the uncertainty set $\hat{T}$, the estimation of the maximum win b and the value related to the weight α.

5.1. Traffic Network Training

The dataset used in this work was provided by DIDI Chuxing[1], which contains detailed tracks of the mobility of taxi vehicles for one month (from 01/11/2016 to 30/11/2016) in Chengdu City (see Section 6.1). Each taxi is characterized by its unique identifier, namely TaxiID. Whenever a taxi receives an order, such an order is assigned to a unique order-number, namely OrderID. During the order, every $2-4$ seconds, the taxi records its current physical location / position (Longitude, Latitude) and time information (Unixtime). For instance, a sample in the data collection displayed in Table 2 expresses the order OrderID=$OOOO$, the taxi with its order-number TaxiID=$TTTT$, which is located at position (Longitude, Latitude)=$(104.07656, 30.69468)$ at time Unixtime=1480496360 (i.e., 30 November 2016 at 08:59:20).

[1]Data source: https://gaia.didichuxing.com.

TaxiID	OrderID	Unixtime	Longitude	Latitude
TTTT	OOOO	1480496360	104.07656	30.69468

Table 2: Illustration of the GPS coordinates of a taxi.

As mentioned in Section 2, the growth of GPS tracking devices and applications has offered new opportunities for modeling traffic flow data. Herein, the transportation network G studied is constructed from OpenStreetMap, where package OSMnx is used (provided by Geoff (2017)). The network contains 4883 nodes and 13933 edges. Thus, the first step is to project the GPS tracks on the real network. Due to uncertain match results caused by GPS sampling error, GPS tracks cannot be directly used to measure the actual traffic. A common technique used for connecting GPS tracks to real road segments is the hidden Markov model; that is a machine learning model designed to regularly integrate noisy GPS' coordinates and road segments.

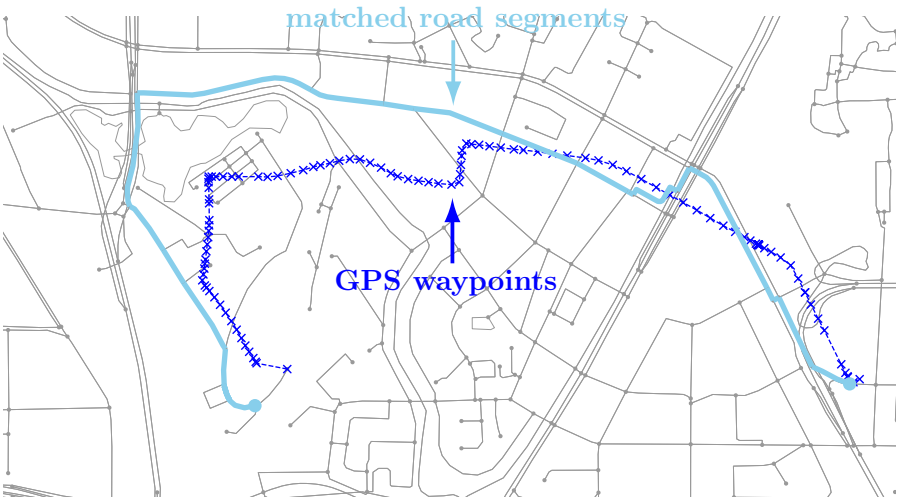


Figure 3: Result of matching road segments to a sequence of GPS' waypoints.

Let $cord^u = \{cord_o^u, \ldots, cord_t^u, \ldots, cord_d^u\}$ be a sequence of GPS' waypoints collected

by individual u when moving from the origin o to the destination d. Thus, the following steps are followed:

1. HMM searches for the road segments that the vehicle traversed in real time for all GPS' coordinates belonging to $cord^u$. Differently stated, the discrete states of HMM are represented by h road segments or edges, i.e., e_l, $l = 1, \ldots, h$.
2. Given a GPS location $cord_t^u \in cord^u$, there is an emission probability for each edge e_l, $p(cord_t^u|e_l)$, $\forall\ l = 1, \ldots, h$. It provides the probability that the measure $cord_t^u$ will be observed if the individual u was actually on the route segment e_l. HMM's objective is to find the set of edges maximizing the total emission probability.
3. In order to train traffic network, DRO applies the HMM-based map matching algorithm proposed in Wannes & Verbeke (2018) to generate the most likely road segments matched to a sequence of GPS tracks.

Figure 3 illustrates an available trajectory computed by HMM for a given sequence of GPS waypoints.

Based on the matching results reached by HMM, DRO uses a Hidden Markov Model-based Generator (HMMG) to provide the uncertainty set for the robust optimization model. Algorithm 2 describes the main steps of HMMG used for generating $\hat{T}$. HMMG can be viewed as a two-stage procedure. The first stage (Steps from 2 to 7) converts the GPS trajectories to the travel time, and creates a sample collection of the travel time for each edge. Indeed, for each edge, the travel time of a vehicle is equal to the length of the edge divided by the average speed of the vehicle. The second stage (Steps from 8 to 11) starts by determining the empirical distribution from the sample collection previously determined and second a set of scenarios for each edge is generated. Finally, HMMG returns the uncertainty set $\hat{T}$, which can be used to represent the traffic network.

