ABSTRACT

Title of Thesis:	OPTIMIZING PHASED DEVELOPMENT OF A PRE-DESIGNED RAIL TRANSIT LINE
	Fei Wu, Master of Science, 2020
Thesis Directed By:	Professor Paul M. Schonfeld, Department of Civil and Environmental Engineering

A novel model is developed for optimizing the phased development of a pre-designed rail transit line. The investment plan and extension phases of the line are optimized over continuous time and under budget constraints to maximize net present value (NPV) over an analysis period. This model determines the maximal allowable train headway while considering demand elasticity. This model is first formulated for a one-directional extension problem, and then modified for its two-directional version. A genetic algorithm with customized operators is developed for optimizing the sequence and grouping of station completions. For each problem version the model is demonstrated with a numerical case and its corresponding optimized solution. The sensitivity of results to several important input parameters is analyzed. Results show that potential demand and in-vehicle time value greatly influence the optimized NPV, while unit construction cost and potential demand are most influential on the optimized extension plan.

OPTIMIZING PHASED DEVELOPMENT OF A PRE-DESIGNED RAIL TRANSIT LINE

by

Fei Wu

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Science
2020

Advisory Committee:
Professor Paul M. Schonfeld, Chair
Professor Ali Haghani
Professor Lei Zhang

Acknowledgements

First, I express my gratitude to Dr. Paul M. Schonfeld, the advisor of my graduate studies, for his technical assistance on my thesis as well as other projects I work on. I also thank CEE professors, including the advisory committee members -- Paul M. Schonfeld, Ali Haghani, and Lei Zhang, for lecturing helpful courses in the field of transportation engineering.

I would like to express my appreciation to the support from the University Mobility and Equity Center, led by Morgan State University, which partially funded the research in this thesis.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Background

In many metropolitan areas worldwide rail rapid transit systems play an important role in serving busy commuting corridors and relieving traffic congestion. Such rail transit systems can significantly reduce congestion, help alleviate traffic-induced air pollution, and reduce travel times for transit users as well as others. Hence, the construction of rail transit systems is widely supported by decision makers in major cities.

Although many new passengers benefit from new links and stations, their construction as well as their regular operation and maintenance impose substantial costs on the operating agencies. Since fares should be affordable for most users, these costs make it difficult for rail transit projects to be profitable. Therefore, it is challenging to design affordable development plans, which are constrained by available budgets. An operator's available funds, including external subsidy and some fraction of revenues, may be needed to pay for operating and maintenance costs, as well as invest in construction. As new stations and links are completed, ridership is redistributed over the operating rail transit line. Revenue, users' total benefit, users' total costs, and supplier's costs increase as a rail transit network is extended. Moreover, while the value of satisfying travel demand favors earlier extensions, budget constraints and the discounting of costs and benefits through the time value of money favor delaying extensions (Sun et al., 2018). Given the above considerations,

it is desirable to optimize a phased development plan. This plan determines which links and stations should be completed at what times in order to maximize the system's discounted net benefit, i.e. its net present value (NPV).

1.2 Literature Review

The design and optimization of rail transit systems has been studied by numerous researchers over decades. Most previous studies focus on the following aspects:

1) Timetabling. Wong et al. (2008) optimize trains' running times, dwell times, turnaround times and headways. Barrena et al. (2014) optimize single line timetabling under dynamic passenger demand. Hassannayebi et al. (2014) optimize a timetable to deal with uncertainty of travel time, demand, and dwell time. In these works, the users' waiting time is minimized. However, for oversaturated passenger flows, Niu and Zhou (2013) optimize timetables with a binary integer programming model that minimizes number of waiting passengers and weighted remaining passengers. Shang et al. (2018) optimize skip-stop scheduling with a passenger equity improvement model.

2) Coordination with other systems. To minimize total system cost in coordinated rail and bus transit systems with location-varying demand, Wirasinghe et al. (1977) optimize station spacings, feeder-bus zone boundaries, and train headways in a network with radial rail lines, while Chien and Schonfeld (1998) optimize rail length, rail station spacings, bus headway, bus stop and route spacings for a single rail route with feeder bus routes in an urban corridor. Gallo et al. (2011) optimize train frequency to minimize operator's cost, users' cost and external cost, considering the

bus system as a feeding and competing system, and the private car system as a competing system.

3) Alignment and network design. These studies involve designing a network or a single line from scratch, or extending existing lines. For alignment design, Samanta and Jha (2011) optimize station locations for a given corridor using three different objective functions, while Lai and Schonfeld (2016) jointly optimize rail transit alignments and station locations under various realistic constraints. Both of those studies evaluate candidate solutions using data from geographic information systems (GIS's). Guan et al. (2006) jointly minimize total line length with given candidate lines and minimize total travel time and total number of passenger transfers with optimized passenger line assignment. Li et al. (2012) develop two models using flat and distance-based pricing to maximize profit with optimized rail line length, station number and locations, train headway, and fare. Saidi et al. (2016) propose a long-term planning method for ring-radial rail transit networks with three steps: exactly optimizing the number of radial lines with minimized total cost, predicting passengers' route choice with a ring line in the network, and identifying optimality and feasibility of the ring line. Canca et al. (2017) formulate a profit-maximizing model for designing rail transit lines based on given demand points, and solve the problem with an adaptive neighborhood search metaheuristic algorithm.

The problem of phased development of a rail transit system is related to network design problems, but has some distinguishing features. It focuses on the timing of improvements for a pre-designed system. Completion times of various system components are decision variables. The objective function value is evaluated and

discounted over an analysis period that includes multiple extension steps, while travel demand grows over time and may be affected by the system's evolving characteristics.

Currently, the phased development problem is still largely unexplored for rail transit lines and even less explored for networks. Only a few studies on this problem have been published. Cheng and Schonfeld (2015) propose the first known model, where the system's NPV is maximized in the analyzed period. Budget constraints, economies of scale (i.e. reducing construction costs by completing multiple links together), and a fixed growth rate of demand are considered. A simulated annealing method is used to search for the optimal extension plan, and the sensitivity of results to budget constraints and interest rates is examined. Sun et al. (2018) improve upon that model by proposing a bi-level program. Fare, headway and train capacity are jointly optimized in the lower-level problem using analytic methods. The extension plan is optimized with dynamic programming in the upper-level problem, where the system's NPV is maximized. An elastic demand function is proposed to incorporate the effects of waiting time, access time and in-vehicle time. They find that a multi-phase plan may be preferable to a single-phase one even without budget constraints since rail segments to outer suburbs may by unwarranted until demand increases sufficiently. Peng et al. (2019) analyze a network with interrelated projects, and capture larger demand growth rates after new station completions. Their travel demand function is time-varying. They use a genetic algorithm to minimize the present value (PV) of total costs. The sensitivity of the results to initial travel demand and annual budget level is examined.

Models proposed by Cheng and Schonfeld (2015) and Sun et al. (2018) specifically for solving the phased development problem are formulated with the analysis period segmented into smaller time steps. Then, the number of possible values of completion times of links and stations is limited, which is somewhat unrealistic and may miss desirable solutions. A model that treats time as being continuous is desirable, so that in the optimized extension plan links can be completed at any time during the analysis period. Peng et al. (2019) formulate the problem with continuous time in the analysis period, but the stations and links to be completed in groups are pre-determined. The extension plan would be more flexible if any feasible sequence of completion of stations and links may be applied.

1.3 Scope of Study

This thesis presents a novel optimization model for the phased development of a rail transit line. The model is first based on a one-directional extension problem, and then modified regarding a two-directional extension problem. Since time is continuously formulated, the model formulation is significantly different from most models proposed in previous studies. In this model, the only group of decision variables is the completion times of new stations and links, whose values can be real numbers between 0 and the duration of analysis period. Demand elasticity is considered by using a linear demand function that shows effects of fare, waiting time (train headway) and in-vehicle time on actual demand. Demand growth rate, economies of completing multiple stations and links together, and funding constraints are also incorporated in the model. The objective is to maximize the overall system NPV (total consumer surplus plus total supplier revenue minus total supplier cost).

While Differential Evolution (DE), a general-purpose heuristic method, is used for solving the one-directional extension problem, a Genetic Algorithm (GA) is applied with customized operators for solving the two-directional extension problem. Constraints on sequence and grouping of station completion are considered in operator settings. The optimization model is modified to fit the usage of GA in the two-directional extension problem.

In this thesis, problem formulations for both problem versions are presented in Chapter 2. Solution methods, including DE and the customized GA, are presented in Chapter 3. Base scenarios and their corresponding optimized extension plans, effects of selected parameters, and sensitivity analysis are presented in Chapter 4, for both problem versions. Conclusions and future improvements are presented in Chapter 5.

Chapter 2: Problem Formulation

2.1 Problem settings for one-directional extension

A planned rail transit line (as shown in Figure 1) connects a central business district (CBD) with outer districts. In the case where it can only be extended in one direction, it includes m ($m \geq 3$) stations, only n ($n \geq 2$) of which are in service. The remaining $(m-n)$ stations and corresponding rail links may be completed in the following T years. Link i is defined as the link between stations $i-1$ and i, and has a length of d_i.

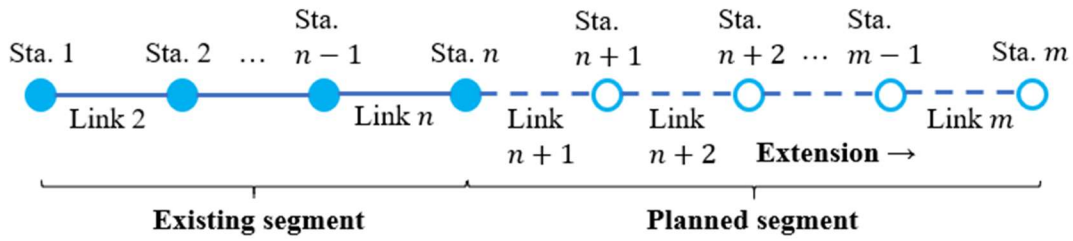


Figure 1 Planned rail transit line discussed in one-directional extension

Decision variables t_k are defined for $k = 1, 2, \ldots, m-n$ to indicate the planned time at which Station $(n+k)$ and Link $(n+k)$ should be completed and start operation. It is defined that Station $(n+k)$ is to be completed t_k years from the start of the analysis period, and that Link $(n+k)$ must be completed simultaneously. The values of t_k are in years and range continuously from 0 to T. Specifically, if $t_k = T$, then Station $(n+k)$ will not be completed within T years. It is also defined that $t_0 = 0$

and $t_{m-n+1} = T$. When extending the line, its continuity should be ensured, by ensuring that $t_k \leq t_{k+1}$ for all $k = 0, 1, \ldots, m-n$.

The time period during which k planned stations will be in operation is defined as "Period k", whose duration is denoted as T_k. Then, $T_k = t_{k+1} - t_k$ $(0 \leq k \leq m-n)$. If two or more stations (along with corresponding links) are completed together, there will be period(s) with zero duration.

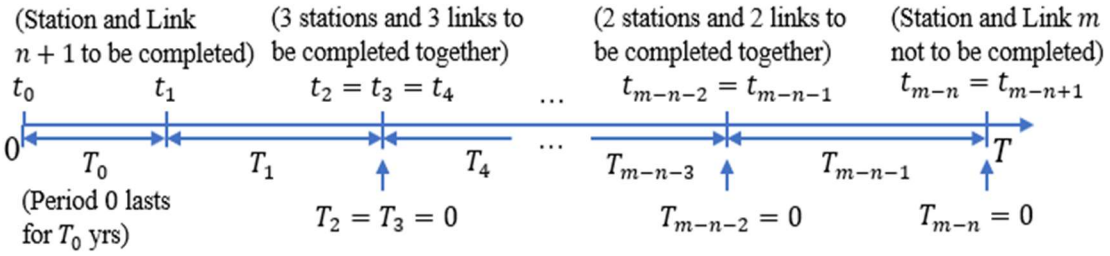


Figure 2 Setting of decision variables and "periods" in one-directional extension

2.2 Notation

The notation and baseline values for variables used in this case with one-directional extension are shown in Table 1.

Table 1 Notation and baseline values for variables in one-directional extension

Symbol	Description	Unit	Baseline Value
b_{ij}	Maximal acceptable impedance (total travel cost) for a passenger from Station i to j	\$	
c_{end}	Cost related to terminal facilities for reversing train direction	\$	3×10^7
c_{in}	Initial cost of a train	\$/train	1.2×10^7
c_m	Avg. hourly maintenance cost per unit length of the rail transit line	\$/mile/hr	200
c_o	Avg. hourly operation cost per train in operation	\$/train/hr	5000
c_{st}	Construction cost of a new station	\$	4×10^7
c_{ln}	Construction cost per unit length of transit line	\$/mile	6×10^7
C_{ij}^k	Impedance per passenger from Station i to j during Period k	\$	
d_i	Length of Link i	miles	
f	Fixed rail transit fare	\$	2.75
F_0	Initial available budget for construction	\$	1×10^8

F	Yearly external budget for construction	\$/yr	5×10^7
g	Constant annual exponential growth rate of potential demand	%/yr	3%
h^k	Train headway during Period k	hours	
h^k_{max}	Maximum allowable train headway during Period k	hours	
H	Number of operation hours per year	hrs/year	6000
K	Capacity of each train	psgrs	960
m	Number of all planned stations in the transit line		
n	Number of existing stations in the transit line		
N^k	Number of trains on the transit line during Period k		
P^k_{co}	Present value (PV) of construction cost incurred at the start of Period k	\$	
P^k_{CS}	PV of consumer surplus incurred during Period k	\$	
P^k_f	PV of fare collected during Period k	\$	
P^k_{in}	PV of initial train cost incurred at the start of Period k	\$	
P^k_m	PV of track maintenance cost incurred during Period k	\$	
P_{NB}	Net PV of social benefit = NPV	\$	
P^k_o	PV of operation cost to be incurred during Period k	\$	
P_{SC}	Total PV of supplier costs	\$	
q^k_{ij}	Actual hourly passenger flow from Station i to j at Period k	psgrs/hr	
q^k_{max}	Largest hourly passenger flow over the operating line at end of Period k	psgrs/hr	
Q_{ij}	Potential hourly passenger flow from Station i to j	psgrs/hr	
r	Constant annual interest rate	%/yr	7%
R^k	Round-trip time on the rail transit line during Period k	hours	
S^k	Approx. avg. hourly consumer surplus during Period k	\$/hour	
t_d	Average dwell time at a station	hours	0.01
t_{dt}	Time needed for reversing direction at each terminal station	hours	0.03
$t^k_{v,ij}$	In-vehicle travel time from Station i to j during Period k	hours	
t_k	The time when Station k and Link k are to be completed	years	
t^k_w	A passenger's avg. waiting time for a train during Period k	hours	
T	Duration of the analysis period	years	
T_k	Duration of Period k	years	
u_v	A passengers' avg. value of in-vehicle time	\$/hour	18
u_w	A passengers' avg. value of waiting time	\$/hour	18
V_{ot}	Avg. speed of alternatives to rail transit	mph	16
V_{tr}	Avg. running speed of a train	mph	40
γ^k	Binary variable indicating whether construction costs are incurred at start of Period k		
δ^k	Binary variable indicating whether terminal facility costs are incurred at start of Period k		
η	Peak hour factor		1.25
ρ	Fraction of collected fare to be used for new construction		15%

2.3 Determining impedance, actual demand and consumer surplus

In the equations presented below, unless otherwise stated, let $1 \le i \le m$, $1 \le j \le m$, and $0 \le k \le m - n$. i, j, and k are integers.

It is assumed here that the initial potential demand (i.e. the largest possible number of rail transit passengers, theoretically occurring at zero travel time and fare) for each OD pair (at the station level and for this line only) is externally given. The initial potential hourly passenger flow from Station i to j is denoted as Q_{ij}. If $i = j$, let $Q_{ij} = 0$. The potential demand for each OD pair is assumed to increase exponentially at an annual rate g. (This is a simplifying assumption for this model. In real world, demand growth over time can be of any other type besides the exponential one. Completion of new stations may also lead to higher growth rate in the short term.)

The impedance for a passenger (user) on each OD pair in Period k is denoted as C_{ij}^k. The impedance equals the rail transit fare f plus the user's time costs, which include waiting cost and in-vehicle cost. (The cost of access time may be neglected here.) Then for $i < j$:

$$C_{ij}^k = f + u_v t_{v,ij}^k + u_w t_w^k, \qquad \forall i < j \ \wedge \ i \le k + n - 1 \tag{1}$$

where u_v is the average value of in-vehicle time, u_w is the average value of waiting time, $t_{v,ij}^k$ is the in-vehicle time for a passenger from Station i to j during Period k, and t_w^k is the average waiting time per transit trip during Period k (assumed here to be half the train headway).

