Cluster-based Energy Load Profiling on Residential Smart Buildings

Alkiviadis Savvopoulos*†, George Kalogeras*, Christos Anagnostopoulos*, Christos Alexakos*, Spyros Sioutas† and Athanasios P. Kalogeras*

*Industrial Systems Institute
ATHENA R.C.
Patras, Greece
savvopoulos@isi.gr, gkalogeras@isi.gr,
anagnostopoulos@isi.gr, alexakos@isi.gr,
kalogeras@isi.gr

†Dept. Computer Engineering and Informatics
University of Patras
Patra, Greece
sioutas@ceid.upatras.gr

***Abstract*—Percentage of population living in cities is expected to reach 60% by 2030, accounting for 60% - 80% of world annual energy needs and making the impact of energy efficient solutions in cities quite significant for environmental protection and fighting climate change. The building sector uses about 40% of European energy and emits approximately 1/3 of greenhouse gas emissions. Black-box measurement based modeling methods allow the estimation of consumption in buildings relying on smart metering devices installed. Vast amount of data generated poses new challenges with reference to their handling and timely processing. The paper presents an approach related to building energy load profiling utilising profile compression and clustering. It discusses the application of different clustering algorithms through their experimental evaluation.**

***Index Terms*—Smart Cities, Smart Energy, Smart Building, Load Energy Profiling**

I. Introduction

The steady increase of worldwide energy consumption imposes several problems in terms of limited energy resources, supply difficulties and environmental impacts. A common challenge is to reduce energy consumption, mitigating at the same time impacts on climate change [1]. The building sector is responsible for approximately 40% of European energy consumption, and 36% of CO_2 emissions. Buildings are therefore the single largest energy consumer in Europe [2]. Residential buildings account for about 27% of EU energy consumption [3]. With these numbers in mind, it is evident that building energy efficiency is extremely important in terms of both the economic and environmental positive effects it induces.

Solutions to the aforementioned challenge can be distinct according to their fields of application: a) technologies for energy efficient building construction or renovation, and b) technologies that can be easily applied to existing infrastructures. Especially with reference to the latter, use of energy profiling is a powerful tool for estimating energy consumption, based on specific static factors (building orientation, occupancy, environmental conditions, etc.). Nevertheless, such profiles are not based on real time data, and are not flexible to changes in such factors as climate changes, building or apartment renovations, and new application of energy technologies.

Nowadays, there is a strong orientation of energy providers towards the utilisation of smart energy networks. Smart energy networks, including smart grids, smart district heating networks, and smart natural gas networks, integrate new control and sensor equipment to the traditional energy transmission network, allowing the collection of data, its processing and related automation in the control of the energy flow. With currently existing technology, it is easy to install smart meters and actuators in any residential building, either in the central energy supply of the building/department, or even inside the residence (smart plugs, smart light bulbs, etc.). Access to accurate data about energy consumption provides the ability of applying data science techniques for the identification of the behavioural profile related to energy consumption of residential buildings. The capacity of continuous monitoring led researchers to design algorithms that can dynamically provide energy profiles using modern methods based on data mining, machine learning and advanced statistics. Outcomes may be applied in various areas in the intelligent energy network domain, such as demand forecasting, energy generation optimization, energy pricing, monitoring and diagnostics [4].

The field of dynamic energy profiling, although witnessing helpful results from the utilisation of conventional clustering methods, experiences new challenges considering the introduction of new technologies in complex energy management systems. The major barrier is the vast amount of data, both real-time and historical, that is generated by smart metering devices and which has to be processed quickly. Such requirements lead the research to new methodologies, like data dimension

We acknowledge support of this work by the project "Enabling Smarter City in the MED area through Networking- ESMARTCITY" (3MED171.1M2022), which is implemented under the "Interreg Mediterranean" programme, co-financed by the European Regional Development Fund(ERDF), Instrument for Pre-Accession Assistance (IPA) andnational sources.

reduction, and timeseries processing.

The motivation underlying in this work resolves around the management of significant energy consumption imbalances that occur between the maximum green energy production time span and the peak energy demand also known as the "Duck Curve" [5]. The proposed approach identifies in energy consumption data, such factors as load demand statistics, peak demand, and energy fluctuations. This space efficient proposed approach is proven via appropriate experimental evaluation, that is able to adequately identify clusters, which peak demand timespan correlates with the production of "green" and renewable energy.

The paper is structured as follows. Section II presents briefly the benefits, challenges, and approaches of energy load profiling. Section III describes the first step of the proposed methodology, that deals with the compression of energy consumption profiles, while Section IV presents the clustering algorithm, that is used for the identification of load profiles. The evaluation is based on experimentation using daily measurements of energy consumption taken from residential buildings, and presented in Section V. Finally, conclusions and future work are presented in Section VI.

II. Residential Building Energy Load

A. Building Energy Efficiency

An energy efficient building is described as having healthy facilities designed and built in a resource efficient manner, using ecologically based principles [6]. This means that a building should have a minimal positive impact on the environment through achieving an as low as possible energy consumption, while maintaining the desirable living and working conditions for its tenants.

