Unsupervised Clustering Method for the Capacited Vehicle Routing Problem

Alina Martínez-Oropeza[1], Marco Antonio Cruz-Chávez[2], Martín H. Cruz-Rosales[3], Pedro Moreno Bernal[2], Jesús Del Carmen Peralta-Abarca[4]

[1]Posgraduate Studies in Engineering and Applied Sciences Research Center,
[2]Engineering and Applied Sciences Research Center, [3]FC, [4]FCQeI, Autonomous University of Morelos State.
Av. Universidad 1001, Col. Chamilpa, 62209, Cuernavaca, Morelos, MÉXICO
mcruz@uaem.mx

Abstract

In this paper an unsupervised clustering method for the Capacited Vehicle Routing Problem is proposed. Advantages and disadvantages of the proposed algorithm are weighed, and comparisons are made to some clustering algorithms commonly used in the literature to tackle routing problems. Experimental tests were performed using Solomon and Hering/Homberger benchmarks applied to different distributions. The proposed algorithm is demonstrated as effective for the attempted problem, with the ability to improve upon weaknesses in some of the clustering algorithms in the literature.

Keywords: *Centroid, Customer, Euclidean Distance, Heterogeneous Population, Cluster.*

1. Introduction

At the present time, there are many optimization problems considered intractable [16], which have been tackled by different heuristic methods in an attempt to improve the best known results. According to literature, an appropriate segmentation of Routing Problems allows them to be addressed more easily by optimization methods. The most useful techniques to perform this segmentation are the unsupervised methods focused on obtaining clusters [8, 13], which allow for disjointed clusters of elements in a heterogeneous population.

In some cases, results of these algorithms are used as initial solutions for another solution method. This is one reason these methods play an important role in different research areas, such as psychology and social sciences [8], biology, statistics, pattern recognition [10], image compression, video and more recently, data mining [15]. They have been applied to combinatorial problems classified as NP-Complete [15], such as Graph Coloring [2] and the Routing Problem and its variants [1, 3], among others.

Clustering Algorithms have been applied to different Optimization Problems. In [25], three clustering algorithms based on k-medoids are proposed. These algorithms are applied to develop an algorithm selector starting from seven different heuristics, evaluating them based on a sample of 2430 instances of the Bin Packing Problem. In [26], the proposed decomposition technique called *Capacited Clustering Algorithm* is applied to the General Distribution Problem. The algorithm divides the problem by applying the classic k-means. An extra centroid is created to assign nodes into new clusters, taking into account the vehicle capacity. In [27], a k-means improved algorithm by phases is applied to the Vehicle Routing Problem with Time Windows. The first phase makes the centroids selection. The second phase distributes customers into clusters using the nearest centroid rule, and the third recalculates centroids.

All of the previous work served as precedent for the development of the proposed algorithm. The proposed algorithm is clearly different from its predecessors; these differences are explained in section four.

The VRP (Vehicle Routing Problem) and its variants, such as the Capacited Vehicle Routing Problem (CVRP) are some of the most studied optimization problems in Computational Sciences. Due to its complexity, this problem has been attempted by clustering methods, which allow the generation of clusters of customers, minimizing costs within each cluster [3]. Based on its efficacy and efficiency for certain sizes of instances, the methods most commonly used for this problem are k-means and its variants [9, 11].

The quality of results for each problem depends on the clustering algorithm applied and the problem and the instances to solve. This is due to the sensitivity of some clustering methods to changes in distribution or large amounts of data.

This paper proposes a clustering method for the Capacited Routing Problem. The proposed method incorporates strengths of some commonly used algorithms in addition to minimizing some of their weaknesses. It also incorporates constraints of the problem, obtaining an effective clustering algorithm.

This paper is divided into seven sections. Section 2 explains the clustering problem, the desirable features of clustering algorithms and the main features of the clustering algorithms most commonly used for Routing Problems. In section 3, a conceptual description of the Capacited Vehicle Routing Problem is presented. Section 4 describes the proposed clustering algorithm. Section 5 reports the experimental results that show the efficacy of the proposed algorithm. Finally, future research and conclusions are presented.

