Clustering and routing for a real automotive parts distribution problem

Cassiano Augusto Isler 1, Antonio Carlos Bonassa 2 and Claudio B. Cunha 2,*
1 Escola de Engenharia de São Carlos, 
2 Escola Politécnica, 
University of São Paulo, São Paulo, Brazil

Abstract. This paper addresses a real problem of an automaker to distribute spare parts from a single distribution center (DC) to its dealers and body shops. More specifically we aim to determine the best number of regions for delivering spare parts and clients assigned to each of them in order to minimize the total transportation cost, and then route a set of vehicles within each region. Our approach divides the problem into two stages: firstly, clients are clustered into groups by solving a Capacitated p-median Problem (CPMP) for different values of p using a heuristic based on a novel Genetic Algorithm (GA); secondly, in order to better estimate these distribution costs, we solve to optimality a series of Capacitated Vehicle Routing Problems (CVRP), one for each cluster. A practical application to a real world problem involving a major automaker in Brazil is also described.

Keywords: automotive parts distribution; capacitated vehicle routing; capacitated p-median; genetic algorithm; heuristics

Introduction

In a context of global competition and decreasing profits from product sales, after-sales services and activities (i.e. those taking place after the purchase of the product and devoted to support customers in the usage, maintenance and disposal of goods) constitute a relevant source of profit as well as a key differentiator for manufacturing companies and resellers. According to Gaiardelli et al. (2007) the automaker industry has one of the most advanced after sales operations among the durable goods; in addition, the after-sale service in the European market contribute for up to 40-50% of total revenue, with a gross profit resulting from service (14%) and parts (39%) that is significantly higher than the one resulting from sales of the new cars (Bohmann et al., 2003). Thus, it is essential that spare parts are timely made available to customers such as dealers, authorized repair shops, spare-part resellers, independent garages, petrol pump stations, etc.

In this paper we deal with a real-world problem of a Brazilian automaker that has to decide on how to distribute spare parts to its customers (mainly car dealers and body shops) from a single distribution center (DC) on a timely and efficient manner. Orders received are fulfilled on a daily basis from a single DC that centralizes the whole inventory and from where spare parts are packed and shipped. Once orders are received and processed, they are then grouped according to their final destinations based on a pre-defined number of regions. These regions are known in advance, and determined based on the different geographic areas covered by various freight carriers or

*Correspondence: Claudio B. Cunha, Department of Transportation Engineering, Escola Politécnica, University of São Paulo, Av. Prof. Almeida Prado, nº1280 - Butantã, São Paulo - SP, 05508-070, Brazil.
E-mail: cbcunha@usp.br
logistics providers specialized in providing efficient and reliable services hired on a long-term basis to deliver the parts to the final clients. Oftentimes, the daily total amount of products to be shipped to a region exceeds the capacity of a single truck; consequently, the required number of trucks to service each region may vary from day to day. Direct shipments from the DC to the final clients depend on the total amount of freight to each region, as well as on the number of stops, their locations and relative closeness in terms of distances, drop sizes, and particularities of each region such as road network, city and parking constraints, among others. It is common for farther areas to have larger trucks performing long distance, line haul transference to a local or regional cross-docking or transit point facility from where products are transferred to smaller vehicles for the immediate last-mile distribution to the final clients. Given that, in times of dealerships and shops holding low inventory levels, spare parts are usually needed as fast as possible to avoid inconvenient delays with repairs of dry docked vehicles, it is considered unacceptable that the delivery of some shipments has to be postponed to the following day due to some vehicle or capacity unavailability.

Determining the proper number of regions in which a country with continental dimensions like Brazil, whose geographic area is larger than the continental United States, should be divided is a matter that has challenged different players in the automotive industry, including third-party logistics providers that are commonly hired to provide comprehensive spare parts services, including procurement, warehousing and distribution. Few regions allow more efficient (in terms of frequencies and costs) long distance, line haul transferences to local transshipment facilities from where the parts are then distributed to the final destinations; on the other hand, it means larger areas to be covered from these facilities, thus increasing the cost of the last mile distribution. Freight transportation costs have recently increased in Brazil due not only to the economic growth and the reduction in unemployment but also and mainly due to new truck drivers' regulations that are in force. These regulations impose that truck drivers will only be allowed to work for eight hours before they would then be required to rest for eleven hours. They will also be required to rest for thirty minutes after each four hours behind the wheel. The goal of these new regulations is to reduce driver fatigue and consequently the number of accidents caused by excessive hours behind the wheel, a common practice especially by independent owner/operators who generally work by themselves and had to stay behind the wheel sometimes for 24 hours straight in order to meet deadlines imposed by company owners, who prefer to contract independent truck owners than to hire employees in order to reduce labor costs, especially taxes and benefits.

In this paper we describe an efficient hybrid Genetic Algorithm (GA) approach to solve a real world problem of determining the best number of regions for delivering spare parts and clients assigned to each of them aiming to minimize the total transportation cost. In our approach, we divide the problem into two stages. Firstly, the clients to be served by the DC are clustered into groups by solving a Capacitated p-median Problem (CPMP) for different values of p using a heuristic based on Genetic Algorithms (GA); the main goal of this stage is to determine clusters of clients that are compact and homogeneous in size. Our GA approach embeds novel mechanisms to allocate customers to the selected medians, as well as a local search procedure that is applied to improve the quality of each generated individual, thus characterizing a hybrid GA, sometimes referred to as memetic algorithm (Moscato, 1989). It also considers restarting the GA whenever it prematurely stagnates in a solution. In order to validate our GA heuristic for the CPMP, we compare our results with instances from the literature.

