Aguilar Castro, J.L., Quintero Gull, C., Rodriguez Moreno, M.D. & Viera, J. 2022, "Analysis of customer energy consumption patterns using an online fuzzy clustering technique", in 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 18-23 July 2022.

Available at http://dx.doi.org/10.1109/FUZZ-IEEE55066.2022.9882764

Analysis of Customer Energy Consumption Patterns using an Online Fuzzy Clustering Technique

Jose Aguilar
Universidad de Alcalá, Escuela Politécnica Superior, ISG,
Alcalá de Henares, 28805, Spain;

CEMISID, Universidad de Los Andes,
Mérida, 5101, Venezuela;

GIDITIC, Universidad EAFIT,
Medellín, 50022, Colombia
jose.aguilar@uah.es

Carlos Quintero Gull
Dpto de Ciencias Aplicadas y Humanísticas. Doctorado en Ciencias Aplicadas Facultad de Ingeniería Universidad de Los Andes
Mérida 5101, Venezuela
carlgull@gmail.com

Maria D. R-Moreno, Juan Viera
Universidad de Alcalá, Escuela Politécnica Superior, ISG,
Alcalá de Henares, 28805, Spain;

TNO, Intelligent Autonomous Systems Group (IAS),
The Hague, The Netherlands
malola.rmoreno@uah.es

***Abstract*—Currently, there is a high rate of generation of new information about the Energy Consumption of customers. It is important the traceability of its consumption pattern evolution to determine in real-time the services of a smart energy management system. This paper analyses the evolution of the Energy Consumption Pattern of customers using the Learning Algorithm for Multivariable Data Analysis (LAMDA). LAMDA is a fuzzy approach for supervised and unsupervised learning, based on the calculation of the Global Adequacy Degree (GAD) of one individual to a class/cluster, through the contributions of all its descriptors. LAMDA can create new classes/clusters after the training stage (online learning). If an individual does not have enough similarity to the preexisting classes/clusters, it is evaluated with respect to a threshold called the Non-Informative Class (NIC) to define if it is a new class/cluster. Particularly, the algorithm of the LAMDA family used in this paper is LAMDA-RD (Robust Distance). In the paper is analyzed the patterns of the initial grouping of the data, as well as, the patterns through their evolution (traceability). For the analysis of the patterns different metrics are considers: Calinski- Harabasz Index and Silhouette Score.**

Keywords—Clustering algorithm, fuzzy systems, LAMDA, Online Learning, pattern evolution

I. INTRODUCTION

The analysis of the evolution of the energy information is relevant for different tasks in a smart energy management system: control, optimization, supervision, among other areas [1, 2]. For that, it is necessary to propose techniques that allow analyzing the evolution of the energy information. For example, it is necessary to identify the patterns that represent the common information in data groups. From these "patterns", their evolution can be analyzed, that is, as they change over time. One domain where that can be interesting is for analyzing the evolution of the energy consumption patterns of the customers.

The interest in the study of the energy consumption patterns is due to the immense global demand for energy, so there is a concern to achieve greater efficiency and optimization of the energy consumption. To do this, it is necessary to identify the consumption pattern of users. With a consumption pattern is possible to characterize the potential energy demand according to the users' needs, and to define specific energy management strategies. Normally, the studies about energy consumption are focused on defining mechanisms to optimize consumption [1, 2], and not to define patterns of consumption, so even less about dynamic patterns of consumption.

This article analyzes the LAMDA algorithm [3, 4, 5] in the problem of tracking the traceability [6, 7, 8] of the energy consumption patterns. To do this, we define the pattern identification problem as a clustering problem, where the centroids represent the pattern of each cluster. From the initial clusters that define the first patterns, it is analyzed how each cluster changes, that is, its pattern evolution. From the evolution of the patterns, aspects such as the re-association of the objects to the closest patterns can be established [9, 10]. Also, another aspect to consider is how to reestablish the relevance of a pattern in a given context, such that if a given characteristic of what is sought changes (for example, its consumption profile) then this pattern can be more or less relevant to what is sought [9, 10].

