Service expansion for chained business facilities under congestion and market competition

Yun Hui Lin[a], Qingyun Tian[b,*], Shaojun Liu[b]

[a]*Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore*
[b]*School of Civil and Environmental Engineering, Nanyang Technological University, Singapore*

Abstract

We study a service expansion problem for chained business facilities under endogenic facility congestion and exogenous market competition. More specifically, we consider a company that operates a chain of facilities and plans to expand service capacities with the objective of maximizing its profit, accounting for revenue and expansion costs. To estimate revenue, the company needs to anticipate customer behaviors. Due to the co-existence of competition and congestion, customer behaviors are explained as a two-stage process. In the first stage, customers make "channel" choices, i.e., they decide whether to seek services from the company. Such a choice reflects the market competition and is predicted by a discrete choice model. Subsequently, customers who select the company will choose one facility to patronize. Owing to congestion, the facility choice will induce "user equilibrium", which in return affects the outcome of market competition. To facilitate the company's decision-making in this complex business environment, we develop a generic modeling framework. Unfortunately, the proposed model is nonconvex. To solve it, we first design an approximate mixed-integer linear programming approach subject to adjustable approximation errors. We then propose a surrogate optimization framework for large-scale instances, which explores the hidden bilevel structure of the model and leverages a "learning-to-optimize" problem and a customer behavior estimation subroutine. Using extensive computational experiments, we demonstrate the effectiveness of the proposed approaches. Finally, we conduct sensitivity analysis and draw practical implications.

Keywords: Chained business operation, Service expansion, User equilibrium, Market competition, Learning-to-optimize

1. Introduction

In operations research and management science, the capacity/service expansion literature studies the size, timing, and location of buying or installing additional capacities for facilities or infrastructures. For an operator who runs a chain of business facilities, capacity expansion is an important strategic decision. When customer demand increases or competitors appear, the existing facilities may not adapt to the new business environment, especially in those scenarios where facilities are subject to congestion. As a countermeasure, the operator can proactively improve its service quality by expanding facility capacities, which could allow it to increase the market share or to stand in an advantageous position when facing exogenous market competition.

The capacity/service expansion problem has been widely studied since Manne (1961) first investigated the problems considering the discount factor and the economies of scale. The early literature focused on deterministic models (Luss, 1982; Herman and Ganz, 1994). Then, there has

*Corresponding author.
Email addresses: `liny@ihpc.a-star.edu.sg` (Yun Hui Lin), `qytian@ntu.edu.sg` (Qingyun Tian), `lius0075@e.ntu.edu.sg` (Shaojun Liu)

been an increasing focus on models with stochastic demand (e.g., Zhao et al. (2018)). In the literature, we have also witnessed extensive research on practical and application-oriented problems, such as manufacturing (Zhang et al., 2004), semiconductor production (Wang and Lin, 2002; Wang and Hou, 2003), electricity generation (Wogrin et al., 2011; Parpas and Webster, 2014; Min et al., 2018), and urban traffic management (Mathew and Sharma, 2009). Detailed surveys can be found in Martínez-Costa et al. (2014); Azeez et al. (2018). We point out that the aforementioned studies were based on the *spatial monopoly* assumption. That is, the operator is the only service provider, and there is no external competition in the market. However, competition exists in most practical business contexts since there may be other options/other providers that offer a similar service to customers. When the operator fails to provide decent service quality, customers may go for other options, leading to the loss of market share and revenue. Therefore, it is certainly worth investigating the effect of market competition on the capacity expansion of chained business facilities.

Motivated by the observation, we investigate a capacity expansion problem for an operator who owns a chain of service facilities. In the market, there are competitors providing equivalent services. We refer to different service providers as different "channels". The operator aims to maximize its profit by expanding service capacities at facilities, which should strike a balance between revenue and expansion costs. In general, increasing service capacities will reduce the congestion effect at facilities. Customers can thus observe higher service quality and will be more willing to seek services from the chained facilities. Accordingly, the operator will observe revenue growth. That is, the revenue of the operator depends on the expansion decision which affects both the congestion level and the likelihood of the chained facilities being selected by a customer. To estimate the revenue, the operator thus needs to anticipate (i) how customers choose between the operator and competitors to seek services from, namely, the *channel choice*, and (ii) how customer choices are affected by congestion at the facility level.

To predict the customer choices and the revenue, probabilistic choice rules or models are frequently used (Drezner and Drezner, 2002; Fernández et al., 2021; Lin and Tian, 2021b; Suhara et al., 2021). Two classical models are the gravity model and the multinomial logit model (MNL). The gravity model belongs to the strict utility theory, which stipulates that the probability of a customer selecting a facility increases with the attractiveness of the facility and decreases with distance (Huff, 1964). The method has been used for several service facility location and design problems (Drezner et al., 2018; Küçükaydın et al., 2011; Lin and Tian, 2021a). By contrast, the MNL is based on the random utility theory, where the utility for an option consists of a deterministic component and a random term. Under the utility-maximizing principle, the probability of choosing an option is then equal to the probability that its utility exceeds that of all other options. Using the MNL model, various facility and channel choice-related studies have been conducted (e.g., see Lin et al. (2020); Zhang et al. (2012)). In this paper, we follow the random utility theory and use it to predict customer channel choices. Since only two channels exist (i.e., the operator channel and the competitor channel. The latter will also be called the "outside option"), the MNL reduces to the binomial logit model (BNL), whose applications can be found in recent studies (Gunawan et al., 2020; Shriver and Bollinger, 2022).

A common limitation of the above works is that the utility of a facility/channel is assumed to be exogenous and does not depend on endogenic factors. However, for a service system, facilities are usually subject to congestion. That is, when the number of customers visiting a facility is high, the arrival rate of customers may approach or even exceed the capacity of the system. Hence, congestion occurs, and customers are forced to wait for a longer time before they get served. This indicates that the channel utility could depend on the expansion decision of the operator and the facility patronizing behavior of the customers. Indeed, facility congestion in terms of waiting time has a major effect on customer choices (Etebari, 2019) and thus should be incorporated into the decision-making process. To date, the congestion effect has been extensively investigated. One stream is to incorporate congestion into the framework of *centralized* systems, wherein users/customers would be assigned by the operator to minimize the system costs (Berman and

Drezner, 2006; Elhedhli et al., 2018; Góez and Anjos, 2017). Another stream is to consider the *decentralized* user choice environment, in which each user selects facilities to minimize his/her own routing time and/or costs, leading to *Wardrop's user equilibrium* (UE) (Fisk, 1980). Examples of the UE in facility and infrastructure planning can be found in road capacity expansion (Chiou, 2005; Wang and Lo, 2010), park-and-ride facility (Ye et al., 2021), charging station deployment (Xiong et al., 2017; Ferro et al., 2021), service network design (Hamzeei and Luedtke, 2018), and probabilistic set covering location (Aboolian et al., 2022), to name a few. However, these studies implicitly impose the spatial monopoly assumption. Consequently, market competition, which could arise in practice, was overlooked.

Indeed, there are several studies that incorporate market competition into congested facility design and planning. Marianov et al. (2008) considered a company planning to build multi-server facilities in a region to compete with a competitor for customers. A market share maximization model was proposed, in which customers take the facility congestion into consideration and select the facilities to patronize following the MNL. A heuristic based on Tabu Search (TS) and a Greedy Randomized Adaptive Search Procedure (GRASP) were proposed to solve the model. In Schön and Saini (2018), a market-oriented service network design model was developed to determine the location and service capacities of facilities, with the goal of maximizing overall profit. Customer choices were predicted by an attraction-based choice model with proportional substitution patterns, similar to MNL. Although the congestion effect was considered in the model, it was only used in a service-level constraint (which requires that the expected waiting time of customers should not exceed a value promised by the company) and did not directly affect customer choices. More recently, Dan and Marcotte (2019) investigated a competitive facility location and design problem. Users/customers are assumed to patronize the facility that maximizes their individual utility, represented as the sum of travel time, queueing delay, and a random term. Customer choices are also predicted by the MNL. In the context of healthcare operations, Zhang and Atkins (2019) studied the network design of "walk-in" congested medical facilities to maximize the total demand volume. Several models were considered by Zhang and Atkins (2019). In particular, the two "user-choice" models are related to our work: one is the probabilistic-choice model, which employs the MNL to predict client choices, whereas the other is the deterministic-choice model, which utilizes the UE to anticipate the flow pattern.

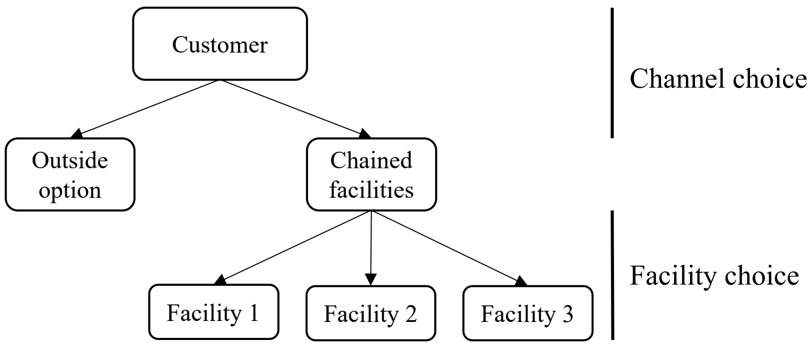


Figure 1: Explanation of customer behaviors.

Essentially, the existing literature represents customer choice as an "single-stage" process, i.e., perceiving congestion, customers either make decisions based on the MNL or the UE. In the former case, it is assumed that all open facilities are considered by the customers since each open facility will have no-zero demand assigned to it. On the other hand, in the UE-based model, customer choices are deterministic, presuming that customers at the same zone are homogenous, and there is no randomness in the choice determination. In this paper, we develop a new modeling perspective, i.e., due to the co-existence of competition and congestion, we visualize the behavior of a customer as a "two-stage" choice as shown in Figure 1. In the first stage, customers will make the *channel choice* to decide from which channel (the operator or the outside option) to seek services. In this stage, we apply the BNL to predict the channel choice probability, which reflects the random preference of customers towards different service channels. The demand split is cast based on the framework of the random utility theory, which allocates the demand to the two channels based on their utility. The utility is, however, affected by the second stage *facility choice*. Specifically, for customers choosing the chained facilities, they will subsequently select one facility to patronize. At the facility choice level, each customer will select a facility that minimizes his/her total time. Owing to congestion, such a choice will induce UE, by which customers at the same zone will experience the same and minimal equilibrium time, and no one can further reduce the time by unilaterally changing his/her facility choice. Notably, the two stages affect each other: the first stage determines how many customers will visit the chained facilities according to the channel utility; whereas based on the current demand split, the equilibrium at the second stage yields the equilibrium time and the channel utility, which in return affects the channel choice probability. Therefore, for the operator, when determining the expansion decision, the inter-dependent two-stage choices of customers must be explicitly and simultaneously considered.

Our contributions. Overall, the contributions of this paper are three-fold.

- Model. We jointly consider the two practical effects, i.e., exogenous market competition and endogenic facility congestion, in a capacity expansion problem. We propose a generic model to facilitate the operator of chained business facilities when making the profit-maximizing expansion decision, upon anticipating the two-stage behaviors of customers under channel competition and user equilibrium. The hierarchical customer behavior representation is a distinct feature of our paper.

- Solution approaches. The proposed model is, unfortunately, a challenging nonconvex program. To solve it, we first design an approximate reformation approach to recast the model into a mixed-integer linear program (MILP) subject to adjustable approximation errors. Such an approach is straightforward to implement and works well for small-scale instances of the problem. For large-scale instances that could arise from real-world applications, we further develop a surrogate optimization framework, which explores the hidden bilevel structure of the model and decomposes the model into a "learning-to-optimize" problem and a lower-level customer behavior estimation subroutine. Our extensive computational experiments clearly demonstrate the efficiency and effectiveness of the proposed approaches.

- Practical implications. Through comprehensive sensitivity analysis, we draw several implications. In particular, for the operator, service expansion is preferred under the scenarios where (i) congestion is moderate, (ii) market competition is moderate, and/or (iii) the initial service capacity is low. For other scenarios, the operator should be conservative in its expansion plan. Moreover, customers tend to experience lower utility when the competition becomes weaker, meaning that customers ultimately benefit from the increasing competition between service providers.