When determining values of C_{ij}^k, for $i < j$:

$$t_{v,ij}^k = \frac{\sum_{l=i+1}^{k+n} d_l}{V_{tr}} + \frac{\sum_{l=k+n+1}^{j} d_l}{V_{ot}} + (k + n - i)t_d, \qquad \forall\, i < k + n < j \tag{2a}$$

$$t_{v,ij}^{k} = \frac{\sum_{l=i+1}^{j} d_l}{V_{tr}} + (j-i)t_d, \qquad \forall\, i < j \le k+n \tag{2b}$$

$$t_w^k = \frac{h^k}{2} \tag{3}$$

where V_{tr} is the average running speed of a train, V_{ot} is the average speed of alternatives to rail transit, t_d is the average dwell time at a station, and h^k is the train headway during Period k. Waiting time for alternatives to rail transit is not specifically considered because V_{ot} has taken it into account. V_{tr}, V_{ot}, and t_d are assumed to be constant over time.

During each period, the in-vehicle time is assumed to be symmetric for each OD pair:

$$t_{v,ij}^{k} = t_{v,ji}^{k}, \qquad \forall i \neq j \tag{4}$$

which makes the impedance symmetric for each OD pair:

$$C_{ij}^{k} = C_{ji}^{k}, \qquad \forall i \neq j \tag{5}$$

It is assumed that there is an underlying linear demand function for determining the actual ridership of each OD pair, as shown in Figure 3. The actual hourly passenger flow from Station i to j at Period k is denoted as q_{ij}^{k}. Its approximate average value during this period is given by:

$$q_{ij}^{k} = Q_{ij}(1+g)^{\frac{t_k + t_{k+1}}{2}} \frac{b_{ij} - C_{ij}^{k}}{b_{ij}} \tag{6}$$

where b_{ij} is the pre-determined maximal acceptable impedance for passengers from Station i to j. The hourly ridership at the midpoint of this period is used as the approximate average.

If the impedance exceeds b_{ij}, the actual ridership for the corresponding OD pair becomes zero.

It is assumed that if a certain OD pair (e.g., from Station $k+n+1$ to Station $k+n+3$) contains no operating links in Period k, passengers of this OD pair will not use rail transit. For $i < j$:

$$C_{ij}^k = b_{ij}, \qquad \forall j > i \geq k+n \tag{7}$$

From equation (5), values of C_{ij}^k with $i > j$ can be determined, given those with $i < j$.

With this demand function the approximate average hourly consumer surplus (CS) during Period k can be calculated. This value is denoted as S^k, and is expressed as:

$$S^k = \sum_i \sum_{j \neq i} q_{ij}^k \frac{b_{ij} - C_{ij}^k}{2} = (1+g)^{\frac{t_k + t_{k+1}}{2}} \sum_i \sum_{j \neq i} Q_{ij} \frac{(b_{ij} - C_{ij}^k)^2}{2b_{ij}} \tag{8}$$

where the hourly CS value at the midpoint of the period is used as the approximate average.

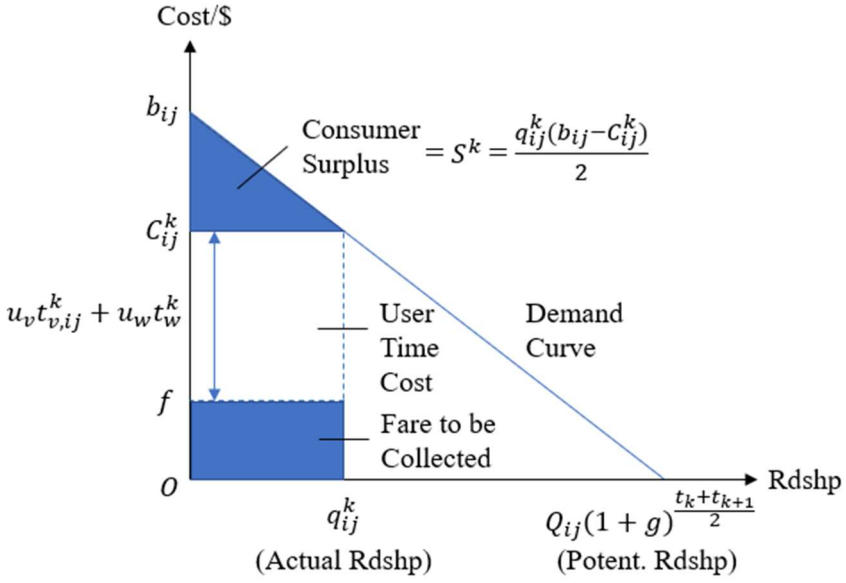


Figure 3 Linear demand curve and related parameters & amounts

2.4 Determining train headway

To determine the maximum allowable train headway in Period k, the capacity of each identical train (denoted as K) and the peak hour factor η are considered. If the higher one-directional hourly passenger flow through Link i at the end of Period k is denoted as q_i^k, then:

$$q_i^k = \max\{q_{i,out}^k, q_{i,in}^k\}, \qquad \forall 2 \le i \le k+n \tag{9}$$

where $q_{i,out}^k$ is the actual hourly outbound passenger flow (from CBD to suburb) through Link i at the end of Period k, and $q_{i,in}^k$ is the corresponding inbound flow (from suburbs to CBD):

$$q_{i,out}^k = (1+g)^{t_{k+1}} \sum_{l=1}^{i-1} \sum_{j=i}^{m} Q_{lj} \frac{b_{lj} - C_{lj}^k}{b_{lj}}, \qquad \forall 2 \le i \le k+n \tag{10a}$$

$$q_{i,in}^k = (1+g)^{t_{k+1}} \sum_{l=1}^{i-1} \sum_{j=i}^{m} Q_{jl} \frac{b_{jl} - C_{jl}^k}{b_{jl}}, \qquad \forall 2 \le i \le k+n \tag{10b}$$

If the highest hourly passenger flow over the operating line at the end of Period k is denoted as q_{max}^k, then:

$$q_{max}^k = \max\{q_2^k, q_3^k, \dots, q_{k+n}^k\} \tag{11}$$

It is assumed that the maximum allowable headway h_{max}^k, and therefore the required fleet size, is determined by the peak hourly passenger flow at the end of Period k:

$$h_{max}^k = \frac{K}{\eta q_{max}^k} \tag{12}$$

Since C_{lj}^k linearly increases as h^k increases, and $q_{i,out}^k$, $q_{i,in}^k$ linearly decrease as C_{lj}^k increases, $q_{1,out}^k$, $q_{i,in}^k$ linearly decrease as h^k increases. Then, to determine the value of h_{max}^k, some quadratic equations must be solved. For each $q_{i,out}^k$ $(2 \le i \le k+n)$ the following equation is used:

$$h_{i,out}^k \eta q_{i,out}^k = K \tag{12a}$$

It is expanded stepwise:

$$h_{i,out}^k \eta (1+g)^{t_{k+1}} \sum_{l=1}^{i-1} \sum_{j=i}^{m} Q_{lj} \frac{b_{lj} - C_{lj}^k}{b_{lj}} - K = 0 \tag{12b}$$

$$h_{i,out}^k \eta (1+g)^{t_{k+1}} \sum_{l=1}^{i-1} \sum_{j=i}^{m} Q_{lj} \frac{b_{lj} - f - u_v t_{v,lj}^k - u_w t_w^k}{b_{lj}} - K = 0 \tag{12c}$$

$$h_{i,out}^k \eta (1+g)^{t_{k+1}} \left[\sum_{l=1}^{i-1} \sum_{j=i}^{m} Q_{lj} \frac{b_{lj} - f - u_v t_{v,lj}^k}{b_{lj}} - \sum_{l=1}^{i-1} \sum_{j=i}^{m} Q_{lj} \frac{u_w h_{i,out}^k}{2 b_{lj}} \right] - K = 0 \tag{12d}$$

$$\frac{u_w \eta (1+g)^{t_{k+1}}}{2} \sum_{l=1}^{i-1} \sum_{j=i}^{m} \frac{Q_{lj}}{b_{lj}} {h_{i,out}^k}^2 - \eta (1+g)^{t_{k+1}} \sum_{l=1}^{i-1} \sum_{j=i}^{m} Q_{lj} \frac{b_{lj} - f - u_v t_{v,lj}^k}{b_{lj}} h_{i,out}^k + K = 0 \tag{12e}$$

For simplicity, coefficients in the quadratic equation (12e) are denoted:

$$\alpha_{i1}^k = \frac{u_w \eta (1+g)^{t_{k+1}}}{2} \sum_{l=1}^{i-1} \sum_{j=i}^{m} \frac{Q_{lj}}{b_{lj}} \tag{13a}$$

$$\beta_{i1}^k = -\eta (1+g)^{t_{k+1}} \sum_{l=1}^{i-1} \sum_{j=i}^{m} Q_{lj} \frac{b_{lj} - f - u_v t_{v,lj}^k}{b_{lj}} \tag{13b}$$

The quadratic equation (12e) with unknown variable $h_{i,out}^k$ has two positive real roots when:

$$\beta_{i1}^{k\,2} - 4K\alpha_{i1}^k \geq 0 \tag{14a}$$

The smaller root is chosen because a longer headway that reduces ridership is undesirable:

$$h_{i,out}^k = \frac{-\beta_{i1}^k - \sqrt{\beta_{i1}^{k\,2} - 4K\alpha_{i1}^k}}{2\alpha_{i1}^k} \tag{15a}$$

Since $q_{i,out}^k$ linearly decreases as $h_{i,out}^k$ increases, the value of $h_{i,out}^k \eta q_{i,out}^k$ reaches the maximum when $h_{i,out}^k = -\beta_{i1}^k/(2\alpha_{i1}^k)$. If (14a) is not satisfied, then the quadratic equation (12e) does not have real roots, which means that $K > (h_{i,out}^k \eta q_{i,out}^k)_{max}$, and the highest possible number of passengers in a train traveling outbound on Link i in Period k is lower than the train capacity. It is assumed that the headway for $q_{i,out}^k$ maximizes the utilization of train capacity. Let:

$$h_{i,out}^k = -\frac{\beta_{i1}^k}{2\alpha_{i1}^k} \tag{15b}$$

Similarly, for each $q_{i,in}^k$ ($2 \leq i \leq k+n$) the equation $h_{i,in}^k \eta q_{i,in}^k = K$ is used. After expansion, coefficients in the resulting quadratic equation are denoted:

$$\alpha_{i2}^k = \frac{u_w \eta (1+g)^{t_{k+1}}}{2} \sum_{l=1}^{i-1} \sum_{j=i}^{m} \frac{Q_{jl}}{b_{jl}} \tag{13c}$$

$$\beta_{i2}^k = -\eta (1+g)^{t_{k+1}} \sum_{l=1}^{i-1} \sum_{j=i}^{m} Q_{jl} \frac{b_{jl} - f - u_v t_{v,jl}^k}{b_{jl}} \tag{13d}$$

Then when

$$\beta_{i2}^{k\,2} - 4K\alpha_{i2}^{k} \geq 0 \tag{14b}$$

there is:

$$h_{i,in}^{k} = \frac{-\beta_{i2}^{k} - \sqrt{\beta_{i2}^{k\,2} - 4K\alpha_{i2}^{k}}}{2\alpha_{i2}^{k}} \tag{15c}$$

If (14b) is not satisfied:

$$h_{i,in}^{k} = -\frac{\beta_{i2}^{k}}{2\alpha_{i2}^{k}} \tag{15d}$$

Then the maximum allowable headway h_{max}^{k} in each period can be determined. To simplify problem in this early model version, it is assumed that h_{max}^{k} is used as h^{k}, without treating h^{k} as an optimizable variable:

$$h_{i}^{k} = \min\{h_{i,out}^{k}, h_{i,in}^{k}\}, \qquad \forall 2 \leq i \leq k+n \tag{16}$$

$$h^{k} = h_{max}^{k} = \min\{h_{2}^{k}, h_{3}^{k}, \dots, h_{k+n}^{k}\} \tag{17}$$

With $(k+n)$ stations in operation, the round-trip time of a train during Period k is:

$$R^{k} = 2[\frac{\sum_{i=2}^{k+n} d_{i}}{V_{tr}} + (k+n)t_{d} + t_{dt}] \tag{18}$$

where t_{dt} is the required time for a train to reverse direction at each terminal station. Then the required number of trains (fleet size) for the rail transit line during Period k is:

$$N^{k} = R^{k}/h^{k} \tag{19}$$

Here a simplification is assumed that the fleet size N^{k} can be non-integer. In a more rigorous way, N^{k} should be limited to integers, thus limiting possible values of h^{k} with given R^{k}.

2.5 Objective function and constraints

The objective of this model is to maximize net present value (NPV), i.e. the discounted net benefit. It is achieved by optimizing completion times of planned stations and links, so that the resulting overall NPV over the analysis period is maximized.

First, the components of the objective function (OF) is explained.

There are H operating hours per year, and Period k lasts for T_k years. When calculating the present value (PV) of consumer surplus, a constant interest rate r is used here. Then the PV of consumer surplus in Period k is:

$$P_{CS}^k = \frac{S^k H T_k}{(1+r)^{\frac{t_k+t_{k+1}}{2}}} = HT_k\left(\frac{1+g}{1+r}\right)^{\frac{t_k+t_{k+1}}{2}} \sum_i \sum_{j \neq i} Q_{ij} \frac{(b_{ij} - C_{ij}^k)^2}{2b_{ij}} \tag{20}$$

An approximation is used that when discounting total consumer surplus in Period k, the original sum is assumed to be concentrated at the midpoint of the period (as shown in Figure 4).

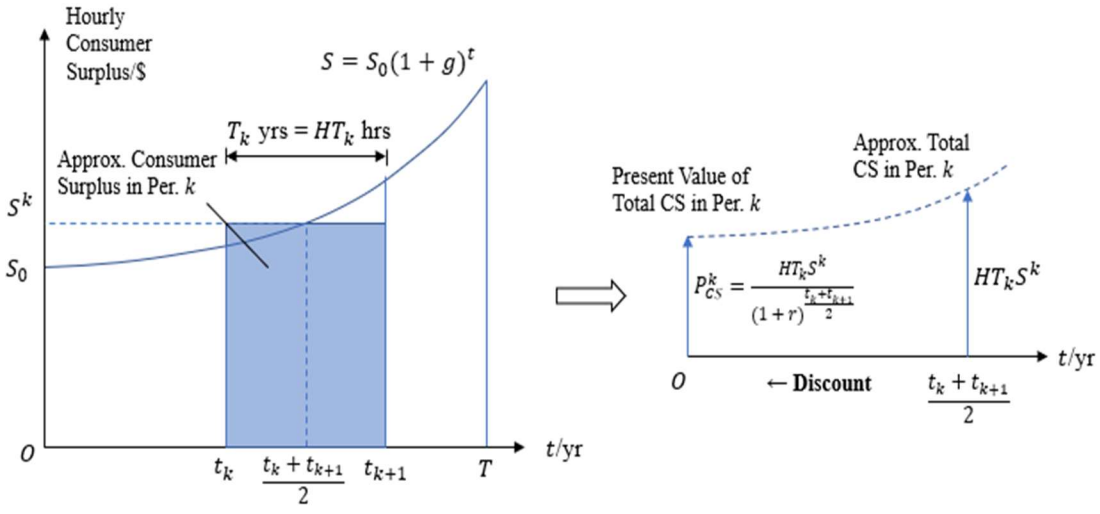


Figure 4 Approximation of total PV of consumer surplus during period k

Using Figures 3 and 4, the approximate PV of fare to be collected from passengers in Period k can be determined similarly:

$$P_f^k = \frac{fHT_k}{(1+r)^{\frac{t_k+t_{k+1}}{2}}} \sum_i \sum_{j \neq i} q_{ij}^k = fHT_k \left(\frac{1+g}{1+r}\right)^{\frac{t_k+t_{k+1}}{2}} \sum_i \sum_{j \neq i} Q_{ij} \frac{b_{ij} - C_{ij}^k}{b_{ij}} \quad (21)$$

Then the PV of various supplier cost components can be determined. The approximate PV of total vehicle operation cost during Period k is:

$$P_o^k = \frac{c_o N^k H T_k}{(1+r)^{\frac{t_k+t_{k+1}}{2}}} \quad (22)$$

where c_o is the average hourly operation cost of each train running on the transit line.

The approximate PV of total track maintenance cost during Period k is:

$$P_m^k = \frac{c_m H T_k \sum_{i=2}^k d_i}{(1+r)^{\frac{t_k+t_{k+1}}{2}}} \quad (23)$$

where c_m is the average hourly maintenance cost per unit length of the rail transit line.

It should be noted that all approximations above that involve $(1+r)^{\frac{t_k+t_{k+1}}{2}}$ or $(1+g)^{\frac{t_k+t_{k+1}}{2}}$ are acceptable when r and g are small and $(1+r)$ is close to $(1+g)$. Integration methods that yield more accurate PV's may be considered in future versions of the model.

Assuming that the fleet size is prepared for the peak demand at the end of each period and is not changed within each period (no new vehicles are added into the system within each period), the PV of initial costs of additional trains at the start of Period k is:

$$P_{in}^{k} = \frac{c_{in}(N^{k} - N^{k-1})}{(1+r)^{t_k}}\gamma^{k}, \qquad \forall 1 \le k \le m-n \tag{24}$$

where c_{in} is the initial cost of a train with a certain number of cars. The binary variable γ^k equals 0 when $t_k = T$, and equals 1 when $t_k < T$. This shows that if a certain station is not completed within the analysis period, then the duration of the corresponding period is zero and the initial train cost is not incurred.

