Energy Profile Clustering with Balancing Mechanism towards more Reliable Distributed Virtual Nodes for Demand Response

Ioannis Koskinas *, Apostolos C. Tsolakis *, Venizelos Venizelou †, Dimosthenis Ioannidis *,
George E. Georghiou †, and Dimitrios Tzovaras*

*Centre for Research and Technology Hellas, Information Technologies Institute, 57001 Thessaloniki, Greece
{jkosk, tsolakis, djoannid, dimitrios.tzovaras}@iti.gr
† Department of Electrical and Computer Engineering, University of Cyprus, 1678 Nicosia, Cyprus,
{venizelou.venizelos, georgiou}@ucy.ac.cy

Abstract—As the energy markets become more dynamic, customers' segmentation has become a major concern, especially for Aggregators that contain Distributed Energy Resources in their portfolio. Furthermore, the management complexity in the direction of insightful Demand Response (DR) actions that will yield high profit margins and will hedge against economical risks has been increased since the incorporation of low and medium customers in DR programs. Grouping customers, as independent accumulated virtual nodes (VNs) to the grid, based on their energy profile and their contractual characteristics, facilitates Aggregators overcoming markets' and network's constraints, as well as the designing of collective price policies and purposeful DR strategies. This paper proposes a fully featured methodology that encompasses a soft clustering approach, based on the Gaussian Mixture Model with Expectation Maximization Algorithm, presenting a Temporal Data Dynamic Segmentation (TDDS) algorithm that not only allocates low and medium customers in VNs that share common energy profiles, but also preserves an internal balance in the VNs' resources, in terms of their ability to satisfy reliably DR requests, exploiting the clusters' intersection points to balance the VNs without disrupting their energy profile purity. Experimental results demonstrate an increase in the reliability of each cluster by up to 17.6% without disrupting the clustering coherence.

Index Terms—Soft Clustering, Gaussian Mixture Model, Expectation Maximization, Demand Response, Reliability

I. INTRODUCTION

The development towards smart distribution grids and the decentralization of the power systems requires the technology, modern buildings and other individual assets (e.g., appliances, HVAC systems, etc.) to be energy efficient, as well as energy flexible. The existence of flexibility in power systems is extremely crucial in order to facilitate integration of the highly volatile Renewable Energy Sources (RES) and cover their intermittency with Demand Response (DR) strategies. The achievement of the aforementioned integration requires efficient data monitoring, which has been achieved through advanced metering infrastructures such as smart meters [1]. However, the wide variety of event information and the large volume of data pose high risks in operation and power distribution between electricity customers, which affects the reliability and the profitability of the power network [2], [3]. In addition, current DR markets, require quite significant amounts of available flexibility per customer (e.g. 1-3 MW), making it extremely difficult for small and medium customers to participate in them. For this reason, clustering electricity customers based on their energy characteristics (consumption, generation, storage, etc.) is necessary, and an upcoming promising solution for risk elimination and introduction of new revenue streams.

Clustering is a data mining technique where electric customers are selected and categorized in various groups (clusters) based on their energy profiles. In addition, this method expedites the specification of intrinsic patterns in the big data sets that have emerged. Essentially, given that all smart meters generate large volumes of data, and in most cases without detailed information, their management can significantly be facilitated by grouping data and customers into smaller groups, offering the extraction of higher level of information and the provision of intuitive understanding of their behaviour. For the energy sector, clustering advantages are mainly identified for those who have access to large amounts of energy-related data such as Transmission System Operators (TSOs), Distribution System Operators (DSOs), retailers, utilities, Aggregators and other decision support systems which are responsible for instant operations and fast decision making. As clusters introduce aggregated information to the distribution nodes, and grouped customers can be handled collectively and not individually, the concept of Virtual Nodes (VNs) is used in the presented work for the created clusters. The VNs are easier to handle entities for both Aggregators and systems operators, especially in the context of DR requests.

