Robust Transit Network Design with Stochastic Demand Considering Development Density

Kun An, Hong K. Lo*

Department of Civil and Environmental Engineering,
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

Abstract

This paper analyzes the influence of urban development density on transit network design with stochastic demand by considering two types of services, rapid transit services, such as rail, and flexible services, such as dial-a-ride shuttles. Rapid transit services operate on fixed routes and dedicated lanes, and with fixed schedules, whereas dial-a-ride services can make use of the existing road network, hence are much more economical to implement. It is obvious that the urban development densities to financially sustain these two service types are different. This study integrates these two service networks into one multi-modal network and then determines the optimal combination of these two service types under user equilibrium (UE) flows for a given urban density. Then we investigate the minimum or critical urban density required to financially sustain the rapid transit line(s). The approach of robust optimization is used to address the stochastic demands as captured in a polyhedral uncertainty set, which is then reformulated by its dual problem and incorporated accordingly. The UE principle is represented by a set of variational inequality (VI) constraints. Eventually, the whole problem is linearized and formulated as a mixed-integer linear program. A cutting constraint algorithm is adopted to address the computational difficulty arising from the VI constraints. The paper studies the implications of three different population distribution patterns, two CBD locations, and produces the resultant sequences of adding more rapid transit services as the population density increases.

Keywords: transit network design; robust; stochastic demand; population density

1. Introduction

The Transit Network Design Problem (TNDP) is to decide the locations of stations, route alignment as well as

* Corresponding author. Tel.: +(852) 2358-8742; fax: +(852) 2358-1534.
E-mail address: cehklo@ust.hk
frequency to serve the travel demands between specific origin-destination (OD) pairs. Due to high construction and operating costs of rapid transit line (RTL), some lines may face low passenger load, and some may even require government subsidy for their operations. Indeed, the population density has a great influence on the sustainability of RTL. The government thus must be prudent in developing RTL and the sequence in constructing different lines to cope with population increases. For a newly developed region, the initial residential density may not be sufficient to support a RTL. Even for regions with high population densities, the travel demand may fluctuate from day to day, making it uneconomical to rely on RTL alone to serve the demand. Dial-a-ride (DAR) services, in contrast, are able to utilize the existing road network, thus having relatively lower capital costs, mainly involving the procurement, operations and maintenance of the vehicle fleet. Meanwhile, they have great flexibility to cater for demand fluctuation. However, the congestion effect of dial-a-ride services cannot be neglected. Thus, it may not be economical and environmentally efficient to rely heavily on dial-a-ride services for areas with a large population producing relatively stable travel demands. The goal of this study is to find out the critical development density when RTL is to be first built and the construction phases as the population density gradually increases.

The Network Design Problem (NDP) can be classified into discrete, continuous and mixed, as discussed in Yang and Bell (1998). The Discrete NDP is concerned with the network topology itself (Wang et al., 2013, Gao et al., 2005, Lai and Lo, 2004). Examples include scheduling or routing of a service network. The Continuous NDP takes the network topology as given and is concerned with optimizing the network parameters. Examples include enhancing the link capacity or setting the toll charges (Gao et al., 2004, Ekström et al., 2012). The Mixed NDP combines the two types to simultaneously determine new links to be added and capacity increases of existing roads (Luathep et al., 2011). The transit network design problem falls in the category of Mixed NDP which involves determining discrete and continuous variables, namely, transit line alignment and frequency. Most existing studies focus on the deterministic TNDP, assuming that the OD demand is fixed and known. It is typically formulated as a mixed integer linear program (MILP) where the station selection, line alignment and frequencies are determined simultaneously to achieve a certain objective, such as cost minimization or coverage maximization (Wan and Lo, 2009, Bruno et al., 1998).

The literature on NDP concerning uncertain demand can be classified into two categories. The first approach is stochastic programming, which assumes known demand distributions and utilizes Monte-Carlo simulation to decompose the random demands into a finite number of scenarios for approximating the cost expectation, and is formulated as a MILP to minimize the total expected cost (Ruszczynski, 2008, Birge and Louveaux, 1988, Benders, 1962) and solved by a commercial software, such as CPLEX, or the L-shaped method. An and Lo (2014a, 2014b) proposed an alternative method, namely, the service reliability (SR)-based approach, which separates the large-size MILP into two phases for solution efficiency. The second approach is robust optimization, which focuses on the min-max problem, namely, optimization of the worst case scenario. It requires that the network design solutions, determined before the demand realization, are feasible for all demand realizations. The side effect is that the solutions may be overly conservative. Some studies thus turned to refining the uncertainty set such that all the realized demand within the set are satisfied while those outside are ignored. It is important to trade off the size of the uncertainty set and the robustness level (Bertsimas and Sim, 2004, Ben-Tal and Nemirovski, 1999).

The aforementioned studies generally specified the OD demand to be satisfied, either deterministic or stochastic (Ben-Tal et al., 2011, Wan and Lo, 2009), only a few traced back to the urban development density which generates the OD demands in the first place. Samanta and Jha (2011) proposed a rail transit line model considering different objectives such as ridership maximization or user cost minimization. Laporte et al. (2007, 2005) integrated trip generation, trip distribution and mode choice into the transit network design problem to produce OD demands for each transit mode. Quadrifoglio and Li (2010, 2009) investigated the feeder transit design problem to find the critical population density for fixed and demand responsive services. Although these studies somewhat examined the relationship between population density and OD demands on network design (Samanta and Jha, 2011, Laporte et al., 2005, Li and Quadrifoglio, 2010), they did not consider the inherent OD demand fluctuation.

In addition to the challenge of including stochastic demand and development density simultaneously, this study also incorporates user equilibrium (UE) passenger flows in a multi-modal transit network. The NDP with UE flows is typically formulated as a bi-level problem, in which the upper level problem focuses on generating the optimal network design; whereas the lower level represents travelers’ travel choices. This bi-level problem is typically non-linear and non-convex. Some studies formulate the UE principle as variational inequality (VI) constraints, which
reduce the bi-level problem into a single level problem with VI constraints. Through applying a cutting constraint algorithm, the single level problem can be solved iteratively, alleviating the onerous task on feasible paths enumeration (Ekström et al., 2012, Luathep et al., 2011). Various solution approaches have been developed to deal with bi-level problems, such as heuristic approach, global optimization approach (e.g., Wang et al., 2013, Wang and Lo, 2010). Marcotte and Nguyen (1998) introduced the hyper-path concept in transit assignment and formulated a user equilibrium model considering passenger strategies on routes selection. Although the TNDP with UE flows has been studied intensively, few studies have investigated the influence of demand uncertainty and population density.

This paper aims at analyzing the influence of urban development density on robust TNDP with stochastic demand under UE flows by considering two types of services, rapid transit line (RTL) and dial-a-ride (DAR) services. The RTL operate on fixed routes and schedules, which may include multiple lines, whereas DAR services are demand responsive to carry the demand realized on a particular day that exceeds the capacity of the RTL. Passengers can transfer between these two modes. The main contribution of this paper is as follows:

1. Instead of assuming exogenous OD demands, we establish the relationship between urban development density and OD demands via the gravity distribution model. We then determine the minimum or critical population density required to financially sustain the first RTL to be built, as well as the construction sequence of adding more RTL services as the population density increases.

2. This study integrates the RTL and DAR services into one multi-modal network and then determines the optimal combination of these two service types under UE passenger flows for a given development density.

3. We utilize the approach of robust optimization to capture the uncertainty demand as a polyhedral uncertainty set. It is reformulated as an MILP by replacing the uncertainty set by its dual problem. The RTL configuration and frequencies, or the here-and-now variables, are determined in the MILP and fixed for the planning horizon, while DAR services deployment for the worst case scenario is determined, with their congestion effect accounted for in the road network.

The reminder of this paper is as follows: Section 2 presents the mathematical formulation. Section 3 illustrates the model and solution algorithm with a case study. Section 4 concludes our work and proposes future extensions.