Organization. The remainder of the paper is structured as follows. We introduce the problem as well as the model formulation in Section 2. Section 3 presents an approximate mixed-integer linear programming approach. Section 4 develops an efficient algorithm based on an intelligent surrogate optimization framework. We then conduct extensive computational experiments in Section 5 to demonstrate the effectiveness and the efficiency of the proposed solution approaches. Sensitivity analysis is presented in Section 6 to investigate the impact of parameters on the solutions and draw insightful observations. Finally, Section 7 concludes the paper and points out potential future research.

2. Problem description

Table 1: Notations.

Sets		
I	:	set of demand zones. The number of customer zones is $\lvert I \rvert$.
J	:	set of facilities. The number of facilities is $\lvert J \rvert$.
Parameters		
f_j	:	cost of expanding one unit capacity of facility j, $\forall j \in J$.
q_j	:	initial capacity of facility j, $\forall j \in J$.
d_i	:	demand at zone i, $\forall i \in I$.
r	:	unit revenue of serving a customer. Set $r = 1$ for the ease of exposition.
T_i	:	minimum routing time of a customer at zone i to the outside option, $\forall i \in I$.
L_j	:	lower bound of the expanded capacity ($L_j > q_j$), $\forall j \in J$.
U_j	:	upper bound of the expanded capacity ($U_j > 0$), $\forall j \in J$.
α_{ij}	:	fixed accessing time of a customer at zone i to facility j, $\forall i \in I, j \in J$.
Variables		
x_j	:	expanded capacity of facility j, $\forall j \in J$.
y_{ij}	:	number of customers at zone i patronizing facility j, $\forall i \in I, j \in J$.
v_j	:	number of customers patronizing facility j, $\forall j \in J$.
z_i	:	number of customers at zone i seeking services from the outside option, $\forall i \in I$.
t_{ij}	:	routing time of a customer at zone i to facility j, $\forall i \in I, j \in J$.
τ_i	:	minimum routing time of a customer at zone i to a facility, $\forall i \in I$.

This section presents the formal problem description and model formulation. We summarize the main notations in Table 1. Additional ones may be introduced when necessary.

Consider a company that operates a chain of facilities (denoted by set J) to serve customers located at different zones (denoted by set I). Facilities are subject to congestion, i.e., if the number of customers visiting facility j (denoted by v_j) increases, then the service time at the facility will become higher. For a customer at zone i seeking the service from facility j, he/she will experience a "routing time" t_{ij} that consists of (i) a constant accessing time α_{ij}, representing the time spent in traveling from zone i to facility j, and (ii) a variable service time at facility j, depending on v_j. We will present the discussion on congestion in Section 2.1.

Besides the company, there exist competitors who provide equivalent services in the market. In this paper, we aggregate the competitors into a single *outside option*, which represents an additional channel, with which the company competes to attract customer demand. Overall, when the routing time to a facility is long, customers perceive a low level of service and may prefer the outside option. Therefore, the company plans to expand the service capacities of its facilities to reduce congestion and, ultimately, attract more customers. Let x_j be the expanded capacity of facility j and f_j be the cost of expanding one unit capacity. Then, the objective of the company is to maximize the profit, which should strike a balance between revenue and the expansion cost. Specifically, we seek to maximize $r \sum_{j \in J} v_j - \sum_{j \in J} f_j x_j$, where the first term is the revenue, which considers the total number of customers patronizing the chained facilities ($\sum_{j \in J} v_j$) and the unit revenue of serving one customer (r), and the second term represents the total expansion cost.

To estimate the revenue, the company needs to anticipate customer behaviors, i.e., how customers choose between the company and the competitor, and how customer choices are affected by congestion. As shown in Figure 1, customers first make the channel choice, where they decide from which channel to seek services. In this paper, we use the BNL to predict the channel choice probability. The mathematical formulation of the BNL will be given in Section 2.3. Note that the choice probability is affected by the congestion at facilities since more intense congestion will lead to lower utility of customers. Therefore, the probability depends on the patronizing behaviors of those customers that seek the service from the company, i.e, the facility choice. At the facility choice level, customer behavior follows the UE principle. Detailed discussions will be presented in Section 2.2.

2.1. Congestion function

We first discuss the facility congestion effect. When the number of customers visiting facility j increases, the service time C_j will increase. Meanwhile, C_j also depends on its initial service capacity q_j and the expanded capacity x_j. We can thus write the service time at facility j as

$$C_j = c_j(v_j, x_j; q_j), \; \forall j \in J, \tag{1}$$

which is a function of variables v_j and x_j, parameterized by the initial service capacity q_j. Naturally, the function should satisfy the following properties: (i) Given an initial service capacity q_j, when v_j is fixed, $c_j(v_j, x_j; q_j)$ should be a decreasing function of x_j, which states that expanding the service capacity will reduce the service time; whereas when x_j is fixed, $c_j(v_j, x_j; q_j)$ should be an increasing and convex function of v_j, which reflects the congestion effect. (ii) When both v_j and x_j are fixed, increasing q_j will reduce the value of $c_j(v_j, x_j; q_j)$, which means if the initial service capacity is higher, then the congestion level will be lower.

Noting the properties, possible choices of service time function includes (but are not limited to)

• Queuing models, such as the M/M/1 queue, taking the form of

$$c_j(v_j, x_j; q_j) = \frac{\kappa}{q_j + x_j - v_j}, \; \forall j \in J, \tag{2}$$

where κ is a positive parameter used to convert the congestion value into measurable time. However, if such a model is assumed, then we need to impose that $v_j < q_j + x_j$.

• BPR-type functions, which is commonly used to model the congestion effect in various research fields (e.g., An et al. (2015); Liu and Meng (2014)). Their generic form can be written as

$$c_j(v_j, x_j; q_j) = s + \gamma \left(\frac{v_j}{q_j + x_j} \right)^{\beta}, \; \forall j \in J, \tag{3}$$

where s is a constant service time at facilities, and $\gamma > 0$ and $\beta \geq 1$ are congestion parameters.

In practice, the choice and the exact form of the service time function should be based on the real-world problem under investigation. In this paper, we adopt BPR-type functions; however, our methodology presented hereafter is not limited to this particular function. Note that, in (3), it is possible to define (s, γ, β) for each facility (i.e., use subscript j to specify their values for facility j). Without loss of generality, we assume they are homogeneous for all facilities.

2.2. User equilibrium in facility choice

With the congestion function defined above, the routing time of a customer at zone i to facility j is given by

$$t_{ij} = \alpha_{ij} + c_j(v_j, x_j; q_j), \; \forall i \in I, j \in J. \tag{4}$$

Define a nonnegative variable y_{ij} to represent the number of customers at zone i patronizing facility j, $\forall i \in I, j \in J$. Then, v_j and y_{ij} have the following relation

$$v_j = \sum_{i \in I} y_{ij}, \; \forall j \in J. \tag{5}$$

To predict the facility choices, we assume that customers are selfish and non-cooperative users. Each customer will select the facility that minimizes his/her routing time. In this setting, we can use the following complimentary constraint to describe the equilibrium condition under facility congestion:

$$(t_{ij} - \tau_i) \cdot y_{ij} = 0, \ \forall i \in I, j \in J, \tag{6a}$$

$$t_{ij} \geq \tau_i, \qquad \forall i \in I, j \in J. \tag{6b}$$

Here, τ_i is the minimum (or equilibrium) routing time of customers at zone i. (6) ensures that all facilities visited by customers at zone i will generate the same and minimal routing time. If the routing time of one facility t_{ij} is larger than τ_i, then no one will visit it, i.e., $y_{ij} = 0$ if $t_{ij} > \tau_i$ according to (6a). In other words, each customer tries to minimize his/her own routing time, and at equilibrium, customers from the same zone will experience the same and minimal equilibrium routing time τ_i, and no one can further reduce his/her routing time by unilaterally changing the facility choice, which induces a stable and equilibrium choice pattern (Aboolian et al., 2022)

2.3. *Channel choice and market competition*

Given the equilibrium behavior of customers in facility choice, we estimate the minimum routing time of a customer at zone i as τ_i. Due to the market competition, if time τ_i is long, then customers may prefer the outside option. As a result, the number of customers at zone i that are attracted to the chained facilities is essentially a function of τ_i. We refer to such a demand function as the *channel choice* function because it reflects how customers choose between providers (i.e., channels).

In practice, channel choices are typically interpreted probabilistically and frequently observed to be in a non-deterministic manner (Avila et al., 2020; Gunawan et al., 2020; Shriver and Bollinger, 2022). That is, not all customers will be attracted to a single channel. Instead, the proportion of customers using a channel should be estimated as a function of the channel disutility/utility. Discrete choice models based on random utility theory thus arise as powerful tools to predict the channel choice. In this paper, the channel disutility is related to the routing time. Specifically, we assume that the routing time of a customer at zone i to any competitor is constant. In this case, for the fixed set of competitors K, we define the the lowest routing time of a customer at zone i to the outside option as $T_i = \min_{k \in K} T_{ik}$. We then predict customers' channel choices by the binomial logit model (BNL), i.e.,

$$\sum_{j \in J} y_{ij} = \frac{d_i}{1 + e^{\theta(\tau_i - T_i - \Psi_i)}}, \ \forall i \in I, \tag{7}$$

where the left-hand-side computes the number of customers at zone i visiting the chained facilities; the right-hand-side is the BNL, and there are two parameters associated with it (in practice, both parameters can be empirically estimated from the data of the original system): (i) θ, a non-negative scaling factor. Essentially, it reflects the sensitivity of customer choices to the disutility. In general, a small value of θ indicates that customers are not sensitive to the difference between disutilities. In one extreme case where $\theta = 0$, customers are assumed to completely ignore the disutilities and will randomly determine whether to use the chained facilities or to seek the service from the outside option (i.e., 50% of the customers will be attracted to the chained facilities regardless of the disutilities); in the other extreme case where $\theta = \infty$, all customers will select the outside option if $t_i > T_i + \Psi_i$, and vise versa. (ii) Ψ_i, a constant, which captures the effects of all factors other than the time difference on the channel choice. For instance, a positive value of Ψ_i can be interpreted as a "facility preference" factor. That is, $\Psi_i > 0$ implies that the share of the operator at zone i is greater than that of the outside option even when $\tau_i = T_i$.

Note that τ_i depends largely on x, and thus the channel choice of customers is also affected by x. In general, when the company invests more in expanding the service capacities, τ_i decreases, and the company has more chances to win in the market competition and is expected to see a growth in its revenue.

2.4. Model formulation

In the context of the above setup, we formulate the service expansion problem for chained business facilities under congestion and market competition as

$$
\textbf{[SEP]} \quad
\begin{aligned}
\max \quad & r\sum_{j\in J} v_j - \sum_{j\in J} f_j x_j && && \text{(8a)}\\
\text{s.t. } & L_j \le x_j \le U_j, && \forall j \in J, && \text{(8b)}\\
& v_j = \sum_{i\in I} y_{ij}, && \forall j \in J, && \text{(8c)}\\
& t_{ij} = \alpha_{ij} + c_j(v_j, x_j; q_j), && \forall i \in I, j \in J, && \text{(8d)}\\
& (t_{ij} - \tau_i)\cdot y_{ij} = 0, && \forall i \in I, j \in J, && \text{(8e)}\\
& t_{ij} \ge \tau_i, && \forall i \in I, j \in J, && \text{(8f)}\\
& \sum_{j\in J} y_{ij} = \frac{d_i}{1+e^{\theta(\tau_i - T_i - \Psi_i)}}, && \forall i \in I, && \text{(8g)}\\
& y_{ij} \ge 0 && \forall i \in I, j \in J. && \text{(8h)}
\end{aligned}
$$

In the objective function (8a), we maximize the profit of the company, accounting for the estimated revenue and the expansion cost. Without loss of generality, we assume throughout the paper that one unit of demand yields one unit of revenue, i.e., $r = 1$. Constraint (8b) defines the domain of the expansion variable, where U_j and L_j are the upper and lower bounds, respectively. Note that if L_j is negative (i.e., takes a value between q_j and 0), then we permit the possibility of capacity reduction. Since this paper positions itself in the literature of capacity expansion, L_j is set to 0. Finally, Constraints (8c)-(8h) formulate customers' facility choices and channel choices under congestion as explained in Sections 2.2 and 2.3.