Going over the literature on the field, the first surveys were done by utilities, system operators and researchers, using the monthly usage and some fixed information (e.g. voltage levels, demand), categorizing households and load profiles based on the following classes: demographics and socio-economic factors, dwelling characteristics, habits (e.g. consumption timing), energy conservation, energy efficiency goals, knowledge about electricity consumption and the attitude of use. Presently, data and detailed measurements for more than tens of thousands

end-users are available and accessible [4].

Over the years, quite a few clustering techniques have been employed to support customer segmentation based on certain key characteristics. From the most well-known k-means [5], and its variations [6], to more sophisticated methods like expectation maximization [7] and spectral clustering [8].

Most research findings focus on consumption-related aspects, without taking into account the emerging prosumer role, that can greatly affect the clustering results, presenting segments with completely different characteristics. On the other hand, the majority of the results presented do not explore the overall DR reliability of the system presented, aiming towards maximizing the consecutive successful delivery of a DR request.

This paper proposes a DR reliability scheme as an evaluation metric for DR from the perspective of each layer: Aggregator, Virtual Node (VN) and individual prosumer. Furthermore, there is an endeavour to formulate VNs consisting of prosumers with similar power flow profile, in order to combine their load, generation and storage capabilities, as all these measurements can potentially affect their DR contribution. Finally, over the clustering results, TDDS enhances the reliability of weak VNs through a balancing mechanism that at the same time respects the clusters' coherence.

The remainder of this paper is structured as follows: Section II presents the methodology followed for the clustering procedure proposed, followed by the actual implementation in Section III. Section V demonstrates experimental results over a specific portfolio, whereas Section VI concludes this manuscript.

II. RELIABILITY EVALUATION OF DISTRIBUTED ENERGY ASSETS

The Reliability concept that is introduced in this paper refers to a layered scheme of evaluation for different entities. In that way, the Aggregator can exploit this information in order to formulate balanced VNs, in terms of their credibility to participate successfully in DR events.

The lurking uncertainty that comes from the small and medium prosumers to deliver DR requests effectively in a distributed energy ecosystem, necessitates the adoption of evaluation metrics as risk quantification tools. In this paper, a reliability scheme is presented that estimates the willingness and effectiveness of an energy asset to deliver DR requests during a period of time. To do so, the concept of DR reliability is introduced, evaluating how reliably can a customer deliver a DR request. In the direction of evaluating this ability in different layers of a decentralized DR architecture, reliability has been arranged in three categories: individual/personal reliability (pr), internal reliability (IR) and external reliability (ER). As the customer and the VN are introduced to the system, they are considered as reliable assets, thus, pr and ER takes the maximum value of 1, while the range of the metrics is enclosed in the [0,1] space. The reliability at the individual customer level is represented through the *pr* metric that actually quantifies the recorded contribution of a user

in corresponding historical DR actions in the past. Equation (eq.1) describes the pr alterations depending on coming DR results. In general, as the customer is unable to deliver a DR request, there is a constant percentage decline in the pr value, while in the opposite scenario, in case of a successful DR participation, pr increases respectively. This process occurs with identical way, at outer level of ER as it is described in equation (eq.3).

$$ pr = \begin{cases} pr_i + alt \cdot pr_i & \text{Customer's success} \\ pr_i - alt \cdot pr_i & \text{Customer's failure} \end{cases} \quad (1) $$

Where pr_i is the individual customer's reliability and alt is a constant percentage of a corresponding alteration (increase or decrease) over the current pr_i value.

The average reliability of a VN, in terms of its assets' *pr* is represented as IR. The following equation (eq. 2) describes the IR of the VN's resources for a P number of cluster participants. A more concise definition of these Reliability metrics is presented in table I.

$$ IR = \sum_{i=0}^{i=P} pr_i / P \quad (2) $$

Where IR is the internal reliability of a cluster, P is the total number of customers, and pr_i is the individual reliability of each customer.

