Optimization Models for Capacitated Clustering Problems

Marcos Negreiros, Pablo Batista, João Amilcar Rodrigues
Universidade Estadual do Ceará (UECE)
Mestrado Profissional em Computação Aplicada
MPCOMP/UECE-IFCE
Av. Silas Munguba, 1700 – Campus do Itaperi
CEP: 60740-000 – Fortaleza – CE – Brazil
negreiro@graphvs.com.br, opablofernandes@hotmail.com,
joaoamilcar@gmail.com

RESUMO

Apresenta-se um estudo de um conjunto de modelos matemáticos para o Problema de Agrupamento Capacitado. O problema geral aqui discutido, considera que são dados indivíduos (itens) com atributos (peso e coordenadas do espaço euclideano), onde se deseja formar grupos de mínima dissimilaridade e capacidades máximas definidas por grupo de tal forma a atender a um objetivo específico. Os modelos consideram uma variação generalizada do problema de p-Medianas Capacitadas (PpMC) com extração da mediana do próprio conjunto de indivíduos, uma nova formulação capacitada para o problema de agrupamento com o menor-maior diâmetro (PCCMMD) e uma nova formulação de agrupamentos com capacidades distintas (heterogêneos) para o problema de agrupamento com centro geométrico (PACHCG). São discutidas também formas de se tirar o maior proveito das formulações usando os ”solvers” mais adequados.

PALAVRAS CHAVE. Agrupamento, Modelagem, “Solvers".

Tópicos: Modelagem Matemática, Otimização Combinatória

ABSTRACT

This work shows a set of models for the Capacitated Clustering Problem. The general problem here discussed consider a given number of items with attributes (weight and coordinates in Euclidean space), where one wishes to determine minimum dissimilar clusters constrained to a given maximum capacity for each cluster. The groups are formed to achieve a specific objective. The models consider a generalized version of the capacitated p-median problem (CpMP) where it is extracted the median from the set of items. A new formulation is introduced for the min-max diameter capacitated clustering problem (MMDCCP) and a new formulation using heterogeneous capacitated centred clustering (HCCCP). We also discuss ways to obtain the best results from the formulations using adequate solvers for each problem.

KEYWORDS. Clustering, Modeling, MP Solvers.

Paper topics: Mathematical Modelling, Combinatorial Optimization

1. Introduction

There are many ways of building clusters from items of any kind and characteristic. Humans are specialized in separating things to perform better decisions when partitions are to be the result of the decision process, although it is not possible when the number of items is high. When the items have demand/offer and limited capacity is involved while building the groups, the resulting decision problem be further more difficult because of its combinatorial counterpart. Also, in this context, there are many ways of doing clusters. To these problems, we will call Capacitated Clustering Problem (CCP) from now on.

Mulvey & Beck (1984) were the first researchers that proposed a model for the Capacitated Clustering problem. In fact, the studied clustering process was related to a Capacitated p-median Problem (CpMP). In the problem, items have attributes (or coordinates and weights). The center of the clusters are medians from a set of possible places that can absorb at maximum capacity a sub-set of the items. The capacity of the clusters may differ for each indicated possible median, Mulvey e Beck [1984].

Negreiros & Palhano (2006) proposed a new problem, considering in the Euclidean plane, a constrained process of doing min-sum clustering from a set of items. The center of each cluster is defined by the coordinates obtained by the average of the coordinates of items in the group. They called the problem as the Capacitated Centred Clustering Problem (CCCP). Different objectives were proposed, indicating variations of the decision process considering a fixed (p) or the best number of clusters. In the model, to find the best number of cluster, it is introduced a cost to open a new cluster. Also, the clusters are of the same capacity. The applications of the problem were indicated for the garbage collection, dengue disease combat, sales force territory design, newspaper subscribers delivery, vehicle routing, etc, Negreiros e Palhano [2003], Negreiros e Palhano [2006].

