Comparison of formulations for the inventory routing problem

Claudia Archetti *Ivana Ljubić*

Department of Information Systems, Decision Sciences and Statistics,

ESSEC Business School, Cergy-Pontoise, France

{archetti,ljubic}@essec.edu

August 4, 2021

Abstract

In this paper we propose an analysis and comparison of the strength of the lower bound, measured as the value of the linear programming relaxation, of different formulations for the Inventory Routing Problem (IRP). In particular, we first focus on aggregated formulations, i.e., formulations where variables have no index associated with vehicles, and we analyse the link between compact formulations and their counterparts involving exponentially many constraints. We show that they are equivalent in terms of value of the linear relaxation. In addition, we study the link between aggregated and disaggregated formulations, i.e., formulations where variables have an index related to vehicles. Also in this case, we show that aggregated and disaggregated formulations are equivalent in terms of the value of the corresponding linear relaxation. To the best of our knowledge, this analysis has never been done for the IRP, which instead is gaining a lot of popularity in the literature. Finally, we propose different exact solution approaches based on the aggregated formulations and we compare them with state-of-the-art exact methods for the IRP. Results show that the approaches based on aggregated formulations are competitive in terms of quality of both upper and lower bounds.

Keywords: Inventory routing; aggregated formulation; linear programming relaxation; polyhedral projection.

1 Introduction

The Inventory Routing Problem (IRP) has attracted a lot of attention from the research community in the last years. The main reason is related to the economical benefits related to the integration of transportation and inventory management. In fact, the IRP is the problem of building a distribution plan covering a given planning horizon where commodities are distributed from the supplier to a set of geographically dispersed customers who face a per period demand. A fleet of vehicles is available to perform the deliveries, and the replenishments have to be such that each customer is always capable of satisfying its demand, in each period of the planning horizon. The goal is to determine the distribution plan that minimizes the total cost, which is given by the sum of transportation (or routing) cost and inventory holding cost.

The benefits coming from integrating transportation and inventory management where analysed in Archetti and Speranza (2016) through a computational study from which it resulted that the average savings related to integration are around 24% of the total cost.

A second reason why the IRP gained such a popularity in the research community is related to the scientific challenges associated with finding its solution. In fact, the IRP combines routing, which is an extremely complex problem in itself, with inventory management, i.e., with replenishment decisions. The first exact approach for the IRP was proposed in 2007 by Archetti et al. (2007) for the single-vehicle case. More recently, the literature has focused on the multi-vehicle case. Exact approaches are mainly based on branch-and-cut schemes (see Coelho and Laporte (2013a), Coelho and Laporte (2013b), Adulyasak et al. (2014), Archetti et al. (2014), Coelho and Laporte (2014), Avella et al. (2015), Avella et al. (2018), Manousakis et al. (2021)) while just one contribution proposes a branch-and-price algorithm (see Desaulniers et al. (2016)). In the latter paper, a nice and complex decomposition of the problem is proposed. The computational results, in which the branch-and-price algorithm is compared with the branch-and-cut proposed in Coelho and Laporte (2013b), show that no method is dominating the other, with branch-and-price performing better for instances with a larger number of vehicles and viceversa. In terms of heuristic approaches, recent contributions are based on metaheuristcs (see Alvarez et al. (2018) and Santos et al. (2016)) or matheuristics (see Archetti et al. (2012), Coelho et al. (2012), Archetti et al. (2017), Chitsaz et al. (2019) and Archetti et al. (2021)). We refer the reader to Bertazzi and Speranza (2012) and Bertazzi

and Speranza (2013) for two tutorials on the IRP and to Coelho et al. (2013) and Roldán et al. (2017) for surveys on scientific contributions to the study of the IRP.

Our contribution: In this article, we study two types of formulations: (a) *aggregated* ones in which the variables describing the feasible routes and quantities delivered to customers in each time period are aggregated over all vehicles, and (b) *disaggregated* ones, in which separate variables are used to describe the routes and delivered quantities of each vehicle. For the aggregated approaches, we study compact formulations and their counterparts of exponential size. To the best of our knowledge, there are no studies in the existing literature that compare these different formulations with respect to the quality of their linear programming (LP) relaxations. The major questions we attempt to answer in our theoretical study are: does the LP-relaxation of one formulation dominates that of another, and whether two LP-relaxations provide the same lower bounds. Our results are based on the methodology of polyhedral projections which has been used for related routing and network design problems, see, e.g. Gouveia (1995); Letchford and González (2006); Ljubić et al. (2006); Chimani et al. (2010). In addition, we apply the same methodology to propose a compact aggregated formulations of the IRP variant of the so-called *multi-star inequalities*.

Contrary to the Capacitated Vehicle Routing Problems where the quantities to be delivered to each customer are pre-determined, they are part of the decisions to be made when solving the IRP. This renders the inventory routing problems more difficult to solve in practice, and it also highly affects the quality of the underlying MIP formulations.

In a computational study performed on benchmark IRP instances, we show that the aggregated formulations studied in this paper are competitive with respect to state-of-the-art exact solution approaches in terms of quality of upper and lower bounds provided.

The paper is organized as follows: In Section 2 we introduce the formal definition of the IRP and some useful notation used in the following sections. In Section 3 we introduce the aggregated IRP formulations and we show the equivalent compact formulations that can replace *fractional capacity cuts (FCC)*, *generalized subtour elimination constraints (GSEC)* and *IRP-multi-star inequalities (MS)*, respectively. In Section 4 we show the link between the LP-relaxation of aggregated and disaggregated formulations. Computational

experiments are shown in Section 6 and conclusions are drawn in Section 7.

2 Problem Definition

The IRP is defined as follows: We are given a complete directed graph $G = (N, A)$ where N is the set of vertices and A is the set of arcs. Set N is composed by vertex 0, representing the supplier (also called depot in the following), and the set N' of customers, with $|N'| = n$. The planning period is $T = \{1, ..., H\}$, where H is the planning horizon. In the following, each element of T is refereed to as 'time period'. Each customer $i \in N'$ is associated with a per period consumption rate $r_{it} \geq 0$, an initial inventory level $I_{i0} \geq 0$ and a maximum inventory capacity $U_i > 0$. The supplier has a per period production rate $r_{0t} \geq 0$ and an initial inventory level $I_{00} \geq 0$, but no limit on inventory capacity. A unitary per period inventory holding cost $h_i \geq 0$ is charged at each vertex $i \in N$. A fleet K of homogeneous vehicles of capacity Q is available to perform deliveries in each time period, with $|K| = m$. Every time a vehicle traverses an arc $(i, j) \in A$, a cost $c_{ij} \geq 0$ is incurred. We assume costs c_{ij} satisfy the triangle inequality. The IRP aims at determining a distribution plan such that:

- vehicle capacity and customer inventory capacities are satisfied;
- no stockout occurs, i.e., customer demands are satisfied in each time period;
- each customer is visited at most once in each time period;
- vehicle routes start and end at the supplier;
- the total cost, given by the sum of inventory holding cost at the customers and at the supplier plus the routing costs, is minimized.

We now introduce some notation that will be used in the following sections. Given a set of vertices $S \subset N$, S^c is the subset of *customers* not in S, i.e., $S^c = N' \setminus S$. Also, $\delta^+(S)$ is the set of arcs going from a vertex $i \in S$ to a vertex $j \notin S$. Similarly, $\delta^-(S)$ is the set of arcs going from a vertex $i \notin S$ to a vertex $j \in S$. For the ease of notation, we write $\delta^+(i)$ $(\delta^-(i))$ instead of $\delta^+(\{i\})$ $(\delta^-(\{i\}))$ when S is a singleton. We also define as $A(S)$ the set of arcs in S, i.e., $(i, j) \in A$ such that $i, j \in S$. Given two subsets $A, B \subset N$,

$(A : B)$ is the set of arcs linking a vertex in A with a vertex in B. Finally, given a set of variables $\mathcal{X}$, we write $\mathcal{X}(S) = \sum_{s \in S} \mathcal{X}_s$.

In addition, let us introduce a definition and notation for the projection of a linear programming formulation that will be used in the theoretical analysis presented in the following two sections.

Given a MIP formulation A, by P_A we denote the polyhedron of its LP-relaxation in which discrete variables are replaced by continuous ones with the lower and upper bounds on their domains defining the valid intervals. Given a formulation A in the *extended* space of (x, g) variables, its *natural projection* into the space of x variables (if not stated otherwise), denoted by $Proj_x(P_A)$, is defined as

$$Proj_x(P_A) = \{x \mid (x, g) \in P_A\}.$$

Given two MIP formulations, A and B, we say that A is at least as strong as B if for any problem instance, the value of the LP-relaxation of the formulation A is at least as good as the value of the LP-relaxation of the formulation B.

3 Aggregated Formulations

For aggregated formulations, we introduce the following decision variables:

- I_i^t (continuous): inventory level at vertex i in time period t, $i \in N$, $t \in T$.
- Q_i^t (continuous): quantity delivered to customer i in time period t, $i \in N'$, $t \in T$.
- Z_i^t (binary): decides whether the customer i is visited in time period t.
- Z_0^t (integer): number of vehicles used in time period t.
- X_{ij}^t (binary): decides whether arc $(i, j) \in A$ is traversed in time period t.

