Using Unsupervised Learning to Determine Geospatial Clusters in Municipalities to Improve Energy Measurements

Italo F. S. Silva*, Polyana B. Costa*, Pedro H. C. Vieira*,
João D. S. Almeida*, Cláudio Baptista† Eliana Monteiro‡
*Applied Computing Group (NCA), Federal University of Maranhão (UFMA), São Luís - MA, Brazil
Email: {francyles, polyanacosta, pedrocarvalho, jdallyson}@nca.ufma.br
†Federal University of Campina Grande (UFCG), Campina Grande - PB, Brazil
Email: baptista@dsc.ufcg.edu.br
‡Companhia Energética do Maranhão (CEMAR)
Equatorial Energia, Brazil
Email: eliana.monteiro@cemar-ma.com.br

Abstract—**This paper presents a tool used to solve the Geospatial capacitated clustering problem applied to an energy company scenario. The billing process of an energy distributor in Brazil is connected to the spatially-aware logistics of collecting energy consumption data. Usually, consumer units are grouped into geospatial clusters that will be covered by meter readers. The process of creating those groups, in general, is carried out manually by analysts, which is an exhaustive process and prone to mistakes. In order to automatize this issue, this work presents a system that automatically generates reading groups for the collection of electrical energy consumption. The approach used to solve the capacitated clustering problem was based on a recursive K-Means. The results obtained with the proposed tool are promising.**

Keywords*–*Geospatial System; Capacitated Clustering Problem; Energy Companies.

I. Introduction

In regards to Brazilian energy companies, the main issues are the management of reading energy consumption and the billing process. The process of reading electrical energy consumption is comprised of two main steps: collecting consumption data from energy metering devices of every spatially distributed consumer unit and delivering the corresponding invoice. Therefore, an energy company must define a reading plan, which is a spatially-aware scheme of reading or collecting the energy consumption of consumer units. This plan changes monthly due to updates in the underlying spatial databases such as including new consumer units or shutting down some of them. In this context, to facilitate and optimize the job of meter readers, it is mandatory to gather consumer units in groups and to define a spatial criterion to create these groups. Groups of geospatial consumer units are called Reading Units (RU), and groups of reading units are called stages.

The requirements for supplying electrical energy in Brazil are regulated by the National Electric Energy Agency (ANEEL). This regulation aims to improve the relationship between utility providers and customers. Among those regulations, ANEEL establishes that the use of geoprocessing is mandatory and an electric power holding company has a certain period to finish the meter reading process [1].

Consequently, energy companies must plan how these readings will be held. The Energy Company of Maranhão (CEMAR) and Power Plants from Pará S.A (CELPA) organize the meter reading task and the delivery of invoices by creating groups of end customers, which can be understood as clusters. Hence, each meter reader must be designated to a group and follow routes to collect consumption data and deliver invoices. In practice, every municipality or region has an individual organization of reading groups and subgroups.

In order to work in this scenario, this paper presents a geographic information system focused on the creation used to optimize the creation of reading plans. This tool focuses on the creation of reading groups in order to reduce costs and optimize the job of meter readers. To achieve this, reading groups must be compact and homogeneous. The result of this work is part of a Research and Development (R&D) project, hired by CEMAR / CELPA (ANEEL PD-00371-0029 / 2016), executed by the Applied Computing Center (NCA) from the Federal University of Maranhão (UFMA).

The compactness of a group refers to its geographical shape and impacts the selection of the consumer units that will be part of each group. Elements of the same group must be close to each other, and the shape of a group should be circular, in order to fully explore a certain area. The homogeneity criterion is used to balance the reading time of the groups. This requirement ensures that the total time required to collect the consumption data from each group will be similar, therefore the workload of meter readers will be balanced. Each group or reading unit must cover the maximum working hours for meter readers, which are 6 hours per day. Therefore, the amount of reading units in a stage corresponds to the number of electrical meters readers required to cover that geographic area.