Most recently, Prata [2015] proposed a variation of the Capacitated p-median Problem, indicating that in the new problem each item requires a type of service and the set of p-medians have an associated capacity. The capacities may be different between types and medians. The author called the problem as the Multi-Capacitated Clustering Problem (MCCP), and reported possible applications like: assignment of students by their level to schools considering their capacity to absorb students by each level, clustering customers demands for products from refineries considering that refineries have different capacities to process each product.

This paper considers different models for the CCP, where there is only one type of service/product by item to be clustered, and the capacity of the clusters may differ. The main goal of our evaluations is to identify the characteristics of each solution obtained, for the same aspect of clustering process. In the context of the difficulty to solve the problem studied, we explain their characteristics considering the aspects related to their mathematical formulation (mixed integer-linear or mixed integer-non linear).

In this work we do not make combination of upper bounds produced by heuristic methods with the solvers resolution, or even we do not include or introduce feasible cuts to achieve better feasible bounds for the different forms of the Capacitated Clustering problem. We have done heuristic and metaheuristic approaches for the HCCCP, Batista et al. [2015], Muritiba et al. [2012], but not for all the models here evaluated. The investigation of these process can be a worth future research to be conducted in this field, mainly because the binary LP approaches returned to the investigated instances unexpected results.

In this work, section 2 consider the mathematical formulations and aspects for the CpMP, in section 3 we consider the Min-Max Diameter Capacitated Clustering Problem, a version for the capacitated clustering problems. In section 4 we discuss a mathematical formulation for the heterogeneous capacitated centred clustering problem (HCCCP). In section 5 we conclude indicating the appropriate use of each formulation.

2. The Capacitated p-Median Problem and a New Variant

Mulvey & Beck (1984) proposed the **CpMP**. It is inherently NP-HARD. The problem consider a capacitated p-median problem, where the capacity of each median can be distinct. The problem can be represented by using the following sets, parameters and variables, **?**:

Sets:

I - is the set of individuals/items (n);

J - is the set of medians (m);

$|J|$ - is the cardinality of the set J, or a fixed number of clusters ($|J| = p$);

Parameters:

p - is the number of clusters/medians;

d_{ij} - is the distance from individual i to its median j;

q_i - is the demand of an individual i;

Q_j - is the maximum capacity of median j;

Variables:

$$x_{ij} = \begin{cases} 1, \text{If the individual } i \text{ is assigned to the median } j; \\ 0, \text{otherwise} \end{cases}$$

$$y_j = \begin{cases} 1, \text{if an individual } j \text{ is assigned to be a median} \\ 0, \text{otherwise} \end{cases}$$

The CpMP can be formulated as:

$$(\textbf{CpMP}) Minimize \sum_{i \in I} \sum_{j \in J} d_{ij}\, x_{ij} \tag{1}$$

$$\text{such that} : \sum_{j \in J} x_{ij} = 1, \forall i \in I \tag{2}$$

$$x_{ij} \leq y_j, \forall i \in I, \forall j \in J \tag{3}$$

$$\sum_{i \in I} q_i\, x_{ij} \leq Q_j\, y_j, \forall j \in J \tag{4}$$

$$\sum_{j \in J} y_j = p \tag{5}$$

$$y_j \in \{0, 1\}, \forall j \in J \tag{6}$$

$$x_{ij} \in \{0, 1\},\ \forall i \in I, \forall j \in J \tag{7}$$

The objective function 1 minimizes the distance between *medians* and items assigned to each median. The constraint 2 defines one item may be assigned to only one median. The constraint 3 indicates an item may be at its median if it is used. The constraint 4 considers the items demand