Remark. Since we consider "static competition" where the competitor is assumed to be irresponsive (i.e., T_i is a constant), we can interpret the competition as "demand elasticity" that characterizes the overall demand allocated by customers to facilities as a variable, depending on the UE routing time. It is then possible to further combine the routing time function (8d) and the BNL function (8g) into a single (but complicated) decision-dependent demand function. Accordingly, the UE conditions (8e) and (8f) can be formulated. This modeling philosophy can be seen in Aboolian et al. (2012, 2022). In this paper, since the congestion effect is represented by a general function, the resultant decision-dependent demand function could be challenging to derive and analyze. Moreover, our current model has a side benefit, which allows us to recast it into an equivalent bilevel formulation. Detailed discussions will be presented in Section 4.

3. Approximate mixed-integer linear programming approach

[SEP] is a nonlinear optimization problem, whose computational difficulty is largely amplified by its non-convexity that arises from the UE complementary condition (8e), the time function (8d), and the logit choice function (8g). This section presents an approximate reformation approach to recast [SEP] into a mixed-integer linear program (MILP) subject to adjustable approximation errors. The reformulated MILP can subsequently be handled by high-performance commercial solvers to derive a (near-)optimal solution. In this section, we discuss the formulation approaches in detail.

3.1. Reformulating UE complementary conditions

We start with the complementary constraints (8e)-(8f). Introduce a binary variable ξ_{ij} such that $\xi_{ij} = 1$ if $t_{ij} > \tau_i$; and $\xi_{ij} = 0$ if $t_{ij} = \tau_i$. Then we can replace Constraints (8e)-(8f) by the following mixed-integer linear constraints

$$\zeta \cdot \xi_{ij} \leq t_{ij} - \tau_i, \quad \forall i \in I, j \in J, \tag{9a}$$

$$t_{ij} - \tau_i \leq M \cdot \xi_{ij}, \ \forall i \in I, j \in J, \tag{9b}$$

$$y_{ij} \leq d_i \cdot (1 - \xi_{ij}), \forall i \in I, j \in J, \tag{9c}$$

$$\xi_{ij} \in \{0, 1\}, \qquad \forall i \in I, j \in J, \tag{9d}$$

where M is a large positive number, and ζ is a small positive number (which is chosen as $\zeta = 10^{-4}$ in this paper). To verify the equivalence, note that if $t_{ij} = \tau_i$, then $\xi_{ij} = 0$ by (9a). In this case, (9c) becomes $y_{ij} \leq d_i$, which trivially holds. If $t_{ij} > \tau_i$, then (9b) imposes that $\xi_{ij} = 1$. As a result, $y_{ij} = 0$ since the right hand side of (9c) is 0. That is, (9) effectively imposes that if the routing time to facility j is higher than the minimum routing time τ_i, then no one will visit facility j from zone i. (9) is thus equivalent to (8e)-(8f).

3.2. *Approximating the nonlinear terms with adjustable error bound*

We then linearize the other nonlinear terms in [SEP], i.e., the routing time function (8d) and the logit choice function (8g). For (8d), under the BPR-type congestion, we define an additional variable μ_j such that

$$\mu_j = \left(\frac{v_j}{q_j + x_j} \right)^{\beta}, \ \forall j \in J. \tag{10}$$

Then (8d) becomes a linear constraint, i.e.,

$$t_{ij} = \alpha_{ij} + s + \gamma \cdot \mu_j, \ \forall i \in I, j \in J. \tag{11}$$

However, (10) states that μ_j is a non-convex function over two-dimensional variables v_j and x_j, which is extremely difficult to approximate by linear functions. To circumvent the difficulty of handling the two-dimensional non-convexity, we take the logarithm on both sides of (10) with a mild assumption that v_j is nonzero (e.g., if $v_j = 0$, then we can set its value to a small number 10^{-4} during the approximation). This gives rise to

$$\ln \mu_j \geq \beta \ln v_j - \beta \ln(q_j + x_j), \forall j \in J, \tag{12}$$

where we also replace the equality in (10) by the "greater-than-or-equal-to" inequality without changing the optimal solution of [SEP] because μ_j should be as small as possible to ensure that the disutility of the chained facilities will optimize the profit of the operator.

Now, we need to deal with the logarithm function. Generically, (12) can be approximated by

$$F(\mu_j; \epsilon) \geq \beta \cdot F(v_j; \epsilon) - \beta \cdot F(q_j + x_j; \epsilon), \ \forall j \in J, \tag{13}$$

where $F(w; \epsilon)$ is defined as a piecewise linear function over w, which approximates $\ln w$ with a maximum relative approximation error of $\epsilon\%$, i.e.,

$$\left| \frac{F(w; \epsilon) - \ln w}{\ln w} \right| \times 100\% \leq \epsilon. \tag{14}$$

In modern MIP solvers, piecewise linear approximation has been embedded into their solution procedures as an internal function. These solvers have provided straightforward, robust, and error-adjustable approaches for automatically generating approximators in a mathematically compact manner (they also handle the potential numerical issue due to the possible occurrence of "zero" in the denominator of (13)). Thus, we prefer to directly use these reliable procedures rather than designing our own approximation scheme (although the latter can be easily done following the linearization techniques described in Sherali and Adams (2013)). In this paper, we employ the solver Gurobi 9.5.0 and leverage its general constraint function *addGenConstrXxx*[1].

Similarly, for the logit choice (8g), we approximate it by

$$\sum_{j \in J} y_{ij} = H_i(\tau_i; \epsilon), \ \forall j \in J, \tag{15}$$

where $H(\tau_i; \epsilon)$ is a piecewise linear function such that

$$\left| \frac{H(\tau_i; \epsilon) - \frac{d_i}{1+e^{\theta(\tau_i - T_i - \Psi_i)}}}{H(\tau_i; \epsilon)} \right| \times 100\% \leq \epsilon. \tag{16}$$

To summarize, [SEP] can be solved by the following approximate mixed-integer linear program subject to a relative error bound of $\epsilon\%$

$$\max \quad r \sum_{j \in J} v_j - \sum_{j \in J} f_j x_j \tag{17a}$$

$$\textbf{[MILP(}\epsilon\textbf{)]} \qquad \text{s.t. } (8b), (8c), (8h), (9), (11), (13), (15). \tag{17b}$$

Parameter ϵ allows for adjusting the error bound. In general, setting a smaller value of ϵ increases the approximation accuracy at the expense of a higher computational burden since the resultant piecewise approximators may contain significantly more numbers of pieces (each one corresponds to a linear segment). To balance the trade-off between accuracy and solution efficiency, in this paper, we set $\epsilon = 1$, meaning that we restrict the maximum relative approximation error to 1%, which is small enough to guarantee the approximation quality on the one hand, while avoiding adding an excess number of pieces into the problem on the other. Note that when conducting the piecewise linear approximation, points are selected from the original nonlinear function to construct linear segments. Such an approach is also implemented by Gurobi within its internal approximation scheme. We, however, cannot guarantee that the original function is approximated from above or from blow since the nonlinear function (e.g., the logit choice function in (8g)) may be neither convex nor concave on its domain. Fortunately, for the company, the only "controllable" variable is the expansion decision x, and applying [MILP(ϵ)], in general, allows us to derive a x that is close to the optimal solution as we will see in our numerical experiments.

4. Surrogate optimization approach

4.1. Motivation

Thanks to the powerfulness of modern solvers, the MILP approach described in Section 3 is straightforward to implement and, oftentimes, works well for small-scale instances of the problem. However, this approach replaces non-convexity terms with piecewise linear functions, and when we desire a high accuracy of approximation, a large number of linear segments needs to be introduced, which dramatically increases the computational difficulty. Consequently, the MILP approach cannot scale well on the problem size and may fail to solve even moderate-sized instances in a reasonable time, signalizing the need for an effective and scalable approach since the problem size in real-world business contexts could be substantially large. This section thus proposes a surrogate optimization approach, which explores the hidden bilevel structure of [SEP] and leverages a "learning-to-optimize" framework that effectively searches for high-quality expansion decisions. We are inspired by the following observation.

[1]See Manual: `www.gurobi.com/wp-content/plugins/hd_documentations/documentation/9.5/refman.pdf`

Lemma 1 (Equivalent bilevel formulation)**.** *[SEP] is equivalent to the following bilevel program*

$$\max\ r\sum_{j\in J} v_j - \sum_{j\in J} f_j x_j \tag{18a}$$

$$\text{s.t. } L_j \le x_j \le U_j, \forall j \in J, \tag{18b}$$

[BP]

$$\min_{v,y,z}\ \sum_{i\in I}\sum_{j\in J}\alpha_{ij}y_{ij} + \sum_{j\in J}\int_0^{v_j} c_j(w,x_j)dw + \sum_{i\in I}\left(\frac{1}{\theta}\left[z_i\ln z_i + (d_i - z_i)\ln(d_i - z_i)\right] + (T_i + \Psi_i)z_i\right) \tag{18c}$$

$$\text{s.t. } \sum_{j\in J} y_{ij} + z_i = d_i, \forall i \in I, \tag{18d}$$

$$v_j = \sum_{i\in I} y_{ij}, \forall j \in J, \tag{18e}$$

$$y_{ij} \ge 0, \forall i \in I, j \in J, \tag{18f}$$

where the expansion decision x is the upper-level decision variable; (18c)-(18f) defines the lower-level problem, in which variable z_i is the number of customers at zone i seeking services from the outside option.

Proof. See Appendix A. □

We interpret [BP] as follows. In the upper-level problem, the operator determines the expansion decision x aiming to maximize the profit; whereas in the lower-level problem, the objective function in (18c) considers four parts: the fixed accessing time of the customers, the congestion effect at the facilities, the entropy functions that characterize customers' dual-channel choices, and the outside option time cost. The key point is that Problem (18c)-(18f) and Constraints (8c)-(8h) are mathematically equivalent, and solving Problem (18c)-(18f) allows the operator to anticipate customer behaviors under the UE-facility choice and the BLM-channel choice, defined by Constraints (8c)-(8h) in [SEP], and subsequently, estimate the revenue. To facilitate readers' understanding, we provide a toy example in Appendix C. In fact, transforming UE conditions into an optimization problem is an effective and widely applied approach in congested network analysis. The following lemma explains our motivation to derive such a transformation.

Lemma 2 (Convexity and uniqueness of the lower-level problem)**.** *Given a feasible expansion decision at $\bar{x}$, the lower-level problem at $\bar{x}$, defined as,*

$$\min_{v,y,z}\ \sum_{i\in I}\sum_{j\in J}\alpha_{ij}y_{ij} + \sum_{j\in J}\int_0^{v_j} c_j(w,\bar{x}_j)dw + \sum_{i\in I}\left(\frac{1}{\theta}\left[z_i\ln z_i + (d_i - z_i)\ln(d_i - z_i)\right] + (T_i + \Psi_i)z_i\right) \tag{19a}$$

[LLP($\bar{x}$)]

$$\text{s.t. } \sum_{j\in J} y_{ij} + z_i = d_i, \forall i \in I, \tag{19b}$$

$$v_j = \sum_{i\in I} y_{ij}, \forall j \in J, \tag{19c}$$

$$y_{ij} \ge 0, \forall i \in I, j \in J, \tag{19d}$$

is a convex optimization problem with a unique optimal solution of (v, y, z).

Proof. See Appendix B. □

According to this lemma, for any $\bar{x}$, [LLP($\bar{x}$)] can be solved efficiently by standard convex optimization approaches or software (e.g., CVXPY and MOSEK). Therefore, the company's profit can be easily evaluated when an expansion decision is given. Meanwhile, [LLP($\bar{x}$)] has a unique optimal solution. In the context of bilevel programming, this property guarantees that the "optimistic" and "pessimistic" formulations coincide (Colson et al., 2005). For any $\bar{x}$, the operator will thus observe a unique profit. To maximize the profit of the operator (i.e., to solve [SEP]), we only need to design an optimization framework that searches for an optimal expansion decision in an efficient and intelligent fashion. This observation inspires our surrogate optimization algorithm, which will be presented next.

4.2. Framework

The basic idea of surrogate optimization methods (also known as response surface or meta-model) is to approximate a complex objective function with a structural surrogate model that is learned and updated on-the-fly. The advantage of the surrogate optimization model is that it is a "learning-to-optimize" approach that efficiently constructs an input-output mapping for both expensive simulations and complex mathematical models (Regis and Shoemaker, 2007). The surrogate model is trained using a data-driven approach, where all previously visited solutions and their corresponding objective function values are kept as the training dataset. During the optimization phase, information from the surrogate model is used to guide the search for improved solutions. As more solutions are visited, the training dataset is enlarged, and the surrogate model will gradually be refitted and updated to better approximate the objective function and to locate a high-quality solution.

