Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

DYNAMIC LOAD USAGE BEHAVIOR SIMULATION IN SMART GRIDS: A DATA-DRIVEN APPROACH IN URBAN BUILDINGS

Shuang Dai[1], Fanlin Meng[2,3]

[1]Dept. of Engineering, University of Exeter, Exeter, UK
[2] Business School, University of Exeter, Exeter, UK
[3] Alliance Manchester Business School, University of Manchester, Manchester, UK

ABSTRACT

Smart grids are essential for sustainable urban energy systems, improving efficiency and integrating renewable sources. Accurately forecasting load demand is key for effective management, but is challenging due to unpredictable behaviors and dynamic consumption patterns. This paper introduces a new data-driven approach using smart meter data from various buildings in Cardiff, UK to better understand electricity consumption behaviors across seasons. Our methodology combines machine learning techniques with an in-depth analysis of physical building characteristics to conduct dynamic load usage behavior simulation. We employ consensus-based clustering to identify buildings with similar consumption behaviors and track dynamic changes in load usage over time. Furthermore, we identify key load-related features that influence consumption patterns, enhancing the precision of load demand forecasting. Empirical validation of our approach underscores its effectiveness in enhancing forecast accuracy and providing robust, sustainable strategies for energy management within the smart grid paradigm.

1 INTRODUCTION

Buildings are integral to the functioning of society and are a significant part of global energy consumption, accounting for one-third of total global energy consumption (González-Torres et al. 2022). Forecasting building load is a critical component in enhancing the flexibility and reliability of energy system operations (Ahmad et al. 2020). The advent of smart meters has ushered in a new era of high-resolution building-level load data, opening up novel opportunities for forecasting applications such as peer-to-peer energy trading and distribution grid operations.

Approaches to building load forecasting typically fall into two categories: physics-based models and data-driven models. The former relies on detailed physical information—such as building material types and ventilation system parameters—which is often difficult to obtain and prone to errors in information collection, leading to unsatisfactory forecasting results. Conversely, data-driven approaches, particularly those utilizing machine learning, leverage real-world data to discern the underlying relationships between variables and have been increasingly favored for their superior performance in the past two decades (Chen et al. 2022). Moreover, the significant influence of occupant-related inputs like occupant schedule, and device usage profile on building energy simulation results and energy demand forecasting further emphasizes the importance of data-driven models (Panchabikesan et al. 2021).

In building energy simulations, default schedules are commonly used to represent occupant behavior. However, these schedules may not accurately represent real-life occupancy patterns, which can change based on the day of the week or the season (Li et al. 2019). This inconsistency can result in inaccurate simulation outcomes (Panchabikesan et al. 2021). The primary reason for using these default schedules is the lack of available data, and in most instances, the actual behavior of occupants remains unknown (Happle et al. 2020). Therefore, when occupancy data is not readily available, it becomes important to

examine the relationship between the physical characteristics of a building and its energy consumption, which can provide valuable insights in refining building energy simulation models.

Building energy usage data at high temporal resolution has become increasingly accessible due to the proliferation of smart meters. However, forecasting models frequently encounter difficulties in achieving high accuracy, mainly because of the unpredictability and randomness of building occupant behavior. This variability introduces significant variance in the energy usage patterns of individual buildings. Studies have shown that substantial energy savings in residential buildings, from 5-25%, can be achieved through strategies targeted at occupants (Li et al. 2019). However, existing recommendations for energy-saving measures have overlooked key aspects related to occupant behavior. Specifically, the diversity of occupant schedules and variability in activity levels across building types have not been adequately accounted for when crafting conservation strategies. Given these circumstances, it becomes imperative to analyze the diversity associated with occupant energy usage behavior in building load forecasting.

This paper introduces a comprehensive, data-driven methodology that capitalizes on smart meter data to simulate dynamic building load usage behavior. Our approach transcends traditional energy simulation techniques by integrating a dynamic load usage behavior simulation that accounts for the temporal clustering of building energy usage, supported by chi-squared analysis to ascertain the influence of various load-related features. The simulations are rooted in real-world load usage data, enabling us to mimic and understand the complex and fluctuating load consumption behaviors of different buildings. Our methodology aims to provide a predictive tool that can adapt to and anticipate the diverse and ever-changing demands of building energy usage, ultimately supporting smarter energy distribution and utilization. The contributions of this paper are as follows:

- By addressing the challenge of reconciling individual behavior unpredictability with the precision of load demand forecasting, we enhance building energy simulations and develop robust and sustainable strategies for energy management within smart grid environments.
- We propose an innovative, data-driven approach that leverages smart meter data to understand electricity consumption behaviors across different seasons.
- Our methodology allows monitoring and analyzing the dynamic clustering trajectories of 167 real-world buildings over the year, offering valuable insights into temporal energy use trends.
- We identify and quantify key load-related features that influence consumption patterns, enhancing the precision of load demand forecasting within the context of smart grids.

The remaining sections of the paper are organized as follows: Section 2 discusses related work, Section 3 presents the proposed data-driven framework, Section 4 covers the experimental analysis, and the paper concludes and points to future work in Section 5.

2 RELATED WORK

The emergence of smart grids and the use of data analytics techniques have enabled new approaches to load demand profiling and forecasting, which have been further explored from the perspectives of system operators and utilities in recent studies.

