Article publicat / *Published paper*:

Albareda Sambola, M., Landete, M., Monge, J.F., Sainz Pardo, J.L. (2017). Introducing capacities in the location of unreliable facilities. European Journal of Operational Research, vol. 259 (1), p. 175-188. DOI: <10.1016/j.ejor.2016.10.013>

Introducing capacities in the location of unreliable facilities

Maria Albareda-Sambola[1], Mercedes Landete[2,*], Juan F. Monge[2], José L. Sainz-Pardo[2]
[1]Universitat Politècnica de Catalunya, 08222 Terrassa (Barcelona) Spain,
[2]Universidad Miguel Hernández de Elche, 03202 Elche (Alicante) Spain.

October the 4th, 2016

Abstract

The goal of this paper is to introduce facility capacities into the Reliability Fixed-Charge Location Problem in a sensible way. To this end, we develop and compare different models, which represent a tradeoff between the extreme models currently available in the literature, where *a priori* assignments are either fixed, or can be fully modified after failures occur. In a series of computational experiments we analyze the obtained solutions and study the price of introducing capacity constraints according to the alternative models both, in terms of computational burden and of solution cost.

Keywords: Location, Discrete location problems, Reliability models, Capacitated facilities

[*]Corresponding author - email: landete@umh.es; Tel: +34 966658718; Fax: +34966658715

1 Introduction

Reliable facility location models are increasingly being studied in the discrete facility location literature, since they allow to make strategic decisions that, without too large increases in the regular operating costs, prevent the systems from severe deteriorations when facilities fail. Although some literature exists dealing with reliability issues in location problems defined on networks(e.g. Puerto et al. [2014]) or on the plane (Fathali [2015]), we will concentrate on the discrete case.

This type of models where first proposed in Drezner [1987], where the authors analyzed the extensions of the classical p-center and p-median problems that are obtained when facility failure probabilities are taken into account, but the number of facilities that can fail is fixed. However, these models were not very much studied until some years later, restarting with Snyder and Daskin [2005]. The reader is referred to Snyder [2006] for a survey on early works concerning this type of models.

More recently, authors have considered different modeling assumptions regarding several aspects of the problem. Some works present heuristics [Shen et al., 2011, Peng et al., 2011, Alcaraz et al., 2012, Aydin and Murat, 2013, Aboolian et al., 2013] while only a few ones are concerned with exact solution methods (see, e.g. O'Hanley et al. [2013], Alcaraz et al. [2015, 2016]). Additionally, there is a vast literature that explores different modeling assumptions, and their effect on the solution structure. On the one hand, a common assumption in the early literature was that all facilities share the same failure probability. This assumption has been relaxed in several papers, as, for instance Berman et al. [2007b], Cui et al. [2010], or O'Hanley et al. [2013]. In a similar way, while all early references concentrated on problems where facility failures take place independently, works addressing correlated failures are being more and more studied [Li and Ouyang, 2010, Berman et al., 2013, Li et al., 2013b].

Other alternatives that can be found in the literature refer to the number of failures considered (as many as located facilities in some cases, like Snyder and Daskin [2005], or just a few of them, like Zhan [2007],Lee and Chang [2007] or Li et al. [2013a]) or the system behavior after a failure takes place. In this case, some authors distinguish between problems with complete information, and problems with incomplete information where customers are only aware of facility failures upon arrival (see, for instance Berman et al. [2009] or Albareda-Sambola et al. [2015]).

Most often, facilities operating in real word have limited capacities. Therefore, introducing capacitated facilities in this type of models is of real interest. To the best of our knowledge, only a few recent references exist where capacities have been taken into account. These include Gade and Pohl [2009], where allocation of customers to facilities is only made in a second stage, once the facilities that have failed are known. For this problem, the authors propose a solution method based on sample average approximation. In contrast to this work, where authors assume that demands are splittable and can thus be partially served from different facilities, Aydin and Murat [2013] address a similar model with unsplittable demands, where excess demand is served from an emergency facility with a higher assignment cost. Similar models can be found in Qin et al. [2013], where only a set of possible scenarios is considered or in An et al. [2014] where the possibility that facility disruptions cause variations on the demands is also modeled. Recently, a capacitated p-center problem with potential failures has been considered in Espejo et al. [2015]. In this case, it is assumed that customers will always be served from their closest available facility and the obtained solutions guarantee that facility capacities will not be exceeded if at most one facility fails.

Another work addressing capacitated facilities is Azad et al. [2014]. Here, the authors consider quite a general model where disruptions can occur both, in the facilities and in the transportation network. Failure probabilities are site-dependent and partial failures are also considered. In their model, on top of location decisions, for each located facility and each assignment it also has to be decided whether it is reliable or unreliable. In case of assignments to unreliable facilities, a backup safe assignment needs to be decided. An interesting issue introduced in this work is the minimization of the conditional value at risk, which requires the use of some risk measure within this context. This model shares some characteristics with some reliability problems defined on networks, although in those cases the measures of system performance are completely different (e.g. Lin and Chang [2012]).

Finally, Lim. et al. [2013] is a rather theoretical paper where the authors evaluate the impact of using inaccurate estimates of failure probabilities and correlation among disruptions in reliable location problems. To this end, they consider a continuously distributed demand on a plane and evaluate the cost associated with a given set of facilities. In the analysis of models with correlated disruptions they include capacities at the facilities, since in this case correlations are more relevant. Furthermore, they impose a unit penalty cost to the reliable facility when it serves additional demand beyond its capacity. In this setting, they conclude that the expected total cost for the capacitated facility reliability problem increases with the degree of correlation and decreases with the facility capacity.

In a different context, some authors have considered the so-called *location problems with stochastic demands and congestion*, where demands arise according to a Poisson process, and the set of located facilities act as servers with limited queue lengths. In these cases, facility congestion can be seen as a facility failure endogenous to the system. In these models, customers are supposed to be more likely to renounce as they have to travel further from their locations in order to get service. The goal in this case is not always to minimize the system costs, but to maximize the expected covered demand. In Berman et al. [2007a], for instance, the authors assume that customers travel to their closest facility, and use the approximations of the actual queue performance measures obtained by ignoring the fact that, in this case, facility congestion events are not independent. A similar problem is heuristically solved in Zhang et al. [2009].

In this work, we consider a fixed charge facility location problem with unsplittable demands. Facilities can fail with homogeneous probability, and these failures occur independently. For each customer, a sequence of assignments to open facilities is defined. An extra dummy non-failing facility with large assignment costs is used to model situations where a customer is either lost or outsourced. Capacity constraints on the facilities are stated as hard constraints for the scenario where no failures occur, but relatively small violations are allowed if they do. In Section 2 we motivate our modeling assumptions, we discuss how the expected demand gives insufficient information about the solution and that the full reallocation of customers when failures occur is not always the best option. In Section 3 we introduce the notation and the basic model for introducing capacities to the so called reliable facility location problem. In Section 4 we give the details of four different models which enlarge the basic formulation with new capacity constraints; this is the main section of the paper. Section 5 is devoted to the analysis of our computational experiments. Finally, some conclusions and future research lines are stated in Section 6. Some extra technical details regarding the proposed models are provided in the two appendices that close the paper.

2 Notation and Modeling assumptions

Let I be a set of customers, and J the set of potential sites for the facilities. For each customer $i \in I$ and each site $j \in J$, d_{ij} represents the transportation cost per unit of demand served from j to i, h_i the demand of customer i, and f_j and Q_j are, respectively, the set-up cost and the regular capacity of facility j. Moreover, facilities are split into two disjoint sets $J = F \cup NF$. Each of the facilities in set F, if open, can fail with probability q (common to all unreliable facilities). As opposite, non-failing facilities (those in NF) are assumed to be fully reliable. We also consider that facility failures take place independently of each other. Moreover, if t facilities from F are opened, 2^t scenarios can occur, depending on which of those facilities are available and which ones have failed.

To model the situations where a customer is lost or outsourced, we include a dummy facility $u \in NF$, with $f_u = 0$, unlimited capacity, and assignment costs equal to the cost of loosing or outsourcing a customer. This facility is forced to be open.

In the Reliable Fixed-charge Location Problem with Capacity constraints (CRFLP) a set of facilities has to be opened, and a sequence of them has to be associated with each customer in such a way that all customers are served by a real facility as their primary assignment, capacity constraints are strictly satisfied in the situation where all open facilities are operative and, at least, nearly satisfied in all other scenarios. The goal of the CRFLP is to minimize the total cost, which includes fixed costs for opening facilities, regular service costs, expected service costs, and expected outsourcing costs.

As mentioned above, the literature on reliable facility location problems is quite extensive and covers a wide range of modelling assumptions. The basic ones refer to the type of objective (center, median, including or not location costs, including extra costs for fortification or penalties, etc.), or the nature of the facility failures (independent or correlated, and having common or site dependent probabilities). Additional assumptions concerning the service strategy must be also made to clearly define how the system will adapt in the scenarios where failures occur.

For instance, capacities can be considered in different ways. To the best of our knowledge, most of the authors that have considered them, stated them as strict limits on the throughput of a facility. However, in many situations, the capacity of a facility is stated in terms of its production in regular conditions, being it possible to increase it in emergency situations. Therefore, in this work we propose to allow assignments for which some facilities might have to serve a demand that is slightly over their capacity if exceptional failures occur. The goal of this paper is to explore different models that allow solutions of this kind, but keep a limit on these excesses. Note that the expected demand at a facility gives some idea on the deviations from the capacity, but some extra information is needed to have a clear view of the risk of overload at a facility.

To illustrate this fact, we consider the example situation of Figure 1. Squares represent open facilities, all with capacity 5 and failure probability equal to 0.2, and circles represent customers, whose demand is given inside the circle. In each configuration, continuous lines

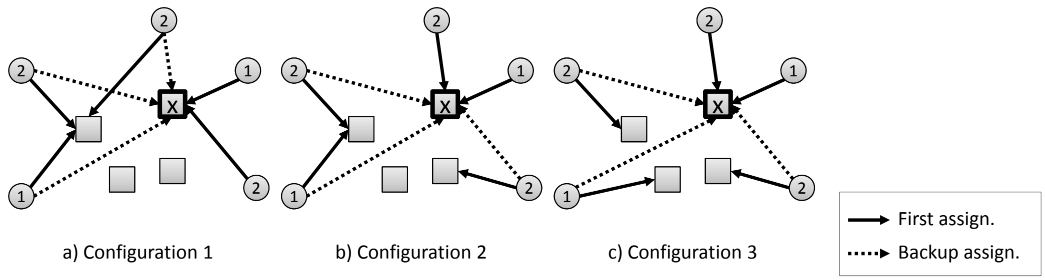
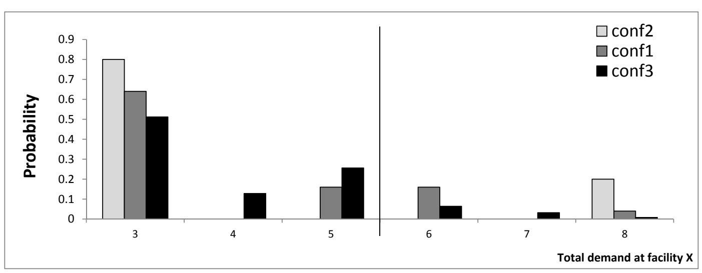


Figure 1: Example 1: Different assignment configurations

show the assignment of customers to facilities in the scenario where no facilities fail. If that facility fails, the dotted line gives the facility where the customer will be served from instead. For instance, in configuration 1, the facility marked with an X, if available, will always serve at least two customers (with demands 1 and 2), but it will also have to serve the other three customers in all the scenarios where the leftmost facility fails, which have a probability equal to 0.2. So, the expected demand at X is $1 \cdot 3 + 0.2 \cdot 5 = 4$.

