Transportation Research Part E 176 (2023) 103216

Contents lists available at ScienceDirect

Transportation Research Part E

journal homepage: www.elsevier.com/locate/tre

The service points' location and capacity problem

Tal Raviv

Faculty of Engineering, Tel Aviv University, Tel Aviv, Israel

ARTICLE INFO

Keywords:
Facility location
Parcel delivery
Stochastic modeling

ABSTRACT

We study the design of a network of automatic parcel lockers to facilitate the last-mile delivery of small parcels where parcels are delivered via service points near their recipients' home addresses. The recipients then pick up their parcels at convenient times. This method saves a substantial share of the handling and transportation costs associated with the parcel delivery process. The deployment of such a network requires decisions regarding the location and capacity of the service points. If a parcel has to be delivered through a service point with no remaining capacity, the parcel is sent directly to the recipient's address at a higher cost, or its delivery is postponed. Hence, there is a trade-off between the fixed setup cost, the variable operational cost, and the quality of the service. In this study, we take a bottom-up approach to the problem. We start by analyzing the dynamics of a single service point and show how to calculate a function that maps the parameters of its environment to the expected number of parcels that will be rejected from service or postponed. We then embed these functions in a mathematical model that optimizes the configuration of the network while considering the trade-offs described above.

1. Introduction

We study the design of a regional network of service points (SPs) equipped with automatic parcel lockers (APLs) to facilitate the last-mile delivery of small parcels. In such a network, parcels are delivered to SPs instead of directly to the recipients' homes. This method saves a substantial share of the handling and transportation costs associated with the parcel delivery process, see for example (Ranjbari et al., 2023). Moreover, while the method requires some effort from recipients who have to go to the SPs to pick up their parcels, it also provides them some advantages. If the system is well-designed and operated, the parcels are delivered to an SP within a short walking distance from the recipient's address. APLs are typically unattended facilities that operate 24/7, so recipients can pick up their parcels at any time. Compared with home-attended delivery, the hassle of synchronizing time with the courier is avoided. Compared with doorstep delivery, the risk of theft or loss of parcels due to misunderstandings is considerably mitigated. Indeed, there are some indications that this mode of parcel delivery is gaining popularity, especially among working-age recipients; see, for example, Lemke et al. (2016). From a public policy perspective, the usage of SPs for parcel delivery may mitigate the negative externalities of the parcel delivery process (e.g., traffic congestion, air pollution, and greenhouse gas emissions). For a thorough discussion of the advantages and opportunities gained from delivering parcels using APL, see Rohmer and Gendron (2020).

In their "Guide to parcel lockers for last-mile distribution", Rohmer and Gendron (2020) identified three important aspects of the good design of an SP network for parcel delivery, namely, the capacity of the APLs installed in the SPs, their locations, and the locker size assortment. This paper presents a mixed-integer linear programming (MILP) model that simultaneously addresses the first two aspects and is sufficiently lightweight to be solvable for large-scale instances. The locker box assortment issue is left for future research. For now, we assume that the APL units consist of lockers of the same size and that their configurations are selected from a short list of options (number of lockers).

E-mail address: talraviv@tau.ac.il.

https://doi.org/10.1016/j.tre.2023.103216
Received 25 November 2022; Received in revised form 9 May 2023; Accepted 29 June 2023
Available online 18 July 2023

The service delivery company replenishes the SPs in the network with new parcels periodically (e.g., every day). Each delivered parcel is left at the SP closest to the recipient's address. Recipients pick up their parcels at their discretion at some later time, possibly after a few replenishment cycles, referred to as *periods*. Parcels that are not picked up for several periods (e.g., a week) are removed from the SP by the delivery company.

The SPs should be deployed densely enough to enable easy access to any potential recipient in the service zone. The capacity of each SP should be sufficiently large to accommodate the periodic number of parcels sent to recipients in the service zone and to hold yet-to-be-picked parcels from previous periods. Since the demand faced by each SP is determined by its location and the location of neighboring SPs, the capacity and location problems should be solved simultaneously. Moreover, since there is uncertainty in both the number of parcels that are delivered to the SP every replenishment cycle and the number of parcels that are collected from it in the period between consecutive cycles, the SP can be viewed as a queue with limited capacity. There is a trade-off between the density of the SPs and the pooling opportunities.

In this study, we adopt a bottom-up approach to this design problem. We begin by analyzing a single SP with a given capacity that faces a stochastic stream of parcel pickups and periodic random replenishment by the parcel delivery company (referred to as *supply*). If the number of parcels that need to be delivered to an SP exceeds its available capacity, the excess is handled in a costlier way (e.g., doorstep delivery). From the single SP perspective, the excess demand is *rejected*. An alternative approach to handle the overflow at the SPs is to postpone the delivery of some packages. In this case, the parcels are queued at the regional depot and delivered to their destination SP in one of the next delivery cycles.

Based on our study of the rejection and postponement phenomena at SPs, we formulate the SP network location and capacity problem as a mixed-integer programming model that aims to minimize the total setup cost of the SPs and the cost of expected rejections.

The rest of this paper is organized as follows: in Section 2, we briefly review the relevant literature. In Section 3.1, we analyze SPs using a discrete-time Markov chain (DTMC) and use it to calculate the number of rejections as a function of capacity, the stochastic process of parcel delivery to its service zone, and the stochastic pickup process. We prove some useful properties of this function, demonstrate that it is robust with respect to the parameters of the pickup process, and show that it can be approximated very well by a piecewise linear function with a small number of pieces. In Section 4, we provide a formal definition of the SP location and capacity problem and formulate it as a MILP. We refer to this model as a piecewise linear (PWL) model (of the stochastic problem) since it tackles the stochasticity indirectly by incorporating the expectation of the objective function value as a known function. In Section 5, we present two additional modeling approaches to the problem inspired by the location theory literature: (1) A purely deterministic facility location model that can be solved with some fixed safety margin to meet the stochastic environment of the problem; (2) a scenario-based stochastic programming formulation. In Section 6, we present the results of our numerical experiments to demonstrate the merits of our proposed PWL model compared with the more traditional alternatives. Additionally, the computational viability of our MILP model for problem instances with realistic dimensions and parameters is demonstrated. Finally, in Section 8, we present some final thoughts and note concrete directions for future research in this fascinating domain of the design and operation of parcel delivery systems using APLs.

2. Literature review

This section reviews the literature on the optimal design of SP networks for parcel delivery. Moreover, since the SP location capacity problem is modeled as a stochastic facility location problem with discrete demand points, we also mention some recent studies in this rich literature.

Deutsch and Golany (2018) were the first to formulate the problem of designing a network of SPs as a facility location model. They assumed that the capacity of the SPs is unlimited; therefore, parcels are always delivered via the closest open SP. The delivery company compensates customers who need to travel a longer distance to collect their parcels. Their model aims to minimize the sum of the total facility setup cost and the sum of the compensation (discounts) given to customers based on travel distance. They show that the model can be cast as a classical uncapacitated facility location model. In their concluding remarks, the authors stress the need to study a capacitated version of the problem.

Bloch and Tzur (2019) studied a location problem of APLs used as an infrastructure for a hyper-connected network for parcel delivery using crowdsourcing. Their model aims to capture as many trips as possible to meet the demand for deliveries.

Rabe et al. (2020b,a) and Rabe et al. (2021) used a deterministic multiperiod capacitated facility location model to plan the future development and expansion of an urban SP network. Their model is based on a forecast of the population growth and the service's penetration over a planning horizon of ten years. Each year additional APLs can be opened in each district of the city. APLs may serve recipients from their current district or neighboring ones for some penalty cost. Future demand scenarios were obtained from a system dynamic simulation model that was applied to generate 20 demand growth scenarios that led to 20 different plans for expanding the SP network. The actual level of service provided by the proposed system was evaluated using a Monte Carlo simulation, but no feedback was returned to the deterministic optimization model. The method was applied in a realistic setting in the city of Dortmund with 68 candidate SP locations.

Che et al. (2022) presented a multiobjective deterministic model for placing SPs in a parcel delivery network when the goals are to minimize the unsatisfied demand, the overlap between the coverage of opened SPs and the total unused capacity of SPs. They present a genetic algorithm to solve this model.

Lin et al. (2020) and Lin et al. (2022) studied a competitive facility location model in the context of a parcel delivery network under different discrete choice models. They presented an optimization model to select the set of locations for SPs in the presence of (static) competition from other player(s) in the market. Their model focuses on the competition question but overlooks the stochasticity of the demand and the capacity of the SPs.

Grabenschweiger et al. (2022) study a location and routing model where APLs are used as an alternative for home delivery for some recipients. The goal is to minimize the cost over a known horizon, where the locations of the APLs are decided once, and the delivery routes are determined for multiple periods in advance. They present an effective mathematical formulation for the problem and a successful heuristic.

A different but related stream of literature aims to design and evaluate the operation of mobile APLs installed on vehicles or trailers that can serve multiple locations. Such mobile facilities can be autonomous or human-driven, see for example Schwerdfeger and Boysen (2022).

One of the most closely related works to the current study is that of Mancini et al. (2023), who model a service point location problem with stochastic demand and locker availability. They consider a problem where a given number of service points with a fixed capacity should be deployed in a service area. Their objectives are to maximize the number of parcels handled by APLS and, secondarily, to minimize the walking distance of recipients. They modeled the problem as a two-stage stochastic optimization integer program based on a set of generated scenarios. Each scenario conveys information regarding the availability of the installed lockers and the demand. They proposed a heuristic method and tested it on randomly generated instances, as well as on a realistic instance with actual locations and demand scenarios from Turin, Italy.

The current study differs from the work of Mancini et al. (2023) in the following aspects: (1) We incorporate the capacity decision into the network design problem; (2) We handle the stochasticity of the demand and locker availability by including in our objective function a component that represents the expected number of rejected parcels; (3) We introduce the setup cost of SPs into the objective function and allow this cost to be location and capacity dependent; (4) We enforce supplying each recipient via its closest SP.

Kahr (2022) presents a scenario-based formulation and a Benders decomposition technique to solve it. The current study differs from the work of Kahr (2022) in the following aspects: (1) We consider a stochastic parcel pick-up process; (2) We handle the stochasticity of the demand and locker availability by including in our objective function a component that represents the expected number rejected parcels; (3) We introduce the setup cost of SPs into the objective function rather than as a budget constraint; (4) We enforce supplying each recipient by its closest SP; (5) We assume that the SP configurations are determined by the number of lockers only, while Kahr (2022) allows an assortment of lockers of various types.

The most significant difference between the current study and other recent works, including Grabenschweiger et al. (2022), Mancini et al. (2023), and Kahr (2022), is the solution method. While the previous literature adopted scenario-based approaches, we model the demand and pickup processes using a DTMC and incorporate its prediction into the mathematical model. The preprocessing step greatly simplifies the optimization problem.

Reviews of earlier contributions to facility location under uncertainty and stochasticity can be found in Owen and Daskin (1998), Berman and Krass (2004), and Snyder (2006). The latter review distinguishes between stochastic models, where the planner assumes full knowledge of the distribution of the unknown parameters, and models with uncertainty, where the planner assumes little information about these distributions (e.g., only their support). The former line of models calls for stochastic optimization, while the latter calls for robust optimization. In the rest of this literature review, we focus on stochastic models, which are more relevant to the topic of this paper. More recent reviews can be found in Correia and Saldanha-da Gama (2019) and Berman and Krass (2019).

Baron et al. (2008) studied a facility location model where demand is assumed to be stochastically evenly distributed on the plane. Each facility serves all the points closest to it. The objectives are to cover the plane such that the maximal distance traveled by the customers is minimized and to set the service capacity of the facilities to minimize customer waiting time.

Albareda-Sambola et al. (2011) presented a capacitated facility location problem with discrete demand points, where the demand of each point follows a Bernoulli distribution. Decisions regarding the locations of facilities with fixed capacity and the allocation of demand to the facility are made ex ante, and an outsourcing alternative serves the surplus. The objective is to minimize the sum of the fixed costs of the open facilities plus the expected cost of outsourcing. The paper presents a close function of the outsourcing cost function for the case where the parameter of the Bernoulli distribution is the same for all demand points. This function is used to formulate a PWL mathematical model.

Pagès-Bernaus et al. (2019) modeled an e-commerce supply chain design as a stochastic capacitated facility location problem with the possibility of outsourcing surplus demand. They presented a scenario-based stochastic program and proposed a heuristic that hybridizes an iterated local search heuristic with simulation to address large-scale instances.

Turkeš et al. (2021) studied a facility location model with stochastic demand, inventory spoilage, and failure risk of transportation links used to supply the demand points. The model is motivated by the need to preposition emergency supplies in preparation for disaster relief. The goal is to minimize the expected unmet demand and response time (lexicographically) subject to a given budget for the facilities, supplies, and transportation. The uncertainty is modeled using a set of scenarios with known probabilities. The authors present a math heuristic that iteratively changes the facility's locations and capacity and evaluates the obtained solution by optimally assigning inventory to the opened facility.

At a superficial level, the stochastic SP location and capacity problem may appear to be a special case of the facility location problem with stochastic demand and capacity constraint. However, a closer look at the problem reveals some complications that stem from the stochasticity of the available capacity due to the randomness of the pickup process.

Many studies on the capacitated facility location problem model the stochasticity of demand using a set of scenarios. Binary decision variables typically represent the ex ante decisions about the location and capacity of the facilities, and the ex post decisions about the allocation of the demand are represented by continuous decision variables defined for each triplet of candidate facility location, demand point, and scenario. Some examples of such models can be found in Correia and Saldanha-da Gama (2019).

The scenario-based stochastic programming approach has two important virtues: (1) it allows formulating the first step (ex ante facility setup) decisions and the second step (ex-post allocation decisions) in the same mathematical program; (2) it allows capturing intricate interdependencies of the random variables based on real data or data that is generated realistically.

However, the scenario-based approach does not lend itself easily to the case of the SP location and capacity problem, where both the ex ante and ex post decisions are made under uncertainty. The uncertainty during the parcel delivery operation stems from the fact that the parcel may spend a random number of future periods in its locker until being picked up and interact with other parcels that are subject to the same type of uncertainty.

On the other hand, assuming that each parcel must be delivered to the SP closest to its recipient address, the demand and pickup processes faced by the SP can be determined ex ante at design time. We capitalize on this assumption to formulate a model that minimizes the expected number of parcels that the SPs cannot handle according to the decision at the design phase.

when compared with recent closely related studies such as Mancini et al. (2023) and Kahr (2022), the main contribution of this study is the analysis of the stochastic availability of lockers as well as the parcel rejection and postponement processes at the SPs rather than modeling the demand stochasticity by a set of scenarios and assume that the parcel pickup process is deterministic. We incorporate the stochastic modeling results into a mixed-integer linear model to optimize the location and capacity of SPs. We show that our model leads to a better design than alternatives borrowed from the facility location literature. Moreover, we demonstrate that, with the proposed formulation, large problem instances based on realistic urban settings can be solved by an off-the-shelf commercial mixed integer programming solver.