Alterations in the power network distribution, such as including new customers, or deactivating consumer units require a redefinition of the reading plan. Those changes can happen monthly and in every city in a certain geographic region. Keeping track of changes and updating reading plans is a time-consuming process, and if done manually, is also prone to mistakes. In order to produce more balanced stages and reading units, this paper presents an interactive tool for generating reading plans automatically. In addition, the proposed tool integrates georeferenced data, since each consumer unit is represented by a pair of latitude and longitude coordinates. This geodata is used to map distances between consumer units, which is an essential part in the creation of reading groups.

The remainder of this paper is organized as follows: Section

II describes the capacitated clustering problem and the applicability of the proposed tool. Section III presents recent work done to solve the capacitated clustering problem. Section IV presents the proposed system and its modules, while Section V presents the results obtained with the system and the discussion of the results.

II. Background

This section addresses the main concepts on the capacitated clustering problem and the application scenario used in this research project.

A. Capacitated Clustering Problem

According to França et. al [2], in capacitated clustering problems, a set of N elements must be subdivided into P clusters of limited capacity. Clusters are mutually exclusive, and the clustering model should maximize the homogeneity within a cluster while it maximizes the heterogeneity between clusters [3]. A generalization of this problem, called Capacitated Districting Problem (CDP), aims to group, under some criterion, an initial set of points into P districts, or to redefine an existing set of districts into P districts [4].

In our model, every geospatial point represents a consumer unit of a particular city, and the districts represent a region where a single meter reader will collect consumption data. Each point has a weight associated to it and must belong to a single cluster, while each cluster has a predefined capacity and the sum of the weights associated with them must not be greater than the capacity previously defined. The weight of each point represents the time required to perform the meter reading in the referred consumer unit. The maximum capacity of each cluster is of 6 six hours, the daily workload of a meter reader. The following requirements must be satisfied in the proposed capacitated clustering problem:

- one weight is associated with each element;
- each element must be associated with a single cluster;
- the elements must be divided into p fixed groups or clusters;
- all the elements must belong to a group;
- the sum of weights for each element of a group must not be greater than the previously defined capacity;
- a criterion to determine the proximity/distance between grouped elements is required;

In clustering problems, it is necessary to define a criterion to measure the similarity or dissimilarity between the elements. In this case, the Euclidean distance between two points was used.

B. The System Application Scenario

This section presents important information about the system application scenario, including characteristics and some requirements for the reading plans creation process.

Some criteria should be considered during the creation of a reading plan. One is the geographic shape of a reading group, because it directly impacts the route traversed by the meter readers. These reading groups should also be homogeneous in relation to the meter's work charge in order to minimize operational costs.

Another criterion to be considered is to follow the main rules for electric energy supply in Brazil. They were defined by Brazilian Electricity Regulatory Agency (ANEEL) in the Resolution 414/2010 in order to improve the relationship between power companies and costumers. According to the rules, a power company must perform a read of a consumer unit at 30-day intervals, but it might happen between 27 or 33-day intervals. Moreover, the bill must be delivered to the costumer 5 working days before the bill due date. In the case of first reading of a consumer unit, or changes on the reading calendar, ANEEL also defines, for these cases, intervals of 15 days minimum and a maximum of 47 days. If a company does not follow these rules, it is liable to pay fines.

The creation of reading plans following these requirements should be performed for all cities served by the CEMAR/CELPA power companies every month because of urban transformations and the expansion of their services. However, doing it manually is a slow process, and the delay might cause financial losses. Therefore, a system that creates optimized reading plans automatically is important because it tends to accelerate that process while satisfies those requirements.

III. Related Work

Several works address the capacitated clustering or redistricting problem; some of them are applied to power meter reading, others to the definition of salesman working zones or garbage collecting, etc. This section presents some of those systems and the techniques used to solve the capacitated clustering problem.

A method to group consumer units from an energy company was proposed by Costa et. al [5], with the aim to reduce the execution time of requested services and to properly distribute tasks among groups. Their approach is based on a capacitated P-medians and a genetic algorithm, which produced better results in comparison with the manual grouping performed by the energy company. However, when comparing the results from both approaches, the genetic algorithm produced better solutions to the problem.