Technological advances in recent years have provided the building sector with new options to achieve a higher level of energy efficiency. It must be noted however that although most intervention options are considered as economically viable, the rate at which investments are being made remains low in many countries. This is attributed to socio-economic reasons, namely recent economic crises, and low level of acknowledgement of the current situation by potential investors and the general public. This phenomenon is described as the energy efficiency gap [7]. Below we try to briefly describe the current state of the art in terms of available technologies for energy efficiency interventions.

A huge part of the building energy consumption is used for controlling its thermal comfort, meaning the operation of the Heating, Cooling and Ventilation systems (HVAC). This can be as high as 76% of the total building consumption according to [8]. Solutions to reduce this consumption include the use of new insulation materials for shielding the building envelope, thus reducing the energy demand, as well as the HVAC distribution system, or the installation of new HVAC systems, such as heat pumps or condensing boilers, which increase the overall efficiency of the system.

A reduction in the electrical consumption of the building is also considered as a good practice for raising its energy efficiency. Renewable energy sources, such as solar and photovoltaic panels, can hugely contribute towards this end, while modern lighting systems and household appliances with better energy ratings are also obvious choices as interventions.

Smart buildings driven by technological advances in the Internet of Things sector are also emerging as technological solutions for increasing energy efficiency in buildings. Smart devices such as sensors can be utilized to monitor human behavior, control building systems to adapt to that behavior and even counter-measure some human actions with adverse effects on energy consumption. When paired with newly developed advanced control algorithms that allow building simulations and energy consumption prediction, smart building interventions represent a highly efficient and economically viable set of interventions focused on building energy management.

Human behavior monitoring and analysis is critical both for finding the best possible intervention to reduce energy consumption and for evaluating the implemented retrofit. This is because human behavior plays a vast role in the energy performance of buildings and creates considerable variations in their efficiency [9]. In fact, it is common for building users to alter their energy consuming habits after a retrofit, even to an extent that partially cancels the gains from the intervention, an effect known as "backfire effect" [10]. This shows the significance of energy profiling to correctly display and predict user behavior for a particular case.

B. Standard Energy Profiles

Energy profiling for buildings is a very useful tool for pointing out possible waste of energy that is attributed to occupancy related actions. Occupant presence and behavior in buildings has shown to have large impacts on space heating, cooling and ventilation demand, energy consumption of lighting and space appliances, and building controls where careless behavior can add one-third to a building's designed energy performance, while conservation behavior can save a third [11]. This underlines the importance of a carefully drafted energy profile that accurately depicts the building energy behavior.

Profiles are constructed using interval data from the building consumption. More data available means a higher accuracy for the generated profile. By comparing the profile with the building occupancy patterns while taking into account the building particular usage, important details can emerge that show discrepancies between the expected and actual energy performance of the building. Another usage of energy profiles is the possibility to correlate them with the building energy demand pattern, which can be obtained through building simulation software that implements well established international standards [12] [13] [14] [15].

Energy profiles for residential buildings can vary greatly for a number of reasons. Number of occupants, behavior patterns and daily schedules can affect the building energy profile. This is particularly true for peaks in electricity demand, a factor that is highly correlated with the presence of users in the building.

Moreover, since, depending on the country, HVAC needs for residential buildings can be mostly relevant during heating or cooling seasons or both, profiles have a large dependence on seasonality. A general norm shows much larger consumption during winter. Finally, residential building profiles are different during weekends and non-working days, since occupants are usually inside the building for a larger portion of the day compared to working days.

Several surveys have been conducted to compare profiles of buildings in different countries and with different usage patterns. In [16] several domestic building profiles are presented and broken down into different appliance type usage during the day. In [17] profiles from residential buildings from different countries are presented, showing the consumption pattern when it comes to domestic hot water usage and electrical loads.

C. Electrical Energy Load Dynamic Profiling

The static energy profiling provides a good estimation of the probable energy consumption for residential buildings, but is rather inadequate for real time energy management. The evolution of intelligent energy networks provides the means for accessing real time energy consumption data, as well as for the identification of accurate energy profiles. Data clustering is usually deployed for creating dynamic energy profiles. Methods like Support vector machine (SVM), Neural networks, K-Means, Gaussian process, Principal component analysis (PCA), Independent component analysis (ICA), Non-negative matrix factorization (NMF) and Learning vector quantization (LVQ), are applied to the collected data, finding consumers with similar behaviours [4]. Al-Wakeel & Wu [18] utilise the K-Means method to identify energy load clusters based on daily customer energy consumption data. They conclude that the minimum clustering ratio is highly correlated with the shortest time window of segmentation, and that small number of clusters can provide more accurate results. A comparative study [19] on the clustering methods for energy profiling reveales that hierarchical algorithms provide better results in data containing daily energy load of a customer, but in terms of execution time they lack against k-means approaches with few clusters. Moreover, the same study concludes that new tools are required to deal with the ”big data” thread, which emerges more evidently as the intelligent monitoring infrastructure grows larger.

Furthermore, some researchers apply hybrid approaches that combine additional data from external resources. In [20], the authors examine energy consumption data in households with the combination of door-to-door surveys related to occupant behaviour. Then, they apply a hierarchical clustering method concluding the three major groups of characteristics that describe each cluster.