3. The single SP model

In this section, we construct and analyze a stochastic model of a single SP with a limited capacity replenished periodically by a random number of parcels that are later picked up according to another stochastic process. Overflow at the SP is handled either by rejecting excessive parcels or by postponing their delivery to the next period. The number of parcels that arrive at the SP in each replenishment cycle is assumed to be a Poisson random variable with rate parameter λ. The number of periods that each parcel spends in the SP until being picked up is drawn from a general distribution denoted by η.

This section is organized as follows. In 3.1, we model the SP with rejection in the case of overflow, assuming η is a geometric distribution with parameter p, as a discrete-time Markov chain (DTMC). In , we define the *rejection function* that maps the SP's capacity and arrival and departure rate to the expected number of parcels that cannot be served at each replenishment cycle. We show how specific points of this function can be calculated from the steady state probabilities of the DTMC and observe some useful convexity properties. In 3.2, we consider the case where the parcel overflow is handled by postponing delivery rather than rejection. For this case, we present an approximation method based on a classical result on $M/M/k$ queuing models.

In Appendices A and B, we use simulation to analyze SP models with rejection and postponement, where some simplifying modeling assumptions made in this section are removed. We demonstrate that our models can be used as good approximations for more realistic settings. In particular, the computational advantages of the DTMC model justify using it compared with the alternative noisy and computationally expensive simulation. For the postponement function, the result is not clear-cut. Under some circumstances, simulation is found to be the preferred method. Finally, in Appendix C we demonstrate that the rejection and postponement functions can be well approximated using piecewise linear functions based on a small number of calculated values.

3.1. The rejection function with memoryless time to pick up

In this section, we consider an SP that faces a periodic parcel arrival (replenishment) and parcel pickup that occurs during the period between consecutive replenishments. The random number of arrivals at each replenishment is $X \sim Pois(\lambda)$. The number of periods between the arrival of a parcel and its pickup is geometrically distributed with a parameter p. Therefore, if there are i parcels at the SP at the beginning of a period, the random number of parcels picked up during the period is $Y_i \sim Bin(i, p)$.

We define a DTMC with two sets of states (A, i) and (B, i) for $i = 0, \dots, C$, where C is the capacity of the SP. The state (A, i) represents the situation where there are i parcels in the SP immediately after its replenishment. Similarly, the state (B, i) represents the situation where there are i parcels in the SP immediately before replenishment. The DTMC is a bipartite-directed graph with transition probabilities:

$$P[(A, i), (B, j)] = P(Y_i = i - j) \quad \forall C \geq i \geq j \geq 0 \tag{1}$$

$$P[(B, j), (A, i)] = P(X = i - j) \quad \forall C > i \geq j \geq 0 \tag{2}$$

$$P[(B, j), (A, C)] = P(X \geq C - j) \quad \forall C \geq j \geq 0 \tag{3}$$

All other transition probabilities of the chain are zero. A graphical representation of this chain is presented in Fig. 1. The chain is periodic with a period length of two; that is, while the chain does not reach a steady state in the regular sense, in the long run, it has a limit for the probabilities of the states in odd and even time steps. For our analysis, we are interested in the distribution of

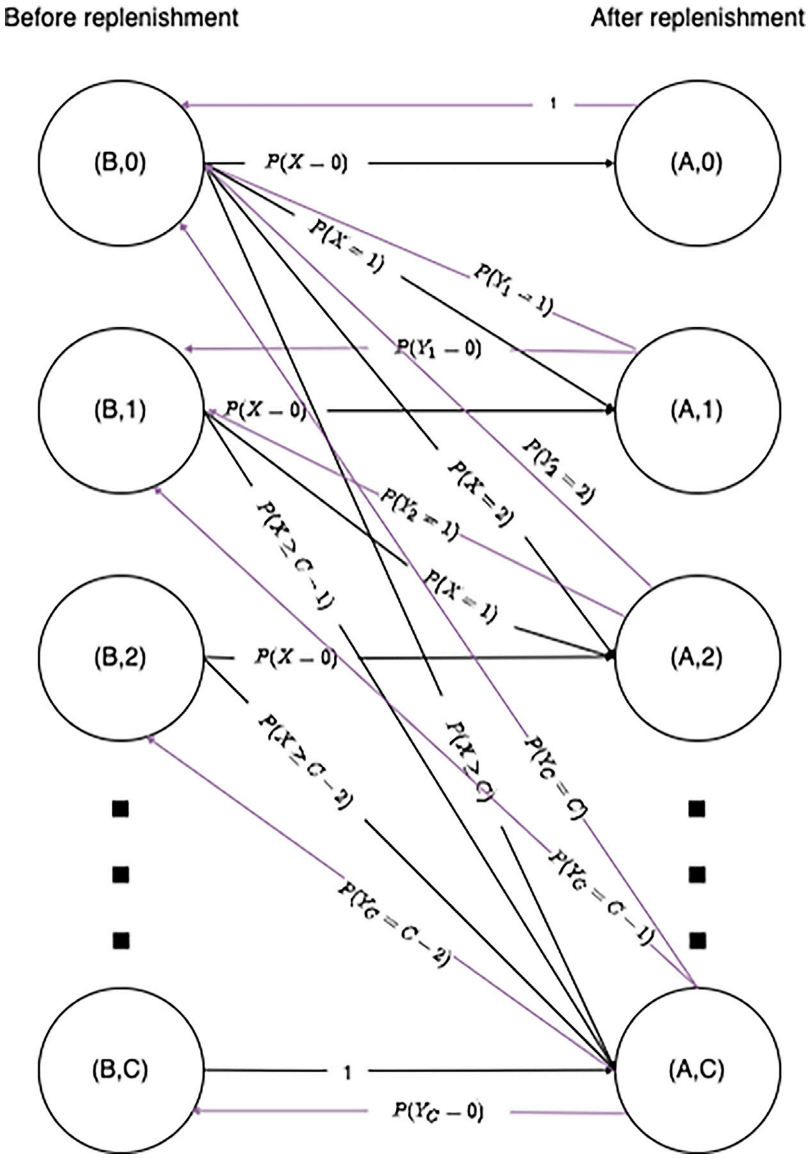

Fig. 1. DTMC of a single SP.

the (B,j) states, representing the long-run mean distribution of the number of parcels immediately before each replenishment. We slightly abuse the standard terminology of DTMC theory and call these probabilities the steady-state probabilities of the SP before replenishment. We denote them by π_j for $j = 0,\dots,C$. One can calculate these probabilities by solving a system of $2(C+1)$ linear inequalities or by looking at a high power of the transition probability matrix.

Given the steady-state probabilities of the DTMC, the expected number of parcels that cannot be accepted by the SP during replenishment, assuming a Poisson distribution of the supply with a rate λ, can be computed as follows.

$$R(C,\lambda,p) \equiv \sum_{j=0}^{C} \pi_j(C,\lambda_1,p) \sum_{k=C-j+1}^{\infty} \frac{\lambda^k e^{-k}}{k!}(k+j-C) \tag{4}$$

where $\pi_j(C,\lambda,p)$ are the steady-state probabilities of the (B,j) states when the parameters are C,λ and p. We refer to $R(C,\lambda,p)$ as the *rejection* function of an SP. The function $R(C,\lambda,p)$ is increasing convex in the supply rate λ. Formal proofs of these intuitive observations require the introduction of some heavy machinery and are out of the scope of this article. Instead, we provide a short sketch of this proof.

Proposition 1. *The rejection function $R(C,\lambda,p)$ is increasing and convex in λ.*

Proof (*Sketch*)**.** The higher the supply rate λ is, the higher the probability of more of the lockers being occupied at the end of each period before replenishment occurs. Strictly speaking, for $\lambda_2 > \lambda_1$ and for all $l \in \{0, \ldots, C\}$, we have

$$\sum_{j=l}^{C} \pi_j(C, \lambda_2, p) \geq \sum_{j=l}^{C} \pi_j(C, \lambda_1, p). \tag{5}$$

Parcels are accepted and rejected during replenishment at the end of each period as a batch. However, without loss of generality, we can assign an ordinal number to each parcel in the batch of a particular period. The probability of the ith parcel in the batch being rejected at the SP when the supply rate is λ_k is

$$\sum_{j=C-i+1}^{C} \pi_j(C, \lambda_k, p)$$

By inequality (5), the rejection probability of each parcel is higher for larger λ. Moreover, by definition, the arrival rate of parcels at the SP grows with λ, which further strengthens our argument about the monotonicity of $R(C, \lambda, p)$.

To prove that $R(C, \lambda, p)$ is convex in λ, we show that for some small $\epsilon > 0$

$$R(C, \lambda_1 + \epsilon, p) - R(C, \lambda_1, p) \leq R(C, \lambda_2 + \epsilon, p) - R(C, \lambda_2, p). \tag{6}$$

The difference $R(C, \lambda_k + \epsilon, p) - R(C, \lambda_k, p)$ is the marginal contribution (of additional ϵ parcels) to the number of rejections, which is, roughly speaking, the rejection probability times ϵ. Since the rejection probability increases in λ, the difference on the left-hand side of (6) is indeed not greater than the difference on the right-hand side, which implies the convexity of $R(C, \lambda, p)$. □

The expected number of parcels that arrive at the SP at the beginning of each period (λ) is an exogenous parameter of our model. However, the expected number of parcels that leave the SP during a period is determined endogenously by the steady-state probability,

$$\sum_{j=1}^{C} \pi_j j p \tag{7}$$

Since $\sum_{j=1}^{C} \pi_j < 1$ and in any term in (7) $jp < Cp$, we observe that Cp is an upper bound on the departure rate of parcels from the SP. Moreover, Cp is the expected number of picked parcels when the SP starts the period fully occupied. On the basis of this bound, we define the load ratio of an SP as

$$\rho = \frac{\lambda}{Cp}.$$

The analysis of the rejection function is intricate because it involves randomness, which is especially prominent when the SP is relatively small and operates near the verge of its capacity, i.e., with ρ close to 1. Indeed, when $\lambda \ll Cp$, the number of rejections is close to zero. When $\lambda \gg Cp$, $R(C, \lambda, p) \approx \lambda - Cp$. Moreover, when the SP capacity and supply are scaled up together, the system shows the same fluid behavior regardless of the ratio between λ and Cp,

$$\lim_{n \to \infty} \frac{R(nC, n\lambda, p)}{n} = \max(0, \lambda - Cp).$$

This phenomenon is visualized in Fig. 2, where we present the rejection rate (relative to the capacity of the SP) as a function of the load, ρ. The graphs are drawn for $p = 0.5$. Unfortunately, the above fluid approximation result is not very useful since SPs are designed to serve a relatively small area with moderate demand. If the network is designed properly, they work under ρ not very far from one.

Fig. 2 also visualizes the convexity of the rejection function in ρ which is proportional to λ for fixed p and C. The decreasing distance between the lines in the figure as C increases hints at the convexity of the function in C (when λ and p are constant), a property that may be useful for some modeling purposes but is out of our scope.

3.2. The postponement function

Until now, we focused on the case where the overflow of parcels at an SP is handled by rejecting them and sending them via a different (more expensive) delivery mode. This method of handling overflow is adopted in the previous literature on the optimal design of SP networks and, to the best of our knowledge, is also widely used in the parcel delivery industry during daily operations.

An alternative approach to handle parcel overflow can be to postpone the delivery of some parcels to a future period. In particular, it is reasonable to assume that overflowing parcels are queued at the regional depot and delivered on a first-come-first-served basis to their desired SP. The delivery company can save the extra cost associated with the external delivery mode at the cost of extending the delivery time, i.e., reducing the quality of service provided to the recipients. Under this method of handling parcel overflow, the delivery company must consider the trade-off between the cost of extending the capacity of its SPs and the expected cost of delivery postponement.

Below, we formalize this tradeoff as a postponement function, and we show how it can be approximated. We make similar assumptions to those we made in 3.1 with regard to the rejection function: (1) the SP is replenished periodically, say every day; (2) the number of new parcels that need to be delivered to the SP at the beginning of each period is a Poisson random variable with

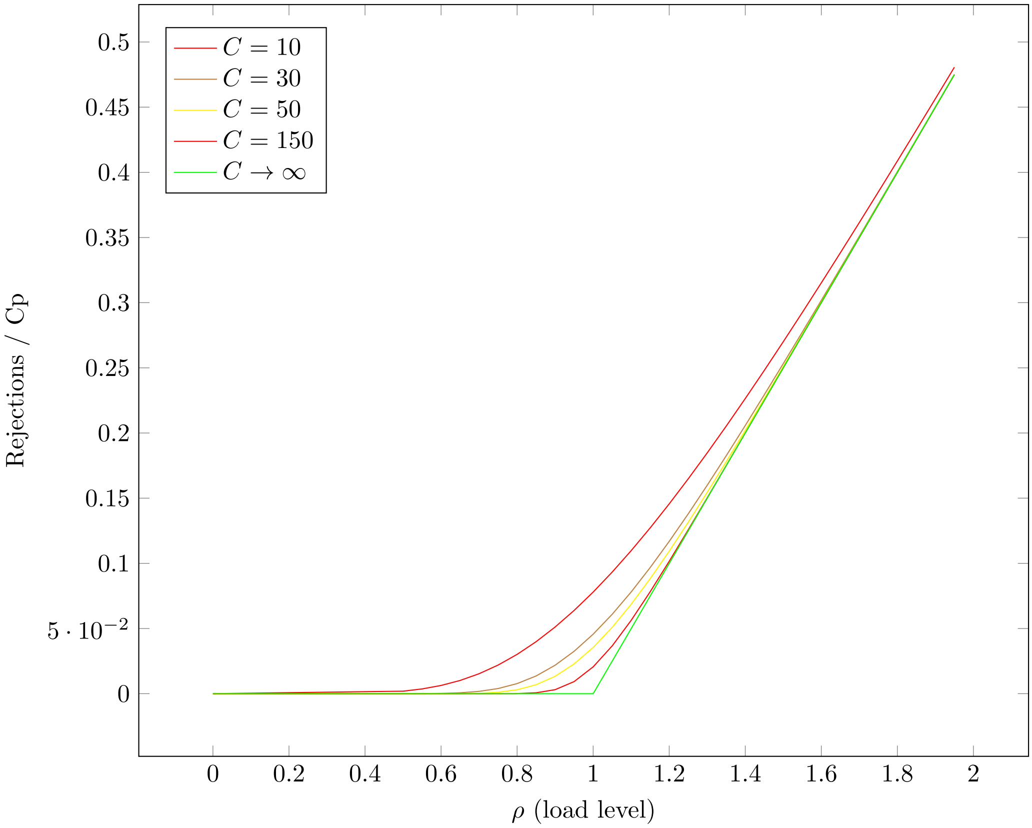

Fig. 2. Convergence to the fluid model (for $p = 0.5$).

rate parameter λ; and (3) The number of periods until a parcel is collected is a discrete random variable with distribution function η.

We define the postponement function

$$D(C, \lambda, \eta)$$

as the expected number of periods until delivery in an SP with a capacity of C lockers, where the arrival rate is λ and the pickup time distribution is η. A first important observation is that $D(C, \lambda, \eta)$ is defined only when the arrival rate is lower than the maximal pickup rate of parcels, i.e., only for $\lambda < \frac{C}{E(\eta)}$. Otherwise, the parcel queue "explodes", and the expected parcel postponement is indefinite.

