



25th ABCM International Congress of Mechanical Engineering
October 20-25, 2019, Uberlândia, MG, Brazil

COB-2019-2397

Capacitated Implementation of Growing Neural Gas Applied to the Electric Distribution Companies Measuring Logistics in Brazil

Artur Albuquerque Silva, , Italo Francyles Santos da Silva, Pedro Henrique Carvalho Viera, Robert Douglas de Araújo Santos, Anselmo Cardoso de Paiva, Geraldo Braz Júnior

Applied Computer Group NCA-UFMA, Federal University of Maranhão - UFMA, Av. dos Portugueses, 1966 - Vila Bacanga, São Luís - MA, 65080-805

astobiro@gmail.com, andsf@live.com, francyles@nca.ufma.br, pettersson.diniz@gmail.com, pedeziq2@hotmail.com, robert@nca.ufma.br, paiva@nca.ufma.br, geraldo@nca.ufma.br

Eliana Márcia Garros Monteiro, Equatorial Energia

eliana.monteiro@cemar-ma.com.br

Abstract. *The main tasks related to the billing process of an electrical energy distribution company are the measuring of client energy meters and invoice delivery. The energy distribution companies have the job of organising the measuring process by creating groups of consumers. Thus, each employee must have routes to perform the measuring and delivering invoices in each group. To solve this clustering problem, the development of computational methods is important, so that the process can be improved. Therefore, this work proposes Capacitated Growing Neural Gas (C-GNG), a modified Growing Neural Gas algorithm using capacitated clustering concepts. The proposed algorithm is applied to the electric distribution companies context in Brazil, aiming to obtain a better clustering than the current one, that was manually generated by company experts. The obtained results, demonstrates improvements in terms of cohesion, separation and silhouette using the proposed C-GNG when compared with the manually crafted solution..*

Keywords: *Capacitated Clustering, Growing Neural Gas, Districting Problem*

1. INTRODUCTION

The main tasks related to the billing process of an electrical energy distribution company are the reading of client energy meters, to measure the client energy consumption and invoice delivery. In Brazil, the regulation of distribution conditions is overseen by National Agency of Electric Energy (ANEEL). This regulation tries to improve the relations between consumers and energy providers.

The electric energy distribution companies have the task of organising the measuring process in consumer groups. Each employee must have routes to perform the measuring and delivering of invoices in each group.

The creation of these groups is named the districting problem. The districting problem consists in grouping small territorial spaces in non superposed adjacent districts, given an objective function and restrictions.

Growing Neural Gas (GNG), proposed by Fritzke (1997), is an algorithm with the objective to generate a graph structure that reflects the topology of input data. The same way a gas fills an area, GNG fills in the input data space, generating a topologically similar structure. It's applications are in clustering, vector quantization and object tracking.

This paper proposes a modification on the original GNG algorithm, applying capacitated concepts, creating the Capacitated Growing Neural Gas (C-GNG). A method to create capacitated clusters applied to the electric energy distribution companies measuring logistics in Brazil. The results and metrics are compared with the manual clustering performed by the companies. GNG was selected because it has the property of reflecting the topology of data, that is why we aim that it maintains the topological characteristics of electrical residential units.

2. METHODS

This section will describe the proposed method to perform the clustering. Figure 1 shows the sequence of steps proposed.

The points used for clustering are the electric energy measuring meters of residential clients of electric distribution companies. Each house has a device that measures the electric energy consumed. Each group of devices that will be visited in a day are called steps.

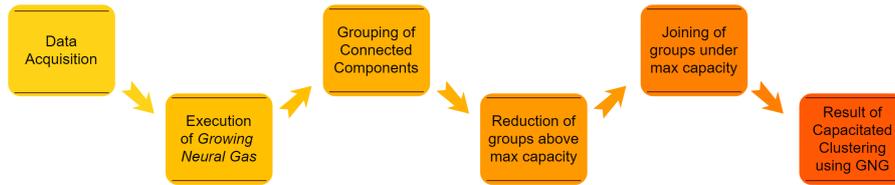


Figure 1. Steps of the proposed methodology

2.1 Data Acquisition

The database was supplied by two electric energy distribution companies in Brazil. It contains the information of the cities of Imperatriz and Açailândia in Maranhão. The database collected has for each residential unit the geographical localization (latitude, longitude), the step and the time. Latitude and longitude are geographical positions of the energy meters, steps are groups of meters that will be visited in a day, and time is the time that the employees take reading a meter in minutes. A sample of the database can be seen in Table 1.

