Residential Electric Energy Consumption in Brazil: An Hierarchical Time Series Clustering Approach

Ana Paula Amazonas Soares¹ and Maria Eduarda Silva²

¹ Universidade Federal Rural de Pernambuco, Recife 52171-900, Brazil
   ana.soares@ufrpe.br

² Universidade do Porto, Porto, Portugal
   mesilva@fep.up.pt

Abstract. Energy consumption plays an important role in day to day basis and it must be seen as a matter of public policy. Increases in residential electric energy consumption represents that residential energy consumption thought time represents growing economic activity and also may represent a consequence of increasing in live standards. Therefore, the study of its trajectory is a matter of Government’s issue. The time series clustering methodology has no prerogative, meaning that no economic principle was used to determine how to form clusters, nor no other economic variable was used to justify homogeneity within groups. Applying the agglomerative hierarchical time series clustering methodology made possible to determine each clusters’ tendency over time. The variable used to characterized series was Residential Electrical Energy Consumption in Brazil, by distributors’ monthly average consumption, from 2003 to 2017. At the results, distributors’ location was applied to verify geographic patterns that may have had economic implications. The results show that the applied methodology identified homogeneous locally groups that had responded to economic factors throughout time. It represents a breach to use such a methodology to enlighten hidden explanations without direct implement traditional econometric model.

Keywords: Time Series Clustering, Energy Consumption, Hierarchical Clustering Time Series.

1 Introduction

Energy has become the most important resource in daily life. Its generalized use ranges from cooking, to workplace, to telecommunications. Nowadays, the use of energy is far-reaching, advancing in ways thought unimaginable years ago – for example, energy-powered cars and buses.

As energy advances in our daily life, so does its production. Currently, it is possible to extract and store electrical energy from the sun, the wind and even sea waves. These new forms of production can be delivered to any home, or even be produced by the homeowner.
The sector of production and distribution of energy has gained importance and will always be considered a strategic sector to Government Policy. Therefore, the study of its trajectory is important, especially in areas where its scarcity is considered a matter of public issue.

Brazil is no different than any other underdeveloped countries: It has a National Policy regarding energy resource; there is a regulation on delivering energy to consumers; and, the Government was the first to investing in the sector. The actual energetic matrix is based upon hydropower plants, 75% according to the regulatory Agência Nacional de Energia Elétrica – ANEEL – reported in 2008, but solar and wind power plants have been increasing in the last few years, and, the Government can guarantee its supply. Although, in 2001, it had occurred a national shortage of supply, due to lack of rain, nonetheless, there are a variety of alternative plants that are being used nowadays to prevent such discontinuity.

Residential energy in Brazil is the second most used, also according to ANEEL (2008). It represents the use of gadgets, appliances and comfort. For short, all wellbeing that residences could use in a house. In other words, residential consumption represents all possible goods that the residence is using.

The paper will focus on the Brazilian Residential Electrical Energy Consumption – REEC – for the period from January 2003 to December 2017, supplied by sixty-two distributors spread throughout the country.

The objective of this paper is to verify increasing consumption patterns in REEC users throughout time, applying the Time Series Cluster – TSC – methodology. The study will aggregate similar patterns over time, in which, it will be possible to identify each cluster’s trajectory.

Since distributors are linked to a local area, it will be possible to plot them on a map and identify if those clusters are geographically related. If members of a cluster are in a continuous area, it will be possible to direct and implement public policy on such areas to guarantee needed consumption.

Also, since all time series will be represented in clusters, it will be possible to find a represented series for each cluster, in which its trajectory over time will represent a tendency.

The study concentrates on areas where consumption is increasing over time because it means that average householders are willing to buy more goods for their comfort. Stable consumption indicates that residents have reach their peak, they are not consuming more goods, while decreasing consumption suggests that residents are reviewing their consumption priorities and deciding which goods to use. Areas where consumption is increasing show that economic activity is increasing as well because consumption is linked to the household budget. Therefore, these are the areas that must be studied.