Assuming that construction costs are incurred at the time of completion of new stations and links, the PV of construction costs of new stations and links at the start of Period k is:

$$P_{co}^{k} = \frac{c_{st} + c_{ln}d_k + \delta^{k}c_{end}}{(1+r)^{t_k}}\gamma^{k}, \qquad \forall 1 \le k \le m-n \tag{25}$$

where c_{st} is the average construction cost of a new station, c_{ln} is the average construction cost per unit length of the transit line, and c_{end} is the cost of removing old terminal facilities and setting new ones when the line is extended. Terminal facilities include depots for storing idle trains and special tracks for reversing train direction, and are used at both ends of the transit line.

The binary variable δ^k is used to indicate whether c_{end} will be incurred at the start of Period k. For all $1 \le k \le m-n$, the default $\delta^k = 1$. For all $1 \le k \le m-n-1$, if $t_k = t_{k+1}$, then Station $(k+n)$ is to be completed together with Station $(k+n+1)$ in a certain extension step of the line. c_{end} will not be incurred at the start of Period k (whose duration is zero), and there will be $\delta^k = 0$. The binary variable γ^k is also used to ensure that construction costs of new track and stations is not incurred if the corresponding station $(k+n)$ is not completed within the analysis period.

Then the total PV of supplier costs incurred in the analysis period is:

$$P_{SC} = \sum_{k=0}^{m-n} (P_o^k + P_m^k) + \sum_{k=1}^{m-n} (P_{in}^k + P_{co}^k) \tag{26}$$

The overall NPV over the analysis period is:

$$P_{NB} = \sum_{k=0}^{m-n} (P_{CS}^k + P_f^k) - P_{SC} \tag{27}$$

In this problem, the objective is to maximize P_{NB} with decision variables t_k. The values of t_k are subject to the sequence and domain constraint:

$$T \geq t_k \geq t_{k-1} \geq 0, \qquad \forall 1 \leq k \leq m-n \tag{28}$$

It is assumed that sources of construction budget include a specified external supply and a fraction of fares collected from passengers. Values of t_k are also subject to the available budget constraint:

$$F_0 + Ft_{k+1} + \rho \sum_{i=0}^{k} P_f^i (1+r)^{\frac{t_i + t_{i+1}}{2}} - \sum_{i=0}^{k} P_{co}^{i+1} (1+r)^{t_{i+1}} \geq 0, \forall 0 \leq k \leq m-n-1 \tag{29}$$

where F_0 is the initial available budget for construction, F is the yearly external budget supply, and ρ is the fraction of collected transit fare that contributes to total available budget. This constraint ensures that over the analysis period, the amount of available budget for construction remains non-negative after deducting construction costs from total available budget at the beginning of each period.

2.6 Problem settings for two-directional extension

A single rail transit line can also be extended in two directions from the existing segment (as shown in Figure 5). There are m ($m \geq 4$) stations in total, among which

n_e ($n_e \geq 2$) stations are existing, n_1 ($n_1 \geq 1$) stations at one end of the existing line (denoted as End 1) and n_2 ($n_2 \geq 1$) stations at the other end (denoted as End 2) may be completed in the following T years. Still, link i is defined as the link between stations $i-1$ and i, and has a length of d_i.

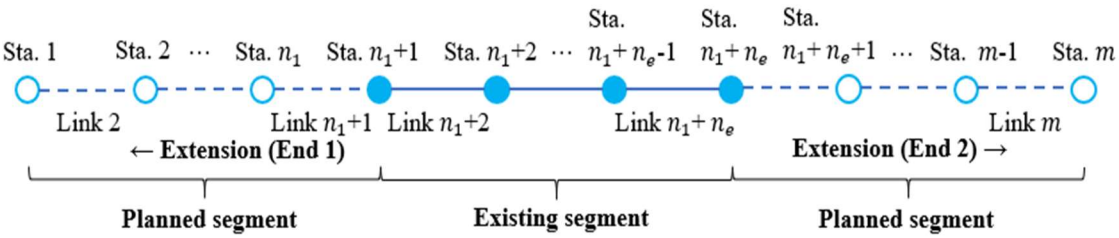


Figure 5 Planned rail transit line discussed in one-directional extension

In the problem with one-directional extension, decision variables t_k denote the planned completion time of each planned station and its corresponding link. In the problem with two-directional extension, however, t_k denote the planned completion time of the kth group of planned stations (which can be one or several) and corresponding links. The number of potential extension steps k_{max} and the stations and links to be completed in each extension step are given by a chromosome from the upper level genetic algorithm (GA). Each chromosome has two rows of integers that represent what groups of stations (and corresponding links) to be completed in a certain sequence in the next following T years. The customized GA and its chromosomes are presented in detail in Chapter 3.2.

When extending the line, its continuity should be ensured, and it is assumed that stations at different ends of the line cannot be completed together. There rules are used as constraints for generating feasible chromosomes in the GA. Each chromosome gives temporary terminal station codes E_1^k (for End 1) and E_2^k (for End 2) after finishing the kth potential extension step. For all $1 \leq k \leq k_{max}$, either $E_1^k <$

E_1^{k-1} and $E_2^k = E_2^{k-1}$, or $E_1^k = E_1^{k-1}$ and $E_2^k > E_2^{k-1}$. Before the first potential extension step is completed, $E_1^0 = n_1 + 1$ and $E_2^0 = n_1 + n_e$. After the last potential extension step is completed, $E_1^{k_{max}} = 1$ and $E_2^{k_{max}} = m$.

The values of t_k range continuously from 0 to T in years. It is assumed that each potential extension step will be completed as soon as the available budget becomes sufficient for this extension within the analysis period. With this assumption, t_k is numerically found for each potential step that can be realized within T years, using the modified formulation for the problem with two-directional extension (to be shown below). When $t_k < T$, $0 < t_k < t_{k+1} \leq T$ for $k \leq k_{max} - 1$. If $t_k \neq T$ and $E_1^k < E_1^{k-1}$ ($E_2^k > E_2^{k-1}$), then stations with codes from E_1^k to ($E_1^{k-1} - 1$) (from ($E_2^{k-1} + 1$) to E_2^k) will be completed t_k years from the start of the analysis period, and their corresponding links will be completed simultaneously. If it is found that the k'th potential extension step cannot be completed within the analysis period ($t_{k'} > T$), then $t_k = T$ is used for $k' \leq k \leq k_{max}$, and the stations and links that should be completed in the k'th and later potential extension steps will not be completed within T years. For all $k = 1, 2, \ldots, k_{max} - 1$, $0 < t_k \leq t_{k+1} \leq T$ is ensured.

The time period between t_k and t_{k+1} (or T) is defined as "Period k", whose duration is denoted as T_k. Then, $T_k = t_{k+1} - t_k$ ($1 \leq k \leq k_{max} - 1$), and $T_{k_{max}} = T - t_{k_{max}}$. For the time period before t_1, "Period 0" is defined with duration $T_0 = t_1$. If $t_k = T$, Period k has zero duration and is not realized within the analysis period with T years.

2.7 Modifications in formulation for two-directional extension

In the equations presented below, unless otherwise stated, let $1 \leq i \leq m$, $1 \leq j \leq m$, and $0 \leq k \leq k_{max}$. i, j, and k are integers.

In the problem with two-directional extension, Period k has its temporary terminal station codes E_1^k and E_2^k. In Period k, the passengers' travel impedance C_{ij}^k and actual rail transit ridership q_{ij}^k of some OD pair (from Station i to j) depend on the rail connection status between Station i and j. It is assumed that at most one transfer between rail transit and its alternative modes is allowed for potential rail transit passengers. In a period, if the operating segment of the rail transit line partially covers the interval between origin and destination stations but none of the OD stations are in operation, passengers of this OD pair will not use rail transit.

Then, for passengers who will use rail transit in Period k, the impedance is given by (for $i < j$):

$$C_{ij}^k = f + u_v t_{v,ij}^k + u_w t_w^k, \qquad \forall i < j \ \wedge \begin{bmatrix} \left(E_1^k \leq i \leq E_2^k - 1 \wedge j \geq E_1^k + 1\right) \\ \vee \left(i \leq E_1^k - 1 \wedge E_1^k + 1 \leq j \leq E_2^k\right) \end{bmatrix} \quad (1')$$

For passengers who will not use rail transit in Period k, the impedance is given by (for $i < j$):

$$C_{ij}^k = b_{ij}, \qquad \forall E_2^k \leq i < j, \forall i < j \leq E_1^k, \forall i \leq E_1^k - 1 \wedge j \geq E_2^k + 1 \qquad (7')$$

Equations (1') and (7') are respectively modified from equations (1) and (7) in the one-direction extension case, with changes on ranges of OD station codes i and j. For the in-vehicle time for a passenger from Station i to j during Period k, the modified version of equations (2a) and (2b) are given by (for $i < j$):

$$t_{v,ij}^{k} = \frac{\sum_{l=E_1^k+1}^{j} d_l}{V_{tr}} + \frac{\sum_{l=i+1}^{E_1^k} d_l}{V_{ot}} + \left(j - E_1^k\right)t_d, \qquad \forall\, i < E_1^k < j \tag{2a'}$$

$$t_{v,ij}^{k} = \frac{\sum_{l=i+1}^{E_2^k} d_l}{V_{tr}} + \frac{\sum_{l=E_2^k+1}^{j} d_l}{V_{ot}} + \left(E_2^k - i\right)t_d, \qquad \forall\, i < E_2^k < j \tag{2a''}$$

$$t_{v,ij}^{k} = \frac{\sum_{l=i+1}^{j} d_l}{V_{tr}} + (j - i)t_d, \qquad \forall\, E_1^k \le i < j \le E_2^k \tag{2b'}$$

Equations (3), (4), (5), (6), (8) are not changed for the two-directional extension problem. The symmetry of in-vehicle time and impedance still holds for each OD pair, and the linear demand function for calculating the actual hourly ridership and the approximate average hourly consumer surplus is still used.

Here a set of OD pairs is defined whose corresponding passengers will use rail transit in Period k. This set is denoted as Ω_k, given by:

$$\Omega_k = \left\{(i,j)\middle| i < j \ \wedge \begin{bmatrix} \left(E_1^k \le i \le E_2^k - 1 \wedge j \ge E_1^k + 1\right) \\ \vee \left(i \le E_1^k - 1 \wedge E_1^k + 1 \le j \le E_2^k\right) \end{bmatrix}\right\}$$

$$\cup \left\{(i,j)\middle| j < i \wedge \begin{bmatrix} \left(E_1^k \le j \le E_2^k - 1 \wedge i \ge E_1^k + 1\right) \\ \vee \left(j \le E_1^k - 1 \wedge E_1^k + 1 \le i \le E_2^k\right) \end{bmatrix}\right\}$$

For determination of the maximum allowable train headway in Period k of the two-directional extension, related equations are modified. Equation (9) is modified to:

$$q_i^k = \max\left\{q_{i,up}^k, q_{i,dn}^k\right\}, \qquad \forall E_1^k + 1 \le i \le E_2^k \tag{9'}$$

where $q_{i,up}^k$ is the actual hourly passenger flow in the direction from Station 1 to m through Link i at the end of Period k, and $q_{i,dn}^k$ is the corresponding flow in the direction from Station m to 1.

Letting $t_{k_{max}+1} = T$, equations (10a), (10b), and (11) are modified to:

$$q_{i,up}^k = (1+g)^{t_{k+1}} \sum_{l<i\le j,(l,j)\in\Omega_k} Q_{lj} \frac{b_{lj} - C_{lj}^k}{b_{lj}}, \qquad \forall E_1^k + 1 \le i \le E_2^k \tag{10a'}$$

$$q_{i,dn}^k = (1+g)^{t_{k+1}} \sum_{l<i\le j,(j,l)\in\Omega_k} Q_{jl} \frac{b_{jl} - C_{jl}^k}{b_{jl}}, \qquad \forall E_1^k + 1 \le i \le E_2^k \tag{10b'}$$

$$q_{max}^k = \max\{q_{E_1^k+1}^k, q_{E_1^k+2}^k, \dots, q_{E_2^k}^k\} \tag{11'}$$

Equation (12) still holds for the two-directional extension version. In subsequent equations regarding the extension and solution of equation (12) (from equation (12a) to (15d)), $q_{i,out}^k$, $q_{i,in}^k$, $h_{i,out}^k$, and $h_{i,in}^k$ are replaced with $q_{i,up}^k$, $q_{i,dn}^k$, $h_{i,up}^k$, and $h_{i,dn}^k$, respectively. The summation operation marks $\sum_{l=1}^{i-1}\sum_{j=i}^{m}$ are replaced with $\sum_{l<i\le j,(l,j)\in\Omega_k}$ in equations (12b) to (13b), and are replaced with $\sum_{l<i\le j,(j,l)\in\Omega_k}$ in equations (13c) and (13d). All other components are unchanged in equations (12a) to (15d), and their underlying rationale still works in the two-directional case.

The maximum allowable headway h_{max}^k in each period (assumed to be h^k, the headway in the operation during each period) is determined with following equations:

$$h_i^k = \min\{h_{i,up}^k, h_{i,dn}^k\}, \qquad \forall E_1^k + 1 \le i \le E_2^k \tag{16'}$$

$$h^k = h_{max}^k = \min\{h_{E_1^k+1}^k, h_{E_1^k+2}^k, \dots, h_{E_2^k}^k\} \tag{17'}$$

The round-trip time of a train during each period is:

$$R^k = 2[\frac{\sum_{i=E_1^k+1}^{E_2^k} d_i}{V_{tr}} + (E_2^k - E_1^k + 1)t_d + t_{dt}] \tag{18'}$$

The required fleet size in each period is still given by $N^k = R^k/h^k$.

The approximate PV of consumer surplus and fare collected from passengers in Period k are still given by equations (20) and (21), respectively. While the

approximate PV of total vehicle operation cost during Period k is given by equation (22), $\sum_{i=2}^{k} d_i$ is replaced with $\sum_{i=E_1^k+1}^{E_2^k} d_i$ in the modified equation (23) that gives the approximate PV of total track maintenance cost during Period k. For the PV of construction costs at the start of Period k (also at the end of Period $(k-1)$, $1 \le k \le k_{max}$), equation (25) is modified:

$$P_{co}^{k} = \begin{cases} \dfrac{c_{st}\left(E_1^{k-1} - E_1^{k}\right) + c_{ln}\sum_{i=E_1^k+1}^{E_1^{k-1}} d_i + c_{end}}{(1+r)^{t_k}}\gamma^k, \ \ if\ E_1^k < E_1^{k-1} \\ \dfrac{c_{st}\left(E_2^{k} - E_2^{k-1}\right) + c_{ln}\sum_{i=E_2^{k-1}+1}^{E_2^{k}} d_i + c_{end}}{(1+r)^{t_k}}\gamma^k, \ \ if\ E_2^{k-1} < E_2^{k} \end{cases}, \qquad (25')$$

$$\forall 1 \le k \le k_{max}$$

When $t_k = T$, $\gamma^k = 0$, and when $t_k < T$, $\gamma^k = 1$. This indicates that if Period k is not realized within the analysis period (has a duration of zero) for some extension plan, the construction costs are not incurred in this period.

To simplify calculation in the two-directional case, the PV of initial cost of additional trains (P_{in}^{k}) is not considered. It is assumed that the initial cost has been first converted into the vehicle operation cost.

In the evaluation of each chromosome (extension plan) in GA, completion time t_k of each potential extension step is numerically found sequentially using the binding constraint of available budget:

$$F_0 + Ft_{k+1} + \rho\sum_{i=0}^{k} P_f^i (1+r)^{\frac{t_i + t_{i+1}}{2}} - \sum_{i=0}^{k} P_{co}^{i+1}(1+r)^{t_{i+1}} = 0, \qquad (29')$$

$$\forall 0 \le k \le k_{max} - 1$$

This assumes that right after completion of each group of stations and links, the amount of available budget for construction reaches zero. With $k = 0$ in equation (29') t_1 is found first, given $t_0 = 0$. Then, with $k = 1$, t_2 is found given t_1, with $k = 2$, t_3 is found given t_2, and so forth. As long as $t_k < T$, this search continues until $t_{k_{max}}$ is found. Once some $t_{k'} > T$ is found, stop the search and let $t_k = T$ for $k' \leq k \leq k_{max}$. Upon finding each $t_k < T$, P_{CS}^{k-1}, P_f^{k-1}, P_o^{k-1}, P_m^{k-1}, and P_{co}^k are calculated. If $t_{k_{max}} < T$, non-zero $P_{CS}^{k_{max}}$, $P_f^{k_{max}}$, $P_o^{k_{max}}$, and $P_m^{k_{max}}$ are calculated. When some $t_{k'} > T$ is found, let $t_{k'} = T$, $T_{k'-1} = T - t_{k'-1}$, $P_{co}^{k'} = 0$, and non-zero $P_{CS}^{k'-1}$, $P_f^{k'-1}$, $P_o^{k'-1}$, and $P_m^{k'-1}$ are calculated. Then, let $P_{CS}^{k-1} = P_f^{k-1} = P_o^{k-1} = P_m^{k-1} = P_{co}^k = 0$ for all $k' + 1 \leq k \leq k_{max}$, and let $P_{CS}^{k_{max}} = P_f^{k_{max}} = P_o^{k_{max}} = P_m^{k_{max}} = 0$. The overall NPV over the analysis period is given by:

$$P_{NB} = \sum_{k=0}^{k_{max}} (P_{CS}^k + P_f^k - P_o^k - P_m^k) - \sum_{k=1}^{k_{max}} P_{co}^k \qquad (27')$$

In the two-directional extension problem, the objective is to find the optimal extension plan of the rail transit line that maximizes P_{NB} for a given analysis period. This optimization search is done by the customized GA method presented in Chapter 3.2.