5.2. Model Training and its Evaluation

In order to evaluate the robustness of the proposed robust optimization model P_{new}, we use the simulation-based robustness analysis method (see Section 6.1). Differently

Algorithm 2 Hidden Markov Model-based Generator (HMMG)

Require: A set of GPS tracks: O.

Ensure: An uncertainty set: $\hat{T}$.

Initialization: Affect each edge $(i, j) \in E$ to a sample collection, noted sam_{ij};
for each $cord^u$ in O **do**
  Apply HMM to compute the most likely path p_{od} associated with $cord^u$;
  Compute the average speed for traversing p_{od}, noted $v_{(o,d)}$;
  For each edge $(i, j) \in p_{od}$, compute the corresponding travel time by using $v_{(o,d)}$;
  For each edge $(i, j) \in p_{od}$, update sam_{ij} according to the provided travel time;
end for
for each (i, j) in E **do**
  Compute the empirical distribution function of sam_{ij}, noted EDF_{ij};
  Generate the set of scenarios $\{\hat{t}_{ij}^k \mid \forall k = 1, \ldots, S\}$ according to the provided EDF_{ij};
end for
return $\hat{T} = \{\hat{t}_{ij}^k \mid \forall (i, j) \in E, \forall k = 1, \ldots, S\}$.

stated, the robustness of a solution p_{od} (of P_{new}) is measured using the estimated probability of winning and the estimated cost of loss, where both estimations can be computed by simulating p_{od} on the traffic network. Therefore, the purpose of the model training is to enable P_{new} to provide the most robust itineraries for simulation experiments. In DRO's framework, the robustness of P_{new} depends on the following parameters: (i) the dimension of the uncertainty set $\hat{T}$, noted S (cf. Section 3.1), (ii) the weight between the probability of winning and the cost of loss α (cf. Section 3.3), and (iii) the best expected travel time b (cf. Section 3.3).

On the one hand, one can observe that the dimension of the uncertainty set $\hat{T}$ may lead to results of variable quality. More its dimension is larger, more the robustness of the model becomes larger, while the runtime effort becomes more significant. Additionally, according to the risk preference of the decision-maker, the weight parameter α is used to give a nice balance between the most promising goal and the worst case. On the other

hand, the estimation of b plays a crucial role for P_{new}. Indeed, it is used to drive the winning goal of P_{new}. In this case, more the value of b is smaller, more the winning goal is matchable, it becomes unattractive otherwise. Finally, with DRO, the estimation of b is based on the empirical distribution associated to each road segment of the network.

Given a network $G = (N, E)$ and OD a pair (o, d) of nodes, let q be the percentile rank and EDF_{ij} be the empirical distribution of edges e_{ij}, $\forall\ (i,j) \in E$; the value of b is equal to the travel time of the shortest path between o and d when the travel time of all road segments is fixed to the $q-$th percentile value of their empirical distribution. Formally, b can be computed by solving the following problem:

$$\begin{aligned} b \ = \ & \min \sum_{(i,j)\in E} t_{ij}^{q} x_{ij} \\ \text{s.t.} \quad & \textit{Constraints}\ (2), \\ & x_{ij} \in \{0,1\}, \forall\ (i,j) \in E, \end{aligned}$$

where t_{ij}^{q} is the $q-$th percentile value of EDF_{ij} for each edge $(i,j) \in E$.

6. Computational Part

The purpose of this section is two-fold: the first is to show how to determine a good trade-off between the quality of the obtained solutions when varying the parameters used by P_{new}; the second is to analyze the robustness of the proposed DRO, where its achieved results are compared to those representing a case study that reflects the peak period's traffic in Chengdu City, China. We note that all proposed methods were coded in C++ combined with Python and run on the Intel Pentium Core i7-4790 with 3.6 GHz.

6.1. Experimental Design

The problem P_{new} is NP-hard and solving it to optimality remains intractable. We then propose an alternative to solve it by using a simulation-based analysis method. For all experimental results, a real traffic network is considered during the morning peak period between 7 am and 9 am (a case study reflecting real data of Chengdu City, China

– cf. Section 5)). The GPS's dataset of the case study considered is divided into two parts:

- The first part: it is related to the training set, noted S_{train}. In this case, the data reflect the period from 01/11/2016 to 29/11/2016.
- The second part: it is related to the training set, noted S_{test}, where the data related to the day 30/11/2016 are considered. In this case, it represents the first day after the data-period used for S_{train}.

We note that S_{train} is used as an input of Algorithm 2 in order to generate the uncertainty set $\hat{T}$ whereas S_{test} is used to generate the simulated traffic network, namely G_{sim}. Of course, G_{sim} is based on the empirical distribution function provided by applying Algorithm 2 to the second set S_{test}.

For both sets S_{train} and S_{test}, in order to reduce the noise in the data, an order that matches one of the following conditions is canceled: (i) an order with a total travel time smallest than 5 minutes[2], (ii) an order whose final-point coincides with its starting-point, and (iii) road segments that are not matched with those of 30/11/2016. In addition, the road segment matched on 30/11/2016, but not for the period from 01/11/2016 to 29/11/2016, its travel time is fixed to the ratio between the length of the road per the maximum speed limit (i.e., 60 K/h). Note that $\hat{T}$ covers 12718 edges and those related to the simulated traffic network G_{sim} are equal to 11610 edges. In this case, G_{sim} has only one new edge doesn't belong to $\hat{T}$. Thus, only edges belonging to G_{sim} are considered; differently stated, 11610 edges are considered for solving both P_w and P_{new}.