Though the resulting direct shipment or the line haul component of the transportation cost (i.e., between the DC and each transshipment point) can thus be determined accurately, the same does not apply to the eventual distribution costs, particularly in the case of LTL shipments from the DC that are consolidated and travel together towards the region, to be then transshipped to smaller vehicles that are more appropriate for the last mile distribution, especially in urban areas. Thus, in order to better estimate these distribution costs, including the required number of trucks, since split deliveries are not allowed, in the second stage we solve a series of Capacitated Vehicle Routing Problems (CVRP) (Toth and Vigo, 2002), one for each cluster. Since the resulting number of clients assigned to each cluster is not relatively large, we are able to model the CVRP as an integer programming (IP) formulation and solve it to optimality. The proposed approach was applied to a real-world problem of distributing spare parts for a major automobile manufacturer in Brazil.

The remainder of this paper is organized as follows: we initially present a brief literature review. The proposed GA approach for the CPMP is then detailed, followed by the section where we present the results of the computational experiments using instances from the literature for validation. In the fifth section, we described the application to a real-world problem of distributing spare parts to an automaker in Brazil. Finally, the last section presents the concluding remarks.
Literature review and problem definition

The capacitated \( p \)-median problem (CPMP) is NP-hard and has wide-ranging applications in clustering, designing transportation and telecommunication networks (Garey and Johnson, 1979). The CPMP is a location problem where a set of \( n \) customers is partitioned into \( p \) disjoint clusters such that the total dissimilarity within each cluster is minimized subject to constraints on maximum cluster capacity (Fleszar and Hindi, 2008). Dissimilarity of a cluster is the sum of dissimilarities between each customer which belongs to the cluster and the median associated with it. If there are no capacity constraints, the problem reduces to the classical \( p \)-median problem (PMP) (see, for instance, Hansen and Mladenovic, 1997).

The CPMP arises in different contexts such as political districting (Bozkaya et al. 2003), design of sales force territories (Mulvey and Beck, 1984), vehicle routing (Koskosidis and Powell, 1992), and the design of a distribution network, where a set of customers is to be supplied from supply points (Fleszar and Hindi, 2008).

The CPMP can be defined on an undirected graph \( G = (V, A) \) where vertices \( \{V\} \) contain all the demand points \( (n) \) connected by arcs of \( \{A\} \) (Correa et al., 2004). The problem aims to find the set of vertices \( V_p \) \( (V_p \subseteq V) \) that minimizes the sum of costs between the vertices of \( \{V-V_p\} \) and its relational vertices on \( \{V_p\} \), always respecting the capacity constraints. The CPMP can then be modeled as (Lorena and Senne, 2003):

\[
\min \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} \cdot w_{ij} \tag{1}
\]

subject to

\[
\sum_{j=1}^{n} w_{ij} = 1, \quad i = 1, 2, \ldots, n \tag{2}
\]

\[
\sum_{j=1}^{n} z_{j} = p \tag{3}
\]

\[
\sum_{i=1}^{n} q_i \cdot w_{ij} \leq C_j \cdot z_j \quad j = 1, 2, \ldots, n \tag{4}
\]

\[
w_{ij} \in \{0,1\}, \quad i, j = 1, 2, \ldots, n \tag{5}
\]

\[
z_{j} \in \{0,1\}, \quad j = 1, 2, \ldots, n \tag{6}
\]

where \( n \) = total number of vertices (demand points) \( (n = |V|) \); \( p \) = number of demand points chosen as medians; \( d_{ij} \) = dissimilarity (distance) between vertexes \( i \) and \( j \); here considered as the Euclidean distance between each pair of clients; \( q_i \) = demand (weight or load) of vertex \( i \); \( C_j \) = capacity of the vertex (median) \( j \). Let \( z_{j} = 1 \) if vertex \( j \) is selected as median, and 0 otherwise; and \( w_{ij} = 1 \) if vertex \( i \) is assigned to median \( j \) and 0 otherwise.

The objective (1) is to minimize the sum of dissimilarities related to the connection between the chosen medians \( \{V_p\} \) and the vertices \( \{V-V_p\} \) allocated to them. Constraints (2) ensure that each demand point of \( \{V-V_p\} \) is allocated to only one median, while constraint (3) imposes that the total number of clusters equals \( p \). Constraints (4) ensure that the sum of demands assigned to each median does not exceed its capacity; they also prevent the assignment of customers to unselected medians. Finally, constraints (5) and (6) provide the binary condition on the decision variables.

A heuristic based on Lagrangean relaxation was proposed by Lorena and Senne (2003) to solve a version of the CPMP. The method improves relaxed solution by swapping \( p \)-medians with demand vertices and also vertices from one median to another one. Lorena and Senne (2004) applied a column generation strategy to solve the problem, in which they also made use of the Lagrangean relaxation in a routine named Lagrange/Surrogate Approach in order to accelerate the column generation process. The authors solved a CPMP with the same approach for large scale linear programming problem and applied it to real instances from the city of São José dos Campos in Brazil. Correa et al. (2004) propose a Genetic Algorithm variation containing a "hyper-mutation" operator that works along with the conventional GA operators. Applications of the method with and without the hyper-mutation in real instances are compared with the results of a Tabu Search algorithm (Tabu Search), which indicate that the new operator considerably improves the final results.
Reese (2006) presents a literature review about the non-capacitated and capacitated $p$-median problems, where mathematical formulation and constraints can be consulted for further information. The author highlighted several different approaches to solve variants of the $p$-median problem using heuristics and meta-heuristics such as Variable Neighborhood Search (VNS), Genetic Algorithm (GA), Simulated Annealing (SA) and Scatter Search (SS). Other researches significantly improved the approaches to solve the PMP and CPMP. Pankratz (2005) applied a variation of GA to cluster a set of vertices in order to deal with a pickup and delivery problem with time window; and Mulvey and Beck (1984) and Osman and Christofides (1994) proposed an hybrid heuristic Simulated Annealing (SA) and Tabu Search (TS) to solve the CPMP. Recent papers include Fleszar and Hindi (2008) with the Variable Neighborhood Search (VNS) heuristic and Chaves et al. (2007) with a Clustering Search (CS) and VNS approach.