Load profiling categorizes consumers into distinct groups based on their energy usage patterns. This categorization enables utilities to tailor services and operations more effectively, such as electricity tariff design (Azarova et al. 2018), engaging consumers in demand response (Parrish et al. 2020). Clustering techniques are fundamental in load profiling, which has been applied in many recent studies to make informed decisions for demand-side management and energy savings initiatives. Tang et al. (2022) focused on load profiling by utilizing K-Medoids clustering to understand residential load consumption patterns and identify the drivers of these patterns from a socioeconomic perspective. While Rafiq et al. (2023) analyzed residential electricity consumption in Dubai based on characteristics of the dwellings and smart meter data. The study adopted K-means clustering to group consumption profiles and used classification algorithms

to predict household consumption patterns based on dwelling and occupant characteristics. These studies underscore the importance of considering diverse factors in creating accurate load consumption profiles.

Dynamic clustering techniques consider changes and transitions in consumption patterns over time, offering a more nuanced analysis compared to traditional static methods. Wang et al. (2016) introduced a novel clustering approach for analyzing electricity consumption behavior dynamics, considering transitions and relations between consumption levels over time. Meng et al. (2023) proposed a multiple dynamic pricing approach for demand response, utilizing customer segmentation and customized demand models to achieve optimal pricing. Chen et al. (2023) utilized a consensus-based clustering approach to group offices based on their indoor temperature profiles for various seasons. By tracking dynamic cluster trajectories, the system can suggest indoor thermal optimization strategies.

Building on the above foundations, our approach integrates seasonal analysis into the dynamic load usage behavior simulation process. Seasonal variations are critical as they can significantly alter consumption patterns. By incorporating a seasonal decomposition approach within clustering techniques, we can refine the load profiles to reflect seasonal shifts in energy use. This enriched profiling not only aids in the design of more effective electricity tariffs but also enhances demand response strategies by aligning them with seasonal consumption trends.

In load demand forecasting, statistical and machine learning techniques applied to smart meter data have shown significant improvements. For instance, Geetha et al. (2021) demonstrated the effectiveness of random forest models for predicting power consumption and peak demand. Bashawyah and Qaisar (2021) successfully applied K-nearest regression and support vector regression for short-term forecasting with high accuracy using London household data. However, the nonlinearity of electrical load data has prompted the use of more sophisticated models. For example, long short-term memory networks (Hochreiter and Schmidhuber 1996), have shown promise in handling such complexities. Kwon et al. (2020) proposed a long short-term memory networks model for day-ahead forecasting in Korea over two years. Moreover, Moradzadeh et al. (2020) combined support vector regression and long short-term memory networks for short-term load forecasting in microgrids, effectively capturing behavioral patterns in input variables and providing more precise load forecasts.

Existing studies have identified gaps in presenting and comparing various methods in load forecasting and a lack of interpretability in the forecasting models, making it challenging for operators to trust the results. To overcome these challenges, this study develops bespoke load forecasting models based on load profiling results and combines load consumption-related features to enhance model interpretability, providing utilities with clearer insights into factors driving load forecasts.

3 PROPOSED DATA-DRIVEN FRAMEWORK

A graphical overview of the system framework is shown in Figure 1 to show the procedure involved in the proposed data-driven framework for dynamic load behavior simulation.

3.1 Building Smart Meter Data Management

Effective preprocessing of raw datasets is essential before conducting data analysis. This preprocessing includes data cleaning, data segmentation, and data normalization.

Specifically, the data is initially analyzed season-wise through seasonal segmentation. Using unsupervised learning techniques, load behavior simulation and forecasting of buildings are performed based on four seasonal segmentation groups. To prioritize shape features of load patterns over amplitude ones for time series clustering, z-normalization is then applied to normalize the seasonal groups.

3.2 Building Load Profiling

In this phase, we conduct clustering on normalized seasonal load consumption data. Clustering involves both distance measurements and cluster prototypes. Recognizing that sole reliance on a single clustering