2. Model formulation

2.1. Demand generation

As mentioned earlier, we investigate the influence of population density on TNDP. To this end and for simplicity of illustration, we adopt the density saturation gradient model to represent the evolutionary process of population density, which states that the intensity of population density declines as the distance or travel time to CBD increases. It can be represented by the basic equation: $\Lambda_\omega = \Lambda_0 e^{-b\omega}$, where $\Lambda_\omega$ is the population density at distance $\omega$ from the city center; $\Lambda_0$ is the density of the central business district (CBD) at the city center; $b$ stands for the density gradient or slope factor. In most urban areas, the higher CBD density and lower suburban density will tend to equalize over time. As shown in Fig. 1, the population density will increase from the bottom red curve to the top horizontal line over time; that is, $b$ will decrease gradually. The trip generation rate in the catchment area of a specific transit station decreases linearly with the walking distance from the station, represented by the black dashed line, plotted on the right Y-axis. Let $\omega_i$ be the coordinate of station $i$ and $r_i$ be the radius of its catchment area. We assume the catchment areas of stations do not overlap with each other. The trip generation rate is defined as $a_i = a_0 - c_0 \omega_i [\omega - \omega_i], \omega_i - r_i \leq \omega \leq \omega_i + r_i$, where $a_0, c_0$, respectively, are the intercept and slope of the trip generation rate function. The total trips generated at station $i$ for a linear network is calculated by

$$P_i = \int_{\omega_i-r_i}^{\omega_i+r_i} a_i \Lambda_\omega d\omega = \int_{\omega_i-r_i}^{\omega_i+r_i} [a_0 - c_0 (\omega - \omega_i)] \Lambda_\omega e^{-b\omega} d\omega + \int_{\omega_i-r_i}^{\omega_i+r_i} [a_0 + c_0 (\omega - \omega_i)] \Lambda_\omega e^{-b\omega} d\omega$$

(1)

The blue dotted line represents the trips generated along the transit line when $a_0 = 1, c_0 = 0.03$. It is formulated as a Christmas tree shaped function with its peak constrained by the population density as represented by the red line.
For the target year, the population density line is specified by the basic equation. The total trip production \( P_i \) from zone \( i \) thus can be calculated by substituting \( \omega, r, a \) into Equation (1). With production \( P_i \) known, we obtain the OD demands by the gravity distribution model, expressed as 
\[ q_{ij} = P_i \left( I_j F_{ij} / \sum_{k \neq j} I_k F_{ik} \right), \]
where \( q_{ij} \) stands for the amount of trips from zone \( i \) to \( j \); \( I_j \) is the attractiveness for zone \( j \), which exponentially decreases with its distance to CBD, i.e., \( I_j = 100 e^{-|d|} \); \( F_{ij} \) is the travel cost friction factor that represents impedance to make trips of various distances and is set proportional to the distance between two stations \( l_{ij} \). In this way, through the trip generation model (1) and gravity distribution model described above, we obtain the expected OD demand. In this paper, we take the planning horizon as the morning peak hour within a period (e.g., a year). To simplify the notation, we use \( d \) to represent a specific OD pair instead of \( ij \) in the following. The OD demand, \( Q^d \), fluctuates from day to day with the mean \( q^d \) or \( q^d \): 
\[ Q^d \in U^d \text{ and } E(Q^d) = q^d. \]
Let \( D^d \) be the destination node of OD pair \( d \). \( U^d \) is known as the uncertainty set in robust optimization. Let \( \theta \) be the uncertainty level. In particular, a polyhedral uncertainty set is defined as:
\[ Q^d \in U^d \equiv \left\{ Q^d : q^d \leq \bar{q}^d, \sum_{d \in \{d \mid d' \neq j\}} Q^{d'} \leq B_j \right\}, \]
where \( q^d = q^d (1 - \theta), \bar{q}^d = q^d (1 + \theta) \). (2)

The polyhedral uncertainty set is less conservative than the box uncertainty set as it includes the joint constraint \( \sum_{d \in \{d \mid d' \neq j\}} Q^{d'} \leq B_j \) (Ben-Tal et al., 2011), which is more realistic as it limits the total travelers heading for the same destination \( j \) by an upper bound \( B_j \). For instance, the total demand arriving at a city center is limited by the amount of jobs, amount of retail shops or parking, etc. The essence of robust optimization is to find a sub-optimal solution for the RTL alignment as well as DAR services deployment such that the solutions are feasible even for the worst case scenario.

2.2. Problem setting

Let \( G(N, A) \) be the candidate transportation network with a node set \( N \) and an arc set \( A \). \( A \) is the set of feasible arcs \((i, j) \in A \) linking stations \( i \) and \( j \) for \( i, j \in N, i \neq j \). Each feasible arc \((i, j) \) is associated with a link distance \( l_{ij} \). DAR services use the same road network as that of RTL. To integrate the two services into one multimodal network, we segregate each node into three sub-nodes, namely, station sub-node, RTL sub-node and DAR sub-node, with their corresponding set denoted as \( N_{ST}, N_{RTL}, N_{DAR} \). These three types of sub nodes are interconnected with each other, allowing for passenger boarding, alighting and transferring between the two modes. OD demands are loaded or unloaded from the station sub-nodes. The RTL (DAR) sub nodes are connected by RTL (DAR) arcs. The feasible arc set \( A \) can thus be separated into two subsets accordingly, RTL arc set \( A_{RTL} \) and DAR arc set \( A_{DAR} \).
Fig. 2 illustrates the multi-modal network with a three station network. $A_0, B_0, C_0$ represent station sub-nodes; $A_1, B_1, C_1$ represent RTL sub-nodes; and $A_2, B_2, C_2$ represent DAR sub-nodes. The purple (orange) arcs connecting $A_i, B_i, C_i$ ($A_2, B_2, C_2$) are RTL (DAR) arcs, and the flows on them represent the amounts of passengers taking RTL (DAR) services. The problem is to determine a set of RTL routes $r \in R$ and their corresponding frequencies $f_r \in f$ as well as the deployment of DAR services to meet the stochastic OD demands so that the total expected cost is minimized. The maximum number of lines $\max R$ is predefined. To select the origin and destination station for each transit route flexibly, we introduce a dummy starting node set $\{rS\}$ and ending node set $\{rT\}$ so that every RTL route $r$ has a fixed dummy origin $rS$ and destination $rT$. The exposition of this network structure formulation can be found in An and Lo (2014a). For brevity and page limitation here, we skip the details and only mention the assumptions for this multi-modal problem:

(a) Each route serves both directions with the same frequency.
(b) DAR services operate on the existing road network, with congestion modeled by the BPR function.
(c) Passengers can transfer between these two modes.

Assumption (a) is a common practice. (b) is how DAR services operate, serving as shuttles to carry the demands not covered by RTL. Meanwhile, they can also serve as another mode for certain congested RTL sections to mitigate crowdedness or as additional segment capacity. The road network for DAR is the same as the candidate network for RTL. The impact of private vehicles on the road network is considered directly through deducting road link capacity by their background traffic flow. The UE principle is upheld while including the congestion of DAR services. While (c) may be a simplification, we can impose a sufficient large penalty for passenger transfers between RTL and DAR so as to limit the number of transfers as in Lo et al (2003, 2004).

The decision variables are as follows. A binary variable set $Y = \{Y_{ij}^r\}, ij \in A_{\text{RTL}}$ denotes whether a link is on a RTL. $Y_{ij}$ is 1 if link $(i, j)$ is on line $r$; 0 otherwise. $f = \{f_r\}$ stands for the RTL route frequency. The binary variable set $W = \{W_i^r\}, i \in N_{\text{RTL}}$ indicates whether station $i$ is on a RTL. $X_{ij}^d, ij \in A_{\text{RTL}}$, represents the passenger flow on RTL from station $i$ to $j$ for OD pair $d$ whose set is denoted as $X$. $Z_{ij}^d, ij \in A_{\text{DAR}}$, represents the passenger flow on DAR from station $i$ to $j$ for OD pair $d$, whose set is denoted as $Z$. $V_i$ stands for the amount of passengers transferring from RTL to DAR at station $i$, whereas $V_i$ stands for the amount of passengers transferring from DAR to RTL. The transfer passenger flow set is denoted by $V$. Note that the first two decision variables have directions. The problem is to determine each transit line for the forward direction, with services for the backward direction included automatically.

2.3. A robust formulation with equilibrium constraints

The goal of the study is to minimize the system cost under the worst case scenario through locating the RTL stations, deciding the line frequencies as well as introducing DAR services as needed. To the company, these two services have different unit costs. To passengers, crowding discomfort on RTL and road congestion of DAR services are considered simultaneously to achieve a UE passenger distribution between these two modes. The network design problem with stochastic demand under UE can be formulated as a robust bi-level program, where the network design variables $\{Y, f, W, Z\}$ are determined in the upper level, and the UE passenger flows $\{X, V\}$ in the lower level. The UE principle can be represented by variational inequality (VI) constraints to reduce the bi-level problem into a single level problem with equilibrium constraints. We formulate the robust TNDP with equilibrium constraints.
in P1. Let $\overline{N} = N_{RTL} \cup S \cup T$ be the set of all RTL nodes including the dummy origin and destination nodes and $\overline{A}$ be the set of all RTL links including the dummy arcs. Since passenger flows on RTL, DAR or transfer links are represented by different variables $X^d_{ij}, Z^d_{ij}, V^d_{ij}, V^d_{i}$, it is redundant to specify the sub-node set that the subscripts of $X^d_{ij}, Z^d_{ij}, V^d_{ij}, V^d_{i}$ belong to. Namely, for $X^d_{ij}$ we must have $i, j \in N_{RTL}$ and for $Z^d_{ij}$, we must have $i, j \in N_{DAR}$. To simplify the notation, we simply use $N$ to represent the station index without specifying the specific mode the flow variables $X^d_{ij}, Z^d_{ij}, V^d_{ij}, V^d_{i}$ belong to.