For the three proposed assignment configurations, the expected demand at the facility marked with an X is equal to 4, but they represent quite different situations. Figure 2 gives the probability distribution of the demand received at facility X, conditional to the event that X has not failed.



	conf. 1	conf. 2	conf. 3
$\mathbb{E}$(demand at X)	4	4	4
$\mathbb{P}$(X's capacity is exceeded)	0.2	0.2	0.104
$\mathbb{E}$(overload at X)	0.6	0.28	0.152

Figure 2: Demand distribution at facility X in Example 1

In this figure it can be seen that in all three cases, even if the expected demand is smaller than the capacity, there are scenarios where the capacity is exceeded, and this happens with different probabilities and by different amounts, depending on the assignment pattern. Detailed information is provided at the bottom of Figure 2. Here, $\mathbb{E}$ stands for expectation, $\mathbb{P}$ stands for probability, and the overload of a facility is defined as the expected value of the amount by which its assigned demand exceeds the facility capacity. Obviously, the difference between the expected demand and the facility capacity provides a lower bound on the expected facility overload, but this lower bound is quite weak, as it can be seen in this example. To attain the goal of this work, tighter bounds and estimates of this excess will be presented which can be expressed linearly in terms of the decision variables. In Example 1, according to Figure 2 if we ignore the costs, we shall choose Configuration 3 among the three configurations since here overload occurs with the smallest probability, and its expected value is also the smallest one.

Of course, since the motivation behind the idea of allowing some capacity violations is the feeling that facilities can serve demands a bit beyond their capacities in some critical situations, in this work, we assume that in regular conditions (that is, in the scenario where no facility fails) capacity constraints are strictly satisfied at all the facilities.

Another main decision that needs to be made to devise a strategy is to define what decisions must be made a priori, and which ones can be made later, once the set of available facilities is known, in order to adapt the a priori solution to the actual situation. In this context, several alternatives can be considered depending on the flexibility of the supply chain, ranging from the high flexibility of Gade and Pohl [2009], Aydin and Murat [2013], An et al. [2014] or Qin et al. [2013], where all assignments can be freely modified in the case of disruptions, to the rigid situation of Azad et al. [2014], where customers only have one assignment for regular operating mode, and one assignment for emergencies.

In this paper, we propose an intermediate strategy that aims at guaranteeing a certain stability in the assignment of customers to facilities while improving the resilience of the system. To this end, each customer has several prioritized assignments, fixed at a first stage and, for a given scenario he will be served from the first available facility from this assignments list. To model these prioritized assignments, we define several assignment levels for each customer, starting from level 0 (primary assignment), and we consider that the customer will be served from the facility it is assigned to at level r, if all the facilities it is assigned to at levels $r' < r$ have failed.

When facilities are uncapacitated, the optimal assignment sequence corresponds to assigning each customer to its closest open facility that is operative, and therefore, the two extreme policies described above, in fact, coincide. However, in presence of capacity constraints, assignments to the closest available facility might be unfeasible, and the two policies (deciding assignments a posteriori, and perform actual assignments according to a predefined order) may yield to different solutions, as it is shown in the following example.

Consider the situation of Figure 3 where each of the five customers (circles) has one unit of demand, three failing facilities (squares) are open, each with a capacity of 3 units, and an extremely large outsourcing cost, ρ. Also, let the failure probability be $q = 0.1$ and assume that service costs are associated with Euclidean distances (a unit grid is also plotted). Assuming that capacity constraints are strict, each row represents the optimal actual assignments for the different scenarios, when assignments can be defined a posteriori. At each scenario, filled squares represent operating facilities, and empty ones, facilities that have failed. For the customers, an arrow is shown to give their actual assignment for that scenario, or they are filled in black if they are outsourced/lost. For brevity, the scenario where all facilities fail has been not included in the picture since, in this case, the assignment cost is trivially $C = 5\rho$. For the remaining scenarios both, the scenario probability and the cost of the actual assignments for that scenario are given at the right hand side of the figure. Being

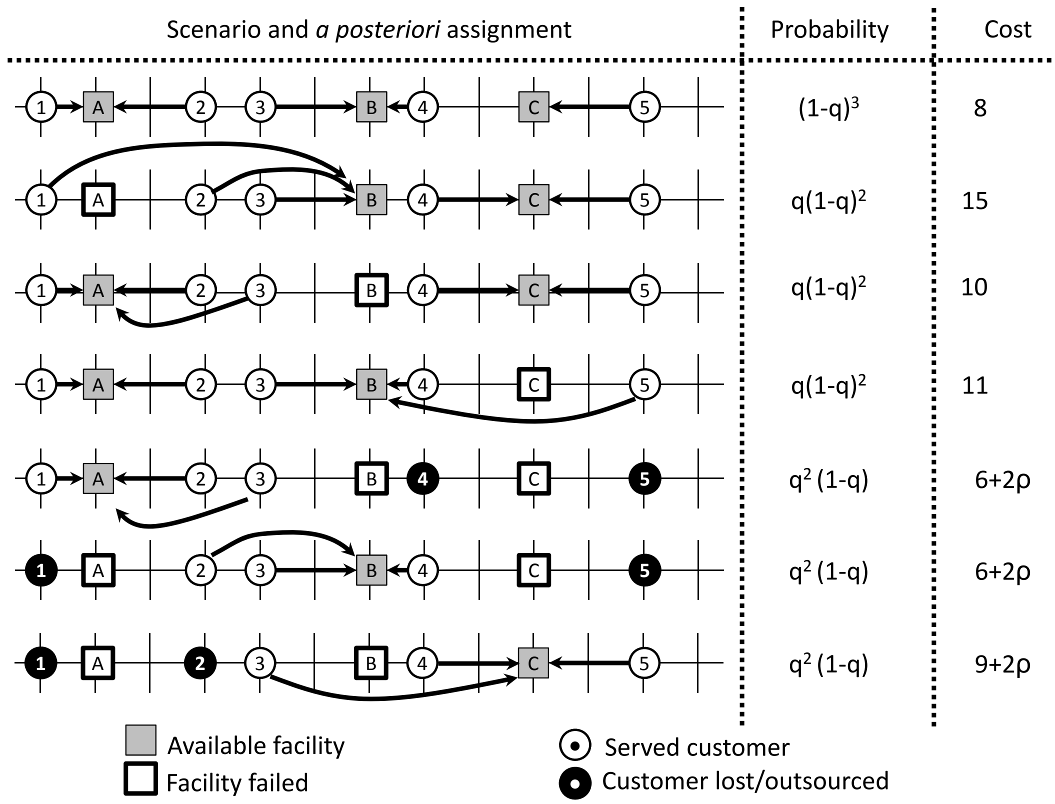


Figure 3: Example 2: Least-cost *a posteriori* assignments

the failure probability 0.1, the expected service cost in this case would be

$$\begin{aligned} & 8 \cdot (0.9)^3 + (15 + 10 + 11) \cdot 0.1 \cdot (0.9)^2 \\ & + ((6 + 2\rho) + (6 + 2\rho) + (9 + 2\rho)) \cdot (0.1)^2 \cdot (0.9) + 5\rho \cdot (0.1)^3 \\ = \; & 8.937 + 0.059\rho. \end{aligned}$$

Note, however, that the above solution does not correspond to any set of predefined assignment sequences for the customers. For instance, the first assignment of customer 4 is to facility B but, under the second scenario, it is being served from C, even if B is available, in order to make it feasible to serve customers 1, 2 and 3 from B. In fact, the optimal solution using our policy, if capacities are set as strict limits, is given by the following lists (D is used to refer to the dummy facility that represents loosing or outsourcing a customer). $1 \leftarrow AD$, $2 \leftarrow ABD$, $3 \leftarrow BACD$, $4 \leftarrow BCD$, and $5 \leftarrow CD$. (For instance, using these lists, in the scenario where facility B fails and the other two facilities are available, customers $1, 2$ and 3 would be served from A, and 4 and 5, from C.) Using these sequences, capacity constraints are satisfied in all scenarios, and the expected distribution cost is $8.469 + 0.221\rho$. Note that, as expected, for reasonable values of ρ this solution has a larger expected cost than the flexible solution since, to guarantee that the demands assigned to each facility never exceed its capacity, customers are more likely to be outsourced. This effect can be somehow reduced by allowing slight capacity violations in some scenarios, as discussed above. For instance, if we consider the solution given by the lists $1 \leftarrow ABD$, $2 \leftarrow ABD$, $3 \leftarrow BACD$, $4 \leftarrow CBD$, and $5 \leftarrow CBD$, then, the capacity constraint is only violated in the scenario where only facility B is available, which has a probability of 0.9%, and the expected excess demand overload is $2 \cdot 0.009 = 0.018$. By allowing this, the overall expected cost is now $10.332 + 0.041\rho$ that, for large outsourcing costs ρ is much smaller than before.

Despite being less flexible than the assumption made in most of the related literature, we believe that this type of strategy is already flexible enough to cope adequately with emergency situations where some facilities fail, and it is much easier to implement, since with the previously used strategies, the decision maker needs to systematically solve generalized assignment problems at an operational level to decide, under each scenario, what are the optimal assignments of the customers. Additionally to the complexity associated with implementing them, such policies propagate the effect of facility disruptions to the entire system whereas, with the proposed policy, this effect is much more local, since it only affects the customers that were originally assigned to a facility that has failed. This quality is specially relevant in systems with multiple agents. For instance, if different carriers are in charge for the distribution, having predefined assignment sequences allows them working autonomously, while planning the distribution according to a completely flexible policy would force them to work coordinately.

Consider the following binary variables:

- X_j that takes value 1 if a facility is open at site $j \in J$
- Y_{ijr} that takes value 1 if customer $i \in I$ is assigned to facility $j \in J$ at level $r = 0, 1, 2, \ldots$

If a given customer is assigned to a facility at level $r > 0$, it will be served from that facility in all scenarios where the facility is available, and all the facilities to which it has been assigned at lower levels have failed. In the assignments provided above for Example 2 ($1 \leftarrow ABD$, $2 \leftarrow ABD$, $3 \leftarrow BACD$, $4 \leftarrow CBD$, and $5 \leftarrow CBD$), the following Y variables are 1: $Y_{1A0}, Y_{1B1}, Y_{1D2}$, $Y_{2A0}, Y_{2B1}, Y_{2D2}$, $Y_{3B0}, Y_{3A1}, Y_{3C2}, Y_{3D3}$, $Y_{4C0}, Y_{4B1}, Y_{4D2}$, Y_{5C0}, Y_{5B1}, Y_{5D2}.