The OD pairs required for the experimental design are generated from the real orders belonging to the morning peak period orders (between 7 am and 9 am) of the day 30/11/2016, namely OD_{sim} (OD_{sim} contains 20295 orders). For each available order, HMM is first called for finding an available path, and then the starting (resp. final) node of the path is setting equal to O (resp. D). Further, the order's real-travel time,

[2]A customer can cancel an order unconditionally within 5 minutes.

namely t^{od}_{real}, is used as a reference-point for analyzing the reliability of the robust optimization models used. Indeed, t^{od}_{real} is equal to the difference between the final-time and the starting-time of the same order. Both P_w and P_{new} are considered for computing itineraries related to all pairs belonging to OD_{sim}, where the set of itineraries provided by P_w (resp. P_{new}) is noted Sol_w (resp. Sol_{new}). For each itinerary belonging to either Sol_w or Sol_{new}, 10000 random runs are performed on G_{sim}. Finally, the robustness of the solutions belonging to Sol_w and Sol_{new} is measured by using the following two criteria:

- Pr_{Av}: the global average win probability over all OD pairs of OD_{sim};
- ts_{Av}: the global average over all OD pairs of OD_{sim}.

For each OD belonging to OD_{sim}, the probability's win related to a single path p_{od} is computed as follows:

$$Pr_{win}(p_{od}) = P\left(t_{sim} \leq t^{od}_{real}\right) = \frac{NB_{better}}{10000},$$

where t_{sim} denotes the travel time provided by simulating p_{od} on G_{sim}, and NB_{better} is the number of times that the simulation results outperforms t^{od}_{real}. Furthermore, during the simulation, the minimum loss is measured by the average of the time saving, where for a single path p_{od} (noted ts^{od}_{Av}) is computed as follows:

$$ts^{od}_{\mathrm{Av}} = \frac{\sum_{k=1}^{10000}(t^{od}_{real} - t^{k}_{sim})}{10000}. \tag{10}$$

6.2. Parameter Settings

It is well-known that when using approximate methods to solve hard problems, different parameter settings lead to a variability in the quality of solutions. For the problem P_{new}, there are three parameters to be tested (cf. Section 5.2) such that for the training model, the following tunings are considered:

- S: the dimension of the uncertainty set, $S \in \{10, 20, 40, 80, 160, 320, 640, 1280\}$.
- α: the weight between the probability of winning and losing, $\alpha \in [0.10, 0.95]$.

- q: the percentile rank used to calculate the best expected travel time b, where $q \in [0.50, 0.99]$.

For each combination regarding the values related to the three parameters used, P_{new} is performed on a training set composed of 100 OD pairs randomly generated from OD_{sim}. The achieved results are illustrated in Figure 4, where the x-axis (resp. y-axis) corresponds to the ts_{Av} (resp. Pr_{Av}) related to the training set when varying the combination of the three parameters. As shown in Figure 4, the following three parameter-settings are chosen (namely $parm1$, $parm2$ and $parm3$, respectively):

- $parm1$: $S = 160$, $\alpha = 0.15$, $q = 0.5$,
- $parm2$: $S = 1280$, $\alpha = 0.10$, $q = 0.53$, and
- $parm3$: $S = 1280$, $\alpha = 0.10$, $q = 0.58$.

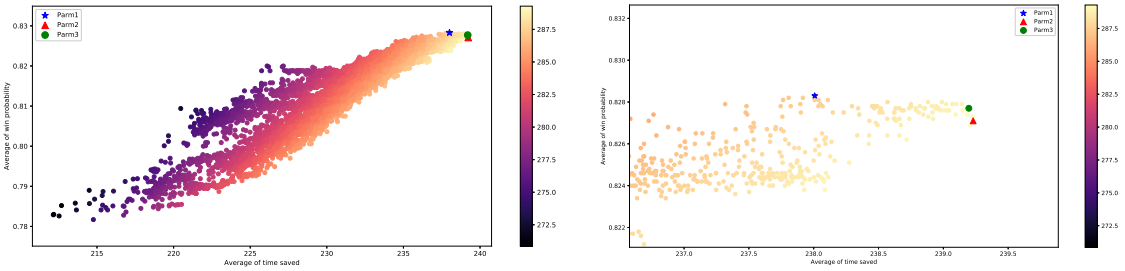


Figure 4: Behavior of the training model with different tunings.

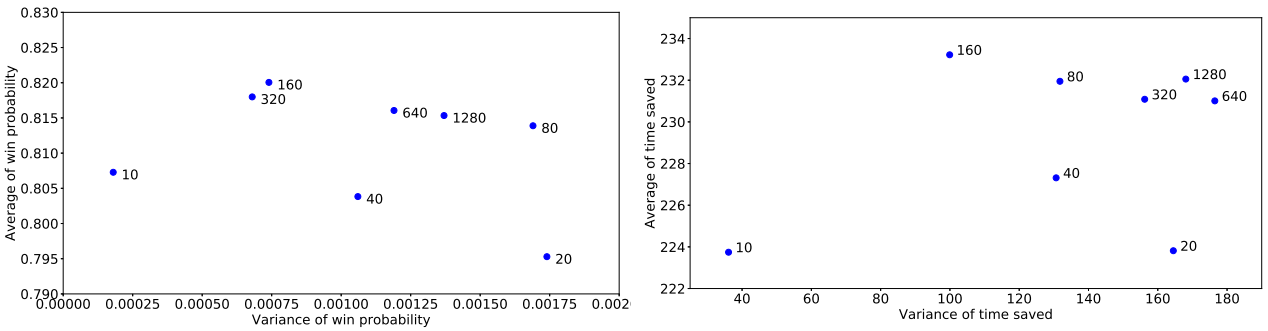
In what follows, we comment on the results reported in Figure 4:

1. First, by using the first parameter-setting $parm1$, one can observe that the solutions provided by P_{new} achieve the highest value for Pr_{Av}, i.e., $Pr_{\mathrm{Av}} = 0.8283$ (the start in blue-color represented on the left-hand and the right-hand of the figure).
2. Second, with the second parameter-setting $parm2$, the solutions reached by P_{new} provides the largest value for ts_{Av}, i.e., $ts_{Av} = 239.2301$ (the triangle in red-color represented on the left-hand and the right-hand of the figure).

3. Third, by applying the third parameter-setting $parm3$, the solutions achieved by P_{new} provides the best value for $(Pr_{\mathrm{Av}} + ts^{norm}_{Av})$, where ts^{norm}_{Av} denotes the normalized value of ts_{Av} (the dot in green-color represented on the left-hand and the right-hand of the figure). In this case, we obtain $Pr_{\mathrm{Av}} = 0.8277$ and $ts_{Av} = 239.1967$.

Because we seek to the settings highlighting solutions with good quality, then the three settings will be used to solve P_{new} for all orders belonging to OD_{sim}.

6.3. Sensitivity analysis



(a) Sensitivity analysis of win probability.

(b) Sensitivity analysis of time saving.

Figure 5: Sensitivity analysis of the main parameters of DRO.

One can observe that the parameter S (i.e., the dimension of the uncertainty set) used by the designed framework may affect the robustness of DRO. In order to study the effect of that parameter, we focussed on two indicators: the average and the variance of the solutions provided by DRO according to different values of S. In this case, Figures 5a and 5b show the effect of S on Pr_{Av} and ts_{Av} respectively, through the use of the solutions reached by the training model (as shown in Figure 4). For both sub-figures, both average “win probability" and “time saved" are represented on the y-axis while the variances related to both “win probability" and “time saved" are represented in the x-axis. According to the results displayed in Figures 5a and 5b, we observe what follows:

1. The larger the average values and the smaller the variance values, the more robust the results become.

2. Using small number of scenarios to simulate the uncertainty generally induces low quality solutions, i.e, whenever S varies in the discrete interval $\{10, 20, 40\}$. For instance, with $S = 10$ (resp. $S = 40$), the corresponding average values (Figure 5a) for Pr_{Av} and ts_{Av} are respectively equal to 0.807 and 224 (resp. 0.804 and 227). While for $S = 10$ (resp. $S = 40$), the corresponding variance (Figure 5b) related to Pr_{Av} and ts_{Av} are equal to 0.00018 and 36.096 (resp. 0.00106 and 130.659) respectively.

   (a) On the one hand, in terms of win probability, the solutions reached with $S = 10$ highlight the stability of the solutions when compared to those achieved by DRO with $S = 40$.

   (b) On the other hand, in terms of time saved, with $S = 40$ the DRO is able to provide less time consuming solutions than DRO with $S = 10$, while all the provided solutions seem less stable.

3. Finally, with $S = 160$, DRO seems to work better since it is able to provide a better stability of solutions with interesting win probability and time saved.

6.4. P_w and its robustness

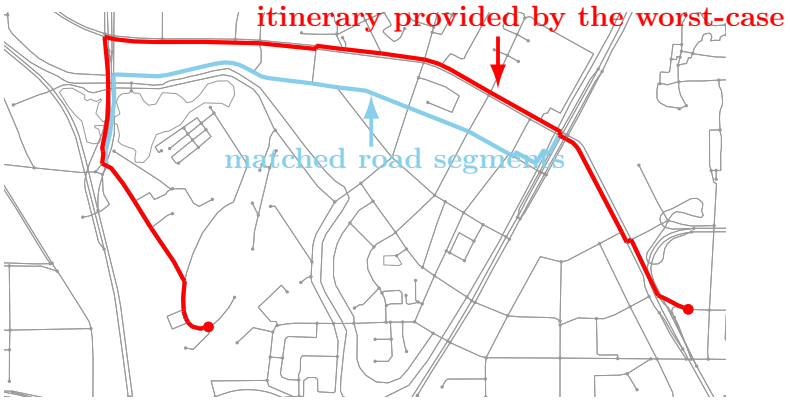


Figure 6: Illustration of the itinerary provided by using the worst-case robust criterion.

In this section, the robustness of the model is analyzed when introducing the worst-criterion. We do it by adapting the solution procedure LRA (cf., Algorithm 1) for approximately solving P_w. On the one hand, Figure 6 illustrates the trajectory matched by HMM and that provided by P_w when using the same order (cf. Figure 3). On the other hand, Figure 7 shows Pr_{Av} and ts_{Av} provided by using different scenarios S for characterizing the travel time uncertainty.