ER represents the outer layer of this architecture and expresses the ability of a Virtual Node (VN) as an undivided asset to deliver DR signals, in terms of the Aggregator's perspective. The hierarchy of these metrics is depicted in Figure 1. It is worth to mention that from the aggregator's perspective, both DR failure and completion does not always conform with the corresponding failure and completion of all the individual assets. In a hypothetical scenario, some individual customers may achieve to complete their DR action, however, the total DR could potentially fail due to other customers' failure and inadequacy to achieve the minimum DR requirements.

$$ ER = \begin{cases} ER + alt \cdot ER & \text{DR success} \\ ER - alt \cdot ER & \text{DR failure} \end{cases} \quad (3) $$

Where ER is the External Reliability and alt is a constant percentage of a corresponding alteration (increase or decrease) over the current ER value.

Reformulating VN's assets, affects directly the IR of the entity, whereas the ER changes will be reflected in long-term periods, as the VN participates in more DR actions. Typical Energy profile clustering approaches that focus on the load measurement, achieve to distinguish customers according to their load behaviour, however this separation is not adequate to formulate balanced VNs with regard to their Reliability metric. As a result, DR strategies that anticipate VNs as individual entities and pursue efficient DR delivery, discriminate over more credible groups and ignore the existence of less reliable assets without reverting factors.
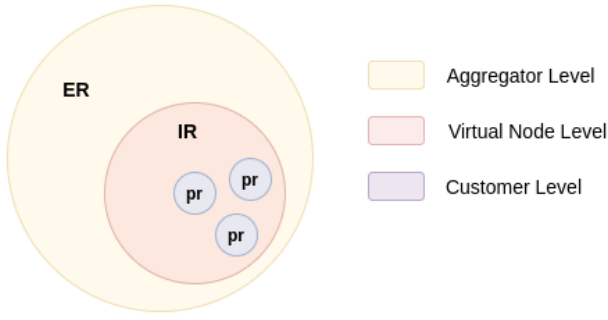
Fig. 1. Reliability namespace per Architecture level.

TABLE I
RELIABILITY METRICS

Types of Reliability	Definition
Internal Reliability (IR)	The average reliability of the VN's assets
External Reliability (ER)	VN's reliability regarding its total contribution to historical DR signals.
Individual Reliability (pr_i)	The reliability of an individual participant i

Fig. 2. Decomposition staging of TDDS algorithm

III. TEMPORAL DATA DYNAMIC SEGMENTATION

This paper proposes a novel methodology to apply clustering over the customers' energy profile, while retaining a balance of the clusters' IR. The Temporal Data Dynamic Segmentation (TDDS) algorithm harnesses the property of a distribution based soft clustering method, in particular the Gaussian Mixture Model with Expectation Maximization Algorithm, to assign each customer in multiple groups with a corresponding probability, in order to apply subsequent correction movements towards the reinforcement of unreliable VNs. In general, GMM's competency to identify clusters and approximate any probability density function with simple gaussian components has found utility in the energy domain [9]–[11]. Compared to other Centroid, Density or Connectivity based algorithms that have been used in the literature for load profile clustering, GMM as a distribution based model approaches the problem from a probabilistic perspective, providing additional information with regard to the strength of association between the data and the corresponding clusters that serves the concept of TDDS.

A fair trade-off between energy profile purity and reliability balance is achieved effectively the exchange pool originates from customers that belong to an intersection space among the clusters. In that way, the purity of the generated energy profiles is not affected from swaps between customers from different clusters, allowing the dynamic creation of, in average, more reliable clusters. The TDDS algorithm comprises of three stages (Fig. 2) which are explained subsequently.