In Appendix B, we present four different methods that we applied to approximate the value of the function $D(C, \lambda, \eta)$ for particular parameters. Here, we focus on two that prove to be successful. The most straightforward one is to estimate the expected postponement by simulating the system for a large number of periods and dividing the simulation into many shorter blocks that represent numerous sample points. The advantages of this approach are that we can represent any pickup process and that we can statistically evaluate the approximation error by establishing a confidence interval on the obtained sample. However, such an approach can be computationally demanding or excessively noisy if the number of simulated periods is too small.

An alternative approach that requires negligible computational effort is based on the observation that the SP can be (roughly) viewed as an $M/G/k$ queuing model. Recall that in the $M/G/k$ model, items (parcels in our case) arrive at the system according to a Poisson process. When a parcel arrives at the system, it starts being served (stored) in a server (locker) for some random time until the service ends (the parcel is picked up). Parcels that arrive when all the lockers are busy (occupied) wait in an unlimited queue (the regional depot) until a locker becomes available.

The reality of the SP is not exactly as described above. Parcels arrive at the system periodically rather than according to a Poisson process. However, the number of parcels arriving in each period is a Poisson random variable, meaning that from a macroscopic perspective, the arrival process is similar to a Poisson process. Moreover, when a locker becomes available, it does not start to be served immediately but only at the beginning of the next period. However, this does not violate the assumptions of the $M/G/k$ model because we can view the pick time as a discrete value; i.e., the service of the parcel ends at the end of the period when it was picked up and not at the moment that it was actually picked up.

There are several classical results on the effective approximation of the waiting time in $M/G/k$ based on the first two moments of the service time distribution. In our experiment, we adopted a classical formula known as Kingman's law of congestion; see, for example, Gans et al. (2003). For completeness of our presentation, we cite the formula here (using our notation)

$$D(C,\lambda,\eta) \approx \frac{1+CV(\eta)}{2} E\left(W_q^{M/M/k}\right)$$

where $CV(\eta)$ is the coefficient of variation of the service time distribution and is the expected waiting time in the queue in an $M/M/k$ queuing model with C servers, service rate $\frac{1}{E(\eta)}$ and arrival rate λ. The latter can be calculated using a closed-form formula known in the classical literature; see, for example, the textbook, Ross (2014), Chapter 8.3.3. Important properties such as monotonicity and convexity of the waiting time were also studied in the literature, e.g., Dyer and Proll (1977) and Harel and Zipkin (1987), and these properties are preserved in the approximated version of $D(C,\lambda,\eta)$ since it is a simple linear transformation of $E\left(W_q^{M/M/k}\right)$.

We note that the SP with postponements model can be viewed as a discrete version of the $M/M/k$ queuing model where the number of servers is C, the arrival rate is λ, and the service rate is p. Recall that in the $M/M/k$ model, customers arrive according to the system according to a Poisson process, wait in an unlimited queue and leave after service with exponentially distributed time. In our model, arrival occurs in every discrete period but is asymptotically equivalent to the Poisson arrival process since each batch consists of a number of parcels that follows a Poisson distribution. Moreover, the number of periods between the arrival of a parcel at an SP and its pickup (equivalent to the service time in the queuing model) is distributed geometrically, again with the same asymptotic behavior as the exponential service time in the $M/M/k$ model.

Finally, we note that both the rejection and the postponement functions are used merely as a surrogate for the actual unknown future implications of the overflow on the operational cost and service quality during the (ex ante) design process of the SP network. At the (ex post) operational level, the resolution of overflow is likely to be more complex. In particular, a policy that mixes rejection and postponement is very sound. The design of such a policy is an interesting topic for a future study focused on the operational aspects of parcel delivery networks.

4. Problem definition and formulation: the SP location and capacity problem

The SP location and capacity network is defined as follows: given finite sets of candidate locations for SPs and possible capacities, select a subset of locations where SPs should be open and determine their capacity. The objective is to minimize the present value of the total setup cost of the SPs and the cost due to capacity shortage.

The model is based on the following assumptions:

1. The service area is represented by sets of discrete demand points and candidate locations for SPs.
2. Before the system begins to operate, the company can set up an SP and determine its capacity at each one of the candidate SP locations. The capacity is selected from a closed list of options.
3. A setup cost is associated with each candidate SP location and each possible capacity.
4. Each demand point represents a set of recipients.
5. An SP must be opened within a predefined walking time (radius) from each demand point.
6. The number of parcels sent to each demand point (via the closest SP) in each period (e.g., every day) is a Poisson random variable with a rate parameter that characterizes the demand point.
7. Parcels that are destined to a demand point must be placed in the closest open SP if there is available capacity. The recipients pick up their parcels after a random number of periods distributed geometrically with a known parameter (referred to as the pickup probability parameter). Only after the parcel is picked up is the capacity in the SP released.
8. Parcels sent to demand points for which the closet SP has no available capacity are considered rejected. These parcels are assumed to be sent to their recipients in a more expensive manner. The delivery company is penalized for each rejection. Alternatively, overflow parcels are kept in the regional depot and delivered to their SP in one of the next replenishment cycles, as soon as a locker in this SP becomes available. Such parcels are considered postponed, and we assume that there is a penalty cost associated with each period of postponement for each postponed parcel.
9. The company selects, in advance (ex ante), a set of locations for SPs and SP capacities so as to minimize the total setup cost and the total expected rejection cost.

We note that with the analysis of a single SP presented in Section 3, we can calculate the expected number of rejections for each SP configuration. Indeed, each SP faces an arrival rate that is the sum of the arrival rates of all its closest demand points.

Next, we present a MILP formulation of the *SP capacity and location problem* starting with our notation. We refer to this model as a piecewise linear approximation (PWLA) model since it is a deterministic model that considers the expectation of the stochastic cost components in its objective function. The model is sufficiently light to enable solving problem instances of realistic size using an off-the-shelf solver, as demonstrated in Section 6. We first introduce the parameters of the model and its decision variables.

Parameters

F Set of candidate locations for SPs.
D Set of demand points.

t_{df} Walking time from demand point $d \in D$ to SP candidate location $f \in F$ or any other measure of inconvenience for a recipient located at demand point d to pick up a parcel from the candidate SP location f.

r The maximum walking time allowed from a demand point to an SP that serves it.

$C_1, \dots, C_{\bar{s}}$ Possible capacities of SPs.

Λ_{ks} Discretized levels of supply rate at SPs with various capacities C_s, where $k = 1, \dots, \bar{k}$. Recall that the service level functions map the arrival rates of parcels to the SPs into the corresponding measure (expected number of rejections or expected delay). For a given SP capacity C_s, the constants $\Lambda_{1,s}, \dots, \Lambda_{\bar{k},s}$ represent the x-axis values of the boundaries of the intervals in the piecewise linear approximation of these functions.

μ_d Expected periodic demand from demand point d (the actual demand during each period is assumed to follow a Poisson distribution).

h_{fs} The cost of opening (and operating during the planning horizon) an SP with capacity C_s at potential location $f \in F$. This cost is amortized to the construction time of the network.

α The amortized cost of a single parcel rejection per period during the planning horizon.

η The probability distribution of the number of periods until a parcel is picked from its locker.

Decision variables

y_{fs} A binary variable that equals "1" if an SP with capacity C_s is opened at location $f \in F$.

x_{df} The expected periodic number of parcels that the model plans to send to demand point $d \in D$ from an SP at location $f \in F : t_{df} < r$. Notably, the interpretation of x_{df} is different from the interpretation in classical facility location models. Here, it only represents a tentative division of the supply among the relevant location, while the actually expected supply may be smaller due to rejections.

z_{fks} Variables used to express the arrival rate of the SP at location f with capacity s. The arrival rate of the SP in location f is obtained as $\sum_{k=1,s=1}^{\bar{k},\bar{s}} z_{fks}\Lambda_{ks}$. This is stipulated by constraint (12) below. We use the fact that the rejection function is convex in the arrival rate to ensure that z_{fks} is assigned with a convex combination of Λ_{ks} points to obtain a linear approximation of the service level function.

The SP location and capacity problem with rejections can now be formulated as MILP (8)–(16). The modifications needed to accommodate the postponements in the objective function are discussed below.

$$\min \sum_{s=1}^{\bar{s}} \sum_{f \in F} \left[h_{fs} y_{fs} + \alpha \sum_{k=1}^{\bar{k}} z_{fks} R(C_s, \Lambda_{ks}, \eta) \right] \tag{8}$$

subject to

$$\sum_{s=1}^{\bar{s}} y_{fs} \le 1 \qquad \forall f \in F \tag{9}$$

$$\sum_{f \in F : t_{df} < r} x_{df} = \mu_d \qquad \forall d \in D \tag{10}$$

$$\sum_{f' \in F : t_{df'} > t_{df}} x_{df'} \le \mu_d \left(1 - \sum_{s=1}^{\bar{s}} y_{fs} \right) \qquad \forall d \in D, f \in F : t_{df} < r \tag{11}$$

$$\sum_{k=1,s=1}^{\bar{k},\bar{s}} z_{fks}\Lambda_{ks} = \sum_{d \in D : t_{df} < r} x_{df} \qquad \forall f \in F \tag{12}$$

$$\sum_{k=1}^{\bar{k}} z_{fks} = y_{fs} \qquad \forall f \in F, s = 1, \dots, \bar{s} \tag{13}$$

$$x_{df} \ge 0 \qquad \forall f \in F, d \in D : t_{df} < r \tag{14}$$

$$y_{fs} \in \{0, 1\} \qquad \forall f \in F, s = 1, \dots, \bar{s} \tag{15}$$

$$z_{fsk} \ge 0 \qquad \forall f \in F, s = 1, \dots, \bar{s}, k = 1, \dots, \bar{k} \tag{16}$$

The objective function (8) minimizes the total setup cost of the facilities (SPs) and the expected number of parcel rejections that will result from the selected configuration of SPs. The second component of the objective function is a piecewise linear approximation of the expected rejections based on an extensive preprocessing phase, where the values of $R(C_s, \lambda_k, p)$ are calculated for $\bar{s}\times\bar{k}$ different combinations of C_s and Λ_k. We can calculate the rejection function very quickly using the DTMC model, developed in Section 3.1, which is handy here. The formulation of (8) is based on the convexity of the rejection function (see Proposition 1) to approximate the function without the need for integer variables or specially ordered sets that could result in a much more intricate model.

Constraint (9) stipulates that at most one size is selected for each candidate SP location. If no size is selected, i.e., $\sum_{s=1}^{\bar{s}} y_{fs} = 0$, no SP is opened at location f. Eq. (10) assures that the total supply rate to demand point d from all the SPs that cover it is equal to the demand of the point. Constraint (11) ensures that a demand point can be served only by its closest SP. This requirement is enforced by requiring that if an SP f' is opened at a closer location than f, the sum of all the $x_{df''}$ values corresponding to SPs at a distance greater than t_{df} must be zero.

Constraints (12) and (13) together assign the correct value to the z variables by requiring that for each SP location f, the values of the z_{fks} variables constitute a convex combination of the Λ_k breakpoints that is equal to the supply rate of the SP at f. Constraint (13) further stipulates that (a) no supply can be made from an SP that was not open and (b) only z_{fsk} variables that correspond to the capacity of the SP C_s that was open in location f can be nonzero. The fact that the values of z_{fsk} for each (f, s) tuple are multiplied by a convex function in the objective function assures that in the optimal solution, only two variables, z_{fsk} and $z_{f,s,k+1}$, can be nonzero. The convex combination $z_{fsk}\Lambda_k + z_{f,s,k+1}\Lambda_{k+1}$ represents the actual supply rate of SP f, and the convex combination $z_{fsk}R(C_s, \Lambda_k, p) + z_{f,s,k+1}R(C_s, \Lambda_{k+1}, p)$ is a piecewise linear approximation of the number of rejections in the SP. Constraints (14)–(16) define the domains of the decision variables.

We note that the combination of Constraints (10), (12), and (13) stipulates that at least one SP is open within a radius of r from each demand point. This constraint may be too restrictive under some conditions because it may be worth not covering some demand points and bearing the cost of their demand as rejections. However, we made a modeling decision not to allow such esoteric solutions because we assume that the company wishes to cover the entire designated service area.

The values of the z_{fsk} variable can be nonzero only if an SP of capacity C_s is opened at location f, i.e., if $y_{fs} = 1$. In the latter case, the sum of the $z_{fs1}, \dots, z_{fs\bar{k}}$ is 1, meaning that the expression in the second component of the objective function is a convex combination of two points on the curve of the service-level function. The argument (independent variable) of this function is the arrival rate of parcels to the SP, and the dependent variable is the service level. The piecewise linear approximation of the service-level function is obtained as a combination of Eqs. (8), (10), (12), and (13).

The model is adapted to accommodate postponement in the objective function by reformulating the objective function as follows

$$\min \sum_{s=1}^{\bar{s}} \sum_{f \in F} \left[h_{fs} y_{fs} + \alpha \sum_{k=1}^{\bar{k}} z_{fks} D(C_s, \Lambda_{ks}, \eta) \right] \tag{17}$$

Here, α represents the amortized penalty cost of one day of postponement instead of the rejection cost. Moreover, the values selected for Λ_{ks} are likely to differ from those in the rejection, and in particular, they should all be lower than the service rate $\frac{C}{E(\eta)}$ since higher values are infeasible for a system with postponements. In addition to avoiding designing a system with insufficient capacity to support stable operation, we add an additional constraint

$$\sum_{d \in D: t_{df} < r} x_{df} \leq \frac{1-\epsilon}{E(\eta)} \sum_{s=1}^{\bar{s}} y_{fs} C_s \qquad \forall f \in F. \tag{18}$$

where ϵ is a small positive number (say 0.01). This constraint ensures that the load on an open SP is safely less than 100% to guarantee stability. The left-hand side of constraint (18) represents the periodic arrival rate of parcels at the SP opened in location f, while the right-hand side represents a value close to its feasible ability to accept parcels without "exploding" its queue. The exact value of ϵ is not particularly important because, in an optimal solution, none of the SPs are likely to operate on the verge of stability when the number of postponements is very high. The largest distribution point for each value of C_s, $\Lambda_{\bar{k}s}$ must coincide with $\frac{1-\epsilon}{E(\eta)} C_s$.