For [BP], when the expansion decision is given at x, we solve [LLP(x)] to obtain v_j and then evaluate the profit $G(x)$. That is, for any x, we can obtain its corresponding profit $G(x)$ by solving the convex optimization problem [LLP(x)]. We refer to this subroutine as *function evaluation*. Certainly, there exists a "mapping" between x and $G(x)$. In this paper, we adopt Radial Basis Function (RBF)-based surrogate model to learn the underlying mapping and to guide us towards optimizing the expansion decision.

In the literature (Mueller, 2014; Eriksson et al., 2019; Tian et al., 2021), the surrogate optimization framework consists of (i) *Initial experiment design*, which generates a set of initial expansion decisions. We set the number of initial expansion decisions as $n = \max\{20, 2(|J| + 1)\}$. Moreover, x is restricted by box constraints (i.e., $L_j \le x_j \le U_j$), and we thus obtain n initial expansion decisions using the symmetric Latin hypercube sampling (SLHS) approach (Kenny et al., 2000); (ii) *Function evaluation*, which receives an expansion decision x as input and derive the true profit $G(x)$ as output; (iii) *Surrogate model training*. The pairs of inputs and their outputs would be used to learn a RBF-based surrogate model, which constructs an approximate mapping from the expansion decision to the profit; and (iv) *candidate point selection*. The basic idea is to generate a set of candidate points and use the surrogate model to predict the function values at these points. The merit function is then adopted to select the most promising candidate points that will go through function evaluation and then be added to the training set in the next iteration.

With these ingredients, we present our algorithm as follows.

Step 1 Initialization.

- **Step 1.1** Generate n initial expansion decision $\{x^1, x^2, ..., x^n\}$ using SLHS.
- **Step 1.2** Initialize the training dataset $\mathbf{X}$ as $\mathbf{X} = \{x^1, x^2, ..., x^n\}$.
- **Step 1.3** Conduct function evaluation under x^i by solving [LLP(x^i)] and derive the corresponding profit $G(x^i), \forall i = 1...n$.
- **Step 1.4** Train the RBF-based surrogate model $Z(x)$ using the data $\{(x^i, G(x^i)), \forall i = 1...n\}$, derived in Step 1.1 and Step 1.2, to fit the initial surrogate model and establish the response surface. That is, we approximate the profit by a RBF model. More specifically, given a set of points x^i and their corresponding profit $G^i(x)$, the surrogate model takes the form of:

$$Z(x) = \sum_{i=1}^{n} \lambda_i \varphi\left(\|x - x^i\|\right) + p(x),$$

where $Z(x)$ approximates the true profit $G(x)$; λ_i is a weight coefficient; $\varphi\left(\|x - x^i\|\right) = (\|x - x^i\|)^3$ is the *cubic kernel* (where the norm is the Euclidean norm); and $p(x)$ is a linear tail, taking the form of $a + b^T x$. In the training process, we use the least-squared approach to find out the value of (λ, a, b).

Step 2 Re-train the surrogate model.

Step 2.1 Use the current surrogate model to predict the profit at candidate points in the variable domain and decide on which point the next function evaluation should be conducted. The weighted-distance merit function (which will be explained later) is adopted (Regis and Shoemaker, 2007; Liu et al., 2018; Eriksson et al., 2019).

Step 2.1.1 Generate a set of candidate points Ω_x with $|\Omega_x| = 100 \cdot |J|$. We adopt the normal random candidate point method, in which the points near the vicinity of the current best solution x^* are obtained by adding random perturbations.

Step 2.1.2 Predict the function values of the candidate points. After generating the set of candidate points Ω_x, the surrogate model $Z(x)$ is used to predict their objective function values, i.e, for $x' \in \Omega_x$, we compute $Z(x')$. Then we can derive $Z^{\max} = \max Z(x'), x' \in \Omega_x$ and $Z^{\min} = \min Z(x'), x' \in \Omega_x$.

Step 2.1.3 Determine the minimum distance from previously evaluated points. Let $d(x')$ denote the distance from x' to the closest evaluated point. Compute $d^{\max} = \max d(x'), x' \in \Omega_x$ and $d^{\min} = \min d(x'), x' \in \Omega_x$.

Step 2.1.4 Compute the weighted score. We first derive the scaled surrogate score by:

$$V^s(x') = \begin{cases} \frac{Z(x') - Z^{\min}}{Z^{\max} - Z^{\min}} & \text{if } Z^{\max} > Z^{\min} \\ 1 & \text{otherwise} \end{cases}$$

which is nonnegative and equals zero at points that have minimal surrogate value among candidate points.

The scaled distance score is computed by:

$$V^d(x') = \begin{cases} \frac{d^{\max} - d(x')}{d^{\max} - d^{\min}} & \text{if } d^{\max} > d^{\min} \\ 1 & \text{otherwise} \end{cases}$$

which is nonnegative and equals zero at points that are most far from evaluated points.

Then, the weighted-distance merit function is defined as $wV^s(x') + (1-w)V^d(x')$, where w is the weight to be selected from [0,1]. The weighted score is used as a criterion for selecting the later function evaluation points. The value of w balances the trade-off between "exploration" and "exploitation". That is, a larger value of w gives higher importance to the surrogate values, leading the search to minimize the surrogate; whereas a smaller value of w favors points that are far from evaluated points, motivating the search to explore new regions.

Step 2.1.5 We then select the top k candidate points from Ω_x according to the value of the weighted-distance merit function. Let $\mathbf{X}'$ be the set of these points (i.e., $\mathbf{X}' \subset \Omega_x$ and $|\mathbf{X}'| = k$).

Step 2.2 Conduct a function evaluation for each x' in $\mathbf{X}'$ to derive the associated profit $G(x')$. In total, there are k function evaluations in this step, and they can done in parallel using the multiprocessing function embedded in modern computer programming languages to reduce the computational time. In this paper, we set $k = 12$, which corresponds to the number of threads of the computer used in our experiment.

Step 2.3 Enlarge the training dataset by setting $\mathbf{X} := \mathbf{X} \cup \mathbf{X}'$.

Step 2.4 Update (re-train) the surrogate model $Z(X)$ using all data $(x^i, G(x^i)), \forall x^i \in \mathbf{X}$. With a gradually enriched dataset $\mathbf{X}$, the approximation accuracy will be improved accordingly.

Step 3 If the maximum number of function evaluations is unmet, return to Step 2; otherwise, stop and return the best solution found during function evaluation.

5. Computational experiments

This section presents computational studies to test the efficiency of the proposed solution approaches. Throughout this section, we set $f_j = f, \forall j \in J$. We fix parameters $f = 1, \gamma = 1, s = 1, \Psi_i = 0, \forall i \in I$. However, we will explore their impacts on the solution structure by varying their values in the next section.

We refer to the approximate mixed-integer linear programming approach in Section 3 as `MILP`, where the reformulated [MILP] is solved directly by Gurobi 9.5.0 using its embedded general constraint modeling functions. The maximum relative approximation error is set at 1%. Since [MILP] is an approximation, the profit derived by it may not match the true one. Therefore, after obtaining an expansion solution from it (say $\bar{x}$), we evaluate the true profit by solving [LLP($\bar{x}$)] and report the associated result. For the surrogate optimization framework proposed in Section 4, we refer to it as `SO`, where the convex subproblem is modeled using CVXPY and solved by calling the solver MOSEK. Moreover, to verify the effectiveness of our solution approach, we introduce a benchmark, `SCIP`, where the global optimization solver SCIP (i.e., PySCIPOpt Version 4.2.0) is applied to solve the original formulation [SEP]. All experiments are conducted using Python on a 16 GB RAM macOS computer with a 2.6 GHz Intel Core i7 processor.

5.1. Performance of `MILP` and `SO`

In our first experiment, we compare the computational performance of `SCIP`, `MILP` and `SO`. Our testbed is generated by the following method: The location of facilities and customers are on the plane and drawn from 2-dimensional uniform distribution $U([0, 20]^2)$. We set α_{ij} to the Euclidean distance between customer i and facility j. The demand of each customer zone d_i is drawn from uniform distribution $U([50, 100])$. Moreover, we set $T_i = \sum_{j \in J} \alpha_{ij}/|J|$, i.e., the average value of vector $\alpha_{i\cdot}$. All datasets used in this section are generated using Python with fixed random seeds and can be provided upon request.

In the context of the above setup, we generate 40 small-scale random instances as follows.

RND-S. Instances with the following combinations of instance sizes and parameters: $|I| = \{15, 20\}$, $|J| = \{4, 5, 6, 8, 10\}$, $\theta = \{0.1, 0.3\}$, and $\beta = \{1, 2\}$. We fix $q = 10$.

For `SCIP` and `MILP`, we impose a time limit of 3600 seconds for solving an instance. For `SO`, the number of function evaluations is 200.

Table 2 reports the computational results of 20 RND-S instances under $\beta = 1$, including the computational time in seconds (CPU[s]), the relative solution gap (rg[%]) reported by the solvers (i.e., Gurobi and SICP), the objective (Profit), and an additional measure Δ, which stands for the relative profit difference between `MILP` (or `SO`) and `SCIP`. It is computed as

$$\Delta = \frac{\text{Profit}(\texttt{MILP}) - \text{Profit}(\texttt{SCIP})}{\text{Profit}(\texttt{MILP})} \times 100\% \quad \text{or} \quad \Delta = \frac{\text{Profit}(\texttt{SO}) - \text{Profit}(\texttt{SCIP})}{\text{Profit}(\texttt{SO})} \times 100\%.$$

A positive value of Δ indicates that the approximation approach `MILP` or `SO` finds a better solution than `SCIP`; whereas a negative value indicates the opposite case. Moreover, for `SCIP` or `MILP`, if an instance cannot be solved within the time limit, we mark the corresponding CPU with "t.l.".

Table 2 shows that for those instances that can be successfully solved by `SCIP`, Δ values for `MILP` and `SO` are very small. This indicates that the solutions found by both approaches are close to the true optimal ones, which justifies the approximation quality of `MILP` and the effectiveness of `SO`. We also notice that `SO` is very fast and consistently finds better solutions than `MILP`.

Table 2: Computational results of the RND-S instances with $\beta = 1$. "t.l." indicates that the computational time exceeds the 3600-second time limit.

θ	$\|I\|$	$\|J\|$	SCIP			MILP				SO		
			CPU[s]	rg[%]	Profit	CPU[s]	rg[%]	Profit	Δ	CPU[s]	Profit	Δ
0.1	15	4	1319.5	0	398.3	7.3	0	398.2	-0.03	5.2	398.3	0
		5	525.4	0	414.6	25.8	0	414.6	0	5.2	414.6	0
		6	2728.5	0	435.0	146.5	0	434.9	-0.02	5.4	435.0	0
		8	978.8	0	497.5	798.5	0	497.2	-0.06	5.6	497.5	0
		10	343.5	0	538.6	587.9	0	538.6	0	6.0	538.6	0
	20	4	587.5	0	542.1	21.1	0	541.8	-0.06	5.4	542.1	0
		5	2073.7	0	563.2	112.0	0	562.7	-0.09	5.2	563.2	0
		6	t.l.	4.72	589.6	719.3	0	589.3	-0.05	5.5	589.7	0.02
		8	t.l.	0.46	683.8	2448.4	0	683.7	-0.01	5.6	683.8	0
		10	t.l.	3.93	729.4	t.l.	1.20	728.9	-0.07	6.1	729.4	0
0.3	15	4	346.9	0	350.2	4.8	0	349.7	-0.14	5.2	350.2	0.00
		5	552.2	0	377.8	9.1	0	377.6	-0.05	5.1	377.7	-0.02
		6	2324.4	0	413.3	35.6	0	412.9	-0.10	5.4	413.3	0
		8	821.9	0	539.4	243.9	0	538.9	-0.09	5.5	539.4	0
		10	1914.7	0	622.0	624.2	0	621.0	-0.16	5.8	621.8	-0.03
	20	4	538.7	0	469.5	15.3	0	468.6	-0.19	5.2	469.5	0
		5	t.l.	8.26	508.9	42.3	0	508.2	-0.14	5.2	508.9	0
		6	t.l.	11.97	559.5	254.0	0	559.2	-0.05	5.4	559.4	-0.02
		8	t.l.	4.30	752.9	1677.7	0	752.1	-0.11	5.5	752.9	0
		10	t.l.	8.47	847.5	t.l.	2.49	847.2	-0.04	6.0	847.5	0

To further compare the performance of these three approaches. We consider another 20 RND-S instances under $\beta = 2$. The results are given in Table 3. Here, `SCIP` may fail to identify any feasible solutions within the time limit. In this case, we mark, in the column of `SCIP`, rg[%] and Profit with "n.a.", and the corresponding Δ in the columns `MILP` and `SO` with "n.a.".