Moreover, the study will present some insights regarding income, since REEC has a positive income elasticity of demand, meaning that increase in income leads to relatively larger increase in demand, which indicates a causality between income and REEC. Subsequently, these areas with increasing REEC could be areas with increasing income.
2 Methodology

Aggregate similar items is a habit that started a long time ago, for instance – separate food in order to prevent degradation, distinguish which animal could be together with other, organize groups of people – to cite but a few. Scientifically, clustering is the idea of put similar things together or, aggregating objects that have a common criterion. Clustering is also a procedure used to find hidden combinations, a procedure which is currently being used in data mining – considering the emergence of big data. As Nadif and Govaert (2010) pointed out, “...cluster analysis can identify several populations within a heterogeneous initial population, thereby facilitating a subsequent statistical study.”

The process of clustering could be divided into supervised and unsupervised. The supervised process has a starting point (ground truth or prior knowledge) and seeks to create a classification (function) for future observations, in other words, create a vector of best responses, given the prior knowledge established. The unsupervised process, which is the one applied in this paper, has the goal to group classified objects by its similarity where no pre-concepts are established.

Liao (2005) pointed out that “The goal of clustering is to identify structure in an unlabeled data set by objectively organizing data into homogeneous groups where within-group-object similarity is minimized and between-group-object dissimilarity is maximized”. This definition considers: (i) a criteria to objectively organize the data, for instance, time; (ii) a measure of similarity, so that objects can be placed together, for example, the number of years lapsed between each object; and (iii) a technique to place them together, or how to combine them closely to each other, starting from a distributor and seeing how close one is to the other, or separating them all and trying them in pairs.

In the case of this research paper, the criteria was a time series of average REEC – measured in Megawatt-hour (MWh) – of a household for each distributor in Brazil. Which means, that for each of sixty-two different providers, in different areas, the total consumption is divided by the total number of consumers, for each month, from January 2003 to December 2017.

The measure of similarity is the distance between the vector points, in this case defined by the difference of averaged MWh consumed by each residence over time. However, to measure it, a matrix of distance for each combination of points is calculated. For example, considerer two distributers: X and Y, each point i (January 2003, February 2003, …) over time T (all observations), two vectors are formed, and a distance measure is implemented. The Euclidian Distance is the simplest, the distance between those two observations (X and Y) is: \( d(X, Y) = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + \ldots + (X_T - Y_T)^2} \). There are other measures, such as Minkowsky or Manhattan.

In addition to measuring the distance, it is important to define how objects will form groups. It will be necessary to define a starting point, all objects belonging to a large group to then be divided – divisive – or each object belonging to a single group which will then be paired – agglomerative –. The method used in this paper will con-
sider an agglomerative method. Meaning that each distributor is a single group and, then, elements will be paired according to their distance.

Also, the technique, by which they will be paired, is the one described by Liao (2005), that uses the Ward (1963) method of linkage. Elements will be paired and grouped considering the minimum similarity among members and maximum dissimilarity between groups.

Ward’s method considers a starting point where there are no groups and then, similar objects are paired considering the change in variance among groups and within groups which are being formed. Initially, each element is considered a group and for each step, the algorithm will consider a joint between objects and a variance within a group and outside the group is calculated and a decision is made. The process continues until one big group is formed. Therefore, Ward’s technique is to create an agglomerative group.

Even though Ward’s technique is based upon an agglomerative technique, it is important to cite that there are two techniques: Hierarchical and Partition. The differences are that the first takes into account that there is no certainty on how many groups can be formed and no object is referred to a specific group, and the second one considers that there are several groups exists and for each group an element is referred to.

For now, the study will use a method that considers: (i) a process to separate and find similarities among members, considering a constant characteristic over time; (ii) that a distance is estimated, a matrix of similarity and dissimilarity is calculated; (iii) a technique to select similar objects is Ward’s; and (iv) an agglomerative hierarchical process is implemented.