Next, we discuss a few possible simple extensions and applications of MILP (8)–(16):

1. In a typical situation, the delivery company does not deploy a new network from scratch. Instead, it extends an existing one by opening several additional SPs at a time and modifying the capacity of existing SPs to meet the gradually growing demand. The above model can be applied in such situations by determining the setup cost of each existing SP, as follows. If the current capacity is C_s, then $h_{fs} = 0$, for $s' > s$ (resp. $s' < s$), $h_{fs'}$ is set to the cost of upgrading (resp., downgrading) the capacity of the SP to $C_{s'}$. In the case of downgrading, $f_{fs'}$ may be negative, e.g., due to saving on rent payments. Moreover, if changing the capacity of some of the SPs in the network is not possible at the current time, their corresponding y_{fs} variables can be fixed.
2. In many facility location models, one component of the objective function represents the transportation cost from opened facilities to the customers. In the SP capacity and location model, the planner is indifferent to this cost since the parcel recipients are responsible for picking up the parcels, as long as they are available at the closest SP. However, if the designer of the network wishes to incorporate the recipients' convenience into her considerations, the following term can be added to the objective function (8):

   $$\sum_{f \in F, d \in D: t_{df} \leq r} t_{df} x_{df}.$$

   Selecting higher values for t_{df} will result in solutions with denser coverage of the service area, probably with smaller capacity. However, a similar outcome could be obtained by reducing the service radius parameter r. In this study, we made a modeling decision to account for the recipient's convenience using a constraint on the service radius, but other approaches are also legitimate and applicable.
3. The model proposed here can also be used for an SP network based on attended facilities, such as grocery stores and kiosks. For example, the delivery company may ask businesses in the candidate locations to place one or more bids for providing the service: each bid specifies the cost and offered capacity. The company can then use the model to select winning bids.

5. Alternative models

Our solution approach cannot be compared directly to the methods presented in previous SP network design studies since previous authors assumed that the demand is deterministic. Their focus was to minimize the setup cost and the distance recipients must walk to the SPs, while deterministically covering all the demand. Therefore, we adopted two existing modeling approaches from the general location theory literature to benchmark our idea. The first approach is to use a model that takes the demand of each point as deterministic and equal to its expectation. The model allows rejection in cases where it is not worthwhile or infeasible to design sufficient capacity to satisfy the entire expected demand. The second approach models the stochasticity of the supply process by optimizing over a large set of demand scenarios.

We are unaware of previous facility location models equivalent to the SP location and capacity problem presented in this paper. However, the deterministic and scenario-based models presented in this Section are in line with previous literature, such as Deutsch and Golany (2018), Rabe et al. (2021) and Pan et al. (2021), and we do not claim any novelty with regard to them. Our aim in this Section is to adapt known methods for the problem to facilitate benchmarking of the PWL model. The results of this comparison are presented in Section 6.

5.1. A deterministic model

An alternative to the PWL SP location and capacity model presented here in the spirit of the classical deterministic location theory literature is a model that opens facilities at locations and sizes that are sufficient to serve the entire expected demand. The stochasticity in such models is not addressed directly. Instead, the plan is based on the expected demand, and at best, some *safety margins* are added to account for the fact that the actual demand varies from day to day. Next, we present such a coverage model using the same parameters and decision variables introduced above for MILP (8)–(16) and add the following:

Parameters

β Safety margin on the demand to account for the stochasticity. It is typically slightly larger than one, and if no safety margins are used, $\beta = 1$

Decision variables

q_f The number of parcels rejected by SP $f \in F$.

$$\min \sum_{s=1, f\in F}^{\bar{s}} h_{fs} y_{fs} + \alpha \sum_{f\in F} q_f \tag{19}$$

subject to

$$\sum_{d\in D: t_{df}<r} x_{df} \le \sum_{s=1}^{\bar{s}} C_s p y_{fs} + q_f \qquad \forall f \in F \tag{20}$$

$$\sum_{f\in F: t_{df}<r} \sum_{s=1}^{\bar{s}} y_{fs} \ge 1 \qquad \forall d \in D \tag{21}$$

$$\sum_{f\in F: t_{df}<r} x_{df} = \beta \mu_d \qquad \forall d \in D \tag{22}$$

$$\sum_{s=1}^{\bar{s}} y_{fs} \le 1 \qquad \forall f \in F \tag{23}$$

$$\sum_{f'\in F: t_{df'}\ge t_{df}} x_{df'} \le \mu_d \left(1 - \sum_{s=1}^{\bar{s}} y_{f's}\right) \qquad \forall d \in D, f \in F : t_{df} < r \tag{24}$$

$$x_{df} \ge 0 \qquad \forall f \in F, d \in D_f \tag{25}$$

$$y_{fs} \in \{0,1\} \qquad \forall f \in F, s = 1, \dots, \bar{s} \tag{26}$$

MILP model (19)–(26) is a limited version of MILP (8)–(16). The objective function (19) consists of the SP setup cost, as in our PWL model, and the total number of planned rejections weighted by the rejection cost α. Constraint (20) stipulates that a facility with sufficient capacity is opened at each candidate location from which parcels are planned to be supplied to demand points. In particular, this constraint eliminates supply from facilities that are not opened. Note that the available capacity of an SP is assumed by this model to be $C_s p$, since, at the steady state, a proportion of approximately p of the lockers become available in every period. The decision variable q_f is added to the right-hand side to represent the number of parcels supplied to the recipients, not via the SP, i.e., rejected parcels. Constraint (21) assures that each demand point is covered by at least one opened SP in the allowed radius.

Constraint (22) stipulates that all the parcels of each demand point are supplied to it. The right-hand side may be inflated by a factor of β to create some safety margin. Indeed, values of β that are greater than one force the model to create a proportionally greater capacity that can absorb some of the supply and demand stochasticity. The rest of the constraints are exactly the same as those in MILP (8)–(16).

We note that the deterministic model assumes that the number of parcels delivered to and picked up by the recipients of each demand point is known and fixed. Consequently, the model opens SPs with sufficient capacity to serve each one of them without rejection or delay. The stochasticity of the demand can be considered indirectly by increasing the safety margin β, but there is now a special distinct treatment for each of the two overflow handling approaches. If one wishes to adjust the value of β, the rejection and postponement functions introduced in Section 3 can be handy.

While we cannot attribute the MILP model (19)–(26) to a particular author, it is written in the spirit of many previous location models, including ones that are presented in textbooks; see, for example, the first chapters of Laporte et al. (2019). Hereafter, we refer to it as the *deterministic model*.

5.2. A scenario-based stochastic model

The SP location and capacity problem does not lend itself naturally to a scenario-based approach because there is uncertainty regarding the demand of each point and the pickup times of parcels from the SPs, where the latter is revealed to the delivery company only a few periods later if the parcel is not rejected or postponed. A scenario cannot include the number of parcels collected from each SP each day because this number depends on the number of parcels waiting at the SP, which is affected by the solution of the location and capacity problem. Therefore, we assume that the number of parcels that an SP with capacity C can accept each day is Cp, as we did in the deterministic model. This scenario includes information about only the daily demand of each demand point.

We note that Mancini et al. (2023) took a different approach toward modeling the stochasticity of locker availability and assume that the number of parcels collected by recipients in each zone is also encoded in the scenarios and is not dependent on the number of rejections.

In this section, we use the same notation as that used in the previous models with the following additions and modifications.

Parameters

$\mathcal{K}$ The set of possible demand scenarios.
μ_{dk} The demand at point d in scenario $k \in \mathcal{K}$
π_k The probability of scenario $k \in \mathcal{K}$.

Decision variables

x_{dfk} The number of parcels allocated from SP $f \in F$ to demand point $d \in F_d$ in scenario $k \in \mathcal{K}$.
q_{fk} The number of parcels rejected by SP $f \in F$ in scenario $k \in \mathcal{K}$.

$$\min \sum_{s=1, f\in F}^{\bar{s}} h_{fs} y_{fs} + \alpha \sum_{k\in\mathcal{K}, f\in F} \pi_k q_{fk} \tag{27}$$

subject to

$$\sum_{d\in D_f} x_{dfk} \le \sum_{s=1}^{\bar{s}} C_s p y_{fs} + q_{fk} \qquad \forall f \in F, k \in \mathcal{K} \tag{28}$$

$$\sum_{f\in F: t_{df}<r} \sum_{s=1}^{\bar{s}} y_{fs} \ge 1 \tag{29}$$

$$\sum_{s=1}^{\bar{s}} y_{fs} \le 1 \qquad \forall f \in F \tag{30}$$

$$\sum_{f\in F_d} x_{dfk} = \mu_{dk} \qquad \forall d \in D, k \in \mathcal{K} \tag{31}$$

$$\sum_{k\in\mathcal{K}} \sum_{f'\in F: t_{df''}\ge t_{df}} x_{df'k} \le \sum_{k\in\mathcal{K}} \mu_{dk} \left(1 - \sum_{s=1}^{\bar{s}} y_{fs}\right) \qquad \forall d \in D, f \in F : t_{df} < r \tag{32}$$

$$x_{fdk} \ge 0 \qquad \forall f \in F, d \in D_f, k \in \mathcal{K} \tag{33}$$

$$y_{fs} \in \{0,1\} \qquad \forall f \in F, s = 1, \dots, \bar{s} \tag{34}$$

The objective function (27) is the weighted sum of the setup cost and the expected number of rejections over all scenarios. Constraint (28) stipulates that all the parcels planned to be sent via an SP in each scenario either fit into the expected available

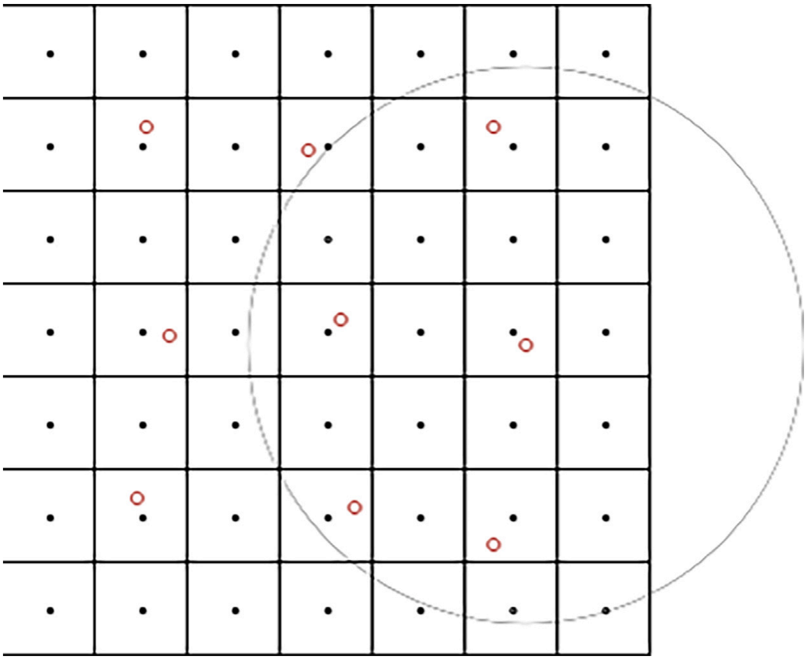

Fig. 3. Synthetic geography.

capacity of the SP or are considered rejections. Constraint (29) assures that each demand point is in the service radius of at least one SP. Constraint (30) which is the same as (9) assures that at most one SP is opened at each candidate location. Constraint (31) stipulates that all the demand is met (or accounted for by rejections) under each scenario. Constraint (32) stipulates that parcels are sent via the closest SP to their recipients. It works in the same way as the (11) in the PWL model. Constraints (33) and (34) define the domains of the decision variables.

Note that the scenario-based model parcel overflow is assumed to be handled by rejections and that the number of parcels picked up each day is assumed to be deterministic. In this setting, the system is reborn each day. Postponements add another dimension of interaction between periods. We cannot conceive a straightforward extension of the scenario-based model to handle overflow with postponement rather than rejections.

6. Numerical experiment

Our numerical study has two goals: to demonstrate the merits of our PWL model compared with the two alternatives presented in Section 5 and to explore the computational limitations of our approach.

To achieve the goals of our numerical experiment, we coded a Python script that generates pseudo-random instances of the problems, including random scenarios for the scenario-based model and a preprocessing step to create a piecewise linear approximation of the rejection function for our PWL model. The three MILP models were implemented in OPL, and the generated instances were solved using IBM Cplex v20.1. Our test machine was an AMD Ryzen 9 5950X workstation with 128 GB of RAM, running Linux Ubuntu 20.04.

All our code and detailed results are available to download from GitHub. The code is well documented and can be readily used to replicate the experiments described in this section and in the appendixes, as well as to generate our input data or to generate different instances with other characteristics and dimensions.

For our experiment, we created a service area with synthetic geography composed of 200 × 200 m pixels with demand points located in the middle of each pixel. The candidate SP locations were placed randomly in all the pixels with even row and column numbers to create some irregularity in the geographic structure that is characteristic of any urban area. In Fig. 3, we provide a toy example of such a geography in a 1.4×1.4 km square. The solid dots represent the demand points, and the red circles represent candidate SP locations. The large dotted circle demonstrates the service area covered by an SP candidate location when the service radius is 600 m.

We focus on the models that handle overflow by rejection rather than postponement, as this alternative is possible in both benchmark models, making comparison possible. However, we believe that similar results (from a computational perspective) could be achieved for the postponement case using our PWL model.

The input of our instance generation script is the dimensions of the service area square (in units of 200 × 200 m pixels) from which the number of candidate SP locations (m) and the number of demand points (n) are derived; the service radius for each facility

(r); the periodic pickup probability (p); the weight of rejections in the objective function (α); a vector C of possible SP capacities; and an integer number to be used as a random seed.

The demand rate, μ_d, of a point is selected randomly from a uniform distribution $U(0.5, 10)$. Next, the setup cost of the SP at each candidate location i and for each possible capacity C_k is generated as follows:

$$h_{fs} = (\phi + \omega C_s) \cdot \zeta_f$$

where ϕ represents the fixed cost component and ω is the variable component that is affected by the capacity of the APL. ζ_f is a random variable representing the real-estate cost of a particular location: we draw it from a normal distribution, $\zeta_i \sim N(1, \frac{1}{6})$.

The script then creates a piecewise linear approximation of the rejection function for each of the capacities in C using 12 values of ρ, as explained in Appendix C, as well as the required number of demand scenarios $\mu_{dk} \sim Poisson(\mu_k)$. For the scenario-based model, the script accepts the number of demand scenarios to generate. All the generated scenarios assume the same probability.

To obtain nontrivial results (i.e., open SPs at all candidate locations), we paid special attention to selecting values for the parameters C, p relative to the other parameters described above. Because there are approximately four times more demand points than candidate SP locations and the mean demand of each point is approximately 5, we created possible SP capacities of $C = (30, 60, 90)$ lockers and set $p = 0.5$. Recall that the effective periodic capacity of an SP is its nominal capacity multiplied by the periodic pickup probability. Allowing SPs with up to 90 lockers means that the maximum possible capacity if the largest possible APL capacity is installed at each of the candidate SP locations is more than twice the periodic demand, which leaves considerable freedom for the model in selecting SP locations and capacities. We also believe that APLs with 30–90 lockers represent the common practice of the delivery industry and result in SPs that can comfortably fit in front of grocery and convenience stores. $p = 0.5$ represents a mean time-to-pickup of two periods, which appears reasonable given that a typical replenishment cycle is a day and many delivery companies require that recipients pick up parcels within several days.

The cost parameters were also selected carefully. We set $\phi = 10, \omega = \frac{1}{6}$ in units of \$1000. We conceive a planning horizon of ten years, which appears to be a reasonable lifetime for the APL hardware and for the length of contracts with the business that will host the SPs. Note that for a 30 locker APL, the above parameters imply an average cost of \$15,000. The market price of an APL with 30 locker boxes and a check-out terminal is in the range of \$2000–5000. To this sum, one should add installation, maintenance, energy, communication, and removal costs, as well as rent to be paid over ten years to the property owner on which the APL is placed. Our assessment of the total lifecycle present value cost of \$15,000 for a 30-locker APL may not be accurate, and it depends on many local factors; however, we believe that it is not far from the true cost.