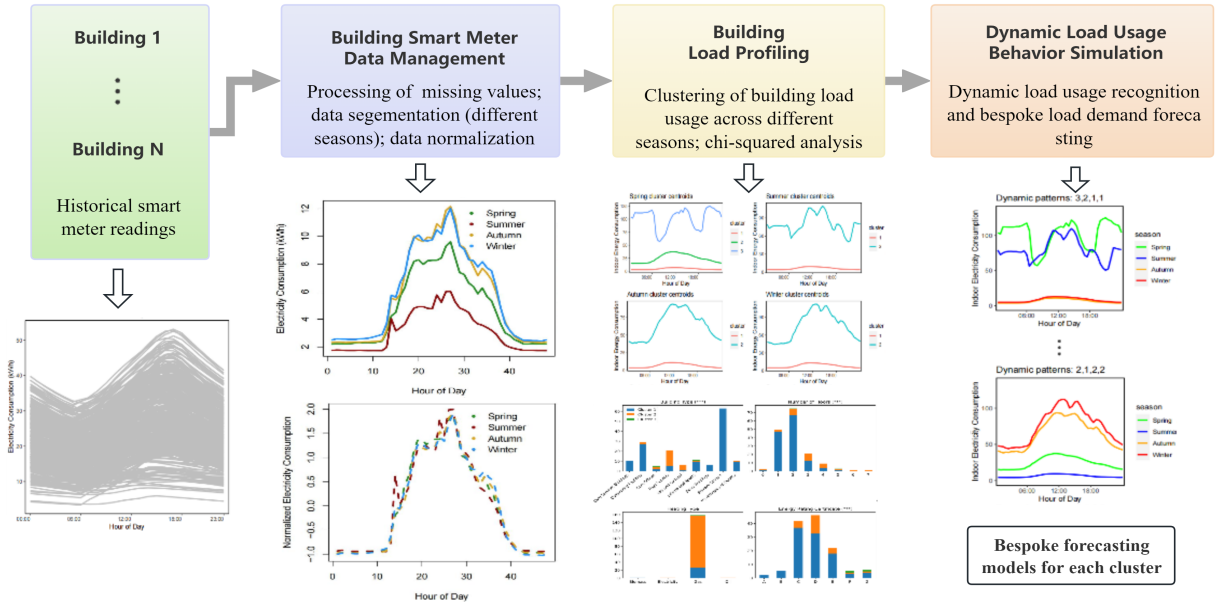


Figure 1: The data-driven framework for dynamic load behavior simulation.

algorithm can yield unstable results (Satre-Meloy et al. 2020), we utilize a consensus-based approach proposed by Chen et al. (2023) that combines multiple distance measurements and cluster prototypes to improve the robustness of building load profiling.

We explore various combinations of distance measurements and prototyping functions to capture the multifaceted nature of energy consumption patterns. The chosen combinations are as follows:

- **Euclidean Distance with *K*-means**. This classic clustering algorithm efficiently identifies spherical clusters by using Euclidean distance (Lloyd 1982). It helps pinpoint central tendencies in load profiles, serving as a baseline for grouping building load usage behavior.
- **Euclidean Distance with Partition Around Medoids (PAM)**. PAM, combined with Euclidean distance, offers robustness to outliers compared to *K*-means (Rdusseeun and Kaufman 1987). It identifies representative objects in each cluster, providing a more resilient characterization of load profiles, crucial for stable dynamic load behavior simulations.
- **Dynamic Time Warping (DTW) with DTW Barycenter Averaging (DBA)**. This clustering algorithm accommodates energy usage variations over time, reflecting temporal shifts in building load profiles (Petitjean et al. 2011). This combination is effective for dynamic load behavior simulation, adapting to energy consumption variability over different periods.
- **Shape-Based Distance (SBD) with Shape Extraction**. SBD focuses on load profile shape, disregarding amplitude and phase (Paparrizos and Gravano 2015). Integrating SBD with shape extraction techniques distills intrinsic waveform patterns of electricity usage, aiding in detecting characteristic load shapes corresponding to electricity consumption behaviors.

The clustering results are quantitatively selected by three cluster validity indices (CVIs): the Silhouette index, Davies-Bouldin (DB) index, and Calinski-Harabasz (CH) index, which help determine the optimal clustering algorithm and cluster number through majority voting. The Silhouette index is employed to assess the similarity of an object within its own cluster compared to other clusters, with higher values indicating more defined clusters. The CH index also points to better clustering outcomes with higher values. Conversely, the DB index functions in an inverse manner, where lower values suggest a more distinct and suitable clustering configuration. A more detailed description of the evaluation metrics can be found in the studies by de Zepeda et al. (2021) and Sardá-Espinosa (2019).

Then, the association between clusters representing different seasons and building physical factors is tested using the chi-square test (Pearson 1900). The final building load profiling outcomes include the building load consumption clusters in different seasons and the association between the seasonal clusters and the building physical factors.

3.3 Dynamic Load Usage Behavior Simulation

Dynamic load usage behavior simulation represents a sophisticated analytical approach to model and predict the fluctuating patterns of load usage within buildings. This process is integral to developing strategies for energy conservation, demand response, and smart grid management. The simulation encompasses two critical facets: recognizing dynamic load usage recognition and bespoke load demand forecasting.

3.3.1 Dynamic load usage recognition

In this part of our methodology, we simulate dynamic load usage behavior based on real-world smart meter data. The simulation is informed by the observed load usage patterns of buildings over time. Specifically, we create a 'dynamic trajectory' that reflects how the electricity consumption of each building evolves over different seasons. This simulated dynamic load usage behavior trajectory allows us to anticipate and understand changes in load demand related to seasonal variations. For instance, increased heating in winter and heightened cooling in summer are expected in the cluster trajectories.

On the other hand, this method enhances our understanding of the variations in seasonal behaviors across different building groups. It may uncover, for instance, that specific buildings are particularly responsive to temperature fluctuations based on their construction characteristics, while others display more consistent consumption profiles throughout the year. Leveraging the insights from dynamic clustering outcomes, stakeholders can formulate tailored approaches for building energy management. This could include optimizing HVAC systems, implementing energy-efficient practices during peak periods, or designing buildings to better adapt to seasonal energy patterns.

3.3.2 Bespoke load demand forecasting

This stage employs bespoke load demand forecasting models to simulate future electricity usage in buildings. This simulation is based on the clusters identified previously, with the goal of finding the forecasting model that best matches the real consumption data within each cluster. To support the forecasting process, we have specifically developed load consumption-related features, which are detailed in Table 1.

Table 1: The developed load consumption-related features for load forecasting.