LAMDA has been considered because is an online unsupervised learning technique (clustering), that is, a good candidate to determine the traceability of patterns [3, 4, 5]. This technique follows an incremental approach of clustering such that can add a new cluster or adjust the current known clusters. Specifically, the LAMDA algorithm is a method based on fuzzy logic that calculates the Global Adequacy Degree (GAD) from an individual to a group (clustering problems) or class (classification problems) [3, 4,5]. LAMDA can create new classes/clusters after the training stage when an individual has not enough similarity to the preexisting classes/clusters. For that, it is evaluated this similarity with respect to a threshold called the NIC. LAMDA algorithm has been modified by different works for different purposes [11, 12, 13] but in this article, we will use the LAMDA RD algorithm from the LAMDA family because it improves significantly the performance of the traditional LAMDA algorithm in the clustering problem [12]. In LAMDA-RD is defined an automatic merge technique to update the cluster partition performed by LAMDA to improve the quality of the clusters, and a new methodology to calculate the Marginal Adequacy Degree (MAD) that improves the individual-cluster assignment [12].

Thus, the main contribution of this paper is the proposition of an approach based on LAMDA-RD to determine the evolution of the energy consumption patterns of the customers, and the analysis of the traceability of the energy consumption patterns based on the patterns determined by LAMDA-RD. This work is organized as follows: Section 2 introduces the fundamentals of LAMDA-RD. Section 3 describes our approach for the definition of dynamic customer energy consumption patterns. Section 4 shows the experiments and presents the analysis of the problem of tracking the traceability of the energy consumption patterns, and finally, section 5 shows the conclusions and future works.

II. LAMDA AND LAMDA-RD

In this section, we will present briefly the basis of the LAMDA-RD algorithm [12]. In general, LAMDA is an algorithm based on fuzzy logic that assigns individuals to a class according to the GAD [3, 4, 5]. The main features of LAMDA are: a) it does not require to define the number of clusters, b) it can generate new clusters with new individuals with not enough similarity with the preexisting clusters (incremental-learning). The analysis of the similarity compares the features of any object/individual $X = [x_1; x_2; \dots; x_j; \dots; x_n]$, where x_j is the descriptor j of the object X, with those of the existing classes/clusters $C = C_1$; C_2; ...; C_k; ...; C_m. The features of the objects are normalized to [0,1] to standardize the individuals, according to the minimum and maximum values (see Eq. (1)):

$$\bar{x}_j = \frac{x_j - x_{jmin}}{x_{jmax} - x_{jmin}} \tag{1}$$

Where: $\bar{x}_j$: Standardized descriptor/feature; $\bar{x}_{jmin}$: Minimum value of descriptor j; $\bar{x}_{jmax}$: Maximum value of descriptor j

Below, we present the main definitions of LAMDA. With normalized values, LAMDA calculates the MAD (see definition 1).

Definition 1: Marginal Adequacy Degree (MAD). It describes the similarity of any feature with the corresponding feature of a given class. MADs are calculated using probability density functions, and one of the most common is the Fuzzy Binomial function, shown in Eq. (2).

$$MAD(\bar{x}_j/\rho_{k,j}) = \rho_{k,j}{}^{\bar{x}_j}\,(1-\rho_{kj})^{(1-\bar{x}_j)} \tag{2}$$

Where: $\rho_{k,j}$: is the average value of the descriptor j that belongs to the class k, calculated using Eq. (3):

$$\rho_{kj} = \frac{1}{n_{kj}} \sum_{t=1}^{n_{kj}} \bar{x}_j(t) \tag{3}$$

Where: n_{kj}: number of observations of class k and descriptor j.

After obtaining the MADs, LAMDA computes the GADs for every class using aggregation functions T- norm and S-norm (see definition 2).

Definition 2: Global Adequacy Degree (GAD). It describes the adequacy of an individual to each class. This value is determined according to Eq. (4):

$$\begin{aligned} GAD_{k,\bar{X}} = \left(MAD_{k,1}, MAD_{k,2}, \dots, MAD_{k,n}\right) &= \alpha T\left(MAD_{k,1}, \dots, MAD_{k,n,}\right) \\ &+ (1-\alpha) S\left(MAD_{k,1,} \dots MAD_{k,n}\right) \end{aligned} \tag{4}$$

Where $\alpha \in [0,1]$ is the exigency parameter; T and S are the aggregation operators. An example of T is $T(a,b) = \min(a,b)$.