Metaheuristics were also used to propose solutions to the capacitated clustering problem applied to power meter reading. De Assis et. al [4] use a greedy randomized adaptive search procedure (GRASP) and multicriteria scalarization techniques to create clusters. Experiments taken on a portion of the city of São Paulo showed the effectiveness of their method.

Capacitated centered clustering was also applied to garbage collecting and definition of salesman working zones, as shown in [6]. The authors present a hybrid data mining heuristic to solve the capacitated centered clustering problem based on a heuristic that combines Clustering Search and Simulated Annealing. The heuristic and the clustering search were used to find the best solutions in the search space, while data mining was used to search for data patterns and improve the searching for newer and better solutions.

In order to collect household water usage data, Smiderle et. al [7] based their approach on operational research techniques in order to find the shortest route between a set of points, leading to a decrease of the time spent by meter readers to collect water consumption data. Their method applied a combination of genetic algorithm and Teitz and Bart algorithm to the P-medians problem, reducing $7,200$ meters in a route that covers a group of 774 houses.

This work presents a system that automatically generates reading plans for collecting electrical energy consumption. The approach used to solve the capacitated clustering problem was based on a recursive K-Means algorithm and a post-processing step was used to improve its results. The reading plan should comply with several restrictions that will be explained in the following sections.

IV. System for Planning of Reading Units

The Consumer Units Reading Planning System uses unsupervised learning to assist the logistic planning creating reading groups, which are organized in Stages (effective reading days) and Reading Units (subdivisions of stages which indicate the necessary amount of power meter readers to perform the reading task).

The system consists of two modules: Manual and Automatic Reading Planning. The first one is responsible for the implementation of the clustering strategy. The second one allows to create or edit reading plans interactively. Figure 1 shows an overview of the system's components.

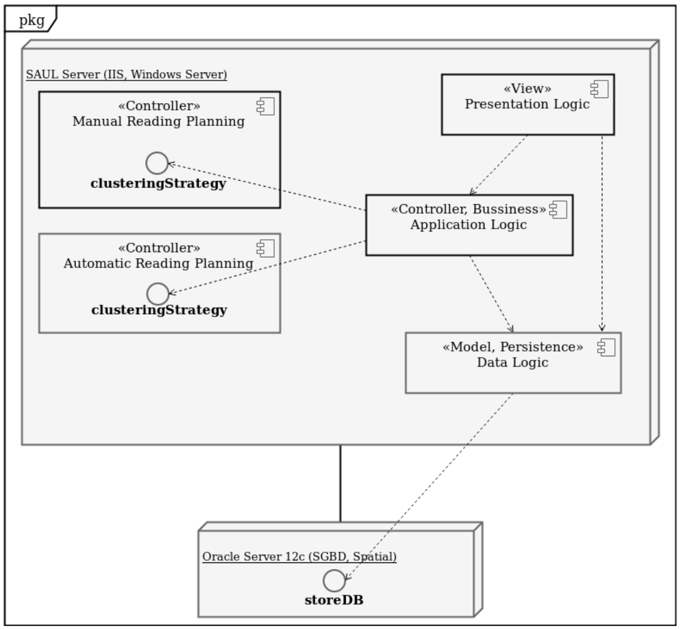


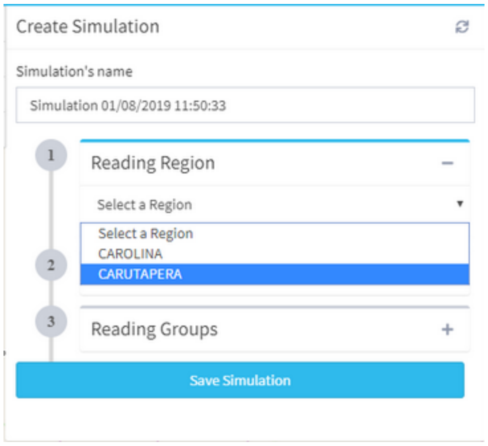
Figure 1. Component diagram of the system.