The following model is a relaxation of the original problem, as its solution does not necessarily respect the vehicle capacity, or it may contain subtours.

$$\min \quad \sum_{t\in T} h_0 I_{0t} + \sum_{i\in N'}\sum_{t\in T} h_i I_{it} + \sum_{(i,j)\in A}\sum_{t\in T} c_{ij} X^t_{ij} \tag{1a}$$

$$\text{s.t.} \quad I_{0t} = I_{0,t-1} + r_{0t} - \sum_{i\in N'} Q^t_i \qquad t \in T \tag{1b}$$

$$I_{it} = I_{i,t-1} - r_{it} + Q^t_i \qquad i \in N',\ t \in T \tag{1c}$$

$$(A) \quad Q^t_i \le U_i - I_{it-1} \qquad i \in N',\ t \in T \tag{1d}$$

$$Q^t_i \le C^t_i Z^t_i \qquad i \in N',\ t \in T \tag{1e}$$

$$X^t(\delta^+(i)) = X^t(\delta^-(i)) \qquad i \in N,\ t \in T \tag{1f}$$

$$X^t(\delta^-(i)) = Z^t_i \qquad i \in N,\ t \in T \tag{1g}$$

$$Z^t_i \in \{0,1\} \qquad i \in N',\ t \in T \tag{1h}$$

$$Z^t_0 \in \{0,1,\dots,|K|\} \qquad t \in T \tag{1i}$$

$$X^t_{ij} \in \{0,1\} \qquad \{i,j\} \in A,\ t \in T \tag{1j}$$

$$Q^t_i \ge 0, I_{it} \ge 0 \qquad i \in N,\ t \in T \tag{1k}$$

The objective function (1a) aims at minimizing the total cost given by the inventory cost at the supplier, the inventory cost at the customers and the routing cost. Equations (1b)-(1c) are inventory balance equations, at the supplier and at the customers, respectively. Constraints (1d)-(1e) provide upper bounds on the quantities that can be delivered to each customer $i \in N'$ in each time period. The constant C^t_i is defined as

$$C^t_i := \min\{U_i, Q, \sum_{t'=t}^{H} r_{it'}\}.$$

Finally, the degree constraints (1f)-(1g) provide a link between Z and X variables guaranteeing that each customer is visited at most once during each time period. We notice that the variables I are auxiliary in this model, and that they can be projected out, but we keep them for simplicity.

The formulation (1) is incomplete in the sense that there is no guarantee that the variables X^t build a feasible set of up to $|K|$ routes, each of which not exceeding the given capacity $\mathcal{Q}$. In the following, we discuss two possible ways proposed in the literature to extend this model and provide a valid formulation. The first one is an extended compact formulation in which we introduce an additional set of *load-based* flow variables, whereas the second one uses an exponential number of constraints to guarantee the feasibility of the routes.

3.1 Load-Based Formulation (LOAD)

To guarantee the feasibility of the routes, one can consider an *extended formulation* in which additional flow variables ℓ_{ij}^t are introduced, to count the *load* of the vehicle, while traversing the arc $(i,j) \in A$ in time period $t \in T$. In addition to the constraints in (1), we add the following constraints:

$$\ell^t(\delta^-(i)) - \ell^t(\delta^+(i)) = \begin{cases} Q_i^t & \text{if } i \neq 0, \\ -\sum_{i \in N'} Q_i^t & \text{if } i = 0. \end{cases} \quad i \in N, t \in T \tag{2a}$$

$$0 \leq \ell_{ij}^t \leq \mathcal{Q} X_{ij}^t \quad (i,j) \in A, t \in T \tag{2b}$$

These constraints guarantee that exactly Q_i^t units of flow are delivered to each $i \in N'$, $t \in T$, and that the vehicle capacity is respected. Together with degree constraints (1f) they also guarantee that subtours are eliminated.

Hence, constraints in (1) together with (2) provide a valid compact formulation for the IRP to which we will refer as LOAD in the remainder of the paper. This model has been originally proposed by Archetti et al. (2014). In a similar fashion, Gavish and Graves (1979) originally proposed to use the load-based variables to derive a compact formulation for the CVRP. Gouveia (1995) later showed that if we project the polyhedron of the LP-relaxation of this model to the space of binary arc variables, we obtain fractional capacity cuts for the CVRP, formulated as in the following section. Contrary to the CVRP, in case of the IRP, the quantities to be delivered to the customers are unknown. Nevertheless we show that a similar result holds for the MIP formulations of the IRP derived from the load-based variables, on one side, and fractional capacity cuts, on the other side.

3.2 Formulation with Fractional Capacity Cuts

The following constraints, called *fractional capacity cuts*, can be alternatively used to guarantee the feasibility of the routes, both in terms of connectivity and capacity (see, e.g., Adulyasak et al. (2014)):

$$X^t(\delta^-(S)) \geq \frac{1}{\mathcal{Q}} Q^t(S) \quad S \subseteq N',\ t \in T. \tag{FCC}$$

By summing up degree constraints (1f) over all $i \in S$, we obtain $X^t(\delta^-(S)) = X^t(\delta^+(S))$. Hence, degree constraints (1f) together with constraints (FCC)

imply the fractional capacity cuts related to arcs going out of S:

$$X^t(\delta^+(S)) \geq \frac{1}{\mathcal{Q}} Q^t(S) \quad S \subseteq N',\ t \in T. \tag{3}$$

Moreover, using the degree constraints (1g), fractional capacity cuts can be equivalently restated as

$$X^t(A(S)) \leq Z^t(S) - \frac{1}{\mathcal{Q}} Q^t(S) \quad S \subseteq N',\ t \in T. \tag{4}$$

To see why this is the case, consider a set $S \subseteq N'$, and let us sum up the degree constraints (1g) for all $i \in S$. We obtain:

$$Z^t(S) - X^t(A(S)) = X^t(\delta^-(S)). \tag{5}$$

Assuming that inequality (FCC) holds, this implies

$$Z^t(S) \geq X^t(A(S)) + \frac{1}{\mathcal{Q}} Q^t(S)$$

which is an alternative way of restating (4). Similarly, assuming that inequality (4) holds, together with (5), it implies (FCC). We remark that constraints (4) are directed counterpart of the cuts used in a formulation proposed by Adulyasak et al. (2014) for the problem with symmetric route costs.

The following result shows that projecting out ℓ variables from the formulation (2), we obtain constraints (FCC). It also establishes a connection between the load based formulation and the model derived from the fractional capacity cuts. In the following, we use A+FCC to denote the formulatoin (1) extended by constraints (FCC).

Theorem 1 *There is a one-to-one correspondence between solutions of the LP-relaxation of the model LOAD, and the solutions of the LP-relaxation of the formulation A+FCC, i.e.:*

$$Proj_{(Z,Q,X)}(P_{LOAD}) = P_{A+FCC}.$$

Proof *The proof consists of two parts:*

- $Proj_{(Z,Q,X)}(P_{LOAD}) \subseteq P_{A+FCC}$*: Let* $(\tilde{Z}, \tilde{Q}, \tilde{X}, \tilde{\ell})$ *represent a feasible LP-solution of LOAD formulation. We only need to show that the vector* $(\tilde{Z}, \tilde{Q}, \tilde{X})$ *satisfies the fractional capacity constraints* (FCC) *(as all*

other constraints in the two formulations remain unchanged). Let us consider a set $S \subseteq N'$ and $t \in T$. After summing up the flow conservation constraints (2a) *over all $i \in S$, we obtain:*

$$\tilde{\ell}^t(\delta^-(S)) = \tilde{\ell}^t(\delta^+(S)) + \tilde{Q}^t(S)$$

After using the capacity constraints (2b) *to bound the flow from above for the arcs $(i,j) \in \delta^-(S)$ and from below for the arcs $(i,j) \in \delta^+(S)$, we obtain*

$$\mathcal{Q}\tilde{X}^t(\delta^-(S)) \geq \tilde{\ell}^t(\delta^-(S)) = \tilde{\ell}^t(\delta^+(S)) + \tilde{Q}^t(S) \geq \tilde{Q}^t(S)$$

i.e.,

$$\mathcal{Q}\tilde{X}^t(\delta^-(S)) \geq \tilde{Q}^t(S),$$

which is another way of writing cuts (FCC). *This derivation shows that* (FCC) *are contained in the projection of the load-based formulation onto the space of (Z, Q, X) variables.*

- *$P_{A+FCC} \subseteq Proj_{(Z,Q,X)}(P_{LOAD})$: Let $t \in T$ be an arbitrary time period. To show that constraints* (FCC) *provide a complete description of the projection of the flow into the space of X variables, we now start with a solution $(\bar{Z}, \bar{Q}, \bar{X})$ for the formulation based on* (FCC), *and we show that there exists a feasible flow $\bar{\ell}^t$ which satisfies constraints* (2a)-(2b). *Such a flow exists if and only if there exists* a feasible flow $\tilde{\ell}^t$ *on the same graph with the lower and upper bounds on the arc capacities ($\underline{\kappa}^t$, $\overline{\kappa}^t$, respectively) defined as follows:*

$$\underline{\kappa}^t_{ij} = 0, \qquad \overline{\kappa}^t_{ij} = \mathcal{Q}\bar{X}^t_{ij}, \qquad i \in N, j \in N', \tag{6}$$

$$\underline{\kappa}^t_{j0} = \bar{Q}^t_j, \qquad \overline{\kappa}^t_{j0} = \bar{Q}^t_j, \qquad j \in N'. \tag{7}$$

Indeed, this follows from the fact that the flow demands of $\bar{Q}^t_j$ of each vertex $j \in N'$ are transformed into fixed arc capacities for backward arcs $(j,0)$ entering the depot (see, e.g., Section 6.7 in Ahuja et al. (1993) for more details). The flow $\tilde{\ell}^t$ is said to be feasible if and only if the following constraints (the flow conservation and the capacity constraints, respectively) are satisfied:

$$\tilde{\ell}^t(\delta^-(i)) = \tilde{\ell}^t(\delta^+(i)), \qquad i \in N, \tag{8}$$

$$\underline{\kappa}^t_{ij} \leq \tilde{\ell}^t_{ij} \leq \overline{\kappa}^t_{ij}, \qquad (i,j) \in A. \tag{9}$$

According to Hoffman (1960), in a digraph (N, A) with arc capacities defined as in (9)*, there exists a feasible flow $\tilde{\ell}^t$ if and only if*

$$\underline{\kappa}^t(\delta^-(S)) \leq \overline{\kappa}^t(\delta^+(S)), \quad S \subset N. \tag{10}$$

Hence, to prove tha there exists a feasible flow $\tilde{\ell}^t$ described above, let us consider a set $S \subset N$. We distinguish the following two cases:

1. *$0 \in S$: in that case, $\underline{\kappa}^t(\delta^-(S)) = \bar{Q}^t(S^c)$ and $\overline{\kappa}^t(\delta^+(S)) = \mathcal{Q}\bar{X}^t(\delta^+(S)) = \mathcal{Q}\bar{X}^t(\delta^-(S^c))$, and so the inequality* (10) *turns into*

$$\mathcal{Q}\bar{X}^t(\delta^-(S^c)) \geq \bar{Q}^t(S^c),$$

   which is the fractional capacity cut imposed for the set S^c.