Using these variables, the CRFLP can be modeled as:

$$\text{(CRFLP)} \min \ \alpha\left(\sum_{j\in J} f_j X_j + \sum_{i\in I} h_i d_{ij} Y_{ij0}\right) + (1-\alpha)\sum_{i\in I} h_i\left(\sum_{j\in NF}\sum_{r\in R} d_{ij} q^r Y_{ijr} + \sum_{j\in F}\sum_{r\in R} d_{ij} q^r (1-q) Y_{ijr}\right) \quad (1)$$

$$\textbf{s.t. } X_u = 1 \quad (2)$$

$$\sum_{j\in F} Y_{ijr} + \sum_{j\in NF}\sum_{s=0}^{r} Y_{ijs} = 1 \qquad i\in I, r\in R \quad (3)$$

$$\sum_{r\in R} Y_{ijr} \le X_j \qquad i\in I, j\in J \quad (4)$$

$$\sum_{i\in I} h_i Y_{ij0} \le Q_j X_j \qquad j\in J \quad (5)$$

$$\text{Capacity constraints mostly hold} \quad (6)$$

$$X_j \in \{0,1\} \qquad j\in J \quad (7)$$

$$Y_{ijr} \in \{0,1\} \qquad i\in I, j\in J, r\in R \quad (8)$$

where $R = \{0, \ldots, |F|-1\}$ and α is a value in $(0,1)$.

Obseve that, for $i \in I$ and $j \in F$, $(1-q)\sum_{r\in R} q^r Y_{ijr}$ is the probability that customer i is served from facility j; if $Y_{ijr} = 1$, then r facilities have to fail before i needs being served from j, which happens with probability q^r. Moreover, for this service to take place, facility j has to be operative, which happens with probability $(1-q)$ (Note that these computations take advantage of the independence of failures). In the case of non-failing facilities, this last factor is not required, since they are always operative. This has been used to compute the expected service costs in the objective function. Note that in this objective function we consider a convex combination ($\alpha \in (0,1)$) of this expected service cost and the full system cost under the scenario where no facility fails. This is a common practice in reliable facility location models (Snyder and Daskin [2005]) that allows the end-user decide to what extent the chosen solution should depend on the potential facility failures.

Constraint (2) guarantees that the dummy facility used to model outsourced customers is indeed open. To force each customer to have one assignment at each level until it is assigned o a non-failing facility we use constraints (3). Furthermore, constraints (4) forbid both, assignments to sites where no facility has been located, and assignments of one customer to the same facility at different levels. Capacity constraints in the scenario where no facility fails are stated as (5). Finally, constraints (7) and (8) set the domains of the different sets of variables.

Constraint (6) is the focus of this paper. As we discussed throughout this section, the idea of this constraint is to forbid potential facility overloads that are too large, or that are too likely to occur. Obviously, there is no straightforward way to state this condition. In the next section we explore alternative sensible ways to deal with it. Later on, in Section 4 we will compare them in a series of computational experiments.

3 Alternatives for modeling capacity satisfaction

From now on, we will denote with QRFLP formulation CRFLP without constraints (6), that is, a CRFLP where capacities are only taken into account in the scenario where no facility fails. In what follows, we explore different alternative ways to state constraints (6). As mentioned above, the role of these constraints is to allow assignment configurations that may yield some facility overloads, but only by small amounts and with small probabilities.

3.1 Limits on the expected loads

A first naive approach to try to avoid capacity violations occurring too often or by large amounts is to keep the expected total demand at each facility below its capacity. This type of approach can be further relaxed by allowing these expected demands to exceed the capacities by a limited amount (V) and in a limited number of facilities, γ. To do so, we will use the following additional variables:

- $\nu_j \in \mathbb{R}_+$: amount by which the expected demand at facility $j \in J$ exceeds its capacity.
- $u_j \in \{0, 1\}$: indicates wether or not the expected demand at facility $j \in J$ exceeds the capacity Q_j.

Then, the above mentioned limits can be established using the following constraints instead of (6):

$$\sum_{i \in I} h_i \sum_{r \in R} q^r Y_{ijr} \leqslant Q_j X_j + \nu_j \qquad j \in J \tag{9}$$

$$\nu_j \leqslant V u_j \qquad j \in J \tag{10}$$

$$\sum_{j \in J} u_j \leqslant \gamma \tag{11}$$

$$\nu_j \geqslant 0 \qquad j \in J \tag{12}$$

$$u_j \in \{0, 1\} \qquad j \in J \tag{13}$$

Constraints (9) are used to compute the values of variables ν_j, while constraints (10) are used both, to identify the facilities where these variables are positive and to establish their limit, V. A limit γ on the number of facilities where ν variables can be positive is set by constraint (11). Finally, constraints (12) and (13) fix the domains of the new variables.

Note that, as opposite to the case of the objective function in formulation CRFLP, in the computation of the expected demand at facility j the factor $(1 - q)$ standing for the probability that facility j is available, is not required. Indeed, here we are computing the expected demand at j given that j is operative, since this demand is only relevant in this case.

In the computational experiments section we will refer to formulation QRFLP enlarged with constraints (9)-(13) as CRFLP-LEL(V, γ). The suffix LEL stands for Limits on Expected Loads.

3.2 Bounds on the expected overloads

As mentioned in Example 1, one global measure of the facility overloads is given by the *expected overload*, which is formally defined next.

Given a feasible solution (X, Y) to QRFLP, let $O(X) \subset J$ be the set of locations where facilities have been placed: $O(X) = \{j \in J : X_j = 1\}$. For each $j \in O(X)$, let ξ_j be the Bernoulli random variable that takes value 1 if facility j is operative, and 0 if it has failed. That is, $\xi_j \sim \text{Bernoulli}(1-q)$ for $j \in O(X) \cap F$, and $\xi_j = 1$ for $j \in O(X) \cap NF$. Then, the expected overload of the solution is defined as:

$$E(X,Y) = \sum_{j \in O(X)} \mathbb{E}\left[\left(\underbrace{\xi_j \cdot \sum_{i \in I} h_i \left(\sum_{r \in R} Y_{ijr} \cdot \prod_{s<r}\left(\sum_{j' \in O(X)} Y_{ij's}(1-\xi_{j'})\right)\right)}_{\text{demand at } j \text{ according to } \xi} - Q_j\right)^+\right], \quad (14)$$

where $(\bullet)^+ = \max\{0, \bullet\}$. That is, the expected overload of the solution is the sum, over all open facilities, of the expected values of the positive difference between the demand allocated to the facility and its capacity. Here, the demand allocated to a facility $j \in O(X)$ is computed as

$$\sum_{i \in I} h_i \left(\sum_{r \in R} Y_{ijr} \cdot \prod_{s<r}\left(\sum_{j' \in O(X)} Y_{ij's}(1-\xi_{j'})\right)\right).$$

That is, for each $i \in I$, its demand h_i will be allocated to j if and only if

- i is allocated to j at some level r ($Y_{ijr} = 1$) and,
- for all levels $s < r$, the facility to which j is assigned at level s (the only j' satisfying $Y_{ij's} = 1$), fails (that is, $1 - \xi_{j'} = 0$).

Accordingly, the demand allocated to an available facility at each scenario is the demand it receives at level 0 plus some of the demands whose primary assignment was made to a facility that has failed in that scenario. Since overloads are defined to be nonnegative, the expected overload increases with both, the probability of a positive overload, and the actual overload in the scenarios where it is positive.

Imposing predefined limits on the expected overload would be, theoretically, a reasonable way to state capacity constraints (6) in the CRFLP. However, as can be seen from expression (14), even the evaluation of $E(X, Y)$ in one solution is computationally expensive. Therefore, solving a variant of the CRFLP that includes a constraint of this type would be computationally unaffordable.

In this section we propose three linear approximations of the expected overload and prove that two of them are actually upper bounds of $E(X, Y)$. Therefore, constraints forcing these bounds to be below some predefined limit guarantee that the same constraints are satisfied in terms of the actual expected overload.

Proposition 3.1. *Given a solution (X, Y) to QRFLP, for $j \in O(X)$ and $r > 0$, let $\nu_{jr} = \left(Q_j - \sum_{i \in I} \sum_{s=0}^{r-1} h_i Y_{ijs}\right)^+$ be the capacity slack available at j if it only had to serve customers* *assigned to it at levels smaller than* r*, and* $\lambda_{jr} = \Big(\sum_{i\in I} h_i Y_{ijr} - \nu_{jr}\Big)^+$ *be the overload at* j *caused by assignments at level* r *after all assignments at lower levels have been considered. Then,*

$$E_1(X,Y) = \sum_{r>0}\left[\sum_{j\in F}\Big((1-q)q^r\Big)\lambda_{jr} + \sum_{j\in NF} q^r\lambda_{jr}\right] \tag{15}$$

is an upper bound of $E(X,Y)$*. That is,* $E_1(X,Y) \geqslant E(X,Y)$*.*

Proof:

In fact, it suffices to prove that for any facility $j_0 \in O(X)$, the contribution of j_0 to $E_1(X,Y)$ is an upper bound on its contribution to $E(X,Y)$.

Case $\mathbf{j_0 \in F}$ Consider a particular solution for which facilities in $O(X)$ can be sorted in a sequence such that, for any customer i assigned to j_0 at a certain level r ($Y_{ij_0r} = 1$), its previous assignments are given by $Y_{ij_1,r-1} = Y_{ij_2,r-2} = \cdots = Y_{ij_r0} = 1$.

In this case, j_0 will receive the requests of the customers assigned at level r whenever facility j_0 is available and facilities $j_1,\ldots,j_r$ fail, which happens with probability $(1-q)q^r$ and, therefore, in this particular solution, the expected overload at j_0 equals $\sum_{r>0}(1-q)q^r\lambda_{jr}$, *i.e.*, in this case, j_0 contributes with exactly the same amount to $E(X,Y)$ and to $E_1(X,Y)$.

In a more general solution, where the previously defined sequence does not exist, more than r facilities need to fail before j_0 completely receives the demand $\sum\limits_{i\in I} h_i Y_{ij_0r}$. Therefore, the overload λ_{j_0r} is incurred with a probability that is bounded above by $(1-q)q^r$ (it has some extra q factors) and, as a consequence, $\sum_{r>0}(1-q)q^r\lambda_{jr}$ is an upper bound on the expected overload at j_0.

Case $\mathbf{j_0 \in NF}$ In the case where $j_0 \in NF$ all the previous arguments apply, except for the probability that r facilities fail and j_0 is operative that now equals q^r, which is the coefficient in $E_1(X,Y)$ of λ_{j_0r} in this case.

□

To illustrate this bound, consider Example 2 depicted in Figure 3 and the already studied solution given by the sequences: $1 \leftarrow ABD$, $2 \leftarrow ABD$, $3 \leftarrow BACD$, $4 \leftarrow CBD$, and $5 \leftarrow CBD$, which as mentioned above, would correspond to

$$Y_{1A0} = Y_{2A0} = Y_{3A1} = Y_{3B0} = Y_{1B1} = Y_{2B1} = Y_{4B1} = Y_{5B1} = 1 = Y_{4C0} = Y_{5C0} = Y_{3C2} = 1.$$

Since the capacity of all three facilities is 3, it is clear that the only facility where there can be some overloads is facility B. Indeed, the only scenario where B is overloaded is the scenario where B is operative and both A and C fail, which happens with probability $(1-q)q^2$. In this case, B receives a total demand of 5, which exceeds its capacity by 2. Therefore, the total expected overload for this solution is $E(X,Y) = 2(1-q)q^2$. On the other hand, it can be clearly seen that, for this solution, the only nonzero capacity slacks are $\nu_{A1} = 1$, $\nu_{B1} = 2$, $\nu_{C1} = 1$, $\nu_{C2} = 1$, and the only nonzero overload, $\lambda_{B1} = 2$ (A total demand of 5 is assigned to B at levels $r \leqslant 1$ which is 2 units over its capacity). Therefore, in this case we have $E_1(X,Y) = 2(1-q)q$, which is not smaller than $E(X,Y)$ for any probability $q \leqslant 1$.