\[
\begin{align*}
\min_{t, y, w, x, z, v} & \sum_{i \in N_{RTL}} \sum_{j \in N_{RTL}} \left(c^1 I_{ij} f X^d_{ij} + c^2 I_{ij} Y^d_{ij} + c^3 \sum_{d \in D} l_{ij} Z^d_{ij}ight) + c \sum_{i \in N_{RTL}} \sum_{j \in N_{RTL}} w_{ij} + c \sum_{d \in D} \sum_{i \in N_{RTL}} l_{ij} Z^d_{ij} \\
& + c^5 \sum_{d \in D} \sum_{i \in N_{RTL}} t_{ij} X^d_{ij} + c^5 \sum_{d \in D} \sum_{i \in N_{RTL}} t_{ij} Z^d_{ij} + c^5 \sum_{d \in D} \sum_{i \in N_{RTL}} t^0 \left(V^d_{ij} + V^d_{i} \right)
\end{align*}
\]  

(P1) 

\[
\begin{align*}
\text{s.t.} & \quad \sum_{i \in N_{RTL}} \sum_{j \in N_{RTL}} Y^d_{ij} \leq 1, \quad \forall i \in N_{RTL} \cup S \\
& \quad \sum_{i \in N_{RTL}} \sum_{j \in N_{RTL}} Y^d_{ij} \leq 1, \quad \forall i \in N_{RTL} \cup T \\
& \quad \sum_{i \in N_{RTL}} Y^d_{ij} = \sum_{j \in N_{RTL}} Y^d_{ji}, \quad \forall r \in R, \forall k \in N_{RTL} \\
& \quad \sum_{i \in N_{RTL}} \left(Y^d_{ij} + Y^d_{ji}\right) \leq 1, \quad \forall ij \in A_{RTL} \\
& \quad f_{\text{min}} \leq f_t \leq f_{\text{max}}, \quad \forall r \in R \\
& \quad Y^d_{ij} = 0 \text{ or } 1, \quad \forall r \in R, \forall ij \in A_{RTL} \\
& \quad W_i - \frac{1}{2R_{\text{max}}} \left(\sum_{i \in N_{RTL}} \sum_{j \in N_{RTL}} Y^d_{ij} + \sum_{i \in N_{RTL}} \sum_{j \in N_{RTL}} Y^d_{ij}\right) \geq 0, \quad W_i = 0 \text{ or } 1, \quad \forall i \in N_{RTL} \\
& \quad \sum_{i \in N_{RTL}} X^d_{ij} + \sum_{j \in N_{RTL}} Z^d_{ij} \geq Q^d, \quad \text{if } k = O^d \\
& \quad \sum_{i \in N_{RTL}} X^d_{ij} + \sum_{j \in N_{RTL}} Z^d_{ij} \geq Q^d, \quad \text{if } k = D^d \\
& \quad \sum_{i \in N_{RTL}} X^d_{ij} - \sum_{j \in N_{RTL}} X^d_{kj} + V^d_{k} - V^d_{k} = 0, \quad k \neq O^d \text{ or } D^d, \quad \forall k \in N, d \in D, Q^d \in U^d \\
& \quad \sum_{i \in N_{RTL}} Z^d_{ij} - \sum_{j \in N_{RTL}} Z^d_{kj} + V^d_{k} - V^d_{k} = 0, \quad k \neq O^d \text{ or } D^d
\end{align*}
\]  

(3) 

\[
\begin{align*}
& t_{ij} = t_0^{I} \left(1 + 0.1 \left(\frac{\sum_{d \in D} X^d_{ij}}{f_{X^d_{ij} + E}}\right)^2 \right), ij \in A_{RTL} \\
& t_{ij} = t_0^{I} \left(1 + 0.15 \left(\frac{\sum_{d \in D} Z^d_{ij}}{C_{Z^d_{ij}}}\right)^4 \right), ij \in A_{DAR}
\end{align*}
\]  

(4) 

\[
\sum_{i \in N_{RTL}} t_{ij}^{*} (x_{ij} - \tilde{x}_{ij}) + \sum_{i \in N_{RTL}} t_{ij}^{*} (z_{ij} - \tilde{z}_{ij}) + \sum_{i \in N_{RTL}} t_{ij}^{*} (v_{ij} - \tilde{v}_{ij}) \leq 0, \quad \forall (\tilde{x}_{ij}, \tilde{z}_{ij}, \tilde{v}_{ij}) \in \Omega
\]  

(5) 

\[
\begin{align*}
X_{ij} &= \sum_{d \in D} X^d_{ij}, \quad z_{ij} = \sum_{d \in D} Z^d_{ij}, \quad v_{ij} = \sum_{d \in D} \left(V^d_{ij} + V^d_{i}\right)
\end{align*}
\]  

(6) 

$c^1, c^2, c^3, c^4, c^5$ are, respectively, the coefficients for RTL operating cost, RTL construction cost, station construction cost, DAR operating cost and passenger value of time. $t^0$ denotes the transfer penalty which is a constant. The RTL link operating and construction costs are proportional to the link distance $l_{ij}$. The objective
function is to minimize the combined RTL operating cost, RTL construction cost, RTL station construction cost, DAR operating cost and passenger cost in order to serve the random demand $Q^d \in U^d$. The RTL connectivity is represented by constraints (4)-(10). Constraints (4) and (5) indicate that only one RTL sub-node is directly connected to RTL sub-node $i$ from upstream and downstream, respectively, if $i$ is on route $r$. Constraints (4) and (5) also ensure that at most one RTL link can be generated from the dummy origin $S_i$ and ended at destination $T_j$. Constraint (6) states that there are exactly two RTL links connecting each RTL node on route $r$. Constraint (7) represents that one link can be occupied by at most one transit line. Constraint (8) sets the frequency boundaries. Constraint (10) ensures that a station is constructed if any line passes through it. The first summation is the total number of outgoing lines of station $i$ and the second summation is the number of incoming lines. The expression in the parentheses calculates the total number of links traversing station $i$, which is no more than two times the maximum route number $R_{\text{max}}$, $W_i = 1$ when there is a line passing through station $i$, and 0 otherwise. Constraints (11)-(14) represent the passenger flow balancing conditions. The demand $Q^d$ in (11)-(14) is stochastic and bounded by the uncertainty set $U^d$. $Q^d \in U^d$ requires that the optimal solution of $\text{P1}$ must be feasible for any demand realization in $U^d$, which leads to a min-max problem, i.e. the worst case or maximum demand scenario within the uncertainty set. $O^d$ and $D^d$, respectively, represent the origin or destination node index of OD pair $d$. To accommodate the approach of robust formulation, the equality constraints for origin and destination nodes are replaced by inequalities. It is easy to see it will be pushed to equality as a consequence of the optimization. The inequality (11) states that the total amount of passengers flowing out from station $k$ either by RTL or DAR is greater than the demand $Q^d$ if $k$ is the origin of OD pair $d$. The second inequality (12) follows the same logic for destination nodes. Constraints (13) and (14) are the flow conservation constraint for RTL and DAR sub nodes, respectively. (15)-(18) are standard VI constraints to achieve UE. The passenger cost on RTL is modeled as a non-linear function (15), where $t_{ij}^d, t_{ij}, \xi_i, \epsilon_i$, respectively, are the free flow travel time, actual in vehicle time on RTL, vehicle capacity and a sufficient small positive number to avoid the case of infeasibility when $Y_{ij} = 0$. The passenger cost on DAR services is represented by the BPR function (16), where $t_{ij}^d, t_{ij}, C_{ij}$ are, respectively, free flow travel time, actual in vehicle time on DAR and road link capacity. In the VI constraints, $t_{ij}$, the link travel times on RTL and DAR are, respectively, calculated by (15) and (16). Link flow is calculated by adding up flows from all OD pairs $d$ as shown in (18). $x_{ij}^d, z_{ij}^d, \nu_i$ are feasible link flows on RTL, DAR, and transfer links, respectively, with their feasible region denoted by $\Omega$. The VI constraints state that for any feasible link flow $\left( x_{ij}^d, z_{ij}^d, \nu_i \right) \in \Omega$, the optimal link flow solution $\left( x_{ij}^d, z_{ij}^d, \nu_i \right)$ must satisfy (17). We note that $\Omega$ is shaped by linear constraints (11)-(14) with stochastic demand $Q^d$, which renders $\Omega$ a polyhedron without simple, explicit boundaries. Namely, it is difficult to determine the boundaries and extreme points of $\Omega$. To deal with this challenge, we reformulate constraints (11) and (12) associated with the demand uncertainty set into a set of linear constraints with deterministic parameters as described in the next section.