2. Clustering Problem

Generating k clusters from a heterogeneous population of N elements is considered an NP-Complete Problem [15], even when the points to be grouped are in a two-dimensional Euclidean space [6]. Heuristic methods are applied to these problems, because there is no known deterministic algorithm that solves them in polynomial time.

An analysis is performed of the characteristic information of the evaluated data in the attempt to solve the clustering problem. In the case of the Routing Problem, the distance between customers is used to assign them to each vehicle. Grouping the most similar elements together will increase the quality of the clusters.

The clustering analysis used to divide the problem could be described as a discovering tool, because it allows relationships to be found among the elements of the population [7].

2.1. Desirable Features for the Clustering Algorithms

In the literature [14], the desirable features for the clustering algorithms can be described in a general way.

- *Scalability.*
- *Randomly formed groups.*
- *Minimal input parameter requirements.*
- *Ability to deal with noise.*
- *Insensitivity to initial order.*
- *Insensitivity to the type of distribution.*

Clustering algorithms have different behavior depending on the population they assess. It is noteworthy that the features expected of an algorithm are based on the problem they are designed for, so that not all clustering algorithms are effective and efficient for all problems.

3. Capacited Vehicle Routing Problem

The Capacited Vehicle Routing Problem (CVRP) is a complex combinatorial problem, classified in the literature as NP-Complete [15].

The CVRP is a very important problem, not only for computer sciences, but for different areas. For example, because it addresses key problems of area such as distribution and logistics, it is important to industry.

Given the complexity of this type of problem, high-performance supercomputing infrastructure, such as Grid computing, is used in the search and improvement of solutions [29].

The problem is formally defined as a set of N customers with a demand d, which must be met by a vehicle k of the set K of available vehicles, so that the sum of the customers' demand assigned to a vehicle k do not exceed the maximum capacity C of the vehicle. It is noteworthy that the starting and ending point of all vehicles is the depot. Once all the customers have been assigned and all vehicles involved have finished their routes, is said that a solution for the problem has been found.

These features are represented by a Binary Integer Lineal Programming Model in [17] (Figure 1).

$$\min f = \sum_{k\in K}\sum_{(i,j)\in A} c_{ij}\, x_{ijk} \tag{1}$$

$s.a.$

$$\sum_{k\in K}\sum_{j\in\Delta^+(i)} x_{ijk} = 1 \qquad \forall i \in N \tag{2}$$

$$\sum_{j\in\Delta^+(0)} x_{0jk} = 1 \qquad \forall k \in K \tag{3}$$

$$\sum_{i\in\Delta^-(j)} x_{ijk} - \sum_{i\in\Delta^+(j)} x_{ijk} = 0 \qquad \forall k \in K \in, i \in N \tag{4}$$

$$\sum_{i\in\Delta^-(n+1)} x_{i,n+1,k} = 1 \qquad \forall k \in K \tag{5}$$

$$\sum_{i\in N} d_i \sum_{j\in\Delta^+(i)} x_{ijk} \le C \qquad \forall k \in K \tag{6}$$

$$x_{ijk} \ge 0 \qquad \forall k \in K \in, (i,j) \in A \tag{7}$$

$$x_{ijk} \in \{0,1\} \qquad \forall k \in K, (i,j) \in A \tag{8}$$

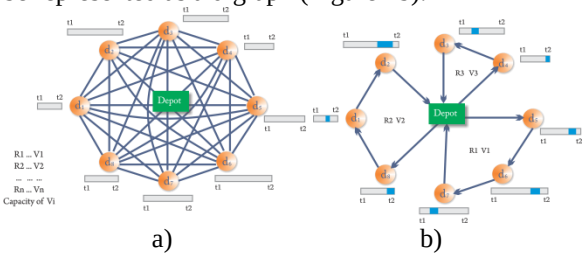
Figure 1. Binary Integer Lineal Programming Model for the CVRP [17]