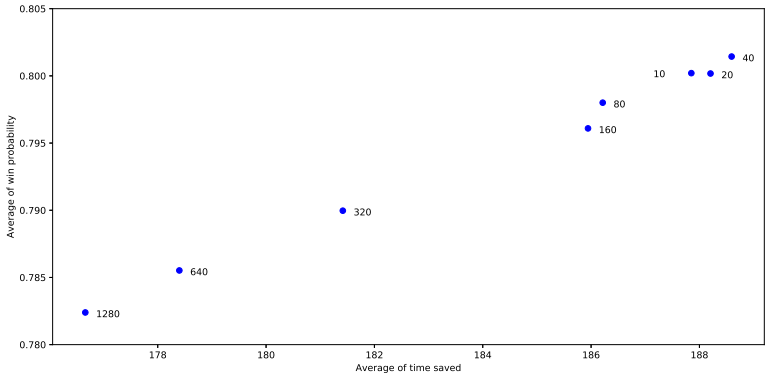


Figure 7: Robustness analysis for the worst-case robust criterion.

In what follows, we comment on the results provided:

1. One can observe that P_w tries to avoid a section of road segments that may be congested according to the worst observation encountered throughout the historical data.
2. According to the simulation results observed in Figure 7, P_w achieves the most robust itineraries with $S = 40$, where $Pr_{\mathrm{Av}} = 0.801$ and $ts_{\mathrm{Av}} = 189$. In this case,
   (a) for all orders belonging to OD_{sim}, the travel time of the solutions provided by P_w (i.e., $S = 40$) has 80% of chance to be better than that related to the real orders.
   (b) P_w is able to save about 189 seconds on average. It is worthy to notice that the number of scenarios used to build the uncertainty set has a direct impact on the runtime required when solving P_w. Indeed, form Figure 7, one can observe

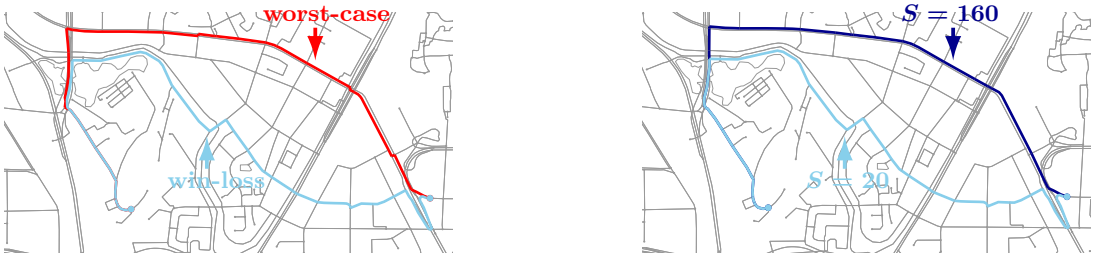
that for $S \geq 80$, LRA's performance decreases proportionally whenever the number of scenarios S increases.

6.5. DRO and its Robustness

In order to analyze the robustness of DRO combined with the win-loss criterion, we perform a comparative study based on the simulation results. Figure 8 displays the itineraries generated by P_{new} when using the same order as Figure 3) and Figure 7). From Figure 8a, one can observe that the win-loss criterion P_{new} with $S = 20$, $\alpha = 0.78$ and $q = 0.1$ tries to build a different itinerary than that achieved by P_{w} (i.e., for $S = 20$), where the taxi has a greater chance of arriving at the expected time. On the one hand, the win probability associated to the itinerary in blue-color provided by P_{new} (i.e., 51%) is better than that provided by P_{w} (i.e., 14%). In detail, P_{new} provide an itinerary such that, the corresponding travel time is better than the real case in 51% of cases over 10000 random simulations, and the average of the travel time is equal to 415 seconds. However, for the itinerary in red-color provided by P_{w}, the travel time is only better than the real case in 14% of cases, and the average of the travel time rose to 450 seconds.

Figure 8b displays the itineraries provided when solving P_{new}, where two parameter-settings are considered: (i) $S = 20$, $\alpha = 0.78$ and $q = 0.1$, and (ii) $S = 160$, $\alpha = 0.50$ and $q = 0.15$. According to the simulation results, the win probability of the itinerary generated by the parameter-settings (i) (resp. (ii)) is equal to 51% (resp. 50%) and the average travel time is equal to 415 seconds (resp. 413 seconds). This means that, P_{new} with the first parameter-settings is able to find an itinerary with a higher win probability, but when it fails, we will also face greater losses.

Based on the simulation results, Figure 9) displays the distribution of the travel time related to P_{w} with $S = 20$, P_{new} with parameter-settings (i) and (ii). We can observe that, P_{new} with parameter-settings (ii) can provide a better solution than the other two. Indeed, the travel time related to P_{new} with parameter-settings (ii) is less than 500 seconds in most cases. Figure 9) shows that we can effectively drive the objective function of P_{new} from the GPS data in order to produce robust itineraries.



(a) Itineraries provided both P_w and P_{new}.

(b) Itineraries provided by P_{new} with different values for S.

Figure 8: Comparison of itineraries provided by different robust criteria.