However, the approach described is known as classical clustering method, which considers the characteristic as constant over time, but since the data is a time series and not a cross-section, the REEC characteristic does change over time. In other words, it will be necessary to adjust the distance measure in order to account for changes in characteristics over time. Time Series Clustering – TSC – differs from the classical approach because it introduces a new measure of distance, which considers changes in the characteristic along the time.

2.1 Time Series Distance Measures

TSC has been, recently, used in many fields, such as motion capture (Li and Prakash, 2011), stock market (Guan and Jiang, 2007), medicine (Wismüller et al., 2002) and speaker verification (Tran and Wagner, 2002), to cite but a few. Focardi (2001) introduced the use of TSC in Economics, it argued that “... concepts of similarity, such as dynamic time-warping, used in a wide range of application domains might be useful in the context of finance and economics.”

However, before considering defining the distance measure, it is important to determine if modifications are needed to be made to the data or if it is necessary to fraction data. In other words, it will be wise to establish requirement for: data transformation; modeling and features extracting; and, forecasting.
Aghabozorgi, Shirkhorshidi and Wah (2015) divided TSC in three approaches: (i) whole time series clustering, which means applying conventional clustering on time series objects with specific distance; (ii) subsequence clustering, clustering of segments from a single long time series; and (iii) time point clustering, clustering of time points based on a combination of their temporal proximity of time points and the similarity of the correspondent values.

Liao (2005) divided TSC into three different approaches, which are: (i) raw based, which is a modification of the static approach (classical approach); (ii) feature extraction, where it transforms a raw time series data into a feature vector of lower dimension and then apply conventional clustering; (iii) model-based approach, modify the vector to a lower dimensional and then extract model parameters.

Ferreira and Zhao (2016) classified TSC into two approaches: (i) data adaption, like model-based, where features from of each time series are extracted and then a clustering algorithm is applied; (ii) algorithm-adaptation, uses a desire clustering algorithm to direct extract TSC. It is important to highlight that in the algorithm-adaptation, the author indicates that “… the major modification rests in the distance function, which should be capable of distinguishing time series” (p. 228).

At this point, the paper will consider the whole time series, raw based data – some data transformation, such as normal standard transformation – to obtain an agglomerative hierarchical methodology with Ward’s linkage technique and the final decision rests at the distance measure.

There are many distance measures, as cited by Łuczak (2016): Dynamic Time Warp – DTW; Longest Common Sub-Sequence Distance – LCSS; Minimal Variance Matching Distance – MVM; and the Euclidian Distance – ED. In addition to those, Paparrizos and Gravano (2017) proposed a new shape-based distance – SBD – which achieves a similar accuracy to DTW but is mostly used in partitioning methods.

The DTW is the typical distance function used for TSC (Li and Prakash, 2011). According to Izakian, Pedrycz and Jamal (2015) the objective of DTW is: to determine an optimal match between two time series in the calculation of their differences (p. 236). Saas, Guitardt and Periñan (2016) use DTW as a minimum distance between two time series. The distance is expressed by: $DTW(X, Y) = \min_{r \in M} \sum_{m=1}^{M} |x_{im} - y_{rm}|$ where the path element $r=(i,j)$ represents the relationship between the two series (Saas et al., 2016). To sum it up, DTW seeks for the best alignment between values of both time series, considering boundary conditions, continuity and monotonicity.

Regarding this paper, TSC methodology will consider an unsupervised process which considers whole time series clustering, the data will be transform into a lower dimension, the distance used will be the DTW and the linkage used is a Ward agglomerative hierarchical clustering.

Finally, the representation of this clustering is a dendrogram, presented in a bi-dimensional diagram. The objects will be placed in the horizontal axe and the distance is measured in the vertical axe. The lower the distance, the higher is the number of clusters. So, the possibility of expanding or lessen the number of clusters will depend upon the distance. A dendrogram is an easy-to-understand form to present the results of a hierarchical clustering method.
Furthermore, the cluster’s size is not specified so it is possible to have a cluster with only one member. Though, having many small clusters is not optimum, since the methodology is to aggregate similar objects; consequently, there is no point on having many clusters of single objects. Studying a cluster of one member is studying the member itself; therefore, clusters with one member will not be analyzed in this the study.