Equation 1 (Figure 1) defines the objective function, which minimizes the total cost of the route, and implicitly the reduction of the required quantity of vehicles. Constraints in 2 specify that each customer must be addressed by at most one vehicle. Constraints in 3 ensure that the number of customers attended by vehicle k from the depot at the beginning

of the route is one. Constraints in 4 specify that the number of vehicles arriving to a customer is the same number of vehicles leaving that customer. Constraints in 5 guarantee that only one node connects with the depot at the end of route. Constraints in 6 ensure that the sum of the demand of all customers assigned to a vehicle *k* must not exceed the maximum capacity of vehicle. Constraints in 7 guarantee the non-negative values of variables *x*, and constraints in 8 define the lineal model as a binary integer lineal model.

3.1. Graphical Representation

A CVRP instance can be represented by a disjunctive graph forming a clique (Figure 2a). On the graph, each vertex corresponds to a customer *i* with a demand *d*, which must be satisfied by a vehicle *k*. Each vehicle *k* corresponds to a route *Rn*, where the maximum capacity of an assigned vehicle must be respected. The objective is to minimize the total route cost and the number of vehicles used. A solution can be represented as a digraph (Figure 2b).



a) b)

Figure 2. a) Disjunctive Graph, b) A possible solution for an instance of 8 customers.

In the digraph corresponding to a solution (Figure 2b), constraints specified in the mathematical model are fulfilled. In this solution three routes are defined, the order of attention to customers is indicated, and the capacity constraints of each vehicle are fulfilled and defined.

3.2. Clustering Algorithms

The CVRP has been approached using various heuristic methods due to its complexity. For instance, clustering algorithms have been used, which allow the division of the problem, creating clusters based on the nearest customer.

There are several features of clustering algorithms which are commonly applied to Routing Problems. The following discussion highlights their advantages and disadvantages.

- **Classic k-means.** This involves the Partitioning algorithm which is based on minimizing the inner distance to the cluster centroid. K-means has been the most widely used algorithm for combinatorial problems due to its easy implementation and relative efficiency for certain instance sizes [18, 19, 20 and 21]. One disadvantage is that the algorithm needs to know the number of clusters to create. Another is that all the centroids are randomly and simultaneously created as a point in space.

  This means that the centroids and distances have to be recalculated, which leads to a decrease in the efficiency of the algorithm when managing populations of thousands or millions of data. Another disadvantage is that the algorithm shows weakness in dealing with clusters of arbitrary shape or of different sizes [20].
- **PAM (Partitioning Around Medians).** This variant of k-means partitions a population into a previously established number of clusters. The advantage is that it is a more robust method than k-means, and according to the literature is very efficient for small databases. For very large populations however, this algorithm is not recommended, due to its scaling problems [21, 22, and 23].
- **ISODATA.** This algorithm uses Iterative Self-Organizing Data Analysis Techniques. Its advantages lie in its flexibility. It requires an initial approximation of the number of clusters to create, which could be adjusted automatically, enabling removal of small clusters [24]. Another advantage is that ISODATA is not sensitive to changes in data distribution. One disadvantage is that the algorithm may become inefficient when evaluating very large instances [24]. It requires tuning values such as maximum number of iterations, threshold value among clusters, and threshold values among elements within the same cluster, which requires a trial and error process in the experimentation [24].
- **Batchelor and Wilkins.** This algorithm creates clusters based on a threshold without specifying the required number of clusters. It demonstrates good efficiency when there are changes in the data distribution. However, the method is very sensitive to threshold values, so it is necessary to perform a sensitivity analysis to determine the appropriate value without compromising the efficacy of the algorithm [24].

4. Proposed Clustering Algorithm

A Clustering Algorithm is proposed to work out the CVRP (CA-CVRP). The algorithm capitalizes on the advantages of the algorithms described in point 3, while minimizing some of their disadvantages. The proposed algorithm is scalable and robust, and able to generate clusters of good quality for the CVRP.