A. TDDS Stages

1) Stage 1: Contains the pre-processing step of the data, the evaluation of the number of clusters optimum and the model design. The pre-processsing step concerns a scaling operation that squeezes the data in a range between [0,1], as well as the application of Princ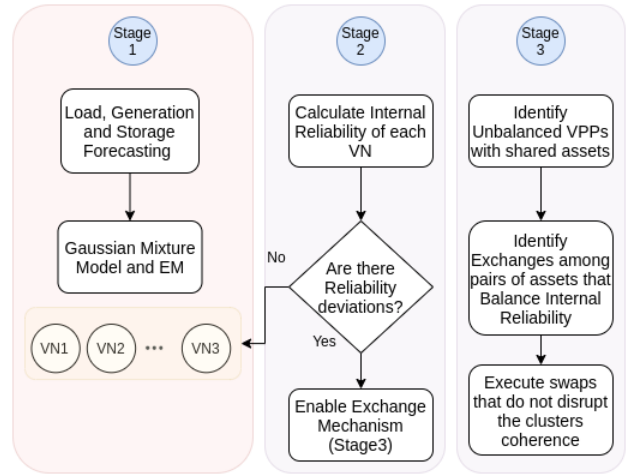ipal Component Analysis (PCA) towards the dimension reduction of our data. Regarding the optimal number of generated clusters, a grid search of the Bayesian Information Criterion (BIC) for multiple candidate clusters is proposed [12], as well as the exploration of the elbow point, that denotes the optimal option regarding the effectiveness and simplicity of the model as density estimator. The final step includes the application of Gaussian Mixture Model (GMM) through the Expectation Maximization (EM) Algorithm and the extraction of the initial clusters (VNs).

2) Stage 2: is a transient layer that concerns the validation of the VNs' reliability balance. TDDS estimates the IR of the generated clusters and examines the case scenario of high IR deviations among the clusters. More specifically, in case that the IR of a specific node deviates from the average IR of the total nodes over a specific margin as it is expressed in equation (eq. 4), an exchange mechanism is triggered and balancing actions are applied, whereas in the opposite case, the already generated VNs' structure is considered as adequate formulation to deliver effectively DR actions.

$$IR_{avg} - IR_{node} > dev \qquad (4)$$

Where, dev is a specific deviation margin, IR_{avg} is the average IR of all nodes and IR_{node} is the IR of the examined node.

3) Stage 3: Describes a balancing mechanism (BM), that exploits the Soft Clustering method that was applied at Stage 1 to identify potential stability candidates. As the segmentation operation focuses primarily on the discovery of prosumers with common energy profiles, hence, balancing operations are inclined towards not disrupting these clusters' identity and coherence. As a result, the BM identifies cluster's points that belong to an intersected area among two groups with high probability and contrary IR, comparing their individual reliability pr_i and applying swaps that equilibrate the IR of the VNs. As it is displayed in Fig. 3 Cluster1 represents a VN with a low IR value, while Cluster2 possesses a high IR value. Additionally, Cluster1 contains a participant point with

low pr_i value and Cluster2 a point with high pr_i. These points belong to an intersection area among the two VNs, as the GMM's results attributes both points, as potential candidates to participate in both VNs with high probability.
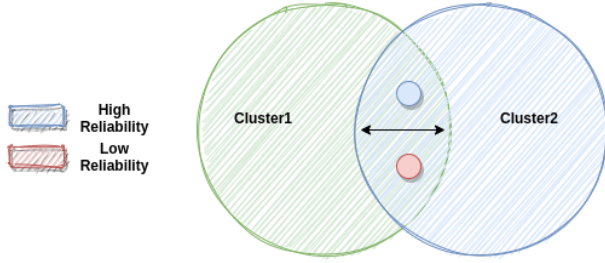

Fig. 3. Balancing Mechanism Representation

B. Gaussian Mixture Models

GMMs are probabilistic models that utilize a Soft Clustering approach to distribute energy profiles among the nodes, indicating at the same time the strength of association between these two entities. They represent a Mixture of several normal distributions with a respective weight π as it is displayed in the equation (eq.5).