Finally, in clustering tasks, the normalized object X is assigned to the group with the maximum GAD (see definition 3).

Definition 3. Let $p = \{1, \dots, m\}$ the number of current clusters. The object $\bar{X}$ is assigned to the cluster with the maximum GAD, where the index corresponds to the number of the cluster.

$$index = \max\left(GAD_{1,\bar{X}}, GAD_{k,\bar{X}}, \dots, GAD_{m,\bar{X}}, GAD_{NIC,\bar{X}}\right)$$

NIC is used to create new clusters, when an object is unrecognized (it is sent to the *NIC*), making the algorithm more adaptive (online learning). It is considered $\rho_{NIC} = 0.5$, because, with this value in the probabilistic function (Eqs. (2)), the $MAD_{NIC} = 0.5$ for any value of the descriptor $\bar{x}_j$.

Below, we present the main definitions of LAMDA-RD [12]:

In some applications, the number of created clusters by LAMDA does not correspond with the number of desired clusters. LAMDA-RD improves the performance of traditional LAMDA in clustering problems with an automatic merge strategy to update the cluster partition performed by LAMDA to improve the quality of the clusters, and a new approach to calculate the MAD [12].

Definition 4: Cauchy Marginal Adequacy Degree (CMAD). It corresponds to the MAD using the Fuzzy Cauchy Function:

$$CMAD = \frac{1}{1 + dist(\bar{x}_j, \rho_{kj})} \tag{5}$$

Where: $dist(\bar{x}_j, \rho_{kj})$ is the distance between descriptor j of individual $\bar{X}$ ($\bar{x}_j$) and the average of descriptor j in class k (ρ_{kj}).

Definition 5: Robust Marginal Adequacy Degree (RMAD). It corresponds to the MAD using a factor that penalizes each cluster:

$$RMAD = k_{\bar{x}k} * CMAD \tag{6}$$

$k_{\bar{x}k}$ is determined using two parameters, the average distance of the individual to the cluster ($d_{k,\bar{X}}$), and the average distance between neighbor clusters $(d_{n,b})$.

Definition 5: density of cluster. Let $d_t \in [0,1]$ a threshold of the density of a cluster obtained through a calibration process.

Definition 6: penalty factor. It is determined using Eq (7) when the distance $d_{k,\bar{X}_r}$ is greater than $d_{n,b}$:

$$k_{\bar{x}k} = \frac{d_{n,b}}{d_{n,b} + dist\left(d_{k,\bar{X}_r}, d_{n,b}\right)} \tag{7}$$

Definition 7: new GAD. GAD is a linear combination of the RMADs, where T and S are the aggregation operators.

$$\mathrm{GAD}_{k,\bar{\mathrm{X}}} = \left(\mathrm{RMAD}_{\mathrm{k,1}}, \mathrm{RMAD}_{\mathrm{k,2}}, \ldots., \mathrm{R\,MAD}_{\mathrm{k,n}}\right) = \alpha T\left(RMAD_{k,1}, \ldots, RMAD_{k,n,}\right) + (1 - \alpha) S\left(RMAD_{k,1,} \ldots RMAD_{k,n}\right) \tag{8}$$

Additionally, LAMDA-RD has a strategy for the automatic fusion of clusters, according to the next definitions [12]:

Definition 8: A cluster C_k is defined by the tuple:

$$C_k = \left(\rho_{kj}, \bar{X}_k, index, n_k\right) \tag{9}$$

Where: ρ_{kj} it's the centroid of descriptor j in C_k, $\bar{X}_k$ it's the set of individuals in C_k, $index$ is the identifier of C_k, and n_k number of elements of C_k.

Definition 9: A neighbor cluster C_{nb} it's defined by the tuple:

$$C_{nb} = \left(\rho_{nb,j}, \bar{X}_{nb}, index, n_{nb}\right) \tag{10}$$

According to the LAMDA bases, the maximum GAD is where the individual is assigned, then, Morales et al. [12] concluded that the second GAD of greater value is the nearest neighbor cluster.