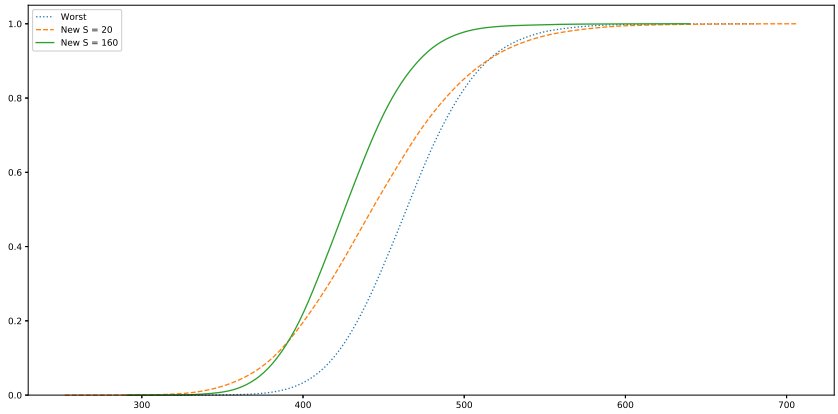


Figure 9: Distribution of the travel time provided by the simulation.

In terms of statistical analysis, using a large number of scenarios to describe the uncertainty set can effectively improve the robustness of the solution. However, a large number of scenarios may significantly increase the average runtimes. Unlike P_w, the proposed model P_{new} can efficiently balance between robustness and runtime.

Figure 10 evaluates the behavior, in term of robustness, between both P_w and P_{new}, where the detailed values are displayed in Table 3. Table 3 displays the average value (Average), the standard deviation (Std Dev), the first quartile (25%), the second quartile (50%) and the third quartile (75%) of both Pr_{Av} and ts_{Av} related to the solutions provided by both P_w and P_{new}. In what follows, we comment on the results provided when applying the DRO-based model:

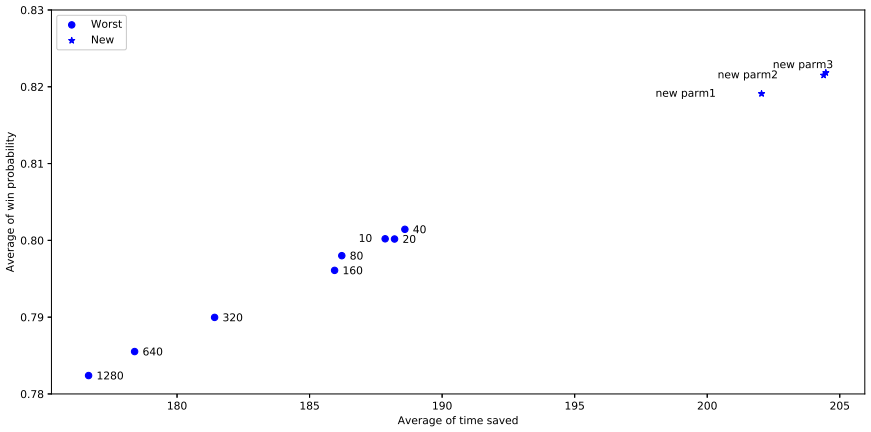


Figure 10: Robustness analysis of P_{new} versus that of P_w: win-loss vs worst-case criteria.

1. First, from Figure 10, one can observe that the simulation results related to P_{new} with $parm1$, $parm2$ and $parm3$ outperforms P_w.
2. Second, when using the same uncertainty set, the time required to calculate P_{new} is usually shorter than the time required to calculate P_w (see CPU_m of Table 3).
3. Third, from the last line of Table 3, one can observe that P_{new} with the third parameter-setting $parm3$ achieves the best win probability and time saving.

Models	Average	Std. Dev	25%	50%	75%	CPU_m
Worst 10	(0.800, 188)	(0.345, 275)	(0.773, 36)	(0.999, 159)	(1, 306)	0.30
Worst 20	(0.800, 188)	(0.348, 277)	(0.779, 38)	(0.999, 161)	(1, 308)	0.37
Worst 40	(0.801, 189)	(0.347, 276)	(0.793, 38)	(0.999, 161)	(1, 307)	0.68
Worst 80	(0.798, 186)	(0.350, 278)	(0.780, 36)	(0.999, 161)	(1, 307)	1.38
Worst 160	(0.796, 186)	(0.352, 278)	(0.771, 34)	(0.999, 161)	(1, 306)	2.47
Worst 320	(0.790, 181)	(0.357, 278)	(0.748, 30)	(0.999, 157)	(1, 302)	4.05
Worst 640	(0.786, 178)	(0.360, 280)	(0.728, 27)	(0.999, 155)	(1, 302)	7.35
Worst 1280	(0.782, 177)	(0.362, 279)	(0.708, 25)	(0.999, 152)	(1, 299)	13.87
new $parm1$	(0.819, 202)	(0.332, 276)	(0.846, 51)	(1, 172)	(1, 320)	1.97
new $parm2$	(0.821, 204)	**(0.329, 275)**	(0.852, 52)	**(1, 174)**	(1, 321)	9.76
new $parm3$	**(0.822, 204)**	**(0.329, 275)**	**(0.855, 53)**	**(1, 174)**	**(1, 322)**	8.86

Table 3: Statistical analysis of computational results: P_w vs P_{new}.