2.2 Data

This paper will consider the REEC in Brazil, obtained in ANEEL (2018), by month and year, distributor and class of consumption. Two sets of data were obtained for the period from 01/2003 to 12/2017, as cited before, for sixty-two electricity distributors, and residential class of consumption: consumption – measured in Megawatt-hour – and numbers of consumers (householders). Those distributors contribute for 99.6% of consumption and consumers, so they are a representative sample for REEC in Brazil. Both sets of data were used to obtain the monthly average consumption by residence to each energy supplier, attaining 62 time series for the range period and average REEC.

It is worth mentioning that each supplier is linked to an area and a state of Brazil. Consequently, it will be possible to obtain the municipality’s average REEC and, in further studies, income can be ascertained for each municipality.

It was observed that the data has seasonality, which could be extracted using a decomposition census method type multiplicative and new series should be obtained for the data. The new series must, again, be transformed into a standard z-normal series, so they all have the same characteristics – zero mean and unit variance.

The data was introduced in R (2017), a statistical package, which is a collaborative project and some functions and algorithms are free to use. The packages used were: dtwclust (Sardá-Espinosa, 2018); ggplot2 (Wickham, 2009); and pastecs (Grosjean and Ibanez, 2018) to transform and obtain the new data set – without seasonality and z-normalized.

3 Results

An algorithm to create a Ward’s agglomerative hierarchical clustering with a DTW distance was programmed. Among the results, a mix of Cluster Validation Indices were calculated (Arbelaitz et al., 2013), the results were obtained for nine different sizes (K=5 to K=13) and seven indices. The indices were ordered by their importance, where the best result is order 9 and the worst is 1. So, the results suggest two possible numbers of clusters, one with five groups and another with twelve.

Considering the CVI results, a dendrogram is presented in Fig. 1. The first cut is made, so five clusters would be observed, the other cut is when twelve clusters are formed, the third cut, made within the first cluster, is a possible division if thirteen clusters were to be formed.
Fig. 1. Dendrogram for REEC Average Residence, distance DTW and Ward Method

In Fig. 1, the clusters could be emphasized, their size could be measured and, also, their members could be identified. Note that, deciding between five and twelve clusters will depend upon: (i) their path through time; and, (ii) their geographic location. Subsequently, for now, both divisions will be taken into consideration.

It should be noted that clusters numbers are presented at the bottom of the figure because they don’t follow any order and their identification will be needed in further analysis. Moreover, their division is also observed at the bottom, for instance, if clusters are to be divided to form twelve clusters, their division is also observed. For instance, if cluster one is divided, it will generate clusters one, two and six. Note that clusters two, three and five will be divided and, as a result, five more clusters of single members will be formed (8, 9, 10, 11 and 12). Also, observe that the only cluster that will not be divided is cluster four. In addition, the labeling given for the first division, into five clusters, is not the same, a new cataloging was given to the second division into twelve groups.

Considering the elements of each group, the series could be plotted, and their tendency could be observed by its centroid, as suggested before. Fig. 2 represents those plots. Unfortunately, the series’ members in each plot could not be identified, but what is important to show, is the path for each group.
As shown in Fig. 2, the divisions into five (a) and twelve (b) clusters are presented. The vertical line in all clusters represents December 2014 and it is mentioned as a break that occurs in most series. Politically, it was a presidential election year, which lead to the reelection of the President in 2015. In 2016, political instability lead to the impeachment of the reelected President. Economically, it was the beginning of a crisis, which lead to high rates of unemployment, inflation and public deficit, three major problems. Both, political and economic, scenarios affected household income and, therefore, REEC in different moments. The crises still haven’t passed by the end of 2017, the end of the time series used in this study. Nonetheless, some economic measures have been taken and a few economic indicators have alleviated, such as inflation and unemployment, in 2017.