2. *$0 \notin S$: in that case, $\underline{\kappa}^t(\delta^-(S)) = 0$ and $\overline{\kappa}^t(\delta^+(S)) = \bar{Q}^t(S) + \mathcal{Q}\bar{X}^t(S : S^c)$, and so the inequality* (10) *is trivially satisfied.*

■

Hence, the compact way of expressing the fractional capacity cuts is given by introducing the load-based variables together with constraints (2a)-(2b). This result is in-line with what is known for the capacitated VRP (see, Gouveia (1995)).

The following result shows that our aggregated model also includes the aggregated vehicle capacity constraints that ensure that in each time period $t \in T$, the total amount of flow delivered by all vehicles does not exceed the number of used vehicles times their capacity $\mathcal{Q}$.

Lemma 1 *The aggregated vehicle capacity constraints*

$$Q^t(N') \leq \mathcal{Q} \cdot Z_0^t, \quad t \in T \tag{11}$$

are implied by the degree constraints (1f)-(1g) *and constraints* (FCC).

Proof *From the degree constraints* (1f)-(1g) *we have*

$$\begin{aligned} Z^t(N') = \sum_{i \in N'} X^t(\delta^-(i)) = X^t(\delta^-(N')) + X^t(A(N')) = \\ = X^t(\delta^+(0)) + X^t(A(N')) = Z_0^t + X^t(A(N')) \Rightarrow \end{aligned}$$

$$\Rightarrow X^t(A(N')) = Z^t(N') - Z_0^t. \tag{12}$$

On the other hand, the FCC for $S = N'$ can be re-written as

$$\begin{aligned} X^t(A(N')) \leq Z^t(N') - \frac{1}{\mathcal{Q}} Q^t(N') &\Rightarrow Z^t(N') - Z_0^t \leq Z^t(N') - \frac{1}{\mathcal{Q}} Q^t(N') \\ &\Rightarrow Q^t(N') \leq \mathcal{Q} \cdot Z_0^t. \end{aligned}$$

■

3.3 Strengthened Load-Based Formulation and Multi-Star Inequalities for the IRP

In this subsection we attempt to strengthen the previously introduced formulations by exploiting the following property of optimal solutions:

Lemma 2 *When input parameters $(\mathcal{Q}, r, U)$ take on integer values, then there exists an optimal solution such that the values of the quantities Q_i^t delivered to each customer $i \in N'$, for each $t \in T$, are integer.*

Proof *Indeed, once the design variables (X, Z) are fixed, the problem becomes a minimum-cost flow in a time-expanded network with vertex- and arc-capacities determined by the values of $(\mathcal{Q}, r, U)$. The problem can be reformulated as a minimum-cost flow problem in a digraph with integral arc capacities and integer supplies/demands. The result follows from the integrality property of the minimum-cost flow, which is guaranteed whenever all arc capacities and supplies/demands of vertices are integer (see, e.g., Theorem 9.10 in Ahuja et al. (1993)).* ■

Corollary 1 *The following constraints are valid for the aggregated formulation:*

$$Z_i^t \leq Q_i^t, \quad i \in N', t \in T. \tag{13}$$

Indeed, these constraints state that if a customer $i \in N'$ is visited in time period $t \in T$, at least one unit of demand is delivered at this vertex. Note that this is valid as we assume that costs c_{ij} satisfy the triangle inequality.

In the following subsections we discuss how the property of Lemma 2 can be exploited to improve the bounds of the load-based formulation and fractional capacity cuts, respectively. We also show that by projecting out flow variables from the strengthened load formulation, we obtain new inequalities for the IRP, that we refer to as *IRP-multi-star inequalities*.

3.3.1 Strengthened Load-Based Formulation (SLOAD)

Due to Lemma 2, we can impose lower and upper bounds on the load transported along each arc $(i,j) \in A$ as in the following lemma.

Lemma 3 *The load-based formulation can be strengthened by replacing constraints* (2) *with the following ones:*

$$X_{ij}^t \leq \ell_{ij}^t \leq (\mathcal{Q}-1)X_{ij}^t \quad i,j \in N',\ t \in T \tag{14a}$$

$$\ell_{j0}^t = 0 \quad j \in N',\ t \in T \tag{14b}$$

$$X_{0j}^t \leq \ell_{0j}^t \leq \mathcal{Q}X_{0j}^t \quad j \in N',\ t \in T \tag{14c}$$

Proof *If arc $(i,j) \in A$ is traversed in time period t, then both customers i and j have to be served. Hence, the load transported along the arc is at least one (cf. Lemma 2), and the maximal load is $\mathcal{Q}-1$ (because at least one unit has to be delivered before traversing the arc).* ■

We will denote by SLOAD the *strengthened load-based formulation*, which is determined by (1), (13), (14). We point out that, following the CVRP literature, much stronger formulations could be obtained by replacing for example constraints (14a) by $Q_j^t X_{ij}^t \leq \ell_{ij}^t \leq (\mathcal{Q}-Q_i^t)X_{ij}^t$. Unfortunately, due to the fact that the quantities Q_j^t are decision variables, we obtain bilinear constraints, and hence the study of the underlying models is out of scope of this article.

3.3.2 IRP-Multi-Star Inequalities

In this section we introduce multi-star inequalities for the IRP, which are adaptation of the well-known multi-star inequalities for the capacitated VRP (see, e.g., Gouveia and Hall (2002); Letchford and González (2006, 2015)). We also show that projecting out the flow variables from the formulation SLOAD, we obtain the IRP-multi-star inequalities.

Definition 1 *Let us consider a set $S \subseteq N'$ and $t \in T$. Then:*

$$\mathcal{Q}X^t(\delta^-(S)) \geq Q^t(S) + X^t(S^c : S) + X^t(S : S^c), \tag{MS}$$

are called IRP-multi-star inequalities.

Based on the value of $Q^t(S)$ which represents (variable) demand served to the customers from the set S, constraints (MS) provide an upper bound on the number of arcs (non-adjacent to the depot) that have exactly one endpoint in S. Indeed, the latter quantity is bounded from above as follows:

$$X^t(S^c : S) + X^t(S : S^c) \leq \mathcal{Q}X^t(\delta^-(S)) - Q^t(S).$$

We notice that inequalities (MS) are a lifted version of constraints (FCC), as the latter ones can be obtained from (MS) by just omitting the term $X^t(S^c : S) + X^t(S : S^c)$ from the right-hand side.

With the following theorem, we show that constraints (MS) are contained in the model SLOAD, and are therefore valid for the IRP.

Theorem 2 *The IRP-multi-star inequalities* (MS) *are implied by the strengthened load-based formulation SLOAD.*

Proof *Let us consider a set $S \subseteq N'$, $S \neq \emptyset$ and $t \in T$. After summing up the flow conservation constraints* (2a) *over all $i \in S$, we obtain:*

$$\ell^t(\delta^-(S)) - \ell^t(\delta^+(S)) = Q^t(S).$$

In the above equation, we can substitute

$$\ell^t(\delta^-(S)) = \ell^t(0 : S) + \ell^t(S^c : S),$$

and

$$\ell^t(\delta^+(S)) = \ell^t(S : 0) + \ell^t(S : S^c).$$

Then, after bounding from above the values of $\ell^t(\delta^-(S))$ using (14a),(14c) *and bounding from below the values of $\ell^t(\delta^+(S))$ using* (14a),(14b)*, we obtain:*

$$\mathcal{Q}X^t(0 : S) + (\mathcal{Q} - 1)X^t(S^c : S) - X^t(S : S^c) \geq Q^t(S). \tag{15}$$

Finally, after adding $X^t(S^c : S)$ to both sides, we obtain the IRP-multi-star inequalities (MS). ■

Theorem 3 *Projecting out ℓ variables from the formulation SLOAD we obtain the inequalities* (MS).

The proof of Theorem 3 is similar o the one of Theorem 1 and is moved to the Appendix.