7. Conclusion

In this paper, a data-driven robust optimization was designed for solving the itinerary planning problem, where the GPS-data, data-driven and robust optimization were combined for highlighting the punctuality rate. First, a new win-loss robust criterion was proposed, which was used for maximizing the probability of achieving the expected travel time and minimizing the longest possible delay. Second, the data-driven robust model is based upon two types of training: (i) generating a traffic network related to the GPS-data, and (ii) training the robust model by using the generated traffic network as an input. Finally, the experimental part showed that the proposed simulation-based protocol, on real-world application, is able to provide high quality solutions. Due to the gap between the GPS waypoints and the actual positions, the estimated travel time may vary from the actual traffic network. Therefore, we believe that Chengdu City may consider installing speed measurement devices on some small forks to improve the estimation of actual traffic.

Because of the flexibility of the new robust "win-lose" criterion, there are plenty possibilities for further investigation involving efficient methods for more complex problems. First, single and multi-objective combinatorial optimization problems with high uncertainty may be potential candidates to explore. In general, we believe that the higher the dimension of the uncertainty set, the more robust the results become. Second, the use of high-dimensional uncertainty sets will generally increase the computational complexity of the model, which may prevent us from designing solution methods capable of achieving good solutions in a reasonable time. Third, in this work, the problem was addressed without considering the correlation between travel times associated with different routes, which is a more complex variant of the problem. Of course, as stated throughout the paper, the objective of our study was related to the use of large-scale GPS data to solve the real traffic problem. Finally, we highlight the literature study that relies on the use of the covariance matrix to simulate the correlation between travel times. In this case, the use of the nonlinear model will make the proposed DRO more difficult to solve. We hope for a future extension of this study to propose a linear model that can be used to

simulate the correlation by applying scenario-based optimization.

Acknowledgment

This work was supported by the General Program of National Natural Science Foundation of China (No. 72071211) and the Special Project of Education and Teaching Reform in Central Universities (No. 31412012201). We also thank DiDi for the data source: Chuxing GAIA Open Data Initiative, and the Innovation and Talent Base for Income Distribution and Public Finance (B20084).

References

Aissi, H., Bazgan, C., & Vanderpooten, D. (2009). Min-max and min-max regret versions of combinatorial optimization problems: A survey. *European Journal of Operational Research*, *197*, 427–438.

Bansak, K., Ferwerda, J., Hainmueller, J., Dillon, A., Hangartner, D., Lawrence, D., & Weinstein, J. (2018). Improving refugee integration through data-driven algorithmic assignment. *Science*, *359*, 325–329.

Ben-Tal, A., El Ghaoui, L., & Nemirovski, A. (2009). *Robust Optimization*. Princeton series in applied mathematics. Princeton: Princeton University Press.

Bertsimas, D., Brown, D., & Caramanis, C. (2011). Theory and applications of robust optimization. *SIAM Review*, *53*, 464–501.

Bertsimas, D., & Brown, D. B. (2009). Constructing uncertainty sets for robust linear optimization. *Operations Research*, *57*, 1483–1495.

Bertsimas, D., Gupta, V., & Kallus, N. (2018a). Data-driven robust optimization. *Mathematical Programming*, *167*, 235–292.

Bertsimas, D., & Sim, M. (2004). The price of robustness. *Operations Research*, *52*, 35–53.

Bertsimas, D., Sim, M., & Zhang, M. (2018b). Adaptive distributionally robust optimization. *Management Science*, *65*, 604–618.

Bertsimas, D., & Thiele, A. (2006). Robust and data-driven optimization: Modern decision making under uncertainty. In *Models, Methods, and Applications for Innovative Decision Making* INFORMS TutORials in Operations Research chapter Models, Methods, and Applications for Innovative Decision Making. (pp. 95–122). INFORMS.

Caserta, M., & Voß, S. (2019). The robust multiple-choice multidimensional knapsack problem. *Omega*, *86*, 16–27.

Chassein, A., Dokka, T., & Goerigk, M. (2019). Algorithms and uncertainty sets for data-driven robust shortest path problems. *European Journal of Operational Research*, *274*, 671–686.

Chen, B. Y., Li, Q., & Lam, W. H. (2016). Finding the k reliable shortest paths under travel time uncertainty. *Transportation Research Part B: Methodological*, *94*, 189 – 203.

Chen, B. Y., Yuan, H., Li, Q., Lam, W. H., Shaw, S.-L., & Yan, K. (2014). Map-matching algorithm for large-scale low-frequency floating car data. *International Journal of Geographical Information Science*, *28*, 22–38.

Cheng, J., Leung, J., & Lisser, A. (2016). New reformulations of distributionally robust shortest path problem. *Computers & Operations Research*, *74*, 196–204.

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

Gabrel, V., Murat, C., & Thièle, A. (2014). Recent advances in robust optimization and robustness: an overview. *European Journal of Operational Research*, *235*, 471–483.

Gabrel, V., Murat, C., & Wu, L. (2013). New models for the robust shortest path problem: complexity, resolution and generalization. *Annals of Operations Research*, *207*, 97–120.

Geoff, B. (2017). Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, *65*, 126–139.

Goerigk, M., & Schöbel, A. (2016). Algorithm engineering in robust optimization. In L. Kliemann, & P. Sanders (Eds.), *Algorithm Engineering*. Springer, Cham volume 9220 of *Lecture Notes in Computer Science*.

Goli, A., Tirkolaee, E. B., Malmir, B., Bian, G.-B., & Sangaiah, A. K. (2019). A multi-objective invasive weed optimization algorithm for robust aggregate production planning under uncertain seasonal demand. *Computing*, *101*, 499–529.