Chapter 3: Optimization Methods

3.1 Differential Evolution – for one-directional extension

For the problem with one-directional extension, the optimization model is coded in Python (Version 3.7.3) and run on Spyder IDE. Due to binary variables γ^k and δ^k, the objective function (OF) is not continuous throughout the feasible solution space. For example, given that the solution vector (with decision variables as elements) is feasible, if t_k is the only changing decision variable and others are fixed, then the OF value (P_{NB}) changes continuously when $t_{k-1} < t_k < t_{k+1}$. However P_{NB} shifts downward when $t_k = t_{k-1}$ or $t_k = t_{k+1}$ due to a reduction in the cost of terminal facilities. These discontinuous shifts divide the feasible solution space into multiple zones, inside each of which the OF is continuous. The global optimum could exist in any of these zones. Therefore, instead of gradient-based methods, heuristic methods are preferable since they can deal with a discontinuous OF.

The synthetic one-directional extension problem (see Chapter 4.1) is discussed in a small problem size. For solving the problem, Differential Evolution (DE), a general-purpose heuristic method, is used by importing package "Optimize" from the Python library "SciPy". This is a stochastic direct search method developed by Storn and Price (1997). In this method, an initial population of solution vectors is randomly generated. In each generation, each vector in the population is a target, for which a mutant vector is generated based on other vectors in the population and the mutation constant. The crossover process swaps elements between the mutant vector and the

target at random positions, and generates a trial vector. If the trial vector improves the OF value, it replaces the target vector and will be passed on to the next generation. Iterations continue until the relative tolerance for convergence of vectors in the population is reached. This method does not guarantee finding the exact global optimum, but Storn and Price (1997) have tested it with various types of multidimensional functions (including discontinuous ones), and it successfully converges to the known global optima in all cases and all test runs.

When coding this model, some modifications are made to facilitate optimum search. Let $T_k = 0$ and $\delta^k = 0$ when $t_{k+1} - t_k < 0.01$, and $\gamma^k = 0$ when $T - t_k < 0.01$, so that the stochastic global search of solution vector is more likely to hit solutions where $\delta^k = 0$ (some stations will be completed together) or $\gamma^k = 0$ (some stations will not be completed within analysis period). To penalize infeasible solutions, their corresponding OF values are set to zero.

3.2 Customized Genetic Algorithm – for two-directional extension

For a problem with a larger scale (e.g., with more than ten planned stations), DE becomes inefficient because the computation time rapidly increases with the number of decision variables. Given the same number of planned stations, when the optimization of the extension plan is considered for both ends, each having multiple planned stations and links, the number of all possible extension plans (with different sequences and combinations of completion of stations and links) is far larger than in the case where extension is considered for only one direction. The DE method and its corresponding model formulation are not applicable in the two-directional version. A customized genetic algorithm (GA) is proposed here for searching a sub-optimal two-

directional extension plan. When applying a GA, the problem formulation is modified as mentioned in Chapters 2.6 and 2.7.

The whole GA module with the modified mathematical model is coded in Python (Version 3.7.3) and run on Spyder IDE. The flowchart of this customized GA is shown in Figure 6.

At first, an initial population (Generation 0) is generated. Individuals in this generation are first evaluated for their fitness values. A small fraction of individuals with best (largest) fitness values are reserved for the next generation. Then some individuals are selected as "parents" based on their fitness values, and "children" are generated using the crossover operator and the mutation operator. The next generation is generated when the total number of individuals (reserved best individuals and newly generated "children") reaches *pop_size.* For each generation thereafter, GA operators (evaluation, selection, crossover, mutation) are applied to individuals so that its next generation is generated. This iteration continues until the best fitness value in a generation remains unimproved for a certain number (denoted as *max_stall*) of generations or the maximal iteration count (denoted as *max_iter*) is reached.

3.2.1 Initialization of Population

The genetic algorithm starts with an initial population with a certain number (denoted as *pop_size*, usually an even number between 20 and 50) of individuals. Each individual is represented by a chromosome with two rows of integer. When a single rail transit line has n_1 planned stations at one end and n_2 planned stations at the other, each row has $(n_1 + n_2)$ locations with integers. Integers in Row 1 represent

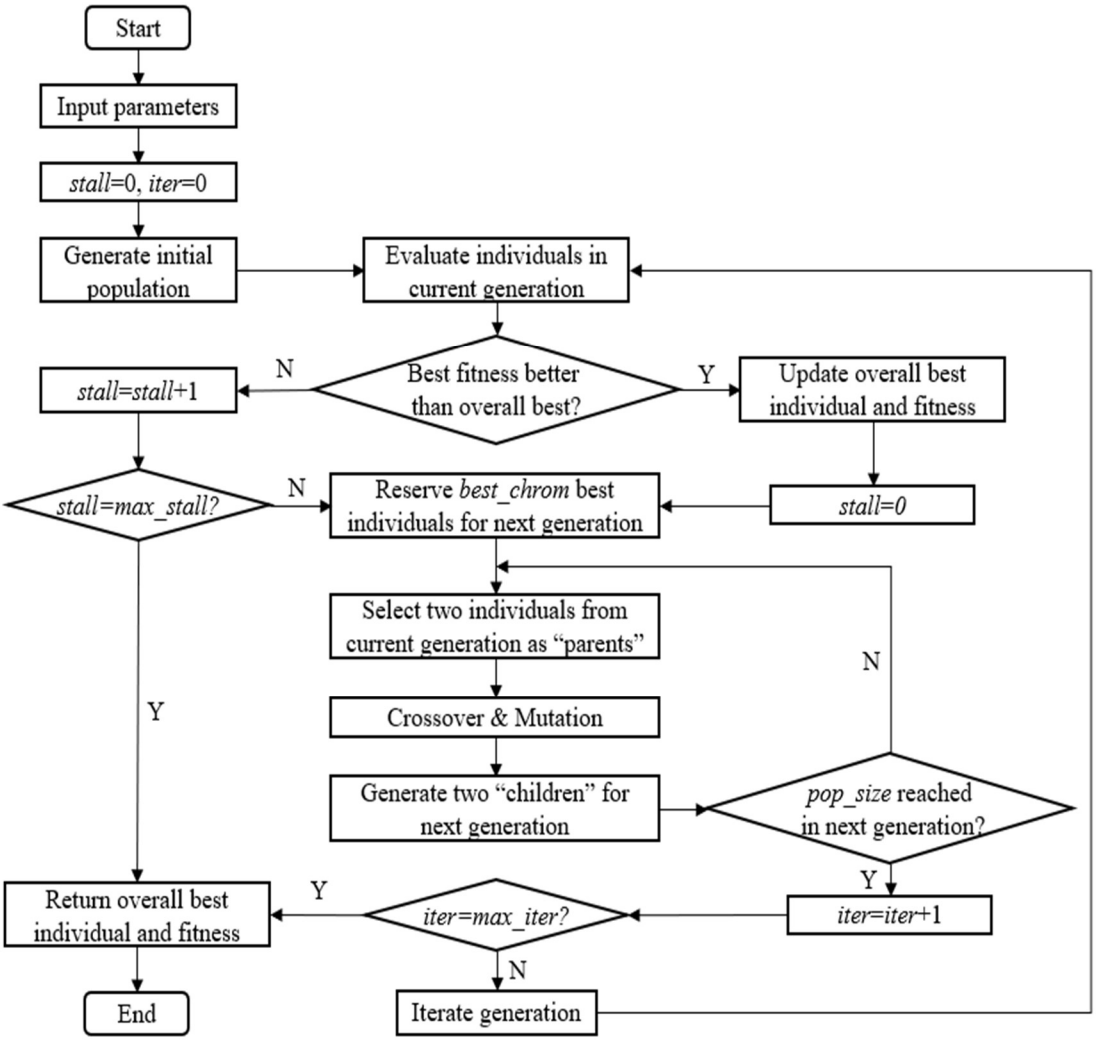


Figure 6 Flowchart of customized GA

the sequence of planned stations to be completed in the future, while the binary ones in the Row 2 indicate groups of stations to be completed. In Row 2, each integer is either 1 or 0. If the integer at a certain location of Row 2 is 1, it indicates that the station represented by the integer at the same location of Row 1 will be completed together with the station represented by the integer at the preceding location. Figure 7 shows an example of a chromosome, given $n_1 = n_2 = 3$ and $n_e = 4$ existing stations. This example indicates an extension plan where Stations 2 and 3 will be completed

together first, then Stations 8 to 10 will be completed together, and finally Station 1 will be completed.

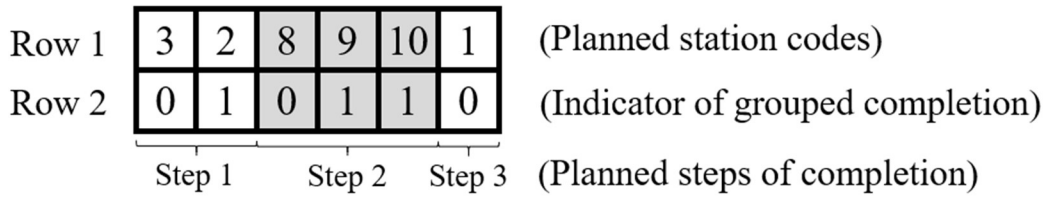


Figure 7 Example of a Chromosome

To randomly generate an individual, n_1 locations randomly selected out of $(n_1 + n_2)$ locations are assigned integer 1, and the remaining n_2 locations are assigned integer 2. Then, for each location (except the first location) in Row 1 that shares an integer with the preceding location, randomly assign either 0 or 1 (with equal probability) to this location in Row 2. Other locations in Row 2 are assigned 0. Here it is assumed that completion of multiple stations can be grouped in one extension step only when they are at the same end of the rail transit line. Finally, from the left to the right, replace n_1 integers 1 in Row 1 with $n_1, n_1 - 1, \ldots, 1$, and replace n_2 integers 2 in Row 1 with $m - n_2 + 1, m - n_2 + 2, \ldots, m$. These steps generate an individual (chromosome) that represents a possible extension plan. These steps are looped for *pop_size* times to initialize the population of Generation 0.

3.2.2 Fitness Value Evaluation

After each generation is created, the GA evaluates the fitness values of its individuals The fitness value of each individual is equivalent to the NPV incurred within the analysis period under the extension plan that individual represents.

The evaluation of each individual takes the following steps. First, each chromosome is decoded into an extension plan with multiple potential periods, each

having temporary terminal stations (with station codes E_1^k and E_2^k) at both sides. The number of potential periods equals the number of integer 0's in Row 2 (which equals k_{max}) plus 1. Figure 3-3 shows an example, given $n_1 = n_2 = 5$ and 4 existing stations. The chromosome below indicates an extension with 6 potential periods (k_{max} = 5). Temporary terminal station codes in Period 0 are E_1^0 = 6 and E_2^0 = 9, which are terminals of the line without any extension. After the first extension, in Period 1 the temporary terminal station at one end is updated to Station 4 (E_1^1 = 4), while the other terminal remains Station 9 (E_2^1 = 9). In each period after each extension step, only one of two temporary terminal stations is changed, as is highlighted in Figure 8.

In the second step, the completion times (t_k, $1 \leq k \leq k_{max}$) is determined for each planned step in an extension plan. The following assumption is applied: upon each completion of stations and links, the budget constraint is binding. That is, the available budget stays non-negative during the whole analysis period and reaches zero at each completion time. Using the "fsolve" function in the Python package of "scipy.optimize", the first completion time t_1 is numerically found such that the available budget (as shown in equation (29)) reaches zero upon the completion. Given

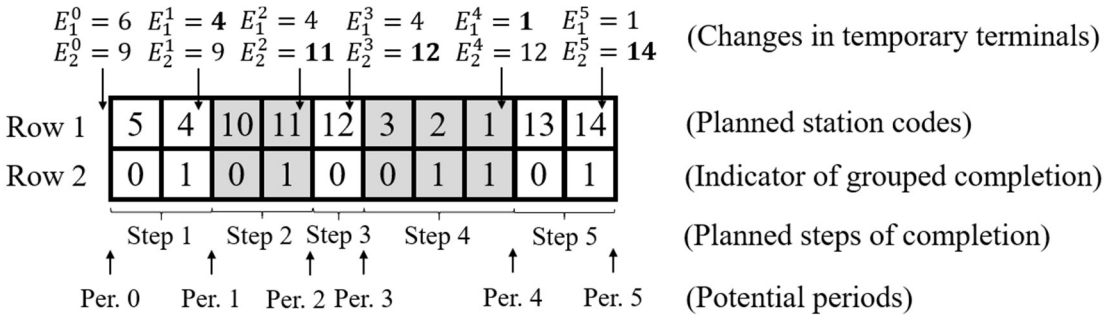


Figure 8 Decoding a Chromosome into Potential Periods and Temporary Terminals

t_1, Period 0 starts at time 0 and ends at time t_1, and has a duration of $T_0 = t_1$. Then, for each t_k given ($1 \leq k \leq k_{max} - 1$), the $(k+1)$ th completion time t_{k+1} is numerically found such that the available budget reaches zero right after this completion. If $t_{k_{max}} \leq T$, then all completion times are within the analysis period. In this case, Period k ($1 \leq k \leq k_{max} - 1$) has a duration of $T_k = t_{k+1} - t_k$, and Period k_{max}, which starts at time $t_{k_{max}}$ and ends at time T, has a duration of $T_{k_{max}} = T - t_{k_{max}}$. Note that the periods in an extension plan are treated as "potential" because the last periods may not be realized within the analysis period. During the numerical search, if there exists some $t_{k'} > T$ ($1 \leq k' \leq k_{max}$), then let $t_k = T$ for $k' \leq k \leq k_{max}$, and end the search. In this case the last period to be realized within the analysis period is Period $(k' - 1)$, which starts at time $t_{k'-1}$ ($t_0 = 0$) and ends at time T. It is given that $T_{k'-1} = T - t_{k'-1}$. If $k' \geq 2$, then for all $0 \leq k < k' - 1$, it is given that $T_k = t_{k+1} - t_k$.

The third step is to calculate the components of NPV (including present values of consumer surplus, fares collected, and supplier's costs) incurred during each of the realized periods within the analysis period using the model presented in Chapter 2, and aggregate to obtain the NPV incurred during the whole analysis period for the extension plan. Note that the construction cost is not counted in the last realized period. The NPV of this extension plan is used as the fitness value of its corresponding individual (chromosome).

3.2.3 Selection of "Parents"

After evaluating all individuals in a generation, "parent" individuals are selected for generating "children" in the next generation. Before selection, the fitness values of

individuals in this generation are sorted in a descending order, and directly succeed a certain number (denoted as *best_chroms,* usually an even number smaller than half of *pop_size*) of individuals with highest fitness values to the next generation. Then, the number of "children" needed in the next generation equals to *pop_size* minus *best_chroms*. The same number of "parents" are selected in the current generation. Each individual may be selected multiple times and some "parent" individuals may be duplicate.

To mimic the natural selection process, individuals with higher fitness values should be given higher probabilities of being selected as "parents". For this optimization problem, the selection probabilities are based on the fitness rankings of individuals. With this method, the selective pressure (in other words, the dominance of individuals with higher fitness value over those with lower fitness value in terms of selection probabilities) keeps constant from generation to generation and is not affected by absolute differences of fitness values (Whitley, 1989). If selection probability is directly based on fitness values rather than the ranking, then as iterations proceed, the absolute gaps of fitness values among individuals tend to shrink, which reduces selective pressure and leads to higher chance of staleness and prematurity.

The ranking-based selection probabilities are determined as follows. It is assumed that the individual with the highest fitness value has the ranking value of 1. Let i be the ranking value (an integer between 1 and *pop_size*) of an individual in the current generation. The selection probability of this individual is given by:

$$p_i = \frac{\alpha(1-\alpha)^{i-1}}{1-(1-\alpha)^{pop\_size}}$$

where $0 < \alpha < 1$. A greater α poses greater selective pressure. The value of α should be carefully determined. If α is too large, there is excessive selective pressure that leads to extremely low selection probability of individuals with lower fitness values and limits the solution searching breadth of GA. If α is too small, the pace of solution improvement is retarded. With a proper value of α, while better individuals are more likely to be chosen and generate potentially better offspring, worse individuals still have a non-negligible chance to pass on their potentially beneficial components to the next generation.