Table 1. Sample of the database

LAT	LONG	STEP	TIME
-5.4927806	-47.4836327	10	1.13
-5.5106509	-47.4862713	8	0.55
-5.5182095	-47.4355126	18	0.58
-5.5188013	-47.4514449	16	0.59

2.2 Execution of Growing Neural Gas

The next step is the execution of the GNG algorithm, but first we have to normalise latitude and longitude so we can use Euclidean distance between residential units. Considering the radius of the earth as 6371 km we apply the following transformations:

$$X = R * \cos(\text{latitude}) * \cos(\text{longitude}) \quad (1)$$

$$Y = R * \cos(\text{latitude}) * \sin(\text{longitude}) \quad (2)$$

Due to the high number of residential units, in order to reduce processing time, a node limit was used as a stopping criterion. We used max capacity as the number of elements, divided by the total number of steps plus a tolerance of 10%.

The parameters used were: $\epsilon_b = 0.1$, $\epsilon_n = 0.006$, $a_{max} = 5$, $\lambda = 20$, $\alpha = 0.5$, $\beta = 0.995$, Maximum number of nodes = 3000, Maximum capacity = time divided by steps * 1.1%.

Figure 2 is an example execution of GNG in a sample of 30 elements (in red). After the algorithm was finished a graph with 26 nodes was generated (in blue).

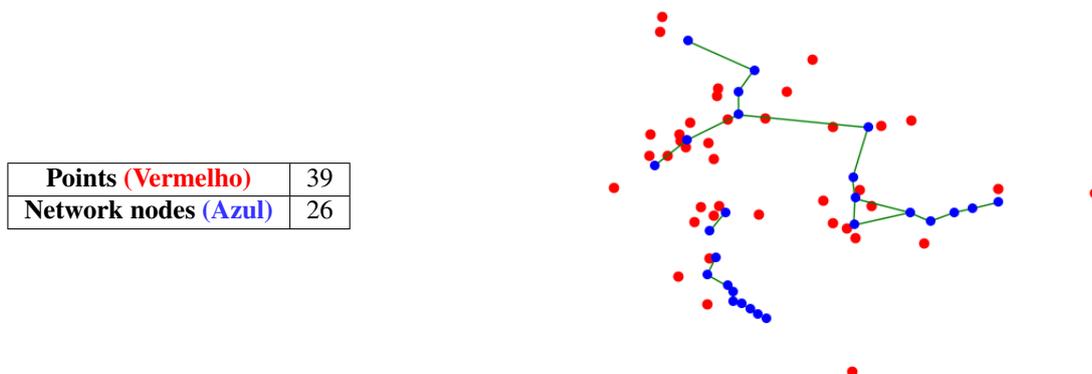


Figure 2. Example execution of GNG (Red: Data points, Blue: Network nodes)

2.3 Grouping of connected components

After GNG is executed we assign a value starting from 0 to each connected component of the output network. Connected nodes are assigned to the same group and each point is assigned to a group. We do that by calculating the distance

between each node and each point, assigning the closest points to the closest node's group. Figure 3 shows an example result of connected component grouping, we use the data shown in Figure 2. After the grouping, Group 0 (in red) has 30 elements, Group 1 (in cyan) has 5 and Group 2 has 4 elements.