Ming and Cheng (2007) point out that subdividing a population in groups (clusters) generates sets with the highest level of similarity among its member than with any individual from outside cluster. Fleszar and Hindi (2008) present similar definition when stating that the solution of a $p$-median problem should contain a certain number of clusters whilst dissimilarity of cluster members in relation to a particular feature is minimized. The capacitated vehicle routing problem (CVRP) is one of the most studied problems in operations research and is an extension of the Traveling Salesman Problem (TSP) and is therefore NP-hard (Cordeau et al., 2007). The books by Toth and Vigo (2002) and Golden et al. (2008) provide comprehensive surveys on this problem and efficient exact methods were proposed by Fukasawa et al. (2006), Baldacci et al. (2008) and Baldacci and Mingozzi (2009). Given the NP-hardness of the CVRP, several heuristics and metaheuristics are proposed to solve it efficiently, as in Prins (2004), Mester and Braysy (2007), Pisinger and Ropke (2007), Nagata (2007) and Szeto et al. (2011).

The CVRP can be defined on a directed graph $G = (V, A)$ where $V = \{0, 1, \ldots, n\}$ is the set of vertices and $A = \{(i, j); i, j \in V, i \neq j\}$ the set of arcs. It is assumed that $|V| = n + 1$, that node “0” represents the depot and the remaining $n$ nodes represent geographically dispersed customers. Each customer has a given demand and the total demand of customers to be visited by a vehicle; $N = C \cup \{0, n+1\} = \text{set of customers of the CVRP}$, where customer “0” and customer “$n+1$” relate to the depot (median of the CPMP); $d_i$ = demand of customer $i$, where $d_0 = 0$ and $d_{n+1}=0$; $E= \{(i, j); i, j \in N; i \neq j, i \neq n+1, j \neq 0\} = \text{nodes connections arcs}$.

Let

\[
X_{ijk} = \begin{cases} 
1, & \text{if vehicle } k \text{ traverses arc } (i, j), \forall k \in K, (i, j) \in E \\
0, & \text{otherwise} 
\end{cases}
\]

The CVRP can then be modeled as (Ball et al., 1995):

\[
\text{Min } \sum_{k \in K} \sum_{(i,j) \in E} c_{ij} \cdot x_{ijk} 
\]

Subject to

\[
\sum_{j \in N} x_{ijk} = 1, \quad \forall i \in C \tag{8}
\]

\[
\sum_{j \in C} x_{ijk} \leq Q, \quad \forall k \in K \tag{9}
\]

\[
\sum_{j \in N} x_{0,j,k} = 1, \quad \forall k \in K \tag{10}
\]

\[
\sum_{j \in N} x_{ijk} - \sum_{j \in N} x_{j,k} = 0, \quad \forall h \in C, \quad \forall k \in K \tag{11}
\]

\[
\sum_{i \in N} x_{i,n+1,k} = 1, \quad \forall k \in K \tag{12}
\]

\[
y_i - y_j + n \cdot x_{ijk} < n - 1, \quad \forall i \in C, \quad \forall j \in C, \quad \forall k \in K \tag{13}
\]

\[
y_i \in \mathbb{R}^+, \quad \forall i \in N \tag{14}
\]

where $c_{ij} =$ cost (Euclidean distance) to traverse arc $(i, j)$; $Q =$ vehicle capacity constraint; $C = \{1, \ldots, n\} =$ set of customers to be visited by a vehicle; $N = C \cup \{0, n+1\} =$ set of customers of the CVRP, where customer “0” and customer “$n+1$” relate to the depot (median of the CPMP); $d_i =$ demand of customer $i$, where $d_0 = 0$ and $d_{n+1}=0$; $E= \{(i, j); i, j \in N; i \neq j, i \neq n+1, j \neq 0\} =$ nodes connections arcs.
In the above formulation, Equation (8) ensures that each customer is served by one, and only one vehicle. Equation (9) represents the vehicle capacity constraints and Equation (10) determines the start point of a vehicle routing as the customer “0”. The route continuity related to the flow balance in each customer is represented by Equation (11) while Equation (12) sets the end point of a route as the customer “n+1” (also the median of a cluster). Equation (13) represents the sub-tour breaking constraints in order to avoid sub-cycles assigned to a vehicle. Equation (14) defines continuous auxiliary variables to avoid sub-cycles in real domain.

Christofides et al. (1981) developed an exact algorithm to solve the CVRP using Lagrangean bounds from minimum q-route sub problems. A q-route is a journey that starts at a depot, traverses a sequence of clients with total demand at most equals to C, and returns to the same initial node. Some clients may be visited more than once, so the set of valid CVRP routes is strictly contained in the set of q-routes. Fukasawa et al. (2006) combine Branch-and-Cut and Lagrange Relaxation to solve to optimality CVRP instances from the literature with up to 135 nodes. Although there are some heuristic approaches to solve the CVRP for large instances, in that paper the authors solved an automaker problem to its optimality using a commercial software.

In the following sections we provide the details about our novel approach to solve the problem which is divided into two stages: firstly, clients are clustered into groups by solving the CPMP for different values of p using a novel GA; secondly, in order to better estimate these distribution costs, we solve to optimality a series of Capacitated Vehicle Routing Problems (CVRP), one for each cluster.