The main differences between these two approaches are related to internal system performance. In the case of manual approach, the system assists the creation and edition tasks performed by user step by step. On the other hand, in the automatic approach, the user just defines the input parameters for the algorithms.

A. The Manual Reading Plan Module

The Manual Reading Planning Module allows the creation of reading plannings interactively. This module uses the metaphor of web maps for consumer units data visualization. The use of maps favors a better comprehension of the location of consumer units and the route that meter readers go through.

In this approach, the definition of stages and reading units is performed manually. The user selects the reading region and the consumer units to be handled on the map. Then, the user must select the option of adding stages to a geographic region or RUs to a stage defined previously. Hence, the manual creation of reading plans is entirely controlled by the user. The application allows to change the order of stages and to swap reading units. Figures 2, 3 and 4 show the steps of manual user interaction to create the reading groups.


Create Simulation

Simulation's name: Simulation 01/08/2019 11:50:33

1 Reading Region

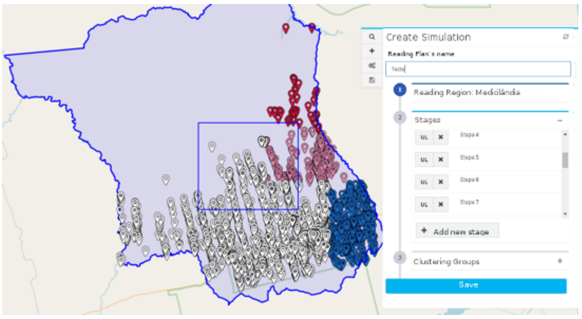
Select a Region

- Select a Region
- CAROLINA
- CARUTAPERA

2

3 Reading Groups

Figure 2. Create simulation interface.


Create Simulation

Reading Plan's name: Teste

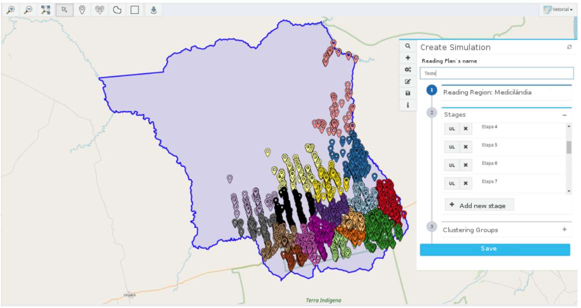
1 Reading Region: Medicilândia

2 Stages

UL	✖	Etapa 4
UL	✖	Etapa 5
UL	✖	Etapa 6
UL	✖	Etapa 7

3 Clustering Groups

Figure 3. Stages and Reading Units Creation interface.



Create Simulation

Reading Plan's name: Teste

1 Reading Region: Medicilândia

2 Stages

UL	✖	Etapa 4
UL	✖	Etapa 5
UL	✖	Etapa 6
UL	✖	Etapa 7

3 Clustering Groups

Figure 4. Generated groups in manual planning.

The reading groups contain all the consumer units and the white groups show undefined units. All the consumer units must be in a unique group to enable the calculation of decision-making reports used by the energy company.

Besides the manual interactive possibility, the system enables the automatic creation of those groups and it is possible to propose different approaches in the automatic clustering process.

B. The Automatic Reading Plan Module

In order to minimize costs related to the elaboration of the reading plans, the system proposed in this work has the Automatic Reading Planning Module that generates clusters automatically. This clustering task follows constraints defined by the company in the scenario of the monthly reading planning of the consumer units.

The grouping of consumer units must consider that they should be spatially close, minimizing displacement and maximizing the number of consumer units read by the power meter readers in their working day. According to the described scenario, it is possible to see this required task may be categorized as a capacitated clustering problem in which generated clusters must not exceed a predefined capacity and also respect other constraints. The next section explains the method used in order to solve the clustering problem.

C. Towards Solving the Capacitated Clustering Problem

This section introduces a method that solves the capacitated clustering problem applied in the power companies scenario. Figure 5 shows the five steps of the proposed method.