Putting together the results from Theorems 2 and 3, we obtain:

Corollary 2

$$Proj_{(Z,Q,X)}(P_{SLOAD}) = P_{A+MS}.$$

3.4 Connectivity Cuts

The following constraints, known as the generalized subtour elimination constraints (GSECs), are frequently used to impose the connectivity and to eliminate subtours in routing (see, e.g., Coelho et al. (2013)) and network design problems (see, e.g., Ljubić (2021)):

$$X^t(A(S)) \leq Z^t(S \setminus \{i\}) \quad S \subseteq N', |S| \geq 2,\ i \in S,\ t \in T. \tag{GSEC}$$

Thanks to the degree constraints (1f)-(1g), GSECs can be equivalently stated as *connectivity cuts*

$$X^t(\delta^-(S)) \geq Z_i^t \quad S \subseteq N',\ i \in S,\ t \in T \tag{CC}$$

enforcing that for every customer i visited at time t, there is a directed path between the depot and i. In the following, we will use interchangeably the terms GSECs and connectivity cuts. Indeed, this transformation follows from the following equations:

$$Z^t(S) = \sum_{i \in S} X^t(\delta^-(i)) = X^t(A(S)) + X^t(\delta^-(S)).$$

Multi-Commodity Flow GSECs can be replaced by a polynomial number of constraints in an extended space in which additional multi-commodity flow variables f^{tl} are introduced for each $t \in T$ and $l \in N'$. For each customer $l \in N'$ such that $Z_l^t = 1$, the following constraints guarantee that there exists a directed path between the depot and l in the solution determined by the vector X:

$$f^{tl}(\delta^-(i)) - f^{tl}(\delta^+(i)) = \begin{cases} Z_i^t & \text{if } i = l, \\ -Z_i^t & \text{if } i = 0, \\ 0 & \text{otherwise.} \end{cases} \qquad l \in N',\ i \in N,\ t \in T \tag{16a}$$

$$0 \leq f_{ij}^{tl} \leq X_{ij}^t \qquad l \in N', (i,j) \in A,\ t \in T \tag{16b}$$

Let in the following MCF refer to constraints (16) and LOAD+MCF refer to the load-based model extended by the multi-commodity flow constraints, i.e., the formulation (1), (2), (16).

From the VRP literature, it is known that constraints (FCC) and (GSEC) are not dominating each other (see, e.g., Letchford and González (2015)), and this result carries over to the IRP. Hence, it is beneficial to include both in the

Formulation	Constraints	Size	Variables
LOAD	(1), (2)	compact	(Z, Q, X, ℓ)
SLOAD	(1), (13), (14)	compact	(Z, Q, X, ℓ)
LOAD+MCF	(1), (2), (16)	compact	(Z, Q, X, ℓ, f)
SLOAD+MCF	(1), (13), (14), (16)	compact	(Z, Q, X, ℓ, f)
A+FCC	(1), (FCC)	exponential	(Z, Q, X)
A+FCC+GSEC	(1), (FCC), (GSEC)	exponential	(Z, Q, X)
A+MS	(1), (13), (MS)	exponential	(Z, Q, X)
A+MS+GSEC	(1), (13), (MS), (GSEC)	exponential	(Z, Q, X)

Table 1: Aggregated formulations considered in this work.

aggregated formulation. We will denote by A+FCC+GSEC the aggregated formulation determined by (1),(13), (FCC) and (GSEC). Similarly, if in the latter formulation constraints (FCC) are replaced by (MS), we will denote it by A+MS+GSEC.

Theorem 4 *Projecting out f variables from the model LOAD+MCF, we obtain the constraints* (GSEC)*, i.e.:*

$$Proj_{(X,Z,Q)}(P_{LOAD+MCF}) = P_{A+FCC+GSEC}.$$

Proof *The result follows from the min-cut max-flow theorem, see, e.g., Ljubić et al. (2006); Ljubić (2021).* ■

Table 1 summarizes all the models considered in this section. Figure 1 provides a hierarchy of these formulations. An arrow pointing from a formulation A to formulation B indicates that the latter is at least as strong as the former one. Correspondingly, with $\leftrightarrow$ we indicate that the two models are equivalent, i.e., projecting out the flow variables (ℓ and/or f) from a given extended formulation results into the corresponding formulation (of exponential size) in the natural space of (Z, Q, X) variables.

4 Disaggregated Formulations

We now turn our attention to disaggregated formulations, in which we disaggregate variables X, Z and Q per vehicle, assuming that an upper bound $|K|$ on the available vehicles is known. Hence we introduce the following variables:

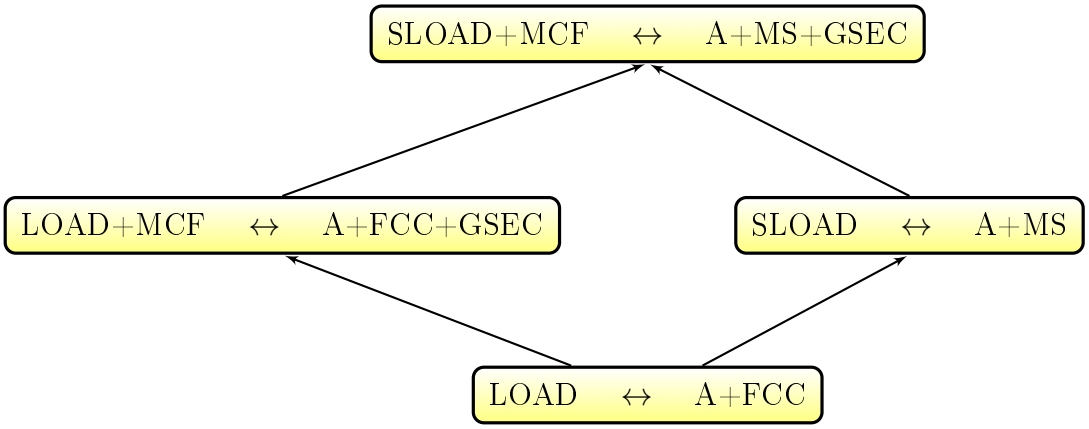

Figure 1: Hierarchy of aggregated formulations. An arrow from a model A to model B indicates that model B is at least as strong as model A.

- I_i^t (continuous): inventory level at vertex i in time period t, $i \in N$, $t \in T$.
- q_i^{kt} (continuous): quantity delivered to customer i in time period t, by vehicle k, $i \in N'$, $t \in T$, $k \in K$.
- z_i^{kt} (binary): decides whether vertex $i \in N$ is visited in time period t by vehicle $k \in K$, $i \in N'$, $t \in T$, $k \in K$.
- x_{ij}^{kt} (binary): decides whether arc $(i, j) \in A$ is traversed in time period t by vehicle k, $(i, j) \in A$, $t \in T$, $k \in K$.

A relaxation of the IRP can then be modelled as:

$$\min \quad \sum_{t\in T} h_0 I_{0t} + \sum_{i\in N'}\sum_{t\in T} h_i I_{it} + \sum_{k\in K}\sum_{(i,j)\in A}\sum_{t\in T} c_{ij} x_{ij}^{kt} \tag{17a}$$

$$\text{s.t.} \quad I_{0t} = I_{0,t-1} + r_{0t} - \sum_{k\in K}\sum_{i\in N'} q_{it}^k \qquad t \in T \tag{17b}$$

$$I_{it} = I_{i,t-1} - r_{it} + \sum_{k\in K} q_{it}^k \qquad i \in N',\ t \in T \tag{17c}$$

$$(D) \quad \sum_{k\in K} q_{it}^k \le U_i - I_{it-1} \qquad i \in N',\ t \in T \tag{17d}$$

$$0 \le q_{it}^k \le C_i^t z_i^{kt} \qquad i \in N',\ k \in K,\ t \in T \tag{17e}$$

$$\sum_{i\in N'} q_{it}^k \le \mathcal{Q} z_0^{kt} \qquad k \in K,\ t \in T \tag{17f}$$

$$\sum_{k\in K} z_i^{kt} \le 1 \qquad i \in N',\ t \in T \tag{17g}$$

$$x^{kt}(\delta^-(i)) = x^{kt}(\delta^+(i)) \qquad i \in N,\ k \in K,\ t \in T \tag{17h}$$

$$x^{kt}(\delta^-(i)) = z_i^{kt} \qquad i \in N,\ k \in K,\ t \in T \tag{17i}$$

$$z_i^{kt} \in \{0,1\} \qquad i \in N,\ k \in K,\ t \in T \tag{17j}$$

$$x_{ij}^{kt} \in \{0,1\} \qquad (i,j) \in A,\ k \in K,\ t \in T \tag{17k}$$

$$I_{it} \ge 0 \qquad i \in N,\ t \in T \tag{17l}$$

Indeed, this models is incomplete and represents only a relaxation of the original problem. Even though the vehicle capacity constraints are respected (thanks to constraints (17e), (17f) and (17i)), the connectivity of the routes to the depot is not guaranteed. Constraints (17b)-(17d), (17h), (17i) represent disaggregated counterparts of constraints (1b)-(1g). With (17e) we guarantee that the quantity delivered at each customer i is at most C_i^t. Finally, constraints (17g) ensure that every customer is visited by at most one vehicle in each time period.

Disaggregated Connectivity Cuts One possible way to impose the connectivity (i.e., eliminate the subtours, and hence obtain a valid formulation) is by inserting the so-called *disaggregated GSECs*:

$$x^{kt}(A(S)) \le z^{kt}(S \setminus \{i\}) \quad S \subseteq N',\ i \in S,\ t \in T,\ k \in K \tag{dGSEC}$$

Formulation	Constraints	Size	Variables
D+FCC	(17), (dFCC)	exponential	(z, q, x)
D+GSEC	(17), (dGSEC)	exponential	(z, q, x)
D+FCC+GSEC	(17), (dFCC), (dGSEC)	exponential	(z, q, x)

Table 2: Disaggregated formulations considered in this section.

Coelho and Laporte (2014) used the undirected counterpart of these cuts for the IRP with symmetric arc costs. (dGSEC) can be equivalently stated as *disaggregated connectivity cuts*:

$$x^{kt}(\delta^-(S)) \geq z_i^{kt} \quad S \subseteq N',\ i \in S,\ t \in T,\ k \in K.$$

In the remainder of the article we will refer to the model (17) with (dGSEC) as D+GSEC. We point out that in a similar fashion as shown for the aggregated formulations, one can derive the compact counterpart of constraints (dGSEC), by using a 5-index formulation, in which flow variables h_{ij}^{ktl} indicate whether the arc (i, j) is traversed to visit customer $l \in N'$ at time t by vehicle k.