Definition 10. The compactness of a cluster C_{nb} is determined by the mean value of all distances between the individuals belonging to the cluster C_{nb}.

$$t_{nb,j} = \frac{\sum_{i=1}^{n_{nb}-1} \sum_{m=i+1}^{n_{nb}} \left| x_{nb,j}^{-i} - x_{nb,j}^{-m} \right|}{n_{nb} x (n_{nb}-1) x \ldots\ldots x 1}; \forall\, j = 1,2, \ldots.n \tag{11}$$

Where, $x_{nb,j}^{-i}$ is the descriptor j of individual i in the cluster C_{nb}.

Definition 11. The number of individuals in the overlapping area is defined by the number of individuals between two clusters C_k and C_{nb} whose distance between their individuals is less than $t_{nb,j}$. It is determined by Eq. (12).

$$N_I = N_{kl} + N_{nbl} \tag{12}$$

Where N_{kl} is the number of individuals of C_k in the overlapping area (see Eq. (13) and N_{nbl} is the number of individuals of C_{nb} in the overlapping area (see Eq. (14)).

$$N_{kl} = \left\{\forall\, \bar{x}_{k,j} \in C_k \mid d\left(\bar{x}_{k,j}, \bar{x}_{nb,j}\right) < t_{nb,j}\,;\ \forall\, j = 1,2, \ldots.n\right\} => N_k = n(N_{kl}) \tag{13}$$

$$N_{nbl} = \left\{\forall\, \bar{x}_{nb,j} \in C_{nb} \mid d\left(\bar{x}_{k,j}, \bar{x}_{nb,j}\right) < t_{nb,j}\,;\ \forall\, j = 1,2, \ldots.n\right\} => N_{nb} = n(N_{nbl}) \tag{14}$$

Definition 12. The density in the overlapping area between the clusters C_k and C_{nb} is defined as:

$$D_{k-nb} = \frac{N_I}{n_{nb} + n_k} \tag{15}$$

Finally, Morales et al. [12] established that two clusters C_k and C_{nb} can be merged when $D_{k-nb} \geq D_t$ (see definition 13), where $D_t \in [0, 1]$ is a density threshold defined by the user.

Definition 13. The cluster resulting after the merging process is:

$$C_{new} = C_k \cup C_{nb} = \left\{\rho_{new,j}\, \bar{X}_k \cup \bar{X}_{nb}, index, n_k + n_{nb}\right\} \tag{16}$$

With:

$$\rho_{new,j} = \frac{1}{n_k + n_{nb}} \sum_{t=1}^{n_k + n_{nb}} \bar{x}_{new,j}^{-t} \tag{17}$$

III. LAMDA-RD FOR THE DEFINITION OF DYNAMIC PATTERNS OF CUSTOMER ENERGY CONSUMPTION

In this work, we propose an approach to determine the evolution of the energy consumption patterns of the clients using an online clustering technique, LAMDA. To do this, the streaming of data that describes the energy consumption behavior of users is analyzed in real-time. Subsequently are generated/updated the different clusters that reflect the different patterns of energy behavior of users based on their consumption data. Finally, to analyze the evolution of the clusters, metrics are calculated, which can be used to determine information of relevance about the energy behavior of the users for a smart management system. Figure 1 describes the previous process.

The expected result is a set of clusters that group the different types of clients. Thus, the centroids of each cluster will represent the behavior pattern of the individuals (clients) that form part of it. In this way, it will be possible to analyze the centroid of each group, to study its trends and characteristics of energy consumption. On the other hand, as we have already explained, the system will work online, therefore, the clusters will be constantly readjusting (the clusters are updated, and may even appear and disappear) to reflect the changing behavior of the energy consumption over time.