The challenge today is the handling of the amount of data that continuously increases. Classical algorithms fail to provide reliable solutions, so new approaches are employed in an effort to tackle this problem. A multi-layered approach is proposed in [21], where two layers of profiling are used: a) a local profile derived from clustering data in a small and specific area (apartment, building or block) and b) a global profile based on the findings of the first layer. This approach ensures an essential reduction in the computational complexity. Ray & Pinson [22] create an adaptive and recursive clustering method based on on-line k-means in order to continuously evaluate and adjust their predictive model with the later measurements. In another research effort [23] related to energy load profiling for demand response (DR) programs, the authors use a modified symbolic aggregate approximation (SAX), which among other benefits, reduces the dimensionality of data, increasing the performance of the clustering algorithms that are used. Finally, a most recent approach [24] uses image processing techniques to suppress excessive sensitivity. Specifically, the authors create load profiles with conventional methods and depict them to two-dimensional images. Then, they apply filtering and thresholding techniques for optimizing the results.

III. Compressing Energy Consumption Profiles

The ultimate goal of the algorithm at hand, is to present a scalable method of handling time series, that represent an energy consumption profile across a period of time, and comprehend its properties in regards to load demand statistics, peak demand, as well as energy fluctuations across a specific time span.

Thus, this work demonstrates a methodology composed by a set of steps, utilizing state of the art tools, in order to achieve the aforementioned. Load profiles, in their essence, are time series, that measure energy-based expenditure. Their span as well as their rate of sampling is ever increasing in terms of velocity and frequency, resulting into vast amounts of data which are difficult to process in their entirety or save into storage and access them efficiently.

It is essential that the scalability of the algorithmic procedure must be ensured. The essential approach for this step, is to develop means of compressing time series. A general challenge for Internet of Things related datasets, is their sheer volume. Handling time series that pertain days or even months, for example with a sampling rate of 2 or even 4 records per minute, could lead to extended processing times, slower and inefficient systems. Compression introduces the concept of loss during the reconstruction, however the utmost ambition of a compressing procedure is to achieve a beneficial trade-off between the error of the algorithm, and the amount of data condensation it bestows to the system.

A. Compression through Auto-Encoders

A compression algorithm, that takes advantage of state of the art solutions such as neural networks, having proven their robustness and efficiency in handling patterns in vast amounts of data, is the Variational Autoencoders, which belongs to the family of unsupervised learning [25] [26].

Variational Autoencoders are fundamentally foundationless of compression. They extend the classic Autoencoders, which consist of two modules: one process that encodes the incoming data into a latent feature space whose size is significantly smaller than the dimensions of the input data, and then another process which consists of a decoding module that extracts the compressed data from the median neural layer, and projects it into a dimensional space identical to the input size [27]. The distinctive feature that separates Variational Autoencoders from classic Autoencoders, is the reparameterization trick that occurs to the data record when it reaches the median layer, in order to better rearrange the latent space and avoid cases of overfitting during the training phase of the algorithm [28].

Thus, the loss function of such an algorithm is conceived through two parts. Initially, since Variational Autoencoders pertain to a reconstruction process, the input data has to be as similar as the output, hence the first part of this loss function consists of this difference between input and output values. The second part of the function, is inevitably introduced through the reparameterization trick. Since the algorithm aims at learning a distribution in the latent space, the difference between the learned distribution and the standard normal distribution represents another part of loss and is best expressed by the Kullback–Leibler divergence [29].

$$L = -D_{KL}(q(z|x^i)||p(z)) + E_{q(z|x^i)}[logp_\theta(x^i|z)] \quad (1)$$

where z represents the latent layer, x^i is the input datapoint and q is the learnt distribution which is compared to distribution p, where θ denotes the characteristic properties of p.

However, due to the fact that the energy consumption profiles are measured as a function of time, an alternative neural network approach for the encoding and decoding modules is considered. More precisely, the introduction of recurrent neural networks (RNNs) in the autoencoding of timeseries data can be deemed very helpful [30]. RNNs can detect patterns of non-coherent timespan in a more accurate manner than classic feedforward neural networks, as well as handle non-stationarity, a characteristic which is often met in timeseries datapoints.

Furthermore, RNNs introduce a flexibility in the size of the input, since they do not demand timeseries of the same dimensions. The training phase also meets an advantage, since RNNs outperform the drawbacks of linearities during the learning process.

B. Auto-Encoding Process Implementation

For the implementation of the aforementioned autoencoding pipeline, a number of neural network schemas were deployed, with different hyperparameter values each time the experiment was reproduced, so as to finetune the process and determine which values lead to a more efficient training of the autoencoding algorithm.

Firstly, the initial neural network deployed was a classic autoencoder with zero additional fine-tuning logic attached to its schema. As a consequence of the data used in this work which maintained a half-hourly sampling rate of residential energy consumption, the outer layers of the autoencoder were of size 48. The inner structure of the encoding and decoding modules were identical and symmetrical, with a hidden layer of size 24 leading to and from a middle layer of latent space that amounted to 5 dimensions, resulting to a total layer size of (48)⟶(24)⟶(5)⟶(24)⟶(48). The decision for the layer sizes was adjudged heuristically. The experiments, as shown in the following sections, were repeated multiple times with various layer sizes for the hidden and middle layers, in order to deduce the size that presents the most beneficial loss; in this case, being 24 and 5 accordingly.