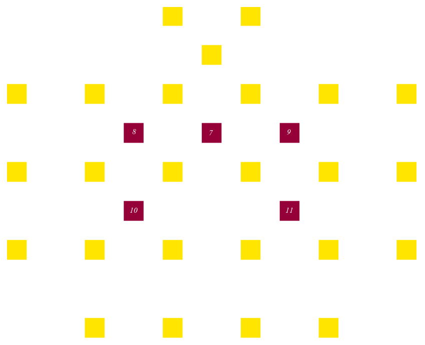


Figure 1: Capacitated p-median instance

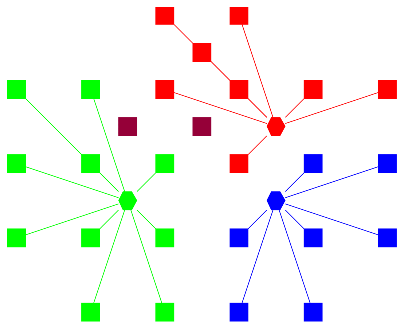


Figure 2: CpMP Solution

may not overpass the limited capacity of the median they are assigned. The constraint 5 limits the number of medians used from the set of medians. The constraints 6 and 7 refer to the decision variables of the problem.

In principle it is considered that the set of medians are disjunct from the set of the items, and observe that each median has its own maximum capacity, Mulvey e Beck [1984]. Figure 1 shows an instance of the CpMP where the set of items is in circles ($q_i = 1, \forall i \in I$) and the set of median appears in squares (maximum demand as shown in the label inside the square, Q = [9, 7, 8, 11, 10]). In Figure 2 we have the solution of the model considering the situation proposed.

If one wants to extract p-medians from all the items by using different capacities, the different values of capacities may be repeated for each item in the set of medians and it is necessary to include a new set of constraints to avoid the use of more than one median between copies of the same item. The same model may be repeated, although the inclusion of this new constraint. Figure 3 shows the situation (instance) where there are no previous defined vertices to be median, instead all of them could be one median with capacity in Q = [9, 7, 8, 11, 10]. Figure 4 shows the result of the model obtained for this instance.

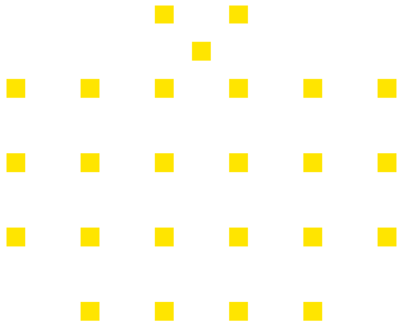


Figure 3: New capacitated p-median instance

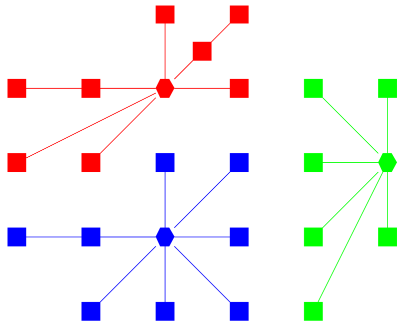


Figure 4: gCpMP Solution

If one wants to consider a direct model that solves the situation of selecting the best median from the set of items, once all the items can absorb all the capacities, a new model may be formulated, that we call generic CpMP or simply (gCpMP). In this model, the variable $x_{jj} \in \{0, 1\}$ may represent the item that will be selected as the median from the set I, $(j \in I)$, that will use the appropriate assigned capacity. It is represented by the variable $g_k \in \{0, 1\}$, when it activates the capacity Q_k. Now, it is not known *a priori* what is the assigned capacity for an item that will be choosen as median, and the model decides by itself.

For the gCpMP consider:

Sets:

K - is the set of groups;

$|K|$ - is the cardinality of the set K, or a fixed number of clusters ($|K| = p$);

Parameters:

d_{ij} - is the distance from individual i to its median j;

Q_k - is the maximum capacity of median k;

Variables:

$$g_k = \begin{cases} 1, \text{if an used cluster } k \text{ from the set of possible groups;} \\ 0, \text{otherwise} \end{cases}$$