Considering the divisions observed in Fig. 2, it is easy to link it to Fig. 1 where clusters have been labeled. Cluster one in Fig. 2 (a) was divided into one, two and six in Fig. 2 (b), the same division observed in Fig. 1. Cluster four in (a) is the same as cluster five in (b). Clusters two, three and five in (a) have a very similar path to clusters three, four and seven in (b), some members were taken to form single clusters in (b), but they have the same original tendency, and a slight change in path because some members were taken.

Recall that the ones to be studied are those with upward tendency over time, especially for recent years. Considering that, clusters two and three in (a) and three and four in (b) are not to be analyzed.

Regarding cluster one’s division, the three clusters (one, two and six) created and presented in (b), showed significantly different paths from cluster one in (a); thus, twelve clusters is a better division than five groups.

The centroids and series for groups one, two, five, six and seven in (b) are presented in Fig. 3. Their centroids are a representation of their tendency, and, as can be noticed, all clusters maintain a crescent tendency for most periods. Towards the end, around the vertical cut, some of them present a downward or constant tendency while group six presents an upward tendency.
Regarding Fig. 3, note that clusters five and seven, present a constant tendency after the break, group one and two suffer a break approximately a year and half later and their centroids present a downward tendency afterwards. Group six is an exception; its tendency is upward, suffering a later break that changes its level but not its upward tendency.

If another division is imposed, cluster six is the one to be divided, it is the third division showed in Fig. 1 above. Fig. 4 shows their series and centroids.

When cluster six is divided and towards to recent data, a difference is noticed between them. The second (6b) suffers a break around 2014, as the others, but recovers fast, presenting as upward tendency. The first one (6a), presents a smoother path after 2014, and suffers a break later, showing an upward tendency towards recent months. Emphasizing once again, group 6a diminishes the intercept and presents a break with an upward tendency and group 6b presents a smaller intercept but preserves the tendency without a break.

Allowing for the geographic location of clusters’ members, it is possible to visualize them on a map (Fig. 5).

**Fig. 3. Clusters’ Centroids for Relevant Clusters, REEC Average Residence**

**Fig. 4. Cluster Six’s Division**
Fig. 5. Geographic Distribution for Average REEC Clusters

Clusters five and seven, which were the ones to have a break and constant tendency afterwards, are closely located, restricted mostly in four states – SP, PR, SC and ES –. Those states are known by its industries and congregate 45% of Brazil’s GDP in 2015 (IBGE, 2017). Cluster one is divided into five states – AM, PA, PE, BA and AL – in North and Northeast Brazil, roughly represent 11% of GDP in 2015 (IBGE, 2017). Cluster two is spread all over Brazil, in six states – AP, AC, RN, SE, GO and part of RS – congregating about 7% of Brazil’s GDP in 2015. Cluster 6 is present in all regions of Brazil, but mostly in Center-west and Northeast Brazil – RO, TO, MA, PI, CE, PB, part of RJ, MT, MS and part of RS –, it represents about 19% of Brazil’s GDP in 2015.

Taking into consideration data from the National Household Sample Survey (IBGE, 2018), 2014 and 2017, it is possible to estimate labor distribution by sectors – Primary; Industry; and Tertiary – for groups, considering not municipalities but states. The average percentage of employment distributed among sectors is presented in Fig. 6. It indicates that all clusters have the tertiary sector holding most of the employment and, during those final years, it has increased. The industry sector employs more in clusters five and seven than in the others. The primary sector employs more in groups one and six than in group two. It is important to notice that the tertiary sector increased in all groups, when comparing 2014 and 2017; so, it is easy to say that, while in crisis, self-employment has a place in the tertiary sector rather than any other.
Regarding income, the United Nations Development Programme – PNUD Brasil, present results for the Human Development Index and its income dimension – HDI-I – which is available in the Atlas do Desenvolvimento Humano no Brasil, by municipality, in 2000 and 2010. For this study, a comparison was calculate between those years and a positive variation higher than 15% is presented in Fig 7. It shows municipalities that had improved their income situation. Therefore, higher percentage represents that families living on those areas are in 2010 better off than in 2000. As it could be seen in Fig. 7, North and Northeast are the areas with higher gain in HDI-I between 2000 and 2010.