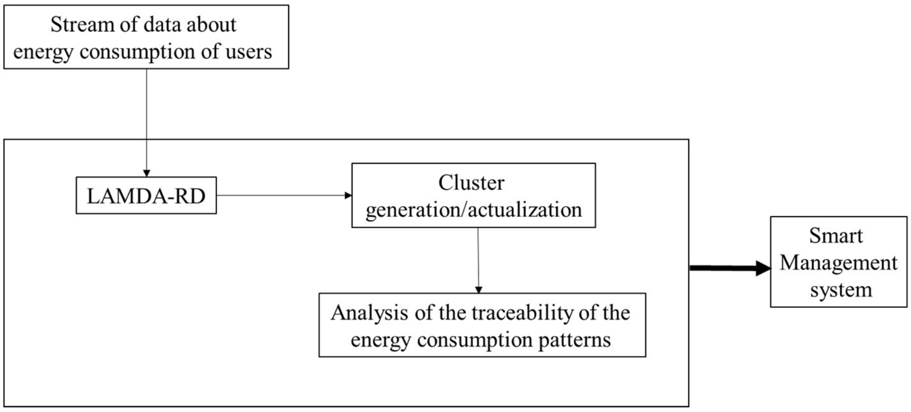
Stream of data about energy consumption of users
LAMDA-RD
Cluster generation/actualization
Analysis of the traceability of the energy consumption patterns
Smart Management system

FIGURE 1. GENERAL SCHEME FIR THE ANALYSIS OF DYNAMIC PATTERNS OF CUSTOMER ENERGY CONSUMPTION USING LAMDA-RD

IV. EXPERIMENTS AND RESULTS

A. Metrics and Dataset

For the evaluation of the results in our context, we use the Calinski-Harabasz Index (used as an intracluster metric), and Silhouette Score (used to define the quality of the clusters).

Calinski-Harabasz Index: This index is defined as the ratio of the sum of between-cluster dispersion and within-clusters dispersion for all clusters, where dispersion is the sum of distances squared (see Calinski and Harabasz [14, 15] for more details). A higher Calinski-Harabasz score is a result with better defined clusters.

Silhouette Score: is a measure of cohesion compared with the separation in the clusters [16]. It determines how similar an object is in its own cluster, compared to other clusters. Values near 1 are desirable, near 0 indicate overlapping clusters, and negative values generally indicate that a sample has been assigned to the wrong cluster.

For this work, the dataset with a simple of the clients of the EPM (Empresas Públicas de Medellín) [17] was used, which contains time series of energy consumption of users, whose most relevant variables are the data of the clients (id, profession, work, address, among other personal data), in addition of his/her daily residential energy consumption. On the other hand, a feature engineering process was carried out to determine the variables that contributed the most to the grouping process focused on energy consumption, with a feature reduction phase as established in the methodology proposed in [18]. On the other hand, the data were grouped into quarterly or semi-annual periods, according to their timestamps.

B. Results

In Tables 1 and 2 we see the results for the clustering problem using LAMDA-RD. Something important to note is that LAMDA-RD initially proposed 5 clusters, and for the data that was arriving from other time periods, it never added new clusters. We can see that the quality of the clusters is quite good through the metrics. According to the silhouette value over time, the quality of the groups worsened, but according to Calinski-Harabasz, the overall result was better. In particular, what the metrics tell us is that over time, LAMDA-RD managed to define clusters that were more separated from each other (indicated by the Calinski-Harabasz values) but with less internal cohesion without overlapping each other (indicated by silhouette values over time). This may be because LAMDA-RD fails to get enough reasons to create new clusters (patterns) but instead tries to add them to the already known ones (it updates its patterns). We discuss that in the next section.

TABLE 1. QUALITY OF THE CLUSTERING RESULTS BY TRIMESTER

Period (by trimester)	**Calinski-Harabasz**	**Silhouette**
First	1559	0.65
Second	1675	0.59
Third	2334	0.45
Four	2676	0.38

TABLE 2. QUALITY OF THE CLUSTERING RESULTS BY SEMESTER

Period (by semester)	**Calinski-Harabasz**	**Silhouette**
First	1754	0.59
Second	2534	0.56

Now, regardless of the results, we see the ability of LAMDA-RD to follow the dynamic behavior of the users' energy consumption patterns. On the other hand, the analysis periods (quarterly or half-yearly) have little influence on the quality of the results. The evolution analysis periods, in this case, would depend more on the interest of the energy management system where we would like to include this service.