Differences between existing algorithms in the literature and the proposed method are clear. They are

most notorious with [25], where a k-means algorithm is used. With this k-means algorithm, as well as those in [26] and [27], there is the disadvantage of needing to set the number of clusters to create. This disadvantage is improved in the proposed algorithm, because it does not require knowledge of the number of cluster to generate.

The proposed algorithm does not require recalculation of the centroids to incorporate isolated customers, unlike [25, 26, and 27]. Instead, the algorithm takes into account the constraints of the problem when performing the clustering. Because of this, feasible solutions are obtained without the application of another method such as insertion heuristics.

4.1. Development of CA-CVRP

The CA-CVRP (*Clustering Algorithm – Capacited Vehicle Routing Problem*) is an unsupervised heuristic method, which emerges from the analysis of advantages and disadvantages of several existing methods.

```
1. Initialize (N, Vh, C)
2. Read Input Data
3. for i=0 : i <N
        for j=0 : j < N
            d[i][j] = √((x2 - x1)^2 + (y2 - y1)^2)
        end-for
   end-for
4. Initialize structure Solution
        vect_tabu, index
5. Repeat
   centroid = 1+(rand()%N)
        if vect_tabu[centroid] == 1 then
           Repeat centroid = 1+(rand()%N)
           Until vect_tabu[centroid] == 0
        end-if
        else
           vect_tabu[centroid] == 1
           Search_nearest_client_at_centroid
           if (di + Σ di) ≤ C then
              if vect_tabu[i] == 1 then
                 Repeat Search_nearest_client_at_centroid
                 Until vect_tabu[i] == 0
                 if (di + Σ di) > C then v++ end-if
              end-if
           end-if
           else
              vect_tabu[i] == 1
              Solution[index] = i
           end-else
        end-else
   Until index == N-1
6. Print Clustering
```

Figure 3. CA-CVRP Algorithm.

The CA-CVRP algorithm is presented in Figure 3. At the beginning, the algorithm reads the input benchmark. Subsequently, it performs the calculation of Euclidean distances corresponding to customers and the depot, whose values are calculated using formula 9 [28] and are stored into an *N x N* matrix.

$$d_{i,j} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \qquad (9)$$

Then a centroid is randomly selected. It is worth mentioning that unlike other clustering methods [26], for this algorithm a centroid is a customer, which eliminates the need for recalculating distances and centroids in each iteration. Once the centroid is selected, the search begins for the nearest customers, evaluating the demand and the capacity constraint of vehicles. If the selected customer exceeds the maximum capacity of the current vehicle, that customer is discarded, and the search for another customer that can be assigned to the current cluster continues. On the other side, another route is generated, until all customers have been assigned to a vehicle.

The approach of the proposed algorithm avoids recalculating distances at each step, thus reducing the computational time. In addition, it is not necessary to know the number of clusters to generate, because the algorithm reduces the total number of routes, as well as the total cost of the solution.

5. Experimental Results

Experimental tests were performed using the benchmarks of Solomon, and Hering / Homberger on their three types of distribution: Clustered data (C type), Random data (R type) and Clustered-Random data (RC type) to prove the efficacy of CA-CVRP. Experimental results allow scalability and sensitivity to change and the type of distribution to be proven.

The tests were performed on a laptop with an Intel Quad-core processor i7 at 1.73 GHz, and a RAM of 6 GB., over Visual C 2008.

For the selected benchmarks, four instances were used, three of 100 customers (C101, R101, and RC101) and one of 1000 customers (RC110_1), it pertaining to the distribution RC. The literature marks RC as the most difficult to treat.