One possibility to ensure that the expected total overload associated with a solution lays below a certain threshold, is to impose some limit V to the value of its upper bound $E_1(X,Y)$. To do so, we can enhance formulation QRFLP by adding the following extra constraints:

$$\sum_{s=1}^{r}\sum_{i\in I} h_i Y_{ijs} \leqslant Q_j + \nu_{jr} \qquad \forall j \in J, r \in R \tag{16}$$

$$\lambda_{j1} = \nu_{j1} \qquad \forall j \in J \tag{17}$$

$$\lambda_{jr} = \nu_{jr} - \nu_{jr-1} \qquad \forall j \in J, r > 1 \tag{18}$$

$$\sum_{j\in F}\sum_{r>0} q^r(1-q)\lambda_{jr} + \sum_{j\in NF}\sum_{r>0} q^r \lambda_{jr} \leq V \tag{19}$$

$$\lambda_{jr}, \nu_{jr} \geqslant 0 \qquad \forall j \in J, r \in R \tag{20}$$

In what follows, we will refer to the resulting formulation as CRFLP-B1(V). Here, B1 stands for Bound of Type 1. In this set of constraints, extra continuous variables λ_{jr} and ν_{jr} are required to compute the value of $E_1(X,Y)$. Constraints (16) are used to compute the value of ν_{jr} that is consistent with the current values of Y variables. Constraints (17) – (18) are required to set the right values to λ_{jr} variables, and constraint (19) is used to compute and limit the value of $E_1(X,Y)$. Finally, the domains of the new variables are set in constraint (20).

In view of the proof of $E_1(X,Y)$ be an upper bound on $E(X,Y)$, and of the previous example, we can see that the bound $E_1(X,Y)$ can be tight if all customers follow the same prospective path, or rather loose if different customers end up at a given facility when different sets of facilities fail. The next bound we present uses this observation to strengthen $E_1(X,Y)$.

Proposition 3.2. *Given a solution to the QRFLP, (X,Y), let $t = |O(X) \cap F|$ and, for $j \in O(X)$ and $k \in O(X) \cap F$, let $\epsilon_{jk} = \left(\left(\sum_{i:Y_{ik0}=1} h_i Y_{ij1}\right) - \nu_{j1}\right)^+$. Then,*

$$E_2(X,Y) = \sum_{j\in J}\sum_{k\in F} \epsilon_{jk} q(1-q)^{t-1} + \sum_{j\in F} \lambda_{j1} q(1-(1-q)^{t-2})(1-q) + \sum_{j\in NF} \lambda_{j1} q(1-(1-q)^{t-1}) + \sum_{r>1}\left[\sum_{j\in F} \lambda_{jr} q^r (1-q) + \sum_{j\in NF} \lambda_{jr} q^r\right] \tag{21}$$

is an upper bound on $E(X,Y)$ that is tighter than $E_1(X,Y)$. That is,

$$E(X,Y) \leqslant E_2(X,Y) \leqslant E_1(X,Y).$$

Proof:

$E_1(X,Y)$ can be split into two terms as:

$$\underbrace{\sum_{j\in F}\lambda_{j1}q(1-q)+\sum_{j\in NF}\lambda_{j1}q}_{A}+\underbrace{\sum_{r>1}[\sum_{j\in F}\lambda_{jr}q^r(1-q)+\sum_{j\in NF}\lambda_{jr}q^r]}_{B}$$

The term B equals the last term of $E_2(X,Y)$; that is, the only difference between $E_1(X,Y)$ and $E_2(X,Y)$ refers to the assignments at level $r=1$, which are taken into account in the first two terms of the definition of E_2 and in part A of the former decomposition of E_1. Therefore, it suffices to prove that

$$A \geqslant \sum_{j\in J}\sum_{k\in F}\epsilon_{jk}q(1-q)^{t-1}+\sum_{j\in F}\lambda_{j1}q(1-(1-q)^{t-2})(1-q)+\sum_{j\in NF}\lambda_{j1}q(1-(1-q)^{t-1}) \quad (22)$$

and that the expression at the right-hand side bounds the contribution to $E(X,Y)$ of assignments at level $r=1$.

Take $j_0\in F\cap O(S)$. Then, the contribution of j_0 to A is $\lambda_{j_01}q(1-q)$, which stands for the excess caused by assignments at level $r=1$ (recall that assignments at level $r=0$ are forced to be feasible with respect to the capacity by (5)) times the probability of one scenario where j_0 is operative and at least one other facility fails. This last probability can be decomposed into the probability that exactly one facility different from j_0 fails (and, therefore, the $t-2$ remaining open failing facilities different from j_0 are operative) plus the probability that more than one of these facilities fail:

$$q(1-q)=q(1-q)(1-q)^{t-2}+q(1-q)\left(1-(1-q)^{t-2}\right)=q(1-q)^{t-1}+q(1-q)\left(1-(1-q)^{t-2}\right)$$

Now, in the case that only one open facility $k\neq j_0$ fails, the corresponding overload will only be $\epsilon_{j_0k}\leqslant\lambda_{j_01}$. Therefore, the contribution of j_0 to A will be

$$\begin{aligned}A_{j_0}&=\sum_{k\in F}\lambda_{j_01}q(1-q)^{t-1}+\lambda_{j_01}q(1-q)\left(1-(1-q)^{t-2}\right)\\&\geqslant\sum_{k\in F}\epsilon_{j_0k}q(1-q)^{t-1}+\lambda_{j_01}q(1-q)\left(1-(1-q)^{t-2}\right),\end{aligned}\quad (23)$$

which still bounds the contribution of j_0 to the expected overload in the solution caused by assignments at level $r\leqslant 1$. Note that this last expression corresponds to the contribution of j_0 to the right-hand-side of (22), which only affects the first and the second term.

Take now $j_0\in NF\cap O(S)$. In this case, the probability that j_0 is operative is 1. For this reason, the weight of λ_{j_01} in A is just q. Now, the probability of one scenario where one particular facility fails and all the others are operative is still $q(1-q)^{t-1}$, but the probability of any scenario where more than one facility fails is $q-\left(1-(1-q)^{t-1}\right)$

Following the same reasoning as before, now we have:

$$\begin{aligned}A_{j_0}&=\sum_{k\in F}\lambda_{j_01}q(1-q)^{t-1}+\lambda_{j_01}q\left(1-(1-q)^{t-1}\right)\\&\geqslant\sum_{k\in F}\epsilon_{j_0k}q(1-q)^{t-1}+\lambda_{j_01}q\left(1-(1-q)^{t-1}\right).\end{aligned}\quad (24)$$

Again, this last expression bounds the contribution of assignments at level $r=1$ to j_0, and equals the part of the right-hand side of (22) corresponding to j_0, which now only affects the first and the third term.

□

To illustrate this bound, let us take once more Example 2 (see Figure 3) and the solution with $Y_{1A0} = Y_{2A0} = Y_{3A1} = 1$, $Y_{3B0} = Y_{1B1} = Y_{2B1} = Y_{4B1} = Y_{5B1} = 1$, $Y_{4C0} = Y_{5C0} = Y_{3C2} = 1$. Recall that in this solution the only facility where overload occurs is B, and the total expected overload is $E(X,Y) = 2(1-q)q^2$.

Since all three facilities $A, B, C \in F$, in this solution $t = 3$. As it happened with E_1, for this solution, the only nonzero capacity slacks are $\nu_{A1} = 1$, $\nu_{B1} = 2$, $\nu_{C1} = 1$, $\nu_{C2} = 1$ and the only nonzero overload is $\lambda_{B1} = 2$. However, if only facility A fails, on top of its 0-level assignments, B receives the demand of customers 1 and 2, so that $\epsilon_{BA} = (2-2)^+ = 0$. In a similar way, $\epsilon_{CA} = 0$. For $j \neq B$ it is straightforward to see that $\epsilon_{jk} = 0$ for any k. Thus, the value of E_2 for this solution is $E_2(X,Y) = 0q(1-q)^2 + 0q(1-q)^2 + 2q(1-q)\left(1-(1-q)^1\right) = 2q^2(1-q) = E(X,Y) < E_1(X,Y)$.

Bounding $E_2(X,Y)$ as it was done for $E_1(X,Y)$ with constraints (16)-(20) involves an extra complexity, since for computing this new bound the number of open failing facilities t, is required. This might be done using a discretized formulation. For simplicity, here we only present the constraints that would be required to do so if $NF = \{u\}$ and a cardinality constraint on the set of open facilities was imposed:

The computation of λ and ν variables is equal to the case of bound E_1. Therefore, constraints (16)-(19) and (20) are again required. Additionally, to limit the value of E_2, the following constraints are also needed

$$\sum_{j \in J \setminus \{u\}} X_j = t \qquad (25)$$

$$D_{ijk} \geqslant h_i Y_{ij1} - h_i M(1 - Y_{ik0}) \qquad i \in I, j \in J \qquad (26)$$

$$\epsilon_{jk} \geqslant \sum_{i \in I} D_{ijk} - \nu_{j1} \qquad \forall j \in J, k \in F \qquad (27)$$

$$\sum_{j \in J} \sum_{k \in F} q(1-q)^{t-1} \epsilon_{jk} + \sum_{j \in F} q(1-(1-q)^{t-2})(1-q)\lambda_{j1}$$
$$+ \sum_{j \in NF} q(1-(1-q)^{t-1})\lambda_{j1} + \sum_{r>1} \left[\sum_{j \in F} q^r (1-q) \lambda_{jr} + \sum_{j \in NF} q^r \lambda_{jr} \right] \leqslant V \qquad (28)$$

$$D_{ijk}, \epsilon_{jk} \geqslant 0 \qquad i \in I, j \in J, k \in F. \qquad (29)$$

In what follows, we will refer to the resulting formulation as CRFLP-B2(V, t) where B2 stands for bound of Type 2. Here, variables D_{ijk} take the value of the demand h_i only for customers whose 0-level assignment is to k, and whose 1-level assignment is j. For all other customers, i, $D_{ijk} = 0$. This variable is used to compute the value of ϵ_{jk} in constraint (27). Finally, constraint (28) imposes a limit on the value of E_2 and (29) set the domains of the new variables.

Note that, even this simplification of the model, where the number of failing facilities to open is fixed, the computational burden associated with limiting bound E_2 is rather large. Indeed, a large number of variables and constraints need to be added to the formulations. Therefore, the general case where this limit is not imposed, which would require using several replicas of some of the variables, is left out of the scope of this work.

Furthermore, the idea of splitting the customers assigned to a facility at a certain level according to their previous assignments pattern might be extended to levels $r > 1$. However, we believe that the complexity that this would involve would make this approach impracticable in realistic size instances.