C. Analysis of the Traceability of the Patterns

This section analyzes the traceability of the energy consumption patterns discovered from the data, using LAMDA-RD. In Figure 2, the results for the different clusters for a quarterly period are shown. Particularly, what is being graphed is the energy consumption value of the centroid of each cluster, but we will go on to detail the rest of the variables that, according to the results of the feature engineering process, were of interest for the clustering process, to describe the energy consumption (which were correlated to it). Those correlated variables were two: age and location.

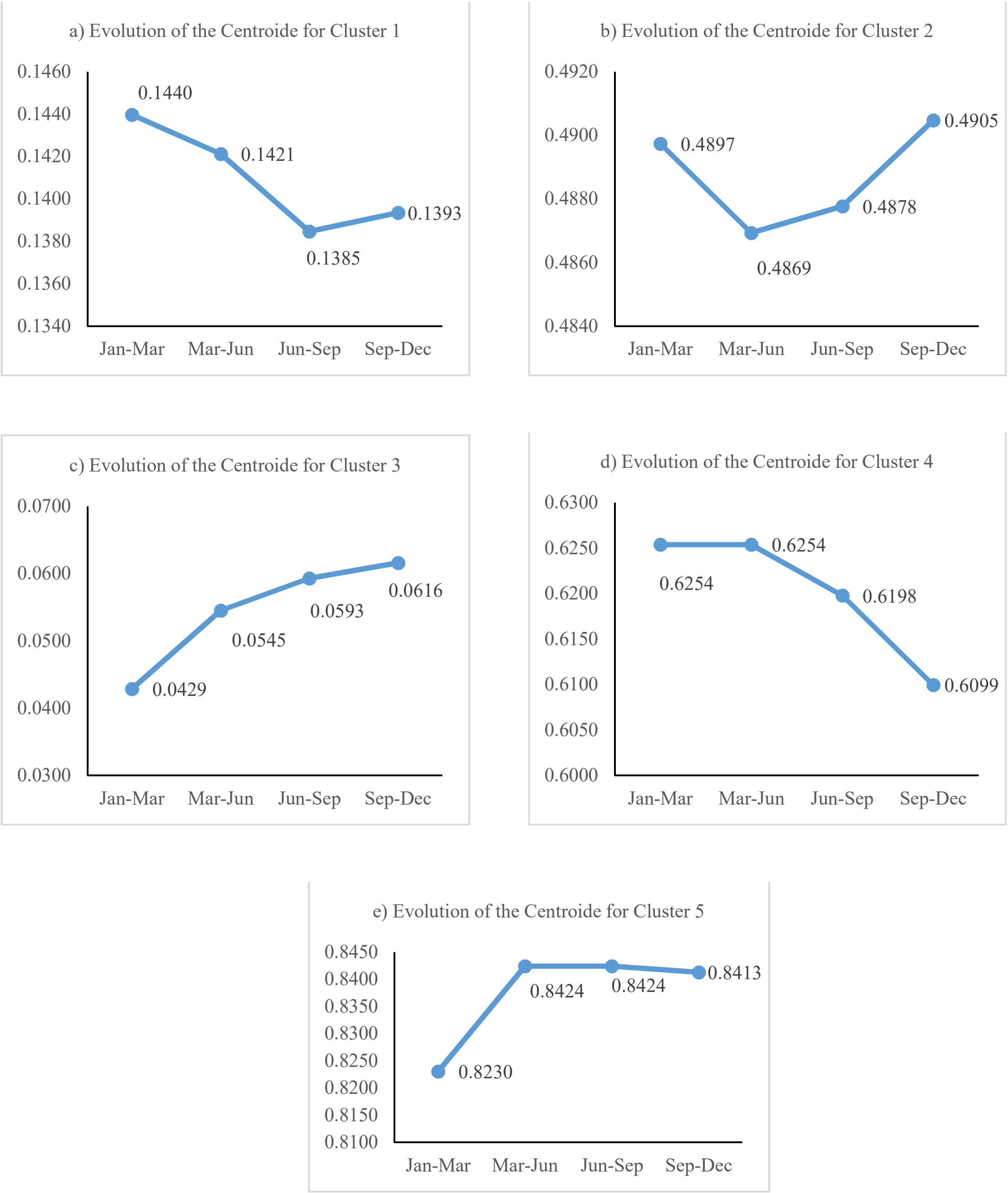

FIGURE 2: EVOLUTION OF THE ENERGY CONSUMPTION VALUE OF THE CENTROIDE OF THE CLUSTERS DURING A YEAR

Figure 2 shows the quarterly evolution of the energy consumption value of the 5 clusters. It is evident that despite the changes that occurred in the rest of the variables, this value is more or less stable, which means that each cluster characterizes very well the different energy consumptions of the users.

Now, we will proceed to identify the characteristics of each cluster taking into account its centroids. Clusters 1 and 3 (see Figures 2.a and 2.c) have centroids whose clients are young (this variable is around 25 years old for cluster 3 and 29 years old for cluster 1), and their homes are in a town area that is a modern urban area of recent development (El Poblado, Medellin, Colombia). On the other hand, the highest consumption is for cluster 5, which are clients whose ages are over 55, but whose areas of residence vary. Finally, clusters 2 and 3 have in their centroid an age value that is above 35 years but does not reach 55 years.

On the other hand, the quarterly consumption of the clusters is very stable (the variations are not large), without any marked trend, with the exception of clusters 4 and 5 (see Figures 2.d and 2.e) with a tendency to decrease (very little in cluster 5) which is surely due to the activities carried out in those months of the year by that age group of the clients. Also, cluster 3 tends to rise, but very slightly ((see Figure 2.c).

What the clusters tell us is that electricity consumption is lower in young populations living in recent modern urban areas, which characterizes the profile of users who spend few hours at home. In general, the clusters identify very well the age range vs. energy consumption. Also, the behavior of the clusters is very stable over time, and it is due to the climatic conditions of the site (Medellin, Colombia), with a tropical climate that means that the weather does not change much over time (rainy or dry season) that increases/decreases the energy consumption.