The schema of the Variational Autoencoder preserved a very similar structure, only differentiating in its fundamental action that requires a double latent layer in order to implement the reparameterization trick.

Regarding the rest of the hyperparameters of the training process, the optimization algorithm selected was the RMSProp based *Adam*, which calculates the gradient's exponential moving average as well as parameters that manage the decay rate of this moving average [31]. For the learning rate, the initial value amounted to $5 - e4$, with the adoption of an adaptive learning approach in order to push the training process out of potential local minima sinkholes.

For the adaptation of the recurrent neural networks inside the autoencoding schema, the encoder and decoder modules had to be altered to a degree. Instead of the classic feed-forward neural networks, the encoding and decoding processes are handled by a specific type of recurrent neural networks called Long-Short Term Memory Neural Networks (LSTMs) [32]. The pipeline of this schema uses also datapoints of the same dimensions, although the hidden layer that is passed to the latent stage is the last layer of the LSTM stack. The encoding to the latent space is still implemented through a linear feed-forward process, and outputs another hidden layer which feeds the decoding LSTM, before the outer dense layer containing the initial 48 dimensions, as also depicted in Figure 1.

The loss function utilized for all the schemes is a Binary Cross Entropy function with logits introduced through a sigmoid activation function [33].

IV. Clustering for Load Energy Profiling

With the conclusion of the encoding step, the algorithmic pipeline presents the time series data compressed to a much smaller feature space, whereas the velocity of processing their characteristics, clustering them or analyzing them in general, significantly increases. The next step of the procedure consists of dividing the available encoded load profiles into separate groups (clusters) which according to the clustering hypothesis, will possess similar features. More specifically, any load profile that is assigned to a cluster must be able to roughly describe the characteristic behaviour of the rest of the cluster's time series.

A. Classic Clustering Algorithms

The first attempt in clustering the encoded time series took place with the exploration of classic clustering solutions such as K-Means or density based clustering algorithms, in order

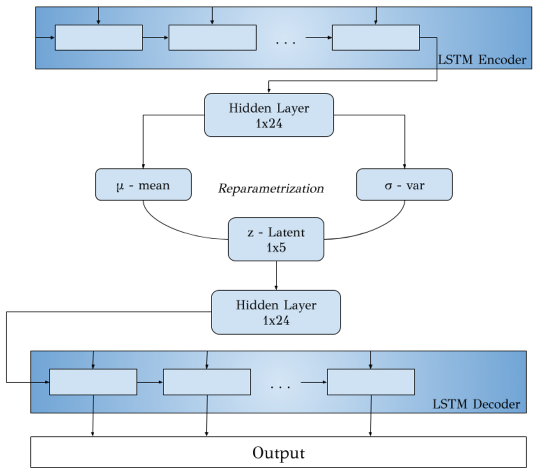

Fig. 1. Variational Recurrent Autoencoder

to pinpoint potential problems that they present, or possibly highlight insightful information.

K-Means was preferred as an algorithm that can quickly produce a rough abstraction on how the data behaves on several different values for the clustering hyperparameters. The clustering was repeated multiple times, each time picking a different value for the number of clusters K, as well as the initial centroids. After each repetition, an analysis was performed regarding the assignment of time series with similar hours of peak demand of electricity [34].

However, predetermining the number of clusters naively as it happens with the K-Means algorithm, without any information that could better dictate the procedure of dividing the input data in clusters, could be characterized as insufficient. The optimal number of clusters to divide the load profiles into, cannot be known $a\ priori$. Thus, the need for a clustering algorithm that does not demand a specific number of clusters emerges in advance.

This problem is efficiently solved with a density based algorithm such as DBSCAN or Spectral Clustering [35]. With DBSCAN, the hyperparameters change, and instead of being the number of cluster centers, the algorithm takes as parameters a) the maximum distance between two samples for one to be considered in the neighborhood of the other, also called $epsilon$, and b) the minimum number of samples in a neighborhood for a point to be considered as a core point ($minPts$), including the point itself [36]. Similarly, Spectral Clustering is a generalized version of DBSCAN, which utilizes eigenvalues for the dimensionality reduction. The usage of such algorithms in timeseries is also highlighted in the work of Jiang et al. [37], where similar challenges are tackled with density-based techniques.

The result of such a procedure, will be a set of clusters that do not have strictly similar silhouettes. In contrast, each cluster will expand accordingly in a more sensible way with the corresponding data points that are in proximity. This process highly depends on the finetuning of the hyperaparameters, and could either result in many clusters, or even in one single cluster in specific occasions, where the datapoints are exceptionally dense.

B. *Predetermined Cluster Representatives*

An alternative to the aforementioned clustering methods, which are presented as part of the contribution of this work, is to carefully predetermine the number of clusters, which the encoded time series will be assigned to, as well as a representative feature vector, that will be the reference point for the assignment of a datapoint to a cluster.