A genetic algorithm for the CPMP

Genetic algorithms (GAs) are one of the most popular heuristic algorithms that represent a powerful and robust approach for developing heuristic for complex and large-scale combinatorial optimization problems. Initially proposed by Holland (1975), GA can be described as a probabilistic search, which imitates the process of natural selection and evolution to evolve a population of initial solutions. Each solution of a problem is treated as an individual, whose fitness is governed by the corresponding objective function value and eventual penalization to infeasibility (Correa et al., 2004). Pairs of individuals of a given population are selected as parents and reproduced to generate the next population of new and hopefully better individuals through a structured yet randomized information exchange known as crossover operator. Diversity is added to the population by randomly changing some genes (mutation operator). As new “offsprings” are generated, unfit individuals in the population are replaced using the concept of survival of the fittest. This evaluation–selection–reproduction cycle is repeated until a satisfactory solution is found or other stopping criteria are met.

GAs can be implemented in a variety of ways. The excellent books by Goldberg (1989), Davis (1991) and Holland (1975) describe many possible variants of GAs. We also refer to these books for various GA definitions and notations as chromosomes, alleles, genes, reproduction, etc., as well as for other problem specific operators. Our GA approach embeds novel mechanisms to allocate customers to the selected medians, as well as a local search procedure that is applied to improve the quality of each generated individual, thus characterizing a hybrid GA, sometimes referred to as memetic algorithm (Moscato, 1989). It also considers restarting the GA whenever it fails to converge.

Let \( H = \{h_1, h_2, \ldots, h_l\} \) be the set of individuals in each of the GA’s generations. An individual is defined by its chromosome \( \{V_p\} = \{v_1, v_2, \ldots, v_p\} \) of dimension \((p \times 1)\) containing the vertices chosen as medians. The set \( \{S_q\} = \{s_1, s_2, \ldots, s_q\} \) \((n \times 1)\) is also associated to each individual where its \(i^{th}\) position refers to the index of a client, and \(s_i\) corresponds to the median to which the client is assigned. In such representation the integer value “-1” is assigned to the vertex chosen as the median such that \(s_i = -1 \forall i = v_1, v_2, \ldots, v_p\), as shown in the example of Figure 1, where \(\{V_p\} = \{v_1, v_2\} = \{4, 6\}\) and \(\{S_q\} = \{6, 4, 4, -1, 6, -1\}\).

We also define the set \(Q_p = \{Q_1, Q_2, \ldots, Q_p\}\) whose dimension is \((p \times 1)\) that contains real numbers that represent the total demand of customers assigned to each median. As each individual has its own fitness value related to the sum of distances between each \(p\)-median and its respective customers, the individuals with the lowest values are considered the strongest candidates to a solution for the CPMP.
The input data are: customers’ coordinates (longitude and latitude) and demand, number of medians \((p)\) and maximum capacity of each median. The capacity of a cluster has been considered as a predefined value to validate the GA from instances of the literature. When applying the algorithm to solve the real distribution problem, this capacity is given by the overall demand of the clients divided by a predefined number of medians. The Euclidean distance between a median and its vertices is the proxy to calculate the fitness function (sum of distances between the medians and their respective vertices) of each individual of the population.

The task of finding the optimal population size \((f)\) that guarantees the variability of the population and that avoids high computation times has been extensively discussed in the literature (Alp et al. 2003; Koumousis and Katsaras, 2006; Alander, 1992; Piszcz and Soule, 2006).

Some authors agree that the population size is directly related to the number of instances (Harik and Lobo, 1999; Yu et al., 2006). Gotshall and Rylander (2000) propose that the optimal size of a population is the point in which the benefits from the convergence to good solutions with a low number of generations are balanced with the benefits of better accuracy of the solutions provided by increasing the size of the population. Equation (15) has been defined from Gotshall and Rylander (2000) conclusions that the optimal population size may assume a logarithmic function of the number of customers to be allocated \((n)\), and has been calibrated from empirical experiments based on the validation data of the following section.

\[
f = \text{EVEN}[17.5, \ln(n)]
\]  

(15)

The logarithmic term of Equation (15) provides a reasonable population size when the number of vertices to be allocated is small (avoiding the premature convergence of the algorithm) and prevent too large populations and, thus, high computation times, when there are too many vertices to be allocated. The “EVEN” operator is considered to ensure that the crossover operator is applied to all the individuals.

The steps involved in our GA heuristic for the CPMP are as follows:

1. Generate an initial population of randomly constructed solutions. For each individual, \(p\) distinct vertices are randomly chosen with a given uniform probability of \(1/(n-p)\), and the corresponding position in the demand array \(Q_p\) is updated. The remaining customers are allocated to the medians as described below. This procedure is repeated until the predefined number of individuals of the population is reached.

2. Assign customers to medians. For each new generated individual, the assignment of the \((n-p)\) customers to the \(p\) selected medians is based on the following empirical rule: the 5% larger customers in terms of demand are sequentially assigned to their nearest median, in descending (or non-increasing) order of demand, subject to capacity constraint; in other words, each node is assigned to closest available median (second, third, fourth, etc.) that satisfies capacity constraint. The remaining customers (i.e., 95% of \((n-p)\)) follow the same procedure, but instead of being sequentially allocated according to this order, they are randomly selected, one by one, with a uniform probability \((1/0.95(n-p))\) and assigned to their closest median that is feasible in terms of capacity. This 5% parameter was empirically determined in preliminary computational experiments using CPMP instances from the literature as described in the following section.

3. Apply the crossover operator. The pairs of individuals to be crossed are randomly chosen with uniform probability of \(1/(f-k)\), where \(k\) is the number of individuals that have already been crossed; in other words, no priority is given to any individual (i.e., high quality individuals can be crossed with those of low quality). Two children solutions are generated for each pair of parents, by applying a crossover operator. This is accomplished by identifying the \(s\) medians that are not common to both parents, then randomly choosing the number of medians to be swapped (an integer between 1 and \(s\) with same probability of being...
chosen) and finally choosing the medians to be swapped with equal probability equal to $1/p$. After the changes of medians, the assignment of the remaining customers of each produced individual follows the same procedure described in Step 1 and their respective quality, given by their fitness functions, are then calculated.