Clearly, the performance of `SCIP` and `MILP` turns out to be rather sensitive to congestion parameter β. Comparing the results in Table 2 and Table 3, we observe that when β increases from 1 to 2, the performance of `SCIP` and `MILP` largely deteriorates as can be seen from the substantial increase in either the rg or the CPU. In particular, `SCIP` cannot solve all instances and, oftentimes, terminates with huge gaps. Worse still, it even fails to find out feasible solutions for 4 instances. For `MILP`, 9 instances are not solved optimally. By contrast, the CPU of `SO` remains at a low level (less than 7 seconds).

In terms of the solution quality, `SO` consistently generates the best solutions, and `MILP` can yield high-quality solutions as well even for unsolved instances. Unfortunately, `SCIP` is practically ineffective since the value of Δ for `SO` and `MILP` can go up to more than 11%, meaning that the solutions derived by `SO` and `MILP` are significantly better.

We thus conclude that our proposed approaches outperform `SCIP`. Moreover, given that `SO` converges to better solutions than `MILP` and requires significantly lower CPU as well, we prefer `SO` over `MILP`.

Notably, the drawback of `MILP` is owing to (i) the increased problem size because linearizing the nonlinear constraints requires additional variables and constraints and (ii) the weak continuous relaxation bound due to the "big-M" introduced for handling the complementary condition in (9). For those unsolved instances, although `MILP` may find high-quality solution in a short time, the convergence of relaxation bound during Gurobi's branch-and-cut process appears to be rather slow. As an example, Figure 2 shows the movement of the incumbent (i.e., the objective under the current best solution) and the best bound of an unsolved RND instance ($I = 20, J = 10, \beta = 2, \theta = 0.3$). We observe that the incumbent quickly grows to 451.9 (which is near-optimal). However, Gurobi struggles in reducing the best bound, and significant CPU is thus consumed on closing the gap.

Table 3: Computational results of the RND-S instances with $\beta = 2$. "t.l." indicates that the computational time exceeds the 3600-second time limit. For `SCIP`, "n.a." indicates that no feasible solution is found upon termination.

θ	$\|I\|$	$\|J\|$	`SCIP` CPU[s]	`SCIP` rg[%]	`SCIP` Profit	`MILP` CPU[s]	`MILP` rg[%]	`MILP` Profit	`MILP` Δ	`SO` CPU[s]	`SO` Profit	`SO` Δ
0.1	15	4	t.l.	76.40	279.7	24.0	0	292.3	4.31	5.9	292.7	4.44
		5	t.l.	98.09	296.0	209.5	0	307.7	3.80	5.8	307.9	3.86
		6	t.l.	83.08	325.3	879.6	0	325.6	0.09	6.0	325.7	0.15
		8	t.l.	69.30	373.0	t.l.	11.66	379.4	1.69	6.0	379.4	1.69
		10	t.l.	80.20	395.6	t.l.	19.82	417.1	5.15	6.5	417.4	5.22
	20	4	t.l.	118.46	357.5	81.5	0	390.0	8.33	6.0	390.5	8.45
		5	t.l.	n.a.	n.a.	695.0	0	408.8	n.a.	6.2	409.2	n.a.
		6	t.l.	111.47	411.1	t.l.	10.36	431.1	4.64	6.2	431.4	4.71
		8	t.l.	107.09	463.0	t.l.	11.70	507.6	8.79	6.6	508.0	8.86
		10	t.l.	81.78	547.6	t.l.	29.93	547.8	0.04	6.9	548.0	0.07
0.3	15	4	t.l.	95.00	231.1	11.6	0	236.6	2.32	5.9	236.9	2.45
		5	t.l.	n.a.	n.a.	47.1	0	258.4	n.a.	5.8	258.7	n.a.
		6	t.l.	141.93	284.4	384.0	0	286.3	0.66	6.2	288.3	1.35
		8	t.l.	118.77	380.4	t.l.	6.07	384.2	0.99	6.3	385.0	1.19
		10	t.l.	n.a.	n.a.	t.l.	16.01	451.9	n.a.	6.9	453.0	n.a.
	20	4	t.l.	192.10	308.9	34.7	0	315.6	2.15	6.2	316.3	2.37
		5	t.l.	197.20	342.6	229.9	0	345.3	0.78	6.1	345.9	0.95
		6	t.l.	191.74	352.6	2367.9	0	384.3	8.25	6.4	384.8	8.37
		8	t.l.	155.46	467.4	t.l.	18.18	528.9	11.63	6.4	530.8	11.94
		10	t.l.	n.a.	n.a.	t.l.	35.28	604.1	n.a.	6.9	604.4	n.a.

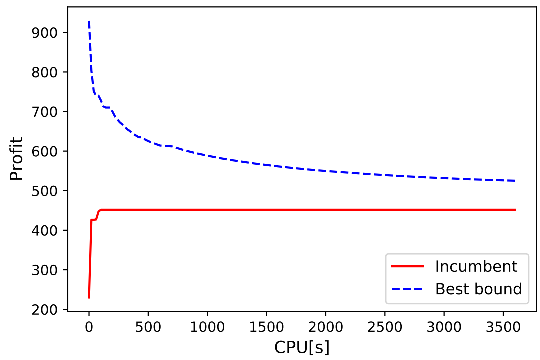


Figure 2: The movement of the incumbent solution and the best bound for an unsolved RND instance ($|I| = 15, |J| = 10, \beta = 2, \theta = 0.3$) by `MILP`.

5.2. *Comparing* `SO` *with other heuristics*

Our second experiment aims to compare the performance of `SO` with other heuristic approaches using large-scale instances. Our benchmark approaches consists of a particle swarm optimization algorithm (`PSO`) and a genetic algorithm (`GA`), where `PSO` and `GA` are used to search for improved expansion decisions and CVXPY-MOSEK is used to solve the convex subproblem to evaluate the profit under a given expansion decision (i.e., the function evaluation is exactly the same as that in `SO`). Both approaches are implemented using *scikit-opt* (available at `https://scikit-opt.github.io`) under default parameter settings. Note that for `PSO` and `GA`, their performance could be affected by the population size in each generation. To be fair in our comparison, we conducted preliminary computational experiments using population sizes of 20, 30, 40, and 50. We found that when the population size is 30, both `PSO` and `GA` perform best in general. Therefore, we will only report the results under the population size of 30 hereafter.

Leveraging the data generation method described in Section 5.1, we generate large-scale random instances as follows.

RND-L. Instances with 100 customer zones ($|I| = 100$). We consider the following combinations of instance sizes and parameters: $|J| = \{20, 30, 40\}$ and $\beta = \{1, 2\}$. We fix $q = 20$ and $\theta = 0.3$.

We start by comparing the performance of SO, PSO, and GA under the same number of function evolutions. Specifically, for all approaches, we set the maximum number of function evaluations as $|J| \times 30$. Meanwhile, for each instance, we perform 10 independent runs. This allows us to (possibly) obtain 10 different values of the objective function (profit) for each approach. We then use "AVG", "MAX", and "ST" to denote the average profit, the maximum profit, and the standard deviation of profits in these 10 independent runs. Moreover, we compute the relative profit difference between PSO (or GA) and SO as

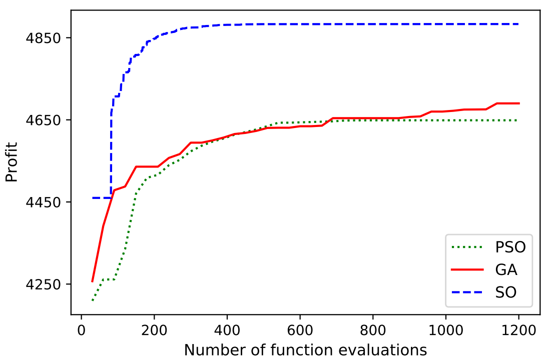
$$\tilde{\Delta} = \frac{\text{AVG}(\texttt{PSO}) - \text{AVG}(\texttt{SO})}{\text{AVG}(\texttt{SO})} \times 100\% \quad \text{or} \quad \tilde{\Delta} = \frac{\text{AVG}(\texttt{GA}) - \text{AVG}(\texttt{SO})}{\text{AVG}(\texttt{SO})} \times 100\%.$$

A negative value of $\tilde{\Delta}$ indicates that, on average, PSO (or GA) finds a solution that is worse than SO by $|\tilde{\Delta}|\%$, and vice versa

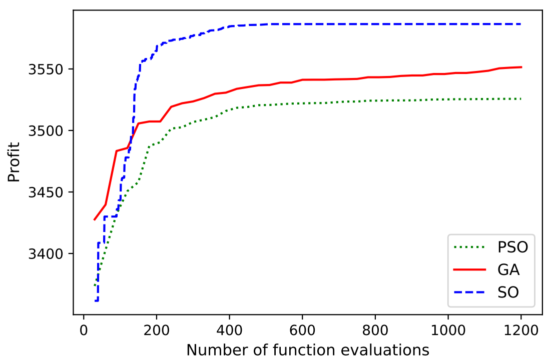
Table 4: Computational results (profit) of RND-L instances under $|J| \times 30$ numbers of function evaluations. Each row shows the statistics for 10 independent runs.

β	$\|J\|$	PSO				GA				SO		
		AVG	MAX	ST	$\tilde{\Delta}$	AVG	MAX	ST	$\tilde{\Delta}$	AVG	MAX	ST
1	20	4253.4	4296.9	50.15	-3.08	4270.5	4321.7	33.67	-2.69	4388.4	4389.2	0.36
	30	4370.5	4572.2	130.83	-6.71	4321.2	4450.1	105.19	-7.76	4684.8	4685.2	0.38
	40	4300.5	4611.3	168.72	-11.92	4258.1	4542.7	157.06	-12.79	4882.4	4883.0	0.62
2	20	3028.9	3060.5	25.12	-2.07	3043.6	3070.0	20.24	-1.60	3093.0	3094.2	1.03
	30	3248.5	3290.9	38.13	-4.12	3281.5	3330.9	29.60	-3.14	3388.0	3389.3	1.32
	40	3363.7	3468.8	79.31	-6.47	3400.3	3486.9	47.12	-5.45	3596.5	3599.1	1.96

Table 4 shows the computational results (profit) of RND-L instances. For all instances, we consistently observe that SO yields the best AVG. In particular, when $\beta = 1$ and $|J| = 40$, AVG by SO is more than 11% higher than that by PSO and GA. Moreover, the performance of SO is very robust because its ST is rather small, whereas the performance of PSO and GA under $|J| \times 30$ numbers of function evaluations turns out to be rather unstable as can be inferred from their larges values of ST.



(a) $\beta = 1, |J| = 40$



(b) $\beta = 2, |J| = 40$

Figure 3: Profit versus the number of function evaluations for two RND-L instances.

In fact, for metaheuristic approaches (e.g., PSO and GA), a large number of function evaluations are generally required to search for high-quality solutions. The poor performance of PSO and GA in Table 4 could thus be attributed to insufficient evaluations. To demonstrate this argument, we plot the evolution of the profit for two RND-L instances in Figure 3. Clearly, `PSO` and `GA` are still far away from convergence despite having performed 1200 function evaluations. On the contrary, `SO` converges rather fast. This result justifies that applying the surrogate optimization framework can largely reduce the number of function evaluations required for convergence and lead to better solutions than the benchmarks as well.

Table 5: Computational results (profit) for RND-L instances under 300-second computational time. Each row shows the statistics for 10 independent runs.