The two last proposals limit the value of an upper bound on $E(X, Y)$ to avoid expected overload overflows. This is quite a conservative strategy, specially if the used bound can be rather loose in some cases, as it happens with $E_1(X, Y)$. Indeed, in this work, limiting the expected overload is only considered as a means of keeping possible overloads under control. Therefore, the imposed limits V are quite subjective and, in general, they do not correspond to a physical limitation. For this reason, we next propose to use an approximation of $E(X, Y)$ that, without being an upper bound, gives a closer estimate of its real value, yielding a less conservative model.

3.3 A linear estimate of the expected overload

In order to obtain a linear estimate of the expected overload, we propose here to fit a linear regression model using the data obtained on a large set of randomly generated instances. For each of these instances, we obtained several solutions, using all the formulations presented in this work with different parameter settings and, for each solution, we computed the corresponding expected overload together with a series of solution characteristics, $\lambda_{\bullet 1}, \ldots, \lambda_{\bullet 4}$, that can be linearly computed as:

$$\lambda_{\bullet r} = \sum_{j \in J} \lambda_{jr} \tag{30}$$

We also computed the average values

$$\overline{\lambda}_{\bullet r} = \lambda_{\bullet r} / \sum_{j \in J \setminus \{u\}} X_j.$$

After analyzing different model proposals, the linear approximation that yielded the best fit ($R^2 = 0.9748$) was

$$\hat{E}(X, Y) = 2.67827q\overline{\lambda}_{\bullet 1} + 1.66348q^2\overline{\lambda}_{\bullet 2} + 1.92325q^3\overline{\lambda}_{\bullet 3} + 4.43350q^4\overline{\lambda}_{\bullet 4} \tag{31}$$

The details on the used instances and the model validation are gathered in Appendix A.

Therefore, the fourth model we propose in this work, that from now on will be referred to as CRFLP-LR(V) (LR stands for Linear Regression) can be formulated by adding to formulation QRFLP constraints (16)-(18) and (20) to set the right values to the λ_{jr} variables, together with:

$$\lambda_{\bullet r} = \sum_{j \in J} \lambda_{jr} \qquad r \in \{1, \ldots, 4\} \tag{32}$$

$$2.67827q\lambda_{\bullet 1} + 1.66348q^2\lambda_{\bullet 2} + 1.92325q^3\lambda_{\bullet 3} + 4.43350q^4\lambda_{\bullet 4} \leqslant V \sum_{j \in J \setminus \{u\}} X_j \tag{33}$$

Note that the complexity of this formulation is similar to that of formulation CRFLP-B1 since, basically, here the upper bound E_1 is replaced by $\hat{E}$, but their computation is pretty similar. However, we expect to obtain different solutions by solving either model, since in the first case the bound can be quite loose, so that overloads will be heavily constrained, while in the second one, where only an approximation is used, the limit V established in the formulation might be even exceeded in a solution, but, in general, the estimate and the real value of the expected excess should be much closer to each other. Therefore, in this second case, the choice of the value of the limit V is much more meaningful.

3.4 Using staggered capacities

All the models proposed previously are rather involved. This is partly due to the need of adding several extra variables to estimate or bound the expected overloads associated with a given solution. In fact, by constraining any of these bounds and estimates, we obtain solutions where the overall amount of demand assigned to a given facility is small enough. Indeed, this overall demand is computed by weighting the assigned demands by smaller factors as the assignment level r increases, to account for the smaller probability that those assignments eventually yield an actual service request.

Intuitively, this can be interpreted as the possibility of assigning a facility amounts of demand larger than its capacity, as long as those assignments are made at high assignment levels. The rationale behind the last model we propose in this section is to use this idea to build auxiliary capacities for the facilities to bound the total amount of demand assigned to them up to a given level r. According to the ideas presented so far, these capacities should increase with r. Therefore, we propose to scale capacities Q_j with different powers of a scale factor $\beta > 1$ and replace constraints (6) with the following set:

$$\sum_{s=0}^{r} \sum_{i \in I} h_i Y_{ijs} \leqslant \beta^r Q_j \qquad j \in J, r \geqslant 1 \tag{34}$$

The formulation obtained by doing so will be denoted CRFLP-S(β) where S indicates Staggered.

CRFLP-S(β) is simpler to state than the three previously proposed models. Nonetheless, we expect that by suitably choosing the scale factor β it can provide high quality solutions as well.

In the next section we compare the performance of the four proposed models with different parameter settings, and analyze the solutions they provide.

4 Computational experiments

In this section we present the results obtained from the computational experience. We have generated several instances from the capacitated p-median instances available at the OR-LIBRARY (Beasley [1990]). In particular, we have generated a total of 400 CRFLP instances, divided in three different sets, based on some of the original p-median instances. The characteristics of the set of instances are given in Table 1. In all cases, the first locations of the original p-median instance have been taken to act both, as customers and as potential facility locations, and the following ones are taken exclusively as potential facility locations. Also, in the objective function (1), a weighting factor $\alpha = 0.5$ has been used and the outsourcing cost has been fixed to $\rho = 400$. In Table 1, column under heading $\sharp$ gives the number of instances of each group, and the following columns give the number of customers, the number of failing and non-failing facilities (excluding the dummy one) and the failure probability q. Regarding the facility set-up costs, we only considered two different costs in each instance; f_F for each of the facilities in F, and f_{NF} for those in NF(except for the dummy one). In the instance groups where several values appear for the same characteristic, all possible combinations have been used.

	$\sharp$	Customers	|F|	$\|NF\|^{(*)}$	q	f_F $(\times 1000)$	f_{NF}/f_F
S20_50_a	180	20	50	0	0.05, 0.10, 0.20	1, 2, 3	1, 2
S20_50_b	180	20	35	15	0.05, 0.10, 0.20	1, 2, 3	1, 2
S50_50_a	10	50	50	0	0.05	2	2
S50_50_b	10	50	35	15	0.05	2	2
S50_75_a	10	50	75	0	0.05	2	2
S50_75_b	10	50	45	30	0.05	2	2

$^{(*)}$: excluding dummy

Table 1: Generated instances

Our experiments were conducted on a PC with a 2.33 GHz Intel Xeon dual core processor, 8.5 GB of RAM, and operating system LINUX Debian 4.0. We use the optimization engine CPLEX v11.0.

We next compare the performance of the different models we have proposed for dealing with facility capacities. Namely, we compare the basic model, QRFLP with CRFLP-LEL(V,γ), CRFLP-B1(V), CRFLP-LR(V) and CRFLP-S(β). Additionally, we have considered different parameter values for each model to evaluate their impact on the solution, as can be seen in Tables 2-6. We do not give comparisons with CRFLP-B2(V, t) since it is the only model where the number of open plants is fixed beforehand. Moreover, in the case of model CRFLP-B1(V) we have replaced (19) with

$$\sum_{j\in F}\sum_{r\leq \ell} q^r(1-q)\lambda_{jr} + \sum_{j\in NF}\sum_{r\leq \ell} q^r\lambda_{jr} \leq V, \tag{35}$$

which only accounts for assignments up to a fixed level ($\ell = 4$ in our experiments) to approximate the upper bound $E_1(X,Y)$ of the total expected overload. This was done this way since using inequality (19) is highly time consuming and this truncation varies very slightly the estimate of the expected overload as we prove in Appendix B.

The following tables show the averages over the corresponding sets of instances of five measures of the solution: v^* stands for the optimal value, $E(X,Y)$ stands for the value of the expected overload (14) in the optimal solution, $\mathbb{P}$*(overload)* represents the probability of having overload computed as the sum of the probabilities of all the scenarios with some positive overload, *Dummy* represents the expected demand at the dummy facility, i.e., the sum of the demands that the dummy facility receives over all the possible scenarios weighted by the probability of the scenario. Finally, *Time* is the time in seconds for solving the problem.

	v^*	$E(X,Y)$	$\mathbb{P}$(overload)	Dummy	Time
QRFLP	8997.20	5.19	0.07	0.26	7.75
CRFLP-LEL(0,$\lvert J\rvert$)	9019.43	4.99	0.07	0.32	29.63
CRFLP-LEL(1,$\lvert J\rvert$)	9010.38	5.13	0.07	0.30	11.94
CRFLP-LEL(2,$\lvert J\rvert$)	9004.69	5.17	0.07	0.27	10.88
CRFLP-LEL(∞,1)	8997.20	5.19	0.07	0.26	7.54
CRFLP-LEL(∞,2)	8997.20	5.19	0.07	0.26	7.45
CRFLP-B1(3)	9378.19	1.64	0.06	1.92	30.20
CRFLP-B1(6)	9143.23	3.90	0.07	0.94	32.33
CRFLP-LR(3)	9287.02	2.53	0.07	1.68	31.71
CRFLP-LR(6)	9051.44	4.65	0.07	0.56	17.82
CRFLP-S(1.1)	9536.01	0.46	0.04	2.39	59.41
CRFLP-S(1.2)	9417.45	1.03	0.05	1.87	42.17
CRFLP-S(1.3)	9327.98	1.75	0.06	1.52	39.59

Table 2: Average values for S20_50 instances with $q = 0.05$

In Tables 2-5 we report the results on the set of smaller instances for failure probabilities $q = 0.05$, $q = 0.10$, and $q = 0.20$, respectively. Results in Table 6 refer to the medium-size instances; in this case, we considered $q = 0.05$. Column *Solved* in Tables 6 and 7 refers to the number of solved instances among the 20 considered within the time limit which is set to 3600 seconds. Note that results regarding instances with and without non-failing facilities have been aggregated in all cases except in the smaller instances with $q = 0.1$ (Tables 3 and 4) since the same behavior was consistently observed.

	v^*	$E(X,Y)$	$\mathbb{P}$(overload)	Dummy	Time
QRFLP	9355.96	10.26	0.15	1.22	5.05
CRFLP-LEL(0,$\lvert J\rvert$)	9444.94	8.84	0.14	1.31	163.68
CRFLP-LEL(1,$\lvert J\rvert$)	9436.18	8.92	0.15	1.27	128.17
CRFLP-LEL(2,$\lvert J\rvert$)	9429.87	9.07	0.15	1.24	69.22
CRFLP-LEL(∞,1)	9355.96	10.26	0.15	1.22	5.38
CRFLP-LEL(∞,2)	9355.96	10.26	0.15	1.22	5.25
CRFLP-B1(3)	10294.36	1.49	0.07	3.02	112.30
CRFLP-B1(6)	9879.32	3.96	0.12	1.82	84.77
CRFLP-LR(3)	10172.13	2.20	0.09	2.56	403.62
CRFLP-LR(6)	9849.81	4.30	0.12	1.75	72.72
CRFLP-S(1.1)	110349.16	0.81	0.07	3.52	315.23
CRFLP-S(1.2)	10040.49	2.14	0.10	2.29	123.55
CRFLP-S(1.3)	9863.07	3.41	0.11	1.70	82.10