Input	Dimensions	Description
L_h^{week}	1	h_{th} hour load on the same day of last week
L_h^{day}	1	h_{th}hour load of yesterday
L_h^{hour}	1	Load of the $(h-1)_{th}$ hour of today
F	2	One-hot code for festival/non-festival day
Y	2	One-hot code for year index
M	12	One-hot code for month index
W	7	One-hot code for day index of a week
H	24	One-hot code for hour index of a day

Selecting the top features is crucial for developing robust, efficient, and interpretable models that can better generalize to new, unseen data. The random forest algorithm is especially valuable for feature selection because it generates an importance distribution for these features. This measure is used to identify the most influential load consumption-related features from the clustered data. The five most significant features, as determined by this distribution, are then used to train forecasting models.

Then, different forecasting algorithms are employed to construct models, each fine-tuned using grid search to optimize their parameters:

- **Support Vector Regression (SVR)**: This algorithm applies the principle of structural risk minimization to effectively tackle non-linear problems (Ahmad et al. 2018).
- **Random Forest (RF)**: As an ensemble method, it creates a multitude of decision trees and integrates their outcomes to make predictions (Breiman 2001).
- ***K*-Nearest Neighbors Regression (KNR)**: It estimates new data points by averaging the k nearest instances, offering a straightforward regression approach (Bhattacharya et al. 2017).
- **Long Short-Term Memory (LSTM)**: This variety of recurrent neural networks excels at recognizing dependencies in sequences, which is crucial for predicting time-series data (Liu et al. 2015).

After tuning, each algorithm is rigorously evaluated. The goal is to identify the best-performing model for each cluster, which is expected to closely align with the unique electricity consumption patterns within that cluster. This process helps in selecting the most effective forecasting tool tailored to the specific load behavior dynamics. In addition, a comparative performance analysis of the selected models provides valuable insights into how the different algorithms correspond to the different consumption patterns.

4 EXPERIMENTAL ANALYSIS

4.1 Dataset and Processing

In the proposed system, we utilized real-world smart meter data from buildings in Cardiff, UK to analyze load variations across different seasons. It is worth noting that this dataset has not been previously used for load consumption analysis. The data, sourced from Cardiff Council , includes various building types such as offices, community facilities, schools, and cultural buildings. Historical load data was preprocessed by replacing missing values below 20% with column means to ensure accurate forecasting during model training. The pre-processed data of each building includes half-hourly load records in 2015 and physical attributes including building type, floor number, heating type, and energy rating certificate.

Table 2: Physical information of the considered 167 buildings.

Building Type	Count	Floor		Heating		Energy Rating			
		0-3	4+	Gas	Other	A-B	C-E	F-G	NA
Care Services Buildings	10	10	0	10	0	0	4	1	5
City Services	6	6	0	-	-	0	0	0	6
Community Facilities	29	27	2	29	0	1	14	2	12
Core Offices	5	3	2	5	0	0	2	2	1
High Schools	21	15	6	20	1	0	19	1	1
Key and Cultural	6	3	3	6	0	1	4	0	1
Leisure and Sports	11	11	0	11	0	3	5	1	2
Parks Buildings	6	6	0	6	0	1	2	0	3
Primary Schools	63	62	1	62	1	1	56	3	3
Workshops and Depots	10	10	0	9	1	0	4	1	5

'-' indicates data not available.

The physical information of the 167 considered buildings is detailed in Table 2. Among the various building types, primary schools emerge as the most prevalent, with 62 buildings characterized by up to three floors and gas heating systems. The distribution of buildings across different energy ratings reveals a concentration in the mid-level categories, specifically falling within the range of C to E ratings. Variations in physical characteristics may lead to different electricity consumption patterns, offering insights for grid operators to optimize energy management strategies. The system aims to analyze the relationship between physical attributes and load behavior, as well as model and forecast dynamic load patterns within buildings.

4.2 Results and Discussion

In this section, we will start by presenting the results of the building load profiling, which include: (1) building load consumption clusters in different seasons, and (2) the correlation between seasonal clusters and the physical characteristics of the buildings. Following that, we will discuss the outcomes of the dynamic load behavior simulation process, which include: (1) the trajectory of clusters across different seasons, and (2) the development of customized load demand forecasting models.

4.2.1 Building load profiling results

We utilized consensus-based clustering to cluster the 167 Cardiff buildings based on their daily load records in different seasons. The optimal cluster number and algorithm for each season were determined through a majority vote from the Cluster Validity Indices (CVIs). We considered cluster numbers ranging from 2 to 10. Based on the majority vote, the CVIs indicated that three clusters were the best clustering number for the spring season, while two clusters were preferred for the other three seasons. The final chosen clustering algorithms and their corresponding CVIs for the four seasons are listed in Table3.

Table 3: Lists of the optimal clustering algorithms and their corresponding CVIs for the four seasons.