Disaggregated Fractional Capacity Cuts Similarly, one can consider *disaggregated fractional capacity cuts* defined as:

$$x^{kt}(\delta^-(S)) \geq \frac{1}{\mathcal{Q}} \sum_{i \in S} q_i^{kt} \qquad S \subseteq N',\ t \in T,\ k \in K. \tag{dFCC}$$

These cuts provide a lower bound on the number of arcs entering each S, for each vehicle k.

In the following we will denote by D+FCC+GSEC the disaggregated formulation (17) with (dFCC) and (dGSEC) constraints. Table 2 provides a summary of disaggregated formulations considered in this article.

Strength of the Disaggregated Formulation. The following result shows that, in terms of the quality of lower bounds, the disaggregated model D+FCC+GSEC studied in this article (and frequently used in the IRP literature, see Archetti et al. (2014)) does not contribute in improving the quality of lower bounds (compared to their disaggregated counterpart).

Theorem 5 *We have*

$$Proj_{(Z,Q,X)}(P_{D+FCC+GSEC}) = P_{A+FCC+GSEC},$$

where the projection $Proj_{(Z,Q,X)}(P_{D+FCC+GSEC})$ *is defined as:*

$$Proj_{(Z,Q,X)}(P_{D+FCC+GSEC}) = \{(Z,Q,X) \mid (z,q,x) \in P_{D+FCC+GSEC} \text{ and } X_{ij}^t = \sum_{k\in K} x_{ij}^{kt}, (i,j) \in A, \quad Z_i^t = \sum_{k\in K} z_i^{kt}, i \in N \text{ and } Q_i^t = \sum_{k\in K} q_i^{kt}, i \in N', t \in T\}. \tag{18}$$

Proof *The proof consists of two parts:*

- $Proj_{(Z,Q,X)}(P_{D+FCC+GSEC}) \subseteq P_{A+FCC+GSEC}$*: To show this result, we consider an arbitrary point feasible for the LP-relaxation of the model D+FCC+GSEC,* $(\hat{z},\hat{q},\hat{x}) \in P_{D+FCC+GSEC}$ *and we consider its projection point* $(\hat{Z},\hat{Q},\hat{X}) \in Proj_{(Z,Q,X)}(P_{D+FCC+GSEC})$ *following the definition from* (18)*. It is not difficult to see that such defined point* $(\hat{Z},\hat{Q},\hat{X})$ *satisfies all the constraints of the model A+FCC+GSEC, i.e., it belongs to* $P_{A+FCC+GSEC}$*. Moreover, the value of the objective function remains unchanged.*

- $P_{A+FCC+GSEC} \subseteq Proj_{(Z,Q,X)}(P_{D+FCC+GSEC})$*: Let* $(\hat{Z},\hat{Q},\hat{X}) \in P_{A+FCC+GSEC}$ *be an arbitrary point satisfying the LP-relaxation of the model A+FCC+GSEC. We will show how to "lift" this point into the* $P_{D+FCC+GSEC}$ *polyhedron without changing the value of the objective function. We construct a vector* $(\hat{z},\hat{q},\hat{x}) \in P_{D+FCC+GSEC}$ *such that for all* $t \in T$*:*

  $$\hat{x}_{ij}^{kt} = \hat{X}_{ij}^t/K, \quad (i,j) \in A \quad \hat{z}_i^{kt} = \hat{Z}_i^t/K, \quad i \in N \text{ and } \hat{q}_i^{kt} = \hat{Q}_i^t/K, \quad i \in N',$$

  and we show that the point $(\hat{z},\hat{q},\hat{x})$ *satisfies all the constraints of the model D+FCC+GSEC. Clearly, constraints* (17b)-(17e) *follow directly from constraints* (1b)-(1e)*, respectively. No-split cuts* (17g) *are satisfied, which follows from the fact that* Z_i^t *variables are bounded by one for any* $i \in N'$*. Moreover, the disaggregated degree constraints* (17h)-(17i)*, cuts* (dFCC) *and* (dGSEC) *all follow from* (1g)-(1h)*,* (FCC) *and* (GSEC)*, respectively. Finally, the validity of* (17f) *follows from Lemma 1.*

■

5 Valid Inequalities

The following valid inequalities are used to strengthen the lower bound for aggregated formulations. They are all inherited from previous works on the IRP and adapted to aggregated formulations. Constraints (19)-(20) were proposed in Archetti et al. (2014), whereas constraints (21)-(23) are from Coelho and Laporte (2014).

$$I_i^\tau \geq (\sum_{t'=t-\tau+1}^{H} Z_i^t)(\sum_{t'=t-\tau+1}^{H} r_{it}) \qquad i \in N', t \in T, \tau = 0, ..., t-1 \qquad (19)$$

$$Z_i^t \leq Z_0^t \qquad i \in N', t \in T \qquad (20)$$

$$\sum_{\tau=1}^{t} Z_i^\tau \geq \lceil \frac{(\sum_{\tau=1}^{t} r_{i\tau} - I_{i0})}{\min\{Q, U_i\}} \rceil \qquad i \in N', t \in T \qquad (21)$$

$$\sum_{\tau=t_1}^{t_2} Z_i^\tau \geq \lceil \frac{(\sum_{\tau=t_1}^{t_2} r_{i\tau} - U_i)}{\min\{Q, U_i\}} \rceil \qquad i \in N', t_1, t_2 \in T, t_2 \geq t_1 \qquad (22)$$

$$\sum_{\tau=t_1}^{t_2} Z_i^\tau \geq \frac{(\sum_{\tau=t_1}^{t_2} r_{i\tau} - I_i^{t_1})}{\min\{Q, U_i\}} \qquad i \in N', t_1, t_2 \in T, t_2 \geq t_1 \qquad (23)$$

We also introduced the following inequalities for breaking symmetries in instances with symmetric costs c_{ij} and directed formulation:

$$\sum_{i \in N'} i X_{0i}^t \leq \sum_{i \in N'} i X_{i0}^t \quad t \in T \qquad (24)$$

In addition, we explicitly consider the subset of GSECs with $|S| = 2$ for initialization of MIP models tested in our empirical study provided below:

$$X_{ij}^t + X_{ji}^t \leq Z_i^t \qquad i, j \in N', t \in T \qquad (25)$$

6 Computational Experiments

In this section we aim at evaluating the performance of the aggregated formulations presented in Section 3. Tests are performed on benchmark instances for the IRP (see Archetti et al. (2014)). The characteristics of the instances are the following:

- Time horizon H: 3 and 6;
- Number of customers n: $5l$, with $l = 1, ..., 6$ for $H = 6$ and $l = 1, ..., 10$ for $H = 3$.
- Number of vehicles m: from 2 to 5.
- High and low inventory cost.

For each combination of the above parameters, 5 instances were generated, thus having 640 instances in total. We refer to Archetti et al. (2014) for more details on the instances.

Computational tests were made on a Windows 64 machine equipped with Intel(R) Xeon(R) CPU E5-1650 v2, 3050 GHz, 64.0 GB RAM. The code was written in C++, compiled with MS Visual Studio 2019 Express Edition in release mode, and CPLEX 12.10 was used as an exact solver and run on 4 threads.

Experiments are organized as follows. We first test the effect of the different valid inequalities and cuts on the lower bound associated with the plain load-based formulation. To do that, we compare the LP-relaxation of the following compact formulations:

1. Plain aggregated formulation with load variables given by (1)–(2) (`LOAD`).
2. Formulation `LOAD` plus subtour elimination constraints on symmetric arcs (25) (`LOAD+S`).
3. Strenghtened aggregated formulation determined by (1), (13), (14), plus subtour elimination constraints on symmetric arcs (25) (`SLOAD+S`).
4. Strenghtened aggregated formulation determined by (1), (13), (14), plus connectivity constraints (16) (`SLOAD+G`).
5. Formulation `SLOAD+G` plus valid inequalities (19)–(24) (`SLOAD+G+VI`).

A second set of tests is aimed then at evaluating the performance of aggregated formulations in solving the problem, thus by retaining integrality constraints. On the basis of the results of the first tests mentioned above, a subset of formulations is used in this second set. Formulations are solved to optimality and compared with state-of-the-art solution approaches, namely the branch-and-cut approach proposed in Coelho and Laporte (2014) (called

CL from now on) and the branch-and-price algorithm in Desaulniers et al. (2016) (called DRC from now on). Some observations about the two latter approaches are needed. Concerning CL, it is a branch-and-cut algorithm where GSECs are formulated as in (GSEC) and separated dynamically using the min-cut algorithm. However, arc variables are undirected and, thus, the formulation is not equivalent to the disaggregated formulation presented in Section 4, having half of the arc variables. Concerning DRC, it uses Dantzig-Wolfe decomposition coupled with column generation. The master problem is a set-partitioning-like formulation and the subproblem is formulated as an elementary shortest path problem with resource constraints.

6.1 Strength of the Relaxation of Aggregated Formulations

In this section we summarize the results of the first set of tests. Table 3 reports, for each of formulation mentioned above except LOAD, the average percentage improvement of the LP-relaxation bound with respect to the value of the LP-relaxation of LOAD (% impr.) and the average computing time (Time) in seconds. The computing time for LOAD is always negligible (few milliseconds) and thus not reported. Instances are divided in two classes according to the value of inventory cost (low and high). Then, they are classified by time horizon H first, number of vehicles m second and number of customers n third.