V. Conclusions

This article studies the capabilities of LAMDA-RD to the problem of tracking the traceability of the energy consumption patterns. Particularly, the paper studies the evolution of the characteristics of the patterns over time. LAMDA-RD has shown the capability to carry out an incremental approach of clustering to adapt the cluster model, via the creation of new clusters or adequacy of the known clusters. The paper analyzes the characteristics of the patterns (their variables) through their evolution (traceability).

For the evaluation, the paper considers two metrics: Calinski-Harabasz Index (used as an intracluster metric) and Silhouette Score (used to define the quality of the definition of the clusters). In general, for the Silhouette Score, LAMDA-RD has good results, with descent in time. In the case of the Calinski-Harabasz Index, the results are very good, with improvement over time. Thus, we observe that the tracking of the traceability of the patterns can be carried out with LAMDA-RD, exploiting its online clustering process.

These are initial results on the use of LAMDA-RD for analysis of the energy behavior of users. Future works should be carried out to make a more exhaustive analysis of the variables of interest for a smart energy management system that describes the energy consumption of its users. For now, the variables of age and location (results of the characteristics engineering process carried out) were considered, but perhaps other variables are required to allow more efficient energy management. For example, user habits, activities carried out in the home, home occupancy rate, etc., are variables that allow optimal planning of the energy consumption in a home. Future works should determine strategies that define how to incorporate them into the processes of building dynamic patterns of energy consumption.

The advantage of this algorithm is that, unlike traditional algorithms, in which the assignment of individuals is based on the distance to the centroids, LAMDA-RD assigns individuals based on their membership degree to each cluster. Additionally, this proposal has a fusion process to improve the quality of the clusters.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska Curie Grant Agreement No. 754382. M.D. R-Moreno is supported by the JCLM project SBPLY/19/180501/000024 and the Spanish Ministry of Science and Innovation project PID2019-109891RB-I00, both under the European Regional Development Fund (FEDER).

References

[1] J. Aguilar A. Garces-Jimenez M. R-Moreno R. García “A systematic literature review on the use of artificial intelligence in energy self-management in smart buildings”, Renewable and Sustainable Energy Reviews, vol. 151, 2021.