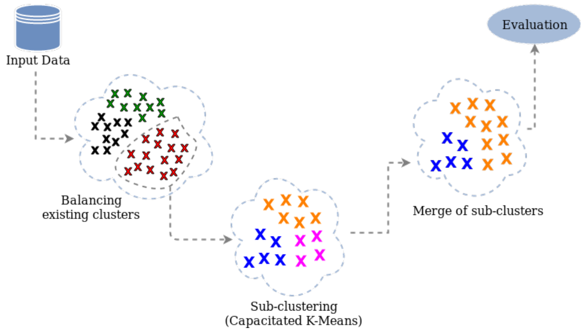

Input Data
Evaluation
Balancing existing clusters
Sub-clustering (Capacitated K-Means)
Merge of sub-clusters

Figure 5. Steps of the proposed method.

Firstly, Latitude and Longitude coordinates, old clustering information and the time of measuring of each consumer unit are collected from the dataset. These data are used as input for the Balancing step and are grouped in order to balance the total time of measuring each group. It is emphasized that the method presented in this work is applied in municipalities that contains a previous reading plan in order to improve it. If a region does not contain reading plan information, consequently there are no measurement time data associated to the consumer units. Thus, in these cases, the automatic creation should consider another set of constraints to be included in this capacitated clustering problem modeling.

After the Balancing step, each balanced group is used as input for the clustering algorithm. In this step, those groups are subdivided into smaller groups. However, some of them contain a small number of points. Hence, these generated clusters are submitted to the merging step, where little clusters are merged based on a proximity criterion. The last step is the evaluation of the generated groups by applying clustering evaluation metrics.

1) Organizing Stages: The dataset contains the reading plans created manually by the company. Geographic coordinates, the reading time of each consumer unit and clusters which they belong to are extracted from the dataset. As described above, these clusters are called Stages. Each stage has a number corresponding to the day when its consumer units will be read by the meter reader.

The balance of clusters step consists of grouping consumer units starting from the previous reading plans in order to balance the sum of the reading times of each Stage, also improving their geographical distribution. Consequently, it also balances the work load of the meter readers. The K-Means algorithm is used in this step.

The K-Means initial cluster centers are the centroids calculated from the previous Stages. The similarity criterion used was the Euclidean distance between the latitude and longitude coordinates.

Another criterion to be considered during the clustering procedure is the minimization of the reading time standard deviation. For this, the average reading time of the new clusters is calculated. These groups are submitted to the clustering procedure until standard deviation reaches the minimum value. Finished the balance of the existing clusters, or Stages, the next step consists of the creation of reading units.

2) Creation of Reading Units using Capacitated K-Means: The Capacitated K-Means is based on the K-Means technique, which is generally used to perform the clustering task. In this one, the data of a set are split into groups according to a similarity criterion. The capacitated version of K-Means, which is used in this work, also includes a capacity constraint for group generation. Figure 6 presents an overview of the clustering method.

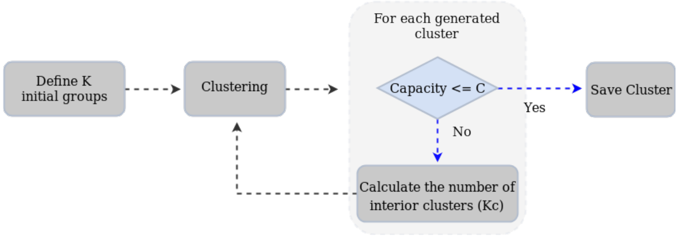

For each generated cluster
Define K initial groups
Clustering
Capacity <= C
Yes
Save Cluster
No
Calculate the number of interior clusters (Kc)

Figure 6. Workflow of Capacitated K-Means algorithm.