4. Replace the individuals of low quality. All the new individuals produced by GA operator are pooled into the same set and sorted in increasing order of the fitness value. We use an elitist generation replacement in which only the $f$ best individuals of this sorted set (with lower values of fitness) are taken to the next generation.

5. Apply the mutation operator. After every five consecutive crossovers, $5\%$ of the individuals (as proposed in Srinivas and Patnaik, 1994) are chosen with equal probability of $1/p$. Each of these individuals are cloned and, for each of them, a number $k$ between 1 and $p$ is randomly chosen with the same probability $(1/p)$ and this amount of medians ($k$) are replaced by other customers that are not medians, chosen with probability of $1/p$ and $1/(n-p)$ respectively. After swapping the medians, the remaining customers are assigned to the new medians as stated in Step 2 and the fitness function of the new individuals are calculated. After every mutation replace the individuals of lower quality as described in Step 4.

6. Apply two local search heuristics to the best individual (i.e., with lowest fitness) in each new generation of the GA. The first procedure aims to find improved solutions by swapping the clients from one cluster to another. All pairs of clients assigned to different medians are tested. The swap is performed if it is feasible in terms of capacity constraints of both medians and the sum of the distances from the two nodes to the new medians is reduced (when compared to the current assignment). In the second local search we search for a new median within the nodes of a given cluster such that the total dissimilarity procedure explores potential improvements from selecting a different median among the nodes that form a cluster if the total dissimilarity is reduced. This is repeated for all $P$ clusters. In order not to slow down the GA, these two improvement procedures are applied sequentially and only once.

7. Renew the current population whenever the GA fails to find a new improved best solution after a certain given number of consecutive generations: a new population of the same size of the original population ($f$) is then generated, keeping only the individual corresponding to the best found solution so far. Half of this new population ($0.5 \cdot f$) is composed by clones of this best individual obtained by randomly changing their features in the following way: for each of these clones, we first randomly determine the number of medians to be replaced by other nodes, being this probability uniformly distributed between 1 and $p$; all non-median nodes are candidates with unbiased equal probabilities and the assignment of nodes to the new set of medians is performed as in Step 2. Finally, the remaining individuals of the population are randomly generated as described in Step 1.

8. Steps 2–7 are repeated until a pre-defined number of generations is reached. As the literature does not provide any rule to set a termination criterion which guarantees the convergence of the GA to a good solution in reasonable computation time, this value has been set empirically as the size of population divided by 15, or $(f/15)$.

**Computational experiments**

Our GA heuristic for the CPMP was coded in Java programming language. In order to validate it, it was used to solve instances from literature described in Lorena and Senne (2003) and available at http://www.lac.inpe.br/~lorena/instancias.html. All experiments were conducted on a Core i3 2.13 GHz computer running under Windows 7. In Table 1, we present our results and compare them with those obtained by a Variable Neighborhood Search (VNS) heuristic by Chaves et al. (2007) executed on a Pentium IV 3.02 GHz.

The results show our GA heuristic yields good solutions. Despite not reaching the optimal solutions for the larger instances, the proposed method provides results close to the best known solutions. One of the main advantages are the lower running times as shown in Table 2.

In most cases the GA reached the termination criterion faster than the Chaves et al. (2007), which indicates that the algorithm provides promising results in terms of computation time given the quality solutions presented before. Actually, the solutions have been reached in reduced time (e.g., just 15.4 seconds for the SJC1 against 50.5 seconds from Chaves et al., 2007), except for the case of SJC2, where the difference is 65.2% (approximately 28 seconds) to reach the best GA solution and 37.4% on average when compared with the VNS.
Table 1. GA application results and best solutions comparison – total cost.

<table>
<thead>
<tr>
<th>Instance</th>
<th>n</th>
<th>p</th>
<th>Best Known Solution</th>
<th>VNS</th>
<th>GA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Best Solution</td>
<td>Best Solution</td>
</tr>
<tr>
<td>SJC1</td>
<td>100</td>
<td>10</td>
<td>17,288.99</td>
<td>17,288.99</td>
<td>17,289.00</td>
</tr>
<tr>
<td>SJC2</td>
<td>200</td>
<td>15</td>
<td>33,270.94</td>
<td>33,270.94</td>
<td>33,453.12</td>
</tr>
<tr>
<td>SJC3a</td>
<td>300</td>
<td>25</td>
<td>45,335.16</td>
<td>45,335.16</td>
<td>45,735.74</td>
</tr>
<tr>
<td>SJC3b</td>
<td>300</td>
<td>30</td>
<td>40,635.90</td>
<td>40,635.90</td>
<td>41,167.38</td>
</tr>
<tr>
<td>SJC4a</td>
<td>402</td>
<td>30</td>
<td>61,925.51</td>
<td>61,925.51</td>
<td>63,234.99</td>
</tr>
<tr>
<td>SJC4b</td>
<td>402</td>
<td>40</td>
<td>52,469.96</td>
<td>52,469.96</td>
<td>53,393.96</td>
</tr>
</tbody>
</table>

1 Chaves et al. (2007) from 10 replications; 2 Error compared to the best known solution calculated as ([GA best solution – VNS best solution]/VNS best solution); 3 Average error from 100 replications; 4 Standard Deviation of error from 100 replications; 5 Between VNS and GA.

Table 2. GA application results and best solutions comparison – running time.