In each selection operation, two different "parent" individuals are selected from the current generation with their corresponding probabilities. After potential crossover and mutation, they produce two "children" for the next generation. These two "parents" are replaced into the population for the next selection. The operation loop of selection-crossover-mutation is executed (*pop_size* - *best_chroms*)/2 times until the number of individuals in the next generation reaches *pop_size*.

3.2.4 Crossover operator

Each pair of selected "parent" individuals (chromosomes) in the current generation is processed by the crossover operator. The crossover operator deals with two "parents" at a time and produces two "children". The probability that crossover between two "parents" actually occurs is given by a parameter *p_c*. Before the crossover operation, a number uniformly distributed between 0 and 1 is randomly generated. Crossover will actually occur only if this number is smaller than *p_c*. Otherwise, no crossover occurs and the "children" are identical to their "parents" before possible mutation.

The whole process of crossover is illustrated in the example (where $n_1 = n_2 = 5$ and there are 4 existing stations coded 6 to 9) shown in Figure 9.

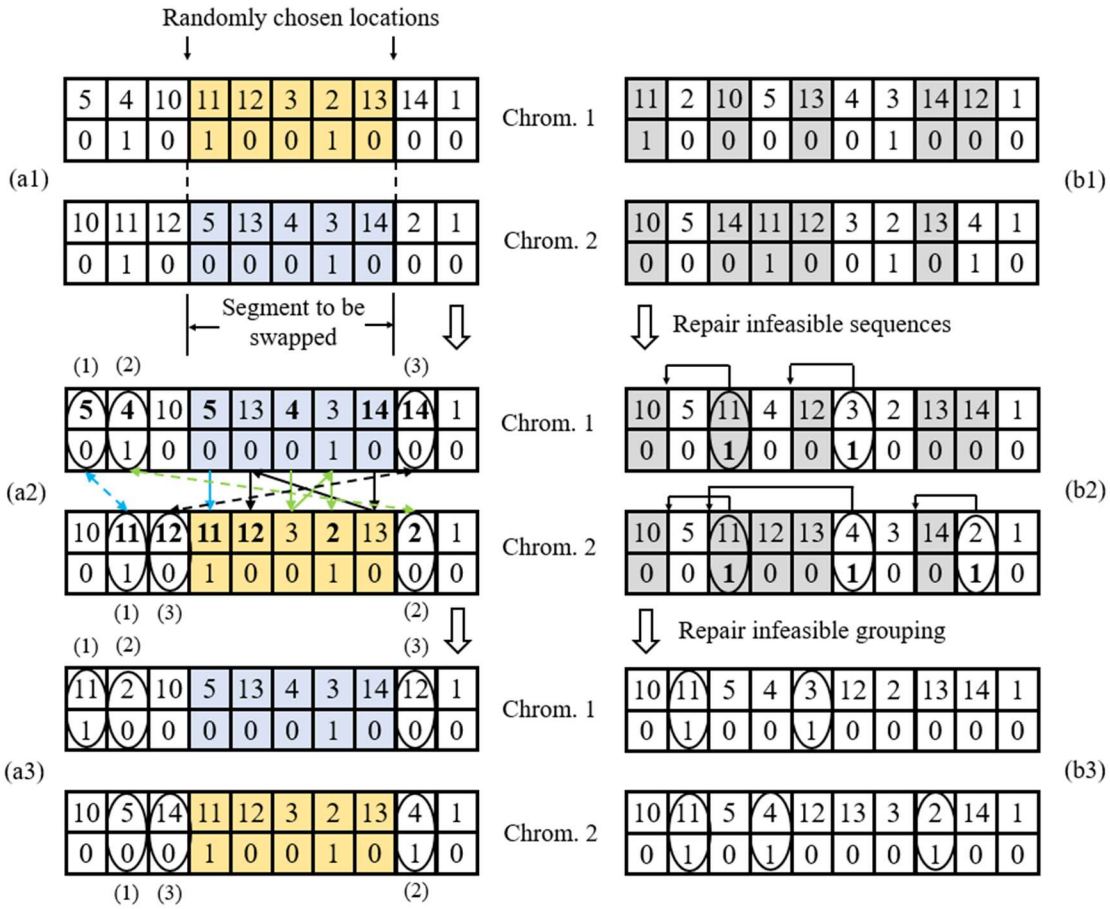


Figure 9 Process of PMX crossover and repairing infeasible results

The first step of crossover is to swap segments in two chromosomes. In a chromosome with a length of $(n_1 + n_2)$ integers in each row, two different locations are randomly chosen in $(n_1 + n_2 + 1)$ possible locations (including two ends). For each of two chromosomes to be operated, the segment between these two locations is swapped with that segment in the other chromosome, as shown in (a1) and (a2) of Figure 9. This may create infeasible chromosomes with duplicate station codes in Row 1, and the Partial Mapped Crossover (PMX) method, proposed by Goldberg and

Lingle (1985), is applied to fix the error. The mapping process is shown in the example in (a2) of Figure 9. In this example, station codes 5, 4, and 14 are duplicate in Chromosome 1, and station codes 11, 12, and 2 are duplicate in Chromosome 2. As shown by solid arrows, the mapping relation is set up by examining station codes at locations within the swapped sections. For duplicate station code 5 in Chromosome 1, the station code at the same location in Chromosome 2 is 11, which is not found in the swapped section in Chromosome 1. Thus, station code 5 is mapped to 11. For duplicate station code 4 in Chromosome 1, the station code at the same location in Chromosome 2 is 3, which is found in the swapped section in Chromosome 1. For the code 3 in Chromosome 1, the station code at the same location in Chromosome 2 is 2, which is not found in the swapped section in Chromosome 1. Thus, station code 4 is mapped to 2. Similarly, station code 14 is mapped to 12. After the mapping relation is determined, the mapping station codes are found outside the swapped segments, as shown by dashed arrows in (a2). Then these station codes are swapped together with the indicators at the same locations in Row 2, between two chromosomes. The result is shown in (a3), without any duplicate station codes in Row 1 in each chromosome.

After these operations, the resulting chromosomes may still be infeasible. These chromosomes should be fixed further to get rid of infeasible completion sequences and infeasible grouping of completion.

First, completion sequences are repaired. In the example in (b1) of Figure 9, station codes that belong to the same end of the line are highlighted with the same color. Then, station codes that belong to the same end are rearranged so that the order of station completion becomes feasible for this end (for example, the order {11, 10, 13,

14, 12} in Row 1 of Chromosome 1 is rearranged to {10, 11, 12, 13, 14}) , while the set of locations these station codes occupy in this chromosome is not changed. The indicator values in Row 2 move together with their corresponding station codes in Row 1. The resulting chromosomes are shown in (b2) of Figure 9.

Next, the grouping of completions (links and stations to be completed at each extension step) is repaired. The 1's in the indicators in Row 2 of each chromosome are checked. If the corresponding station code and that station code at the previous location do not belong to the same end of the line, this indicator 1 is infeasible because it is assumed that multiple links and stations to be completed in one extension step must belong to the same end. Each infeasible indicator 1 in Row 2 is moved together with its corresponding station code in Row 1 to a destination location such that this station code, together with that station code at the previous location of this destination location, belong to the same end of the line, as shown in (b2). The resulting chromosomes in (b3) are the final products of the crossover of two "parents", if crossover occurs.

3.2.5 Mutation operator

"Children" individuals may experience mutation before they are finally passed on to the next generation. The probability that mutation of a "child" actually occurs is given by a parameter p_m. Before the mutation operation, a number uniformly distributed between 0 and 1 is randomly generated. Mutation will actually occur only if this number is smaller than p_m. All possible mutation cases are illustrated in examples (where $n_1 = n_2 = 5$ and there are 4 existing stations coded 6 to 9) in Figure 10.

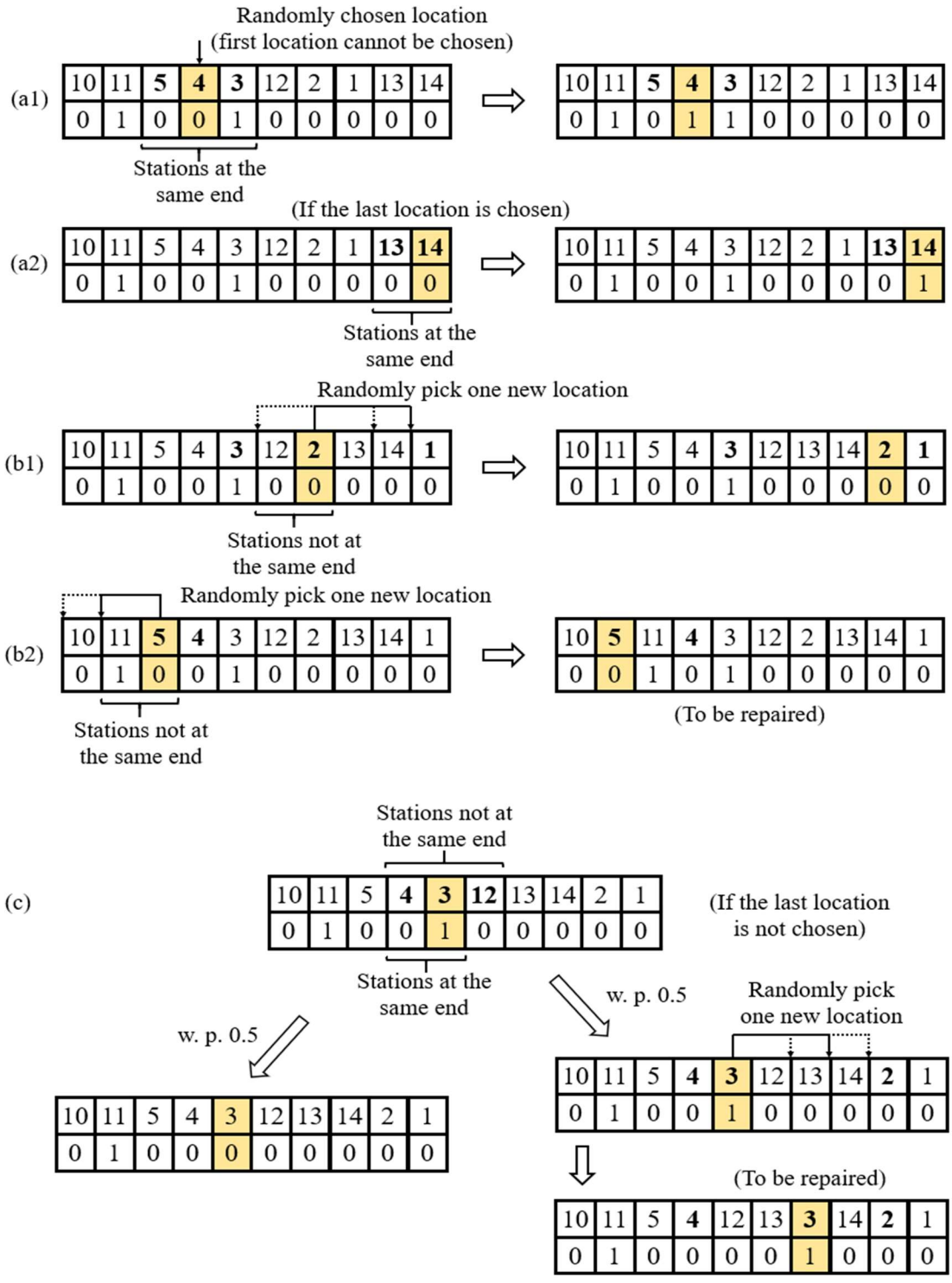


Figure 10 Types of operations in the mutation operator

In a mutation operation, a location (except the first location) is randomly selected in a chromosome. Depending on whether the station code at the chosen location shares one end of the line with codes at neighboring locations, there will be three types of possible operations:

1) If the chosen station code shares one end of the rail transit line with the previous code and the next code, then the chosen indicator in Row 2 is changed from 0 to 1 (or from 1 to 0), as is highlighted in (a1) in Figure 10. If the last location is chosen, then the indicator is changed if the chosen station code shares one end with the previous code, as shown in (a2). This operation type yields the final result of mutation.

2) If the chosen station code does not share one end of the line with the previous code, then the chosen station code and indicator are moved to a randomly chosen new location such that the station completion sequence that the chromosome represents is still feasible after this move. As illustrated in (b1) and (b2) of Figure 10, possible moves are shown by dashed arrows while the actual move is shown by the solid arrow.

3) If the last location is not chosen, and the chosen station code shares one end of the line with the previous code but not the next one, as shown in (c), then a random number uniformly distributed between 0 and 1 is generated. If it is smaller than 0.5, then the chosen indicator in Row 2 is changed from 0 to 1 (or from 1 to 0). The operation type in 1) is used, and the final result of mutation is obtained. If it is larger than 0.5, then the chosen elements are moved to a new feasible location. The operation type in 2) is used.

The operation type in 2) may produce chromosomes that represent infeasible groups of stations to be completed. These chromosomes are repaired using the method shown in Figure 11. First, the infeasible indicator 1's that group stations in different ends into one completion step are counted and located. Then, the indicator 0's that can be changed into 1's without producing infeasible completion groups are counted and located. If the number of changeable 0's (denoted here as a) exceeds that of infeasible 1's (denoted here as b), then b randomly chosen changeable 0's out of a are changed into 1's, and all b infeasible 1's are changed into 0's. If $a \leq b$, all changeable 0's are changed into 1's and all infeasible 1's are changed into 0's. After this correction the final result of mutation is obtained.

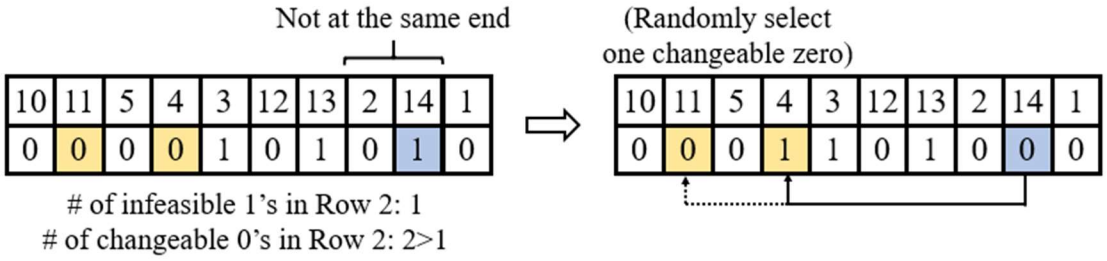


Figure 11 Repairing infeasible results of mutation

Chapter 4: Numerical Results

4.1 For one-directional extension problem

4.1.1 Solving the problem in a base scenario

A numerical case is synthesized to demonstrate the model for one-directional extension problem, with a single rail transit line similar to that line shown in Figure 1. In this small-scale problem, there are 9 stations and 8 links in a rail transit line. 4 stations (Station 1 to 4) and 3 links (Link 2 to 4) in the CBD are currently in operation. 5 stations (Station 5 to 9) and 5 links (Link 5 to 9) may be completed in the upcoming analysis period of 10 years. Link lengths and potential demand values in the base scenario are listed below.

Table 2 Link lengths and potential demand values in the base scenario (for one-directional extension)

j	1	2	3	4	5	6	7	8	9
d_j/mi		1	0.8	1.2	1.2	1.5	2	2.5	2.2
Q_{1j}	0	950	1030	1120	1070	1310	840	670	990
Q_{2j}	950	0	880	1150	930	1100	880	670	900
Q_{3j}	1030	880	0	990	820	1150	740	700	870
Q_{4j}	1120	1150	990	0	850	1080	890	720	850
Q_{5j}	1070	930	820	850	0	870	750	580	780
Q_{6j}	1310	1100	1150	1080	870	0	700	600	790
Q_{7j}	840	880	740	890	750	700	0	540	750
Q_{8j}	670	670	700	720	580	600	540	0	600
Q_{9j}	990	900	870	850	780	790	750	600	0

Values of b_{ij} are given by:

$$b_{ij} = 4 + 1.75 \sum_{l=i+1}^{j} d_l, \quad \forall i < j$$

$$b_{ij} = b_{ji}, \qquad \forall j < i$$

$$b_{ij} = 4, \qquad \forall i = j$$

The optimization model coded in Python 3.7.3 is run on a personal laptop with an Intel® Core™ i7-8750H CPU @ 2.20GHz. For this numerical case, the parameters of the DE method are default values (given by the SciPy package "Optimize") except for the relative tolerance, which is set at 0.001 instead of 0.01. When the standard deviation of fitness values (NPV) in the population is not greater than the relative tolerance times the absolute value of mean of fitness values, the search stops. The initial vector population has a size of 400 and each individual is randomly generated under the sequence and domain constraint (28) in each model run.

The model is run 10 times on the base scenario. The average computation time per run is 211.8 seconds. With different initial vector populations, all the runs returned results with nearly identical "optimal" solutions, which indicates that the vector population is very likely to have converged at the global optimum. The slight differences in 10 optimized solutions are due to the random nature of DE search. These solutions can be treated as the same solution.