Combining the two maps, one with the clusters and one with results from HDI-I – Fig. 8 – it could be easily noticed that cluster 6a was the one with the majority number of municipalities with higher increases in HDI-I over most of its territory.
As shown above, most municipalities are in Northeast Brazil, where cluster 6a and part of cluster one are located. It is also easy to visualize, that they are concentrated inland, where the primary sector is a majority. So, it is likely to be true that the improvement of the standard of living for families in those areas, is related to gains from the primary sector.

4 Conclusion

Applying time series clustering methodology to average REEC proves to be a resourceful mechanism in finding similar patterns over time. This methodology exposes the possibility of identifying homogeneous populations – household consumers – that likely respond to some exogenous economic facts throughout time. Since population is linked to a geographic location, it is possible to verify if such groups belong to neighboring area.

The methodology identifies homogeneous groups as well as groups’ tendency over time, which is very helpful and desirable to public decision makers by whom policy can be directed and implemented in areas where such policies are needed. The implementation of said methodology, serves not only to discriminate, but also to identify tendency patterns linked to neighboring geographic areas – a useful advance if such methodology could be implemented on other economic fields as well.

After groups were formed, it was possible to link economic factors to them and verifies if such factors are related to some deviation of pattern. For instance, it was possible to recognize the impact of the economic and political crises, occurring since 2014, in all groups. Furthermore, it was possible to link some economic factors, such as wealthy and employment, to directions taken by groups.
Results for clusters were presented showing that it is possible to link their average REEC behavior to some external variables. However, further study will be prudent. For instance, to verify the influence of income on REEC over time using such methodology is possible, but both variables must be measured over time and municipality in order to implement the time series clustering methodology. Although, it was possible to have some insight about the influence of income over the average REEC, as will be further analyzed.

Results for clusters five and seven reveal that: residents in such groups adapted their average REEC faster than the others clusters’ residents. They imposed changes in such a way that, the average REEC tends to its series’ average. Clusters five and seven were the first to suffer when crisis took place, residents could quickly adapt by cutting down some house conveniences when crises occurs, probably because these residents are employed in the industry sector, and this sector is the first one affected in a crisis. Consequently, households are more likely to refrain from using some home appliances – such as, dishwasher or the air conditioners, for example – while waiting for better economic conditions to return.

All clusters suffer with the crises, but its impact took place latter and in a downward tendency for clusters one and two. Cluster one has an important primary sector, represents a significant percentage of PIB but, its member does have an important industry sector, that were significant, and they did endure the crisis’ impact. Cluster two, on the other hand, suffer less with the crisis, possible because its members are employed mostly on the tertiary sector and their industry sector is linked to primary goods. Both clusters’ members decided to consume fewer conveniences over the time, for example, some time ago not use dishwasher, another time, not use air condition, and so on. The result was that, over time, the average REEC was decreasing. It was faster to group one and their consumption nowadays is closer to average REEC over time. However, for group two, their consumption still above the average.

Cluster six is the one with a very different pathway for the average REEC. It has an upward tendency, with important primary sector, their states are wealth, their industry sector is also linked to the primary sector. But, what differentiates it from the rest is the change on income over the years. Considering the HDI-I, it is visually easy to prove that, municipalities on that area were the ones with higher improvement on that index. Meaning that, more members in that cluster have changed their standard of living, no other cluster had so many members changing their standard of living. The change was preserved in what concerns average REEC, they did not diminish in the face of a crisis. Rather, they prefer to increase.

When cluster six is divided into clusters 6a and 6b, there was a slight difference between them. Cluster 6a, where a great number of members appears, they prefer to continue consuming and only diminished their level towards the end of the series. Members from cluster 6b, on the other hand, decided to diminish their level when the crisis was implemented, but continues to increase average REEC. Those are the areas where public policy should focus and should guaranty increasing supply of REEC.
References