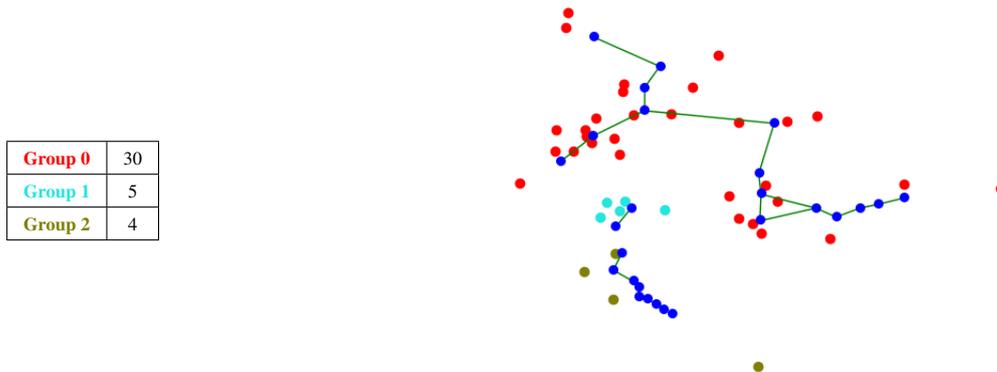


Figure 3. Connected Component Grouping

2.4 Reduction of groups above max capacity

After the initial clustering it is necessary to check each cluster's capacity. Because GNG doesn't assign each point to groups using capacity, there is no guarantee that the groups are under or equal to maximum capacity. So we execute two steps, reducing the groups that are over maximum capacity and then joining groups that are under maximum capacity.

Group reduction algorithm works in the following way, for each group that is over maximum capacity, we assign each point to the closest node, so each node represents a group consisting of each closest point. After that we execute the Breadth-First Search (BFS) algorithm as described in Cormen *et al.* (2009), to iterate in the network.

To exemplify group reduction we select Group 0 from Figure 3. Assuming a max capacity of 10 and that each point has capacity of 1, Group 0 then has a total capacity of 30. For each node of the network we check the closest points and group them. After this process what remains is more than 7 groups, as seen in Figure 4.

Group a	4
Group b	6
Group c	2
Group d	2
Group e	2
Group f	2
Remaining	12

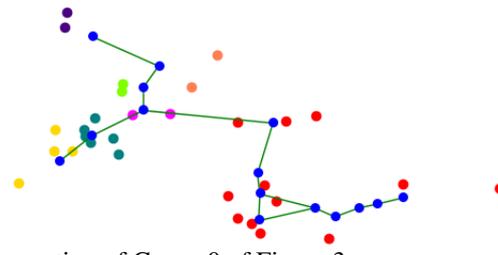


Figure 4. Node by node separation of Group 0 of Figure 3

The next step is selecting the first node of the group that is over capacity. So we select the node that contains group *a* (in Figure 4 in yellow) and execute BFS starting from it. The output of the algorithm is: $a \rightarrow b \rightarrow c \rightarrow d \rightarrow e \rightarrow f \rightarrow \text{Restante}$. Group *a* has capacity equal to 4, we check if its capacity plus the next group's capacity, group *b*'s, exceeds max capacity ($4 + 6$). As this does not occur we add *b*'s elements to a new Group 0 and then check its capacity plus group *c*'s capacity ($10 + 2$). As the result exceeds max capacity, we increment the group counter and add element *c* to the current group, Group 1 and like that we go through each element in the graph. Resulting in groups *d*, *e* *e* *f* joining Group 1 and the remaining 12 elements form Groups 2 and 3, as Figure 5 shows.

Group 0	10
Group 1	8
Group 2	10
Group 3	2

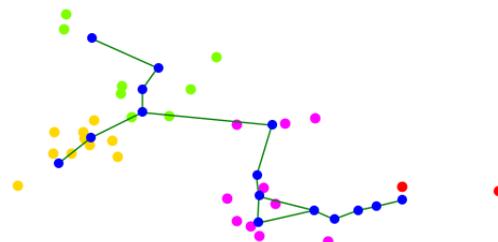


Figure 5. Result of node by node separation of Group 0 of Figure 3

2.5 Joining groups under max capacity

After reducing the groups we can guarantee that no group is over the maximum capacity, but we can't guarantee that the groups are balanced or close to the maximum capacity. So we execute a group joining algorithm proposed in Costa *et al.* (2018).

Besides capacity, it is necessary that a group is joined with groups next to it. So we create a graph that connects them. To define each edge we use Delaunay Triangulation proposed in Delaunay *et al.* (1934). Creating a complete graph would require high computational power and wouldn't be practical because of the time needed to calculate the weights of each node, that is euclidean distance. It is also worth noting that connecting each node would result in distant groups joining. So a non complete graph is sufficient. In Chew (1986) it is discussed that it is possible to generate a graph as good as a complete graph using Delaunay Triangulation.