<table>
<thead>
<tr>
<th>Instance</th>
<th>n</th>
<th>p</th>
<th>VNS</th>
<th>GA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Time (s)</td>
<td>Time (s)</td>
</tr>
<tr>
<td>SJC1</td>
<td>100</td>
<td>10</td>
<td>50.50</td>
<td>15.4</td>
</tr>
<tr>
<td>SJC2</td>
<td>200</td>
<td>15</td>
<td>44.08</td>
<td>72.8</td>
</tr>
<tr>
<td>SJC3a</td>
<td>300</td>
<td>25</td>
<td>8,580.30</td>
<td>199.8</td>
</tr>
<tr>
<td>SJC3b</td>
<td>300</td>
<td>30</td>
<td>2,292.86</td>
<td>129.5</td>
</tr>
<tr>
<td>SJC4a</td>
<td>402</td>
<td>30</td>
<td>4,221.47</td>
<td>362.7</td>
</tr>
<tr>
<td>SJC4b</td>
<td>402</td>
<td>40</td>
<td>3,471.44</td>
<td>426.9</td>
</tr>
</tbody>
</table>

1 Chaves et al. (2007) to obtain the best solution from 10 replications; 2 Time to obtain the best solution; 3 Compared to the VNS, calculated as ([GA running time - VNS running time] / VNS running time); 4 Average difference from 100 replications compared to the VNS; 5 Standard Deviation from 100 replications compared to the VNS.

The overall results evidence that our GA heuristic yields to good results and is sufficiently fast to be worth being used to solve the CPMP. A more precise comparison of computing times would require both heuristics to be run on the same computer, something that was not possible in this case. Based on the comparative estimates by Dongarra (2008), we can roughly infer that computing times are slightly lower than the VNS heuristic by Chaves et al. (2007), particularly for larger instances.

Real world application of spare parts distribution

We applied our GA heuristic to solve a real problem of automotive spare parts distribution in Brazil. Initially a database with shipping information for 401 valid locations of Brazilian dealers and body shops was made available to the authors, each of which containing the address, ZIP code and state, and the daily demand (in cubic meters or m³) for the 25 working days of January 2011. ZIP codes for these 401 destinations were used to obtain their geographic locations in terms of latitude and the longitude, thus allowing to determine the Euclidean distances between all the pairs of points, including the central distribution center from where all parts are shipped countrywide.
Nine scenarios were created to determine the number of medians (regions or clusters) that minimize the total distribution cost, comprise the transportation costs from the DC to the medians and the last mile routing costs. Each scenario differs from the others with respect to the number of medians \( p \), which is set initially to 10 and increased by 5 to a maximum of 50 medians. This range was set based on discussions with top regional sales and supply chain executives from the company, who were not sure on how to determine the proper number of regions to a country with continental dimensions similar to the United States, but it is also so different in terms of amount of road infrastructure, no other alternative modes of transportation such as railroads to rely on for long distance ground transportation, port inefficiencies that prevent maritime transportation to be used to link coastal capitals, distinct regional availability of truckload (TL) and less-than-truckload (LTL) freight carriers and costs, to mention a few. Truck transportation, until recently regarded as unexpensive and affordable since most carriers relied on independent truck owners that oftentimes charged less than their total costs (disregarding mainly capital costs) due to fierce market competition caused by the nation’s poorly performing economy and a high unemployment rate, has faced significant rises in freight prices in the recent years. This is due not only to significant economic growth but also to new truck drivers’ regulations that have recently been put in force. As mentioned before, these regulations impose that truck drivers will only be allowed to work for eight hours before they would then be required to rest for eleven hours. Prior to that, independent truck owners drivers that are hired on a trip-by-trip basis, usually had to drive very long journeys to cope with deadlines imposed by company owners.

Truck maximum loading capacity has been assumed as the same as the most frequent truck used in the current operation (24.3 m\(^3\)). As it is practically impossible to reach 100% capacity (given the operational constraints as the space required for closing the doors and managing the cargo, for example) this capacity has been reduced to 20 m\(^3\), based on past historic data of shipments. The maximum capacity of a cluster is defined by the total demand of the clients to be serviced divided by the number of medians of the respective scenario, which in turn can be converted to a number of equivalent trucks based on their standard capacity. Thus, the number of trucks to be assigned to a cluster is defined in terms of the capacity of a standard truck and the average demand per cluster. Let \( C \) the practical capacity of a truck, \( c_i \) the demand of customer \( i \); \( p \) the number of \( p \)-medians, and \( \lfloor . \rfloor \) the operator that defines the largest integer less than or equal to the value in it, then the number of trucks assigned to the cluster of the \( j \textsuperscript{th} \) \( p \)-median (\( NT \)) is defined by Equation (16) below.

\[
NT_j = \max \left\{ \left\lfloor \frac{1}{C} \left( \sum_{i=1}^{p} c_i \right) \right\rfloor : \left\lfloor \frac{1}{C} \max \{c_j\} \right\rfloor \right\}
\]

(16)

For example, considering a scenario with 35 \( p \)-medians and average demand per cluster equals to 68 m\(^3\) (from the total demand of all the costumers divided by the number of medians), the number of trucks per cluster is established as 4 (68/20 = 3.40 \( \approx \) 4 trucks). In the same scenario, if the average demand is 85 m\(^3\), then 5 trucks should be assigned to each cluster (85/20 = 4.25 \( \approx \) 5 trucks). Table 3 below exhibits the resulting minimum number of trucks required to serve each cluster in each of the different 9 scenarios analyzed.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median (( p ))</td>
<td>10</td>
<td>15</td>
<td>20</td>
<td>25</td>
<td>30</td>
<td>35</td>
<td>40</td>
<td>45</td>
<td>50</td>
</tr>
<tr>
<td>Capacity (m(^3))</td>
<td>140</td>
<td>100</td>
<td>80</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>Truck equivalence</td>
<td>7</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