### 2.3.1. Linearization of stochastic demand constraints

We note that $\text{P1}$ is a non-linear model with stochastic demand. It can be reformulated as a two-stage stochastic program with complementary constraints to minimize the expected cost as described in Lo et al. (2013) and An and Lo (2014a, 2014b). However, the problem size depends on the sample size needed to conduct the stage-two cost expectation, which limits its application for large networks. In this study, we turn to robust optimization which focuses on the worst case scenario. For stochastic demand described by a polyhedral uncertainty set as in (2), $\text{P1}$ can be linearized by reformulating the constraints related to the stochastic demand as a LP via its dual problem. The worst case scenario (highest demand combination) associated with constraints (11) and (12) are equivalent to find the maximum $Q^d$ in $U^d$, which still satisfies (11) and (12):

$$\max_{Q^d} Q^d, \quad \text{s.t. } Q^d \leq q^d, -Q^d \leq -q^d, \quad \sum_{\forall d \in \left[ D^d - j \right]} Q^d \leq B_j$$

(19)

We rewrite the polyhedral uncertainty set $U^d$, i.e. the constraints in (19), as $A\hat{Q}^d \leq b$ for simplicity, where $\hat{Q}^d = \left\{ Q^d, \forall d \in \left\{ d : D^d = j \right\} \right\}$ is the vector of $Q^d$ involved in $U^d$; $A$ and $b$ are, respectively, the parameter matrix...
and the RHS vector of the linear constraints. An equivalent constraint can be obtained by its dual problem (Ben-Tal et al., 2011, Bertsimas and Sim, 2004).

\[
\max_{Q^d} \quad s.t. \quad A Q^d \leq b \Leftrightarrow \min_{\lambda^d} b^T \lambda^d \quad s.t. \quad A^T \lambda^d = 1, \quad \lambda^d \geq 0
\]  

(20)

where \( \lambda^d = \{\lambda^d_1, \lambda^d_2, \lambda^d_3, \lambda^d_{op}\} \) is the vector of dual variables corresponding to the three constraints in (19), and \( D^d \) is the destination node index of OD pair \( d \). The constraint objective of (11) changes from finding the maximum \( Q^d \) that is less than \( \sum_{k \in N} x^{d}_{ik} + \sum_{j \in N} z^{d}_{kj}, \) if \( k = O^d \) to finding the minimum \( b^T \lambda^d \) that is less than \( \sum_{j \in N} x^{d}_{kj} + \sum_{i \in N} z^{d}_{ik}, \) if \( k = O^d \). Constraint (12) follows the same logic. It enables us to directly add the dual problem to \( P1 \) as constraints. By applying this method, constraints (11) and (12) are replaced by:

\[
\lambda^d_q - \lambda^d_{z} q^d + \lambda^d_{op} B_{y^d} \leq \sum_{j \in N} x^{d}_{y^d j} + \sum_{j \in N} Z^{d}_{y^d j}, \quad k \text{ is the origin of OD } d, \forall d
\]  

(21)

\[
\lambda^d_q - \lambda^d_{z} q^d + \lambda^d_{op} B_{y^d} \leq \sum_{i \in N} x^{d}_{h^d i} + \sum_{i \in N} Z^{d}_{h^d i}, \quad k \text{ is the destination of OD } d, \forall d
\]  

(22)

\[
\lambda_{q}^d - \lambda_2^d q^d + \lambda_{op}^d B_{y^d} = 1, \forall d
\]  

(23)

\[
\lambda_1^d, \lambda_2^d, \lambda_{op}^d \geq 0, \forall d
\]  

(24)

After the linearization, the feasible passenger flow set \( \Omega \) is shaped by a set of linear constraints (13), (14) and (21)-(24). Now \( \Omega \) is a bounded polyhedron with finite vertexes. This attribute of \( \Omega \) will assist us in developing efficient solution algorithm.

2.3.2. Linearization of the VI constraint

In the VI constraint, we seek to find passenger link flow \( x_{ij}, z_{ij}, v_{ij} \) so that for any feasible flow \( (\tilde{x}_{ij}, \tilde{z}_{ij}, \tilde{v}_{ij}) \in \Omega \), (17) is satisfied. However, it is computationally formidable to enumerate all the feasible flows in \( \Omega \). Since \( \Omega \) is a bounded polyhedron, the feasible flows can be calculated by the convex combination of the vertices or extreme points of \( \Omega \). Take \( \tilde{x}_{ij} \) for example, \( \tilde{x}_{ij} = \sum_{e} \eta_{e} \tilde{x}_{ij,e}, \sum_{e} \eta_{e} = 1, 0 \leq \eta_{e} \leq 1 \), where \( \tilde{x}_{ij,e} \) represents the \( e \)-th extreme point or vertex of \( \Omega \). Let \( z_{ij,e} \) and \( \tilde{v}_{ij} \) be defined in the same way and the number of extreme points be denoted by \( E \). For RTL passenger flow, we have:

\[
t_{ij}*(x_{ij} - \tilde{x}_{ij}) \Leftrightarrow t_{ij}*(x_{ij} - \sum_{e} \eta_{e} \tilde{x}_{ij,e}) \Leftrightarrow t_{ij}*(\sum_{e} \eta_{e} x_{ij} - \sum_{e} \eta_{e} \tilde{x}_{ij,e}) \Leftrightarrow \sum_{e} \eta_{e} \left( t_{ij}*(x_{ij} - \tilde{x}_{ij,e}) \right)
\]  

(25)

The VI constraint (17) is reformulated as:

\[
\sum_{e} \eta_{e} \left( \sum_{q \in 4_{IN}} t_{ij}*(x_{ij} - \tilde{x}_{ij,e}) + \sum_{q \in 4_{IN}} t_{ij}*(z_{ij} - \tilde{z}_{ij,e}) + \sum_{i \in N} t_{ij}^0*(v_{ij} - \tilde{v}_{ij,e}) \right) \leq 0
\]  

(26)

It is obvious that if the main bracket on the LHS of (26) is less than or equal to zero for any \( e = 1...E \), then (26) must hold. (17) is equivalent to the following constraints:

\[
\sum_{q \in 4_{IN}} t_{ij}*(x_{ij} - \tilde{x}_{ij,e}) + \sum_{q \in 4_{IN}} t_{ij}*(z_{ij} - \tilde{z}_{ij,e}) + \sum_{i \in N} t_{ij}^0*(v_{ij} - \tilde{v}_{ij,e}) \leq 0, \quad \forall e = 1,...,E
\]  

(27)

After the reformulation, we only need to ensure the feasibility of (27) for all extreme points of \( \Omega \) instead of all the points in the whole region of \( \Omega \), which substantially reduce the actual number of constraints to be added to
problem P1. Note that P1 is a mixed integer non-linear program, with all its nonlinear terms involved in the objective function and the VI constraints. The following section describes the procedure to linearize the nonlinear terms.

2.3.2.1. Linearization of \( f, Y^r \)

A real variable \( \check{y}_y \) is introduced to replace the product of frequency \( f \) and RTL construction variable \( Y^r \), i.e. \( \check{y}_y = f \cdot Y^r \). \( \check{y}_y \) can be interpreted as the RTL link frequency which is 0 when the link is not covered by RTL and equal to \( f \) when a line \( r \) traverses it. A set of mixed integer linear constraints are employed to realize the transformation.

\[
\begin{align*}
0 \leq \check{y}_y - \check{y}_y - \check{y}_y \leq 0, & \quad \forall \check{y}_y \leq 0 \leq \check{y}_y \\
\end{align*}
\]

where \( \check{y}_y \) is an extremely large positive number (28).

2.3.2.2. Linearization of \( t_y \) and \( t_y \) on dial-a-ride services, \( ij \in A_{DAR} \)

A continuous real variable \( \check{t}_y \) is introduced to represent the total passenger travel time on link \( ij \) using DAR services, i.e., \( \check{t}_y = t_y \cdot z_y \). The link travel time \( \check{t}_y \) and total travel time \( \check{t}_y \) only depend on link flow \( z_y \). We adopt a piecewise linear function to approximate the nonlinear functions of \( t_y \) and \( t_y \). The idea is to partition the passenger flow \( z_y \) into \( m \) segments first. The passenger flows at the breaking points are denoted as \( \kappa_m^0, \kappa_m^m \). The link travel time \( \check{t}_y \) and total link travel time \( \check{t}_y \) at breaking points are obtained through plugging \( \kappa_m^0 \) into the corresponding travel time functions. The are between two break points is approximated by a straight line connecting the two adjacent breaking points, with the slope of \( \check{t}_y - \check{t}_y \) for \( t_y \) function and \( \check{t}_y - \check{t}_y \) for \( t_y \) function. Now we are ready to formulate the piecewise linear functions for each link \( ij \in A_{DAR} \):

\[
\begin{align*}
t_y &= t_y + \sum_{m=1}^{M} \frac{t_y - t_y}{z_y - z_y} \mu_y^m, & \check{t}_y &= \frac{t_y}{z_y} + \sum_{m=1}^{M} \frac{t_y - t_y}{z_y - z_y} \mu_y^m \\
0 \leq \nu_y \leq \nu_y \leq \nu_y, \quad \forall m = 1, ..., M ; & \kappa_m^0 = 1, \kappa_m^m = 0, \kappa_y^m \in \{0,1\}, \forall m = 1, ..., M - 1
\end{align*}
\]

\( \mu_y^m \) is the length of the segment \( m \) covered by \( z_y \). For instance, if \( \check{t}_y \) falls in the \( k \)th segment, \( \mu_y^m \) is equal to the length of the segment for \( 1 \leq m \leq k - 1 \), i.e. \( \mu_y^m = z_y^m - z_y^{m-1} \), and is less than the length of the last segment, i.e. \( \mu_y^m \leq z_y^m - z_y^{m-1} \) for \( m = k \). For \( k + 1 \leq m \leq M \), \( \mu_y^m = 0 \). This segment scheme is ensured by constraints (30)-(32).