The method consists in using a Delaunay Triangulation on each cluster's centroid to create a network. Check the mean distance of each vertex and cut those that are over, then, for each cluster under capacity, traverse the network using BFS starting on that cluster. If the current cluster's (Cluster A) capacity plus the next cluster's capacity (Cluster B) is under the maximum capacity, join both of them and proceed to the next one, marking Cluster B as visited. If the sum is over the maximum capacity, get the elements closest to Cluster A from Cluster B that fill it till to reach the maximum capacity.

Continuing the example from the previous section Figure 6 shows the result of Delaunay Triangulation using the same data from Figure 5. Black points represent the centroids and the blue edges represent the result graph's connections.

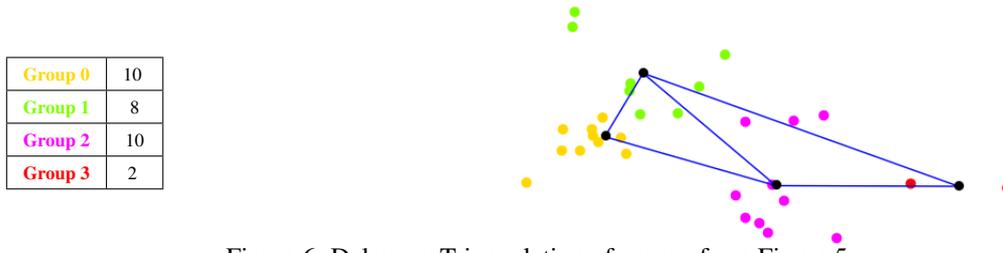


Figure 6. Delaunay Triangulation of groups from Figure 5

The next step after the triangulation is to check the distance of the edges and calculate the mean distance. The edge between groups 1 and 2 has a distance bigger than the mean distance, so it is removed, as depicted in Figure 7.

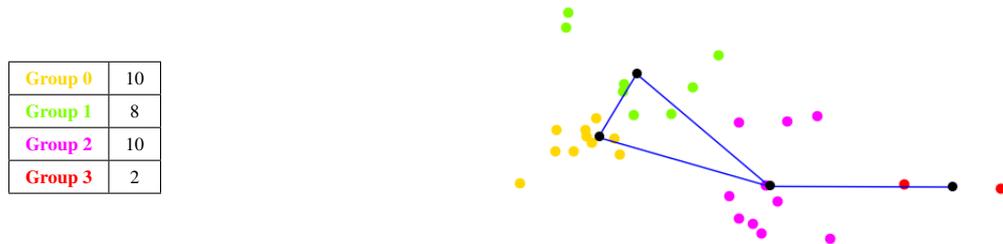


Figure 7. Delaunay Triangulation of groups from Figure 5 after edge removal

Then we check the capacity of Group 0 and that it has capacity equal to max capacity. We mark it as visited and then check Group 1, it has capacity equal to 8, lower than max capacity, we execute BFS starting from its centroid obtaining the following output: $1 \rightarrow 0 \rightarrow 2 \rightarrow 3$.

We check the node after the initial one, Group 1's centroid, and see that it has been visited. Then we proceed to the next node, Group 2's centroid. Because it hasn't been visited yet we check the sum of its capacity plus Group 1's capacity ($8 + 10$), it is larger than max capacity. We then select the first point closest to Group 1's centroid, check if its capacity plus the group's capacity does not exceed max capacity ($8 + 1$). Then we select the next closest point and check its capacity plus the group's capacity ($9 + 1$). We do the same for the next group and check that it now exceeds the max capacity, so we change Group 1's status to visited and check Group 2 the same way.

Because the process removed 2 elements from Group 2 its capacity is now 8, we check that its capacity plus the next Group's capacity doesn't exceed the max ($8 + 2$). We then add Group 3 to Group 2. Seeing that there are no more nodes to visit, the group joining is completed. Figure 2.5 shows the final result.