The gCpMP can be formulated as:

$$(\mathbf{gCpMP}) Minimize \sum_{i \in I} \sum_{j \in I} d_{ij}\, x_{ij} \tag{8}$$

$$\text{such that}: \sum_{j \in I} x_{ij} = 1, \forall i \in I \tag{9}$$

$$x_{ij} \leq x_{jj}, \forall i \in I, \forall j \in I, \tag{10}$$

$$\sum_{i \in I} x_{ii} = \sum_{k \in K} g_k, \tag{11}$$

$$\sum_{i \in I} q_i\, x_{ij} \leq Q_k\, g_k, \forall j \in J \tag{12}$$

$$\sum_{j \in I} x_{jj} = p, \tag{13}$$

$$g_k \in \{0, 1\}, \forall k \in K \tag{14}$$

$$x_{ij} \in \{0, 1\},\ \forall i \in I, \forall j \in J \tag{15}$$

The objective function 8 minimizes the distance between *medians* and each item assigned to its median. The constraints 9 indicate each item may be assigned to only one median. The constraints 10 define that once a median is used an item can be assigned to it. The constraints 11 indicate that the number of medians is the same as the number of groups to be opened. The constraints 12 consider the sum of items demand assigned to a median may not overpass its limited capacity. The constraints 6 limit the number of medians used. The constraint 14 and 15 refer to the decision variables of the problem.

The gCpMP problem is NP-Hard. The model can be solved fast for instances where the capacities of the clusters are homogeneous. For the case of instances with heterogeneous capacities, our experiments did not return a feasible solution in a reasonable time by using LINGO version 13, or even Gorubi and Xpress from Argonne Labs, Neos [2015]. It is worth to observe that this model is O(n^2), in the number of binary variables and constraints, where n is the cardinality of I. This is a better formulation than CpMP formulation that is O($(n + m)^2$) in the number of variables and constraints, where m is the cardinality of J.

3. The Min-Max Diameter Capacitated Clustering Problem

In this version of the CCP, the clusters must be done considering the minimization of the maximum internal distance of a cluster between its individuals. The unconstrained version of this problem was proposed by Roi [1971], and now we add the capacity constraint. The min-max diameter capacitated clustering problem (MMDCCP) can be proposed as follows:

Sets:

I - is the set of individuals ($|I| = n$);

M - is the set of clusters ($|M| = m$);

Parameters:

p - is the number of clusters;

d_{ij} - is the distance from individual i to individual j;

q_i - is the demand of an individual i;

Q_j - is the maximum capacity of a cluster j;

Variables:

Z - is the greater diameter of a cluster between all m clusters;

$$x_{ik} = \begin{cases} 1, \text{If the individual } i \text{ is assigned to the cluster } k; \\ 0, \text{otherwise} \end{cases}$$

The MMDCCP can be formulated as:

$$(\textbf{MMDCCP}) Minimize \;\; Z \tag{16}$$

$$\text{such that}: \tag{17}$$

$$d_{ij}x_{ik} + d_{ij}x_{jk} - Z \leq d_{ij}, \;\; \forall i \in I, \forall j \in I, \; i \neq j, \; \forall k \in M \tag{18}$$

$$\sum_{k \in M} x_{ik} = 1, \;\; \forall i \in I \tag{19}$$

$$\sum_{i \in I} q_i \, x_{ik} \leq Q_k, \;\; \forall k \in M \tag{20}$$

$$Z \geq 0, \;\; x_{ik} \in \{0, 1\}, \; \forall i \in I, \forall k \in M \tag{21}$$

The objective function 16 minimizes the diameter of the greater cluster. The constraint 18 indicates an item i may be assigned to a cluster k if it does not overpass the greater diameter of the clusters Z. The constraint 19 assign one item to one cluster. The constraint 20 considers the sum of items demand assigned to a cluster may not overpass its limited capacity. The constraint 21 refers to the decision variables of the problem.

The MMDCCP is also NP-HARD. Its model is of O($n \; m$) in the number of constraints as for the number of binary variables. It is a good and fast way of doing capacitated clustering once of its mixed-integer formulation is easier than previous capacitated median problems.

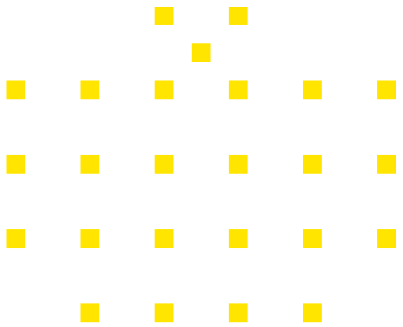


Figure 5: Min-max Diameter Capacitated Clustering instance.