The piecewise linear functions for \( t_y \) and \( \check{t}_y \) are represented by (29).

2.3.2.3. Linearization of \( t_y \) and \( t_y \) on rapid transit services, \( ij \in A_{RTL} \)

The RTL link travel time function is more complicated than that of DAR services since it involves two variables, link capacity \( \sum f \cdot Y^r + \xi \) and link flow \( x_y \). This two-dimensional function requires a different approximation method. We make use of the same piecewise linear approximation method for multi-dimensional functions as described in Luathep et al. (2011). Similarly, the total travel time on link \( ij \) through rapid transit services is denoted by \( \check{t}_y = t_y \cdot x_y \). The link capacity is denoted by \( \check{Y}_y = \sum f \cdot Y^r + \xi \cdot x_y \) is segmented into \( M \) intervals while \( Y^r \) is segmented into \( N \) intervals. They partition the domain into \( M \times N \) rectangles. Each rectangle can be further separated into two triangles by the upward diagonal line. The feasible domain of \( t_y \) and \( \check{t}_y \) is thus partitioned into a set of triangles as illustrated in Fig. 3.
Fig. 3. Piecewise linear approximation for two-dimensional functions

The key issue is to determine the active triangle that \((x, y)\) falls into. Let \((x^m, \tilde{y}^m, \tilde{x}^n, \tilde{y}^n)\) be the coordinate of an arbitrary corner point \((m, n)\), \(0 \leq m \leq M, 0 \leq n \leq N\). \(t\) and \(\tilde{t}\) are represented by the convex combination of the corner point coordinates of the active triangle. Special ordered sets (SOS) variables are introduced in constraints (35)-(41) to identify the active triangle that \((x, y)\) belongs to. \(\{\alpha^m\}, \forall m = 0, \ldots, M\text{ and }\{\beta^n\}, \forall n = 0, \ldots, N\) are Special Ordered Set of type One (SOS1) variables, which require that at most one member from the set may be non-zero. SOS1 variable is to represent a set of mutually exclusive alternatives. \(\{\gamma^\tau\}, \forall \tau = 0, \ldots, M, M+1, \ldots, M+N\) is Special Ordered Set of type Two (SOS2) variables, which requires that at most two adjacent members from the set are non-zeroes.

\[
x_y = \sum_{m=0}^{M} \sum_{n=0}^{N} \delta^{m,n}_{y} x^m_y, \quad \tilde{y}_y = \sum_{m=0}^{M} \sum_{n=0}^{N} \delta^{m,n}_{y} y^n_y
\]

\[
t_y = \sum_{m=0}^{M} \sum_{n=0}^{N} \delta^{m,n}_{y} t^m_y, \quad \tilde{t}_y = \sum_{m=0}^{M} \sum_{n=0}^{N} \delta^{m,n}_{y} \tilde{t}^n_y
\]

\[
\sum_{m=0}^{M} \sum_{n=0}^{N} \delta^{m,n}_{y} = 1, \quad \delta^{m,n}_{y} \in [0,1], \forall m = 0, \ldots, M, \forall n = 0, \ldots, N
\]

\[
\sum_{m=0}^{M} \sum_{n=0}^{N} \delta^{m,n}_{y} \leq \alpha^m_y + \alpha^{m+1}_y, \quad \forall m = 0, \ldots, M
\]

\[
\alpha^0_y = \alpha^{M+1}_y = 0, \quad \alpha^m_y \in [0,1], \forall m = 1, \ldots, M
\]

\[
\sum_{m=0}^{M} \sum_{n=0}^{N} \delta^{m,n}_{y} \leq \beta^n_y + \beta^{n+1}_y, \quad \forall n = 0, \ldots, N
\]

\[
\beta^0_y = \beta^{N+1}_y = 0, \quad \beta^n_y \in [0,1], \forall n = 1, \ldots, N
\]

\[
\gamma^\tau = \sum_{m=0}^{M} \sum_{n=0}^{N} \delta^{m,n}_{y} \gamma^{m,n+\tau}_y, \quad \gamma^\tau \in SOS2, \forall \tau = 0, \ldots, M, M+1, \ldots, M+N, \max \{0, M - \tau\} \leq m \leq \min \{N+M - \tau, M\}
\]

\[
\sum_{\tau=0}^{M+N} \gamma^\tau = 1
\]

**Proposition 1:** A feasible solution for the SOS1 and SOS2 variables identifies a unique triangle in the domain of \((x, y)\).

Proof: Without loss of generality, \(\alpha^1_y, \beta^2_y\) are assumed to be the positive elements in SOS1 set of variables. Substitute the values of SOS1 variables into constraints (36) and (38), we obtain

\[
\sum_{n=0}^{N} \delta^{m,n}_{y} = 0, \quad \forall m = 2, 3, \ldots, M, \sum_{n=0}^{N} \delta^{m,n}_{y} \leq \alpha^1_y, \sum_{n=0}^{N} \delta^{m,n}_{y} \leq \alpha^1_y
\]
Combining (42) and (43), we can get
\[
\delta_{ij}^{0,1} + \delta_{ij}^{0,2} \leq \alpha_{ij}^{1}, \quad \delta_{ij}^{1,1} + \delta_{ij}^{1,2} \leq \alpha_{ij}^{2}, \quad \delta_{ij}^{0,1} + \delta_{ij}^{1,1} \leq \beta_{ij}^{1}, \quad \delta_{ij}^{0,2} + \delta_{ij}^{1,2} \leq \beta_{ij}^{2}
\]  
(44)

All the other \(\delta_{ij}^{m,n}\) which are not stated in (44) are zero.

\(\delta_{ij}^{m,n}\) takes on a positive value only at the corner point of the rectangle \(\begin{pmatrix} \delta_{ij}^{0,1} & \delta_{ij}^{0,2} \\ \delta_{ij}^{1,1} & \delta_{ij}^{1,2} \end{pmatrix}\) as shown in Fig. 3.

The active rectangle is thus determined by the two sets of SOS1 variables. The next question is whether the SOS2 variable \(\gamma_{ij}^{r}\) can identify the active triangle. In constraint (40), \(\gamma_{ij}^{r}\) is defined as the sum of \(\delta_{ij}^{m,n}\) along each diagonal line. Hence the nonzero \(\gamma_{ij}^{r}\) can only occur at the three possible diagonal lines, \(\tau = M, \tau = M + 1, \tau = M + 2\), passing through points \((1,1), (0,1)\) & \((1,2)\), and \((0,2)\), respectively, as shown in Fig. 3. It is easy to show that the two feasible solutions \(\gamma_{ij}^{r}, \gamma_{ij}^{r+1} > 0\), or \(\gamma_{ij}^{r}, \gamma_{ij}^{r+1} > 0\) identify the upper and lower triangle, respectively. This finishes the proof. Reversely speaking, to determine the target point as shown in Fig. 3, the positive elements in SOS1 and SOS2 variables can only be \(\gamma_{ij}^{r}, \gamma_{ij}^{r+1}\), \(\gamma_{ij}^{r}, \gamma_{ij}^{r+1}\) while all the other elements are zero.