More specifically, the definition of the cluster representatives takes place while taking into consideration the algorithm's fundamental objective, which is to efficiently manage energy's peak demand in residential smart buildings. Thus, ideally, there should be as many clusters created as the number of time zones that best describe advantageous periods of time, additionally to the timespans whereas the peak demand resides in non-optimal places.

The foremost time period that calls for a well defined cluster representative is the period, which consists of the late morning hours, noon hours, until late afternoon, when the percentage of energy that is generated through renewable sources, such as solar energy, is high. In the contest of this paper analysis, other renewable energy sources, such as wind energy or geothermal energy are not being taken into account. Considering further renewable energy sources might add further cluster representatives in our approach.

Furthermore, after subtracting the aforementioned period of high given green energy, there are two periods left within the 24-hour span of the day: the one that begins at midnight, when the day begins until the beginning of the high green energy timespan, and from the end of the high green energy timespan until the end of the corresponding day. Statistically, the vast majority of residential buildings have traditional behaviours during these hours. During the hours after midnight, the energy consumption as well as the energy demand remains low, as a consequence of their sleep schedule. Demand side management scenarios are not taken into account in our approach; yet this does not affect its genericity, as in the case of their presence the time interval referred to in the next paragraph would increase.

The last period of the day is the most ambiguous, and is the one that presents the most load balancing issues. The foundation of these problems, lies in the following events. Firstly, the sheer load demand upright increases during the beginning of this period, due to the fact that most inhabitants of residential building have fulfilled their daily work and have time to invest in household activities. Moreover, the beginning of this period signifies the end of the daily generation of green energy, thus this increase in the energy demand is completely covered by non-renewable sources of energy.

C. *Custom Clustering Representatives Implementation*

The fundamental idea behind the custom cluster implementation heavily relied on the manner of creating the representa-

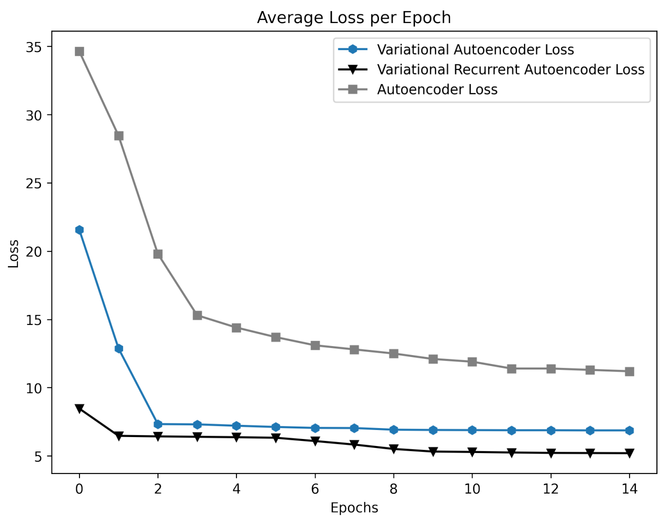

Fig. 2. Training Evolution of selected Neural Network Schemas

tives for each cluster. As mentioned in the previous sections, the constitutional timespans within one day can be broken down to the period during which the renewable energy is plentifully available in the power grid, and the second period being the rest of the day hours, from late afternoon till the next dawn.

V. Experimental Evaluation

A. Experiments and Results

B. Dataset

Taking into account statistics regarding the daily solar radiation, the first cluster profile is modelled accordingly [38]. Firstly, the median value of the dataset's maximum registered energy is extracted. A single value is needed that will serve as the maximum of the custom cluster's profile peak demand absolute value. The index of the peak demand will be correlated with noon time, and the values for the rest of the profile will be a percentage of the maximum value with an exponential decay, spanning between 7:00 am and 18:00 pm. The rest of the values will be set to zero.

The second profile will be the complementary of the first, with the same max values but in different time zones for the peak demand, this time during the late afternoon and the hour around dawn. Afterwards, these vector profiles will be folded together with the rest of the dataset in order to be properly and similarly scaled.

After extracting the two scaled profiles, and having finished with the encoding process, the profiles must be encoded as well, in order to have the same dimensions with the input data points. By having both profiles and the list of time series in a comparable form, the next step of the algorithm pertains a similarity measure between each timeseries and the two profiles, in order to determine its class. The function that was utilized was the Mean Square Error, meaning that the comparison with the smaller value decides whether the