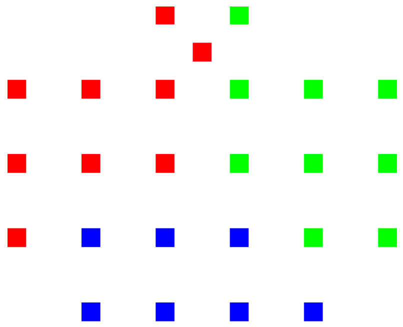


Figure 6: MMDCCP associated Solution

MMDCCP can be of enthusiastic use, if the set of items are formed by homogeneous positions in the Euclidean space ($\Re^2$) and the groups are of enough capacity to maintain individuals as close as possible, Figure 5. The groups, in this situation, may not overlap, and reasonable final geometric solutions may be obtained, as can be seen in Figure 6. In other direction, if the individuals are disperse and/or the capacity of the clusters introduce slackness, the solution may result in bad clustering, but in good mixture of groups, Figure 3. In this case, the objective may be changed to attend the size of each particular group to be formed.

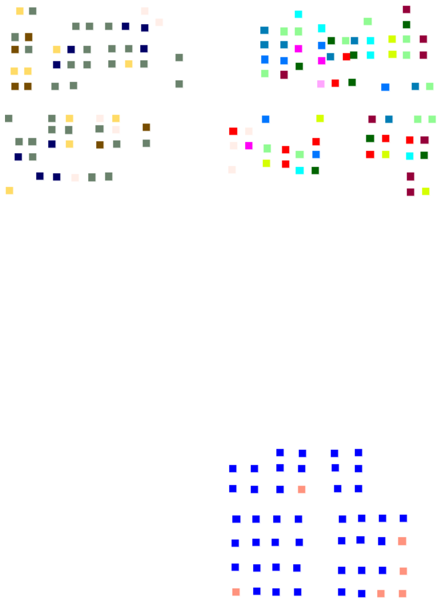


Figure 7: Min-max Diameter Capacitated Clustering instance and solution degeneration

4. The Heterogeneous Capacitated Centred Clustering Problem

In the original pCCCP model proposed by Negreiros e Palhano [2006], the capacity of the clusters are homogeneous, and the objective function consider the minimization of the total variance between items and cluster center. Here we consider the clusters of distinct capacities and the minimization of the total distance between items and assigned cluster centers. For this case the problem can be represented by using the following sets, parameters and variables:

Sets:

r - is the dimension of the space ($r = 2$, in our case);

I - is the set of individuals;

J - is the set of clusters centers;

$|J|$ - is the cardinality of the set J, or a fixed number of clusters ($|J| = p$);

Parameters:

p - is the number of clusters;

a_i - is a vector of dimension r with the coordinates of the individual i;

q_i - is the demand of an individual i;

Q_j - is the maximum capacity of a cluster;

n_j - is the number of individuals in cluster j;

Variables:

$$\bar{x}_j = \begin{cases} 1, \text{is a vector of dimension } r \text{ representing the center coordinates of the } \textit{cluster } j \\ 0, \text{otherwise} \end{cases}$$

$$y_{ij} = \begin{cases} 1, \text{if an individual } i \text{ is assigned to the cluster } j \\ 0, \text{otherwise} \end{cases}$$

$$(\textbf{pHCCCP}) Minimize \sum_{i \in I} \sum_{j \in J} ||a_i - \bar{x}_j|| \, y_{ij} \tag{22}$$