The results show that (25) are effective in improving the bound with a negligible burden on computing time. IRP-multi-star inequalities (i.e., their compact counterpart given by (14)) and valid inequalities (19)–(24) are less effective but the computational time increase caused by their inclusion is tiny. On the other side, connectivity cuts (16) cause a remarkable increase in computational time. Thus, in the following tests we decided to keep (14) and (19)–(25). Connectivity cuts are instead either disregarded or dynamically separated as described in the following section.

6.2 Comparison with State-of-the-Art

In this section we compare the performance of aggregated formulations versus CL and DRC. Given the results from the previous section, valid inequalities (14), and (19)–(25) are added to the basic formulation (1)–(2). Note that,

Table 3: Average results on solutions of linear relaxations

		LOAD+S		SLOAD+S		SLOAD+G		SLOAD+G+VI	
		% impr.	Time	% impr.	Time	% impr.	Time	% impr.	Time
Low inventory cost									
H	3	3.19	0.20	3.25	0.20	4.30	127.97	4.79	140.25
	6	5.23	0.17	5.28	0.43	7.32	23.22	7.32	27.93
m	2	7.30	0.15	7.33	0.26	10.46	85.09	10.66	95.70
	3	4.82	0.18	4.87	0.25	6.63	85.88	6.74	96.25
	4	3.37	0.23	3.43	0.30	4.53	91.29	4.74	99.56
	5	2.38	0.20	2.44	0.31	3.14	92.50	3.36	100.99
n	5	0.85	0.00	0.94	0.00	0.96	0.00	1.59	0.00
	10	1.90	0.00	1.96	0.00	2.09	0.00	2.44	0.00
	15	3.52	0.00	3.59	0.00	5.21	1.38	5.37	1.93
	20	5.56	0.00	5.61	0.00	7.19	6.83	7.28	8.38
	25	5.60	0.00	5.64	0.43	8.50	22.23	8.59	25.50
	30	4.84	0.50	4.89	0.85	6.41	60.48	6.57	70.33
	35	6.80	0.00	6.83	0.00	9.14	64.00	9.14	67.45
	40	7.11	0.80	7.14	0.00	10.91	136.35	10.91	154.20
	45	7.31	1.20	7.35	0.95	11.05	336.75	11.05	345.40
	50	5.67	0.00	5.70	1.00	7.20	700.10	7.20	790.70
Av. *Low*		4.46	0.19	4.52	0.28	6.19	88.69	6.37	98.13
High inventory cost									
H	3	1.56	0.21	1.59	0.20	2.12	131.68	2.86	143.24
	6	1.90	0.17	1.92	0.36	2.64	23.58	2.64	27.94
m	2	2.72	0.16	2.73	0.23	3.85	85.85	4.16	97.59
	3	1.91	0.20	1.94	0.26	2.63	88.86	2.87	95.81
	4	1.42	0.20	1.45	0.25	1.90	93.59	2.19	102.09
	5	1.05	0.20	1.08	0.30	1.39	96.26	1.67	104.51
n	5	0.53	0.00	0.59	0.00	0.60	0.00	1.12	0.00
	10	1.03	0.00	1.07	0.00	1.13	0.00	1.58	0.00
	15	1.70	0.00	1.74	0.00	2.53	1.35	2.89	1.80
	20	2.43	0.00	2.46	0.00	3.14	6.95	3.44	8.70
	25	2.37	0.00	2.39	0.35	3.62	21.93	3.93	25.98
	30	1.80	0.50	1.82	0.73	2.43	62.75	2.71	70.73
	35	2.35	0.00	2.37	0.00	3.18	68.95	3.18	70.90
	40	2.36	0.85	2.37	0.00	3.46	143.75	3.46	155.30
	45	2.15	1.20	2.16	1.00	3.27	337.35	3.27	352.70
	50	1.76	0.00	1.77	1.00	2.30	722.25	2.30	806.70
Av. *High*		1.77	0.19	1.80	0.26	2.44	91.14	2.72	100.00
Total av.		3.12	0.19	3.16	0.27	4.32	89.91	4.55	99.06

as they are polynomial in number, no separation is needed. The following exact approaches are tested:

- `Compact`: `SLOAD+S+VI` is passed to Cplex and solved.
- `B&C`: `SLOAD+S+VI` is augmented by GSECs (GSEC), which are dynamically separated on fractional points through the classical min-cut algorithm (see Padberg and Rinaldi (1991)). Notice that subtours at integer points are already eliminated by the constraints of the formulation `SLOAD`.
- `Benders`: `SLOAD+S+VI` is augmented by connectivity constraints (16). In order to avoid the sharp computational burden caused by constraints (16) and observed in the results presented above, a Benders decomposition is proposed where variables f and the associated constraints (16) are moved to the subproblems. Thus, with this setting, we replace the combinatorial separation of constraints (GSEC), by internal separation provided by Cplex using the annotated Benders setting. Benders subproblems are separated by $t \in T$ and $l \in N'$ as they are fully independent.

A time limit of two hours has been set. For the two competing approaches, `CL` was run on a grid of Intel Xeon^{TM} processors running at 2.66 GHz with up to 48 GB of RAM installed per node, with the Scientific Linux 6.1 operating system and a time limit of 2 hours. Cplex 12.5 was used as exact solver. `DRC` was run on an Intel Core i7-2600 processor clocked at 3.4 GHz with 8 cores and 16 GB RAM, with a time limit of 2 hours. Cplex 12.2 was used as exact solver.

We point out that we did not separate inequalities (MS) and we kept them explicitly in the model using (14). There are two reasons for this: 1) projecting out load variables ℓ does not reduce the size of the model significantly (as we keep the same order of magnitude of X variables), and 2) by keeping ℓ variables in the model, a complete information about the structure of feasible solutions is provided to the general-purpose solver. Hence, generic heuristics integrated in modern solvers can be more effective in finding high-quality solutions earlier in the branching tree. This property might be lost when ℓ variables are eliminated and the solver learns the feasibility constraints "on the fly".

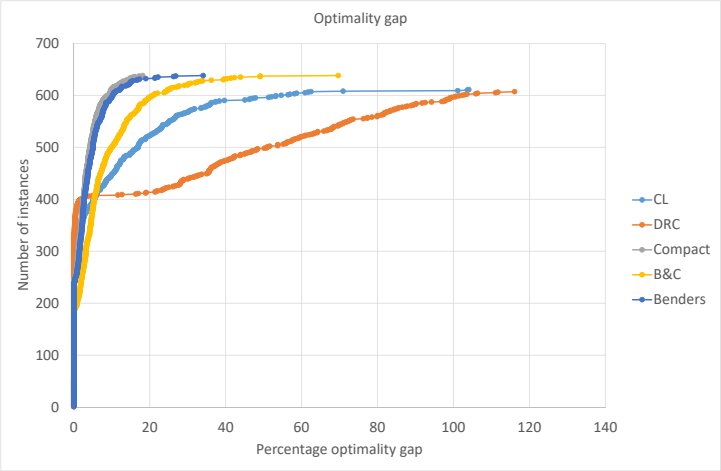

Figure 2: Optimality gap at termination.

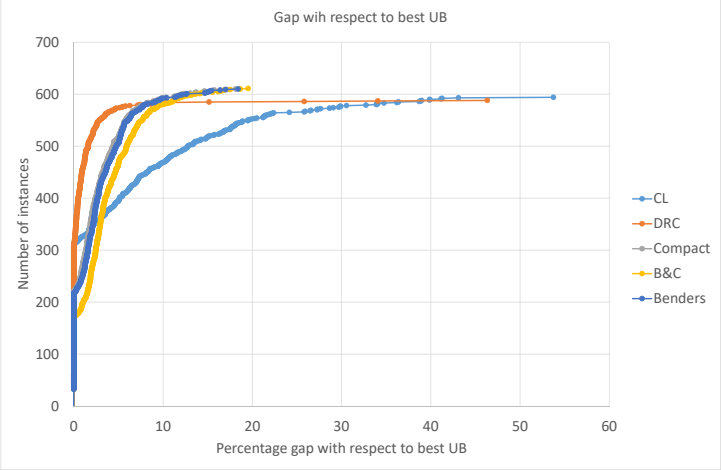

Figure 3: Gap between lower bound at termination and best upper bound.

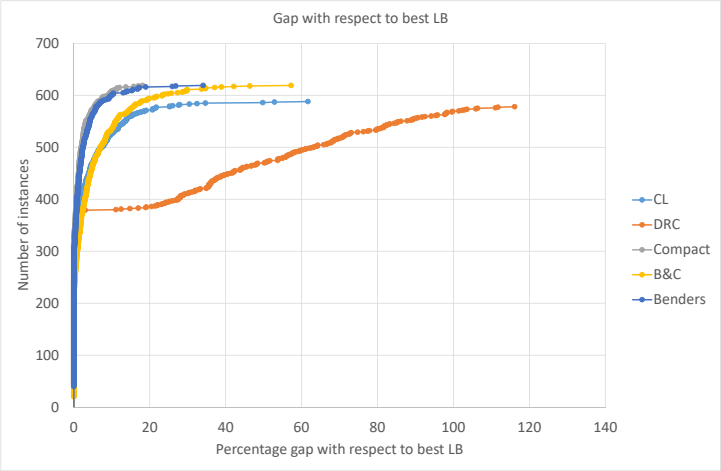

Figure 4: Gap between upper bound at termination and best lower bound.