Next, we set the present value of the cost of one rejection per day over the conjectured ten-year APL lifecycle to $\alpha = 5$ or $\alpha = 20$. Assuming a 3650 day lifecycle of an APL and an annual interest rate of 4%, these values of α imply a single rejection present cost in the range of \$1.59 to \$6.34. In our view, these sums soundly span the loss due to the need to ship the item directly to the recipient when there is a shortage of lockers.

Finally, to create instances with various opportunities for serving each SP, we tested two values of service radius, $r \in \{401, 601\}$ m, where each SP covers approximately 12 and 25 demand points (respectively) on average. Such radii ensure considerable flexibility in locating SPs.

As a preliminary test, we generated 10 different small geographies with 5×5 = 25 SP candidate locations and 21×21 = 121 demand points with random demand rates at each demand point. Each such geography was combined with the four possible combinations of $r \in \{401, 601\}$ and $\alpha = \{5, 20\}$ to create 40 problem instances in total.

In our experiment, we tried to solve each of these 40 problem instances with the three solution methods presented in Sections 4 and 5 using Cplex with a time limit of one hour. For the scenario-based formulation, we generated 50 scenarios for each of the problem instances using the true distribution assumed by our PWL model.

Since the estimates of the expected number of rejections made by the deterministic and scenario-based models are not accurate, we reevaluate the values of their solutions in a post-processing phase using our piecewise linear approximation of the rejection function.

Subject to the one-hour time limit, we could solve all the instances of the deterministic and PWL models to optimality. The average solution time for the deterministic model was 1.3 s (the maximum over all 40 instances was 3.3 s). Similarly, the average (resp., maximal) solution time of our PWL model was 1.0 (resp., 3.3) seconds. The situation was different for the scenario-based model. The instances with $r = 401$ could be solved in 57.9 s on average (212.8 s maximum). However, none of the instances with the larger service radius, $r = 601$, could be solved within the one-hour time limit. Cplex produced feasible solutions for all these instances with an average optimality gap of 10.1% (14.7% max).

More interesting is the comparison between the objective function values of the three models. In some sense, it is not a fair comparison because only the PWL model optimized the “true” function, while the other two models used a surrogate function. However, we can see how good this surrogate is when plugging the obtained solution into the true objective function. In Fig. 4, we present the average loss in terms of the total cost (i.e., the setup costs of the SP plus the expected rejection cost) of the two alternative models compared with our PWL model.

It is apparent from Fig. 4 that the deterministic model performs almost as well as our PWL when the cost of each rejection is small ($\alpha = 5$); the PWL model yields a substantial improvement only when the rejection cost is high. This result is not surprising because as long as the rejection cost constitutes only a minor portion of the total cost, a more accurate account of the rejections in the objective function does not yield great potential for improvement.

More surprising are the poor results obtained from the scenario-based model. It may be the case that considering 50 scenarios is not sufficient to capture the stochastic nature of the rejection phenomenon. Considering the randomly generated scenarios only

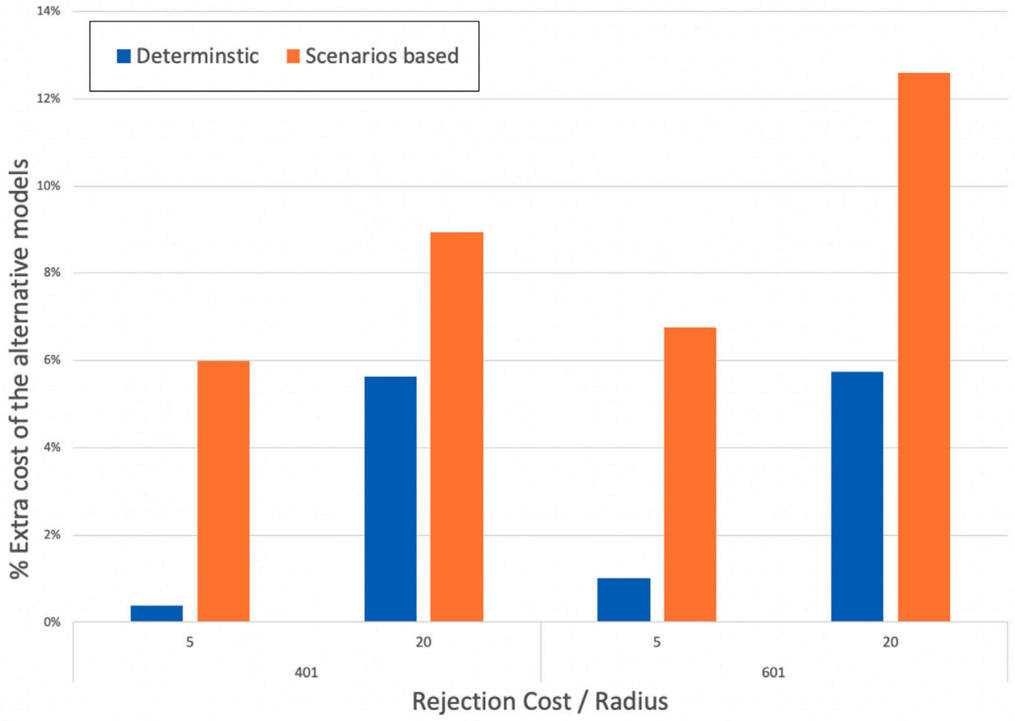

Fig. 4. Comparison with the two alternative models, small instances with $m = 25, n = 121$.

introduces noise to the model that diverts it from placing SPs at the correct locations and with the correct capacities. Theoretically, a much larger number of scenarios should result in a better approximation of rejection and, hence, better solutions. However, the scenario-based model is already very hard to solve, even with only 50 scenarios and instances of small dimensions.

Looking more closely into the results (available as an electronic companion to this paper on GitHub), we observe that the deterministic model consistently underestimates rejection and thus sets too few and too small SPs. The scenario-based model overestimates rejection in some instances and underestimates it in others, without any consistent pattern. This observation supports our conjecture that a much larger set of scenarios is needed to obtain a sufficiently accurate representation of the stochastic demand.

The observation that the deterministic model underestimates rejection calls for the usage of a safety margin, as we suggested in Section 5.1. However, the solution times of the deterministic and PWL models are similar, and the solution of PWL is at least as good as that obtained from the deterministic model. Therefore, it is always better to use the PWL model instead of trying to fine-tune the safety margin parameter of the deterministic model for each particular instance. Moreover, to accomplish such safety margin parameter fine-tuning, one must possess a tool to evaluate the expected number of rejections at each SP. However, compared to using such a tool, creating the input needed for the PWL model is a simple alternative.

Based on our experiment described above, we ruled out the scenario-based model as a practical solution approach to the SP location and capacity problem. Indeed, with the ILP formulation and solution method at hand, it is impossible to solve instances of interesting size with the scenario-based model. Moreover, the values of the solutions of the small instances that we could solve with this model are not competitive with those of the deterministic and PWL models. Therefore, we are not encouraged to devise better solution methods for the scenario-based model.

We further compared the deterministic and PWL models using larger synthetic instances generated as described above. Our results, reported in detail in Appendix D, support our observations discussed above.

7. Case study

The case study presented in this section is adapted from Kahr (2022), who focused on different aspects of the design of the SP network. We created problem instances based on the three largest Austrian cities: Linz, Graz, and Vienna. For these three cities, Kahr (2022) has provided census data updated to 2020. The geographical granularity of these data sets is $100\ \mathrm{m} \times 100\ \mathrm{m}$ squares. We use the center of each such square as a demand point, with demand for parcel delivery proportional to the population registered in it. Following Kahr (2022), we selected the set of candidate SP locations as the locations of all the bus stops, tram stops, train stations, markets, supermarkets, post offices, and fuel stations in each city. In addition, some centers of census squares that are more than

Table 1
Summary statistics of the use case instances.

City	Population	Number of DPs	Candidate locations	Locations in radius
Linz	206,724	3,351	658	3.55
Graz	291,245	6,777	1459	4.46
Vienna	1,911,274	17,701	4260	22.0

Table 2
Results of the use case instances.

Instance		PWL model				Cover model			
City	Demand	SP Cost	Rej. cost	Total cost	Opt. Gap	SP Cost	Rej. cost	Total cost	Opt. Gap
Linz	2%	6008	158	6166	0.65%	5825	616	6441	0.01%
Linz	4%	7481	291	7772	1.25%	7088	1227	8315	0.01%
Graz	2%	9897	261	10,158	2.38%	9600	976	10,576	0.26%
Graz	4%	12,042	382	12,424	2.49%	11,342	1934	13,276	0.96%
Vienna	2%	32,827	1019	33,846	12.51%	29,569	6408	35,977	4.14%
Vienna	4%	47,240	7290	54,530	7.60%	43,660	15,264	58,924	4.94%

300 m from any of the above locations were also considered as candidate locations in order not to degenerate the problem and leave some remote areas of the cities unserved.

We assumed four possible sizes of SPs: 30, 60, 100, and 150. The setup cost for each possible size was randomized as described in Section 6. That is, the mean setup cost of an SP with 30, 60, and 100, 150 lockers was set to 15, 20, 33.33, and 45 respectively. For these mean values, we added some white noise, as described in 6. The time-to-pickup was assumed to follow a geometric distribution with $p = 0.5$, meaning that a parcel remains in a locker for an average of two periods. We considered two levels of expected daily demand for parcel deliveries. In the low-demand setting, the expected demand was 2% of the population of the demand point; in the high-demand setting, it was 4%.

The service radius for SPs was set to 300 m (as in Kahr (2022)), and the rejection cost was set to $\alpha = 10$, which means that one unit of average daily rejection over the entire lifetime of the system is equivalent to approximately half the mean cost of an SP with 60 lockers. Assuming all the prices are given in $1000 s and the operational lifetime of an APL is approximately 10 years, this amounts to a cost of approximately $3.18 per rejected parcel.

In Table 1, we provide some summary statistics on the dimension of the problem instances of the three cities. The table presents the total population (obtained from the census data). The number of demand points (DPs), which is the number of $100\text{ m} \times 100\text{ m}$ cells with a nonzero population in the municipal area, the total number of candidate SP locations in our data set, and the average number of candidate SP locations within 300 m radius from each SP.

As described above, we created two instances based on the data of each city. One with low demand (the expected number of parcels per day is 2% of the population) and one with an expected high demand level (4% of the population). We solved these six instances with both the PWL and the covering model using Cplex in the same environment described in Section 6. We allocated 3 h for the solution of each instance. Our testing environment is described in Section 6.

In Table 2, we report the results of this experiment. For each instance and each solution method, we report the total cost of the SPs in the solution, the expected cost of rejections, and the total cost, SP Cost + $\alpha\cdot$ number of rejections. In addition, we report the optimality gap of the solution obtained by the solver when the solution process was stopped after three hours. The optimality tolerance was set to 0.01%, so the results with this value in the "Opt. Gap" columns of the cover model represent instances for which the solver stopped due to convergence.

It is apparent from the optimality gap columns in Table 2 that while large realistic instances of both models cannot be solved to optimality within a reasonable time budget, it is possible to obtain an approximate solution with an optimality gap of a few percentage points. The PWL model is solved with larger gaps. In Fig. 5, we present the breakdown of the solution values to their two components: setup costs of the SPs (in blue) and expected rejection cost (in orange). It is apparent from the table and the figure that the PWL model yields consistently lower costs (in all six instances compared). The advantage of the PWL is more prominent when the demand is higher. Moreover, the PWL model generally spends somewhat more than the cover model on the setup of SPs but saves significantly on rejection costs. These results are in line with the result observed with our synthetic instances presented in 6. A closer look at the results indicates that the PWL model opens SPs in approximately the same number of locations but with a higher average capacity.

We provide the input files and Python code needed to replicate this numerical experiment, along with a detailed solution, at GitHub repository. We also provide a graphical representation of the solutions for Linz, Graz, and Vienna in Google Maps (click the name of the city and select the corresponding layer to view the solution). Each Map contains four layers with different sets of SP locations for the two demand levels and based on the two models (PWL and Cover). Each SP is represented by a locker icon with a color code representing the capacity. The demand points are presented on the map as tiny black circles. We partitioned the demand points into several layers due to a technical limitation of Google Maps that allows up to 2000 locations in each layer. An example of a section of the Map of Linz with high demand is presented in Fig. 6.

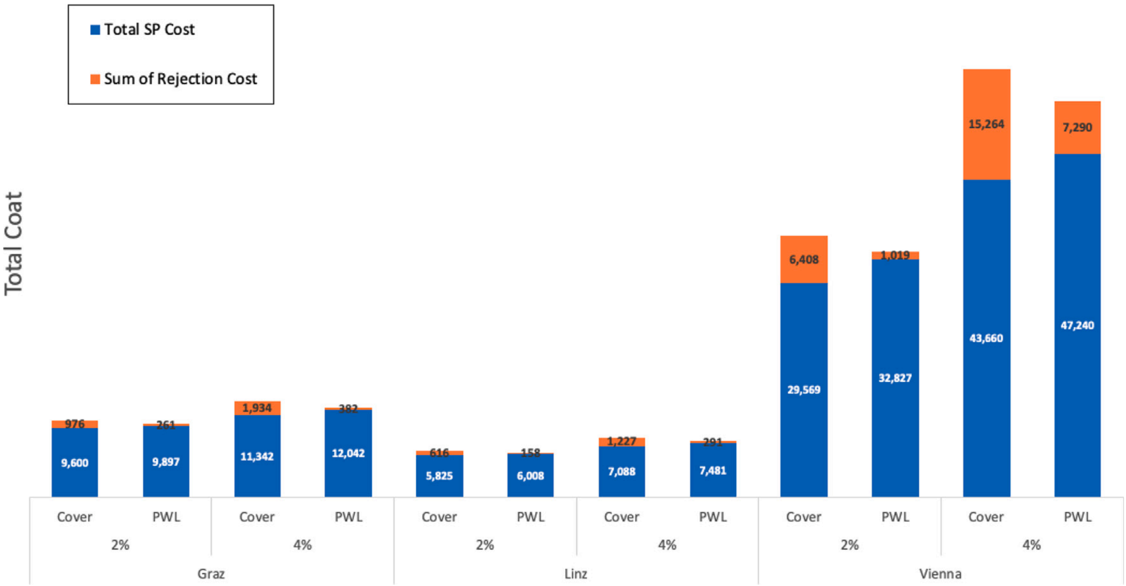

Fig. 5. A breakdown of the cost under the PWL and cover models for the three cities and two demand levels. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

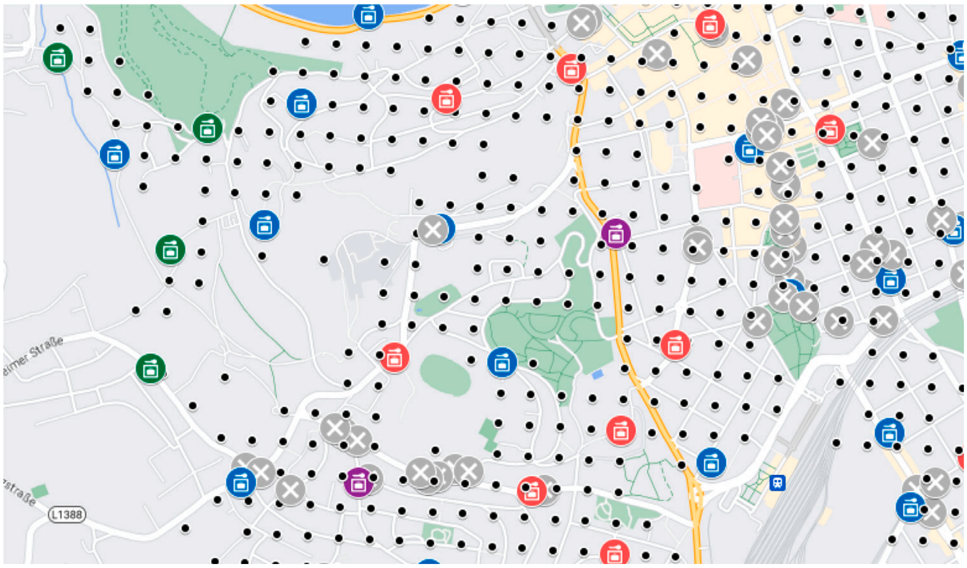

Fig. 6. A geographical representation of a section of the solution for the Linz use case with high demand (4% of the population). Black dots represent demand points. Green, blue, red, and purple locker icons are SPs with 30,60, 100, and 150 lockers, respectively. Gray circles with an "x" are unoccupied candidate SP locations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

8. Conclusions

This paper's main contribution is the introduction of a new piecewise linear approximation facility location and capacity model in the context of a network of SPs for parcel delivery. We took a bottom-up approach by first analyzing the stochastic phenomenon of each SP separately and then using the results to design the entire network optimally. The problem is formulated as a MILP model, and its solution, which can be obtained from a commercial solver in a reasonable time, significantly outperforms solutions obtained using alternative models.