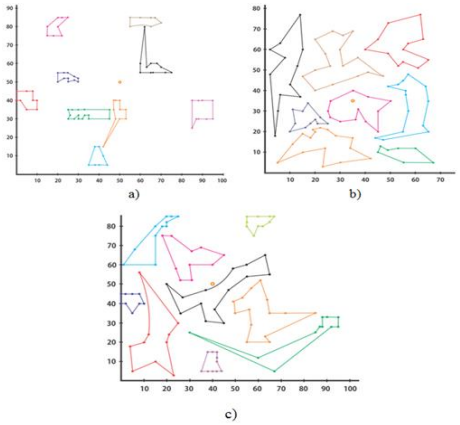


Figure 4. Example of results obtained by CA-CVRP for instances a) C101, b) R101, c) RC101

There were 30 tests conducted for each instance. The obtained results correspond to clustering (feasible solutions) generated for each instance. An example of the results obtained for each instance are shown below in Figures 4a (a, b, c) and 5.

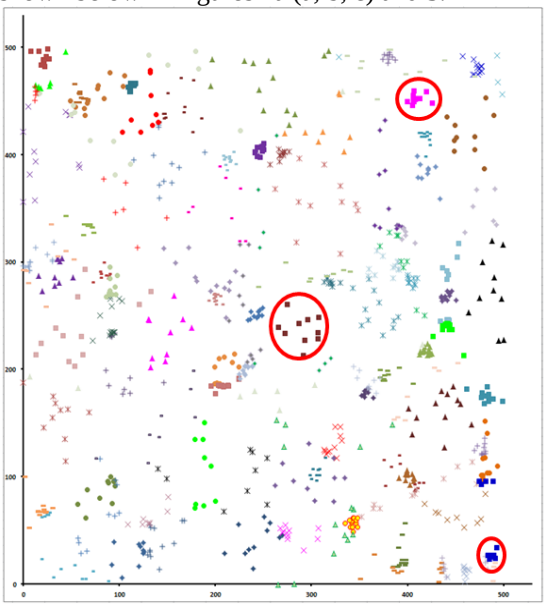


Figure. 5. Example of results obtained by CA-CVRP for the instance RC110_1

In Figure 5, three of the 94 obtained clusters are highlighted because the number of clusters is too large to indicate all of them. According to the results obtained for instances of 100 and 1000 customers, clusters of high quality can be observed, because there are no outliers (isolated customers). In addition, the time needed to generate a solution is minimal. Therefore, it is considered a fast algorithm.

Table 1 presents the average time for 30 tests per instance, as well as the best and worst clustering obtained by proposed algorithm. In [27], the fewest number of clusters obtained for the instances are presented, results which are similar to those obtained by the proposed algorithm. It is noteworthy that the results are similar because the problem used in [27] is a variant of the problem used in this work.

Table 1. Average of obtained results for each executed instance.

Benchmark	***Min_Num Clusters***	***Max_Num Clusters***	***Time (secs.)***
C101	9	9	$6 * 10^{-5}$
R101	8	8	$6 * 10^{-5}$
RC101	8	9	$7 * 10^{-5}$
RC110_1	93	94	$6 * 10^{-3}$

The obtained results show that the algorithm meets the proposed objective because it shows good development for instances of any size, and is insensitive to changes in the type of data distribution. The advantages presented by the proposed algorithm that compare to those presented in [20, 21, 22, 23, and 24] were:

- No need to know the number of clusters to generate.
- Insensitivity to changes in the type of distribution.
- Scalability.
- Reduction of the distance among customers within a cluster.
- No need to recalculate distances or centroids.
- Incorporation of the capacity constraint of CVRP.

CA-CVRP is fast, as shown in experimental tests, where the average of 30 test per instance shows that for 100 customers a solution is obtained in 7 x 10-5 seconds, while for 1000 customers it spend 6 x 10-3 seconds, yielding good quality clusterings.

It is noteworthy that the benchmarks used have different distributions, but obtain similar results in quality and speed. In this case, there is no discussion of efficiency, because is necessary to perform comparative tests with other clustering algorithms to ensure the efficiency of the proposed algorithm.