$$p(x) = \sum_{k=1}^{k=K} \pi_k N(x|\mu_k, \Sigma_k) \qquad (5)$$

where $N(x|\mu_k, \sigma_k)$ is the univariate Gaussian Distribution for variable x,

$$N(x|\mu_k, \sigma_k) = \frac{1}{\sigma_k \sqrt{2\pi}} \exp(\frac{-(x-\mu_k)^2}{2\sigma_k^2}) \qquad (6)$$

σ_k represents the standard deviation, μ_k the means, π_k the mixing coefficients for each component k with the constraint as it is described in equation (eq.7) and K the number of mixture Components.

$$\sum_{k=1}^{k=K} \pi_k = 1 \qquad (7)$$

When the data have multiple features, the d-dimensional Gaussian Distribution is defined as:

$$N(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2} |\Sigma^{1/2}|} \exp(\frac{-1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)) \qquad (8)$$

where, Σ is the Covariance Matrix that describes the relationship between the features.

Determination of the optimal parameters for GMM is commonly occurred through an iterative algorithm EM [13] that is used to identify the maximum likelihood estimation with the presence of latent variables. An iterative sequence of Expectation and Maximization steps is applied until the convergence of the algorithm. The estimation of the latent variables is achieved through the Expectation step, while Maximization step is responsible to optimize the models' parameters. Finally, the parameters' initialization of EM algorithm could highly impact the quality of the solution [14].

IV. EXPERIMENTAL SETUP

A. Dataset

The Energy Profile Clustering introduced in this manuscript is examined over the calculated power flow of 81 households consisted exclusively by low/medium prosumers for different periods of time. An overview of their general profile is reflected through their nominal capacities: the portfolio has an average consumption capacity of 9kW, around 20% of the portfolio has generation capacity of 10kW, while approximately 30% of the portfolio has storage capacity which ranges between 3kWh and 7kWh. In order to calculate the overall power flow from and to the grid, of each customer, consumption, generation, and storage measurements are combined according to the following equation (eq.9). Data are managed as univariate time series data.

$$pf = load - generation \pm ess \qquad (9)$$

where, pf is the overall power flow, $load$ refers to the power attributed to load consumption, $generation$ refers to power attributed to energy generation, and ess refers to power attributed to energy storage systems, while the sign refers to the state of the battery (charge, discharge).

The DR reliability pr of each household is considered by taking into account historical information of DR signals, that have been made under a simulated environment after approximately one year of DR requests. This presents the necessary heterogeneity of reliability indexes to support the evaluation of the presented model, covering a reliability range from 0.23 up to 0.98.

B. Scenarios

Different periods of time have been selected towards the replication of these conditions that point out two possible use cases (UC). Specifically, the examined scenarios are: UC1: generated clusters contain households with intersected energy profile, thus, swaps among the nodes are feasible, and UC2: there are no possible households swaps that could balance virtual nodes' IR despite their energy profile similarity.

The following table (Table II) displays the Nodes' IR over three time periods, that reflect the two use cases. Red background colour expresses the nodes' weakness in terms of the IR metric and the need for balancing actions in case that there are available resources to satisfy the aforementioned criteria. Such weakness, could be attributed to multiple factors, one of most common being the fact that most clustering methods in the literature do not consider DR reliability but only energy-related measurements.

V. Results and Discussion

Experiments and results are inclined towards the proof of validity for clustering process in conjunction with the balancing mechanism's activation. The possible scenarios that can be anticipated from TDDS have been categorized with regard to the clustering results and the balancing mechanism's decision making. In the conducted experiments, different parameters' initialization strategies were examined: Agglomerative Hierarchical Clustering, random starting values and KMeans, while the latest was finally selected, as it showed the most promising results. Regarding the type of the covariance matrix that dictates the spread and the orientation of each distribution, diagonal covariance was selected instead of spherical or full type, in order to avoid overfitting issues.