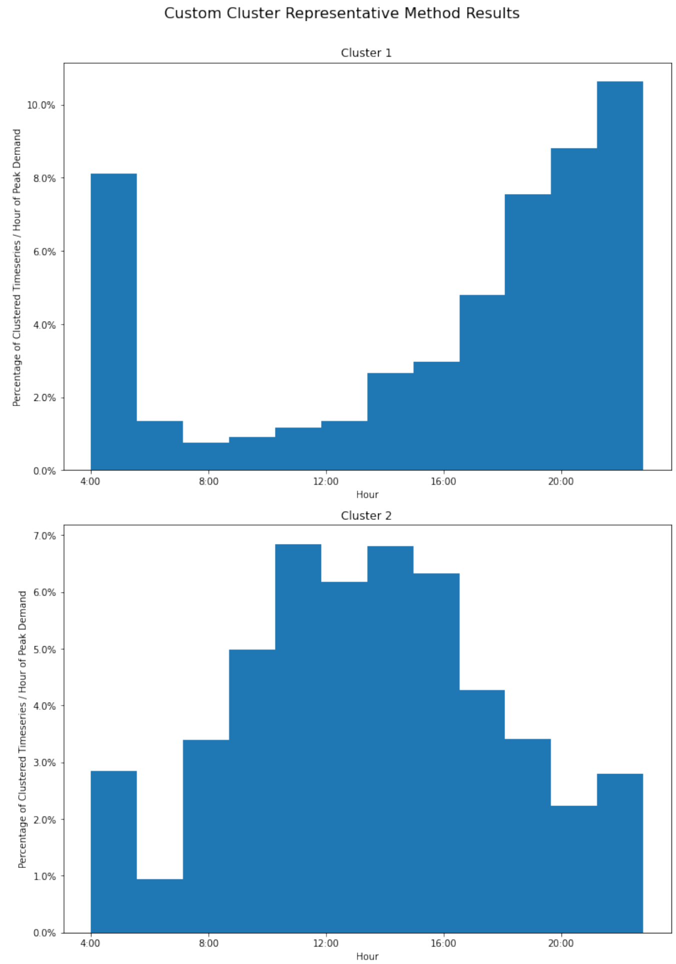

Fig. 3. Custom Cluster Representative Method Results

timeseries' peak demand belongs to the renewable energy compliant cluster or to the second one.

Finally, after concluding the extraction of the classes for each timeseries, the encoded timeseries can be pipelined into a feed-forward neural network with 2 hidden layers of 100 and 10 nodes each. The outer layer that pertains to the input must have the same dimensions as the latent space, since the encoded timeseries are directly fed to this classifier, while the outer layer of the neural network maintains two nodes, one for each class of the aforementioned representatives. Between each layer there are interset ReLU activation functions, a dropout hyperparameter set to 0.2, as well as batch normalization layers [39] [40] [41].

The data that was utilized in the current work, represents daily measurements of energy consumption taken from residential buildings in London, whereas the dataset contains unique identifiers for each residence, the specific date and time of the measurement as well as consumption in kilowatt-hour (kWh) [1]. The sampling rate of the available data amounted

[1] https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households

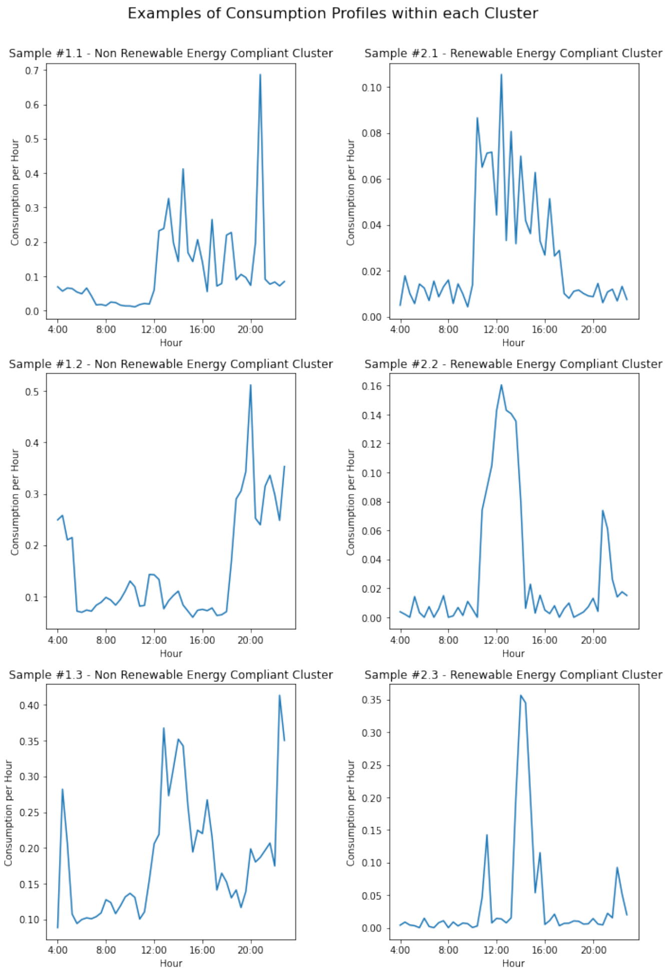

Fig. 4. Examples Retrieved from each Representative Cluster

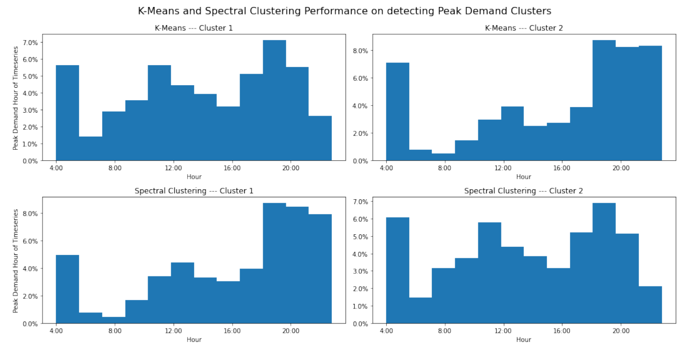

Fig. 5. Classic Clustering Algorithms Results

to a halfhourly rate of two samples per hour, thus resulting in vectors of size 48 in order to model the profile of one single day for each residency. After the aforementioned pre-processing steps, the final dataset resulted in containing three million daily energy consumption profiles across the span of two years, representing approximately five thousand and five hundred residential buildings.