Table 3: Average values for S20_50_a instances with $q = 0.1$

	v^*	$E(X,Y)$	$\mathbb{P}$(overload)	Dummy	Time
QRFLP	9187.60	7.18	0.10	0.61	8.78
CRFLP-LEL(0,$\|J\|$)	9243.07	6.03	0.10	0.65	88.80
CRFLP-LEL(1,$\|J\|$)	9236.05	6.23	0.10	0.67	90.63
CRFLP-LEL(2,$\|J\|$)	9230.16	6.31	0.10	0.64	78.32
CRFLP-LEL(∞,1)	9187.60	7.18	0.10	0.61	9.23
CRFLP-LEL(∞,2)	9187.60	7.18	0.10	0.61	9.28
CRFLP-B1(3)	9690.77	0.90	0.05	1.54	48.52
CRFLP-B1(6)	9473.40	2.47	0.07	0.91	43.72
CRFLP-LR(3)	9625.14	1.27	0.05	1.28	84.72
CRFLP-LR(6)	9455.83	3.31	0.08	0.96	38.28
CRFLP-S(1.1)	19727.21	0.50	0.04	1.82	168.18
CRFLP-S(1.2)	9565.06	1.22	0.06	1.19	55.93
CRFLP-S(1.3)	9472.12	1.85	0.06	0.86	43.42

Table 4: Average values for S20_50_b instances with $q = 0.1$

First of all, from the comparison of the results in Tables 3 and 4 we can see that the possibility of including in the system fully reliable facilities allows to reduce the overloads in all cases, both, in terms of its expected value, and of the probability that they occur, at the same time that it allows to reduce the outsourced demand (assigned to the dummy facility). In terms of computational burden, with the only exception of the basic model QRFLP, where the only scenario where capacities arise is the one with no failures, solving problem instances with fully reliable facilities results in smaller computation times. We further analyzed the results used to build these tables and, in the case of the instances in Table 4, the number of opened facilities in the optimal solution ranges between 2 and 6 and, among them, between 0 and 4 belong to NF, therefore, it becomes clear that, depending on the cost structure, and on the tightness of capacity constraints, it might pay or not to open facilities in NF. As for the total solution costs, we believe that the smaller costs observed in S20_50_b instances are strongly related with the cost structure of the generated instances.

Analyzing Tables 2-5, we can see that QRFLP obviously yields the cheapest solutions and the ones with larger overloads and overload probabilities. Also, the demand received at the dummy facility is smaller in QRFLP solutions for small failure probabilities ($q = 0.05$ and $q = 0.1$) although it takes intermediate values when $q = 0.2$. QRFLP is also nearly always the one which requires the least computational time (except for $q = 0.05$ where some CRFLP-LEL variants are solved slightly faster). Thus, reducing the overloads as done by the other models increases both, the solution costs and the computational burden. According to Tables 2-5 model CRFLP-S(1.1) is the one yielding the smallest overloads and overload probabilities and in contrast it is the one with largest solution costs and larger demands for the dummy facility: model CRFLP-S(1.1) strongly forces the reduction of overloads in exchange of losing clients and increasing the investment. Models CRFLP-B1(3) and CRFLP-LR(3) are also effective for reducing the expected overload and the overload probability as compared with QRFLP and they lose less clients. For each block of rows in Tables 2-5, form top to bottom the parameters go from stricter to more relaxed: for CRFLP-LEL(V,$|J|$) which represents the situation in which all the facilities in the system may receive demands whose expected value exceeds their

	v^*	$E(X,Y)$	$\mathbb{P}$(overload)	Dummy	Time
QRFLP	9995.34	12.73	0.21	2.44	7.25
CRFLP-LEL(0,$\|J\|$)	10066.36	8.63	0.17	1.58	68.19
CRFLP-LEL(1,$\|J\|$)	10064.61	8.69	0.18	1.58	71.06
CRFLP-LEL(2,$\|J\|$)	10063.03	8.71	0.18	1.58	83.51
CRFLP-LEL(∞,1)	9995.34	12.72	0.21	2.44	7.49
CRFLP-LEL(∞,2)	9995.34	12.72	0.21	2.44	7.43
CRFLP-B1(3)	10975.59	1.27	0.05	2.63	236.17
CRFLP-B1(6)	10512.67	4.60	0.12	1.62	954.25
CRFLP-LR(3)	10636.21	3.48	0.09	1.16	744.68
CRFLP-LR(6)	10503.05	4.49	0.11	1.55	944.17
CRFLP-S(1.1)	10879.70	1.30	0.07	3.38	1252.76
CRFLP-S(1.2)	10451.35	3.68	0.14	2.29	344.71
CRFLP-S(1.3)	10215.46	5.76	0.15	1.41	49.83

Table 5: Average values for S20_50 instances with $q = 0.20$

capacities, $V = 0$ is the situation in which no facility can receive any excess, $V = 1$ the situation in which each facility can receive an overload of at most one unit, and $V = 2$ for two units; for CRFLP-LEL(∞, γ) which represents the situation in which exactly γ facilities can receive demands whose expected value may exceed their capacities by any quantity, the option $\gamma = 2$ only allows to freely allocate to two facilities and the option $\gamma = 3$ allows to freely allocate to three facilities.

Models CRFLP-B1(3) and CRFLP-B1(6) impose that the upper bound of the expected overload is smaller than 3 and 6, respectively while models CRFLP-LR(3) and CRFLP-LR(6) limit to the same values the linear estimate of this overload. Model CRFLP-S(1.1) is also more exigent than CRFLP-S(1.2) which, in turn, is more restrictive than CRFLP-S(1.3). So, at each block when reading the values from top to bottom, optimal values decrease, expected overloads and overload probabilities increase and dummy demands decrease. Considering the most permissive options of each block (CRFLP-LEL(2,$|J|$), CRFLP-LEL(∞,2) CRFLP-B1(6), CRFLP-LR(6) and CRFLP-S(1.3)), the options CRFLP-LEL(2,$|J|$) and CRFLP-LEL(∞,2) only reduce the overload of QRFLP slightly. Comparing CRFLP-B1 with CRFLP-LR both, with $V = 3$ and $V = 6$, model CRFLP-LR(V) proves to take advantage of the goodness of fit of our estimate of the expected overloads; actual expected overloads seldom exceed the requirements (only in the case with larger demand probability we observe 3.48>3) and they are much closer to the requirements than they are with model CRFLP-B1(V): the expected overloads with CRFLP-B1(3) are always smaller than the expected overloads with CRFLP-LR(3) and the same happens with CRFLP-B1(6) and CRFLP-LR(6). This is consistent with the design of the models; recall that CRFLP-B1 limits de value of an upper bound of the expected overload, while what is limited in the case of CRFLP-LR is a linear approximation of this expected overload, which can be either an underestimate or an overestimate. This should be taken into accout by decision makers before setting the value of V in either model.

In terms of computational time all the models perform quite well: when $q = 0.05$ or $q = 0.1$ nearly all the instances require less than four minutes. Even when $q = 0.2$ computational times for models CRFLP-B1(V), CRFLP-LR(V) and CRFLP-S(β), although being larger, are still affordable. Finally, it can be seen in Tables 2-5 that when q increases all the measures taken into account change, but not uniformly. Of course, the optimal value increases uniformly with the value of q but this does not hold for the expected overload, the overload probability, the dummy demand or the computational time. In most cases the expected overload, the overload probability and the dummy demand increase with q and in half of the cases the increase in q implies an increase in computational time. Summarizing, if we consider only the above results we shall advice the use of CRFLP-LR(V) or CRFLP-S(β) depending on the importance we give to each measure.

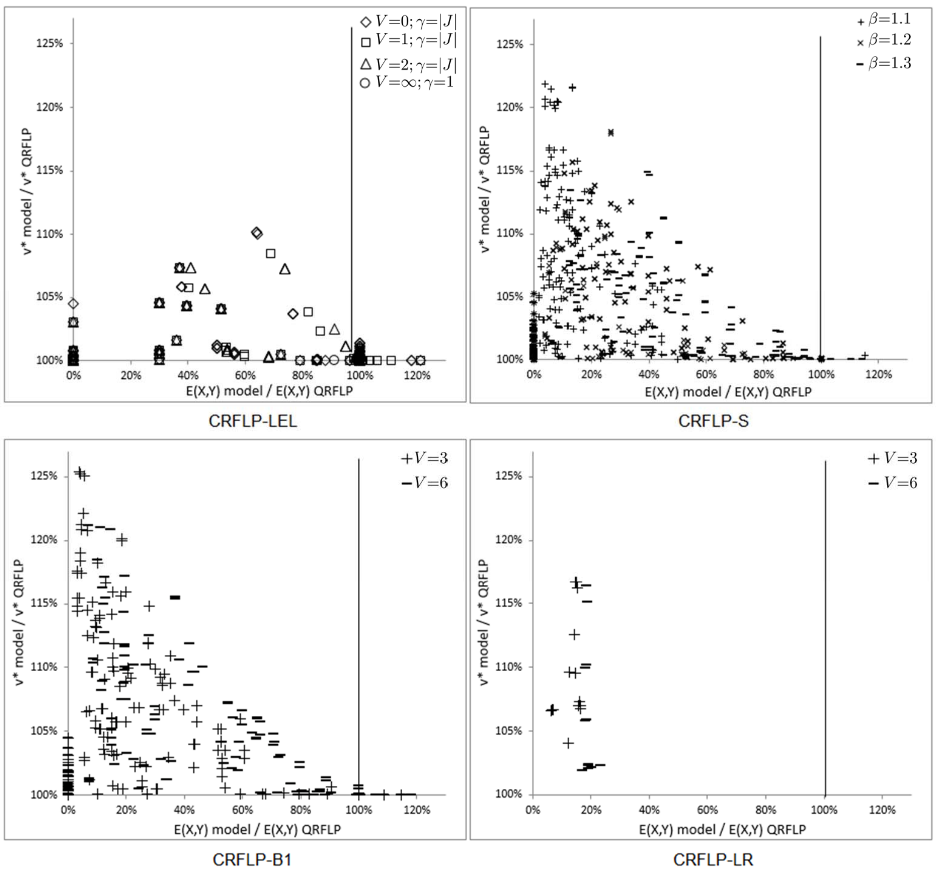


Figure 4: Relative cost increase vs. relative expected overload reduction

In order to learn more about the four different models and the sensitivity to their parameter values, among the five measures in Tables 2-5 we further analyze the optimal value and the expected overload. To this end, in Figure 4, we plot one point for each of the smaller 360 instances as follows. For each obtained solution we have computed the ratio between its value and the optimal QRFLP value for the same instance, and the ratio between its expected overload, and the expected overload of the optimal QRFLP solution. The figure shows the expected overload's ratio against the optimal value's ratio. For each of the four models CRFLP-LEL(V,γ), CRFLP-B1(V), CRFLP-LR(V) and CRFLP-S(β) we have considered the same parameters as in Tables 2-5 except CRFLP-LEL(∞,2) because all the corresponding points overlap those of CRFLP-LEL(∞,1) and most of them coincide at point (100,100). Point $(100, 100)$ represents an instance for which the new model obtains the same expected overload as QRFLP at the same cost, i.e., for that instance, either the new model does not contribute to our goal, or the capacity is not binding. Points to the left of the vertical straight line at 100 correspond to instances where we reduced the overload, and their ordinate indicates the relative cost of doing so. Points to the right of the vertical straight line in 100 represent instances for which although some constraint aiming at reducing the overload was imposed, it has increased, possibly because we have not allowed enough margin. Points on the vertical axis represent situations in which we have totally reduced the expected overload; the closer they are to the intersection of the axis, the the smaller is the cost increase of the new solution with respect to the QRFLP solution. Points on the horizontal axis represent instances with no increase in the optimal value. Ideally, one would like to have points near the intersection of the axes, which represent high reduction of the expected overload and small cost increase. Considering the different symbols for the different parameters we observe that, as expected, as we become more strict, the points tend to move closer to the vertical axis. The cloud of points for model CRFLP-LEL(V,γ) lies below 110% of cost, but it is not close to the intersection of the axes. The cloud of points for models CRFLP-B1(V) and CRFLP-S(β) are quite similar and promising because they try to concentrate in the intersection of the axes. The cloud of points for model CRFLP-LR(V) is much more structured because all points concentrate on few locations and they always show an overload ratio below 30% and cost ratio below 120%. Thus, model CRFLP-LR(V) performs homogenously and is consistently successful in reducing the expected overloads. This illustrates the goodness of the approximation used for the expected overloads. The information in these graphs evidences the superiority of CRFLP-LR(V), which successfully achieves the goal of the additional constraints. Furthermore, the required parameter V has a direct interpretation.