Season	Optimal clustering algorithm	Optimal cluster number	CVIs		
			Silhouette	DB	CH
Spring	DTW+DBA	3	0.75	0.71	178.38
Summer	DTW+DBA	2	0.86	0.82	156.46
Autumn	DTW+DBA	2	0.82	0.52	192.69
Winter	DTW+DBA	2	0.82	0.51	185.74

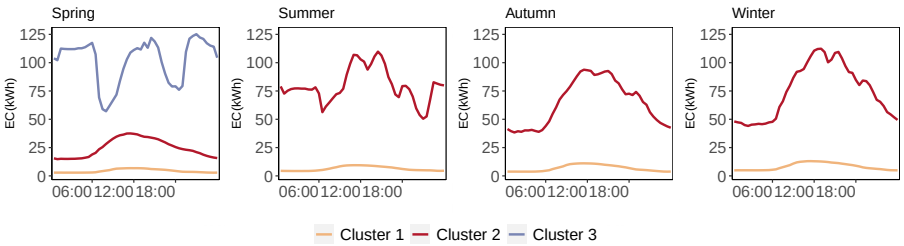
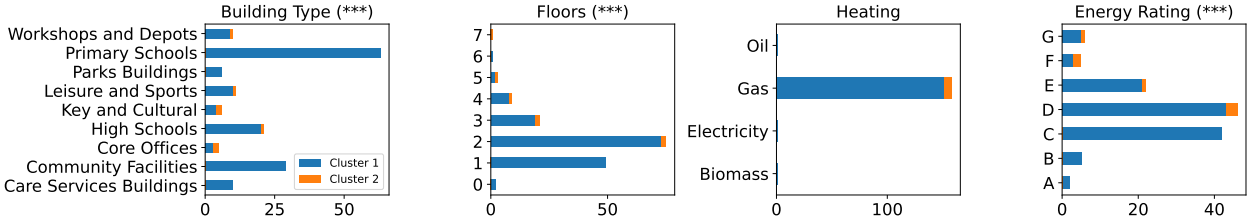
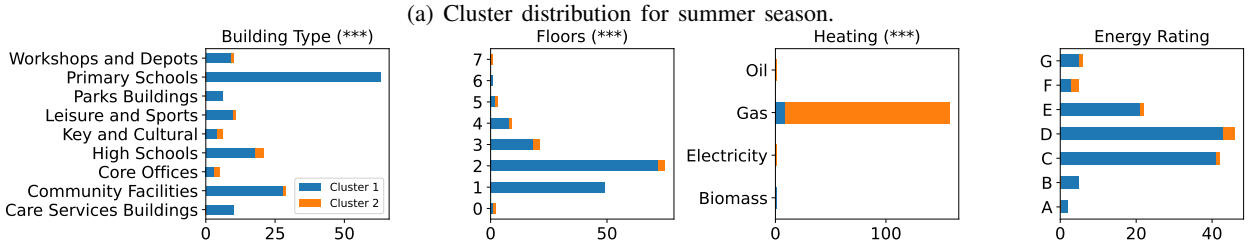


Figure 2: The cluster centroids of clusters in different seasons.



(a) Cluster distribution for summer season.



(b) Cluster distribution for winter season.

Figure 3: Cluster distribution and chi-square test results in representative seasons.

Table 4: Category summary for the dynamic cluster trajectories over the year.

Spring cluster	Summer cluster	Autumn cluster	Winter cluster	Spring Mean	Summer Mean	Autumn Mean	Winter Mean	Number of Buildings	DT No.
1	1	1	1	4.39	3.74	4.56	5.49	137	1
2	1	1	1	19.05	20.18	24.89	27.64	17	2
		2	1	23.91	32.27	37.36	36.23	2	3
		2	2	30.51	31.78	42.85	49.52	3	4
	2	2	2	41.40	53.25	60.88	66.31	5	5
3	2	1	1	104.94	94.38	0.21	0.53	1	6
		2	2	100.75	139.03	132.93	134.96	2	7

Dynamic time warping clustering is selected as the optimal clustering method for all four seasons. Figure 2 plots the cluster centroids for the clusters in the four seasons. Each cluster represents a distinct daily electricity consumption pattern by season. Warmer seasons (spring and summer) show peaks around noon and afternoon valleys. Colder seasons (autumn and winter) have stable patterns with peaks around noon and late afternoon.

Following the consensus results, a chi-square test was conducted to explore the association between clusters concerning the categorical physical factors of buildings across different seasons. To highlight these seasonal variations in influencing factors and save space, we specifically plotted the cluster distribution and chi-square test results for the contrasting seasons of summer and winter in Figure 3.

According to the results, it is worth noting that the clusters are imbalanced, the rationale behind this could be due to the nature of the building stock, where educational buildings are more prevalent than others. The results of our analysis show a significant consistency between the clusters identified in summer and winter, except for heating types. This consistency highlights stable, season-independent factors like building type and number of floors that influence load consumption patterns. These factors were also found to be statistically significant in relation to clusters throughout the year, indicating their importance as variables that continually differentiate between the clusters. Moreover, in the warm season, it was the energy rating that showed a significant association with clusters, which indicates that buildings with similar energy ratings might exhibit similar characteristics or behaviors that define the clusters during the warm season. Conversely, in the cold season, the type of heating emerged as a statistically significant factor associated with clusters. This underscores the heightened dependence on heating appliances during the cold seasons as a primary influence on building load consumption.

4.2.2 Dynamic load usage behavior simulation results

Following the load profiling analysis, a category summary is conducted to monitor the dynamic cluster trajectories (DTs) of the buildings. Table 4 provides a summary of all the cluster trajectories observed in the buildings, along with the average load profile associated with each trajectory in the four seasons. The table also includes the count of buildings corresponding to each DT category, which allows for understanding the distribution of buildings across different trajectories.