However, our solution approach has a computational limitation. Problem instances with a large number of candidate SP locations and demand points cannot be solved to optimality with an off-the-shelf commercial solver. Therefore, developing more effective exact and heuristic methods to solve the model is an interesting objective for future research. Heuristic methods can use our characterization and approximation method of the rejection function and perhaps solve our mathematical model as a subroutine, e.g., in a large neighborhood search (LNS) algorithm.

Additional important directions for future research, in our view, include the following:

1. The formulation of richer models with APLs that consist of lockers of different sizes. This problem is intricate because the assignment of parcels to lockers must be solved simultaneously with the design problem. A large locker may accommodate small parcels but not the other way around. Such a model can integrate the results presented here and in Kahr (2022).
2. Our model assumes the ex ante allocation of parcels to SPs based on their demand points. Indeed, since each parcel is delivered to the SP closest to its demand point, the parcel allocation decisions are stipulated by the system's design. While we hope that this assumption is sufficient for the design phase, during the operation phase, some additional degrees of flexibility can be exercised. For example, if the closest SP is occupied, it may be possible to deliver parcels to SPs that are farther from the demand point or to postpone delivery by one or more periods. As demonstrated in Orenstein et al. (2019), such flexibility may improve the performance of a given system configuration and reduce the number of rejected parcels. This improvement is achieved thanks to the pooling effect between neighboring SPs and consecutive periods. Returning to the SP location and capacity problem, with ex post parcel allocation decisions, the optimal design of the SP network is likely to be different from the one proposed here. Neither our PWL model nor the set covering and scenario-based models presented in Section 5 can be naturally adapted to these settings. The problem is more intricate than previously studied stochastic facility location problems with ex post allocation decisions, e.g., Correia and Saldanha-da Gama (2019). In the SP network setting, the ex post parcel allocation decisions are also made under uncertainty regarding the pickup times of the parcels and the number of parcels that will be sent in future periods. Therefore, a straightforward implementation of bilevel stochastic programming is not adequate.
3. Allowing flexibility in parcel allocation also calls for new operational models. For example, suppose that the recipient can choose the delivery SP for her parcel (e.g., during the check-out process at the e-store or once the parcel arrives at the regional depot). An interesting question then is how to dynamically construct assortments of alternatives that the delivery company should display to recipients to economize the operation of an existing SP network. Such an assortment is built online based on the system's current state and on a discrete choice model that characterizes the recipients' preferences.

CRediT authorship contribution statement

Tal Raviv: Conceptualization, Methodology, Implementation, Writing – original draft, Writing – review & editing, Software, Visualization, Validation.

Acknowledgments

This research was supported by a grant from the Israel Science Foundation (ISF) [Grant no. 1460/20].

Appendix A. Evaluation of the rejection function approximation

One possibly justifiable criticism of our DTMC model is that it is based on the assumption that the number of parcels picked up by recipients in each period is a random variable whose distribution is determined by only the number of parcels in the SP. In other words, the parcels are memoryless, and the number of periods they spend in the SP follows a geometric distribution. In practice, it is likely that the probability of each parcel being picked up in a given period is affected by its seniority. For example, many parcels are likely picked up shortly after arriving at the SP. Moreover, most delivery companies limit the time recipients have to pick up their parcels and remove the parcels from the SP after this limit expires. Therefore, the parcels in their last period before expiry are removed from the SP before the next replenishment (either by the recipient or the company). Representing such a stochastic process using a DTMC would require the states to encode rich information about the seniority of the parcels in the SP. While this representation is theoretically feasible, it would lead to an explosion in the number of states and render the DTMC modeling approach impractical for the problem.

In this appendix, we report a numerical experiment demonstrating that while our DTMC model may be wrong (in its assumptions), it is still able to produce a very accurate approximation of the rejection function. To this end, we constructed a simulation model that removes the abovementioned simplifying assumption and compared the number of rejections to that predicted by the DTMC model.

In our simulation model, each parcel is picked up after Z periods, where Z is a random variable generated from a general empirical distribution with support $\{1, \dots, E\}$ and a probability function $P(Z = d) = p_d \quad \forall z = 1, \dots, E$. The parcel supply process is governed by the same Poisson random variable as in the DTMC, and parcels that arrive at the SP when it is full are rejected. The mean *time-to-pickup* under this model is $\sum_{d=1}^{E} d p_d$.

For comparison, we pair each simulation experiment, based on an empirical time-to-pickup distribution, with an exact solution of the DTMC model by selecting p to represent a time-to-pick-up distribution with the same mean value; that is,

$$p = \frac{1}{\sum_{d=1}^{E} d p_d}.$$

Table 3
The time-to-pickup distributions.

Empirical distribution					Geometric dist.	
Support	p_d	Mean	Var.	C.V	p	Var.
$\{1, 2, 3\}$	$(0.5, 0.2, 0.3)$	1.8	0.76	2.06	0.5556	1.44
$\{1, 2, \dots, 7\}$	$(0.4, 0.2, 0.1, 0.08, 0.05, 0.02, 0.15)$	2.84	4.63	1.32	0.3521	5.23

In our numerical experiment, we assumed two time-to-pickup distributions: one with $E = 3$ periods, where $p_d = (0.5, 0.2, 0.3)$, and one with $E = 7$ periods, where $p_d = (0.4, 0.2, 0.1, 0.08, 0.05, 0.02, 0.15)$. Details of the empirical distribution and the corresponding geometric distribution are provided in Table 3.

The capacity of the stations was selected such that $C \in \{20, 50, 100\}$, and λ was selected to obtain $\rho \in \{0.9, 0.95, 1, 1.05, 1.1\}$. That is, we simulated the system with $2 \times 3 \times 5 = 30$ configurations. We note that we have decided to focus on the above ρ value (around 100% utilization) because, in our preliminary experiments, we noticed that this is the range where the rejection function is the hardest to calculate. When the utilization is very low or very high, the fluid model (assuming $C \to \infty$) is a good approximation anyway. See, for example, Fig. 2.

We run the simulation under each configuration for one million periods. The first 100 periods were excluded as warmup time, and the expected number of rejections per period in the steady state was estimated with a 99% confidence interval based on the remaining periods.

Both the DTMC and simulation models were implemented in Python 3, with the NumPy package for fast linear algebra calculations and the SciPy package for probability mass and cumulative functions. The solution times reported below were obtained on an Apple MacBook Air notebook with an ARM M1 processor (2020).

The results of our experiment are reported in Fig. 8, where the expected number of rejections is plotted for each of the 30 configurations in 6 groups, one for each combination of E (the maximum period to pick up) and C (the SP capacity). The most important observation from the figure is that the prediction of the DTMC model conforms to the prediction of the simulation, even though the simulation is based on a more realistic general empirical distribution of the time-to-pickup. Indeed, in all 42 configurations, the rejection rate predicted by the DTMC model falls within the narrow 99% confidence interval estimated from the rather long simulation. The margin of error of the simulation was 0.008 on average (maximum 0.018).

The computation time of the simulation is approximately 12 s for each of the instances, and it is not affected by the configuration characteristic: it grows linearly with the number of simulated periods. The DTMC model takes approximately 0.08 s to compute for $C = 20$, 0.5 s for $C = 50$, and 2 s for $C = 100$.

Both our simulation and the DTMC models are viable alternatives for approximating the rejection function. Indeed, even when one must calculate hundreds of points on the rejection function for various values of p and λ as a preprocessing step for the PWL model, this can be done in a relatively short time. Moreover, we observe that the simplifying assumption that the pickup process is memoryless has a negligible effect on the rejection function, at least for the two empirical pickup distributions that we tested.

The rest of our observations from Fig. 7 are not surprising and conform to our intuition and analysis above. The rejection rate increases with λ when C and p are fixed. Furthermore, the pooling effect is strong in our SP model. Indeed, for a given load level, the rejection rate can be reduced drastically when the capacity of the SP and the supply rate are inflated simultaneously.

Notably, for the same load level ρ, the expected number of rejections is smaller when the pickup time is longer. This may appear surprising at first, but note that the arrival rate of the parcel is adjusted for the pickup probability per period in order to keep the same utilization level. Another peculiarity in Fig. 7 is that the plots do not appear convex. This is only because the gaps (of the ρ value) are not uniform on the horizontal axis.

Appendix B. Evaluation of the postponement function approximation

In this appendix, we report our computational experiment with several methods to estimate the postponement function $D(C, \lambda, \eta)$ introduced in Section 3.2. As with the rejection function, the motivation for finding a fast and accurate approximation method of this function stems from the fact that we need to calculate numerous values of this function with different C and λ arguments as a preprocessing step for our PWL model presented in Section 4. In this experiment, we consider four different evaluation methods discussed in Section 3.2.

1. Running a simulation assuming the empirical distribution of the parcel pickup process as demonstrated in Table 3.
2. Running a simulation assuming that the pickup process is memoryless; i.e., each parcel is picked up with probability p in each period regardless of its seniority. As above, the value of p is selected such that the expected pickup time is identical to the one of the corresponding empirical distribution.
3. Evaluating the mean postponement based on the steady-state probability of a truncated DTMC. The number of states in the chain (on each side of the bipartite graph) was selected to be five times the capacity of the SP. Thus, the probability of the system being in a state that the DTMC does not represent is negligible, unless it is working under an extremely heavy load. A setting that in any case is not likely to be of interest to a planner of an SP network.

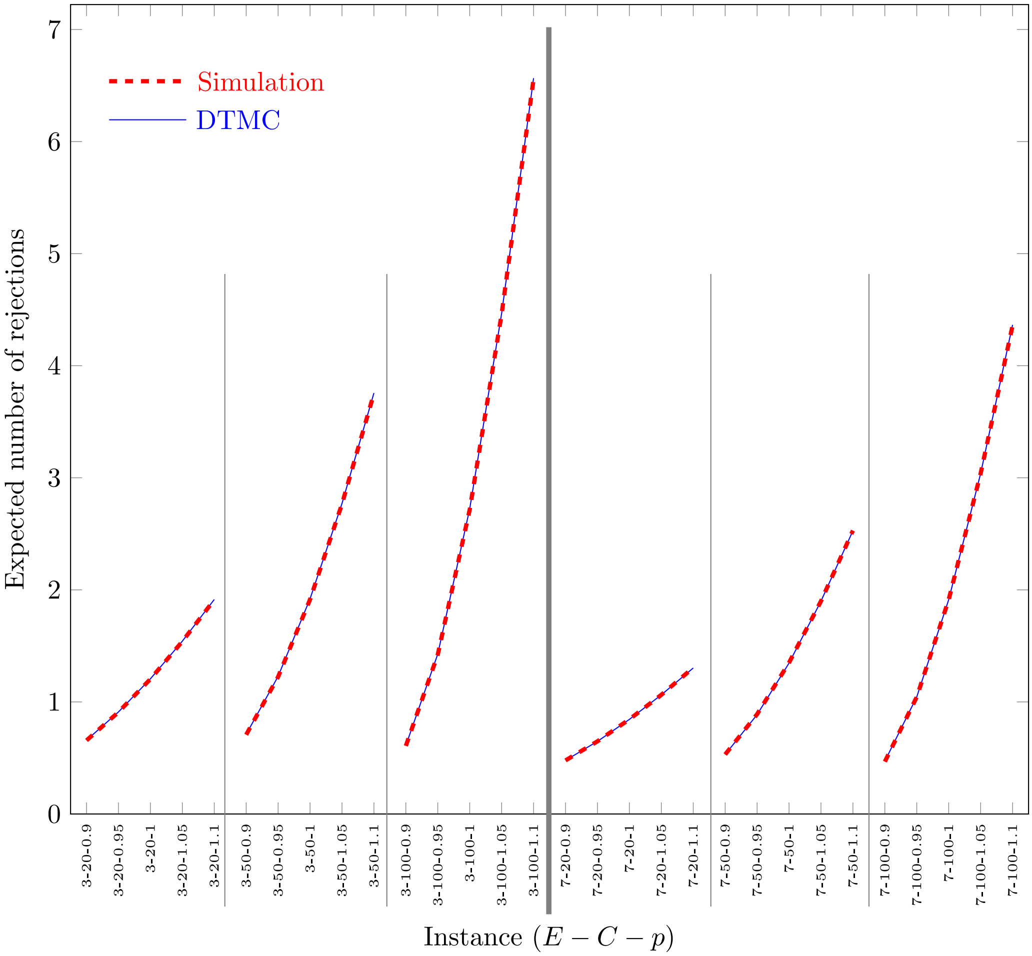

Fig. 7. Calculation methods of the rejection function. . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4. Evaluation of the postponement function as the waiting time in a $M/G/k$ queuing system, where the number of servers is C, the arrival rate is λ, the service rate is p, and the coefficient of variation is calculated based on the mean and variance in Table 3. Recall that we can apply Kingman's law of congestion to obtain an approximation of the waiting time based on the first two moments of the service time distribution.

In our experiment, we estimated postponement for 24 different configurations with the two pickup time distributions (and corresponding calculated pickup probability), three levels of SP capacity ($C \in \{20, 50, 100\}$), and four levels of system utilization ($\rho \in \{0.8, 0.9, 0.95, 0.98\}$), from which the corresponding arrival rates are derived ($\lambda = \frac{Cp}{\rho}$). The solution times reported below were obtained on an Apple MacBook Air notebook with an ARM M1 processor (2020).