After the reformulation, the original robust optimization problem with equilibrium constraints is reduced into a mixed integer linear problem (MILP) which can be readily solved. The equivalent MILP is summarized as follows:

\[
(P2) \quad \min_{t, Y, W, X, Z, V} \sum_{r \in R} \sum_{y \in A_{RT}} \left( c_1 t_y Y_r + c_2 l_y Y_r \right) + c_3 \sum_{i \in N_{RT}} W_i + c_4 \sum_{y \in A_{RT}} l_y z_y + c_5 \sum_{\tilde{t} \in \tilde{t}} \tilde{t} + c_5 \sum_{\tilde{t} \in \tilde{t}} \tilde{t} + c_5 \sum_{l \in L} l_v
\]

s.t.  (46)-(49) and

\[
\sum_{y \in A_{RT}} \left( I_y - t_y \right) \leq 0, \quad \forall e = 1, \ldots, E
\]

Denote \(f^{*}, Y^{*}, W^{*}, t^{*}, \tilde{t}^{*}, x^{*}, z^{*}, v^{*}\) as the optimal solution of the reduced problem \(P3\) with a smaller number of
extreme points. Additional extreme points can be found by identifying any feasible solution \( \{x_{ij}, z_{ij}, v_i\} \in \Omega \) that satisfies \( \sum_{j \in A_{ij}} (t_{ij} - t_{ij}^* x_{ij}) + \sum_{j \in A_{ij}}(t_{ij} - t_{ij}^* z_{ij}) + \sum_{i \in N} t^* (v_i^* - v_i) > 0 \) (Luathep et al., 2011). Adding this extreme point into \( P3 \) will make the current optimal solution infeasible, which thus leads to a new solution. An equivalent optimization problem \( P4 \) is formulated to find the extreme points:

\[
(P4) \quad \max_{x, z, v} F = \sum_{ij \in A_{ij}} (t_{ij} - t_{ij}^* x_{ij}) + \sum_{ij \in A_{ij}} (t_{ij} - t_{ij}^* z_{ij}) + \sum_{i \in N} t^* (v_i^* - v_i)
\]

s.t. (47)

\( (53) \)

If \( F > 0 \), its optimal solution \( \{\tilde{x}_{ij}, \tilde{z}_{ij}, \tilde{v}_i\} \) will formulate a new VI constraint (50) which is then added into \( P3 \). Otherwise we can claim that the global optimal solution for \( P2 \) has been found.

After solving \( P2 \), the line alignment and frequency of rapid transit lines \( f', Y', W' \) are fixed for the whole studying horizon (say, a year). Moreover, \( P2 \) also calculates the DAR services needed under the worst-case (or highest) demand scenario. The exact deployment of dial-a-ride services for a particular day will depend on the demand realization. Meanwhile, the passenger cost under UE will change with demand as well. We calculate the average DAR operating and passenger costs by drawing samples of the uncertain demand. With the RTL capacity fixed by \( f', Y', W' \), for a specific demand realization, \( P2 \) is reduced to a traditional UE traffic assignment problem. A variety of efficient solution approaches, such as Frank-Wolfe algorithm, Gradient Projection algorithm, etc. (Chen et al., 2002) can be applied. We adopt the Frank-Wolfe algorithm in this paper. The procedure is summarized as follows:

**Step 0.** Define the boundaries of the study area, location of CBD, initial population density in CBD, and population density evolutionary pattern over time.

**Step 1.** Determine the population in the catchment area of each candidate transit station. For a specific year, the density distribution pattern is determined by \( \Lambda_{o} = \Lambda_{o} e^{-\omega o} \). Given a candidate transit network topology (say, a linear network with a certain number of nodes), the study area could be partitioned into a set of disjoint segments according to the walking distance to the candidate transit station. Each segment is defined as the catchment area of the candidate station and the total population resided in the catchment area can be calculated by integrating \( \Lambda_{o} \) over distance \( \omega \).

**Step 2.** Determine the OD demand matrix. The total trip amount \( P_{ij} \) generated in the predefined catchment area is calculated by integrating the product function of population and trip generation rate over distance \( \omega \) as shown in (1). Trip distribution is conducted by the gravity model \( q_{ij} = q_{ij} = P_{ij} \left( I_{ij} F_{ij} / \sum_{o} I_{o} F_{io} \right) \) to obtain the expected OD demand matrix.

**Step 3.** Determine the uncertainty level \( \theta \) of stochastic demand \( Q' \) and formulate the MILP problem \( P3 \) without the VI constraint (52). Namely, the system optimal solution is adopted as the initial solution.

**Step 4.** Solve the relaxed MILP \( P3 \) with a reduced set of extreme points. The optimal solution is denoted as \( f', Y', W', t', x', z', v' \).

**Step 5.** Solve the linear program (LP) problem \( P4 \) for a new extreme point \( \{\tilde{x}_{ij}, \tilde{z}_{ij}, \tilde{v}_i\} \in \Omega \).

**Step 6.** Convergence check. If \( F \leq \varepsilon \), terminate the procedure and the optimal solution under UE flows are maintained as \( f', Y', W', t', x', z', v' \). Otherwise, add \( \{\tilde{x}_{ij}, \tilde{z}_{ij}, \tilde{v}_i\} \) into \( P3 \) through the VI constraint (52) and repeat Step 4 until the convergence criteria is satisfied.

**Step 7.** Calculate the average passenger cost under UE by sampling the stochastic demand given the RTL network \( f', Y', W' \).

**Step 8.** Decrease \( b \) to simulate the population density increase over time and repeat Steps 2-7 to find the critical population density for the first financially sustainable RTL and its construction sequence over time.

3. Case study

We apply the model formulation and solution algorithm to an eight-node linear network to illustrate its properties.
and performance, as shown in Fig. 4. The density $\Lambda_0$ at the city center CBD is 1000 people per km$^2$. The candidate stations are uniformly distributed along the corridor with a distance of 1 km between two adjacent stations. There are 8*8 OD pairs in total. The inter-zonal OD trip distribution is conducted through the gravity model. In addition, we assume that the inner zonal demand can walk to their destinations and thus would not contribute to road congestion. The trip generation rate is defined as $a_i = 1 - 0.02 \left( \omega - \omega_1 \right)$, $\omega_1 = 5 \leq \omega \leq \omega_2 + 5$, in Section 2.1. The robust parameter is $\theta = 0.3$, and the trips heading for destination $j$ is no more than 1.2 times its expectation: $Q^d \in U^d \equiv \{ Q^d : q^d (1 - 0.3) \leq Q^d \leq q^d (1 + 0.3), \sum_{q^d \neq j} Q^d \leq (1 + 0.2) \sum_{q^d \neq j} q^d \}$. The frequency boundaries are $f_{\min} = 20 / \text{hr}$ and $f_{\max} = 3 / \text{hr}$.

The transit unit capacity is $C = 80$ persons/h. The road links have a capacity of $C_g = 200$. At most 1 transit line is allowed in the network, i.e. $R_{\max} = 1$. The unit RTL operating cost per transit unit is $c^1 = 1$; the unit line construction cost is $c^2 = 4$; the unit station construction cost is $c^3 = 1$; the unit DAR operating cost is proportional to the unit RTL operating cost per passenger $c^4 = c^3 / \xi$; the passenger value of time is $c^5 = 0.01$. The transfer penalty is: $t^0 = 3$. In this study, we are interested in the critical population density that the first financially sustainable RTL can be constructed and the RTL construction sequence as the population increases over time. Intuitively, many factors are influential to the RTL construction sequence, such as the population distribution pattern, the number of CBDs, and DAR services cost. In the following, sensitively analysis is employed to investigate the system performance under different combination of the influencing factors.

For a fixed population, planners decide how to spread the population density over the city, and residents decide their housing or activity locations based on their accessibility to work places, and amenities, etc. Hence, there are different population distribution patterns (PDP), such as decreasing density from CBD to the suburb as is commonly observed, uniform distribution pattern, or high density at CBD and in the suburb but low density midway between them as in certain new towns in Hong Kong. We select three representative PDPs, referred to as Type I, II, and III in Fig. 5. The solid lines show the basic shape of the population distribution and the dotted lines represent population density changes over time. The area under each curve is the total population, which can be calculated by integrating the PDP function over the X-axis. For Types I and III, the population density is represented by an exponential function $\Lambda_\omega = \Lambda_0 e^{-\omega}$ with the saturation density $\Lambda_0$ at CBD or suburb. $\omega$ is distance from CBD. For type II, we define $\Lambda_\omega = b * \Lambda_0$, which is uniformly distributed along the city. $b$ represents the degree of saturation.

![Fig. 4. Network setting](image)

![Fig. 5. Population distribution pattern](image)

### 3.1. Illustration of solution procedure

Before we embark on discussing the system performance, to illustrate the formulation and solution procedure, the solutions for the scenario of PDP Type I, with CBD located at $\omega=0$ and a total population of 3000 are presented
below. We first calculate the expected OD demand matrix according to Steps 2 and 3, as shown in Table 1. The last column shows the total trips produced from each node, as calculated by (1). The amount of trip productions $P_i$ exponentially decreases from CBD to the suburb, i.e. from node 1 to node 8. The attractiveness of destination $A_j$ is expressed as an exponentially decreasing function from CBD. Now we are ready to formulate $P_3$ with the parameters and demand information as determined in Step 3. We repeat Steps 4-6 until the UE condition is satisfied. The algorithm took 1 to 9 iterations to converge to the UE solution. The robust solution for the worst case scenario is obtained in Step 6. With the RTL alignment, frequency and the provided DAR services fixed at the worst case scenario, we calculate the expected passenger cost in Step 7. The computational times for different scenarios vary from 9 to 151 seconds, with an average of 56 seconds.