A total of seven dynamic trajectories (DTs) were identified in the clustering results for the 167 buildings. By analyzing the load usage behavior of each DT and combining the cluster labels for each season, we can observe potential changing points within the trajectories. The imbalanced distribution, where the majority of buildings in DT1 show no significant changes in behavior over time, could indicate that a significant portion of buildings in the dataset have consistent and predictable energy consumption profiles, suggesting efficient energy use or well-maintained systems.

However, it is also important to address the needs of buildings with more dynamic load trajectories. Exploring ways to adapt energy management approaches to accommodate these buildings could lead to more tailored and effective energy efficiency interventions. For instance, DT2 shows a medium load consumption in spring, followed by a decrease in load consumption over the subsequent seasons. This shift may be attributed to the heating systems used in buildings within DT2, which could be more energy-efficient, such

as electricity, gas, or oil. Similarly, DT3 experiences an increase in load consumption during autumn, while DT4 shows an increase in load consumption during both autumn and winter due to biomass heating.

Understanding these changing points within the DTs is crucial for smart grid operators to anticipate variations in energy demand. Specifically, DTs 3-5 should be closely monitored during the winter season for potential higher demand, while DT5 may require attention during the summer months due to increased electricity consumption. In contrast, buildings in DT6 exhibit higher electricity usage in spring and summer, with a decrease during autumn and winter. On the other hand, buildings in DT7 demonstrate high load consumption throughout the year, mainly due to large floor areas and biomass heating during cold months. Smart grid operators should allocate sufficient load budgets for buildings in DT6 during summer and those in DT7 throughout the year, considering these changing patterns in load consumption behavior.

To visually illustrate patterns in how different building types correspond to various DTs, Figure 4 plots the heatmap to compare the relative distribution across different building types and DTs.

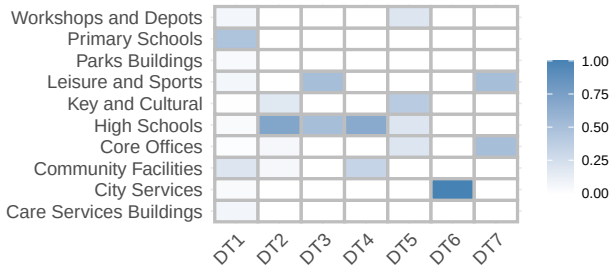


Figure 4: Building type percentage distribution by DTs.

The heatmap reveals distinct energy consumption patterns for various building types across different DTs. For example, primary schools mostly consume electricity within DT1, indicating a stable load usage pattern, likely tied to the regular school schedule. High schools, on the other hand, show significant energy use in DT2 and DT3, suggesting variable consumption that may correspond to seasonal school operations and activities. In contrast, city services are uniquely associated with DT6, reflecting a specialized load usage pattern possibly linked to consistent municipal operations. Leisure and sports facilities have substantial load demands in both DT3 and DT7, indicating high usage that could be associated with specific seasonal events or year-round activities. Core offices display energy consumption in DT1 and DT7, suggesting a mix of steady and peak usage times throughout the year.

In the bespoke load forecasting phase, we consider two metrics to evaluate forecasting performance, including the mean absolute percentage error (MAPE), and the R^2. MAPE offers a comparative percentage error across different scales of electricity consumption:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{\hat{y}_i}\right| \times 100\% \tag{1}$$

Meanwhile, R^2 quantifies how well the predicted values from the model fit with the actual values, with a range between 0 and 1, where 1 indicates a perfect fit:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(\bar{y}_i - y_i)^2} \tag{2}$$

Here, $\hat{y}_i$, y_i, and $\bar{y}_i$ are the observed i-th load consumption, the predicted electricity consumption, and the observed mean consumption, respectively. N denotes the size of the testing dataset, and i is the index of test observations.

Based on the feature importance reported by RF, for each season, the first five most important features are selected and displayed in Table 5.

In spring, 'last hour' energy usage is the most important feature for predicting consumption in most clusters, typically representing primary schools, with specific times like 9:00 AM also being significant due

Table 5: Feature importance distribution in different clusters across seasons.

Season	Cluster	Top Feature 1	Top Feature 2	Top Feature 3	Top Feature 4	Top Feature 5
Spring	C1	last hour (0.8904)	last day (0.0395)	last week (0.0358)	9:00 (0.0024)	0:00 (0.0024)
	C2	last hour (0.9017)	last day (0.0450)	last week (0.0280)	9:00 (0.0024)	Monday (0.0019)
	C3	last day (0.6992)	last hour (0.2258)	7:00 (0.0217)	last week (0.0158)	20:00 (0.0061)
Summer	C1	last hour (0.8999)	last day (0.0584)	last week (0.0251)	Saturday (0.0011)	7:00 (0.0011)
	C2	last hour (0.7624)	last day (0.1297)	last week (0.0552)	6:00 (0.0133)	7:00 (0.0085)
Autumn	C1	last hour (0.8831)	last day (0.0502)	last week (0.0464)	Monday (0.0019)	Saturday (0.0012)
	C2	last hour (0.6390)	last week (0.2630)	last day (0.0513)	Wednesday (0.0053)	Saturday (0.0039)
Winter	C1	last hour (0.9124)	last day (0.0350)	last week (0.0306)	23:00 (0.0016)	9:00 (0.0014)
	C2	last hour (0.5174)	last week (0.4033)	last day (0.0396	20:00 (0.0029)	18:00 (0.0028)

to school start times. During summer, 'last hour' remains the key feature for cluster 1, while in cluster 2, 'last day' becomes more prominent, suggesting a shift in energy use patterns, perhaps due to daily cooling needs. Specific morning hours gain importance, reflecting the start-up of cooling systems. The distribution of feature importance from fall to winter indicates a persistent trend of 'last hour' usage being a crucial determinant, with signs that weekly cycles also have a substantial impact. This shift with the seasons highlights the effect of heating needs and operational timings on electricity use.