Guignard, M. (2003). Lagrangian relaxation. *Sociedad de Estadistica e Investigacion Operativa Top*, *11*, 151–228.

Hanks, R. W., Lunday, B. J., & Weir, J. D. (2019). Robust goal programming for multi-objective optimization of data-driven problems: A use case for the united states transportation command's liner rate setting problem. *Omega*, *90*.

Hummel, B. (2006). Map matching for vehicle guidance. In R. Billen, E. Joao, & D. Forrest (Eds.), *Dynamic and Mobile GIS: Investigating Space and Time* chapter 10. (pp. 1–12). CRC Press: Florida. (1st ed.).

Kasperski, A., & Zieliński, P. (2016). Robust discrete optimization under discrete and interval uncertainty: a survey. In M. Doumpos, C. Zopounidis, & E. Grigoroudis (Eds.), *Robustness Analysis in Decision Aiding, Optimization, and Analytics* (pp. 113–143). Springer, Cham volume 241 of *International Series in Operations Research & Management Science*.

Kouvelis, P., & Yu, G. (1997). *Robust discrete optimization and its applications*. Boston: Kluwer Academic.

Lee, L., Jones, M., Ridenour, G. S., Bennett, S. J., Majors, A. C., Melito, B. L., & Wilson, M. J. (2016). Comparison of accuracy and precision of gps-enabled mobile devices. In *2016 IEEE International Conference on Computer and Information Technology (CIT)* (pp. 73–82).

Newson, P., & Krumm, J. (2009). Hidden markov map matching through noise and sparseness. In *17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL GIS 2009), November 4-6, Seattle, WA* (pp. 336–343).

Roy, B. (2010). Robustness in operational research and decision aiding: A multi-faceted issue. *European Journal of Operational Research*, *200*, 629–638.

Shao, H., Lam, W. H. K., Sumalee, A., Chen, A., & Hazelton, M. L. (2014). Estimation of mean and covariance of peak hour origin–destination demands from day-to-day traffic counts. *Transportation Research Part B: Methodological*, *68*, 52–75.

Stabler, B., Bar-Gera, H., & Sall, E. (2016). Transportation networks for research. URL: `https://github.com/bstabler/TransportationNetworks`.

Su, H., Yang, J., & Yang, C. (2019). A robust optimization approach to multi-interval location-inventory and recharging planning for electric vehicles. *Omega*, *86*, 59–75.

Tirkolaee, E. B., Aydin, N. S., Ranjbar-Bourani, M., & Weber, G.-W. (2020). A robust bi-objective mathematical model for disaster rescue units allocation and scheduling with learning effect. *Computers & Industrial Engineering*, *149*.

Tirkolaee, E. B., Mahdavi, I., & Esfahani, M. M. S. (2018). A robust periodic capacitated arc routing problem for urban waste collection considering drivers and crew's working time. *Waste Management*, *76*, 138–146.

Wannes, M., & Verbeke, M. (2018). Hmm with non-emitting states for map macthing. In *European Conference on Data Analysis (ECDA)*. Paderborn, Germany.

Wu, X. (2015). Study on mean-standard deviation shortest path problem in stochastic and time-dependent networks: A stochastic dominance based approach. *Transportation Research Part B: Methodological*, *80*, 275–290.

Xia, F., Rahim, A., Kong, X., Wang, M., Cai, Y., & Wang, J. (2017). Modeling and analysis of large-scale urban mobility for green transportation. *IEEE Transactions on Industrial Informatics*, *14*, 1469–1481.

Xing, T., & Zhou, X. (2011). Finding the most reliable path with and without link travel time correlation: A lagrangian substitution based approach. *Transportation Research Part B: Methodological*, *45*, 1660–1679.

Xing, T., & Zhou, X. (2013). Reformulation and solution algorithms for absolute and percentile robust shortest path problems. *IEEE Transactions on Intelligent Transportation Systems*, *14*, 943–954.

Yang, C., & Gidofalvi, G. (2017). Fast map matching, an algorithm integrating hidden markov model with precomputation. *International Journal of Geographical Information Science*, *32*, 547–570.

Yang, L., & Zhou, X. (2014). Constraint reformulation and a lagrangian relaxation-based solution algorithm for a least expected time path problem. *Transportation Research Part B: Methodological*, *59*, 22–44.

Yu, G., & Yang, J. (1998). On the robust shortest path problem. *Computers & Operations Research*, *25*, 457–468.

Zeng, W., Miwa, T., Wakita, Y., & Morikawa, T. (2015). Application of lagrangian relaxation approach to α-reliable path finding in stochastic networks with correlated link travel times. *Transportation Research Part C: Emerging Technologies*, *56*, 309–334.

Zhang, Y., Max Shen, Z.-J., & Song, S. (2017). Lagrangian relaxation for the reliable shortest path problem with correlated link travel times. *Transportation Research Part B: Methodological*, *104*, 501–521.

Zhang, Y., Shen, Z.-J. M., & Song, S. (2016). Parametric search for the bi-attribute concave shortest path problem. *Transportation Research Part B: Methodological*, *94*, 150–168.

Zhang, Y., Song, S., Shen, Z. M., & Wu, C. (2018). Robust shortest path problem with distributional uncertainty. *IEEE Transactions on Intelligent Transportation Systems*, *19*, 1080–1090.