TABLE II
Nodes' IR in different scenarios after stage2 and before stage3

IR	Node0	Node1	Node2	Node3	Node4	Node5
UC1-Period1	0.85	0.86	0.51	0.58	0.92	0.74
UC1-Period2	0.87	0.89	0.55	-	-	-
UC2-Period3	0.84	0.46	0.91	0.87	0.79	-

1) Use Case 1 - Period 1: In this period, the analysis of BIC metric for a range of clusters, indicated the formulation of six clusters as optimal number for more efficient clusters separation. The IR of Node2 and Node3 can be considered low reliability (LR) assets with 0.51 and 0.58 IR respectively as they deviate from the average IR of the portfolio according to (eq.4). Thus, TDDS is responsible to identify potential households' swaps between LR and HR nodes' assets that will mitigate reliability issues. Regarding the energy profile of each nodes' portfolio, Fig. 4 and Fig. 5 provide an intuitive representation of the clusters' position in a two dimensional space and a heatmap that displays each profile's strength of association with existing nodes. Thus, through the heatmap, it is possible to discern profiles that belong to the intersection zone between two nodes, as they are illustrated with intermediary colours (i.e. light green). In this scenario, the balancing mechanism over the clustering results applied four assets swaps (three for Node2 and one for Node3). These exchanges between the nodes affected the silhouette score [15] of the clusters, reducing its value from 0.43 to 0.42, while at the same time the value of the node with the lowest IR increased from 0.51 to 0.6.

2) Use Case 1 - Period 2: The minimum value of BIC metric indicated the formulation of three clusters. Node2 is the only node with LR that needs support in terms of reliability. From Fig. 4 and Fig. 5 it is discernible that Node1 is the only Node that contains four households in the intersection area with Node2, while Node0 assets do not share common energy profile with other nodes. In Stage 3 (see Section A.3), the algorithm endeavours to identify pairs of households in the intersection zone with opposite IR in favour of the weak node. Since, the are four pairs that conform to the criteria, the balancing mechanism is activated. The results of these

TABLE III
Clustering results and IR metrics for 3 Periods of time.

TDDS Results	Period1	Period2	Period3
Clustering Results without BM	0.43	0.57	0.39
Lowest IR without BM	0.51	0.56	0.54
Clustering Results with BM	0.42	0.55	-
Lowest IR with BM	0.60	0.65	-
BM	Y	Y	N
Increase IR	17.6%	16%	-

exchanges reduced the silhouette score from 0.53 to 0.52 while the IR of the weakest node increased from 0.56 to 0.65 (i.e. increase by 16%).

3) Use Case 2 - Period 3: In this period, there are 5 clusters as they are displayed in Fig. 4 and Fig. 5 and Node1 can be characterized as LR asset according to table (Table II). In stage 2, the balancing mechanism activated and identified two nodes with similar energy profile. However, the pairs of households that belong to the intersection area cannot contribute to the empowerment of LR node as they have similar IR values. This scenario, presents a use case that the balancing mechanism cannot be applied on top of the clustering results.

VI. Conclusion

To conclude, the incorporation of low-medium customers in DR programs, necessitates the adoption of ancillary services that mitigate the absence of data in real life scenarios and hedge against the risks of unpredictable responses. This paper proposes a multilevel reliability scheme in order to manage and formulate credible VNs. On top of this scheme, TDDS is an algorithm based on the soft clustering technique that aims to identify groups of assets with common energy profile, while at the same time there is a balancing mechanism that is triggered to empower unreliable nodes. The results of this algorithm were examined over three separate periods that contained at least one unreliable node. In Period1 and Period2 the algorithm achieved to extract clusters and apply balancing actions on top of these groups, increasing the IR of the weakest node by 17.6% and 16% respectively (Table V-3), without disrupting the clustering coherence, as the silhouette scored had a minor decrease. On the contrary, in Period3, the balancing mechanism could not identify pairs of households that could participate in balancing actions because of the lack of assets with opposite IR value inside the intersection zone.