$$\text{such that}: \sum_{j \in J} y_{ij} = 1, \forall i \in I \tag{23}$$

$$\sum_{i \in I} y_{ij} \leq n_j, \forall j \in J \tag{24}$$

$$\sum_{i \in I} a_i \, y_{ij} \leq n_j \bar{x}_j, \forall j \in J \tag{25}$$

$$\sum_{i \in I} q_i \, y_{ij} \leq Q_j, \forall j \in J \tag{26}$$

$$\bar{x}_j \in \Re^r, n_j \in \aleph, \forall j \in J \tag{27}$$

$$y_{ij} \in \{0, 1\}, \; \forall i \in I, \forall j \in J \tag{28}$$

The objective function 22 minimizes the Euclidean distance between *clusters* centers ant their assigned item. The constraint 23 assign one individual to just one *cluster*. The constraint 24 consider the number of individuals per *cluster*. The constraint 25 defines the geometric center of the *clusters*. The constraint 26 limits the assigned individuals to the maximum capacity of the *cluster* j. The constraints 27 and 28 refer to the decision variables of the problem.

pHCCCP is non-linear and binary. It is NP-Hard as the previous problems, once its unconstrained version is also NP-Hard, Hansen e Jaumard [1997]. It seems that the problem is further more difficult then the other previous mentioned above. The formulation can return quality solutions of any kind, being disperse or not the set of items.

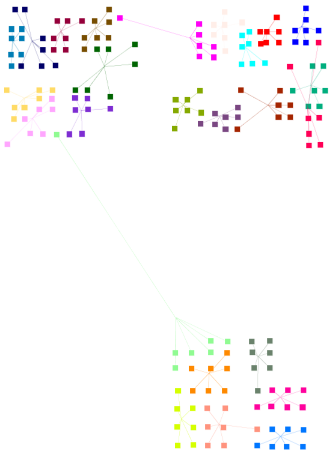
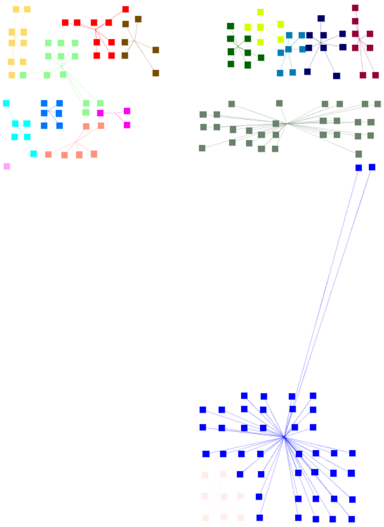


Figure 8: Solution for homogeneous cluster capacity.

This model is O(n m) in the number of binary variables and O(n) in the number of constraints. It is necessary a non-linear-{0,1} solver to run this model. Solvers like FILMINT from NEOS Server, and LINGO from LINDO Systems, can solve this model close to optimality for some instances, Neos [2015], Lindo [2014].



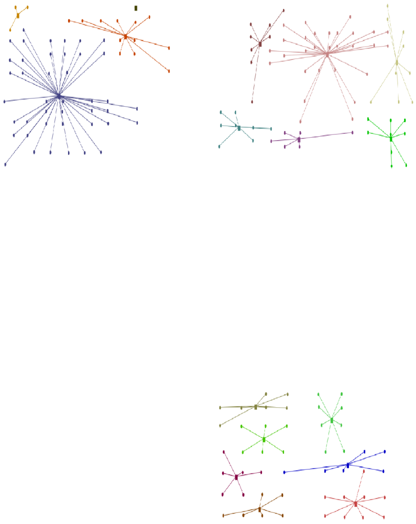


Figure 9: FILMINT solution for Hoterogeneous cluster capacity

Figure 10: Heuristic solution for Hoterogeneous cluster capacity

Although its inherent complexity, this problem is solved close to optimality by heuristic methods with very good upper bounds, Chaves e Lorena [2010], Chaves e Lorena [2011], Muritiba et al. [2012]. Batista et al. [2014] showed the results of the best known heuristics in comparison with the NEOS solvers resolution for the homogeneous version, with good results for heuristics and weak results for the solvers, that achieved at most close to the best upper bound already obtained from the benchmark instances. From the experiments, this model can also be used with FILMINT with reasonable upper bound at the end. Important to notice is that the solution produced by the heuristic method costs 30031.9218 while FILMINT indicates an "optimal cost" for the same instance with cost 35440.3, the heuristic is 15,26% better than the upper bound produced by FILMINT.