The initial number of clusters (K) is defined as 2. Thus, a big cluster is divided into two big parts and it allows these ones to be sub-divided into more parts according to the capacitated clustering strategy. An initial clustering is performed based on the number K. The capacity of each created group is evaluated and, if it exceeds the maximum capacity added to a threshold, the points of a group will be separated for a recursive clustering process. The new generated groups will also be evaluated. The new Kc values are based on the total of points and on the established capacity as seen in 1. The steps of re-clustering, evaluation and group split are repeated recursively until all the generated groups satisfy the constraints. It is important to note that clusters whose capacity satisfies the constraint are preserved.

$$Kc = \frac{N}{C_m} \tag{1}$$

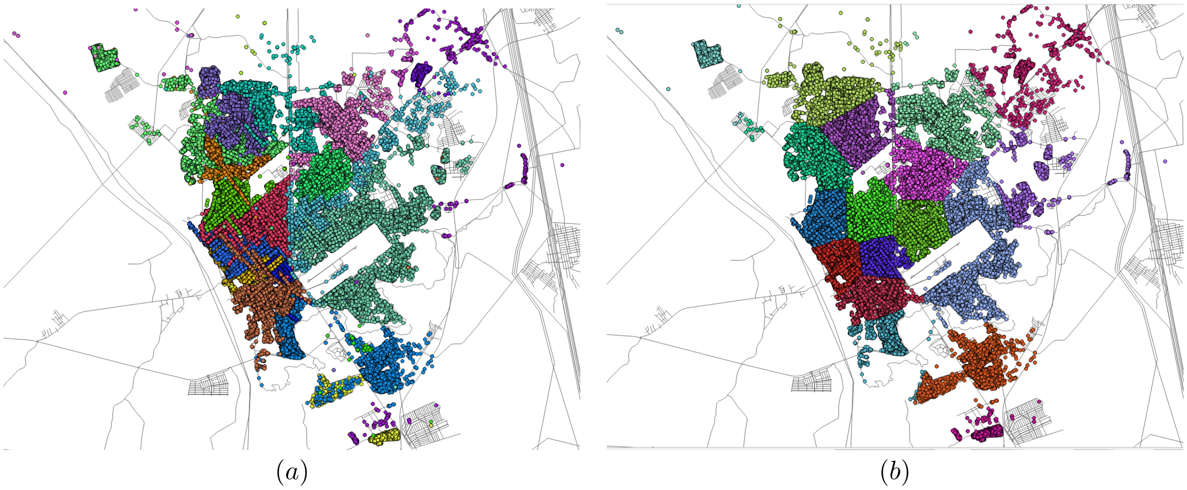


Figure 7. Comparison between results of (a) Current Reading Planning and (b) Automatic Reading Planning.

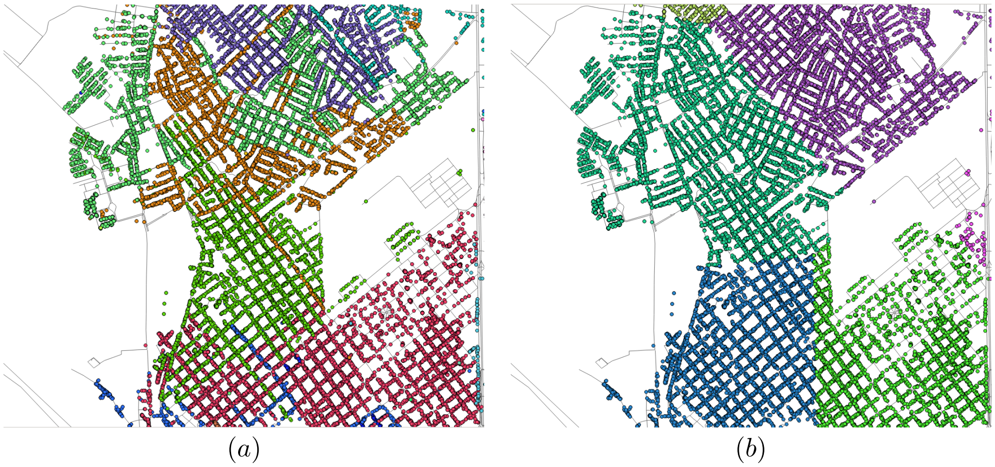


Figure 8. A closer view comparing Stages in (A) Current Reading Planning and (B) Automatic Reading Planning.