References

[1] J. F. Martins, A. G. Pronto, V. Delgado-Gomes, and M. Sanduleac, "Smart meters and advanced metering infrastructure," in *Pathways to a Smarter Power System*. Elsevier, 2019, pp. 89–114.

[2] Q. Gong, Y. Jin, Y. Han, C. Sang, R. Wang, and B. Zhou, "Abnormal electricity customer clustering method based on electricity big data," in *2019 IEEE Sustainable Power and Energy Conference (iSPEC)*. IEEE, 2019, pp. 2221–2226.
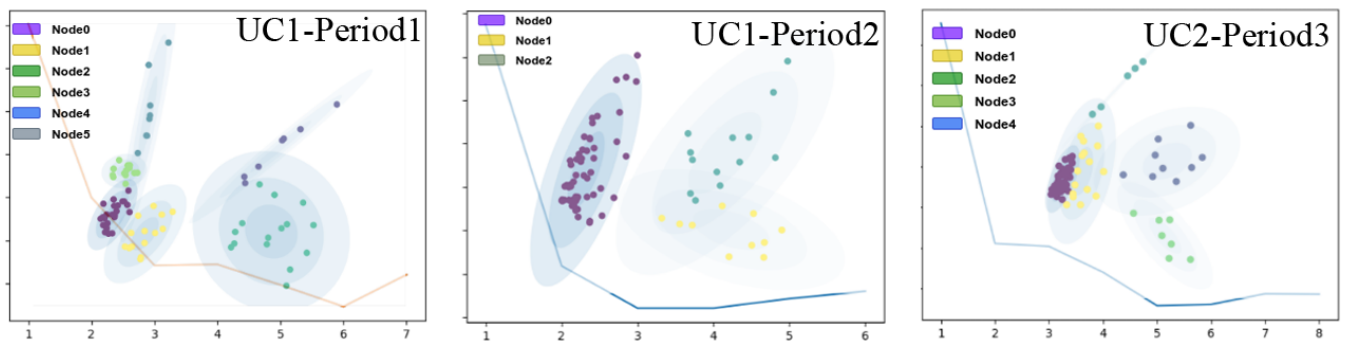
Fig. 4. Clusters' representation in two dimensional space combined with the BIC values for several number of clusters illustrated in a curve for the two UC and the three Periods.
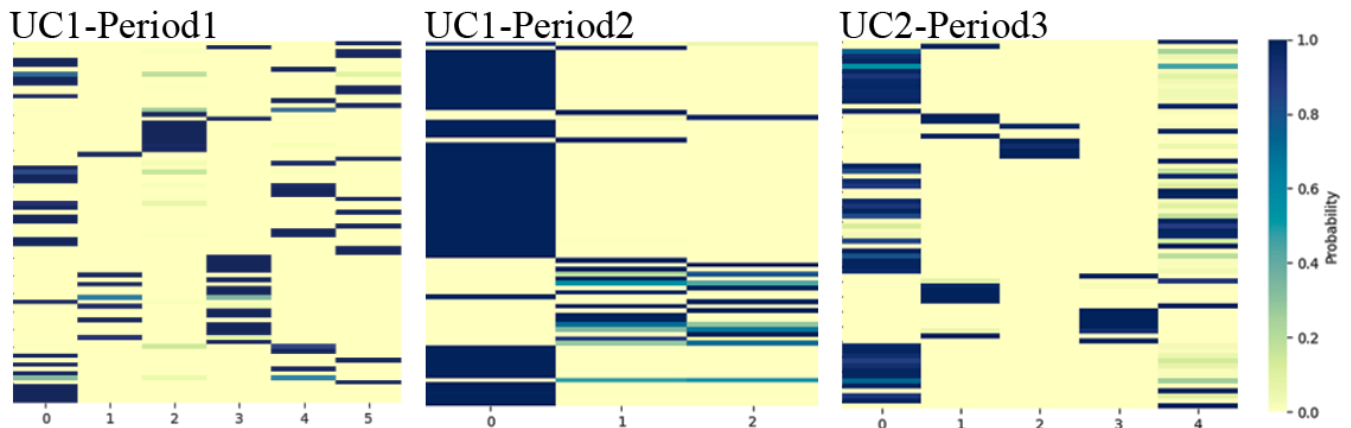