After selecting the five most important features for each cluster, forecasting models were built for each cluster of each season based on the selected algorithms, and a total of 45 models were built. The forecasting results outlined in Table 6 enable the tailored model selection across different seasons and clusters. The bespoke approach aims to identify the most appropriate model for each unique scenario, thereby optimizing forecasting accuracy.

Table 6: Forecasting results for each model.

Model	Evaluation Indicators	Spring			Summer		Autumn		Winter	
		C1	C2	C3	C1	C2	C1	C2	C1	C2
SVR(Linear)	*MAPE* (%)	8.52	8.40	19.43	9.22	7.41	10.26	6.75	8.36	6.18
	R^2	0.89	0.91	0.79	0.94	0.82	0.92	0.84	0.94	0.83
SVR(RBF)	*MAPE* (%)	9.27	7.91	17.44	8.69	7.24	9.12	6.01	8.24	**5.53**
	R^2	0.91	0.94	0.87	0.96	0.83	0.95	0.83	0.96	**0.89**
RF	*MAPE* (%)	8.59	5.10	6.81	8.19	**5.51**	7.26	**4.81**	8.46	5.17
	R^2	0.93	0.95	0.92	0.97	**0.97**	0.96	**0.88**	0.96	0.87
KNR	*MAPE* (%)	8.39	**4.47**	**5.60**	6.90	6.46	7.63	4.96	7.95	5.29
	R^2	0.93	**0.95**	**0.96**	0.96	0.90	0.94	0.88	0.96	0.85
LSTM	*MAPE* (%)	**8.13**	9.41	8.68	**5.11**	6.87	**6.22**	6.55	**6.55**	7.02
	R^2	**0.94**	0.94	0.91	**0.96**	0.82	**0.97**	0.85	**0.97**	0.84

Model performance fluctuates across clusters and seasons, highlighting the absence of a one-size-fits-all solution and therefore reinforcing the necessity for a bespoke approach to load forecasting. For the spring season, the LSTM model demonstrates the lowest MAPE and highest R^2 for cluster 1, suggesting its superior ability to capture the load patterns in this cluster. Conversely, KNR outperforms other models in clusters 2 and 3. During summer, LSTM again proves to be the best fit for cluster 1, while RF excels in cluster 2 with the best MAPE and an impressive R^2 of 0.97. In autumn forecasting, RF performs best in cluster 2, while LSTM offers the most accurate forecasts for cluster 1. In winter, SVR(RBF) emerges as the optimal model for cluster 2, indicating its strength in capturing winter load variations in this cluster. In winter, SVR(RBF) excels for cluster 2, while LSTM performs best for cluster 1. These insights are valuable for effective energy management and strategic planning, enabling operators to optimize resource allocation and maintain grid stability.

5 CONCLUSION AND FUTURE WORK

This study utilized a unique dataset of smart meter readings from buildings in Cardiff to investigate seasonal load variations and the impact of building attributes on energy consumption. The results highlight distinct consumption patterns identified through cluster analysis, emphasizing the correlation between building characteristics and energy usage. The dynamic load behavior simulation offers a detailed insight into load consumption trends and aids in the development of load demand forecasting models. Building on these findings, future research could extend the dataset to include a broader time range, enriching the analysis with longitudinal trends. Integrating additional variables such as weather patterns and building occupancy rates could offer a more comprehensive view of energy consumption dynamics. Moreover, exploring the synergy between electricity consumption and renewable energy generation presents an opportunity to further optimize grid operations and support the transition to sustainable energy systems.

REFERENCES

Ahmad, M. W., M. Mourshed, and Y. Rezgui. 2018. "Tree-based ensemble methods for predicting PV power generation and their comparison with support vector regression". *Energy* 164:465–474.

Ahmad, T., H. Zhang, and B. Yan. 2020. "A review on renewable energy and electricity requirement forecasting models for smart grid and buildings". *Sustainable Cities and Society* 55:102052.

Azarova, V., D. Engel, C. Ferner, A. Kollmann and J. Reichl. 2018. "Exploring the impact of network tariffs on household electricity expenditures using load profiles and socio-economic characteristics". *Nature Energy* 3(4):317–325.

Bashawyah, D. A. and S. M. Qaisar. 2021. "Machine learning based short-term load forecasting for smart meter energy consumption data in london households". In *2021 IEEE 12th International Conference on Electronics and Information Technologies (ELIT)*, 99–102. IEEE.

Bhattacharya, G., K. Ghosh, and A. S. Chowdhury. 2017. "Granger causality driven AHP for feature weighted kNN". *Pattern Recognition* 66:425–436.

Breiman, L. 2001. "Random Forests". *Machine Learning* 45(1):5–32.

Chen, H., S. Dai, and F. Meng. 2023. "Smart Building Thermal Management: A Data-Driven Approach Based on Dynamic and Consensus Clustering". *Sustainability* 15(21):15489.