The simulation with the empirical pickup distribution was run for one million periods, and the first 100 were discarded from the calculation as a warmup. The mean postponement time was estimated with very small 99% confidence intervals, typically within 1%–4% of the simulation average. Therefore, we consider these values (plotted in blue in Fig. 8) as the ground truth for our analysis. We note that the simulation of each of the configurations took 12–13 s.

The methods based on the assumption that the pickup time distribution is geometric, i.e., the truncated DTMC and the simulation with geometric pickup distribution, were ruled out from further consideration as both deviated significantly from the solutions obtained via simulation. Moreover, none is significantly easier to compute than the more realistic simulation. In particular, the truncated DTMC required the definition of a relatively large chain for which the calculation of the steady-state probability requires long computation time and large memory.

The $M/G/k$ model is computed in negligible time (less than 1 ms) and provides a relatively good approximation for the postponement function, as observed in Fig. 8. It fails sometimes when the SP works at a very high utilization ($\rho = 0.98$), which is not likely to be the operational domain of the system if optimally configured.

We conclude that, unlike the rejection model, the postponement model is highly sensitive to the particular nature of the pickup process. Since, in practice, the pickup process is not likely to be similar to a memoryless process, we recommend using simulation

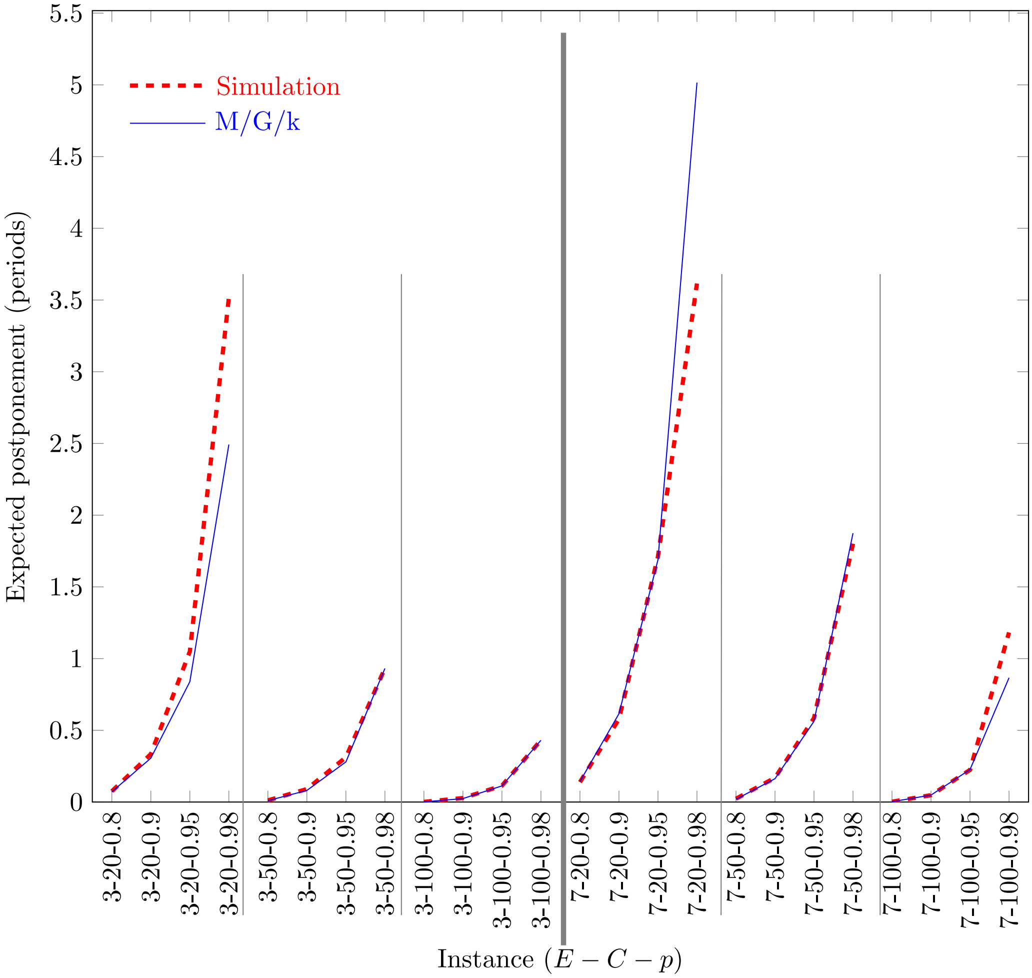

Fig. 8. Approximation methods of the postponement function. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to accomplish the preprocessing, but if computation time is an issue, approximation of the function based on $M/G/k$ is a viable alternative if the system is not expected to work under a very high load.

Appendix C. Piecewise linear approximation of the rejection and postponement functions

As discussed in Sections 3.1 and 3.2 and numerically demonstrated in Appendices A and B, we can approximate the values of the rejection and postponement functions for any set of parameters. We are still interested in a method to analytically describe the number of rejections as a function of the supply rate λ for some fixed combinations of the SP capacity and the time-to-pick-up distribution. Such a functional description can be embedded into a mathematical model that designs a network of SPs. Note that the location and capacity of each opened SP, as well as the location of its neighboring SPs, affects the supply rate it faces. Unfortunately, closed analytical descriptions of the rejection and postponement functions are out of our reach at this time. However, we will show that a piecewise linear approximation of these functions is a viable alternative. We demonstrate this concept with respect to the rejection function, but the same idea is applicable to the postponement function as well.

In Fig. 9, we plot the rejection function for an SP with capacity $C = 30$ and periodic pickup probability of $p = 0.5$ for $\lambda \in [0, 30]$, which is equivalent in this system to $\rho \in [0, 2]$, a range that spans any realistic scenario for a working SP. The red line is a very fine description of the function based on 101 calculations of its value in horizontal distances of 0.3. In contrast, the blue line is a piecewise linear approximation based on ten segments of the same length (2.72 each). The figure visualizes the strength of the approximation. The largest absolute vertical distance between the blue and red lines is obtained at $\lambda = 12.3$ $(\rho = 0.82)$, and it is only 0.11. Similar phenomena were observed for several other relevant system settings that we checked. In fact, for larger SPs, the linear approximation is even better. Therefore, we recommend using this granularity level of the piecewise linear approximation, where each "piece" represents a range of 0.2 load level units ρ.

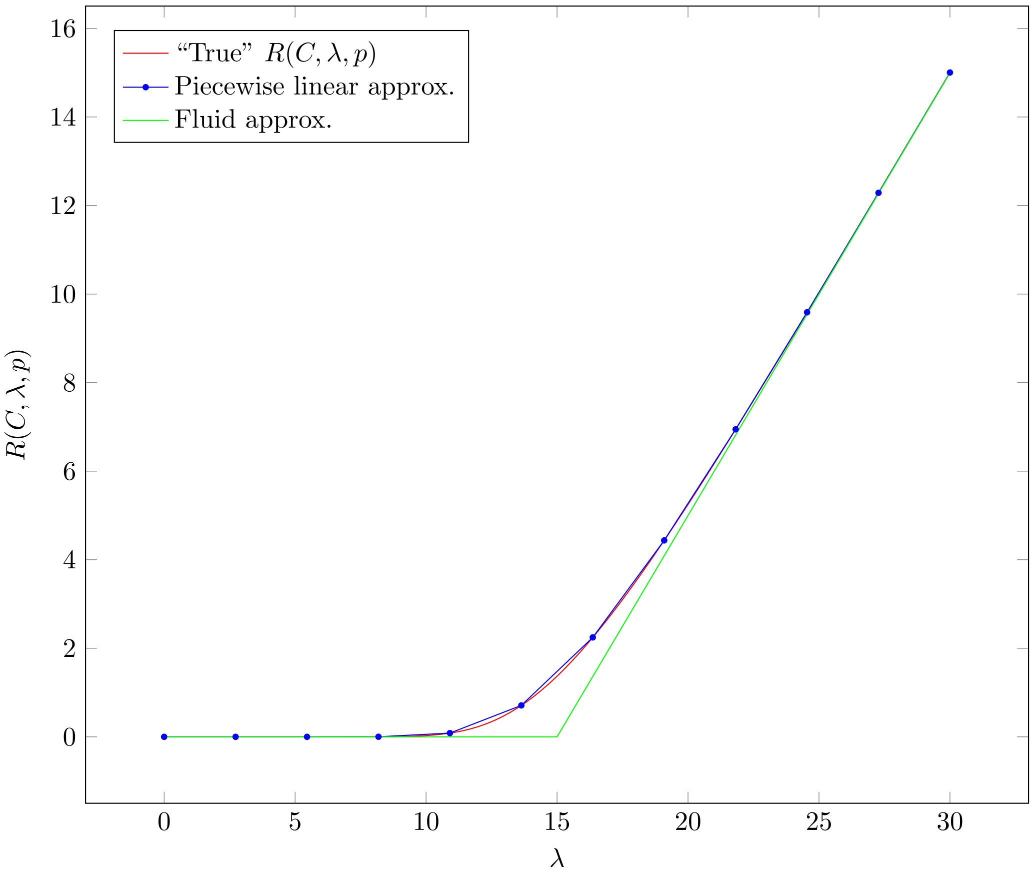

Fig. 9. Demonstrating the accuracy of the piecewise linear approximation in an SP with $C = 30, p = 0.5$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The green line in Fig. 9 represents the number of rejections predicted by a fluid approximation of the system. The figure demonstrates why this is not an acceptable approximation method for an SP that operates with ρ close to 1. With such an approximation of the rejection function, any optimization method will push all the SPs to full utilization when the approximation of the number of rejections is the least accurate; that is, no slack capacity will be built to cope with the stochasticity of the system.

We note that a better approximation of the rejection function at the same computational effort could be achieved by dividing the range into nonuniform segments. It is best to have shorter "pieces" around $\rho = 1$ and longer ones farther from this point, where the rejection function is nearly linear. While we leave a thorough investigation of this improvement for future research, a few trials indicated that setting 12 breakpoints of the piecewise linear approximation to $\rho = 0, 0.6, 0.7, 0.75, 0.85, 0.9, 0.95, 1, 1.1, 1.25, 1.5, 2$ resulted in a plot in which the piecewise linear approximation and the "true" function are almost indistinguishable by the naked eye: the maximum vertical distance between the two functions is less than 0.05 at $\lambda = 1.36$ (see Fig. 10).

Appendix D. Additional results with large synthetic instances

To further explore the computational limitations of our solution method and to support our insights from the experiments on smaller instances reported in Section 6, we created a test set of additional 40 instances with the same structure but with $10\times10 = 100$ candidate SP locations and $21 \times 21 = 441$ demand points. These instances represent a service area of 17.64 km^2. The dimensions of these instances are still not equivalent to a large city; however, they are representative of a new area to which a delivery company may consider expanding its coverage. As in Section 6, we created 10 random instances for each of the four combinations of $r \in \{401, 601\}$ and $\alpha \in \{5, 20\}$.

We tried to solve each of these instances with Cplex using the deterministic model and our PWL model with a time limit of one hour. While for both models, approximately half of the instances could be solved to optimality within this time limit, we obtained feasible solutions for all instances with relatively small optimality gaps.

In Fig. 11, we present the average optimality gap in each of the four groups (determined by the service radius and the rejection cost) for the two models. For each group, we report an average of 10 random instances.

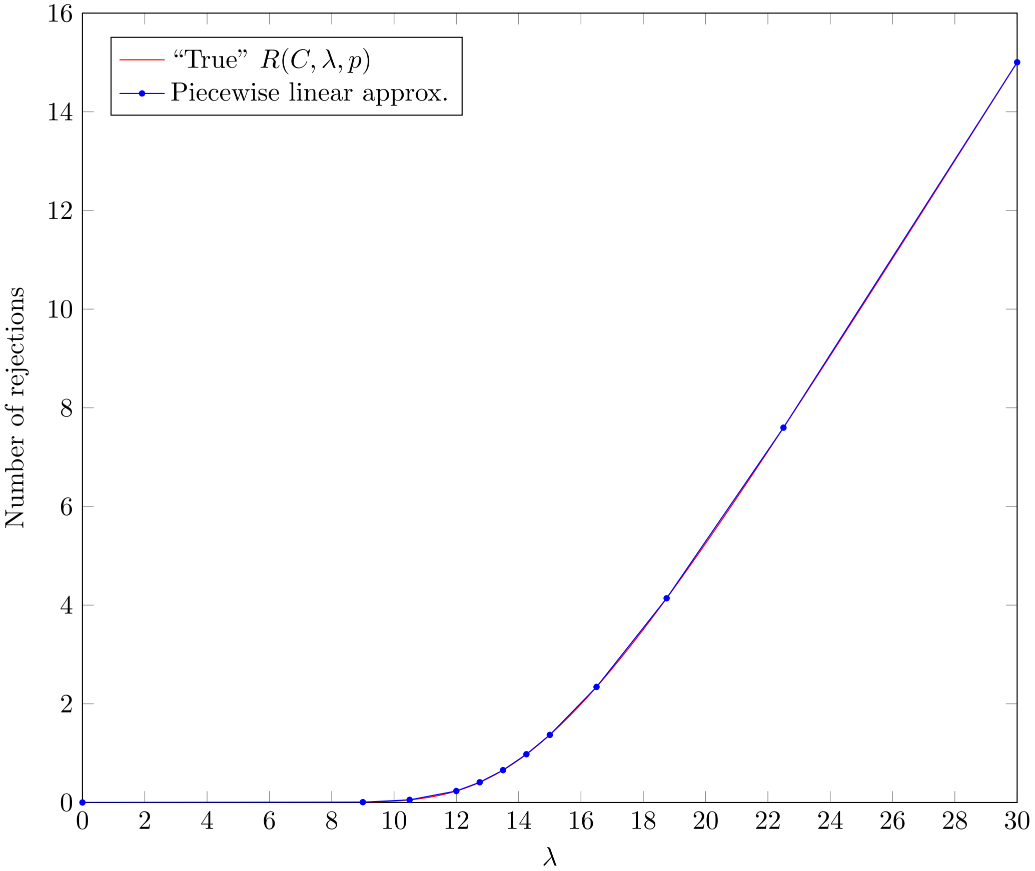

Fig. 10. Demonstrating the accuracy of the piecewise linear approximation in an SP with $C = 30, p = 0.5$, with nonuniform pieces.

It is apparent from Fig. 11 that the two models are harder to solve for larger service radii, which is not surprising. In the $r = 401$ instances, each SP covers approximately 12–13 demand points, while for $r = 601$, the average coverage of each SP is almost double that value. Furthermore, higher rejection costs make the approximation of the optimal solution more challenging, at least for the PWL model.

One can easily observe in Fig. 11 that for the harder instances (i.e., $r = 601$), the PWL model is easier to approximate than is the deterministic model, and it can consistently approximate the solutions of all the instances in our test set within an optimality gap of a few percent, subject to the one-hour time limit. The highest optimality gap over all the instances obtained for the deterministic (resp., PWL) models was 8.43% (resp., 4.43%). Both were obtained with $r = 601$ and $\alpha = 20$. These maximal values are not substantially higher than the average values depicted in the two rightmost bars of Fig. 11.