As described before, after determining the regions and the customers assigned to each of them by means of solving the CPMP, we then solve to optimality, for each cluster, a CVRP using the formulation aforementioned. We assume that the demand of a customer is split into equal-sized buckets for the cases in which it is greater than the capacity of one standard truck, and the related vertex is split into fictitious nodes (equals to the number of buckets) such that the client may be visited more than once by different trucks in the CVRP solution. Despite being a NP-hard problem, we were able to solve the CVRP to optimality given the resulting small sizes of the instances using the commercial solver Gurobi (http://www.gurobi.com/) under its academic license.
CPMP results

When solving the CPMP applied to the auto parts distribution problem, 100 replications of the proposed GA algorithm have been performed for each scenario and the best solution among them has been selected as starting point of the CVRP. The best solution is the one with the lowest total distribution cost, which concerns to the sum of distances between the DC and each median of a scenario plus the sum of distances between these p-medians and the respective clients assigned to them. When the demand of a cluster requires more than one truck based on Equation (16), the connecting distance between the DC and the cluster median must be multiplied by the number of trucks.

Table 4 presents the best solution for each instance and their related computation times. These results evidence that the scenario with 35 different regions, i.e., $p=35$ results in the lowest total distance to be traveled.

Table 4. GA results applied to the CPMP phase of the auto parts distribution problem.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$P$</th>
<th>Cost GA ($\times 10^8$ km)</th>
<th>Time (s)</th>
<th>Cost from DC to medians ($\times 10^8$ km)</th>
<th>Total Cost ($\times 10^5$ km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>401</td>
<td>10</td>
<td>9.073   9.273   0.137   77.7   80.2   5.6   3.810   12.882</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>401</td>
<td>15</td>
<td>6.515   6.712   0.099   89.5   87.7   4.1   4.337   10.852</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>401</td>
<td>20</td>
<td>5.255   5.363   0.054   96.2   94.6   7.0   4.505   9.760</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>401</td>
<td>25</td>
<td>4.742   4.918   0.069   111.1  102.4   4.0   4.513   9.254</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>401</td>
<td>30</td>
<td>3.814   3.962   0.056   103.9  102.4   4.3   5.024   8.838</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>401</td>
<td>35</td>
<td>3.176   3.380   0.038   102.2  104.9   4.4   5.148   8.324</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>401</td>
<td>40</td>
<td>2.892   2.954   0.028   119.5  108.4   5.5   5.645   8.537</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>401</td>
<td>45</td>
<td>2.586   2.625   0.020   97.2   105.2   5.0   6.549   9.135</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>401</td>
<td>50</td>
<td>2.311   2.354   0.016   117.5  112.4   7.9   6.919   9.230</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From 100 replications; * Standard Deviation from 100 replications; † To reach the minimum cost.

With respect to the computational times, Table 4 shows that, as expected, these values increase as the number of medians increases since the CPMP becomes more complex. However, the observed values can be considered low for practical and planning purposes, reaching a maximum of 112 seconds on average.

Figure 2 depicts the geographical distribution of p-medians and associated vertices for the best GA solution for the extreme scenarios with 10 and 50 p-medians. A visual comparison reveals that when clients are clustered into 10 groups, the clusters cohesiveness is low (Figure 2 - a). On the other hand, for $p=50$ the average number of customers per median is low (Figure 2 - b), the total dissimilarity is lower and the corresponding cohesiveness within the groups seems to be higher but the costs to reach the DC are also higher as the number of required trucks increases when compared to the previous scenarios.

Based on Figure 2, one can visually notice that there might exist an intermediate number of p-medians for which the total costs associated with linking cluster to DC added to the total cost of connecting vertices to its related p-median is minimal. The relationship between the number of p-medians and their dissimilarities, i.e., the trade-off between the total connecting distances from DC to the p-medians and the total distribution distance within the cluster – defines a concave total distance curve, as shown in Figure 3, in which the minimum total cost point relates to the 35 p-medians scenario.

The reduction of total distances in the scenarios from 10 to 35 medians can be explained by the relationship between distribution and connecting distances. The less dispersed the vertices within a cluster, the shorter the distances between them but the higher the total distance to connect the medians to the DC. This can be directly figured out from Figure 3 since from the scenario with 35 medians to the one with 50 medians the sum of distances from the DC to the p-medians decreases and progressively increases distances from these medians to their respective clients.

As can be observed from Figure 3, the scenario among all nine tested for which the total distance is minimized corresponds to 35 medians, and it is geographically represented in Figure 4. One can observe that the demand points are not so dispersed as they are in the 10 p-medians scenario, reducing the distribution costs. On the other hand, there are not so many p-medians as in the scenario with 50 p-medians which reduces the DC to p-medians connecting costs.
Fig. 2. Representation of the GA’s best solutions for 10 (a) and 50 medians (b)
Fig. 3. Cost as a function of the number of medians

Fig. 4. Representation of the 35 \( p \)-medians best solution
Finally, Figure 5 illustrates the size of the clusters that have been obtained from the 35 \( p \)-medians’ scenario which corresponds to the best results in terms of the overall distances between the DC and the medians and between these clusters’ central points and their related customers.

![Cluster sizes of the best solution for 35 \( p \)-medians](image)

**Fig. 5.** Cluster sizes of the best solution for 35 \( p \)-medians

**CVRP results and analysis**

Based on the solution obtained from the GA applied to the CPMP (\( p = 35 \) medians), we could find the optimal solutions of the CVRP for each of the resulting 35 regions. Sometimes we could not find a feasible solution to this problem given that deliveries cannot be split, thus requiring that the resulting number of required trucks from the CPMP had to be increased. In other words, it was not possible, for the last-mile distribution within each region, to find any feasible combination of shipment loads that could fit into the resulting number of trucks used in the CPMP, thus requiring that, in some cases, additional trucks to distribute the shipments to the final clients.