β	$\|J\|$	PSO				GA				SO		
		AVG	MAX	ST	$\tilde{\Delta}$	AVG	MAX	ST	$\tilde{\Delta}$	AVG	MAX	ST
1	20	4283.2	4339.5	49.29	-2.42	4383.0	4386.0	2.79	-0.15	4389.5	4389.6	0.09
	30	4379.0	4589.0	135.47	-6.54	4611.1	4663.2	31.29	-1.58	4685.3	4685.5	0.12
	40	4318.4	4611.3	169.76	-11.55	4701.8	4781.9	57.14	-3.70	4882.6	4883.2	0.48
2	20	3049.4	3077.4	24.26	-1.45	3077.2	3089.2	6.85	-0.55	3094.3	3094.7	0.26
	30	3257.5	3315.3	40.78	-3.88	3340.0	3359.1	13.97	-1.44	3388.9	3390.0	0.77
	40	3370.1	3470.3	80.13	-6.35	3518.3	3549.2	21.38	-2.23	3598.6	3600.8	1.35

Next, we compare the performance of `SO`, `PSO`, and `GA` under the same computational time. Specifically, we set the maximum computational time to 5 minutes (300 seconds). For each instance, we again perform 10 independent runs. Table 5 reports the results. Similar to the previous experiment, `SO` finds better (higher AVG and MAX) and more robust (smaller ST) solutions than the others.

To summarize, it is reasonable to conclude that `SO` is very effective in finding high-quality expansion solutions for [SEP].

6. Sensitivity analysis and implication

This section conducts sensitivity analysis to draw managerial implications. We use a randomly generated instance from RND-L with $|I| = 100$ and $|J| = 20$. Due to a large number of parameters in our problem, we fix $s = \beta = 1$ in the congestion function to simplify our discussion. Meanwhile, as can be seen from (7), the impact of T_i and Ψ_i is similar, i.e., a larger value of T_i or Ψ_i tends to increase the number of customers being attracted to the chained facilities. Therefore, we will hold T_i fixed. The remaining parameters are (i) cost parameter $f_j = f, \forall j \in J$; (ii) congestion parameter γ; (iii) initial service capacity $q_j = q, \forall j \in J$; and (iv) choice model parameters θ and $\Psi = \Psi_i, \forall i \in I$. Their impacts will be explored in this section. Now, we introduce the notations that will be used in our later discussion.

- δ, increased profit, defined as the difference between the profit with and without service expansion.
- *Expansion*, summation of all expansion units, i.e., $\sum_{j \in J} x_j$.
- *Usage [%]*, defined as the proportion of customers using the service from the chained facilities, i.e., *Usage* $= \sum_{j \in J} v_j / \sum_{i \in I} d_i \times 100\%$.
- $\bar{\tau}$, (weighted) average routing time of all customers seeking services the chained facilities, defined as
$$\bar{\tau} = \frac{\sum_{i \in I} \sum_{j \in J} y_{ij} \tau_i}{\sum_{i \in I} \sum_{j \in J} y_{ij}}.$$

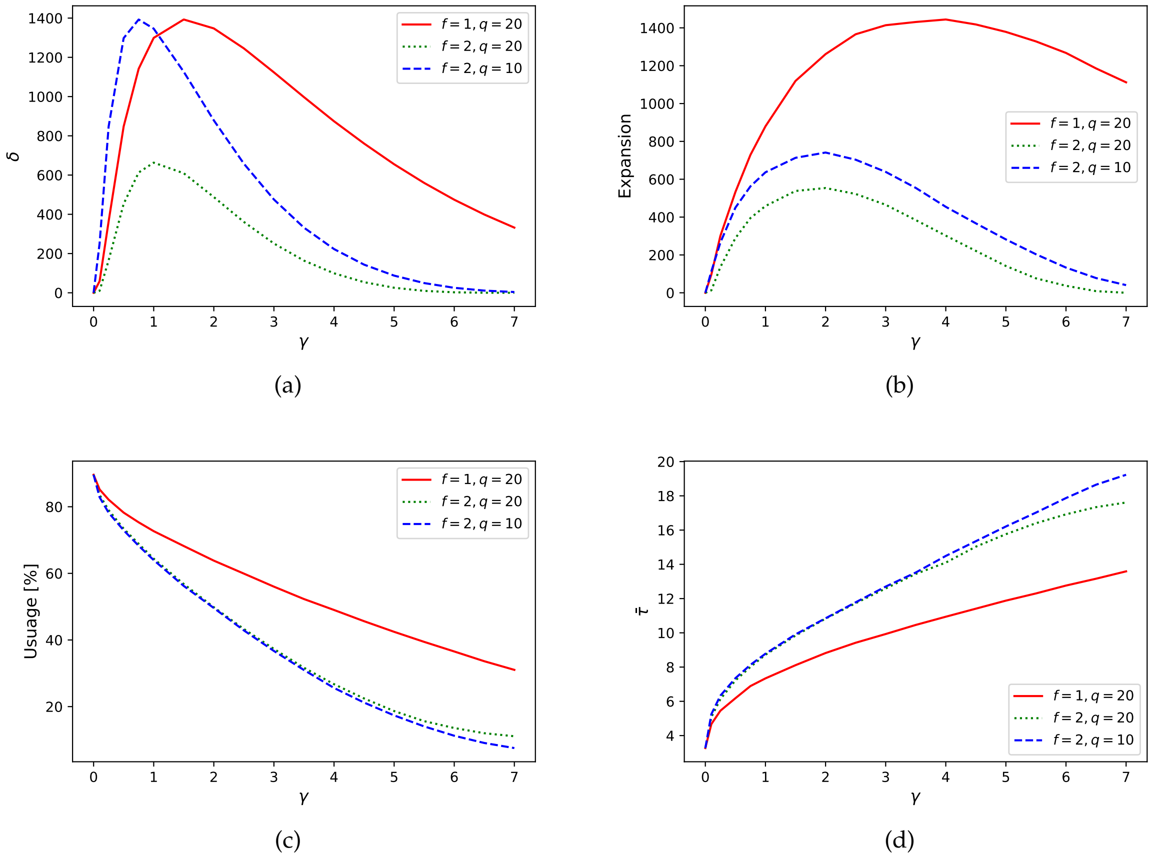


Figure 4: Sensitivity analysis on congestion parameters. We fix $\theta = 0.3$, $\Psi = 0$.

6.1. Impact of congestion parameter γ

We start with the sensitivity analysis on the congestion parameter γ. For this experiment, we fix $\theta = 0.3$, $\Psi = 0$.

We plot the results in Figure 4. Figure 4(a) reveals that for all combinations of (f, q), δ first increases with γ and then decreases. Based on (3), γ affects the magnitude of facility congestion. As revealed in Figure 4(d), $\bar{\tau}$ increases with γ. Essentially, when γ is very small, the congestion effect is marginal. Then, expanding services at facilities cannot effectively reduce customer disutility since the disutility has already been relatively low. In this case, the company can only attract a limited number of additional customers and thus should be conservative in the expansion decision. In the extreme case where the congestion effect is not presented (i.e., $\gamma = 0$), service expansion will bring no benefit to both the company and the customer, and thus *Expansion* is zero (see Figure 4(b)).

On the other hand, a very large value of γ indicates that the congestion effect is extremely significant. As shown in Figure 4(d), $\bar{\tau}$ is high when γ is large. In this case, additional customers attracted through service expansion may not compensate for the cost of expansion since customer disutility remains at a relatively high level. The company thus prefers not to invest substantially and should stay conservative when facing market competition. As a result, we observe that, when γ is high, both *Usage* and *Expansion* are low (see Figures 4(b) and 4(c). Only when γ is at intermediate levels, the benefit of service expansion is more likely to dominate the cost. The company is motivated to become more aggressive in service expansion. This explains the bell-shaped curves in Figures 4(a) and 4(b) and leads us to the first implication.

Implication 1. *Service expansion is preferred under moderate congestion parameters. For other cases, the* *company should be conservative in its expansion plan.*

6.2. *Impact of initial service capacity q*

From Figure 4, we observe that under the same value of f (i.e., $f = 2$), both δ and *Expansion* are higher when the initial service rate of facilities q is smaller (see the green and the blue plots). To further investigate the impact of q, we fix $\gamma = 1$ and vary q from 10 to 100.

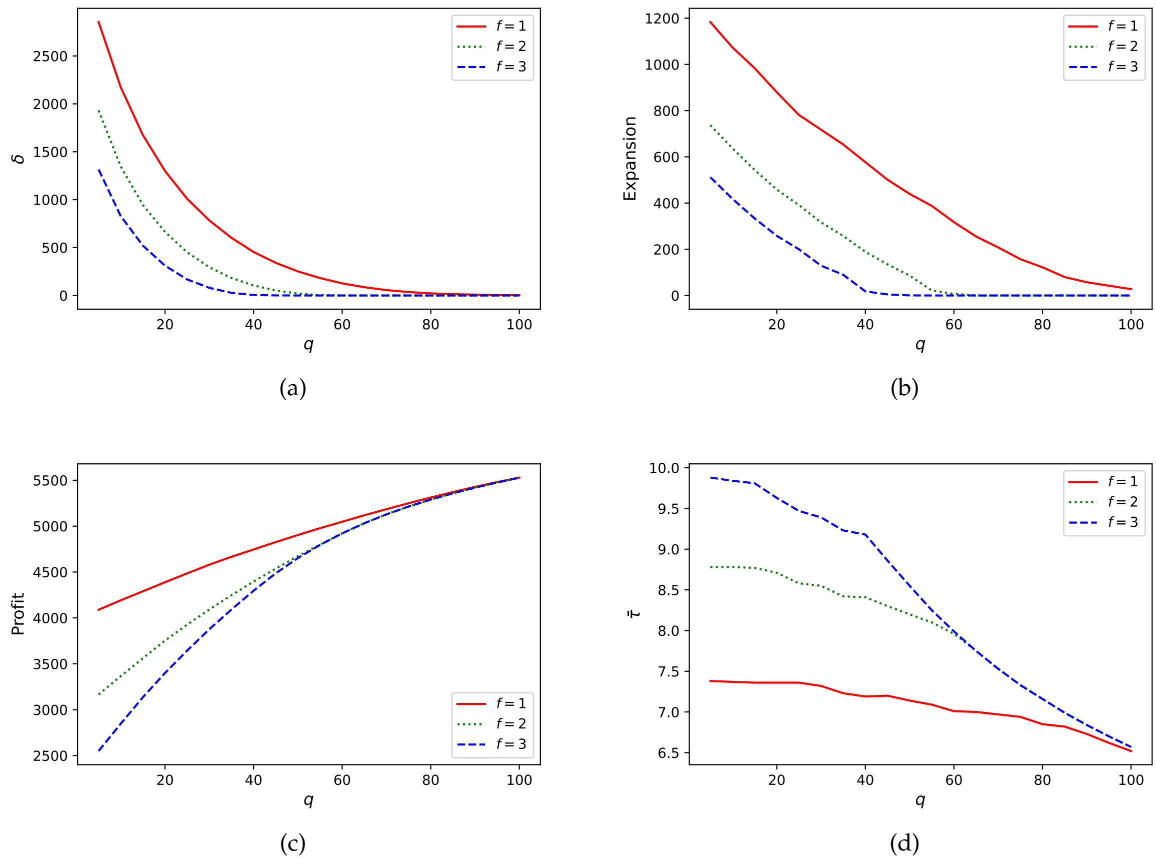


Figure 5: Sensitivity analysis on parameters q and f. We fix $\gamma = 1, \theta = 0.3, \Psi = 0$.

Figure 5 summarizes the main results. Notably, both δ and *Expansion* show non-increasing trends with q. Intuitively, if the initial service rate q is very high, then the initial congestion will not be significant, meaning that the disutility will be at a low level (e.g., see Figure 5(d) when $q \geq 80$), and the company has nearly achieved its maximum possible profit. As a result, expanding the service capacities of facilities cannot effectively reduce customer disutility to an extent where the additional revenue will exceed the expansion cost. This argument can be confirmed by the almost zero *Expansion* in Figure 5(b) and the nearly overlapping of lines in Figure 5(c) when $q \geq 80$.

Implication 2. *Service expansion is preferred when the initial service capacity is low.*

In effect, this implication is somewhat intuitive as the service expansion naturally favors the case where the initial service capacity is limited. Nevertheless, the results effectively demonstrate that a proper service expansion plan can indeed boost the profit but should be carefully determined, taking the congestion effect and the initial service capacity into consideration.

6.3. *Impact of market competition*

Next, we investigate the impact of choice model parameters θ and Ψ on the solution structure. For this experiment, we fix $\gamma = 1, q = 20, f = 2$.