A more interesting comparison between the solutions of the two models was obtained by considering the total cost (setup + rejection costs) calculated using our rejection function for the (non-optimal) solutions obtained by running the two models. Recall that this calculation is built into the PWL model and can be performed as a post-processing step for the deterministic model. In Fig. 12, we present the extra total cost of the configurations obtained from the solutions of the deterministic model relative to those of the PWL model. Clearly, the PWL model yields configurations with lower total costs. This is the case for all 40 instances in our test set, including the cases where the solution obtained from the models is not optimal. However, as observed in the cases of smaller instances, the difference is substantial when the payment for each rejection is high ($\alpha = 20$), while it is relatively minor in the instances with $\alpha = 5$. This difference can be explained by the fact that the deterministic model underestimates the number of rejections, which has a greater impact on the solution when rejections are more expensive. Indeed, in all 40 instances, the deterministic model missed the true expected number of rejections by approximately 67 on average. The system-wide average number of rejections calculated by the deterministic model was 20, while the true expected number of rejections calculated using our rejection function was approximately 87. The estimation error was slightly higher for $\alpha = 20$ and lower for $\alpha = 5$.

Moreover, when comparing Fig. 4 with 12, we observe that the advantage of the PWL model over the pure deterministic one increases with the dimensions of the problem (i.e., the number of demand points and candidate SP locations).

A closer look at the results of our experiment with the medium-size instances indicates that in the solutions obtained from the PWL model, the rejections cost accounts for 10%–11% of the average total cost. The ratio is not significantly affected by the rejection cost parameter (α). The PWL model appears to "know" how to adjust the investment in SP infrastructure to the rejection

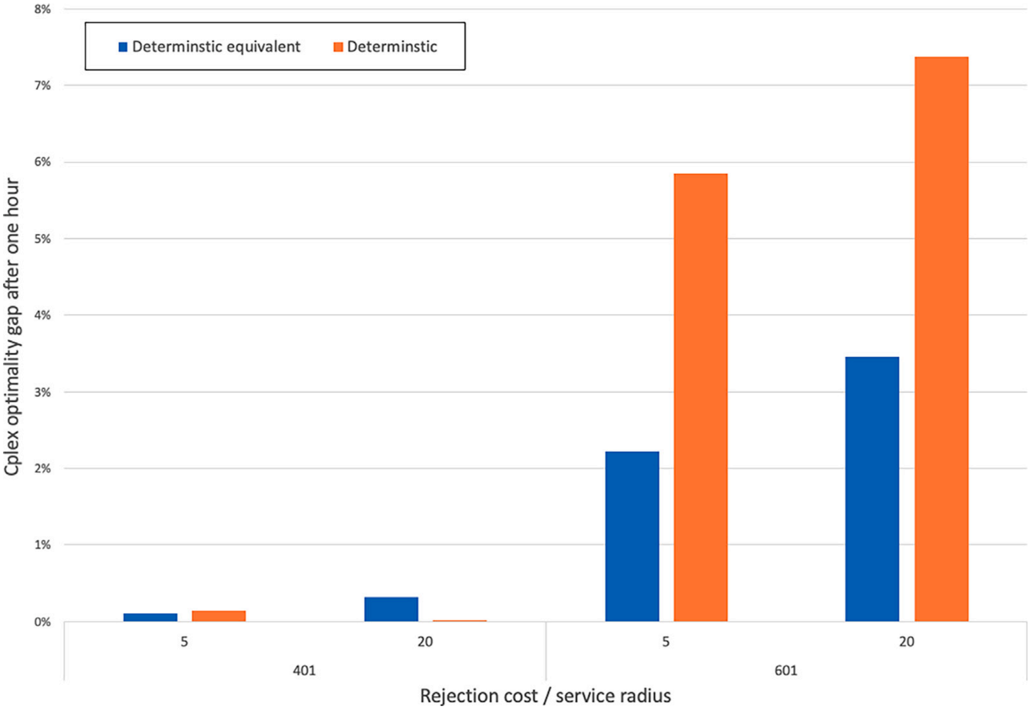

Fig. 11. Optimality gap of the deterministic and PWL models, medium instances with $m = 100, n = 441$ with one-hour time limit.

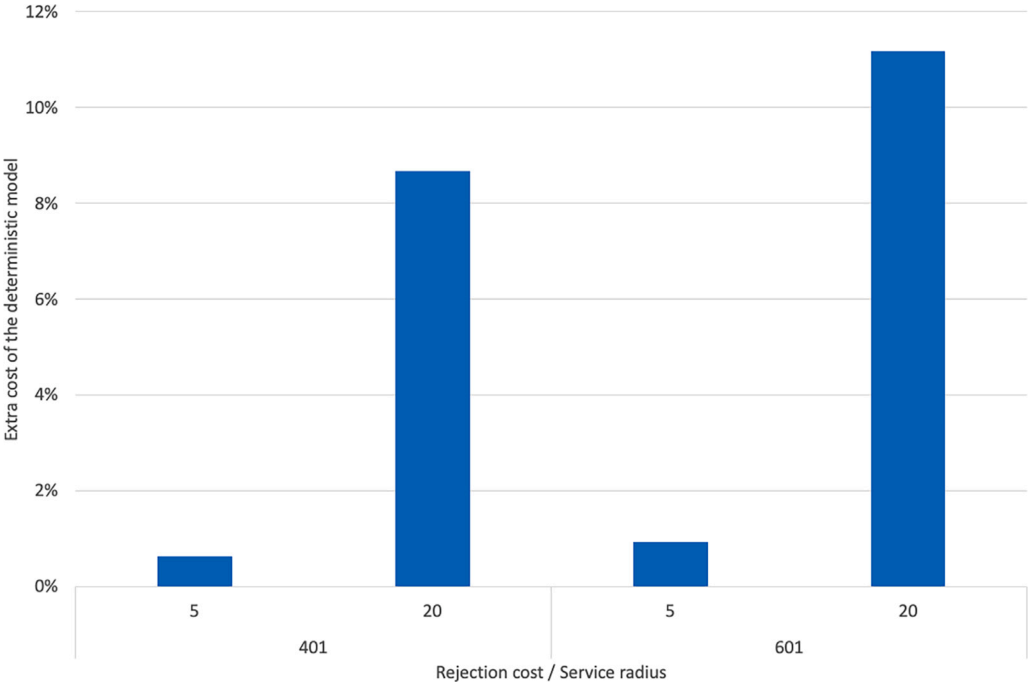

Fig. 12. Extra total cost of the deterministic model, medium instances with $m = 100, n = 441$.

cost parameter. However, in the deterministic model, when increasing α from 5 to 20, the share of the rejection cost in the total cost doubles from 13% to 26%.

Additionally, when applying the solution obtained from the PWL model, the average utilization level of the SPs in the system ranges from 81.7% for the $\alpha = 20$ instances to 92.6% for $\alpha = 5$. In contrast, the solutions obtained from the deterministic model work at much higher utilization and are not affected as much by the rejection cost: their ρ ranges from 90.3% for $\alpha = 20$ to 95.1% for $\alpha = 5$.

Thus, we conclude, for now, that the proposed PWL model is better suited to the task of determining the capacity and locations of the SPs than the two baseline models presented above. Our model is easier to solve (or approximate) and yields better solutions because it minimizes rejection directly instead of using a surrogate function. Its advantage is more prominent when the cost associated with the stochastic component of the problem (i.e., the rejection cost) is higher.

Facility location problems and, in particular, the SP location and capacity problems are solved to support an organization's strategic decisions. As such, decision-makers are likely ready to allocate substantial computational resources to solve them. However, as an NP-Hard problem, the SP location and capacity problem is bound to face some computational limitations that may require creative heuristic approaches to overcome.

The last phase of our numerical effort was to draw the border of the problem instance dimensions and characteristics, for which the exact solution method using state-of-the-art solvers and hardware is still practical. By "practical", we mean that a good approximation can be obtained within a reasonable time. Of course, both "good" and "reasonable" in this context are in the eye of the beholder. However, in our view, problem instances that can be solved consistently within three hours with optimality gaps of approximately five percent may be considered satisfactory. Harder instances may call for heuristic methods that are not within the scope of this paper.

To accomplish our last goal, we created 10 large and challenging instances. We focused on larger service radius and rejection cost parameters that were shown to create more challenging instances in the previous experiments. That is, we set $r = 601, \alpha = 20$. We created a new synthetic geography composed of $31 \times 31 = 961$ pixels of 200 × 200 m each for a total area of 38.44 km^2, which may represent the total built-up area of a city. A total of $15 \times 15 = 225$ candidate SP locations were generated, as described above and illustrated in Fig. 3. To further push the limits, we extended the variety of SP capacities from three to four and considered SPs with capacities of 30, 60, 90, and 120. Ten random instances with different demand data and SP candidate locations were generated and attempted to be solved by Cplex with a time limit of three hours.

None of the ten instances were solved optimally, but feasible solutions with average optimality gaps of 5.89% were obtained. The range of the optimality gaps was 4.91–6.98%. These results demonstrate that solving our PWL model using a commercial solver is a viable option for planners of SP networks in relatively large areas. However, very large networks designed at one time call for different solution methods. Interestingly, some of the realistic problem instances solved in Section 7 are much larger than the synthetic instances that we solved here, and yet we could solve them with smaller optimality gaps. It may be the case, that the irregularity and asymmetry that characterize a realistic urban environment lend themselves to easier problem instances.

References

Albareda-Sambola, M., Fernández, E., Saldanha-da Gama, F., 2011. The facility location problem with Bernoulli demands. Omega 39 (3), 335–345.

Baron, O., Berman, O., Krass, D., 2008. Facility location with stochastic demand and constraints on waiting time. Manuf. Serv. Oper. Manag. 10 (3), 484–505.

Berman, O., Krass, D., 2004. 11 Facility location problems with stochastic demands and congestion. In: Drezner, Z., Hamacher, H.W. (Eds.), Facility Location: Applications and Theory. Springer Science & Business Media, chapter 11.

Berman, O., Krass, D., 2019. Stochastic location models with congestion. In: Location Science. Springer, pp. 477–535.

Bloch, A., Tzur, M., 2019. The Location Problem in Crowd Delivery Systems. Technical report, Industrial Engineering Department, Tel Aviv University.

Che, Z.-H., Chiang, T.-A., Luo, Y.-J., 2022. Multiobjective optimization for planning the service areas of smart parcel locker facilities in logistics last mile delivery. Mathematics 10 (3), 422.

Correia, I., Saldanha-da Gama, F., 2019. Facility location under uncertainty. In: Location Science. Springer, pp. 185–213.

Deutsch, Y., Golany, B., 2018. A parcel locker network as a solution to the logistics last mile problem. Int. J. Prod. Res. 56 (1–2), 251–261.

Dyer, M., Proll, L., 1977. Note—on the validity of marginal analysis for allocating servers in m/m/c queues. Manage. Sci. 23 (9), 1019–1022.

Gans, N., Koole, G., Mandelbaum, A., 2003. Telephone call centers: Tutorial, review, and research prospects. Manuf. Serv. Oper. Manag. 5 (2), 79–141.

Grabenschweiger, J., Doerner, K., Hartl, R., 2022. The multi-period location routing problem with locker boxes. Logist. Res. 15 (8).

Harel, A., Zipkin, P.H., 1987. Strong convexity results for queueing systems. Oper. Res. 35 (3), 405–418.

Kahr, M., 2022. Determining locations and layouts for parcel lockers to support supply chain viability at the last mile. Omega 113, 102721.

Laporte, G., Nickel, S., Saldanha-da Gama, F., 2019. Introduction to location science. In: Location Science. Springer, pp. 1–21.

Lemke, J., Iwan, S., Korczak, J., 2016. Usability of the parcel lockers from the customer perspective–the research in polish cities. Transp. Res. Procedia 16, 272–287.

Lin, Y.H., Wang, Y., He, D., Lee, L.H., 2020. Last-mile delivery: Optimal locker location under multinomial logit choice model. Transp. Res. E 142, 102059.

Lin, Y., Wang, Y., Lee, L.H., Chew, E.P., 2022. Profit-maximizing parcel locker location problem under threshold luce model. Transp. Res. E (ISSN: 1366-5545) 157, 102541.

Mancini, S., Gansterer, M., Triki, C., 2023. Locker box location planning under uncertainty in demand and capacity availability. Omega 102910.

Orenstein, I., Raviv, T., Sadan, E., 2019. Flexible parcel delivery to automated parcel lockers: models, solution methods and analysis. EURO J. Transp. Logist. 8 (5), 683–711.

Owen, S.H., Daskin, M.S., 1998. Strategic facility location: A review. European J. Oper. Res. 111 (3), 423–447.

Pagès-Bernaus, A., Ramalhinho, H., Juan, A.A., Calvet, L., 2019. Designing e-commerce supply chains: a stochastic facility–location approach. Int. Trans. Oper. Res. 26 (2), 507–528.

Pan, S., Zhang, L., Thompson, R.G., Ghaderi, H., 2021. A parcel network flow approach for joint delivery networks using parcel lockers. Int. J. Prod. Res. 59 (7), 2090–2115.

Rabe, M., Chicaiza-Vaca, J., Gonzalez-Feliu, J., 2020a. Concept for simulation-optimization procedure model for automated parcel lockers as a last-mile delivery scheme: A case study in the city of dortmund. In: 13th International Conference of Research in Logistics and Supply Chain Management. July 4th-7th, Le Havre, France. pp. 765–774.

Rabe, M., Chicaiza-Vaca, J., Tordecilla, R.D., Juan, A.A., 2020b. A simulation-optimization approach for locating automated parcel lockers in urban logistics operations. In: 2020 Winter Simulation Conference. WSC, pp. 1230–1241. http://dx.doi.org/10.1109/WSC48552.2020.9384087.

Rabe, M., Gonzalez-Feliu, J., Chicaiza-Vaca, J., Tordecilla, R.D., 2021. Simulation-optimization approach for multi-period facility location problems with forecasted and random demands in a last-mile logistics application. Algorithms 14 (2), 41.

Ranjbari, A., Diehl, C., Dalla Chiara, G., Goodchild, A., 2023. Do parcel lockers reduce delivery times? Evidence from the field. Transp. Res. E 172, 103070.

Rohmer, S., Gendron, B., 2020. A Guide to Parcel Lockers in Last Mile Distribution: Highlighting Challenges and Opportunities from an OR Perspective. Cirrelt Montreal, Number 11.

Ross, S.M., 2014. Introduction to Probability Models. Academic Press.

Schwerdfeger, S., Boysen, N., 2022. Who moves the locker? A benchmark study of alternative mobile parcel locker concepts. Transp. Res. C 142, 103780.

Snyder, L.V., 2006. Facility location under uncertainty: a review. IIE Trans. 38 (7), 547–564.

Turkeš, R., Sörensen, K., Cuervo, D.P., 2021. A matheuristic for the stochastic facility location problem. J. Heuristics 27 (4), 649–694.