The first line of experiments pertains to the training of the autoencoding neural networks which has to be performed before proceeding to the rest of the algorithm. Thoroughly represented in Figure 2, is the comparison between three versions of the autoencoder neural network : the classic Autoencoder, the Variational Autoencoder, and Variational Recurrent Autoencoder, which is a Variational Autoencoder enhanced with LSTM neural networks in the encoding and decoding modules.

The performance of each algorithm is within the expected results, as perceived by the theoretical foundations of each corresponding algorithm. The Variational Autoencoder outperforms its classic version in terms of loss, while the LSTM-enhanced encoder succeeds as the most efficiently trained neural network. All measurements and training results are averaged over five different folds of the training data, while further repeating the training experiment three times per fold.

Subsequently, the percentage of timeseries data from each fold that was withheld as a testing set, is now being encoded into latent space of five dimensions and clustered with their corresponding predetermined manner. Taking into consideration the Custom Cluster Representative logic, the experimental clustering is being performed with a number of clusters $K = 2$, as many as the energy consumption profiles.

In Figure 3, the level of distinguishability between the characteristics of each cluster is being highlighted. Our proposed algorithm of assigning each timeseries to a predetermined vector of elements modelled against daily solar radiation and encoded in the same manner as the input data, is able to efficiently disambiguate which cluster has captured the patterns of timeseries which peak demand timespan correlates with the production of "green" and renewable energy, and which is not.

As also seen in Figure 4, a few representative example timeseries from each cluster clearly correlate with their assigned cluster profile.

In contrast, the performance of the classic clustering algorithms didn't manage to deliver an efficient separation of the timeseries, in terms of detecting their specific hour of peak demand and grouping them together in a cluster of well-determined characteristics. As seen in Figure 5, neither K-Means nor Spectral Clustering algorithms managed to captivate a cluster that evidently represents timeseries of users with renewable energy compliant patterns.

VI. Conclusions

The paper presents a method for managing load profile of time series that represent building energy consumption data, and for dynamically detecting fundamental characteristics such as their peak energy demand through state-of-the-art deep learning techniques. Smart metering solutions make this feasible at a cost of vast datasets that need to be efficiently dealt with on real time. To this end, our work focuses on compression algorithms and clustering encoded time series. Experimentation shows that Variational Recurrent Autoencoder enhanced with LSTM neural networks in its encoding and decoding modules produces less loss per epoch with reference to classic and Variational Autoencoders. This paper's Predetermined Cluster Representatives algorithm performance, proves its efficiency and is able to accurately disambiguate peak demand hours in timeseries thus managing to detect load profiles that heavily overload a regional grid's "Duck Curve" problem. The concentration of future work for this project, will pertain to the inclusion of new dimensions for each timeseries such as weather statistics, or the socioeconomic status of the residential buildings. These new features, will further aid the machine learning models in their objective of properly identifying homogenous groups of similar consumption behaviours.

REFERENCES

[1] V. Amicarelli, C. Bux, G. Lagioia, and T. Lagioia, "Building energy efficiency. the economic campus of the university of bari aldo moro," *New Trends in Sustainable Business and Consumption*, p. 739, 2019.

[2] E. Commission, "Energy performance of buildings directive," May 2019. [Online]. Available: https://ec.europa.eu/energy/topics/energy-efficiency/energy-efficient-buildings/energy-performance-buildings-directive_en

[3] M. Economidou, B. Atanasiu, C. Despret, J. Maio, I. Nolte, and O. Rapf, "Europe's buildings under the microscope. a country-by-country review of the energy performance of buildings," *Buildings Performance Institute Europe (BPIE)*, pp. 35–36, 2011.

[4] Z. Ma, J. Xie, H. Li, Q. Sun, Z. Si, J. Zhang, and J. Guo, "The role of data analysis in the development of intelligent energy networks," *IEEE Network*, vol. 31, no. 5, pp. 88–95, 2017.

[5] H. O. R. Howlader, M. M. Sediqi, A. M. Ibrahimi, and T. Senjyu, "Optimal thermal unit commitment for solving duck curve problem by introducing csp, psh and demand response," *IEEE Access*, vol. 6, pp. 4834–4844, 2018.

[6] C. J. Kibert, *Sustainable construction: green building design and delivery*. John Wiley & Sons, 2016.

[7] C. Camarasa, C. Nägeli, Y. Ostermeyer, M. Klippel, and S. Botzler, "Diffusion of energy efficiency technologies in european residential buildings: A bibliometric analysis," *Energy and Buildings*, p. 109339, 2019.

[8] "Directive 2010/31/EU of the European Parliament and of the Council of 19 May 2010 on the energy performance of buildings," p. 23.