The following strategy is used to obtain the "processed" optimized solution. For one of 10 "raw" solutions, its completion time values are checked first. If some completion times are larger than (T-0.01), the corresponding stations will not be completed within the analysis period (see Chapter 3.1), and these values are adjusted to T. For two neighboring completion times that differ by less than 0.01, the corresponding stations will be completed together (see Chapter 3.1), and these values are adjusted to the larger one. Next, the remaining available budget after completion

time of each extension step is checked. If the remaining available budget after an extension step is less than some value (say, $\$1.0\times10^6$) that is far smaller than the optimized NPV, it is assumed that the available budget constraint is binding for this step, its completion time value is decreased by a tiny margin until the remaining available budget is smaller than a value much closer to zero (say, $\$1.0\times10^3$). If the resulting NPV becomes larger than that of the "raw" solution, these adjustments are justified. This strategy is used for all optimized solutions in the one-directional extension problem.

For the base scenario the "processed" optimized solution is:

$$P_{NB} = \$4.530 \times 10^9, t_1 = 0.523, t_2 = 2.224, t_3 = 3.978, t_4 = 5.778, t_5 = 10$$

which means that, Station 9 (and Link 9) should not be completed within the analysis period, while Stations 5 to 8 and their corresponding links should be completed in year 0.523, 2.224, 3.978, and 5.778 into the analysis period, respectively. Periods 0 to 4 should last 0.523, 1.701, 1.754, 1.800, and 4.222 years, respectively. The maximal overall NPV over 10 years is $4.530 billion. Under this extension plan, the available budget will be very close to zero (below $\$1 \times 10^3$) after each extension, showing that in this base scenario, construction of a certain station (except Station 9) should be completed as soon as the available budget is sufficient to cover the construction costs of the station, with its corresponding link and terminal facilities. In other words, constraint (29) is binding for all planned extension steps. Based on this observation an assumption is made for determining completion times in the two-directional extension problem: upon completion of each extension step, the constraint of available budget is binding.

4.1.2 Effects of terminal cost and analysis period duration

In the base scenario optimum, Station 9 is not completed within 10 years. However, if the analysis period is extended to 15 years with other conditions unchanged, the optimized solution suggests that Station 9 should be completed 9.915 years into the analysis period, while the completion times of other stations are unchanged. It is noted that when Station 9 is completed, the available budget constraint (29) is not binding for this extension step. When the hourly train operation cost is halved (i.e., let c_o=2500), however, the available budget constraint becomes binding and the optimized value of t_5 reaches the minimum 7.277. When c_o is smaller than 2500, the available budget constraint remains binding. When c_o =3000, this constraint is not binding again. As c_o increases above 3000, the optimized value of t_5 increases. Here is an explanation for this observation. It is assumed in this model that the required fleet size in each period is decided by the peak passenger flow at the end of the period. Within the last period (Period 5), while the total hourly train operation cost remains constant (with constant fleet size, without discounting to PV), the total hourly ridership at the beginning of this period is lower than that at the end (given constant annual growth of ridership). An earlier start of the last period leads to a lower hourly ridership at the beginning, and therefore a lower hourly value of consumer surplus and collected fare (without discounting to PV). The total hourly train operation cost (without discounting to PV), however, does not change with the completion time of Station 9. As a result, under the objective of maximizing the NPV within the analysis period, a higher hourly train operation cost (c_o) discourages the supplier from completing the last planned station earlier. In the two-directional extension problem, the assumption that all planned

extension steps have the constraint (29) binding is more justified with a lower value of c_o.

Station 5 to 8 should be separately completed in the base scenario optimum. With other conditions unchanged, the value of c_{end} is changed to $\$1.8\times10^8$, six times of that in the base scenario. The optimized extension plan, with such high terminal facility costs, is to complete Station 5 and Station 6 together 3.860 years into the analysis period, and to complete Station 7 6.857 years into the analysis period. If T is further changed to 15, then Station 5 and 6 should be completed together 3.860 years into the analysis period, and Station 7 and 8 should be completed together 8.447 years into the analysis period. For all these extension steps, the available budget constraint is binding.

These results are shown in Figure 12. In can be learned that the analysis period duration can affect the optimized extension plan for stations and links. It can also be confirmed that higher costs of terminal facilities increase the economic advantage of completing multiple neighboring stations in a single step.

4.1.3 Analysis of Sensitivity to Selected Parameters

Next, the sensitivity of the optimized solution to various parameters is examined. The selected parameters include: c_{ln} (unit construction cost of rail line), F (yearly external budget supply), g (annual growth rate of potential demand), Q_{ij} (potential hourly

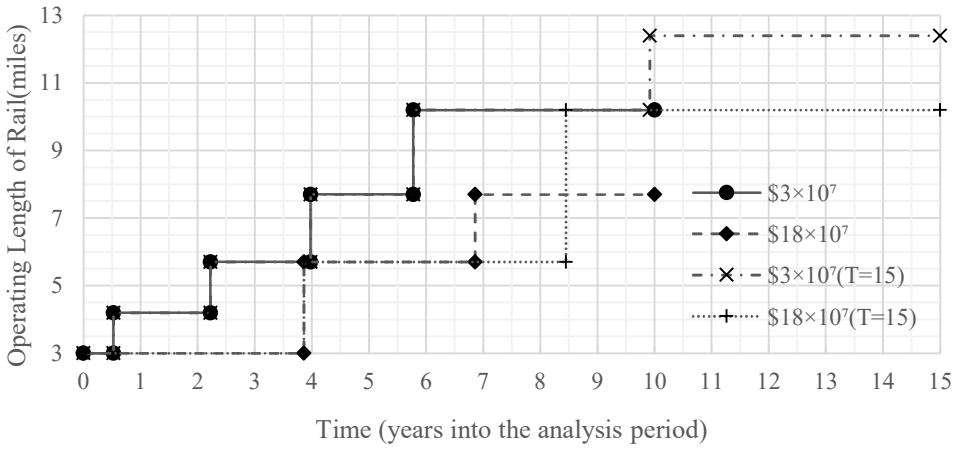


Figure 12 Optimized extension steps for different values of c_{end} and T

ridership), u_v (value of in-vehicle time), u_w (value of waiting time), and ρ (fraction of collected fare to be used for further construction). For each of these parameters, its value is slightly changed from its base scenario value (within ±20%), with other parameters unchanged (except that c_{st}, the construction cost of a station, changes proportionally with c_{ln}). The model is run to obtain the "raw" optimized solution which is then processed using strategies in Chapter 4.1.1. In all the modified scenarios, the optimized extension plan is to separately complete Stations 5 to 8 with the available budget constraint being binding and to not complete Station 9.

Despite similar extension steps, the maximized NPV (P_{NB}) shows different sensitivity to different parameters, and so does the completion time of Station 8 (t_4). Table 3 lists all changes of parameters in modified scenarios, the corresponding NPVs and their change rates from the base scenario, the corresponding t_4 values and their change rates from the base scenario, and the proportional change (elasticity, calculated from the ratio of percentage changes) of NPV and t_4 in response to the change of each parameter. (For each parameter the elasticity calculation uses the two

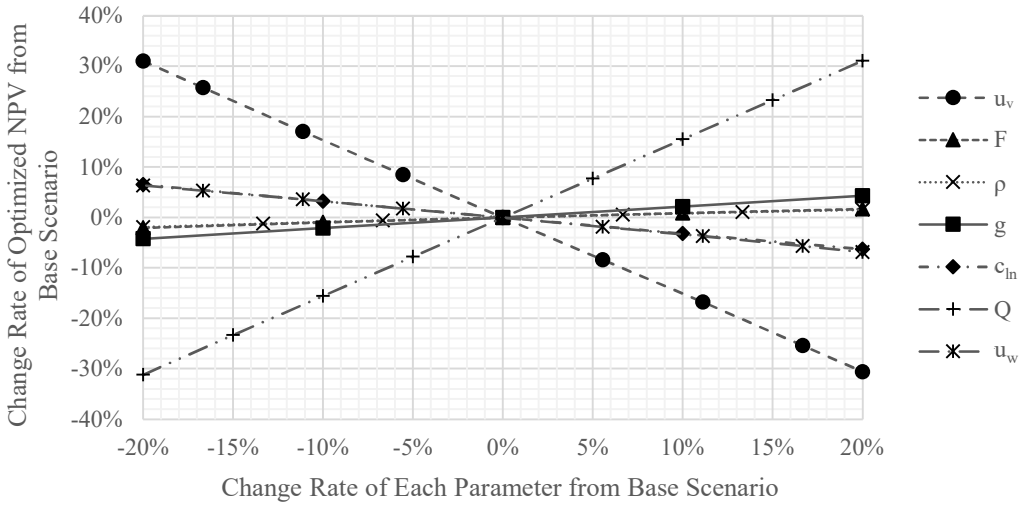
scenarios closest to the base value. For example, the elasticity of NPV to c_{ln} is calculated as follows. All other parameters unchanged, a decrease of c_{ln} by 10% from its base value leads to a 3.22% increase of NPV, while an increase of c_{ln} by 10% from its base value leads to a 3.18% decrease of NPV. The elasticity of NPV to c_{ln} is given by [-3.18%-3.22%]/[10%-(-10%)] = -0.320.)

The differences in sensitivity of optimized NPV and t_4 to various parameters are clearly presented in Figure 13. It is found that both the optimized NPV and the optimized t_4 are fairly sensitive to changes in hourly potential demand (Q_{ij}). In addition to the estimation of Q_{ij}, the rail transit operator should carefully determine the users' value of time, especially for in-vehicle time (u_v), whose small change significantly affects the estimated overall social benefit and the pace of line extension. While the optimized NPV is most sensitive to Q_{ij}, the optimized t_4 is most sensitive to c_{ln}, which means that a small uncertainty in major construction costs of links and stations leads to significant changes in the construction plan. The external budget supply (F) and the re-investment fraction of revenue (ρ) are also important parameters for scheduling phased development. Among all selected parameters, the effect of the yearly growth rate of demand (g) is the slightest.

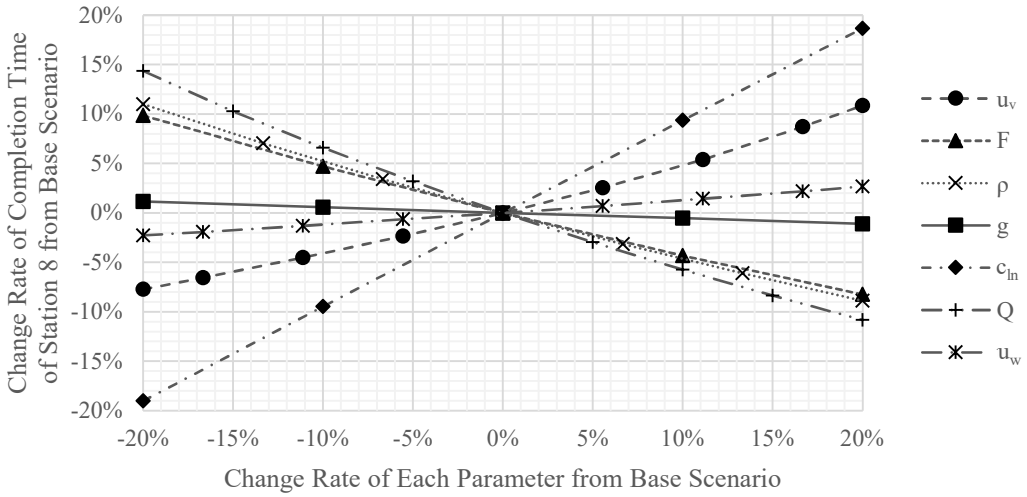
Table 3 List of changes of parameters in modified scenarios and corresponding changes of NPV and completion time of Station 8

Para-meter	Value	Change Rate	NPV (×\$$10^9$)	Change Rate	Elasticity of NPV	t_4 Value	Change Rate	Elasticity of t_4
c_{ln}	4.8×10^7	-20%	4.826	6.53%	-0.320	4.679	-19.01%	0.941
	5.4×10^7	-10%	4.676	3.22%		5.231	-9.45%	
	6.6×10^7	10%	4.386	-3.18%		6.318	9.36%	
	7.2×10^7	20%	4.245	-6.29%		6.856	18.68%	

F	4.0×10^7	-20%	4.436	-2.08%	0.094	6.347	9.87%	-0.452
	4.5×10^7	-10%	4.485	-0.99%		6.050	4.73%	
	5.5×10^7	10%	4.570	0.88%		5.528	-4.31%	
	6.0×10^7	20%	4.606	1.68%		5.301	-8.24%	
g	2.4%	-20%	4.338	-4.24%	0.214	5.844	1.16%	-0.057
	2.7%	-10%	4.433	-2.14%		5.811	0.59%	
	3.3%	10%	4.627	2.14%		5.746	-0.54%	
	3.6%	20%	4.726	4.33%		5.713	-1.11%	
Q	0.8 Q	-20%	3.117	-31.19%	1.552	6.607	14.37%	-0.613
	0.85 Q	-15%	3.473	-23.33%		6.372	10.30%	
	0.9 Q	-10%	3.826	-15.54%		6.159	6.61%	
	0.95 Q	-5%	4.178	-7.77%		5.961	3.19%	
	1.05 Q	5%	4.881	7.75%		5.607	-2.94%	
	1.1 Q	10%	5.233	15.52%		5.446	-5.73%	
	1.15 Q	15%	5.585	23.29%		5.294	-8.36%	
	1.2 Q	20%	5.938	31.08%		5.152	-10.82%	
u_v	14.4	-20%	5.935	31.02%	-1.516	5.331	-7.72%	0.441
	15	-16.67%	5.696	25.74%		5.398	-6.56%	
	16	-11.11%	5.301	17.02%		5.517	-4.50%	
	17	-5.56%	4.913	8.45%		5.642	-2.34%	
	19	5.56%	4.150	-8.39%		5.925	2.56%	
	20	11.11%	3.770	-16.78%		6.089	5.40%	
	21	16.67%	3.378	-25.43%		6.281	8.72%	
	21.6	20%	3.145	-30.57%		6.405	10.87%	
u_w	14.4	-20%	4.818	6.36%	-0.328	5.646	-2.27%	0.118
	15	-16.67%	4.771	5.32%		5.667	-1.90%	
	16	-11.11%	4.692	3.58%		5.702	-1.30%	
	17	-5.56%	4.611	1.79%		5.741	-0.62%	
	19	5.56%	4.446	-1.85%		5.817	0.69%	
	20	11.11%	4.361	-3.73%		5.860	1.44%	
	21	16.67%	4.273	-5.67%		5.904	2.20%	
	21.6	20%	4.219	-6.87%		5.932	2.68%	
ρ	12%	-20%	4.443	-1.92%	0.086	6.413	11.01%	-0.492
	13%	-13.33%	4.474	-1.24%		6.184	7.05%	
	14%	-6.67%	4.503	-0.60%		5.974	3.41%	
	16%	6.67%	4.555	0.55%		5.595	-3.15%	
	17%	13.33%	4.578	1.06%		5.425	-6.09%	
	18%	20%	4.600	1.55%		5.264	-8.88%	



(a)



(b)

Figure 13 Sensitivity of optimized NPV and optimized completion time of Station 8 to influential parameters

4.2 Results for two-directional extension problem

4.2.1 Solving the problem in a base scenario for different terminal costs

Another numerical case is synthesized for the two-directional extension problem. This problem has a larger scale, with m=20 stations and 19 links in a rail transit line similar to that line shown in Figure 5. n_e=4 stations (Stations 9 to 12) and 3 links

(Links 10 to 12) in the CBD are currently in operation. n_1=8 stations at End 1 (with codes 1 to 8), n_2=8 stations at End 2 (with codes 13 to 20) and their corresponding links may be completed in the upcoming analysis period of T=30 years. Link lengths and potential demand values in the base scenario are listed in Table 4. The synthetic potential demand matrix assumes that the existing segment (with 4 stations and 3 links) of the rail transit line is located in the CBD of the city, and the planned segments connect suburb residential areas. Commuting trip between the CBD and residential areas is assumed to be the dominating trip purpose, and stations closer to the CBD have higher rates of trip production and attraction. For most stations (especially those in the CBD), the potential demands of rail transit trips to nearby stations tend to be lower than those to farther stations, because for travels of shorter distances, using rail transit tends to save less travel costs (fare plus time) over walking and cycling, especially given the waiting time for trains.

In the two-directional extension problem, values of some parameters (as listed in Table 5) are different from those in the one-directional case. c_{in} is not used in the two-directional case. Other parameters use the same values as in the one-directional case.

The GA optimization model coded in Python 3.7.3 is run on a personal laptop with an Intel® Core™ i7-8750H CPU @ 2.20GHz. For this numerical case, GA parameters are set as shown in Table 6.