2.6 Validation metrics

To validate the method, metrics that calculate the efficiency of groups in relation to the current grouping were selected. What we desire from capacitated clustering, besides the clusters being under or equal to max capacity, is balancing. Using standard deviation we can calculate if the groups are balanced or not, the bigger the deviation, less balanced, and the

Group 0	10
Group 1	10
Group 2	10

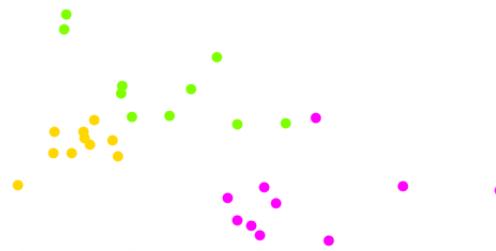


Figure 8. Final groups after join

smaller, more balanced.

The metrics used to validate the method were mean, standard deviation, cohesion, separation and silhouette coefficient. Proposed in Tan *et al.* (2007), cohesion calculates how close each element in a cluster are to each other, separation calculates how distant a cluster is to every other cluster. The silhouette coefficient combines both cohesion and separation and outputs a number between -1 and 1 the closest to 1 the more compact a cluster is.

$$Cohesion(C_i) = \sum_{x \in C_i} euclidean_distance(x, c_i) \quad (3)$$

$$Separation(C_i, C_j) = euclidean_distance(c_i, c_j) \quad (4)$$

$$SC(C_i) = \frac{b_x - a_x}{max(a_x, b_x)} \quad (5)$$

Where a_x is the result of the mean distance between a random element x and the other elements in a cluster C_i . b_x is the mean distance of the element x to each point of the cluster closest to C_i that does not contain x . The function $max(a, b)$ returns the maximum value between a and b .

3. RESULTS AND DISCUSSION

To validate the method tests were performed in two situations, The city of Imperatriz and the city of Açailândia. The program QGIS from QGIS Development Team (2019) was used to plot the data and obtain geospatial information.

For Imperatriz the maximum capacity is of 8135.07, 182297 measuring meters and a total of 18 steps numbered 2-18 and 31. Table 2 shows mean time, standard deviation, shortest time and longest time for current data that is used today and the result of the C-GNG approach. We can see in Table 2 that the cluster with longest time did not exceed maximum capacity. But the cluster with shortest time is 12 times smaller than the maximum capacity. We can also see that the resulting standard deviation is smaller than the current one. Which shows that the resulting clustering of C-GNG is better than the current clustering.

Table 2. Results in Imperatriz

	CURRENT	C-GNG	DIFFERENCE
Mean	7395.52	7395.52	0
Standard Deviation	2563.62	2030.86	532.75
Shortest time (min)	449.9	645.21	195.3
Longest time (max)	10869.88	8135.06	2734.82

Table 3 shows mean Cohesion, Separation and SC values for Imperatriz, we can see again that the proposed method is better than the current one. The resulting cohesion is almost two times bigger, the separation is almost two times smaller and the SC is closer to 1 than the current method.

Table 3. Mean Cohesion, Separation and SC for Imperatriz

	CURRENT	C-GNG	DIFFERENCE
Cohesion	371995.56	662704.73	290709.17
Separation	51.46	30.24	21.21
SC	-0.501419	-0.006277	0.495141

Figure 9 illustrates steps defined manually and the clusters generated using C-GNG. We can observe that clusters generated using C-GNG do not possess a circular shape, unlike conventional clustering methods such as K-Means. The clusters maintain this characteristic despite the fact that the method used to join groups uses cluster centroids.

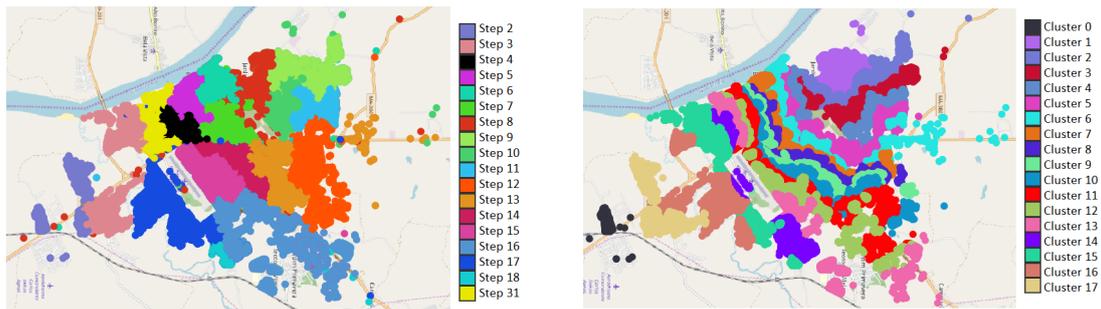


Figure 9. Current clustering in Imperatriz and C-GNG results in Imperatriz

The interesting thing in this result is that the clusters obey a linear shape that could correspond to streets. This characteristic could be interesting when organising access logistics based on the streets that is desired.