Conclusions

The algorithm CA-CVRP proposed in this paper, was developed specifically for the Capacited Vehicle Routing Problem. The algorithm is based on the approach of the nearest customer, taking customers as centroids, and evaluating specific constraints of the problem, which avoids the need to recalculate distances and centroids at each iteration.

The proposed algorithm does not need to know the number of clusters to generate. In addition, it does not require another heuristic to evaluate the problem constraints, because the algorithm obtains feasible solutions.

According to experimental tests using benchmarks of 100 and 1000 customers, using three different types of distribution, there were effective results obtained in a minimal time. In addition, the algorithm showed itself to be robust.

Future Work

A comparative analysis with other clustering algorithms applied to CVRP, such as k-means, PAM, ISODATA, and Batchelor and Wilkins algorithm will be performed. The three types of distribution (C, R, and RC), will be taken into account to prove the efficiency of the proposed algorithm.

References

[1] Ganesh K., Dhanlakshmi R., Tangavelu A., Parthiban P. Hybrid Artificial Intelligence Heuristics and Clustering Algorithm for Combinatorial Asymmetric Traveling Salesman Problem. Utilizing Information Technology Systems across Disciplines: Advancements in the Application of Computer Science. pp.1–36. ISBN13: 9781605666167, ISBN10: 1605666165, EISBN13: 9781605666174. 2009.

[2] Bozdag Doruk. Graph Coloring and Clustering Algorithms for Science and Engineering Applications. Doctoral Thesis, Ohio State University. USA, 2008.

[3] Nallusamy R., Duraiswamy K., Dhanalaksmi R. Pathiban P. Optimization of Multiple Vehicle Routing Problems Using Approximation Algorithms. Intenational Journal of Engineering Science and Technology.Vol.1(3). ISSN.0975-5462. 2009.

[4] Lenstra J. K., Rinnooy Kan A. H. G. Complexity of Vehicle Routing and Scheduling Problems. Networks. Vol. 11, Issue: 2. ISSN. 00283045. pp. 221-227. 1981.

[5] Tan, Steinbach, Kumar. Data Mining Cluster Analysis: Basic Concepts and Algorithms. Lecture Notes for Chapter 8. Introduction to Data Mining. pp. 487-488. Stanford University, 2009.

[6] González Teofilo F. On the Computational Complexity of Clustering and Related Problems. System Modeling and Optimization. Lecture Notes in Control and Information Sciences. Volume 38/1982, 174-182. 1982.

[7] Anderberg M. R. Cluster Analysis for Applications. Academic Press, New York, 1973.

[8] Fung Glenn. A Comprehensive Overview of Basic Clustering Algorithms. IEEE Citeseer. 2001.

[9] Nallusamy R., Duraiswamy K., Dhanalaksmi R. and Parthiban P. Optimization of Multiple Vehicle Routing Problems Using Approximation Algorithms. International Journal of Engineering Science and Technology. Vol 1 (3). pp. 129 -135. ISSN. 0975-5462. 2009.

[10] Kanungo Tapas, Netanyahu Nathan S., Wu Angela Y. An Efficient k-Means Clustering Algorithm: Analysis and Implementation. IEEE Transactions on Pattern Analysis on Machine Intelligence. Vol. 24, No. 7. pp. 881 – 892. 2002.

[11] Wu Chih Sheng. A Study of Heuristic Algorithms for Optimization and Clustering Problems. Master's Thesis, Department of Information Management. University Tec. de Chaoyang. Taichung, Taiwan, China Republic, 2003.

[12] Mohamed Jafar O. A., Sivakumar R. Ant-based Clustering Algorithms: A Brief Survey. International Journal of Computer Theory and Engineering, Vol. 2, No. 5. pp. 1793 – 8201. 2010.

[13] Jain A. K., Murty M. N., Flynn P. J. Data Clustering: A Review. ACM Computing Surveys. Vol. 31. No. 3. pp. 264 – 323. 1999.