The model is run 10 times on the base scenario. The average computation time per run is 712.49 seconds, and the average iteration count is 48. Each iteration takes 14.84 seconds on average. With the *max_stall* of 30, the average iteration count

Table 4 Link lengths and potential demand values in the base scenario (for two-directional extension)

j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
d_j		2.0	2.0	1.5	1.2	1.1	1.0	0.8	1.0	0.8	0.8	1.0	1.0	1.1	1.2	1.2	1.4	1.0	1.6	1.8
Q_{1j}	0	50	60	120	100	150	200	305	335	440	470	360	305	255	200	135	110	70	90	55
Q_{2j}	50	0	45	90	70	140	190	270	370	490	465	365	350	295	225	180	135	85	80	65
Q_{3j}	60	45	0	90	90	160	225	305	375	575	505	375	335	280	240	160	145	95	95	90
Q_{4j}	120	90	90	0	75	110	185	240	350	490	480	340	280	255	185	145	105	75	130	110
Q_{5j}	100	70	90	75	0	95	185	315	400	585	490	405	295	280	210	150	140	105	110	120
Q_{6j}	150	140	160	110	95	0	145	245	370	495	510	375	280	270	210	185	145	125	145	135
Q_{7j}	200	190	225	185	185	145	0	185	335	450	510	400	310	270	230	210	150	175	175	185
Q_{8j}	305	270	305	240	315	245	185	0	280	425	415	310	305	305	265	230	240	225	250	270
Q_{9j}	335	370	375	350	400	370	335	280	0	335	370	295	280	335	350	315	360	330	350	335
$Q_{10,j}$	440	490	575	490	585	495	450	425	335	0	305	280	290	310	385	390	455	450	510	495
$Q_{11,j}$	470	465	505	480	490	510	510	415	370	305	0	265	310	350	360	430	440	505	560	520
$Q_{12,j}$	360	365	375	340	405	375	400	310	295	280	265	0	250	265	345	370	400	410	530	410
$Q_{13,j}$	305	350	335	280	295	280	310	305	280	290	310	250	0	210	270	265	280	350	375	360
$Q_{14,j}$	255	295	280	255	280	270	270	305	335	310	350	265	210	0	230	210	250	215	250	250
$Q_{15,j}$	200	225	240	185	210	210	230	265	350	385	360	345	270	230	0	175	195	150	130	120
$Q_{16,j}$	135	180	160	145	150	185	210	230	315	390	430	370	265	210	175	0	130	95	80	85
$Q_{17,j}$	110	135	145	105	140	145	150	240	360	455	440	400	280	250	195	130	0	70	70	55
$Q_{18,j}$	70	85	95	75	105	125	175	225	330	450	505	410	350	215	150	95	70	0	50	75
$Q_{19,j}$	90	80	95	130	110	145	175	250	350	510	560	530	375	250	130	80	70	50	0	65
$Q_{20,j}$	55	65	90	110	120	135	185	270	335	495	520	410	360	250	120	85	55	75	65	0

Table 5 Parameter values modified in the two-directional extension case

Parameter	Baseline Value	Unit	Parameter	Baseline Value	Unit
c_m	500	\$/mile/hr	c_{end}	1.5×10^8	\$
c_{ln}	1.4×10^8	\$/mile	ρ	25%	
c_{st}	6×10^7	\$/mile	K	1280	psgrs

Table 6 GA parameters used for the base scenario

Parameter	Value	Parameter	Value
pop_size	40	*best_chroms*	4
max_iter	1000	*max_stall*	30
p_c	0.8	*p_m*	0.5
α	0.06		

needed for GA to attain the optimized chromosome is 18, which requires 720 evaluations of fitness values (NPV) of chromosomes. With different initial populations, 6 of 10 runs return the same optimized chromosome with the best (largest) fitness value in these 10 runs. The best chromosome is shown as:

13	8	7	14	15	6	5	16	17	4	3	18	19	2	1	20
0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0

which represents an extension plan with 9 potential extension steps. This plan can be denoted as the following array, where station codes inside each pair of round brackets are to be completed together, and extension steps are shown chronologically from left to right, separated by commas.

[(13), (8 7), (14 15), (6 5), (16 17), (4 3), (18 19), (2 1), (20)]

With the binding constraint on available budget, all these extension steps can be realized within the analysis period. Results show following completion times:

t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9
2.873	6.881	10.021	12.297	14.455	16.371	18.055	20.084	21.140

This means that, extension steps 1 to 9 should be completed in year 2.873, 6.881, 10.021, 12.297, 14.455, 16.371, 18.055, 20.084, and 21.140 into the analysis period, respectively. Periods 0 to 9 should last 2.873, 4.008, 3.140, 2.276, 2.158, 1.916, 1.684, 2.029, 1.056, and 8.860 years, respectively. For this optimized extension plan, $P_{NB} = \$15.781 \times 10^9$, which means the overall NPV over 30 years is \$15.781 billion.

4.2.2 Effects of terminal cost

Next, the effect of terminal cost (c_{end}) on the model and the optimized extension plan is examined. The original value of c_{end} is halved to 7.5×10^7 first, and then

doubled to 3.0×10^8, with other parameters unchanged. In each modified scenario the model is run 10 times.

When c_{end}=7.5×10^7, the average computation time per run is 850.40 seconds, and the average iteration count is 40.1. Each iteration takes 21.21 seconds on average. All 10 runs return the same optimized chromosome that represents the following extension plan:

[(13), (8), (14), (15), (7 6), (16), (5), (17), (4), (3), (18), (19), (2), (20), (1)]

The GA attains the optimized chromosome in fewer iterations, because most chromosomes in the initial population have much more 0's than 1's in Row 2, and attaining the optimized chromosome with 15 0's and only one 1 in Row 2 needs fewer mutations than attaining that with 9 0's and 7 1's in Row 2. Average calculation time per iteration increases, because as iterations proceed, chromosomes with more 0's in Row 2 are favored. Since these chromosomes represent more extension steps, more completion times need to be numerically determined.

All 15 extension steps can be realized within the analysis period at the following completion times:

t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}
2.022	4.256	6.124	7.754	9.825	10.999	12.030	13.129	14.071	15.084	15.851	16.813	17.877	18.826	19.809

For this optimized extension plan, the overall NPV over 30 years (P_{NB}) is \$17.067 billion.

When c_{end}=3.0×10^8, the average computation time per run is 698.50 seconds, and the average iteration count is 61.5. Each iteration takes 11.36 seconds on average. 9 out of 10 runs return the same optimized chromosome that represents the following extension plan:

[(13 14), (8 7 6), (15 16 17), (5 4 3), (18 19 20), (2 1)]

Compared to the case where c_{end}=1.5×10^8, the GA terminates after more iterations due to more mutations needed for attaining the optimized chromosome with 6 0's and 9 1's in Row 2. Average calculation time per iteration, however, becomes shorter than in the case where c_{end}=1.5×10^8. As iterations proceed, chromosomes with more 1's in Row 2 are favored. Since these chromosomes represent fewer extension steps, fewer completion times need to be numerically determined.

All 6 extension steps can be realized within the analysis period at the following completion times:

t_1	t_2	t_3	t_4	t_5	t_6
6.359	11.437	15.336	18.325	21.045	23.225

For this optimized extension plan, the overall NPV over 30 years (P_{NB}) is \$14.265 billion.

The optimized extension steps under different terminal cost values are shown in Figure 14. It explicitly reveals that higher costs of terminal facilities lead to fewer extension steps and more stations to be completed together in each step. Consistent with the one-directional case, a higher c_{end} increases the economic advantage of completing multiple neighboring stations in a single step. A higher c_{end} also leads to later completion of the rail transit line. For almost any given operating length, the optimized extension plan with higher c_{end} achieves this length later. The delayed coverage of the operating segment on OD pairs reduces total consumer surplus and total collected fare over the analysis period, resulting in a lower NPV for an extension plan with a higher c_{end}.

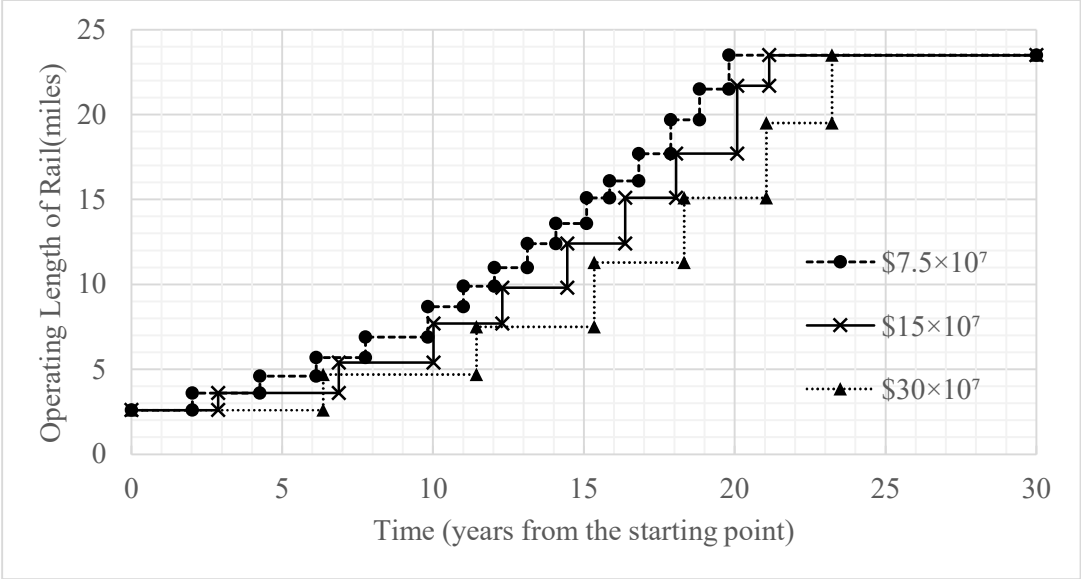


Figure 14 Optimized extension steps under different values of c_{end}

For the numerical case where $n_1 = n_2 = 8$, the total number of all possible permutations of chromosome is given by:

$$2 \times \left[\sum_{i=0}^{7} \binom{7}{i} \cdot \binom{7}{i} \cdot 2^{14-2i} + \sum_{i=0}^{6} \binom{7}{i} \cdot \binom{7}{i+1} \cdot 2^{13-2i} \right] = 3{,}968{,}310$$

For each numerical case in 4.2.2 with different values of c_{end}, 50,000 different chromosomes are randomly sampled from the full set of all 3,968,310 possible permutations and are evaluated. The distribution of fitness values (P_{NB}) regarding these sampled chromosomes in cases with c_{end} value of 7.5×10^7, 1.5×10^8, and 3.0×10^8, are shown in histograms (a), (b), and (c) in Figure 15, respectively.

It appears that no specific distribution can generalize the sample distribution of P_{NB} under three scenarios with different c_{end}. Hence, statistical tests using probability distribution fitting are not appropriate for these cases. However, the best (highest) fitness value among sampled chromosomes in each case with a c_{end} value of 7.5×10^7,

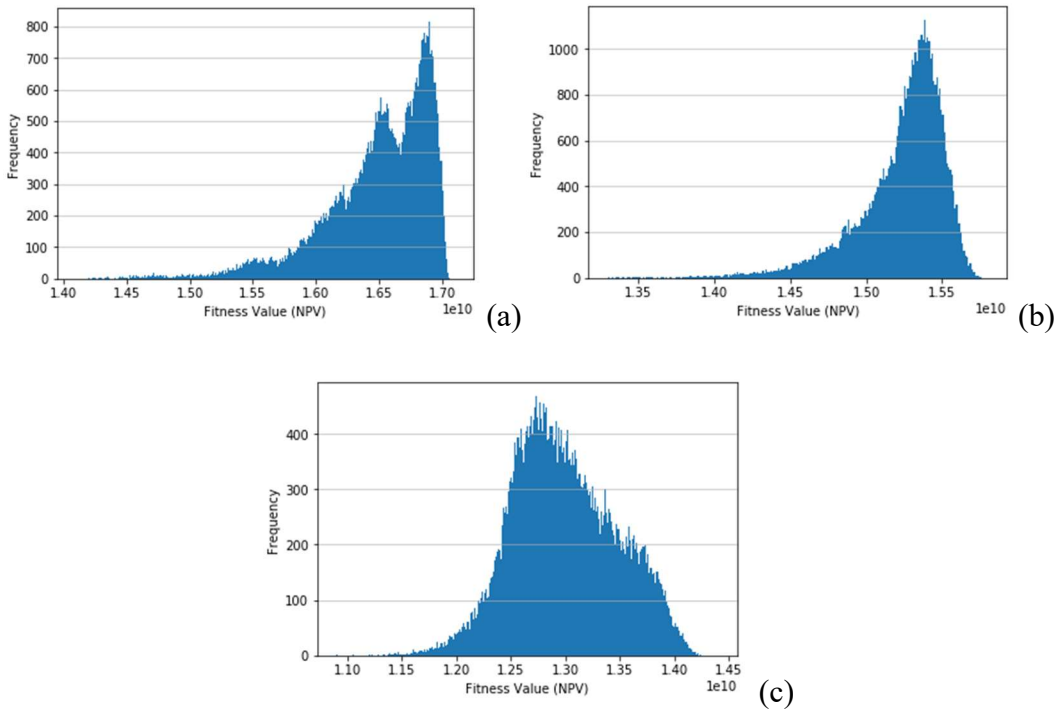


Figure 15 Distribution of fitness values of sampled chromosomes with different c_{end}

1.5×10^8, and 3.0×10^8 is 17.058×10^9, 15.759×10^9, and 14.235×10^9, respectively. Each one is lower than the fitness value of the optimized chromosome obtained through GA for the same c_{end} value.

In each case, since the 50,000 chromosomes are randomly selected from all 3,968,310 unique chromosomes, the probability that at least one chromosome from the 0.01% of chromosomes with the best fitness value is selected is: $1-\prod_{i=1}^{50000}\frac{3968310-396-i+}{3968310-i+1}=0.9934$. This means, given that the best fitness value among 50,000 randomly selected chromosomes is lower than that of the GA-optimized chromosome, it can be claimed that with over 99% confidence that the fitness value of the GA-optimized chromosome dominates 99.99% of all possible chromosomes. This demonstrates the effectiveness of the proposed GA framework and operators customized for this problem. Heuristic methods such as GA cannot

guarantee global optimality of the optimized solution, and the globally maximal fitness value is unknown (unless all possible chromosomes are exhaustively evaluated, which is expected to take more than a week), but a 99.99% dominance in fitness value is acceptable for optimizing an extension plan in this problem. Moreover, in practice the uncertainties in input parameters (e.g., u_v, u_{ln}, F, g, Q_{ij}) outweigh the uncertainty in optimality of solutions under given input parameters.

4.2.3 Effects of analysis period duration

In numerical cases shown above with analysis period length of T=30 years, all potential extension steps can be realized within 30 years, with binding constraint of available budget. If a shorter analysis period is used, some later potential extension steps could not be completed within T years, and the optimized extension plans could be affected.

For each of c_{end} values 7.5×10^7, 1.5×10^8, and 3.0×10^8, shorter analysis period durations of T=25 and T=20 are applied. Other parameters are unchanged. For each numerical case, the GA model is run multiple times until the best fitness value of the optimized extension plan (chromosome) is replicated in at least three runs. *max_stall* is adjusted to 50. The optimized results are shown in Table 7.

Given that c_{end} equals to 7.5×10^7 (1.5×10^8), if T is reduced from 30 to 25, all potential extension steps can still be completed and the first 10 (6) steps in optimized extension plans are identical in terms of sequence and station grouping, but several later steps are combined, with more stations to be completed in each step (as is underlined in Table 7). When T=20, or c_{end}=3.0×10^8 and T=25, the last potential steps in optimized extension plans cannot be completed, and only the steps that can

Table 7 Optimized extension plans under different values of c_{end} and T

c_{end}/$	T/year	Optimized plan (only showing steps that can be completed within T years)
7.5×10^7	30	[(13), (8), (14), (15), (7 6), (16), (5), (17), (4), (3), (18), (19), (2), (20), (1)]
	25	[(13), (8), (14), (15), (7 6), (16), (5), (17), (4), (3), (18 19), (2 1), (20)]
	20	[(13), (8), (14), (15), (7), (6), (16), (5), (17), (4), (3), (18), (19), (2), (20)]*
1.5×10^8	30	[(13), (8 7), (14 15), (6 5), (16 17), (4 3), (18 19), (2 1), (20)]
	25	[(13), (8 7), (14 15), (6 5), (16 17), (4 3), (18 19 20), (2 1)]
	20	[(13 14), (8 7), (15 16), (6 5), (17), (4), (18 19)]*
3.0×10^8	30	[(13 14), (8 7 6), (15 16 17), (5 4 3), (18 19 20), (2 1)]
	25	[(13 14), (8 7 6), (15 16), (5 4), (17 18), (3), (19), (20)]*
	20	[(13 14), (8 7 6), (15 16 17)]*

be completed within T years are shown in Table 7 with a star mark (*). Under each value of c_{end}, compared to optimized extension plans that can be fully completed in a longer T, those marked with (*) show more differences in the sequence and station grouping of extension steps. When the last extension steps cannot be completed, stations that can be completed in some extension steps (mostly later ones) may be completed with different groupings. New extension steps to be completed at later times tend to be smaller than previous later steps containing same stations under a longer T. Also note that given that T=20, as c_{end} increases, the number of extension steps as well as stations that can be completed within the analysis period decreases sharply.