In 1, C_m value is the ratio between the defined capacity and the average reading time of consumer units of a sub-group.

There are two guarantees provided by this algorithm: (1) all points belong to a single cluster, and (2) all groups satisfy the desired capacity. To avoid maintaining groups with small capacities, a merge of adjacent clusters is performed in order to ensure the creation of more homogeneous groups.

An under capacity cluster can be merged with other clusters until they reach the maximum capacity, which is of 6 hours per cluster. Besides the capacity, a small cluster must be merged with a close cluster, otherwise the compactness of the group will decrease. In order to merge clusters based on their capacity and proximity, a graph that connects them is created. Each cluster on the graph will be a vertex represented by the cluster's centroid; the edges represent the connection between two centroids, and the weights of the edges are the Euclidean distance between the points. The graph was built based on a Delaunay Triangulation algorithm [8]. After building the graph, a Breadth-First searching algorithm (BFS) [9] was used to search the graph in order to merge adjacent vertices. This post processing step results in less clusters, with more elements in each cluster.

V. Results and Discussion

To analyze the obtained results, the generated stages and reading units must conform to the application's scenario presented in Section II-B. Additionally, the efficiency of the automatic generation of reading plans was evaluated in regards to the homogeneity of the groups, their geographical shape and if they comply with ANEEL's regulation.

Tests were performed based on Imperatriz data, a mildly populated municipality in Maranhão, Brazil. Figure 7 shows the comparison between the manually defined stages and the stages generated by the proposed tool. The image shows that the new groups have become more compact and homogeneous than the manually defined ones. Table I confirms this result and presents the comparison for the mean area of the clusters and the standard deviation of the reading time.

Table I also shows the average Silhouette Coefficient for the entire clustering. The Silhouette Coefficient (SC) is a measure that evaluates a cluster in terms of cohesion and separation [10]. Cohesion quantifies how close the objects within a cluster are, and it expresses the compactness of a group, while separation determines how isolated a cluster is from other clusters. The SC has a range of values that vary from $[-1, 1]$. If the coefficient has a negative value, it means that the clustering is sparse. The closer to 1 the coefficient is, the more compact a cluster is. In a good clustering, all groups should have a positive silhouette coefficient.

TABLE I. RESULTS OF THE CALCULATED METRICS.

	Current State	A. R. Plan
Std. Times	2872.60	2991.09
Average Area (Km^2)	62.38	29.15
Silhouette Coefficient	-0.38	0.26

Manually defined groups tend to be line-shaped, the proposed system, on the other hand, produces circular groups. This happens because of the intricacies of K-means, the algorithm groups the consumer units that are closer to the centroid of each stage. This can be seen in Figure 8, which shows a closer look at a specific region of Imperatriz.

In regards to reading groups, the results of the clustering performed by the capacitated K-Means are similar to ones that the energy company already has, as shown in Figure 9. However, the proposed system assures that the reading groups will have balanced workloads due to the merging clusters step.

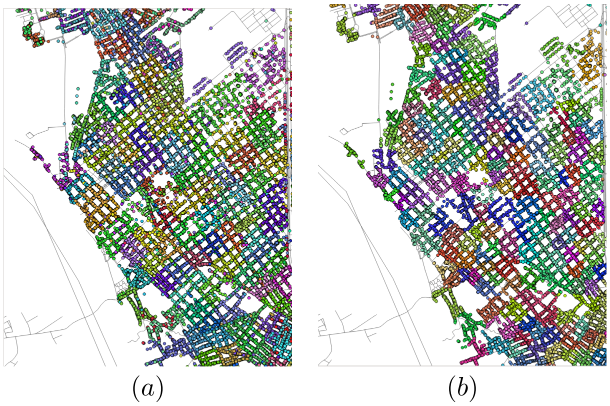


Figure 9. A closer view comparing Reading Units in (A) Current Reading Planning and (B) Automatic Reading Planning.