In Table 6 we report the results for the medium-size instances. Here we can observe the same behavior as in Tables 2-5, but more exaggerate. Again, QRFLP yields the cheapest solutions, which have the largest expected overloads and overload probabilities. Moreover, it again gives intermediate values of demand allocated to the dummy facility and of CPU times. Model CRFLP-S(1.1) is the one with smaller expected overloads and overload probabilities and, consequently, it is again the one with the largest solution values which are once more related to large demands for the dummy facility. Models CRFLP-LEL(2,$|J|$) and CRFLP-LEL(∞,2) reduce only slightly the expected overload of QRFLP. Comparing CRFLP-B1(3) with CRFLP-LR(3) (analogously CRFLP-B1(6) with CRFLP-LR(6)) model CRFLP-LR(V) seems to use a more accurate estimate of $E(X, Y)$, since the deviations of this expected overload from the limit set in the model are much smaller in this case ($|3-2.33| > |3-3.31|$ and $|6 - 4.71| > |6 - 6.41|$). Apparently, model CRFLP-S(1.3) tends to give better solutions than CRFLP-LR(6) since, on the average, solution costs are smaller as well as expected overloads and demand allocated to the dummy facility, while the overload probability is almost the same. However, if we become more restrictive the small advantage of the solution costs obtained with CRFLP-S(1.2) with respect to those obtained with CRFLP-LR(3), comes along with an increase of both, the expected overload and the overload probability. As for the CPU requirements, model CRFLP-S is much more demanding than the rest, followed by CRFLP-LR which, surprisingly, is much more expensive than CRFLP-B1. Summarizing, if

	v^*	$E(X,Y)$	$\mathbb{P}$(overload)	Dummy	Time	Solved
QRFLP	17955.67	22.06	0.23	0.00	719.03	18
CRFLP-LEL(0,$\lvert J\rvert$)	18024.36	19.35	0.23	0.00	800.75	19
CRFLP-LEL(1,$\lvert J\rvert$)	18015.69	19.43	0.23	0.00	751.15	20
CRFLP-LEL(2,$\lvert J\rvert$)	18009.50	19.54	0.23	0.00	728.45	20
CRFLP-LEL(∞,1)	17955.67	22.00	0.23	0.00	716.95	18
CRFLP-LEL(∞,2)	17955.67	22.00	0.23	0.00	703.05	18
CRFLP-B1(3)	21428.69	2.33	0.15	16.37	187.00	20
CRFLP-B1(6)	20870.85	4.71	0.19	13.37	215.00	20
CRFLP-LR(3)	21198.76	3.31	0.16	15.12	1152.60	18
CRFLP-LR(6)	20449.32	6.41	0.21	11.07	1188.75	16
CRFLP-S(1.1)	21357.77	1.75	0.15	14.66	2514.90	10
CRFLP-S(1.2)	20682.61	3.40	0.20	8.97	2554.80	11
CRFLP-S(1.3)	20016.06	5.07	0.22	3.75	3600.00	0

Table 6: Average values for S50_50 instances with $q = 0.05$

computational time is not an issue, model CRFLP-S may be used to obtain good solutions both in terms of cost and capacity utilization. However, in general, CRFLP-LR gives the best compromise between computational effort and solution quality.

In order to have an idea of how the previous conclusions extend to larger instances, we have chosen one parameter for each of the three best performing models (CRFLP-B1(3),CRFLP-LR(3) and CRFLP-S(1.2)) and compared their behavior on the set of larger instances. The obtained results are summarized in Table 7

	v^*	$E(X,Y)$	$\mathbb{P}$(overload)	Dummy	Time	Solved
CRFLP-B1(3)	11090.57	1.32	0.10	2.44	1412.03	19
CRFLP-LR(3)	10885.10	2.57	0.12	1.91	1337.10	17
CRFLP-S(1.2)	10737.63	2.35	0.13	2.09	1287.10	17

Table 7: Average values for S50_75 instances with $q = 0.05$

According to the results obtained with this last set of instances, the large differences in computation time observed for the small instances tend to become smaller. Indeed, CRFLP-B1 now takes larger times than the other two methods, while it was the least demanding one for smaller instances. As for the obtained solutions, again, when the same parameter is used CRFLP-LR becomes less demanding than CRFLP-B1, and the staggered model obtains comparable solutions.

In general all the proposed models, except, maybe, for CRFLP-LEL, are quite effective at reducing expected overloads and overload probabilities. Finally, to illustrate how the overload distribution varies in different solutions, in Figure 5 we have represented the distribution function of the overload for the optimal solution of the fourth instance in S50_50_b corresponding to models QRFLP and CRFLP-LR(6). The solution of QRFLP has an overload probability equal to 0.23 and an expected overload equal to 19.62 while the solution of CRFLP-LR(6) has larger overload probability 0.26 and smaller expected overload 5.40. Light grey bars correspond to the solution of QRFLP and dark grey bars to the solution of CRFLP-LR(6). We

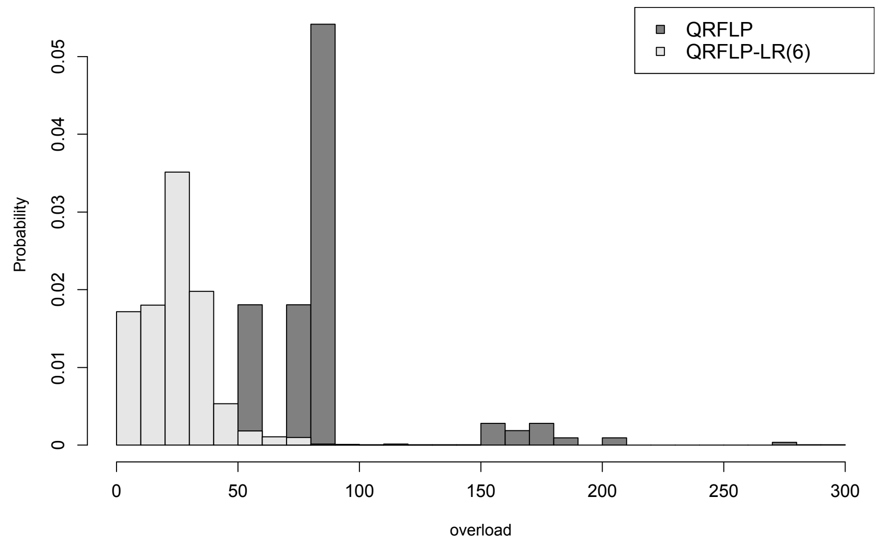


Figure 5: Probability distribution of the "overload" random variable in two different solutions

have not drawn the bar for overload equal to zero because it has a huge probability with respect the others. We observe that the mode for CRFLP-LR(6) is 25 while it is 85 for the QRFLP. Likewise, the interquartile range for CRFLP-LR(6) is on the left of the interquartile range for QRFLP. If we also draw the overload probability distribution for CRFLP-LEL, CRFLP-B1 or CRFLP-S they are similar to the light gray one. So, Figure 5 illustrates how the optimal solution changes, in terms of overloads when we impose the kind of constraints that we have introduced in this paper.

5 Conclusions

In this paper we have proposed and analyzed four alternative models for including capacity constraints in the reliable facility location problem, namely CRFLP-LEL, CRFLP-B1, CRFLP-LR and CRFLP-S. The proposed models are obtained following different rationales, although they all pursue the same goal: to minimize the solution costs while keeping facility capacity overloads and the probabilities that they occur small.

RFLP-LEL focuses on the expected value of the demands allocated to each facility and, depending on the used parameters, allows these expected values to exceed the capacity by a limited amount, on a limited number of facilities. In contrast, the intention in both, CRFLP-B1 and CRFLP-LR is to limit the expected overload value. Since this expected overload is highly nonlinear on the decision variables, they use two different linear estimates of its value; CRFLP-B1 uses an estimate which we proved to be an upper bound, and CRFLP-LR uses a linear approximation found by using linear regression. Finally, CRFLP-S is based on a completely different idea, and simply uses auxiliary capacities that become larger as higher assignment levels are considered, accounting for the smaller probabilities that these assignments yield actual service requests.

In a series of computational experiments we have analyzed the evolution of the solutions obtained with the different models and using different parameters, and compared them with the solution of the problem with capacity constraints only for the scenario where no facility fails. We have shown that, in many cases, it is possible to obtain relevant reductions of the expected overloads without incurring dramatic cost increases, which in our opinion, is a strong motivation to use these models. Regarding the comparison of the models, CRFLP-LR and CRFLP-S have provided the most promising results. On the one hand, CRFLP-LR uses quite an accurate estimate of the expected overload and, therefore, when the corresponding capacity constraints are binding, the actual expected overloads are quite close to the imposed limits, so that the imposed limits have a direct interpretation. In contrast, the more demanding model, CRFLP-B1, uses an upper bound that in some occasions is rather loose, forbidding thus solutions that might be indeed interesting for the decision maker. On the other hand, CRFLP-S yields solutions with similar expected overloads to those obtained with CRFLP-B1, but associated with solutions with slightly smaller costs. Unfortunately, this model becomes computationally expensive as the number of customers increases. Therefore, CRFLP-LR can be regarded as being superior to the others in terms of usability: it obtains high quality solutions with affordable computational times.

We consider to develop heuristic methods for solving larger instances as a future research line. Additionally, since the idea of controlling the expected overloads has provided very interesting solutions in this case, we also plan to extend this idea to other reliability location models.

Acknowledgments

This work was partly supported by the Spanish Ministry of *Economy and Competitiveness* through grants MTM2015-68097-P(MINECO/FEDER) and MTM2015-63779-R.

References

R. Aboolian, T. Cui, and J.-J. Shen. An efficient approach for solving reliable facility location models. *INFORMS Journal on Computing*, 25:720–729, 2013.

M. Albareda-Sambola, Y. Hinojosa, and J. Puerto. The reliable p-median problem with at-facility service. *European Journal of Operational Research*, 245:656–666, 2015.

J. Alcaraz, M. Landete, and J. F. Monge. Design and analysis of hybrid metaheuristics for the reliability p-median problem. *European Journal of Operational Research*, 222(1):54–64, 2012.

J. Alcaraz, M. Landete, J. F. Monge, and J. L. Sainz-Pardo. Strengthening the reliability fixed-charge location model using clique constraints. *Computers and Operations Research*, 60(1):14–26, 2015.