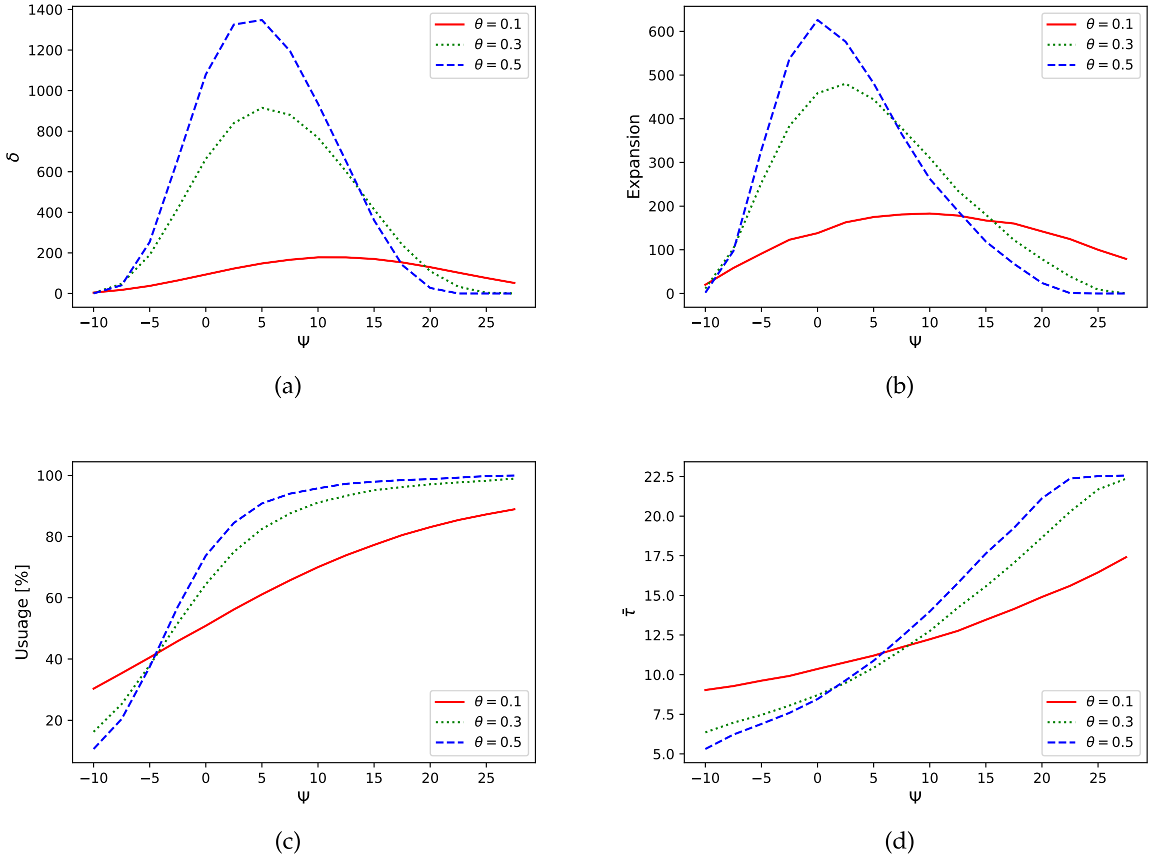


Figure 6: Sensitivity analysis on choice model parameters (θ, Ψ). We fix $\gamma = 1, q = 20, f = 2$.

We plot the results in Figure 6. Note that Ψ reflects the market competition level. When Ψ_i increases, the market competition from the competitor becomes weaker, and the company is expected to attract more customers.

For all values of θ, we observe bell-shapes in both Figure 6(a) and Figure 6(b), indicating that when market competition is intermediate, the company should invest more in service expansion to generate substantial additional profit. In other words, service expansion favors the case where market competition is intermediate. Certainly, when the market is very competitive (i.e., $\Psi < -5$ and we can interpret that the disutility of the outside is very low), the company can hardly win customers through the competition, and the majority of the market share will be taken by the competitor (see the low percentage of usage in Figure 6(c)). Under such fierce competition, it does not make sense for the company to aggressively expand the service capacity since the competitor will almost always stand in an advantageous position, and the investment will hardly yield rewards. On the other hand, when market competition is marginal, the company can attract most of the customers even without improving its service. In this case, there is limited motivation to engage in any expansion activity. As a result, we observe from Figure 6(d) that when Ψ is high (e.g., $10 < \Psi < 20$), the average time of chained facilities $\bar{\tau}$ increases sharply, which is owing to both the growing number of visitors to facilities and the reduced service expansion (the combined effect of them significantly intensifies the facility congestion). In other words, customers can benefit from the increasing market competition between service providers. Therefore, we have the following observation.

Implication 3. *Service expansion is preferred under moderate market competition. Moreover, customers can be better off if the market competition becomes more intense.*

7. Discussion and conclusion

This paper investigated a service expansion problem for chained business facilities under facility congestion and static market competition, wherein customer behaviors were virtually characterized as a two-stage process. More specifically, in the first stage, customers make channel choices, which were predicted by a discrete choice model. If customers choose the chain facilities, then they make facility choices at the second stage, following the user equilibrium principle. Both stages are inter-dependent since the channel choice provides the number of customers visiting the chained facilities based on the channel utility; whereas the equilibrium at the second stage yields the channel utility, which in turn affects the channel choice probability. The two-stage choices of customers thus must be explicitly and simultaneously considered. To this end, we proposed a modeling framework to optimize the expansion decision, accounting for estimated revenue and expansion costs. We subsequently provided two solution approaches and then conducted extensive computational experiments to demonstrate the effectiveness of them. Through comprehensive numerical studies, we observed that service expansion can generate substantial additional profit if congestion and market competition are moderate, or the initial service capacity is low. In these cases, the operator can be more aggressive in service expansion. Moreover, customers will ultimately benefit from the increasing competition between service providers.

Associated with this paper are a few limitations and future research directions. Above all, when solving large-scale instances, we utilized the surrogate optimization approach, which is a learning-based heuristic. Although the approach works well in practice, it is still worth trying to develop an exact solution algorithm for the model (if possible). Moreover, the surrogate optimization approach requires function evaluations (i.e., solving a subproblem to estimate customer behaviors). Through our computational experiments, we demonstrated that it can largely reduce the number of required function evaluations compared to the two metaheuristics. We, however, have to admit that the number of function evaluations for large-scale instances still needs several hundred. In the future, a faster convergence heuristic can be developed to expedite the solution process. Finally, this paper focused on static market competition. The utility of the outside option/the competitor was fixed and kept unchanged. In many business contexts, when the operator expands service capacities at its facilities, the competitor may react accordingly by locating new facilities or changing its services. Such a non-static problem requires game-theoretical modeling approaches, such as the Leader-follower game (Drezner et al., 2015; Küçükaydın et al., 2012; Lin et al., 2022) and the Nash game. In the area of capacity expansion/management, games-theoretical approaches have not received enough attention thus far, partly owing to the difficulty in the modeling and algorithmic development. This creates a gap that is interesting and worth investigating in the future.

Appendix A. Proof of Lemma 1

We show that Constraints (8c)-(8h) are equivalent to the following optimization problem

$$\min_{v,y,z} \sum_{i\in I}\sum_{j\in J}\alpha_{ij}y_{ij} + \sum_{j\in J}\int_0^{v_j} c_j(w,x_j)dw + \sum_{i\in I}\left(\frac{1}{\theta}\left[z_i \ln z_i + (d_i - z_i)\ln(d_i - z_i)\right] + (T_i + \Psi_i)z_i\right) \tag{A.1a}$$

$$\text{s.t.} \sum_{j\in J} y_{ij} + z_i = d_i,\ \forall i \in I, \tag{A.1b}$$

$$v_j = \sum_{i\in I} y_{ij}, \forall j \in J, \tag{A.1c}$$

$$y_{ij} \geq 0, \forall i \in I, j \in J. \tag{A.1d}$$

Let τ_i, μ_j, and φ_{ij} be the dual variables associated Constraints (A.1b), (A.1c), and (A.1d), respectively. We construct the Lagrangian function as follows

$$L(y,z,v,\tau,\mu,\varphi) = \sum_{i\in I}\sum_{j\in J}\alpha_{ij}y_{ij} + \sum_{j\in J}\int_0^{v_j} c_j(w,x_j)dw + \sum_{i\in I}\left(\frac{1}{\theta}\left[z_i\ln z_i + (d_i - z_i)\ln(d_i - z_i)\right] + (T_i + \Psi_i)z_i\right)$$
$$+\sum_{i\in I}\tau_i\left(d_i - \sum_{j\in J}y_{ij} - z_i\right) + \sum_{j\in J}\mu_j\left(\sum_{i\in I}y_{ij} - v_j\right) - \sum_{i\in I}\sum_{j\in J}\varphi_{ij}y_{ij}. \tag{A.2}$$

The KKT conditions then read

$$\frac{1}{\theta}\ln\frac{z_i}{d_i - z_i} + T_i + \Psi_i - \tau_i = 0, \tag{A.3a}$$
$$c_j(v_j, x_j) - \mu_j = 0, \tag{A.3b}$$
$$\alpha_{ij} - \tau_i + \mu_j - \varphi_{ij} = 0, \tag{A.3c}$$
$$\sum_{j\in J} y_{ij} + z_i - d_i = 0, \tag{A.3d}$$
$$v_j = \sum_{i\in I} y_{ij}, \tag{A.3e}$$
$$\varphi_{ij}y_{ij} = 0, \tag{A.3f}$$
$$\varphi_{ij} \geq 0, \tag{A.3g}$$

where (A.3a)-(A.3c) are the stationary conditions for variables z_i, v_j, and y_{ij}; (A.3d)-(A.3e) are primary feasibility; (A.3f) is the complementary slackness condition; finally, (A.3f) states dual feasibility.

By (A.3a), we can derive the solution of z_i as

$$z_i = \frac{d_i e^{\theta(\tau_i - T_i - \Psi_i)}}{1 + e^{\theta(\tau_i - T_i - \Psi_i)}}. \tag{A.4}$$

Plugging this condition into (A.3d) gives rise to

$$\sum_{j\in J} y_{ij} = \frac{d_i}{1 + e^{\theta(\tau_i - T_i - \Psi_i)}} \tag{A.5}$$

which exactly recovers Constraint (8g).

Now, plug (A.3b) into (A.3c). We have

$$\alpha_{ij} + c_j(v_j, x_j) - \tau_i = \varphi_{ij}. \tag{A.6}$$

Note that $\alpha_{ij} + c_j(v_j, x_j) = t_{ij}$. Therefore, the above equation becomes $t_{ij} - \tau_i = \varphi_{ij}$. Based on (A.3f)- (A.3g), we arrive at

$$(t_{ij} - \tau_{ij})\cdot y_{ij} = 0, \tag{A.7a}$$
$$t_{ij} \geq \tau_{ij}. \tag{A.7b}$$

These two equations are exactly Constraints (8e) and (8f). We thus complete the proof.

Appendix B. Proof of Lemma 2

The objective function of [LLP($\bar{x}$)] is strictly convex. To check this, note that when the expansion decision is fixed at $\bar{x}$, according to (3), the second term of the objective function (19a) can be expressed as a function of v as

$$\sum_{j\in J}\int_0^{v_j} c_j(w,\bar{x}_j)dw = \sum_{j\in J}\left(sv_j + \frac{\gamma\cdot v_j^{\beta+1}}{(\beta+1)\cdot(q_j+\bar{x}_j)^{\beta}}\right) \tag{B.1}$$

which is strictly convex since $\beta > 0$. Moreover, it is easy to see that the third term is also a strict convex function over z_i. Now, since the objective function of is strictly convex and all constraints are linear and feasible, the optimal solution of [LLP($\bar{x}$)] is unique.

Appendix C. A toy numerical example

We consider 1 customer zone with demand volume d and 2 service facilities. Now, since there is only one customer zone, we drop subscript i and thus variable y and the associated constraints. For the routing time parameters, we set $\alpha_1 = 1$ and $\alpha_2 = 4$ for Facility 1 and 2, respectively. Also, we set $\gamma = s = \beta = \theta = 1$, $T = 4$, and $\Psi = 0$.

Suppose that an expansion has been given at $\bar{x}$ and that the resultant service capacities are 2 and 4 for Facility 1 and 2, respectively. Then, the routing times of customers to Facility 1 and Facility 2 are $t_1 = 2 + v_1/2$ and $t_1 = 5 + v_2/4$, respectively.