In order to find a feasible solution to all clusters, an auxiliary procedure was developed. Whenever a feasible solution could not be reached from the first CVRP model’s run, we increase the number of trucks assigned to a cluster by one and the solver is executed again, until a feasible solution to the respective median problem is found. Table 5 exhibits, for the best scenario with 35 \( p \)-medians (or regions), four examples of adjustments in the number of trucks required to deliver the loads within a set of 15 clusters which incurred in infeasible solutions in the first execution of the CVRP.

After adjustments on the distribution capacity, the CVRP was able to find a feasible and optimal solution using 90 trucks to satisfy all the 401 demand points. This solution has resulted in higher total costs than those provided by the solution of the CPMP, which required 75 trucks to distribute all the products to their related clients.

To illustrate this, Figure 6 shows the resulting routes for the cluster numbered “69”, in which 3 trucks are required (blank square in Figure 5) to satisfy all customers’ demand. All routes start and end at the \( p \)-median. One truck services 4 customers (270 → 22 → 307 → 91), the second deliveries to 10 clients (378 → 291 → 215 → 363 → 300 → 180 → 129 → 275 → 65 → 175) and the third truck must visit 2 customers (398 → 307). Given that the demand of customer 307 is higher than the capacity of a truck its vertex is visited by two vehicles in order to have its total demand satisfied.
Table 5. Number of trucks of some clusters required by the CPMP and the CVRP.

<table>
<thead>
<tr>
<th>Vertex selected as median</th>
<th>Vertices clustered to median</th>
<th>Vehicles required by CPMP</th>
<th>Vehicles required by CVRP</th>
<th>Vehicle requirements comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>55</td>
<td>18</td>
<td>3</td>
<td>5</td>
<td>+ 2</td>
</tr>
<tr>
<td>255</td>
<td>11</td>
<td>3</td>
<td>4</td>
<td>+ 1</td>
</tr>
<tr>
<td>263</td>
<td>21</td>
<td>3</td>
<td>5</td>
<td>+ 2</td>
</tr>
<tr>
<td>310</td>
<td>17</td>
<td>3</td>
<td>5</td>
<td>+ 2</td>
</tr>
</tbody>
</table>

Following the increase in terms of the number of trucks, the CVRP solution has yielded to larger total distances to be travelled than the total sum of distances from the medians to their related customers in the best scenario of 35 regions. While clustering has generated a total distribution cost of 296,548 km, the trucks routed by solving the CVRP satisfied all the customers with total cost of 315,929 km, which indicates an increasing of 6.5% (19,380 km) on the total distance traveled. One reason for this is that in some cases the loads demanded by the clients of a cluster have been split into parts because they were not able to be carried by a single truck due to its capacity constraint.

Fig. 6. Cluster number “69” distribution routes

Given the average size of the clusters as illustrated in Figure 6, the solution times for each of the groups from the commercial solver range from less than 1 second to more than 60 minutes. From the 35 clusters, 30 of them (85.7%) reached the best solution in the routing phase in less than 1 minute (average of 2.4 seconds, standard deviation of 6.0 seconds, and maximum of 25.6 seconds), the 31st slowest reached the solution in 75.1 seconds (cluster 1 with 18 customers), the 32nd in 94.4 seconds (which corresponds to cluster 16 with 17 customers), the 33rd in 293.2 seconds (19 customers), the 34th in 2137.8 seconds (cluster 33 with 22 customers) and finally the 35th in 3848.1 seconds (cluster 2 with 20 customers).

It should be noted that a single routing model applied to all the 401 customers would not provide an optimal solution to the problem if the clustering step is not considered. Moreover, although the medians provided by the CPMP are still simple customers, from the strategic point of view the decision maker can eventually convert these locations to depots of products coming from the DC and later distributed to the related customers.
Concluding remarks

The results from the validation of the Genetic Algorithm presented in this paper show that our approach provides good solutions to the CPMP. Although it does not reach the optimal solutions for all the instances of validation, the results from Table 1 and Table 2 show that its performance is fairly enough to provide good solutions to the real auto parts distribution problem in Brazil.

Some GA performance tests indicate that the task of allocating demand vertices top-medians satisfying the related constraints have great influence on processing time. The allocation strategy presented in section that describes the proposed GA guarantees the variability of the population to prevent the stagnation in local optimum solutions, proving to be an efficient procedure in order to provide less computational time to obtain good quality solutions. The procedure of random selection of individuals to be crossed over and genes to be mutated also has proved to be efficient to prevent the algorithm to remain in local minima. It should be noted that the strategy of renewing the population influences the quality of the solutions, particularly by keeping the variability of individuals in the population.

The GA solution’s quality seems to be negatively impacted by reductions in the number of individuals involved in the crossover routine, without improvements when the number of individuals mutated increases. Regarding the CVRP step, it was proved that the routing after clustering provided optimal solutions within the groups in low processing times and that the approach presented in this paper is appropriate since the application of a single routing procedure is not able to provide an optimal solution to the routing formulation in reasonable computation times. The results presented in this paper closely represents the distribution strategy used by many companies. For ease of use, the models can be considered as a Decision Support tool to plan the monthly fleet size and its allocation in relation to demand considering volumes of sales forecasting.

As an attempt to provide more refined results, the medians of the CPMP can be considered as depots of products and the costs related to opening these facilities could be added to the model. Considering further modeling improvements, startup procedures can be added to the GA as a way to create higher quality initial solutions implemented as a local search method – such as GRASP (Greedy Randomized Adaptive Search Procedure) – in order to try to reach the best solutions for the test instances presented here and so for the CPMP application case.

References