Finally, the resulting reading plan is evaluated in relation to compliance with the regulation defined by ANEEL. Figure 10 shows a graph that relates the number of consumer units of the reading plan and the number of days necessary to collect its consumption data. According to the figure, the proposed tool was capable of grouping consumer units in a way that all of them get the invoice within the period of time specified by ANEEL. In comparison with the manually defined reading plan, more consumer units will get the invoices in a period of 30 days, and less of them in a period of 32 days.

VI. Conclusion

This paper presented a geospatial tool for generating automatic reading plans applied to the Brazilian energy companies CEMAR and CELPA. Machine Learning and optimization techniques were used to produce reading groups in an unsupervised way. The generated reading plan should balance the workload of meter readers, completely explore the same region and comply with ANEEL's regulations. The results achieved with the proposed tool were promising, more consumer units were covered within the period of 30 days, the generated reading groups were more compact and homogenous, at the same time, their configuration did not contrast from the manually defined groups. The workload of the groups was balanced and the reading plan complied with the constraints These promising results bring a perception that the proposed method can be applied in domains with analogous constraints, e.g., garbage collection or measurement of water consumption.

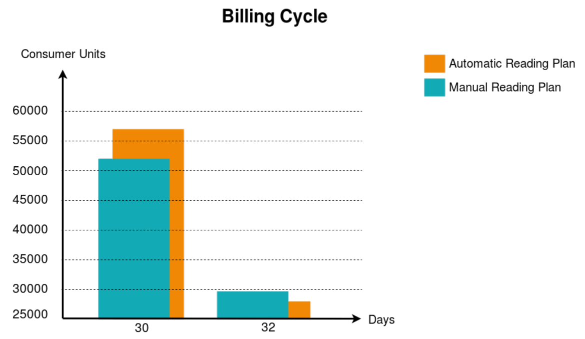


Figure 10. Graph relating the number of consumer units and the number of days necessary to collect their consumption data.

For future work, the presented tool should be tested on more cities, especially more populated ones. Along with balancing the workload within reading groups, it is also desirable to balance the workload within the stages of the reading plan.

Acknowledgements

The authors would like to thank UFMA, IFMA, FAPEMA, CEMAR/CELPA for making this work possible through the project PD-00371-0031/2017 and CNPq for financial support. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

[1] ANEEL. Resolução normativa nº 414. [Online]. Available: http://www.aneel.gov.br/documents/656877/14486448/bren2010414.pdf [retrieved: jan, 2019]

[2] P. M. França, N. M. Sosa, and V. Pureza, "An adaptive tabu search algorithm for the capacitated clustering problem," International Transactions in Operational Research, vol. 6, no. 6, 1999, pp. 665–678.

[3] J. M. Mulvey and M. P. Beck, "Solving capacitated clustering problems," European Journal of Operational Research, vol. 18, no. 3, 1984, pp. 339 – 348.

[4] L. S. De Assis, P. M. Franca, and F. L. Usberti, "A redistricting problem applied to meter reading in power distribution networks," Computers & Operations Research, vol. 41, 2014, pp. 65–75.

[5] C. Costa, D. Costa, and A. Góes, "Determinação de setores de atendimento em uma concessionária de energia," Trends in Applied and Computational Mathematics, vol. 8, no. 3, 2007, pp. 381–390.

[6] M. Guerine, M. B. Stockinger, I. Rosseti, and A. Plastino, "Heurística híbrida com mineração de dados para o problema de agrupamento capacitado com centro geométrico," XLIX Simpósio Brasileiro de Pesquisa Operacional, 2017.

[7] A. Smiderle, M. T. A. Steiner, and C. Carnieri, "Problema de cobertura de arcos – um estudo de caso," XXIII Encontro Nacional de Engenharia de Produção, 2003.

[8] D.-T. Lee and B. J. Schachter, "Two algorithms for constructing a delaunay triangulation," International Journal of Computer & Information Sciences, vol. 9, no. 3, 1980, pp. 219–242.

[9] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, Introduction to algorithms. MIT press, 2009.

[10] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Boston, MA, USA: Pearson Education India, 2007.