Now, based on [SEP], we estimate the customer behavior by the UE condition and the logit choice, i.e.,

$$(2 + v_1/2 - \tau)\cdot v_1 = 0 \tag{C.1a}$$

$$(5 + v_2/4 - \tau)\cdot v_2 = 0 \tag{C.1b}$$

$$v_1 + v_2 = \frac{d}{1+e^{(\tau-4)}} \tag{C.1c}$$

$$v_1, v_2 \geq 0 \tag{C.1d}$$

which we refer to as the UE-based approach. On the other hand, [BP] describes the behavior by an optimization model. According to our setup, we have

$$\min\ 2v_1 + v_1^2/4 + 5v_2 + v_2^2/8 + z\ln z + (d-z)\ln(d-z) + 4z \tag{C.2a}$$

$$\text{st. } v_1 + v_2 + z = d \tag{C.2b}$$

$$v_1, v_2 \geq 0 \tag{C.2c}$$

Note that the objective function is strictly convex and thus, (C.2) has a unique optimal solution.

Next, we proceed to verifying the solutions of both approaches. For illustration purposes, we consider two representative demand scenarios.

Scenario 1 (low demand volume). We set $d = 8$. By inspection, $v_1 = 4, v_2 = 0, \tau = 4$ is a solution to the UE-based approach (C.1). Here, no customer visits Facility 2 since the associated routing time is 5, which is longer than that of Facility 1 (i.e., $t_1 = \tau = 4$). For the optimization-based approach, we use MOSEK to solve (C.2) and generate the same solution $v_1 = 4, v_2 = 0$.

Scenario 2 (high demand volume). We set $d = 40$. For the UE-based approach, we solve the nonlinear system (C.1) using Python-library scipy and derive $v_1 = 6.722$, $v_2 = 1.443$, $\tau = 5.361$. Here, both facilities attract certain volume of customers since $t_1 = t_2 = \tau$. For the optimization-based approach, we again use MOSEK to solve (C.2) and obtain $v_1 = 6.722$, $v_2 = 1.443$ as well.

This example shows that the two approaches are mathematically equivalent, despite having very different formulations.

References

Aboolian, R., Berman, O., and Karimi, M. (2022). Probabilistic set covering location problem in congested networks. *Transportation Science*, 56(2):528–542.

Aboolian, R., Berman, O., and Krass, D. (2012). Profit maximizing distributed service system design with congestion and elastic demand. *Transportation Science*, 46(2):247–261.

An, S., Cui, N., Bai, Y., Xie, W., Chen, M., and Ouyang, Y. (2015). Reliable emergency service facility location under facility disruption, en-route congestion and in-facility queuing. *Transportation research part E: logistics and transportation review*, 82:199–216.

Avila, M., Larco, J., Antonini, C., Ortíz, M., and Argueta, C. M. (2020). A discrete choice model for retailer selection in emerging markets. In *Supply chain management and logistics in emerging markets*. Emerald Publishing Limited.

Azeez, K. A., DongPing, H., and Mahmood, M. A. (2018). Capacity expansion decisions into asymmetric cost behaviour: reviews and search for new determinants. *International Journal of Services Operations and Informatics*, 9(2):139–159.

Berman, O. and Drezner, Z. (2006). Location of congested capacitated facilities with distance-sensitive demand. *IIE Transactions*, 38(3):213–221.

Chiou, S.-W. (2005). Bilevel programming for the continuous transport network design problem. *Transportation Research Part B: Methodological*, 39(4):361–383.

Colson, B., Marcotte, P., and Savard, G. (2005). Bilevel programming: A survey. *4or*, 3(2):87–107.

Dan, T. and Marcotte, P. (2019). Competitive facility location with selfish users and queues. *Operations Research*, 67(2):479–497.

Drezner, T. and Drezner, Z. (2002). Validating the gravity-based competitive location model using inferred attractiveness. *Annals of Operations Research*, 111(1-4):227–237.

Drezner, T., Drezner, Z., and Kalczynski, P. (2015). A leader–follower model for discrete competitive facility location. *Computers & Operations Research*, 64:51–59.

Drezner, T., Drezner, Z., and Zerom, D. (2018). Competitive facility location with random attractiveness. *Operations Research Letters*, 46(3):312–317.

Elhedhli, S., Wang, Y., and Saif, A. (2018). Service system design with immobile servers, stochastic demand and concave-cost capacity selection. *Computers & Operations Research*, 94:65–75.

Eriksson, D., Bindel, D., and Shoemaker, C. A. (2019). pysot and poap: An event-driven asynchronous framework for surrogate optimization. *arXiv preprint arXiv:1908.00420*.

Etebari, F. (2019). A column generation algorithm for the choice-based congested location-pricing problem. *Computers & Industrial Engineering*, 130:687–698.

Fernández, P., Pelegrín, B., Lančinskas, A., and Žilinskas, J. (2021). Exact and heuristic solutions of a discrete competitive location model with pareto-huff customer choice rule. *Journal of Computational and Applied Mathematics*, 385:113200.

Ferro, G., Minciardi, R., Parodi, L., and Robba, M. (2021). Optimal planning of charging stations in coupled transportation and power networks based on user equilibrium conditions. *IEEE Transactions on Automation Science and Engineering*, 19(1):48–59.

Fisk, C. (1980). Some developments in equilibrium traffic assignment. *Transportation Research Part B: Methodological*, 14(3):243–255.

Góez, J. C. and Anjos, M. F. (2017). Second-order cone optimization formulations for service system design problems with congestion. In *Modeling and Optimization: Theory and Applications*, pages 97–120. Springer.

Gunawan, H., Kusuma, A., Abdullah, S., et al. (2020). Parcel locker as the alternative of parcel delivery service of online shopping. *Journal of Computational and Theoretical Nanoscience*, 17(2-3):1311–1317.

Hamzeei, M. and Luedtke, J. (2018). Service network design with equilibrium-driven demands. *IISE Transactions*, 50(11):959–969.

Herman, O. and Ganz, Z. (1994). The capacity expansion problem in the service industry. *Computers & Operations Research*, 21(5):557–572.

Huff, D. L. (1964). Defining and estimating a trading area. *Journal of marketing*, 28(3):34–38.

Kenny, Q. Y., Li, W., and Sudjianto, A. (2000). Algorithmic construction of optimal symmetric latin hypercube designs. *Journal of statistical planning and inference*, 90(1):145–159.

Küçükaydın, H., Aras, N., and Altınel, İ. (2011). Competitive facility location problem with attractiveness adjustment of the follower: A bilevel programming model and its solution. *European Journal of Operational Research*, 208(3):206–220.

Küçükaydın, H., Aras, N., and Altınel, İ. (2012). A leader–follower game in competitive facility location. *Computers & Operations Research*, 39(2):437–448.

Lin, Y. H., He, D., Wang, Y., and Lee, L. H. (2020). Last-mile delivery: Optimal locker location under multinomial logit choice model. *Transportation Research Part E: Logistics and Transportation Review*, 142.

Lin, Y. H. and Tian, Q. (2021a). Exact approaches for competitive facility location with discrete attractiveness. *Optimization Letters*, 15(2):377–389.

Lin, Y. H. and Tian, Q. (2021b). Generalized benders decomposition for competitive facility location with concave demand and zone-specialized variable attractiveness. *Computers & Operations Research*, 130:105236.

Lin, Y. H., Tian, Q., and Zhao, Y. (2022). Locating facilities under competition and market expansion: Formulation, optimization, and implications. *Production and Operations Management*.

Liu, T., Ceder, A., and Rau, A. (2018). Deployment planning of single-line modular-vehicle semi-rapid transit system. In *Proceedings of the Conference on Advanced Systems in Public Transport and TransitData, Brisbane, Australia*, pages 23–25.

Liu, Z. and Meng, Q. (2014). Bus-based park-and-ride system: a stochastic model on multimodal network with congestion pricing schemes. *International Journal of Systems Science*, 45(5):994–1006.

Luss, H. (1982). Operations research and capacity expansion problems: A survey. *Operations research*, 30(5):907–947.

Manne, A. S. (1961). Capacity expansion and probabilistic growth. *Econometrica: Journal of the Econometric Society*, pages 632–649.

Marianov, V., Ríos, M., and Icaza, M. J. (2008). Facility location for market capture when users rank facilities by shorter travel and waiting times. *European Journal of Operational Research*, 191(1):32–44.

Martínez-Costa, C., Mas-Machuca, M., Benedito, E., and Corominas, A. (2014). A review of mathematical programming models for strategic capacity planning in manufacturing. *International Journal of Production Economics*, 153:66–85.

Mathew, T. V. and Sharma, S. (2009). Capacity expansion problem for large urban transportation networks. *Journal of Transportation Engineering*, 135(7):406–415.

Min, D., Ryu, J.-h., and Choi, D. G. (2018). A long-term capacity expansion planning model for an electric power system integrating large-size renewable energy technologies. *Computers & Operations Research*, 96:244–255.

Mueller, J. (2014). Matsumoto: the matlab surrogate model toolbox for computationally expensive black-box global optimization problems. *arXiv preprint arXiv:1404.4261*.

Parpas, P. and Webster, M. (2014). A stochastic multiscale model for electricity generation capacity expansion. *European Journal of Operational Research*, 232(2):359–374.

Regis, R. G. and Shoemaker, C. A. (2007). A stochastic radial basis function method for the global optimization of expensive functions. *INFORMS Journal on Computing*, 19(4):497–509.

Schön, C. and Saini, P. (2018). Market-oriented service network design when demand is sensitive to congestion. *Transportation Science*, 52(5):1253–1275.

Sherali, H. D. and Adams, W. P. (2013). *A reformulation-linearization technique for solving discrete and continuous nonconvex problems*, volume 31. Springer Science & Business Media.

Shriver, S. K. and Bollinger, B. (2022). Demand expansion and cannibalization effects from retail store entry: A structural analysis of multichannel demand. *Management Science*.

Suhara, Y., Bahrami, M., Bozkaya, B., and Pentland, A. S. (2021). Validating gravity-based market share models using large-scale transactional data. *Big Data*.

Tian, Q., Wang, D. Z., and Lin, Y. H. (2021). Service operation design in a transit network with congested common lines. *Transportation Research Part B: Methodological*, 144:81–102.

Wang, D. Z. and Lo, H. K. (2010). Global optimum of the linearized network design problem with equilibrium flows. *Transportation Research Part B: Methodological*, 44(4):482–492.

Wang, K.-J. and Hou, T. (2003). Modelling and resolving the joint problem of capacity expansion and allocation with multiple resources and a limited budget in the semiconductor testing industry. *International Journal of Production Research*, 41(14):3217–3235.

Wang, K.-J. and Lin, S.-H. (2002). Capacity expansion and allocation for a semiconductor testing facility under constrained budget. *Production Planning & Control*, 13(5):429–437.

Wogrin, S., Centeno, E., and Barquín, J. (2011). Generation capacity expansion in liberalized electricity markets: A stochastic mpec approach. *IEEE Transactions on Power Systems*, 26(4):2526–2532.

Xiong, Y., Gan, J., An, B., Miao, C., and Bazzan, A. L. (2017). Optimal electric vehicle fast charging station placement based on game theoretical framework. *IEEE Transactions on Intelligent Transportation Systems*, 19(8):2493–2504.

Ye, J., Jiang, Y., Chen, J., Liu, Z., and Guo, R. (2021). Joint optimisation of transfer location and capacity for a capacitated multimodal transport network with elastic demand: a bi-level programming model and paradoxes. *Transportation Research Part E: Logistics and Transportation Review*, 156:102540.

Zhang, F., Roundy, R., Cakanyildirim, M., and Huh, W. T. (2004). Optimal capacity expansion for multi-product, multi-machine manufacturing systems with stochastic demand. *IIE Transactions*, 36(1):23–36.

Zhang, Y. and Atkins, D. (2019). Medical facility network design: User-choice and system-optimal models. *European Journal of Operational Research*, 273(1):305–319.

Zhang, Y., Berman, O., and Verter, V. (2012). The impact of client choice on preventive healthcare facility network design. *OR spectrum*, 34(2):349–370.

Zhao, S., Haskell, W. B., and Cardin, M.-A. (2018). Decision rule-based method for flexible multi-facility capacity expansion problem. *IISE Transactions*, 50(7):553–569.