[14] Hernández Valadez Edna. Algoritmo de Clustering Basado en Entropía para Descubrir Grupos en Atributos de Tipo Mixto. Master Thesis in Sciences with option in Computing. Centro de Investigación y de Estudios Avanzados del IPN. Departament of Electrical Engineering Computing Section. 2006.

[15] M. Garey, R. Jonson, D. S. and SEIT, R.: The Complexity of Flow Shop and Job Shop Scheduling, in Mathematics of Operation Research, Vol. 1, No. 2. 1976. pp. 117-129.

[16] Papadimitriou C. H., Steiglitz Kenneth. Combinatorial Optimization. Algorithms and Complexity. ISBN. 0-486-40258-4. USA. 1998

[17] Toth Paolo, Vigo Daniele. The Vehicle Routing Problem. ISBN. 0-89871-579-2. Ed. Siam, 2002.

[18] Fu Jie, Lei Lin, Zhou Guohua. A Parallel Ant Colony Optimization Algorithm with GPU-Acceleration Based on All-in-Roulette Selection. Third International Workshop on Advanced Computational Intelligence. China, 2010.

[19] Mount David M. Kmlocal: A Testbed for k-means Clustering Algorithms. University of Maryland and David Mount. Partially supported by the National Science Foundation. 2005.

[20] Karypis George, Hong Eui, Kumar Vipin. CHAMALEON: A Hierarchical Clustering Algorithm using Dynamic Modeling. IEEE Computer. Vol. 32. ISBN. 0018-9162. 1999.

[21] Stan Salvador, Chan Philip. Determining the Number of Clusters / Segments in Hierarchical Clustering / Segmentation Algorithms. 16th. IEEE International Conference on Tools with Artificial Intelligence. ISBN. 0-7695-2236-X. 2004.

[22] Huang Zhe Xue. Extensions to the k-means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery. pp. 283 – 304. Netherlands, 1998.

[23] Wu Chih-Sheng. A Study of Heuristic Algorithms for Optimization and Clustering Problems. Master's Thesis in Information Management. China, 2003.

[24] Memarsadegui Nargess, Mount David M., Netanyahu Nathan S., Moigne Jacqueline Le. A Fast Implementation of the Isodata Clustering Algorithm. Internatinal Journal of Computational Geometry & applications. pp. 71 – 103, 2007.

[25] Cruz Reyes Laura. Clasificación de Algoritmos Heurísticos para la Solución de Problemas de Bin Packing. PhD Thesis in Computing Sciences. Cenidet. June 2004.

[26] Lian L. Castelain E. A Decomposition-based Heuristic Approach to Solve General Delivery Problems. Proceedings of the World Congress on Engineering and Computer Science. Vol. II. ISBN:978-988-18210-2-7. WCECS 2009, San Francisco, USA. 2009.

[27] Díaz Parra Ocotlán, Ruiz Vanoye Jorge A., Zavala Díaz José C. Population Pre-selection Operators used for Generating a Non-random Initial Population to Solve Vehicle Routing Problem with Time Windows. Scientific Research and Essays Vol. 5 (22). pp. 3529 – 3537. ISBN. 1992-2248. Academic Journal 2010.

[28] Krajewski Lee J., Ritzman Larry P. Administración de Operaciones: Estrategia y Análisis. 5th Edition. Vol. 1. Ed. Addison-Wesley Iberoamericana. pp. 376. ISBN. 9789684444119. Mexico, 2000.

[29] Cruz Chávez M. A., Rodríguez León A., Rivera López R., Juárez Pérez F., Peralta Abarca Carmen, Martínez Oropeza Alina. Book Chapter. Chapter 3: Grid Platform Applied to the Vehicle Routing Problem with Time Windows for the Distribution of Products. Logistics Management and Optimization through Hybrid Artificial Intelligence Systems. ISBN 978-1-4666-0297-7. ISBN 978-1-4666-0298-4 (ebook). ISBN 978-1-4666-0299-1.