Fig. 5. Heatmap representation that illustrates the probability of each household to belong to each node according to GMM results for the two UCs and the three Periods.

[3] Y. Zhang, T. Huang, and E. Bompard, "Big data analytics in smart grids: a review," *Energy Informatics*, vol. 1, 08 2018.

[4] A. Rajabi, M. Eskandari, M. J. Ghadi, L. Li, J. Zhang, and P. Siano, "A comparative study of clustering techniques for electrical load pattern segmentation," *Renewable and Sustainable Energy Reviews*, vol. 120, p. 109628, 2020.

[5] P. Nystrup, H. Madsen, E. M. Blomgren, and G. de Zotti, "Clustering commercial and industrial load patterns for long-term energy planning," *Smart Energy*, vol. 2, p. 100010, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666955221000101

[6] S. Cembranel, F. Lezama, J. Soares, S. Ramos, A. Gomes, and Z. Vale, "A short review on data mining techniques for electricity customers characterization," 03 2019, pp. 194–199.

[7] S. Haben, C. Singleton, and P. Grindrod, "Analysis and clustering of residential customers energy behavioral demand using smart meter data," *IEEE transactions on smart grid*, vol. 7, no. 1, pp. 136–144, 2015.

[8] B. Auder, J. Cugliari, Y. Goude, and J.-M. Poggi, "Scalable clustering of individual electrical curves for profiling and bottom-up forecasting," *Energies*, vol. 11, no. 7, p. 1893, 2018.

[9] Y. Zhang, Q. Ai, and Z. Li, "Improving aggregated baseline load estimation by gaussian mixture model," *Energy Reports*, vol. 6, pp. 1221–1225, 2020, 2020 The 7th International Conference on Power and Energy Systems Engineering. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352484720314761

[10] Y. Huang, Q. Xu, X. Jiang, Y. Yang, and G. Lin, "An analytic approach to probabilistic load flow incorporating correlation between non-gaussian random variables," *Elektronika ir Elektrotechnika*, vol. 24, 06 2018.

[11] J. Xu, X. Kang, Z. Chen, D. Yan, S. Guo, Y. Jin, T. Hao, and R. Jia, "Clustering-based probability distribution model for monthly residential building electricity consumption analysis," *Building Simulation*, vol. 14, 09 2020.

[12] E. Grall-Maes and D. T. Dao, "Assessing the number of clusters in a mixture model with side-information," in *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods*, ser. ICPRAM 2016. Setubal, PRT: SCITEPRESS - Science and Technology Publications, Lda, 2016, p. 41–47. [Online]. Available: https://doi.org/10.5220/0005682000410047

[13] J. P. Vila and P. Schniter, "Expectation-maximization gaussian-mixture approximate message passing," *IEEE Transactions on Signal Processing*, vol. 61, no. 19, pp. 4658–4672, 2013.

[14] E. Shireman, D. Steinley, and M. Brusco, "Examining the effect of initialization strategies on the performance of gaussian mixture modeling," *Behavior Research Methods*, vol. 49, 12 2015.

[15] L. Jin, D. Lee, A. Sim, S. Borgeson, K. Wu, C. A. Spurlock, and A. Todd, "Comparison of clustering techniques for residential energy behavior using smart meter data," in *AAAI Workshops*, 2017.