In this solution instance, the RTL services cover links from node 1 to node 5 with a frequency of 4 vehicles per hour. The resultant multi-modal network together with the link flows are shown in Fig. 6. The top nodes and links indicate the RTL services; the middle layer nodes in gray indicate station nodes where demands enter; the bottom layer nodes and links represent DAR services; and the vertical links represent transfer links between RTL and DAR with arrows indicating their directions. The figure on each link stands for the amount of passengers on specific services and the one in parenthesis for its link travel time.

To ascertain that the UE flow pattern is achieved by the cutting constraint algorithm, we depict the passenger assignment result for 6 representative OD pairs under the worst case scenario, which is sufficiently simple to track down the details. Note that the demand for $P_2$ under the worst case scenario shown in the second column of Table 2 is higher than the expected demand in Table 1. Column 3 in Table 2 shows the path and transport mode for each OD pair; the number stands for the node and alphabet for the transport mode, i.e. R for RTL, D for DAR, and T for Transfer. Passengers may choose different paths, and the used paths have essentially the same travel time, with a miniscule difference due to the linearization error. For OD pair 1-2, paths 1R-2R and 1D-2D have the same minimum travel time of 10.1. The unused path has longer travel time, i.e. for OD pair 2-6, the travel time on the unused path 2R-3R-4R-5T-6D is 43.1, higher than 40.1, as consistent with the UE principle. Passengers on OD 5-6 have no choice but to take DAR services since they are the only available services. We observe that most passengers choose direct services; only a few would transfer between the two modes owning to the high transfer penalty, which greatly prohibits their willingness to transfer. Table 3 shows the cost comparison between the worst case scenario and the expected cost. They have identical RTL alignment and frequencies hence have the same RTL cost. The expected cost of DAR services is 21% lower than the cost estimated by the worst case scenario, indicating that the worst case scenario overestimated the cost as expected. We anticipate that this overestimation is larger for a higher demand uncertainty. How to decide the level of demand uncertainty that the RTL services are planned for such that the expected system cost is minimized is an interesting extension to be explored in future studies.

![Fig. 6. RTL alignment, passenger flows and link travel time on RTL, DAR services and transfer links](image-url)
### Table 3 Cost component for solutions of the worst case scenario and the sample average

<table>
<thead>
<tr>
<th>Alignment</th>
<th>Freq.</th>
<th>RTL cost</th>
<th>DAR cost</th>
<th>Passenger cost</th>
<th>Total cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worst Case</td>
<td>1-2-3-4-5</td>
<td>4</td>
<td>565</td>
<td>97</td>
<td>190</td>
</tr>
<tr>
<td>Expectation</td>
<td>1-2-3-4-5</td>
<td>4</td>
<td>565</td>
<td>77</td>
<td>124</td>
</tr>
<tr>
<td>Comparison</td>
<td>--</td>
<td>--</td>
<td>0%</td>
<td>21%</td>
<td>35%</td>
</tr>
</tbody>
</table>

#### 3.2. Sensitivity analysis of population distribution pattern (PDP) with one CBD

In this section, only one CBD located at the origin \( \alpha_{CBD} = 0 \) is considered. The RTL alignment, frequency, and cost components under the three PDPs are shown in Table 4 and Fig. 7. Twenty different population sizes varying from 200 to 8000 are tested to investigate the construction pattern of RTL over time. The step size is 200 for population ranging from 200 to 3000 and is scaled up to 1000 for population from 3000 to 8000. We can observe an obvious inflection point at population of 3000 in Fig. 7, due to the change of step size. Critical Population 1 (CP1) is defined as the population size at which the RTL is introduced for the first time whereas Critical Population 2 (CP2) is defined as the population at which the RTL is constructed to cover the whole city range. CP1 and CP2 are marked by the two arrows respectively in Fig. 7. The first RTL appears at 2400, 1200, 1200, respectively. It indicates that a dispersive population distribution will be more favorable for constructing RTL as compared with concentrating the population at CBD. CP2 for the three PDP types are 6000, 3000, and 2200, respectively. The distance between the two arrows represents the RTL construction time duration. It shows that for a uniformly distributed pattern (PDP Type II), it requires a longer time to provide RTL for the city. On the contrary, a much shorter time is needed for the PDP dispersed to both ends. The blue curve represents the total cost under the worst scenario calculated by \( P_3 \) while the bars stand for the expected system cost calculated in Step 7 of the solution algorithm. The gap can be interpreted as the protection level offered by robust optimization to hedge against the stochastic demand. We can see that the protection level is much higher for a denser city with large population. The RTL constitutes a major cost component once introduced, whereas the DAR only constitutes a small fraction of the cost. The passenger cost linearly increases with population for the three PDP types.

![Graph Type I](image1.png)

![Graph Type II](image2.png)
In Table 4, we select five representative population sizes 600, 1200, 2400, 3000 and 8000 out of 20 scenarios to show the RTL alignment and cost components. The last column states the total cost comparison (%) of the three PDP types with the Type II cost selected as the benchmark. A more negative % indicates a lower cost as compared with the Type II counterpart. Type I outperforms the other types for all populations in terms of total social cost while Type III yields the highest cost. For population less than 600, the costs follow from II>III>I. When population is greater than 600 and less than 800, the rank is III>II>I. It indicates that the uniform population distribution pattern (Type II) is not beneficial to reduce total system cost for low population but will become more cost effective than concentrated development in the suburb for larger populations. The RTL capacity is enlarged to handle the increase in population either through constructing more lines or providing higher frequency. The RTL cost jump happens when a new line is constructed: i.e. for Type I, the RTL cost increases from 263 to 565 when the population changes from 2400 to 3000. In contrast, the increase in frequency does not incur that much increase in RTL cost. i.e., for Type II, the RTL cost increase from 907 to 967 when population changes from 2400 to 3000. It indicates the capital cost for line construction is the highest while operating cost increases due to increase in frequency is relatively lower. Utilization of RTL greatly reduces the use of DAR services and thus mitigates the road congestion. We can also observe that the DAR cost increases when there is no RTL, and drops dramatically when the first RTL is introduced (refer to Type I from the population of 2200 to 2400 in Fig. 7). The RTL cost, passenger cost and total cost increase with the population as expected.

As for the construction sequence, for Type I, the RTL extends gradually from CBD to the suburb; whereas for Type III, a reverse sequence is observed, i.e. from the suburb to CBD; and for Type II, the RTL starts from the middle and then expands to CBD and the suburb gradually. For Type I, the construction sequence follows the same trend as the population density increases, as expected that the RTL would be constructed on the most congested road segment first. When residents are uniformly distributed along the study area, i.e., Type II, road segments in the middle will attract most passengers. This phenomenon is incurred by the exponential distribution of the destination attractiveness to CBD from the suburb. It indicates that the demand generated from nodes 4 to 8 will head for nodes 1-3 while for the demand generated from nodes 1-3 will mostly be absorbed by themselves (intra-zonal demand). Hence most OD demands have to traverse the middle road segments but not the segments at both ends. Under Type III, the high residual population in the suburb will make the road segments away from the CBD congested. Hence the construction will start from the right end nodes 4-8.

Table 4. Line alignment and the corresponding cost component with the change of total population

<table>
<thead>
<tr>
<th>PDP Type</th>
<th>Pop.</th>
<th>RTL Alignment</th>
<th>Freq.</th>
<th>RTL cost</th>
<th>DAR cost</th>
<th>Pass. cost</th>
<th>Total Cost</th>
<th>Comparison (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>600</td>
<td>--</td>
<td>--</td>
<td>25</td>
<td>15</td>
<td>40</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1200</td>
<td>--</td>
<td>--</td>
<td>80</td>
<td>49</td>
<td>129</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2400*</td>
<td>1-2-3</td>
<td>3.0</td>
<td>263</td>
<td>154</td>
<td>211</td>
<td>628</td>
<td>-55</td>
</tr>
<tr>
<td></td>
<td>3000</td>
<td>1-2-3-4-5</td>
<td>4.0</td>
<td>565</td>
<td>77</td>
<td>309</td>
<td>951</td>
<td>-41</td>
</tr>
<tr>
<td></td>
<td>8000</td>
<td>1-2-3-4-5-6-7-8</td>
<td>14.0</td>
<td>1688</td>
<td>86</td>
<td>1579</td>
<td>3353</td>
<td>0</td>
</tr>
</tbody>
</table>
3.3. Sensitivity analysis of population distribution pattern (PDP) with two CBDs