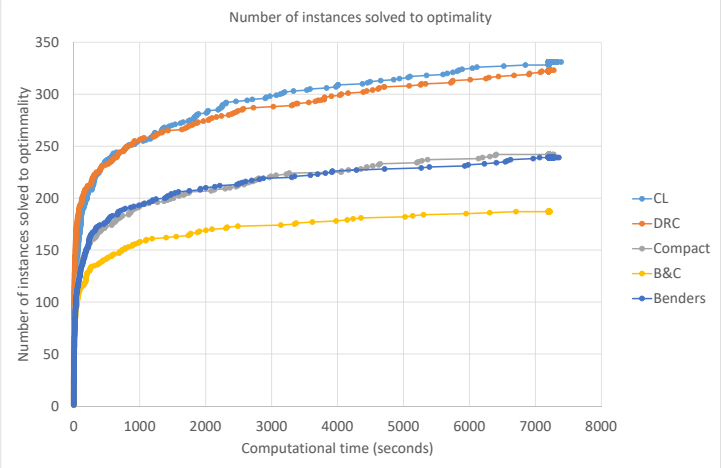

Figure 5: Number of instances solved to optimality vs. computing time.

Results are reported in Figures 2–5. Figure 2 shows the optimality gap at termination. In particular, the figure shows the number of instances (on the vertical axis) for which the gap at termination is smaller than the value reported on the horizontal axis. We see that the approaches proposed in this paper largely outperform `CL` and `DRC`: in fact, for the latter two the optimality gap goes up to more than 100%. Instead, it remains below 20% for `Compact`, below 40% for `Benders` and below 70% for `B&C`.

Similar considerations can be observed with respect to Figure 3 which reports the average gap between the lower bound at termination and the best upper bound among all the five compared approaches: `Compact`, `Benders` and `B&C` are always below 20% while `CL` and `DRC` go above 40% and 50%, respectively. This means that our approaches are more effective in producing good lower bounds.

A symmetric analysis is done to verify the quality of the upper bounds produced by the different approaches. Figure 4 reports the average gap between the upper bound at termination and the best lower bound among all approaches. Again, `Compact` is by far the best approach while `DRC` behaves much worse than the second worst, which is `CL`. Specifically, `Compact` produces upper bounds which are always within at most 20% of the best lower bound. The maximum gap increases to at most 35% for `Benders` and to slightly less than 60% for `B&C`. Instead, for `CL` and `DRC` the maximum gaps are slightly more than 60% and almost 120%, respectively.

Overall, Figures 2–3 show that the approaches proposed in this paper are much more effective than `CL` and `DRC` in producing good upper and lower bounds in a shorter computing time. `DRC` is largely outperformed on both upper and lower bound analysis. Focusing on upper bounds, `CL` behaves similar to `B&C` but is outperformed by `Compact` and `Benders`. Finally, `Compact` is the best methodology on all former statistics, which seems to be contradicting the outcome of Table 3, where it was shown that connectivity constraints are indeed helpful in strengthening the formulation. We will provide a more detailed analysis on this aspect afterwards.

Lastly, Figure 5 reports the number of instances solved to optimality within a given computational time. Here the picture changes: Both `CL` and `DRC` dominate the three approaches proposed in this paper in terms of number of instances solved. Among the approaches we propose, `Compact` and `Benders` are comparable and dominate `B&C`. This seems to be contradicting what has been seen in Figure 2–3. We will explain the reason afterwards.

Table 4 provides some statistics about computations performed with

Compact, B&C and Benders. In particular, the average computational time (in seconds) and the number of branch-and-bound (B&B) nodes explored are reported for each method. In addition, the average number of GSECs inserted is reported for B&C and, similarly, the average number of Benders cuts is reported for Benders. The table is organized as Table 3. From Table 4 we can see that Compact explores a much larger number of B&B nodes than B&C and Benders in a lower or comparable time, which is expected as no separation of GSECs or Benders cuts is done, so the solution of each B&B node is faster. This explains why, even if from Table 3 we know that GSECs and connectivity constraints strengthen the relaxation, still Compact behaves better in terms of lower bound quality (see Figure 4).This, together with the fact that the load based model includes the complete information about the structure of feasible solutions, also explains a better quality of the upper bounds.

We unfortunately do not have statistics related to the number of B&B nodes explored by CL, but we can reasonably argue that it is much lower than the number of nodes explored by Compact given that the corresponding formulation is disaggregated so the corresponding linear relaxation is much heavier. Instead, in Desaulniers et al. (2016) the number of B&B nodes explored by DRC is reported and we can notice that it is nearly one order of magnitude lower than the one associated with Compact. On one side, this explains the better quality of primal and dual bounds associated with the approaches proposed in this paper. On the other side, this also explains why, despite providing better lower and upper bounds, Compact, Benders and B&C close a lower number of instances to optimality: as the size of the B&B tree increases much faster, a larger number of B&B nodes are open so it becomes more difficult to close all of them and to prove the optimality.

Finally, we note that a new exact approach for the IRP has been proposed in Manousakis et al. (2021). It is based on a two-commodity formulation strengthened through a set of valid inequalities, some of them being in exponential number and thus, giving rise to a branch-and-cut algorithm. In addition, the authors implemented a Tabu Search (TS) heuristic to compute an initial upper bound which is provided as warm start for the branch-and-cut. We decided not to include this algorithm, called from now on MRZT, in the comparison done above as the initial heuristic upper bound plays a crucial role and makes the comparison unfair with respect to all other approaches included in the comparison. Indeed, the advantages are multiple:

Table 4: Statistics on solvers' behavior

		Compact		B&C		Benders	
		Time	# nodes	Time	# nodes	Time	# nodes
Low inventory cost							
H	3	4212	72627	4925	30785	4302	59947
	6	5693	132127	5885	106194	5636	83728
m	2	3106	48762	4151	19451	3087	38788
	3	4907	99582	5236	54369	4879	70634
	4	5455	112316	5883	78831	5585	85366
	5	5601	119097	5870	83602	5658	80670
n	5	354	133584	500	176053	36	20055
	10	2679	154436	2894	136399	2849	161495
	15	4334	146341	5205	102437	4543	130347
	20	5686	105096	6405	24590	5677	81828
	25	6194	83725	6475	14413	6211	69229
	30	6287	45043	6896	7755	6386	29403
	35	5614	50868	6542	7787	5688	35651
	40	6427	51007	7205	6580	6552	36830
	45	6321	42922	6851	4405	6303	26295
	50	6849	37781	7208	2944	6890	18344
Av. *Low*		4767	94939	5285	59063	4802	68865
High inventory cost							
H	3	4285	58191	4888	25035	4293	42318
	6	5472	105209	5775	105415	5536	82245
m	2	3142	41926	4161	16940	3086	34482
	3	5077	74283	5231	35198	5012	54046
	4	5309	88474	5772	70280	5488	71426
	5	5394	98608	5719	98293	5449	69208
n	5	131	63587	196	136223	16	8187
	10	2347	131793	2761	167639	2712	152743
	15	4346	124775	5212	90126	4192	94886
	20	5580	82655	6238	19175	5733	66977
	25	6292	67229	6556	11980	6371	51994
	30	6289	44266	6895	7241	6238	25301
	35	5558	51588	6382	6640	5577	30713
	40	6512	51769	7206	5149	6616	37978
	45	6761	46888	7016	3641	6533	29956
	50	6892	34311	7209	2649	6890	17828
Av. *High*		4731	75823	5221	55178	4759	57291
Total av.		4749	85381	5253	57121	4781	63078

- the upper bound found by the initial heuristic might be better than any solution found by the pure branch-and-cut,
- a good initial upper strongly helps in pruning the branch-and-bound tree and thus improving the lower bound and the overall branch-and-cut performance.

Anyway, `MRZT` is the new state-of-the-art approach for the solution of the IRP. We briefly mention that it solves to optimality 394 instances out of 640, the average optimality gap is 0.85% and the average computing time is 2973 seconds (a time limit of two hours was set in the experiments). Thus, `MRZT` largely outperforms all former exact solution approaches. We refer the reader to Manousakis et al. (2021) for details on computational results.

7 Conclusions

We presented a polyhedral study of aggregated formulations for the IRP, i.e., formulations in which vehicle index is discarded. We show that compact formulations using flow variables that ensure capacity and connectivity constraints provide a lower bound, in terms of value of the LP-relaxation, which is equivalent to the one provided by modelling capacity constraints through fractional capacity cuts, which are exponential in number. We also provide a strengthening of the load-based formulation corresponding to the adaptation to the IRP of the multi-star inequalities proposed for the capacitated VRP and we again show that a compact formulation of the multi-star inequalities provides the same lower bound as the formulation using exponentially many constraints. In addition, we show that no advantage is gained when considering disaggregated formulations, i.e., formulations where variables have vehicle index: the value of the LP-relaxation is equivalent to the one of aggregated formulations. To the best of our knowledge, this is the first study that shows the link between aggregated vs. disaggregated and compact vs. exponential formulations for the IRP. Most of the formulations analysed in this paper have been used in former contributions to the IRP. Thus, this analysis provides a picture of the links between them. It also shows that nothing is gained when moving from an aggregated to a disaggregated formulation and, thus, there is a clear advantage in using a smaller number of variables (and constraints) involved in the aggregated formulations.

We perform exhaustive computational tests on benchmark IRP instances comparing different aggregated formulations with two state-of-the-art approaches. The results show that the aggregated formulations are competitive in terms of values of both upper and lower bounds. In particular, the best aggregated formulations is the compact formulation.

As a future research direction, it might be interesting to study the link between the aggregated formulations presented in this paper and the two-commodity flow formulation recently proposed in Manousakis et al. (2021), that provides excellent computational results. Also, the analysis of the link between directed and undirected formulations (as the one used in Coelho and Laporte (2014)) could provide further hints in formulations' performance.

References

Adulyasak, Y., Cordeau, J.-F., and Jans, R. (2014). Formulations and branch-and-cut algorithms for multivehicle production and inventory routing problems. *INFORMS Journal on Computing*, 26(1):103–120.

Ahuja, R. K., Magnanti, T. L., and Orlin, J. B. (1993). *Network Flows: Theory, Algorithms, and Applications*. Prentice hall.