[9] Z. T. Gill and M. Pegg, "I., & allan, n.(2010)," *Low-energy dwellings: The contribution of behaviours to actual performance. Building Research & Information*, vol. 38, no. 5, pp. 491–508.

[10] A. Druckman, M. Chitnis, S. Sorrell, and T. Jackson, "Missing carbon reductions? exploring rebound and backfire effects in uk households," *Energy Policy*, vol. 39, no. 6, pp. 3572–3581, 2011.

[11] T. A. Nguyen and M. Aiello, "Energy intelligent buildings based on user activity: A survey," *Energy and buildings*, vol. 56, pp. 244–257, 2013.

[12] EnergyPlus, "U.S. Department of Energy's (DOE) Building Technologies Office (BTO) and National Renewable Energy Laboratory (NREL)," 2015. [Online]. Available: https://energyplus.net/

[13] C. Koulamas, A. Kalogeras, R. Pacheco-Torres, J. Casillas, and L. Ferrarini, "Suitability analysis of modeling and assessment approaches in energy efficiency in buildings," *Energy and buildings*, vol. 158, pp. 1662–1682, 2018.

[14] L. Ferrarini, E. Fathi, S. Disegna, and S. Rastegarpour, "Energy consumption models for residential buildings: a case study," in *2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*. IEEE, 2019, pp. 673–678.

[15] G. Kalogeras, S. Rastegarpour, C. Koulamas, A. Kalogeras, J. Casillas, and L. Ferrarini, "Predictive capability testing and sensitivity analysis of a model for building energy efficiency," in *Building Simulation*, vol. 13, no. 1. Springer, 2020, pp. 33–50.

[16] R. Hendron and C. Engebrecht, "Building america house simulation protocols (revised)," National Renewable Energy Lab.(NREL), Golden, CO (United States), Tech. Rep., 2010.

[17] K. Darcovich, E. Entchev, and P. Tzscheutschler, "An international survey of electrical and dhw load profiles for use in simulating the performance of residential micro-cogeneration systems," *ECB Annex*, vol. 54, 2014.

[18] A. Al-Wakeel and J. Wu, "K-means based cluster analysis of residential smart meter measurements," *Energy Procedia*, vol. 88, pp. 754 – 760, 2016, cUE 2015 - Applied Energy Symposium and Summit 2015: Low carbon cities and urban energy systems.

[19] 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.

[20] J. P. Gouveia and J. Seixas, "Unraveling electricity consumption profiles in households through clusters: Combining smart meters and door-to-door surveys," *Energy and Buildings*, vol. 116, pp. 666 – 676, 2016.

[21] O. Y. Al-Jarrah, Y. Al-Hammadi, P. D. Yoo, and S. Muhaidat, "Multi-layered clustering for power consumption profiling in smart grids," *IEEE Access*, vol. 5, pp. 18 459–18 468, 2017.

[22] G. L. Ray] and P. Pinson, "Online adaptive clustering algorithm for load profiling," *Sustainable Energy, Grids and Networks*, vol. 17, p. 100181, 2019.

[23] A. Rajabi, M. Eskandari, M. J. Ghadi], S. Ghavidel, L. Li, J. Zhang, and P. Siano, "A pattern recognition methodology for analyzing residential customers load data and targeting demand response applications," *Energy and Buildings*, vol. 203, p. 109455, 2019.

[24] K.-J. Park and S.-Y. Son, "A novel load image profile-based electricity load clustering methodology," *IEEE Access*, vol. 7, pp. 59 048–59 058, 2019.

[25] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[26] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of the International Conference on Learning Representations (ICLR)*, vol. 1, 2014.

[27] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proceedings of ICML workshop on unsupervised and transfer learning*, 2012, pp. 37–49.

[28] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[29] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[30] J. T. Connor, R. D. Martin, and L. E. Atlas, "Recurrent neural networks and robust time series prediction," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 240–254, 1994.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*. Ithaca, NY: arXiv.org, 2015.

[32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[33] J. Han and C. Moraga, "The influence of the sigmoid function parameters on the speed of backpropagation learning," in *International Workshop on Artificial Neural Networks*. Springer, 1995, pp. 195–201.

[34] H. Steinhaus, "Sur la division des corp materiels en parties," *Bull. Acad. Polon. Sci*, vol. 1, no. 804, p. 801, 1956.

[35] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[36] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "Dbscan revisited, revisited: why and how you should (still) use dbscan," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, pp. 1–21, 2017.

[37] D. Jiang, J. Pei, and A. Zhang, "Dhc: a density-based hierarchical clustering method for time series gene expression data," in *Third IEEE Symposium on Bioinformatics and Bioengineering, 2003. Proceedings.* IEEE, 2003, pp. 393–400.

[38] A. Brunger, F. Hooper, and D. Wardle, "A comprehensive sky radiance data base," in *Intersol Eighty Five*. Elsevier, 1986, pp. 2415–2419.

[39] Y. Gal and Z. Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," in *Advances in neural information processing systems*, 2016, pp. 1019–1027.

[40] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 448–456.

[41] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for lvcsr using rectified linear units and dropout," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 8609–8613.