Chen, Y., M. Guo, Z. Chen, Z. Chen and Y. Ji. 2022. "Physical energy and data-driven models in building energy prediction: A review". *Energy Reports* 8:2656–2671.

de Zepeda, M. V. N., F. Meng, J. Su, X.-J. Zeng and Q. Wang. 2021. "Dynamic clustering analysis for driving styles identification". *Engineering Applications of Artificial Intelligence* 97:104096.

Geetha, R., K. Ramyadevi, and M. Balasubramanian. 2021. "Prediction of domestic power peak demand and consumption using supervised machine learning with smart meter dataset". *Multimedia Tools and Applications* 80(13):19675–19693.

González-Torres, M., L. Pérez-Lombard, J. F. Coronel, I. R. Maestre and D. Yan. 2022. "A review on buildings energy information: Trends, end-uses, fuels and drivers". *Energy Reports* 8:626–637.

Happle, G., J. A. Fonseca, and A. Schlueter. 2020. "Context-specific urban occupancy modeling using location-based services data". *Building and Environment* 175:106803.

Hochreiter, S. and J. Schmidhuber. 1996. "LSTM can solve hard long time lag problems". *Advances in neural information processing systems* 9.

Kwon, B.-S., R.-J. Park, and K.-B. Song. 2020. "Short-term load forecasting based on deep neural networks using LSTM layer". *Journal of Electrical Engineering & Technology* 15(4):1501–1509.

Li, J., K. Panchabikesan, Z. Yu, F. Haghighat, M. El Mankibi and D. Corgier. 2019. "Systematic data mining-based framework to discover potential energy waste patterns in residential buildings". *Energy and Buildings* 199:562–578.

Li, J., Z. J. Yu, F. Haghighat, and G. Zhang. 2019. "Development and improvement of occupant behavior models towards realistic building performance simulation: A review". *Sustainable Cities and Society* 50:101685.

Liu, P., X. Qiu, X. Chen, S. Wu and X.-J. Huang. 2015. "Multi-timescale long short-term memory neural network for modelling sentences and documents". In *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2326–2335.

Lloyd, S. 1982. "Least squares quantization in PCM". *IEEE transactions on information theory* 28(2):129–137.

Meng, F., Q. Ma, Z. Liu, and X.-J. Zeng. 2023. "Multiple dynamic pricing for demand response with adaptive clustering-based customer segmentation in smart grids". *Applied Energy* 333:120626.

Moradzadeh, A., S. Zakeri, M. Shoaran, B. Mohammadi-Ivatloo and F. Mohammadi. 2020. "Short-term load forecasting of microgrid via hybrid support vector regression and long short-term memory algorithms". *Sustainability* 12(17):7076.

Panchabikesan, K., F. Haghighat, and M. El Mankibi. 2021. "Data driven occupancy information for energy simulation and energy use assessment in residential buildings". *Energy* 218:119539.

Paparrizos, J. and L. Gravano. 2015. "k-shape: Efficient and accurate clustering of time series". In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, 1855–1870.

Parrish, B., P. Heptonstall, R. Gross, and B. K. Sovacool. 2020. "A systematic review of motivations, enablers and barriers for consumer engagement with residential demand response". *Energy Policy* 138:111221.

Pearson, K. 1900. "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50(302):157–175.

Petitjean, F., A. Ketterlin, and P. Gançarski. 2011. "A global averaging method for dynamic time warping, with applications to clustering". *Pattern recognition* 44(3):678–693.

Rafiq, H., P. Manandhar, E. Rodriguez-Ubinas, J. D. Barbosa and O. A. Qureshi. 2023. "Analysis of residential electricity consumption patterns utilizing smart-meter data: Dubai as a case study". *Energy and Buildings* 291:113103.

Rdusseeun, L. and P. Kaufman. 1987. "Clustering by means of medoids". In *Proceedings of the statistical data analysis based on the L1 norm conference, neuchatel, switzerland*, Volume 31.

Sardá-Espinosa, A. 2019, 06. "Time-Series Clustering in R Using the dtwclust Package". *The R Journal* 11:22 https://doi.org/10.32614/RJ-2019-023.

Satre-Meloy, A., M. Diakonova, and P. Grünewald. 2020. "Cluster analysis and prediction of residential peak demand profiles using occupant activity data". *Applied Energy* 260:114246.

Tang, W., H. Wang, X.-L. Lee, and H.-T. Yang. 2022. "Machine learning approach to uncovering residential energy consumption patterns based on socioeconomic and smart meter data". *Energy* 240:122500.

Wang, Y., Q. Chen, C. Kang, and Q. Xia. 2016. "Clustering of electricity consumption behavior dynamics toward big data applications". *IEEE transactions on smart grid* 7(5):2437–2447.

AUTHOR BIOGRAPHIES

SHUANG DAI is a Postdoctoral research associate in the Department of Engineering at University of Exeter, UK. Her research interests cover Machine Learning and Big Data Analysis focused on Sustainability, Smart Cities, and Artificial Intelligence. Her email address is s.dai@exeter.ac.uk.

FANLIN MENG is a Senior Lecturer at University of Exeter Business School, UK. His primary research interests include Energy Data Analytics, Energy Market, Carbon Market, Smart Energy and Mobility, Machine Learning, Game Theory and Optimisation. He is Fellow of British Computer Society (FBCS), member of EPSRC Peer Review College, and Senior Member of IEEE. His email address is f.meng2@exeter.ac.uk.