Figure 8 shows solution for the homogeneous case for the DATAPREV instance - (layout for IT team development in a software factory), and figures 9 show prints of two solutions from heterogeneous capacitated instances respectively obtained with FILMINT and figure 10 with the heuristic procedure proposed by Batista et al. [2015]

5. Conclusions

The capacitated clustering models can be adjusted to achieve different objectives. Adaptations on the basic models are needed to achieve the best way to reproduce the desired solution. Here we showed three ways of doing capacitated clustering, where the clusters capacity may differ. We introduce and evaluate solutions obtained for models related to: p-medians, in this case one wishes to extract from the set of individuals the cluster and the center of the cluster, that arises in applications like: layout teams of software development, location, etc; most compact groups, where the formulation consider only the groups and the best partition; and a new alternative formulation for the heterogeneous capacitated centred clustering problem.

The gCpMP and MMDCCP are mixed binary-linear models, which complexity is highly dependent on the sets of clusters. The first model was tested in the set of instances of IT Teams and does not achieve any solution for the heterogeneous cases but works well in all homogeneous instances. The second model is shorter and may be convenient for many reasons, for it the solvers work well in the majority of the situations also for the tested instances arose from layout IT teams in a software factory.

For the case of the pHCCCP the model is much more difficult, although the number of constraints and binary variables are less then the previous models. This model cannot be used in general, because the optimality proof obtained by the evaluated solvers is not always correct. The use of heuristics and metaheuristic is now a better way to achieve good results in reasonable time.

The main contributions of this work are two: first we showed binary LP forms of heterogeneous capacitated clustering which does not work well as a more difficult non linear binary model for capacitated clustering problem; second, we propose and evaluate new formulations for the problems also here introduced for the first time to the literature.

As future work we intend to perform evaluations to a new set of instances from sector design in garbage collection, sanitary agents coverage, wholesales, and others; which could be considered as short, medium and large scale instances from the point of view of the mathematical programming approaches.

References

Batista, P., Negreiros, M., Muritiba, A., e Palhano, A. (2015). New framework of metaheuristics for the capacitated centred clustering problem. In *Annals of XI MIC*. MIC.

Batista, P., Negreiros, M., e Palhano, A. (2014). Solvers resolution perspective to the capacitated centred clustering problem. In *Anais do XLVI SBPO*, p. 2458–2468. SOBRAPO.

Chaves, A. A. e Lorena, A. (2010). Clustering search algorithm for the capacitated centred clustering problem. *Computers and Operations Research*, 37:552–558.

Chaves, A. A. e Lorena, A. (2011). Hybrid evolutionary algorithm for the capacitated centered clustering problem. *Expert Systems with Applications*, 38:5013–5018.

Hansen, P. e Jaumard, B. (1997). Cluster analysis and mathematical programming. *Mathematical Programming*, p. 191–215.

Lindo, S. (2014). *LINGO 14.0 - Optimization Modeling Software for Linear, Nonlinear, and Integer Programming*. LINDO SYSTEMS.

Mulvey, J. e Beck, M. (1984). Solving capacitated clustering problems. *European Journal of Operational Research*, 18:339–348.

Muritiba, A., Negreiros, M., Souza, M., e Oria, H. (2012). A tabu search for the capacitated centred clustering problem. In *Annals of SBPO/CLAIO 2012*. ALIO/SOBRAPO.

Negreiros, M. e Palhano, A. (2003). Uma aplicação para o problema generalizado de percurso de veículos. In *Annals of the XXV SBPO-Natal/RN*. SOBRAPO.

Negreiros, M. e Palhano, A. (2006). The capacitated centred clustering problem. *Computers and Operations Research*, 33:1639–1663.

Neos (2015). Neos (network-enabled optimization system). `http://www.neos-server.org/neos/solvers/index.html`.

Prata, B. (2015). The multi capacitated clustering problem. Technical report, Universidade Federal do Ceará.

Roi, M. (1971). *Cluster analysis and mathematical programming*, volume 66. Journal of the American Statistical Association.