Alvarez, A., Munari, P., and Morabito, R. (2018). Iterated local search and simulated annealing algorithms for the inventory routing problem. *International Transactions in Operational Research*, 25(6):1785–1809.

Archetti, C., Bertazzi, L., Hertz, A., and Speranza, M. (2012). A hybrid heuristic for an inventory routing problem. *INFORMS Journal on Computing*, 24(1):101–116.

Archetti, C., Bertazzi, L., Laporte, G., and Speranza, M. G. (2007). A branch-and-cut algorithm for a vendor-managed inventory-routing problem. *Transportation Science*, 41(3):382–391.

Archetti, C., Bianchessi, N., Irnich, S., and Speranza, M. G. (2014). Formulations for an inventory routing problem. *International Transactions in Operational Research*, 21(3):353–374.

Archetti, C., Boland, N., and Speranza, M. G. (2017). A matheuristic for the multivehicle inventory routing problem. *INFORMS Journal on Computing*, 29(3):377–387.

Archetti, C., Guastaroba, G., Huerta-Muñoz, D. L., and Speranza, M. G. (2021). A kernel search heuristic for the multivehicle inventory routing problem. *International Transactions in Operational Research*.

Archetti, C. and Speranza, M. G. (2016). The inventory routing problem: The value of integration. *International Transactions in Operational Research*, 23(3):393–407.

Avella, P., Boccia, M., and Wolsey, L. (2015). Single-item reformulations for a vendor managed inventory routing problem: Computational experience with benchmark instances. *Networks*, 65(2):129–138.

Avella, P., Boccia, M., and Wolsey, L. (2018). Single-period cutting planes for inventory routing problems. *Transportation Science*, 52(3):497–508.

Bertazzi, L. and Speranza, M. G. (2012). Inventory routing problems: An introduction. *EURO Journal on Transportation and Logistics*, 1(4):307–326.

Bertazzi, L. and Speranza, M. G. (2013). Inventory routing problems with multiple customers. *EURO Journal on Transportation and Logistics*, 2(3):255–275.

Chimani, M., Kandyba, M., Ljubić, I., and Mutzel, P. (2010). Orientation-based models for $\{0, 1, 2\}$-survivable network design: theory and practice. *Mathematical Programming*, 124(1-2):413–439.

Chitsaz, M., Cordeau, J.-F., and Jans, R. (2019). A unified decomposition matheuristic for assembly, production, and inventory routing. *INFORMS Journal on Computing*, 31(1):134–152.

Coelho, L. C., Cordeau, J.-F., and Laporte, G. (2012). Consistency in multi-vehicle inventory-routing. *Transportation Research Part C: Emerging Technologies*, 24(1):270–287.

Coelho, L. C., Cordeau, J.-F., and Laporte, G. (2013). Thirty years of inventory routing. *Transportation Science*, 48(1):1–19.

Coelho, L. C. and Laporte, G. (2013a). A branch-and-cut algorithm for the multi-product multi-vehicle inventory-routing problem. *International Journal of Production Research*, 51(23-24):7156–7169.

Coelho, L. C. and Laporte, G. (2013b). The exact solution of several classes of inventory-routing problems. *Computers & Operations Research*, 40(2):558–565.

Coelho, L. C. and Laporte, G. (2014). Improved solutions for inventory-routing problems through valid inequalities and input ordering. *International Journal of Production Economics*, 155:391–397.

Desaulniers, G., Rakke, J. G., and Coelho, L. C. (2016). A branch-price-and-cut algorithm for the inventory routing problem. *Transportation Science*, 50(3):1060–1076.

Gavish, B. and Graves, S. (1979). The traveling salesman problem and related problems. Technical report, Graduate School of Management, University of Rochester, NewYork. Working Paper.

Gouveia, L. (1995). A result on projection for the vehicle routing problem. *European Journal of Operational Research*, 85(3):610–624.

Gouveia, L. E. N. and Hall, L. A. (2002). Multistars and directed flow formulations. *Networks*, 40(4):188–201.

Hoffman, A. (1960). Some recent applications of the theory of linear inequalities to extremal combinatorial analysis. In Bellman, R. and Hall, M., editors, *Combinatorial Analysis*, pages 113–128. American Mathematical Society, Providence, RI.

Letchford, A. and González, J. J. S. (2006). Projection results for vehicle routing. *Mathematical Programming*, 105:251–274.

Letchford, A. N. and González, J. J. S. (2015). Stronger multi-commodity flow formulations of the capacitated vehicle routing problem. *European Journal of Operational Research*, 244(3):730–738.

Ljubić, I. (2021). Solving Steiner trees: Recent advances, challenges, and perspectives. *Networks*, 77(2):177–204.

Ljubić, I., Weiskircher, R., Pferschy, U., Klau, G. W., Mutzel, P., and Fischetti, M. (2006). An algorithmic framework for the exact solution of the prize-collecting Steiner tree problem. *Mathematical Programming*, 105(2-3):427–449.

Manousakis, E., Repoussis, P., and Zachariadis, E..and Tarantilis, C. (2021). Improved branch-and-cut for the inventory routing problem based on a two-commodity flow formulation. *European Journal of Operational Research*, 290(3):870–885.

Padberg, M. and Rinaldi, G. (1991). A branch-and-cut algorithm for the resolution of large-scale symmetric traveling salesman problems. *SIAM Rev.*, 33:60–100.

Roldán, R., Basagoiti, R., and Coelho, L. C. (2017). A survey on the inventory-routing problem with stochastic lead times and demands. *Journal of Applied Logic*, 24:15 – 24.

Santos, E., Ochi, L., Simonetti, L., and González, P. (2016). A hybrid heuristic based on iterated local search for multivehicle inventory routing problem. *Electronic Notes in Discrete Mathematics*, 52:197–204.

8 Appendix

Proof of Theorem 3.

Proof *To show that MS cuts provide a complete description of the projection of the flow into the space of X variables, we start with a solution $(\bar{Z}, \bar{Q}, \bar{X})$ for the formulation based on MS cuts, and we show that there exists a feasible flow $\bar{\ell}$ which satisfies constraints* (14)*. Such a flow exists if and only if there exists* a feasible flow $\tilde{\ell}$ *on the same graph with the lower and upper bounds on the arc capacities ($\underline{\kappa}$, $\overline{\kappa}$, respectively) defined as follows:*

$$\underline{\kappa}_{ij}^t = \bar{X}_{ij}^t, \qquad \overline{\kappa}_{ij}^t = (\mathcal{Q}-1)\bar{X}_{ij}^t, \qquad i,j \in N', t \in T \tag{26}$$

$$\underline{\kappa}_{0j}^t = \bar{X}_{0j}^t, \qquad \overline{\kappa}_{0j}^t = \mathcal{Q}\bar{X}_{0j}^t, \qquad j \in N', t \in T \tag{27}$$

$$\underline{\kappa}_{j0}^t = \bar{Q}_j^t, \qquad \overline{\kappa}_{j0}^t = \bar{Q}_j^t, \qquad j \in N', t \in T \tag{28}$$

As in the proof of Theorem 1, this follows from the fact that the flow demands of $\bar{Q}_j^t$ of each vertex $j \in N'$ are transformed into fixed arc capacities for backward arcs $(j,0)$ entering the depot. The flow $\tilde{\ell}$ is said to be feasible if and only if the following constraints (the flow conservation and the capacity constraints, respectively) are satisfied:

$$\begin{aligned} \tilde{\ell}^t(\delta^-(i)) &= \tilde{\ell}^t(\delta^+(i)), && i \in N, t \in T \\ \underline{\kappa}_{ij}^t \le \tilde{\ell}_{ij}^t &\le \overline{\kappa}_{ij}^t, && (i,j) \in A, t \in T. \end{aligned} \tag{29}$$

Then, according to Hoffman (1960), there exists a feasible flow $\tilde{\ell}$ if and only if

$$\underline{\kappa}^t(\delta^-(S)) \leq \overline{\kappa}^t(\delta^+(S)), \quad S \subset N, t \in T \tag{30}$$

To prove this result, let us consider $t \in T$, and a set S, $\emptyset \neq S \subset N$. We distinguish the following two cases:

1. *$0 \in S$: in that case, $\underline{\kappa}^t(\delta^-(S)) = \bar{Q}^t(S^c) + \bar{X}^t(S^c : S \setminus \{0\})$ and $\overline{\kappa}^t(\delta^+(S)) = \mathcal{Q}\bar{X}^t(0 : S^c) + (\mathcal{Q}-1)\bar{X}^t(S\setminus\{0\} : S^c)$, and so the condition (30) turns into*

   $$\mathcal{Q}\bar{X}^t(0 : S^c) + (\mathcal{Q}-1)\bar{X}^t(S \setminus \{0\} : S^c) \geq \bar{Q}^t(S^c) + \bar{X}^t(S^c : S \setminus \{0\}),$$

   which is the MS cut imposed for the set S^c.

2. *$0 \notin S$: in that case, $\underline{\kappa}^t(\delta^-(S)) = \bar{X}^t(\delta^-(S))$ and $\overline{\kappa}^t(\delta^+(S)) = \bar{Q}^t(S) + (\mathcal{Q}-1)\bar{X}^t(S : S^c)$, and so we have:*

   $$\begin{aligned}\bar{Q}^t(S) + (\mathcal{Q}-1)\bar{X}^t(S : S^c) \geq \\ \geq \bar{Q}^t(S) \geq \bar{Z}^t(S) = \bar{X}^t(\delta^-(S)) + \bar{X}^t(A(S)) \\ \geq \bar{X}^t(\delta^-(S)).\end{aligned}$$

   The second inequality follows from (13) and the equation above follows from (1g).

■