J. Alcaraz, M. Landete, J. F. Monge, and J. L. Sainz-Pardo. On relaxing the integrality of the allocation variables of the Reliability Fixed-Charge Location Problem. *Journal of Global Optimization*, DOI: 10.1007/s10898-016-0439-z.

Y. An, B. Zeng, Y. Zhang, and L. Zhao. Reliable p-median facility location problem: two stage robust models and algorithms. *Transportation Research Part B*, 64:54–72, 2014.

N. Aydin and A. Murat. A swarm intelligence based sample average approximation algorithm for the capacitated reliable facility location problem. *International Journal of Production Economics*, 145:173–183, 2013.

N. Azad, H. Davoudpour, G. Saharidis, and M. Shiripour. A new model for mitigating random disruption risks of facility and transportation in supply chain network design. *International Journal of Advanced Manufacturing Technology*, 70:1757–1774, 2014.

J. E. Beasley. Or-library: distributing test problems by electronic mail. *Journal of the Operational Research Society*, 41:1069–1072, 1990. URL `http://www.brunel.ac.uk/~mastjjb/jeb/info.html`.

O. Berman, R. Huang, S. Kim, and M. Menezes. Locating capacitated facilities to maximize captured demand. *IEE Transactions*, 39:1015–1029, 2007a.

O. Berman, D. Krass, and M. Menezes. Facility reliability issues in network p-median problems: Strategic centralization and co-location effects. *Operations Research*, 55(2):332–350, 2007b.

O. Berman, D. Krass, and M. Menezes. Locating facilities in the presence of disruptions and incomplete information. *Decision Sciences*, 40(4):845– 868, 2009.

O. Berman, D. Krass, and M. Menezes. Location and reliability problems on a line: Impact of objectives and correlated failures on optimal location patterns. *Omega*, 41:766–779, 2013.

Cui, T., Ouyang, Y., and Shen, Z. (2010). Reliable facility location design under the risk of disruptions. *Operations Research*, 58:998–1011.

Z. Drezner. Heuristic solution methods for two location problems with unreliable facilities. *Journal of the Operations Research Socieety*, 38:509–514, 1987.

I. Espejo, A. Marín, and A. M. Rodríguez-Chía. Capacitated p-center problem with failre foresight. *European Journal of Operational Research*, 247:229–244, 2015.

J. Fathali. Backup multifacility location problem with ℓ_p norm. *OPSEARCH*, 52(2):382–391, 2015.

D. Gade and E. Pohl. Sample average approximation applied to the capacitated-facilities location problem with unreliable facilities. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, pages 259–269, 2009.

S.-D. Lee and W.-T. Chang. On solving the discrete location problems when the facilities are prone to failure. *Applied Mathematical Modelling*, 31:817 – 831, 2007.

Q. Li, B. Zeng, and A. Savachkin. Reliable facility location design under disruptions. *Computers & Operations Research*, 40(4):901–909, 2013a.

X. Li and Y. Ouyang. A continuum approximation approach to reliable facility location design under correlated probabilistic disruptions. *Transportation Research Part B*, 44:535–548, 2010.

X. Li, Y. Ouyang, and F. Peng. A supporting station model for reliable infrastructure location design under interdependent disruptions. *Transportation Research, part E*, 60:80–93, 2013b.

Y.-K. Lin and P.-C. Chang. Evaluation of system reliability for a cloud computing system with imperfect nodes. *Systems Engineering*, 15(1):83–94, 2012.

K. Lim., A. Bassamboo, S. Chopra, and M. Daskin. Facility location decisions with random disruptions and imperfect estimation. *Manufacturing & Service Operations Management*, 15(2):239–249, 2013.

J. R. O'Hanley, M. P. Scaparra, and S. García. Probability chains: A general linearization technique for modeling reliability in facility location and related problems. *European Journal of Operational Research*, 230:63–75, 2013.

P. Peng, L. Snyder, A. Lim, and S. Liu. Reliable logistics networks design with facility disruptions. *Transportation Research Part B: Methodological*, 45(8):1190–1211, 2011.

J. Puerto, F. Rica, and A. Scozzari. Unreliable point facility location problems on networks. *Discrete Applied Mathematics*, 166:188–203, 2014.

X. Qin, X. Liu, and L. Tang. A two-stage stochastic mixed-integer program for the capacitated logistics fortification planning under accidental disruptions. *Computers & Industrial Engineering*, 65:614–623, 2013.

Z. Shen, R. Shan, and J. Shang. The reliable facility location problem: formulations, heuristics and approximation algorithms. *INFORMS Journal on Computing*, 23(3):470–482, 2011.

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

L. Snyder and M. S. Daskin. Reliability models for facility location: The expected failure cost case. *Transportation Science*, 39(3):400–416, 2005.

R. Zhan. *Models and algorithms for reliable facility location problems and system reliability optimization*. PhD thesis, Graduate School of the University of Florida, 2007.

Y. Zhang, O. Berman, and V. Verter. Incorporating congestion in preventive healthcare facility network design. *European Journal of Operational Research*, 198:922–935, 2009.

Appendix A

Validation of the regression model used in CRFLP-LR

For estimating the coefficients of the regression model we have used the first 5 instances with 50 nodes in the OR-LIBRARY. We have taken the number of non-failing facilities $|NF|$ ranging in $\{1, 11, 21\}$, the fixed cost of failing facilities f_F in $\{500, 1500, 2500\}$ and the fixed cost of non-failing facilities in $\{f_F, 2f_F, 3f_F\}$. We have also considered different failure probabilities, $q \in \{0.05, 0.15, 0.25, 0.35, 0.45\}$, and we have set $\alpha = 0.5$. For these combinations of parameters we have registered the values of $\lambda_{\bullet 1}, \ldots, \lambda_{\bullet 4}$ for the optimal solutions of models CRFLP-LEL($(V, \gamma) = \{(0, |J|), (3, |J|), (6, |J|), (2, 1), (2, 3)\}$), CRFLP-B1($V = \{0.5, 1\}$) and CRFLP-S($\beta = \{1.1, 1.3\}$). Note that we have used here parameters different from those used in the computational experiments to check the behavior of the model obtained in this learning phase, on different problem instances.

We have adjusted with the statistical R-Project[1] the model:

$$\hat{E}(X, Y) = \beta_1 q \lambda_{\bullet 1} + \beta_1 q^2 \lambda_{\bullet 2} + \beta_3 q^3 \lambda_{\bullet 3} + \beta_4 q^4 \lambda_{\bullet 4}. \tag{36}$$

The obtained output is shown in Figure 6. Additionally to the estimated coefficient values,

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.9413 -0.7236 -0.0099  0.3217 12.0733

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
L1    2.67827    0.01056  253.57   <2e-16 ***
L2    1.66348    0.01828   91.02   <2e-16 ***
L3    1.92325    0.03783   50.84   <2e-16 ***
L4    4.43350    0.11368   39.00   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.575 on 5561 degrees of freedom
Multiple R-squared:  0.9748, Adjusted R-squared:  0.9748
F-statistic: 5.373e+04 on 4 and 5561 DF,  p-value: < 2.2e-16
```

Figure 6: R output

it illustrates the validity of the model with $R^2 = 0.9748$ and tiny p-values.
The residual plots required to validate the model are shown in Figure 7. All of them corroborate the adequacy of the model.

Apart from the traditional analysis of residuals, we also analyzed the quality of the estimate provided by this model in the instances used in Section 4 which, as mentioned above, are different from the set of instances used to collect the data and fit the model. To this end, we have counted the number of instances in which the estimate of the expected overload has exceeded the actual expected overload over the 120 instances in Tables 2-5 and over the 20 instances in Table 6. The results of the count are given in Tables 8 and 9 respectively: for the instances with 20 customers around $40\% (= (293/720) * 100)$ of times we overestimate the expected overload, for the instances with 50 customers this percentage is $80\% (= 32/40) * 100)$.

[1] https://www.r-project.org

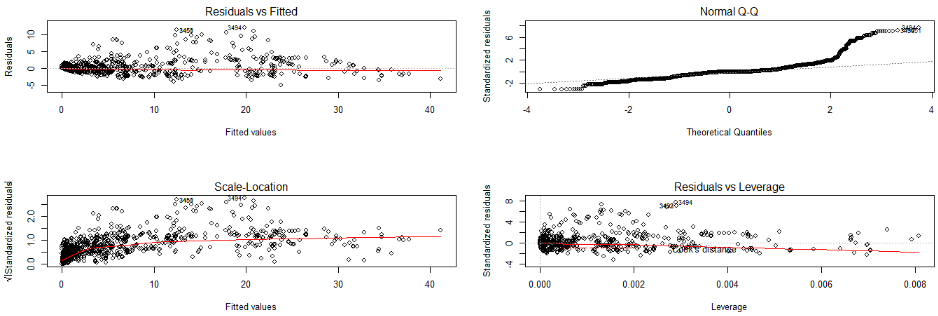


Figure 7: Residual plots

	$\beta = 3$	$\beta = 6$	Total
$q = 0.05$	61/120	51/120	112/240
$q = 0.10$	20/120	32/120	52/240
$q = 0.20$	75/120	54/120	129/240
Total	156/360	137/360	293/720

Table 8: Count of overloads in 20 customers instances

	$\beta = 3$	$\beta = 6$	Total
$q = 0.05$	16/20	16/20	32/40

Table 9: Count of overloads in 50 customers instances

Apparently these values are large, however Tables 10 and 11 show that the excess is small: on average we exceed 0.61 units of 3 and 0.44 units of 6 when solving instances with 20 customers and we exceed 0.41 units of 3 and 0.59 units of 6 when solving instances with 50 customers.

	$\beta = 3$	$\beta = 6$	Total
$q = 0.05$	0.49	0.65	0.57
$q = 0.10$	0.11	0.41	0.26
$q = 0.20$	1.21	0.26	0.74
Total	0.61	0.44	0.52

Table 10: Average overload in case of overload for 20 customers

	$\beta = 3$	$\beta = 6$	Total
$q = 0.05$	0.41	0.59	0.50

Table 11: Average overload in case of overload for 50 customers

Appendix B

Truncation error in CRFLP-B1

In order to reduce the computational effort required for solving CRFLP-B1, we can simplify this model by truncating (19) as in (35). This implies that in some cases where the expected overload limit is reached we can make an error and have an actual expected overload above the imposed limit by the amount:

$$E = \sum_{j \in F} \sum_{r > \ell} q^r (1-q) \lambda_{jr} + \sum_{j \in NF} \sum_{r > \ell}^{l} q^r \lambda_{jr} \tag{37}$$

which is bounded by:

$$E \leq \sum_{r > \ell} \sum_{j \in J} \lambda_{jr} q^r \leq q^{l-1} \sum_{i \in I} h_i.$$

In our case, we took $\ell = 4$ and we have $\sum_{i \in I} h_i \leqslant 203$ when $|I| = 20$. Thus, the maximal potential error by truncation is smaller than 0.00127, 0.0203 and 0.3248, respectively, for the values of $q = 0.05$, $q = 0.10$ and $q = 0.20$. In the other case, when we use $|I| = 50$ and $q = 0.05$, $\sum_{i \in I} h_i \leqslant 490$ then the possible truncation error is bounded by 0.003 (Recall that, in this case, we only considered $q = 0.05$).