In Table 4 data of each group is detailed next to the current group data, that were made manually by the electric distribution companies. It shows the number of elements, the time and the total capacity of each group.

The resulting Group 0 is the one that has the least number of elements and the least total time, but we can see that it has time and elements bigger than step 18, which is the step with the least number of elements and least time of the current data. Group 3 is the one that has the highest time and highest number of elements. It has less elements than the step 16, but it has capacity less than the max capacity.

Table 4. Specific data Imperatriz - MA

CURRENT			RESULT		
Step	Time (min)	Quantity	Groups	Time (min)	Quantity
2	1863.25	3568	0	645.21	1847
3	7963.49	11961	1	8134.73	11938
4	9127.98	14989	2	8135.03	11535
5	6993.61	8750	3	8135.06	12243
6	7546.64	10243	4	8134.99	11625
7	8556.27	11325	5	8135.04	9765
8	7179.23	11392	6	8135.04	10682
9	6951.05	9260	7	8134.68	10644
10	7501.9	10315	8	8134.92	11065
11	6752.19	10060	9	8134.04	10347
12	7521.65	8063	10	8134.85	10125
13	9205.71	10726	11	8134.87	10968
14	7943.22	11980	12	8134.76	10571
15	9319.26	10647	13	8134.72	11428
16	10869.88	13137	14	8134.76	11617
17	10212.42	12243	15	8134.54	12194
18	449.91	1023	16	7088.6	9376
31	7161.8	12615	17	3363.62	4327

To explain why Group 0 is the group with the least time we selected an area of Figure 9 that illustrates an enlarged view of this group.

As we can analyse from Figure 10, Group 0 is the most distant from the others and, possibly, because of that it was the last group to be checked on the BFS in the joining step. So its elements were removed and added to the nearest group, because it was the last group and all the others were marked as visited. Another possibility is that the edges that connected Group 0 to the others had distances bigger than the mean distance and were excluded, putting Group 0 out of the graph and without removing near elements to complete it.

For the clustering in Açailândia the maximum capacity value is 1092.03, a total of 34760 measuring meters and 17 steps, numbered 2-18. Table 5 shows the mean and standard deviation as well as shortest and longest time. We see that the mean found using C-GNG is bigger than the current method (the clusters are better grouped). In Table 6 we can see that the separation is bigger than the current one. This means that the clusters found are more distant to each other.

In Figure 11 we can see the current steps made manually and the result of C-GNG in Açailândia. The first thing we see is that the clusters do not have that wave shape, as seen on Fig. 11, these clusters are intertwined with each other. We can also see that used method found less clusters, which in practice means one less day of work, but these clusters do not



Figure 10. Section of Figure 9 showing Group 0, Group 17 and Group 16

Table 5. Results in Açailândia

	CURRENT	C-GNG	DIFFERENCE
Mean	992.76	1054.80	62.04
Standard Deviation	174.82	117.69	57.13
Shortest Time (min)	677.3	631.14	46.16
Longest Time (max)	1439.56	1092.03	347.53

Table 6. Mean Cohesion, Separation and SC for Açailândia

	CURRENT	C-GNG	DIFFERENCE
Cohesion	54652.95	69799.38	15146.43
Separation	435.81	474.71	38.89
SC	-0.340341	-0.054340	0.286000

have a large compactness.