Central business district (CBD) is the commercial heart of a city, typified by a concentration of retail, resident and office buildings. Owing to the fast population growth and economic development, many metropolises face great pressure on providing sufficient housing, employment and medical services in urban areas. In addition, transportation congestion is another impending issue to be addressed. The concept of "satellite town" was thus proposed to avail more lands to satisfy the increasing housing need. Satellite towns are smaller municipalities that are built in the vicinity of a metropolitan area. It is typically self-contained in the sense that the internal employment bases are sufficient to support their residential populations. However, a satellite town may produce some longer commute trips with the CBD. To investigate the benefits of satellite town in mitigating transportation congestion and its influence on population distribution pattern, a satellite town is assumed to be located at the city suburb, i.e., node 8, in this section. It functions as another employment and commercial centre, similar to the CBD. The attractiveness for nodes 5-8 is modified to be symmetric with no des 4-1. The attractiveness around the satellite town should be lower than the CBD in general. In this study, the attractiveness of the CBD and satellite town are set to be identical to amplify the influence of the satellite town on RTL construction and to better contrast with the case of having only one CBD. We thus call this scenario as 2-CBD in the following. Introducing two CBDs substantially alters the RTL construction phases. Table 5 shows the total costs under 9 selected population scenarios for the three PDPs. The last column calculates the average cost across the 9 population scenarios from 200 to 8000 for each PDP. The last two rows show the cost differences (in %) as compared with Type II. We can observe that under most population scenarios, the total cost follows from Type II > I > III. Type III generates the lowest system cost as population are concentrated on the two CBDs where most employment and entertainment are provided. The average cost difference between Type II and III is as high as 52%, indicating that a large amount of resources could be saved if CBD location is properly planned.

<table>
<thead>
<tr>
<th>PDP Type</th>
<th>Pop. 200</th>
<th>1000</th>
<th>2000</th>
<th>3000</th>
<th>4000</th>
<th>5000</th>
<th>6000</th>
<th>7000</th>
<th>8000</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total cost</td>
<td>I</td>
<td>7</td>
<td>175</td>
<td>627</td>
<td>1390</td>
<td>1614</td>
<td>1869</td>
<td>2153</td>
<td>2444</td>
<td>2659</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>51</td>
<td>333</td>
<td>1187</td>
<td>1419</td>
<td>1652</td>
<td>1656</td>
<td>2173</td>
<td>2451</td>
<td>2673</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>0.5</td>
<td>99</td>
<td>345</td>
<td>810</td>
<td>1451</td>
<td>1712</td>
<td>2039</td>
<td>2367</td>
<td>2683</td>
</tr>
<tr>
<td>Cost Diff. w.r.t II (%)</td>
<td>I</td>
<td>-86</td>
<td>-47</td>
<td>-47</td>
<td>-2</td>
<td>-2</td>
<td>13</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>-99</td>
<td>-70</td>
<td>-71</td>
<td>-43</td>
<td>-12</td>
<td>3</td>
<td>-6</td>
<td>-3</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5. Total cost difference between the three PDPs

Table 6. Total cost difference and RTL construction sequence between 1-CBD and 2-CBD scenarios
Table 6 compares the total costs and critical densities between 1-CBD and 2-CBD. Under PDP Type I, the total cost for 2-CBD increases 46% as compared with 1-CBD. Since people are concentrated in the original CBD, it will possibly motivate some people to the other far end CBD with longer distance and travel time. Under Types II and III, the total costs are substantially reduced as most people living in the suburb or the middle of the city are attracted by the other CBD with shorter travel time. In summary, it does not necessarily bring down the total system cost through introducing another commercial centre in the other end of the city. It suggests that only if sufficient employments are provided in a CBD and residential areas are located nearby, the total cost would be reduced. However, this example neglects the land use issue, i.e., the land resources for commercial and residential uses in CBD are limited. In this study, we did not impose any limitations on the residential land use in CBD, which is left for future studies.

According to the last six columns related to the critical population, CP1 under 2-CBD is much higher than that under 1-CBD except for one case, PDP Type I. This shows that 2-CBD could greatly postpone RTL construction. For Type I, trips surge from the original CBD area to the other CBD impels the RTL construction. The RTL construction duration is prolonged for PDP Type I; shortened for Type II and maintained the same for III.

### 3.4. Sensitivity analysis of DAR cost on Critical Population 1 (CP1)

Intuitively, higher DAR operating cost will encourage the introduction of RTL. On contrary, low DAR operating cost should be able to detain RTL construction. To verify this statement, we calculate the population that the first rapid transit line appears (CP1) under different DAR cost ratios for the three PDP types. DAR cost ratio represents the unit operating cost for DAR services, which is proportional to the RTL capacity. For instance, ratio=0.2 stands that the unit DAR cost is \( \frac{0.2}{\xi} \), \( \xi \) for RTL vehicle capacity. From Table 7, we can see that when ratio<1, CP1 under the five PDP Types does not change with cost ratio, which seems to be counter intuitive. However, it can be interpreted in this way: recall that the congestion cost for DAR services is represented by the BPR function. The DAR cost will increase with a power of 4 when demand exceeds the road capacity. When population density reaches a certain amount, it is not economical to rely on DAR alone even for a low DAR unit cost. When ratio>1 in Type I, increase in cost ratio leads to decrease in CP1. It coincides with our expectation. In summary, when road congestion is severe, RTL construction is imperative regardless of the cost of DAR. However, appropriate pricing on DAR will help in introducing RTL services.

### Table 7. CP1 under various DAR cost ratio and PDP types with one CBD

<table>
<thead>
<tr>
<th>PDP Type</th>
<th>DAR cost ratio</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td></td>
<td>2400</td>
<td>2400</td>
<td>2400</td>
<td>2400</td>
<td>2400</td>
<td>2000</td>
<td>1600</td>
<td>1200</td>
<td>1000</td>
</tr>
<tr>
<td>II</td>
<td></td>
<td>1200</td>
<td>1200</td>
<td>1200</td>
<td>1200</td>
<td>1200</td>
<td>1000</td>
<td>1000</td>
<td>800</td>
<td>600</td>
</tr>
<tr>
<td>III</td>
<td></td>
<td>1200</td>
<td>1200</td>
<td>1200</td>
<td>1200</td>
<td>1200</td>
<td>1000</td>
<td>1000</td>
<td>800</td>
<td>800</td>
</tr>
</tbody>
</table>

### 4. Conclusion

In this paper, we formulated the transit network design problem under demand uncertainty through robust
optimization. Road congestion and transit crowdness are taken into account to formulate the problem under the UE principle. To solve this highly complex formulation, we developed linearization procedures combined with a cutting constraint algorithm to achieve substantial gains in computation time, averaged 56 seconds per solution. The approach developed brings promises for applying it to larger networks.

As far as studying the relationships between population distribution and provision of financially sustainable RTL, we investigated three population distribution patterns to determine the critical population densities to afford the first RTL. Under the scenario of having one CBD, the results showed that the critical population was lower for population concentrating in the suburb due to the longer travel demands for a larger population in the suburb. On the other hand, a PDP with the population concentrated in the center (Type I) could postpone RTL development, i.e. with a high critical density, which also had the lowest total system cost. In addition, Type I is capable of handling a higher population with the same ground road network. On the other hand, under the scenario of having two CBDs, the total system cost of Type I was not always lower than the other types of developments. And in this case, Type III (with housing developments at both ends) outperformed the other two.

Although we modeled the TNDP for two types of services, rapid transit services and dial-a-ride services, there are certain limitations that need to be addressed in future studies. The first issue is that passenger waiting time for a transit line should be incorporated, which can be extended by the approach developed in An and Lo (2014a). The second is that the model allows unlimited transfers between transport modes, which is not realistic. In this study, we imposed a heavy transfer penalty to mitigate the problem, but we cannot totally avoid it as it is inherent in link based formulations for the network design problem. This problem can be entirely overcome by a path-based formulation, as developed in Lo et al. (2003). The third is that the travel cost friction factor $F_{ij}$ is proportional to distance but not directly to travel time in trip distribution. Using the actual path travel times for trip distribution will incur interactions between the gravity model and the robust optimization model, which will add complexity but should be feasible and certainly is a worthy extension. This paper aimed at minimizing the total system cost. Conceivably, the optimal RTL sequence will differ according to different objectives, such as those of the passengers, RTL companies, DAR companies, or the government. Combining these various objectives in the formation provides another direction for extensions. Finally, the influence of network topologies on the line construction sequence can be further studied by the proposed method, as can be the incorporation of other sources of variations (Watling and Cantarella, 2013). Hopefully, some general insights can be obtained in the relationship between development density and patterns and sustainable transit network developments.

Acknowledgements

The study is supported by the Public Policy Research grant HKUST6002-PPR-11, General Research Fund #616113 of the Research Grants Council of the HKSAR Government, the Hong Kong PhD Fellowship, and Strategic Research Funding Initiative SRF111EG15.

References


Quadrifoglio, L., Li, X., 2009. A methodology to derive the critical demand density for designing and operating feeder transit services. Transportation Research Part B: Methodological 43(10), 922-935.