It can be learned from above that, if all potential steps in the optimized extension plan can be completed within the analysis period, a shorter analysis period (T) could decrease the fraction of completion of optimized extension plans and affect station grouping in steps, with late steps likely to be smaller. On the other hand, a longer T yields more and smaller extension steps, with most steps unchanged. One possible explanation is given as follows. When a late extension step with multiple stations is decomposed into smaller steps without changing any other steps, inner stations in the

original step will be completed earlier, while the completion of outermost stations in this original step as well as stations in succeeding steps could be delayed. Given the same T, earlier completion of stations leads to longer time in operation during the analysis period and therefore increased PV of consumer surplus (P_{CS}) and collected fares (P_f) from related OD pairs, while delayed completion of stations affects P_{CS} and P_f reversely. Typically, if c_{end} is not too large, the completion advance of each inner station is greater than the completion delay of each outer and succeeding station. More extension steps also leads to higher PV of terminal cost. Moreover, in equations (20) and (21) a factor $(\frac{1+g}{1+r})^{\frac{t_k+t_{k+1}}{2}}$ is used for approximating PV of P_{CS} and P_f incurred in Period k based on its time midpoint $\frac{t_k+t_{k+1}}{2}$. In the numerical case, $g < r$ makes $\frac{1+g}{1+r} < 1$, and $(\frac{1+g}{1+r})^{\frac{t_k+t_{k+1}}{2}} > (\frac{1+g}{1+r})^{\frac{t_{k_{max}}+T}{2}}$ $(k < k_{max})$. Given that all potential steps can be completed within T, when the value of T moves farther from the final completion time ($t_{k_{max}}$), the midpoint of Period k_{max} also moves farther from that of any other period, leading to a larger difference between $(\frac{1+g}{1+r})^{\frac{t_k+t_{k+1}}{2}}$ and $(\frac{1+g}{1+r})^{\frac{t_{k_{max}}+T}{2}}$. As a result, a longer T more largely favor PV of P_{CS} and P_f from OD pairs related to stations that start operation in earlier periods than Period k_{max}. Positive effects of decomposing steps on NPV -- increased P_{CS} and P_f due to advanced completion of some stations in earlier periods -- are therefore more likely to outweigh its negative effects, which include decreased P_{CS} and P_f due to delayed completion of some other stations, as well as increased PV of terminal cost.

It should be noted that with a smaller T, the proposed GA method is more susceptible to prematurity. A shorter duration of analysis period means that more extension steps cannot be realized within T years, and more different chromosomes will have the same fitness value (NPV). Thus the optimization search becomes more likely to be trapped in local optima.

4.2.4 Analysis of sensitivity to selected parameters

From what is found in the one-directional case, five parameters that have major impacts on NPV or completion time are selected: c_{ln} (unit construction cost of rail line), F (yearly external fund supply), Q_{ij} (potential hourly ridership), u_v (value of in-vehicle time), and ρ (fraction of collected fare to be used for construction). The sensitivity of the optimized solution to these parameters is examined. For each of these parameters, its value is slightly changed from its base scenario value (within ±20%), with other parameters unchanged (except that c_{st}, the construction cost of a station, changes proportionally with c_{ln}). The model is run to obtain the optimized solution. In each of the modified scenarios, the optimized extension plan is fully completed.

The optimized extension plans with changes of various parameters are shown in Table 8. Table 9 lists all changes of parameters in modified scenarios, the corresponding optimized NPVs and their change rates from the base scenario, the corresponding $t_{k_{max}}$ values under optimized plans and their change rates from the base scenario, and the elasticity of NPV and $t_{k_{max}}$ in response to the change of each parameter.

Recall that the optimized extension plan in the base scenario is denoted as:

[(13), (8 7), (14 15), (6 5), (16 17), (4 3), (18 19), (2 1), (20)]

It can be learned from the results that, similarly to the case in one-directional extension problem, the optimized NPV is fairly sensitive to Q_{ij} and u_v. An increase of Q_{ij} or a decrease of u_v by a small percentage leads to an increase of NPV by a greater percentage. It should also be noted that the change of optimized extension plans is highly correlated with the change of the optimized NPV. The optimized extension plans denoted as [(13 14), (8 7), (15 16), (6 5), (17), (4 3), (18 19), (2 1), (20)] all correspond to decrease of NPV. While slight increases of NPV (caused by decrease of c_{ln}, increase of F, or increase of ρ) corresponds to slight changes in extension plans (with a few stations regrouped), significant increases of NPV (caused by increase of Q_{ij} or decrease of u_v) correspond to greater changes in extension plans (with more stations regrouped and more extension steps). Here is an explanation for this: both increase of Q_{ij} and decrease of u_v increase actual hourly ridership for all OD pairs served by the operating segment. By completing some neighboring stations in multiple steps instead of one, some stations can be completed earlier, and the increase of PV of total consumer surplus and collected fare incurred over T years could overcome the increase of PV of terminal cost.

Unlike the case in one-directional extension problem, the final completion time $t_{k_{max}}$ is most sensitive to Q_{ij} and ρ. $t_{k_{max}}$ is less sensitive to c_{ln} (and c_{st}) possibly because of the smaller percentage of station and link costs in overall construction costs in this numerical case. $t_{k_{max}}$ is much less sensitive to F, because in this large-scale numerical case, the future ridership as well as the reservation rate of revenue is much higher, making internal funding (fare collected from passengers) dominant to

the external funding. Note that while $t_{k_{max}}$ is similarly sensitive to Q_{ij} and ρ, the changes of optimized extension plans in response to Q_{ij} and ρ do not appear to be similar, which implies that changes of optimized extension plans are more correlated to those of optimized NPVs than to those of optimized $t_{k_{max}}$.

Table 8 Changes of parameters in modified scenarios and corresponding changes of optimized extension plans

Para-meter	Value	Change Rate	Optimized Extension Plan
c_{ln}	1.12×10^8	-20%	[(13), (8 7), (14 15), (6 5), (16 17), (4 3), (18), (19), (2), (20), (1)]
	1.26×10^8	-10%	[(13), (8 7), (14 15), (6 5), (16 17), (4 3), (18 19), (2), (20), (1)]
	1.54×10^8	10%	[(13 14), (8 7), (15 16), (6 5), (17), (4 3), (18 19), (2 1), (20)]
	1.68×10^8	20%	[(13 14), (8 7), (15 16), (6 5), (17), (4 3), (18 19), (2 1), (20)]
F	4.0×10^7	-20%	[(13 14), (8 7), (15 16), (6 5), (17), (4 3), (18 19), (2 1), (20)]
	4.5×10^7	-10%	[(13 14), (8 7), (15 16), (6 5), (17), (4 3), (18 19), (2 1), (20)]
	5.5×10^7	10%	[(13), (8 7), (14 15), (6 5), (16 17), (4 3), (18 19), (2), (20), (1)]
	6.0×10^7	20%	[(13), (8 7), (14 15), (6 5), (16 17), (4 3), (18 19), (2), (20), (1)]
Q	0.8 Q	-20%	[(13 14), (8 7), (15 16), (6 5), (17), (4 3), (18 19), (2 1), (20)]
	0.9 Q	-10%	[(13), (8 7), (14 15), (6 5), (16 17), (4 3), (18 19), (2 1), (20)]
	1.1 Q	10%	[(13 14), (8 7), (15 16), (6 5), (17), (4 3), (18 19), (2), (20), (1)]
	1.2 Q	20%	[(13 14), (8 7), (15 16), (6 5), (17), (4), (3), (18), (19), (2), (20), (1)]
u_v	14.4	-20%	[(13 14), (8 7), (15 16), (6 5), (17), (4), (3), (18), (19), (2), (20), (1)]
	16.2	-10%	[(13 14), (8 7), (15 16), (6 5), (17), (4 3), (18), (19), (2), (20), (1)]
	19.8	10%	[(13 14), (8 7), (15 16), (6 5), (17), (4 3), (18 19), (2 1), (20)]
	21.6	20%	[(13 14), (8 7), (15 16), (6 5), (17), (4 3), (18 19), (2 1), (20)]
ρ	20%	-20%	[(13 14), (8 7), (15 16), (6 5), (17), (4 3), (18 19), (2 1), (20)]
	22.5%	-10%	[(13 14), (8 7), (15 16), (6 5), (17), (4 3), (18 19), (2 1), (20)]
	27.5%	10%	[(13), (8 7), (14 15), (6 5), (16 17), (4 3), (18 19), (2 1), (20)]
	30%	20%	[(13), (8 7), (14 15), (6 5), (16 17), (4 3), (18 19), (2 1), (20)]

Table 9 Changes of parameters in modified scenarios and corresponding changes of NPV and final completion time

Para-meter	Value	Change Rate	NPV ($\times\$10^9$)	Change Rate	Elasticity of NPV	$t_{k_{max}}$ Value	Change Rate	Elasticity of $t_{k_{max}}$
c_{ln}	1.12×10^8	-20%	16.758	6.19%	-0.297	19.344	-8.50%	0.492
	1.26×10^8	-10%	16.265	3.07%		20.244	-4.24%	
	1.54×10^8	10%	15.326	-2.88%		22.325	5.61%	
	1.68×10^8	20%	14.896	-5.61%		23.467	11.01%	

F	4.0×10^7	-20%	15.420	-2.29%	0.117	21.945	3.81%	-0.110
	4.5×10^7	-10%	15.598	-1.16%		21.538	1.88%	
	5.5×10^7	10%	15.966	1.17%		21.074	-0.31%	
	6.0×10^7	20%	16.137	2.26%		20.697	-2.10%	
Q	0.8 Q	-20%	11.009	-30.24%	1.654	24.356	15.21%	-0.626
	0.9 Q	-10%	13.363	-15.32%		22.634	7.07%	
	1.1 Q	10%	18.582	17.75%		19.987	-5.45%	
	1.2 Q	20%	21.266	34.76%		19.378	-8.34%	
u_v	14.4	-20%	22.553	42.91%	-1.986	19.861	-6.05%	0.425
	16.2	-10%	19.244	21.94%		20.523	-2.92%	
	19.8	10%	12.976	-17.78%		22.318	5.57%	
	21.6	20%	10.525	-33.31%		23.527	11.29%	
ρ	20%	-20%	15.297	-3.07%	0.141	23.873	12.93%	-0.569
	22.5%	-10%	15.554	-1.44%		22.410	6.01%	
	27.5%	10%	15.999	1.38%		20.006	-5.36%	
	30%	20%	16.184	2.55%		19.001	-10.12%	

Chapter 5: Conclusions

A novel optimization model that features a continuous time formulation is developed for solving two versions (one-directional and two-directional extension) of the phased development problem of a single rail transit line. Demand elasticity is considered, and the closed form of maximal allowable headway is derived. The objective is to maximize system NPV over the analysis period, while line continuity and the available budget at the start of each period serve as constraints. The economies of completing multiple links together and the option of not completing some links during the analysis period are captured in the model. The model is coded in Python 3.7.3, and two heuristic methods – Differential Evolution (DE) and Genetic Algorithm (GA) -- are used for solving the one-directional and the two-directional version of the problem, respectively. Customized operators of GA are developed for solution search in the two-directional extension problem. For the base scenario of the one-directional extension problem, the optimized extension steps (development phases) indicate that first 4 of 5 planned links (stations) should be completed in the analysis period, with the available budget constraint being binding at each extension. Under the assumption of a binding available budget constraint, an optimized extension plan is obtained with the customized GA for the base scenario of the two-directional extension problem. With other parameters unchanged, when the analysis period is lengthened, the completion of outermost planned links within the analysis period becomes justified, and when the construction cost of terminal facilities is increased, the optimized extension plan increasingly favors completion of multiple links in one step (one joint completion). Sensitivity of the maximized NPV and the

optimized completion time of Station 8 to seven selected parameters are examined in the one-direction extension problem. Sensitivity of the maximized NPV and the optimized completion sequence and grouping of planned stations to five selected parameters are examined in the two-direction extension problem. Sensitivity analysis reveals that decision makers should be especially careful in determining future potential demands, value of users' in-vehicle time, and the unit construction cost of the rail transit line before making extension plans.

The model presented here may be improved in several ways in the future:

1) Due to the difficulty in formulating the closed form of the optimal train headway that maximizes total net social benefit in each period, the headway used in each period is assumed to be the maximal allowable headway. Some numerical method may be developed to optimize headways in each period. Fare, train capacity, and train speed may also be optimizable in more complex versions of this model.
2) Some additional demand features, such as faster growth due to new station completion, effect of access time, and nonlinear demand functions, may be developed.
3) Land use development induced by rail line extensions may be considered.
4) The computations of total PV of consumer surplus and supplier's revenue, operation cost and maintenance cost include approximations, which may be replaced with a more precise integration method.
5) Integer fleet sizes may be imposed.
6) Cyclical operations (e.g. peak and off-peak) may be considered.

7) Uncertainties regarding demand, budget and construction costs may be considered.
8) This model could be further extended beyond single rail lines to solve phased development problems for rail transit networks and connecting bus routes.

References

Barrena, E., D. Canca, L. C. Coelho, and G. Laporte. Single-line Rail Rapid Transit Timetabling under Dynamic Passenger Demand. *Transportation Research Part B: Methodological*, 2014. 70: 134-150.

Canca, D., A. De-Los-Santos, G. Laporte, and J. A. Mesa. An Adaptive Neighborhood Search Metaheuristic for the Integrated Railway Rapid Transit Network Design and Line Planning Problem. *Computers & Operations Research*, 2017. 78: 1-14.

Cheng, W. C., and P. Schonfeld. A Method for Optimizing the Phased Development of Rail Transit Lines. *Urban Rail Transit*, 2015. 1(4): 227-237.

Chien, S., and P. Schonfeld. Joint Optimization of a Rail Transit Line and Its Feeder Bus System. *Journal of advanced transportation*, 1998. 32(3): 253-284.

Gallo, M., B. Montella, and L. D'Acierno. The Transit Network Design Problem with Elastic Demand and Internalisation of External Costs: An Application to Rail Frequency Optimisation. *Transportation Research Part C: Emerging Technologies,* 2011. 19(6): 1276-1305.

Guan, J. F., H. Yang, and S. C. Wirasinghe. Simultaneous Optimization of Transit Line Configuration and Passenger Line Assignment. *Transportation Research Part B: Methodological*, 2006. 40(10): 885-902.

Hassannayebi, E., A. Sajedinejad, and S. Mardani. Urban Rail Transit Planning Using a Two-stage Simulation-based Optimization Approach. *Simulation Modelling Practice and Theory*, 2014. 49: 151-166.

Lai, X., and P. Schonfeld. Concurrent Optimization of Rail Transit Alignments and Station Locations. *Urban Rail Transit*, 2016. 2(1): 1-15.

Li, Z. C., W. H. Lam, S. C. Wong, and A. Sumalee. Design of a Rail Transit Line for Profit Maximization in a Linear Transportation Corridor. *Transportation Research Part E: Logistics and Transportation Review*, 2012. 48(1): 50-70.

Niu, H., and X. Zhou. Optimizing Urban Rail Timetable Under Time-dependent Demand and Oversaturated Conditions. *Transportation Research Part C: Emerging Technologies*, 2013. 36: 212-230.

Peng, Y. T., Z. C. Li, and P. Schonfeld. Development of Rail Transit Network Over Multiple Time Periods. *Transportation Research Part A: Policy and Practice,* 2019. 121: 235-250.

Saidi, S., S. C. Wirasinghe, and L. Kattan. Long-term Planning for Ring-radial Urban Rail Transit Networks. *Transportation Research Part B: Methodological*, 2016. 86: 128-146.

Samanta, S., and M. K. Jha. Modeling a Rail Transit Alignment Considering Different Objectives. *Transportation Research Part A: Policy and Practice*, 2011. 45(1): 31-45.

Shang, P., R. Li, Z. Liu, L. Yang, and Y. Wang. Equity-oriented Skip-stopping Schedule Optimization in an Oversaturated Urban Rail Transit Network. *Transportation Research Part C: Emerging Technologies*, 2018. 89: 321-343.

Storn, R., and K. Price. Differential Evolution–a Simple and Efficient Heuristic for Global Optimization Over Continuous Spaces. *Journal of global optimization*, 1997. 11(4): 341-359.

Sun, Y., P. Schonfeld, and Q. Guo. Optimal Extension of Rail Transit Lines. *International Journal of Sustainable Transportation*, 2018. 12(10): 753-769.

Whitley, L. D. (1989, June). The GENITOR algorithm and selection pressure: why rank-based allocation of reproductive trials is best. In Icga (Vol. 89, pp. 116-123).

Wirasinghe, S. C., V. F. Hurdle, and G. F. Newell. Optimal Parameters for a Coordinated Rail and Bus Transit System. *Transportation Science*, 1977. 11(4): 359-374.

Wong, R. C., T. W. Yuen, K. W. Fung, and J. M. Leung. Optimizing Timetable Synchronization for Rail Mass Transit. *Transportation Science*, 2008. 42(1): 57-69.