[2] L. Hernández, A. Hernández-Callejo, Q., Zorita-Lamadrid, F. Duque-Pérez, A. Santos García, “Review of strategies for building energy management system: Model predictive control, demand side management, optimization, and fault detect & diagnosis”, Journal of Building Engineering, vol. 33, 2021.

[3] J. Aguilar-Martin, R. L. De Mantaras, “The process of classification and learning the meaning of linguistic descriptors of concepts,” Approximate reasoning in decision analysis, vol. 1982, pp. 165–175, 1982.

[4] T. Kempowsky, A. Subias, J. Aguilar-Martin, “Process situation assessment: From a fuzzy partition to a finite state machine”. Engineering Applications of Artificial Intelligence,vol. 19, pp. 461-477, J. 2006.

[5] J. Waissman R. Sarrate T. Escobet J. Aguilar B. Dahhou “Wastewater treatment process supervision by means of a fuzzy automaton mode”l, In Proc. IEEE International Symposium on Intelligent Control, pp. 163-168, 2000.

[6] J. Aguilar, C. Salazar, J. Monsalve-Pulido, E. Montoya, H. Velasco, “Traceability Analysis of Patterns using Clustering Techniques”, Advances in Artificial Intelligence and Applied Cognitive Computing (H. Arabnia et al. (eds.)), Transactions on Computational Science and Computational Intelligence book series, Springer , pp. 235-250, 2021.

[7] G. Aiello, M. Enea, and C. Muriana, “The expected value of the traceability information,” European Journal of Operational Research, vol. 244, no. 1, pp. 176–186, 2015.

[8] C. Mills, J. Escobar-Avila, A. Bhattacharya, G. Kondyukov, S. Chakraborty, and S. Haiduc, “Tracing with Less Data: Active Learning for Classification-Based Traceability Link Recovery,” in Proc. IEEE International Conference on Software Maintenance and Evolution (ICSME), pp. 103–113, 2019.

[9] H.. Zadeh and R. Boostani, “A novel clustering framework for stream data,” Canadian Journal of Electrical and Computer Engineering, vol. 42, no. 1, pp. 27–33, 2019.

[10] Md Arafatur, N. Zaman, A. Taufiq, F. Al-Turjman, Md. Z. Alam, M. Zolkipli, “Data-driven dynamic clustering framework for mitigating the adverse economic impact of Covid-19 lockdown practices”. Sustainable Cities and Society vol, 62, 2020.

[11] L. Morales, J. Aguilar, D. Chavez, C. Isaza, “LAMDA-HAD, an extension to the Lamda classifier in the context of supervised learning”. International Journal of Information Technology & Decision Making, vol. 19, pp. 283-316, 2018.

[12] L. Morales, J. Aguilar, "An Automatic Merge Technique to Improve the Clustering Quality Performed by LAMDA," IEEE Access, vol. 8, pp. 162917-162944, 2020.

[13] L. Morales Escobar, J. Aguilar, A. Garcés-Jiménez, J. A. Gutierrez De Mesa and J. M. Gomez-Pulido, "Advanced Fuzzy-Logic-Based Context-Driven Control for HVAC Management Systems in Buildings", IEEE Access, vol. 8, pp. 16111-16126, 2020.

[14] J. Aguilar, I. Bessembel, M. Cerrada, F.; Hidrobo, F. Narciso, “Una Metodología para el Modelado de Sistemas de Ingeniería Orientado a Agentes”, Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial, vol. 12, no. 38, pp. 39-60, 2008.

[15] T. Caliński and J. Harabasz, “A dendrite method for cluster analysis,” Communications in Statistics-theory and Methods, vol. 3, no. 1, pp. 1–27, 1974.

[16] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” Journal of computational and applied mathematics, vol. 20, pp. 53–65, 1987.

[17] https://cu.epm.com.co/

[18] F. Pacheco C. Rangel J. Aguilar M. Cerrada A. Altamiranda “Methodological framework for data processing based on the Data Science paradigm" In Proc. XL Latin American Computing Conference, 2014.