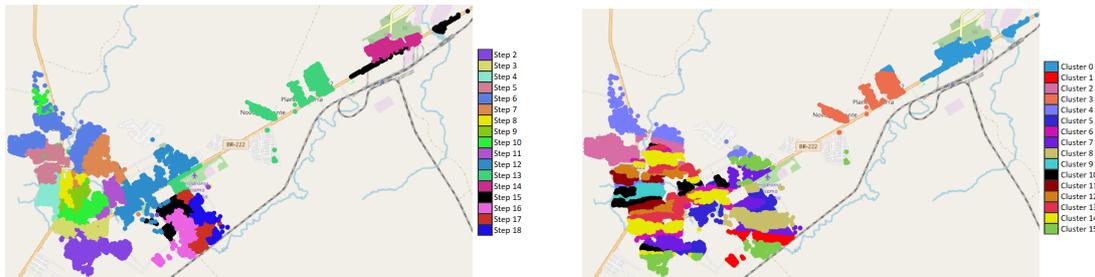


Figure 11. Current clustering in Açailândia C-GNG results in Açailândia

Table 7 shows detailed data for each cluster. In it we can see that Group 15 is the one with the least number of elements, and the least time. Table 7 also shows that C-GNG has the capacity to generate less number of clusters (which would equal one less day of work in the field). But these groups do not have a high compactness.



Figure 12. Section of Figure 11 showing a closer view of the groups

Table 7. Specific data Açailândia - MA

CURRENT			RESULT		
Step	Time (min)	Quantity	Groups	Time (min)	Quantity
2	892.08	2119	0	1091.86	2437
3	964.65	1733	1	1091.94	1972
4	897.66	1944	2	1091.95	1614
5	1097.55	2084	3	959.66	1692
6	1439.56	2472	4	1091.89	1990
7	1076.22	2197	5	1091.99	2800
8	1040.76	2105	6	1091.86	2041
9	767.51	1975	7	1091.96	2666
10	1100.96	2371	8	1091.74	2328
11	677.3	1697	9	1092.02	2789
12	1095.25	1932	10	1092.03	2275
13	1162.61	2191	11	1091.69	2373
14	858.37	1859	12	1091.88	2213
15	900.8	1790	13	1091.72	2332
16	970.81	1726	14	1091.62	2255
17	1066.39	2287	15	631.14	983

4. CONCLUSION

This work presented a method for electrical installation clustering to optimise the management of data collection and billing for electricity distribution companies. The work is based on an adaptation of the Growing Neural Gas algorithm to solve the problem modelled as a capacitated cluster in graphs. The propose method achieve a clustering result more efficient than the the defined manually by the company analysts. The obtained result improved the separation and cohesion indexes, and also improved the silhouette coefficient, providing an optimised version with balanced load and requiring less days for the reading of all units. For future work , we intend to modify the cluster union strategy to achieve more homogeneous clusters, and also to test the algorithm using rural unities.

5. REFERENCES

- Chew, P., 1986. "There is a planar graph almost as good as the complete graph". In *Proceedings of the Second Annual Symposium on Computational Geometry*. ACM, New York, NY, USA, SCG '86, pp. 169–177. ISBN 0-89791-194-6. doi:10.1145/10515.10534. URL <http://doi.acm.org/10.1145/10515.10534>.
- Cormen, T.H., Leiserson, C.E., Rivest, R.L. and Stein, C., 2009. *Introduction to algorithms*. MIT press.
- Costa, P., da Silva, I.S., Vieira, P.C., Santos, R.A., Silva, M., Lima, C.C., Júnior, D.G., Garros, E., da Silva, I.S. and Pinheiro, L.A., 2018. "Rcmeans: A recursive capacitated means for districting problem". *Journal on Advances in Theoretical and Applied Informatics*, Vol. 4, No. 1, pp. 28–35. ISSN 2447-5033. doi:10.26729/jadi.v4i1.2745. URL <http://revista.univem.edu.br/jadi/article/view/2745>.
- Delaunay, B. *et al.*, 1934. "Sur la sphere vide". *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk*, Vol. 7, No. 793-800, pp. 1–2.
- Fritzke, B., 1997. "Unsupervised ontogenetic networks; handbook of neural computation".
- QGIS Development Team, 2019. "Qgis geographic information system". URL <https://qgis.org/en/site/>.
- Tan, P.N. *et al.*, 2007. *Introduction to data mining*. Pearson Education India.

6. ACKNOWLEDGEMENTS AND RESPONSIBILITY NOTICE

The authors thank to ANEEL and Equatorial Energia for the support of this work. The authors are the only responsible for the printed material included in this paper.